Geochemical databases

Marthe Klöcking[a], **Kerstin A Lehnert**[b], **and Lesley Wyborn**[c], [a]Institute for Mineralogy, University of Münster, Münster, Germany; [b]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, United States; [c]ANU Research School of Earth Sciences, Acton, ACT, Australia

© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

This chapter has been reviewed by the section editor Anat Shahar.

https://doi.org/10.1016/B978-0-323-99762-1.00123-6

Abstract

Geochemistry is a data-driven discipline. Modern laboratories produce highly diverse data, and the recent exponential increase in data volumes is challenging established practices and capabilities for organizing, analyzing, preserving, and accessing these data. At the same time, sophisticated computational techniques, including machine learning, are increasingly applied to geochemical research questions, which require easy access to large volumes of high-quality, well-organized, and standardized data. Data management has been important since the beginning of geochemistry but has recently become a necessity for the discipline to thrive in the age of digitalization and artificial intelligence. This paper summarizes the landscape of geochemical databases, distinguishing different types of data systems based on their purpose, and their evolution in a historic context. We apply the life cycle model of geochemical data; explain the relevance of current standards, practices, and policies that determine the design of modern geochemical databases and data management; the ethics of data reuse such as data ownership, data attribution, and data citation; and finally create a vision for the future of geochemical databases: data being born digital, connected to agreed community standards, and contributing to global democratization of geochemical data.

Keywords

Artificial intelligence; CARE; Community standards; Data ethics; Data management; Databases; FAIR; Geochemistry; Machine learning; Machine readable data; Repository; TRUST

Key points

- Geochemical databases have played a critical role in geochemical research, facilitating access and analysis of geochemical data that has led to new discoveries and a more efficient research enterprise.
- Geochemical databases are only one aspect of the geochemical data ecosystem that also encompasses repositories and other types of data systems to support the geochemical data life cycle. The geochemical data ecosystem is dynamic and continues to evolve as new analytical techniques and advances in computational and data science methods continue to emerge.
- Geochemical data management today recognizes the importance of data principles, standards, and best practices that optimize their value for science and society and ensure ethical reuse. To maximize use and repurposing of geochemical data and open new horizons, the geochemistry community needs to come together to define the international standards required for machine readability of data, including minimum variables, formats, and controlled vocabularies.
- The adoption of international standards will enable the creation of global data networks and help overcome the currently fragmented landscape of geochemical data resources.

Introduction

At the heart of geochemistry is the acquisition and interpretation of compositional data—observations that describe the elemental and isotopic composition of terrestrial and planetary materials, including, for example, rocks, minerals, meteorites, sediments, soils, water (oceans, rivers, lakes), air, and cosmic dust. Geochemical data are key to understanding the natural processes that shape our planet, our solar system, and the universe, from the differentiation of protoplanets to magma genesis in the Earth's mantle to the evolution of life on land to changes in climate today and in the past. For nearly two centuries now, researchers have generated geochemical data in laboratories around the globe and used them in their research. Throughout this time, analytical methodologies, protocols, and instrumentation have changed and improved radically. Today, geochemists can measure compositions at a level of precision and resolution that was unimaginable just a decade ago. The automation of analytical workflows and computerization of analytical instruments have accelerated the speed of data creation and thus exponentially increased the volume of data. Researchers today face a rising tide of data that not only offers enormous opportunities for new scientific discoveries but also challenges established practices, capabilities, and capacity for organizing, analyzing, publishing, preserving, and accessing these data. Emerging data-driven and computational research paradigms demand easy online access to large volumes of standardized, high-quality, digital data. Today, data management is no longer just an option; it needs to be an intrinsic component of research and education in geochemistry. It is therefore essential for researchers to understand the landscape of geochemical data management, including the different types of data structures and systems and their purposes; the relevance of and requirements for data and metadata standards; the process and benefits of publishing data; the ethics of data reuse; and the distribution of responsibilities, including their own, in the global geochemical data ecosystem.

The scientific relevance and impact of geochemical databases were emphasized by the Geochemical Society in its policy on geochemical databases that recognized "*development of databases as a very positive direction for our profession*": "*. . . are there examples of studies where the compilation of a large amount of data has resulted in good science and has moved the field forward? There are many*"

(Geochemical Society, 2007). The policy highlights articles by Dziewonski and Anderson (1981), Zindler and Hart (1986), and Rudnick and Gao (2003) as examples for "paradigm-shifting" studies that used large data compilations and refers specifically to new web-accessible databases that emerged in the late 1990s. Since the policy of the Geochemical Society was written and adopted in 2007, the creation and use of databases in geochemistry have proliferated. This "digital" approach to geochemistry is increasingly driven by large-scale science questions, new paradigms in computational data analysis, and the growing implementation of Open Data policies by funding agencies and publishers as part of the Open Science movement.

This chapter reviews the landscape of geochemical databases: It provides the rationale for the creation of databases and their evolution in a historic context (Section "Historic context for modern geochemical databases"); applies the life cycle model of geochemical data and defines the different types of databases based on their purpose and technical implementation (Section "The ecosystem of geochemical data management"); introduces international principles for legal and ethical aspects of databases and explains the framework of current standards, practices, and policies that determine the design of modern geochemistry databases and their operational environment (Section "Principles, standards, and best practices for modern research data management"); describes currently available data systems, their features, and limitations (Section "Status of today's geochemistry data ecosystem"); and finally creates a vision for the future of geochemical databases and outlines requirements for turning that vision into reality (Section "Future opportunities for digital geochemical data").

Historic context for modern geochemical databases

The appearance and evolution of databases in geochemistry are intimately linked to the progression of analytical techniques in geochemical laboratories that have allowed researchers to acquire an ever-growing volume of data, and to the changes in the way researchers have (or have not) shared these data with their peers and the general public. The geochemical data ecosystem has radically changed since the first geochemical analyses were performed in the 19th century: from the exponential growth of data volumes (e.g., Hey and Trefethen, 2003); to the advent of electronic publishing (e.g., Smith, 2001); to the rise of the Open Science movement (e.g., Woelfle et al., 2011; Vicente-Saez and Martinez-Fuentes, 2018); to the rapidly expanding application of computational methodologies such as artificial intelligence and machine learning for extracting new knowledge from geochemical data (e.g., Pignatelli and Piochi, 2021; He et al., 2022).

Evolution of analytical methodology

The term "geochemistry" was first introduced by Schönbein in 1838 for "*the investigation of the chemical and physical properties of all geological formations and of their age relationships*" (Schönbein, 1838; Manten, 1966; Kragh, 2001). For the first 120 years, the majority of geochemical analyses were undertaken on powdered bulk rock samples or on mineral separates using "wet" or classical chemical techniques to determine the chemical constitution of unknowns against a series of known compounds and observe reactions. Geochemical laboratories were dominated by Bunsen burners and glassware, and only major elements and a few select trace elements could be measured routinely. The analytical processes were labor-intensive and time-consuming, generating small numbers of observations. An average PhD student may have barely completed 15 analyses in three years.

Around 1950, geochemical techniques became more physics-based as a result of the "Instrumental Revolution" (Morris, 2002; Reinhardt, 2006, 2019). Researchers started to use instruments that allowed the discrimination of chemical components through their physical properties in order to determine chemical compositions. A variety of new techniques, such as X-ray fluorescence, mass spectrometry, infrared and ultraviolet spectroscopy, and nuclear magnetic resonance spectroscopy, were introduced, and instruments rapidly dominated geochemistry laboratories (Baird, 1993; Borg, 2020). This change caused the first steep increase in the volumes of geochemical data (Fig. 1). With time, the variety of analytical techniques grew, and analytical instruments became more sensitive, accurate, and precise, while detection limits were progressively lowered. Microanalytical in situ techniques such as atom probe tomography were progressively developed and allowed sample sizes to diminish to atomic scale. From the 2000s, increasing computerization and automation of instruments subsequently led to an even bigger, exponential increase in the volumes of data generated. The diversity of elements and isotopes analyzed has expanded and now covers the periodic and isotopic table; the data deluge began. Fifteen bulk analyses are now routinely completed within a few days and an average PhD student may produce 1000s of in situ data points. Synchrotron techniques, as well as 2D and 3D imagery, are becoming essential components of geochemical analysis and increasingly challenge long-established practices of data storage, publication, and access.

Evolution of digital data infrastructure

In the past, geochemical data did not require a central data management facility such as is common in other research communities—for example, geodesy or astronomy—to store and disseminate large volumes of data acquired by sensor networks or satellites. Researchers usually stored and managed their data in paper laboratory notebooks and printouts, especially the comparatively small volumes of data generated by individual researchers in their laboratories. As personal computers (PCs) became available and affordable, and easy-to-use spreadsheet software could run on these PCs (e.g., Lotus 1-2-3, EXCEL), researchers were able to locally store and manage their geochemical data digitally and share it with colleagues on disks, CD-ROMs, and thumb drives that were exchanged by hand or mail.

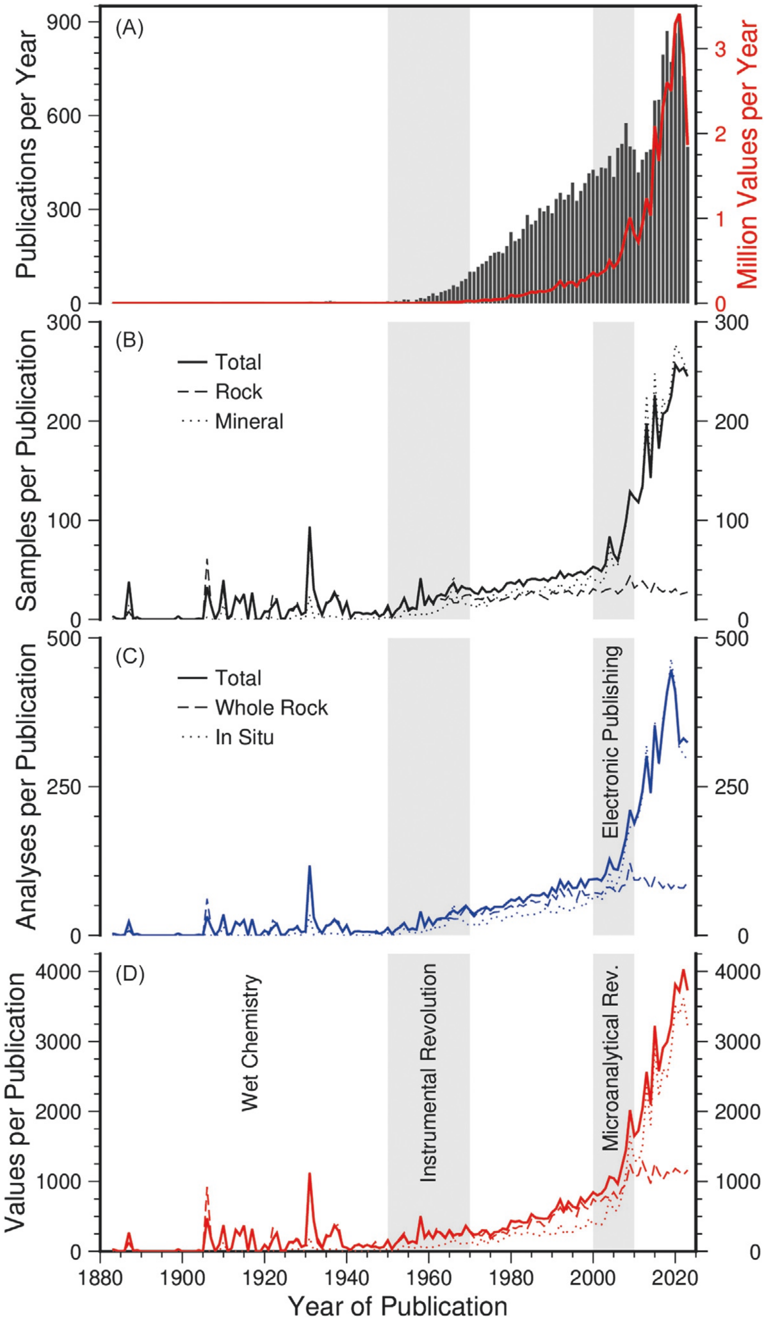


Fig. 1 Increase in volumes of geochemical data published in the peer-reviewed literature, based on data from the GEOROC database, version 2024-03-01, as a proxy for geochemical compositions of igneous rocks and minerals. (A) Rise in publications and data volumes. *Gray bars* = total number of publications per year containing geochemical data; *red line* = total number of single data values published per year. (B) Number of samples per publication per year. *Solid line* = rock and mineral samples normalized by total number of publications each year; *dashed line* = whole rock samples normalized by number of papers describing whole rock samples; *dotted line* = mineral samples normalized by number of papers describing mineral samples. (C) and (D) Same as (B) for number of analyses and single data values, respectively.

Until the 1990s, only large organizations such as geological surveys could afford to operate digital databases for geochemical data as they could only be developed on mainframe computers and required specialist expertise in relational database modeling and software (e.g., Wyborn and Ryburn, 1989). In the mid- to late 1990s, technological advancements such as the emergence of the Internet created new opportunities for managing and accessing geochemical data. Forward-thinking members in the global geochemical community took advantage of these technologies and created databases that synthesized large volumes of published geochemical data and made them easily accessible and searchable over the Internet. Major online databases emerged that compiled geochemical data from publications and provided online interactive user interfaces for searching, filtering, and retrieving data (e.g., Lehnert et al., 2000; Walker et al., 2006). These databases revolutionized access to geochemical data for the igneous rock geochemistry community and enabled and supported a vast number of studies and discoveries that would have otherwise not been possible. They in turn inspired the development of many other geochemical data systems (e.g., Niu et al., 2011).

Evolution of publications

Geochemical data have, until recently, been shared primarily as part of scholarly publications. Prior to the 1970s, a research paper in the literature would rarely contain more than 15–20 analyses within the paper as a typeset table. As the instrument revolution progressed, the increase in data volumes and in the number of samples studied in a research project made it very difficult to incorporate the complete dataset of a study as typeset tables in scholarly publications (e.g., Le Bas and Durham, 1989; Smith, 2001). It became the norm to report only "representative data" and/or report data in paper supplementary files that could only be retrieved through direct contact with the author. In either case, many of these data are no longer accessible today (Vines et al., 2014; Tedersoo et al., 2021). In the early 2000s, as electronic publishing expanded, digital electronic supplements became an integral part of articles, allowing larger volumes of data to be included (Hinze, 2001). Fig. 1 shows very clearly that the inability to publish complete datasets acted as a major bottleneck, and the volume of data published increased dramatically after the widespread adoption of electronic supplements. These electronic supplements were stored and made available on FTP servers operated by publishers or could be obtained from the author, most commonly as spreadsheet files (XLS, CSV). At the same time, the "microanalytical revolution" triggered a multiplication of the number of in situ analyses on mineral samples, in particular, that is still ongoing today. Unfortunately, electronic supplements of many older publications are also no longer available as FTP servers were abandoned and authors retired or passed away. Another limitation on reuse was that the hardware, software, and media that were used to store and read the data became obsolete (e.g., cartridges, 8 inch and 5¼ inch floppy disks) and many digital datasets from 1970 to 2010 have been lost forever as they can no longer be accessed.

Today, an increasing number of publishers endorse and support Open Science policies and require authors to deposit the data of their article in publicly funded, openly accessible data repositories, preferably those with disciplinary focus, to ensure their quality for reuse and their long-term preservation for persistent access (Stall et al., 2019, e.g., AGU Data Sharing Agreement, n.d., Geochimica et Cosmochimica Acta Guide for Authors, n.d., Journal of Petrology Information for Authors, n.d.). This change was triggered by the rise of the Open Science paradigm and further supported by international governmental policies and mandates from funding agencies. Disciplinary-focused repositories are more likely to migrate data to new hardware and software and remaster and recurate the data as standards and vocabularies evolve.

Driven by the need to support and incentivize data sharing, publishers have established a growing number of peer-reviewed journals where researchers can describe and publish specific datasets (e.g., Candela et al., 2015; Walters, 2020). Articles in these data journals provide detailed information about the provenance and purpose of a given dataset and the methodologies used to generate it. Examples of data journals that publish geochemical data include Earth and Space Sciences (Wiley/AGU), Earth System Science Data (Copernicus, EGU), Scientific Data (Springer/Nature), Data in Brief (Elsevier), and the Geoscience Data Journal (Royal Meteorological Society, Wiley).

Rise of Open Science

The broad availability of the Internet in the late 20th century for the first time ever offered the chance to constitute a global and interactive representation of human knowledge and cultural heritage with the guarantee of worldwide access. This ability led to the rise of the Open Science movement (UNESCO, 2021), which aims to ensure transparent and accessible knowledge, for the scientific community and the general public alike, to the whole scientific workflow rather than just the final research product (Gentemann, 2023). Open Science encompasses all aspects of access to and communication of scientific knowledge, from publishing open research to campaigning for open access to openly sharing data and code to broader dissemination of and engagement in science. Open Science has been embraced by governments around the world and has triggered national and international initiatives and policies to establish and promote principles, practices, and infrastructure for openly sharing knowledge, data, and other research outputs, including software and samples (e.g., UNESCO, 2021, OECD Recommendation of the Council Concerning Access to Research Data From Public Funding, n.d.), the EU Open Science Policy, n.d., the Global Open Science Cloud, n.d., the Nelson Memo, n.d. of the US government, and the Year of Open Science, n.d. in the United States). Geochemical data systems have gained increased attention and usage due to the Open Science movement as funding agencies and publishers have implemented policies that require researchers to openly share their data generated with public funds, following discipline- and technology-agnostic principles such as FAIR (Findability, Accessibility, Interoperability, and Reusability), TRUST (Transparency, Responsibility, User Focus, Sustainability, and Technology) and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics; Wilkinson et al., 2016; Lin et al., 2020; Carroll et al., 2020, see Section "Principles, standards, and best practices for modern research data management"). A swath of new geochemistry data repositories and databases has emerged to support data sharing on national, thematic, and programmatic levels. This in turn has created an urgent demand for best practices and standards for geochemical data and for a change in the data culture in geochemistry (Chamberlain et al., 2021; Klöcking et al., 2023), see Sections "Principles, standards, and best practices for modern research data management" and "Future opportunities for digital geochemical data").

Rise of computational geochemistry

The fundamental value of applying statistical methods to geochemistry was already raised by Shaw and Bankier (1954), but it would take another half a century before the growth of data volumes and advancements in information technologies allowed large-scale statistical data science in geochemistry. Before the rise of the Internet, databases were either personal or institutional, such as the comprehensive collections at geological surveys and government agencies in Australia, Japan or the United States, and very difficult to share and aggregate. The establishment of international synthesis databases as well as the rapidly growing data volumes made it possible for researchers worldwide to access, explore, and analyze geochemical "big data" over the Internet and start using computational methods to generate serendipitous opportunities for new scientific discoveries through reuse of geochemical data. PetDB and GEOROC are prominent examples of two databases in igneous geochemistry that have provided access to "big" geochemical data since the turn of the century, enabling new research paradigms such as "statistical geochemistry" (Keller and Schoene, 2012) that were not possible when individual researchers needed to spend months or years wrangling data from hundreds of publications into personal databases in order to develop or test their hypotheses. Although data volumes in geochemistry do not currently compare to "big data" disciplines like climate science or seismology, the complexity of geochemical data brings its own challenges. A single rock sample can be subdivided into a whole-rock portion as well as sections for in situ microanalysis and tens of individual mineral separates. These mineral grains might in turn contain melt inclusions. Each of these subsamples could then be analyzed for the entire suite of geochemical and isotopic analytes using a variety of different analytical techniques. Geochemical databases play a fundamental role in preserving these complex hierarchies while providing access to large-scale, harmonized, and quality-controlled geochemical data. Today, these databases are increasingly accessed directly by computational tools via Advanced Programming Interfaces (APIs) and mined with "data science" and machine learning techniques to explore spatial, temporal, and compositional patterns (e.g., Keller and Schoene, 2018; Ueki et al., 2018; Hasterok et al., 2019; Hazen et al., 2019; Liu et al., 2019; Keller and Harrison, 2020). Geochemical databases have gained recognition as "*drivers in the trend toward computation as researchers incorporate large datasets into their studies*" (Hwang et al., 2017).

The ecosystem of geochemical data management

Research data management requires research data infrastructure, which is defined as a "*body of complex social, technical and socio-technical attributes and relationships*" (Parsons, 2013). This definition encompasses physical infrastructure, facilities and services, digital assets, including data and software, standards, policies, expertise, governance, and funding. In the following section, we introduce the concept of the geochemical data life cycle and the different components of the geochemical data infrastructure that support the phases in this data life cycle. We explain different types of data systems as well as the data structures, metadata, and the conceptual model that are applied to store, manage, process, curate, disseminate, and access geochemical data. A data system is defined here as an architecture for storing and managing data in various forms such as numbers, text, images, and videos. See Table 1 for a glossary of database-related terms used in this paper.

Three types of data are broadly relevant to modern geochemistry: numerical data, image data, and simulation data. Complementing numerical geochemical data, the recent advances in analytical instrumentation have enabled quantitative in situ measurements of elemental concentrations and isotopic compositions of samples across large areas or at increasingly higher spatial resolution. 2D and 3D imagery are becoming pervasive in geochemistry in the form of quantitative maps that connect geochemical and textural data of the mineral phases in a sample. At the same time, geochemical data are increasingly augmented with machine learning techniques or in preparation for numerical modeling applications, such as the prediction of high-resolution data from lower-resolution images or the imputation of missing values in legacy datasets. With the rise of computational capabilities and their growing number of applications within geochemistry, the boundaries between analytical and computed data are being diminished and both data types are increasingly connected within a single research workflow. Data management approaches are largely identical for all three data types and for simplicity, this chapter primarily focuses on numerical analytical data produced in geochemical laboratories. However, many concepts discussed below can easily be translated into imagery, geochemical software, and computational models. The unique challenges posed by high-volume image data and ethical considerations around "data cleaning" are discussed in Section "Limitations of today's geochemistry data ecosystem."

The life cycle of geochemical data

All data go through a sequence of stages known as the "Data Life Cycle" from their initial generation or capture to their eventual archival and/or deletion at the end of their useful life (Wing, 2019; Badia, 2020). Geochemical research starts with a decision on the hypotheses to be tested by a project, including the decision on which samples and data to collect or generate. The various processes of sample collection, data generation, quality assurance, data analysis, description, publication, and reuse of geochemical data, often occurring over the course of multiple research projects conducted over many decades, are summarized in the geochemical data life cycle (Fig. 2). This framework highlights the interconnectedness of the individual processes and the need to carefully document each stage in the data life cycle in order to generate geochemical data that can be reused with confidence and that comply with quality expectations such as defined by the FAIR principles (Wilkinson et al., 2016). However, the complexity of the modern full data life cycle model makes it clear that the responsibility for these different aspects of ensuring data quality and reusability cannot

Table 1 Glossary of terms describing data types, data structures and data systems. Most definitions are taken from the CODATA Research Data Management Terminology (CODATA RDM Terminology Working Group, 2024).

Term	*Definition*	*Examples*	*Source*
Data	Facts, measurements, recordings, records, or observations about the world, collected by researchers, that are yet to be processed/interpreted/analyzed. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records.		CODATA RDM Terminology
Metadata	Data about data. It is data (or information) that defines and describes the characteristics of other data. It is used to improve the understanding and use of the data.		CODATA RDM Terminology
Dataset	Organized collection of data or objects in a computational format, that are generated or collected by researchers in the course of their investigations, regardless of their form or method, that form the object on which researchers test a hypothesis. This includes the full range of data: raw, unprocessed datasets, proprietary generated and processed data and secondary data obtained from third parties. The presentation of the data in the application is enabled through metadata.		CODATA RDM Terminology
Data Compilation	A single-file dataset in which previously unconnected data are harmonized and integrated.	\citet{Class2012, Rasmussen2022}	this paper
Structured Data	Data whose elements have been organized into a consistent format and data structure within a defined data model such that the elements can be easily addressed, organized and accessed in various combinations to make better use of the information, such as in a relational database.		CODATA RDM Terminology
Data File Format	Layout of a file in terms of how the data within the file are organized and encoded for storage.		CODATA RDM Terminology
Data Model	Model that specifies the structure or schema of a dataset. The model provides a documented description of the data and thus is an instance of metadata. It is a logical, relational data model showing an organized dataset as a collection of tables with entity, attributes and relations.		CODATA RDM Terminology
Linked Open Data	Data where relationships/connections between them are available to allow easy data access. A typical case of a large Linked dataset is DBPedia (http://dbpedia.org/), which essentially makes the content of Wikipedia available in RDF. This related collection of interrelated datasets is stored on the Web and available via a common format -RDF.	DBPedia	CODATA RDM Terminology, http://www.w3.org/standards/semanticweb/data#summary
Data Infrastructure	Geochemical data infrastructure is part of the broader global research data infrastructure that is often viewed as a body of complex social, technical and socio-technical attributes and relationships; including all types of data systems.		Parsons 2013
Data System	Architecture/means of storing and managing data in various forms such as numbers, text, images, and videos. Could be composed of one or multiple components, including all layers within a software architecture.	database, data repository, LIMS, etc.	This paper
Repository	Physical or digital storage location that can house, preserve, manage, and provide access to many types of digital and physical materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the physical and digital material to be authentic, reliable, accessible and usable on a continuing basis.		CODATA RDM Terminology
Trusted Digital Repository	Infrastructure component that provides reliable, long-term access to managed digital resources. It stores, manages, and curates digital objects and returns their bit streams when a request is issued. Trusted repositories undergo regular assessments according to a set of rules such as defined by TRAC (ISO 16363:2012) or CoreTrustSeal. Such an assessment has the potential to increase trust from its depositors and users. Certain quality criteria need to be met to distinguish trusted repositories from other entities that store data, such as notebooks or lab servers.	EarthChem Library, PANGAEA	CODATA RDM Terminology

(*Continued*)

Table 1 (Continued)

Term	*Definition*	*Examples*	*Source*
Synthesis Database	A collection of data that is organized according to a conceptual structure/model describing the characteristics of these data and the relationships among their corresponding entities, supporting one or more application areas. A database may house one or many datasets.	PetDB, GEOROC	CODATA RDM Terminology
Data Portal	A web-based interface designed to make it easier to find reusable information. In combination with specific search functionalities, they facilitate finding datasets of interest. Application programming interfaces (APIs) are often available as well, offering direct and automated access to data for software applications. â€˜Openâ€™ data portals provide public access to any data, without the need to reply to individual requests for access to data.	DataONE, EarthChem Portal	European Commission, Shaping Europe's Digital Future (https://digital-strategy.ec.europa.eu/en/policies/open-data-portals)
Curation	Managing and promoting the use of assets from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials.		CODATA RDM Terminology
(Controlled) Vocabulary	An organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. A controlled vocabulary is a list of standardized terminology, words, or phrases, used for indexing or content analysis and information retrieval, usually in a defined information domain.	AGU Index of Terms	CODATA RDM Terminology
Ontology	Shared and standardized list of words, terms and phrases to describe components of a particular discipline or domain, along with a taxonomy of their relations. Compare this to a controlled vocabularies, which tend not to include a structure of relations between their terms. Ontologies are typically developed by domain-specific institutions or communities to aid in the precise referencing of elements.	The Environment Ontology (ENVO)	CODATA RDM Terminology
Glossary	Alphabetical list of terms with definitions	Glossary of Geology	Neunendorf et al. 2011
Taxonomy	Divisions of (preferred) terms into ordered, hierarchical groups or categories based on particular characteristics. A form of classification.	Goldschmidtâ€™s Classification of the Elements (Goldschmidt 1937)	Zeng 2008
Thesaurus	Sets of terms representing concepts and the hierarchical, equivalence, and associative relationships connecting them	The USGS Thesaurus	Zeng 2008
Protocol	A formal or official record of scientific experimental observations.		CODATA RDM Terminology
Data Product Specification	Detailed description of a dataset or dataset series together with additional information that will enable it to be created, supplied to and used by another party. A data product specification provides a description of the universe of discourse and a specification for mapping the universe of discourse to a dataset. It may be used for production, sales, end-use or other purposes.		CODATA RDM Terminology
Standard	Set of agreed-upon and documented guidelines, specifications, accepted practices, technical requirements, or terminologies that have been prepared by a standards developing organization or group, and published in accordance with established procedures. These can be mandatory or voluntary.	Periodic Table of the Elements	CODATA RDM Terminology
Data Standard	Technical specification that defines how data should be structured and formatted to ease interoperability across different systems, publishers and users.	GeoSciML, ISO19156:2023	CODATA RDM Terminology

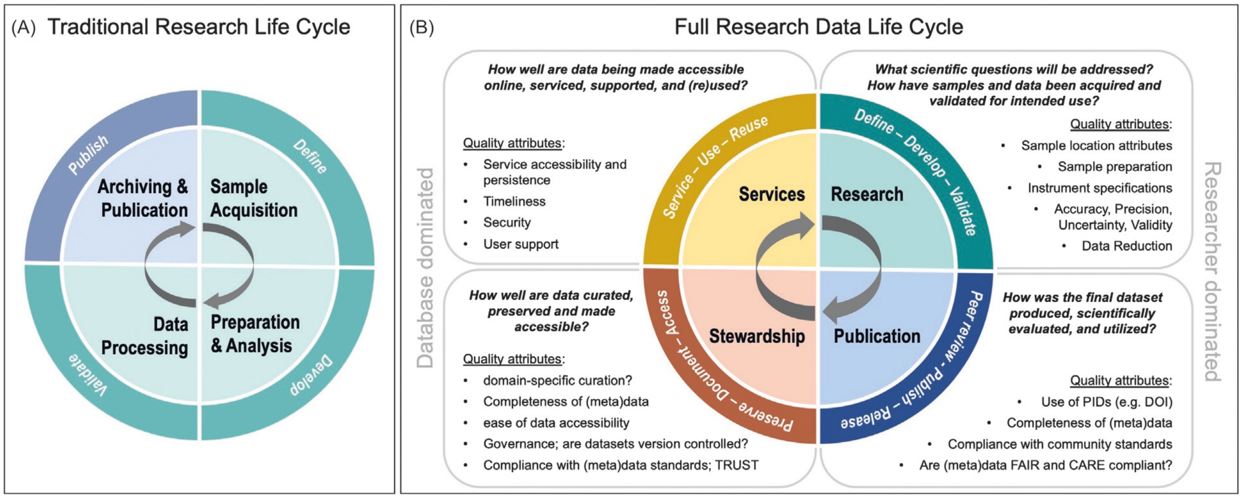

(A) Traditional Research Life Cycle
Publish
Define
Develop
Validate
Archiving & Publication
Sample Acquisition
Data Processing
Preparation & Analysis
(B) Full Research Data Life Cycle
How well are data being made accessible online, serviced, supported, and (re)used?
Quality attributes:
• Service accessibility and persistence
• Timeliness
• Security
• User support
What scientific questions will be addressed? How have samples and data been acquired and validated for intended use?
Quality attributes:
• Sample location attributes
• Sample preparation
• Instrument specifications
• Accuracy, Precision, Uncertainty, Validity
• Data Reduction
Service – Use – Reuse
Define – Develop – Validate
Services
Research
Stewardship
Publication
Preserve – Document – Access
Peer review - Publish - Release
Database dominated
Researcher dominated
How well are data curated, preserved and made accessible?
Quality attributes:
• domain-specific curation?
• Completeness of (meta)data
• ease of data accessibility
• Governance; are datasets version controlled?
• Compliance with (meta)data standards; TRUST
How was the final dataset produced, scientifically evaluated, and utilized?
Quality attributes:
• Use of PIDs (e.g. DOI)
• Completeness of (meta)data
• Compliance with community standards
• Are (meta)data FAIR and CARE compliant?

Fig. 2 Geochemical data life cycle and quality attributes. (A) Traditional research cycle/workflow from sample acquisition or synthesis, to laboratory analysis, data processing and computational analysis, and finally archiving of data and samples and publication of research results. (B) Full modern research data life cycle that also includes long-term stewardship and service by dedicated data systems to guarantee preservation and reuse. For definitions of FAIR, CARE, TRUST, and a description of standards, see Section "Principles, standards, and best practices for modern research data management." Modified after Peng G, Lacagnina C, Downs RR, Ganske A, Ramapriyan HK, Ivánová I, Wyborn L, Jones D, Bastin L, Shie C-L, Moroni DF (2022) Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. Data Science Journal 21. https://doi.org/10.5334/dsj-2022-008.

lie with one person alone. The geochemical researcher is primarily concerned with the collection of high-quality data ("research"), their interpretation and their initial publication as part of a scientific article. Various types of data systems have been developed to cover the remainder of the data life cycle, including curation and archiving of primary data and metadata and their acquisition into a data repository (another aspect of "publication"); maintenance of the dataset and other data products over time ("stewardship"); recasting of the data into more usable compilation databases, followed by the development of tools that make the data accessible online ("services"). Although these additional steps and layers may at times seem cumbersome and superfluous, digital data preservation means that 19th century manuscripts and data from the beginnings of geochemistry can still be read and accessed today.

Management of geochemical data follows the sequence of phases in their life cycle from sample collection in the field to data acquisition in the lab to sharing data as part of scholarly publishing to long-term archiving of and online access to the data. Each phase in the geochemical data life cycle has its specific workflows and metadata that need to be recorded to fully document data provenance so that others can replicate or reproduce the results. Historically, data publication, stewardship, and any related services were taken care of by journals, research institutions, and physical libraries. Geochemists could focus on the planning, execution, and documentation of their research (Fig. 2A). Since the digital revolution, however, new infrastructure was required to support traditional libraries and to maintain more control over the curation and preservation of research data. Increasingly, these additional aspects of the data life cycle are also becoming the researchers' responsibility. Discipline-focused data systems now work closely with researchers to ensure appropriate curation of data that preserves all original features but also provides easy access for meaningful reuse by the community. The following briefly summarizes the different life cycle stages from a researcher's perspective, highlighting where there are services and workflows in place to support the process. In the subsequent sections, critical concepts and components of the geochemical data infrastructure are discussed that support the management of geochemical data through the various stages of its life and ensure its immediate and long-term value for science and society.

Data generation and publication

Sample collection or synthesis, the generation of data, and their interpretation remain the primary objectives of geochemical research. Many other chapters in this Treatise volume are dedicated to research methodologies, and there is no need to discuss these in detail here. However, regardless of the specific methodology followed in any one project, establishing and following a research data management plan can provide guidance and structure for recording all relevant aspects of the research workflow and save time at the publication stage. Where they exist, best practice guidelines and templates should be followed to compile and store ancillary information alongside the primary analytical data, such as geographic location of sample collection, instrument setup, and analytical parameters.

Publication is an opportunity for external peer review of both a research study and the data collected and applied within it. As outlined in Section "Evolution of publications," at present there are several ways to publish geochemical research data. It is still common to publish data alongside research findings as tables or electronic appendices to journal articles. Since these data tables can often be difficult to access (e.g., they reside behind the paywall of the respective journal), many journals are now recommending or even requiring data publication in an open repository. This data publication can then be cited as part of the research article.

However, its publication independently from the journal publisher means that the dataset is also accessible separately and can usually be updated through versioning. In fact, data publication in a repository does not need to be linked to any research manuscript. Generally, this outsourcing of the data publication provides an additional aspect of peer review since domain repositories provide curation of submitted data as well as offering guidance and specific data templates. Journal reviewers are usually asked to review any associated data publication as well as the scientific manuscript itself. Finally, there are now more than five specific data journals suitable for geochemical data that publish short papers describing and documenting a particular dataset. Often, these data journals also require authors to separately publish their data in a repository, which will then be cited in the data paper. The Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) provides an overview of data policies for different journals as well as general information and guidelines on data publication.

Data curation, preservation, accessibility, and reuse

Data repositories and other tools in the geochemistry data ecosystem support researchers in gathering, organizing, publishing, and archiving geochemical data and the relevant metadata. Through review and curation of both data and metadata, the value of a dataset can be substantially increased. An important aspect of this curation is the standardization of data formats and the language used to describe a dataset, which ensures that datasets are compatible with any previous or future data publications and can easily be compared or synthesized. Discipline-specific curation of data and metadata can ensure use and reuse within that domain for a long time. At the same time, data storage comes at a cost and decisions have to be made about which data are worthy of preservation and which are redundant. For example, derived trace element ratios are often reported alongside absolute concentrations of the same elements in research articles. However, the long-term preservation of these ratios is not required since they can easily be recalculated. In contrast, isotopic ratios are usually reported as normalized values relative to a standard reference material. The preservation of this reference material value is therefore very important for any future recalculation of the data. During data publication, this decision on what data to preserve largely lies with the authors, guided by community best practices and recommended data templates, and is driven by the immediate needs of the research problem in question. For long-term preservation, in contrast, driving factors are reusability and sustainability so that derived parts of a dataset may be excluded to avoid duplication and to optimize cost and performance of the data system.

The significant cost of data and metadata curation, as well as the maintenance of software, hardware, and data, used to be carried by journals/publishers but is now largely outsourced to separate scientific infrastructure that is often funded in the same way as research projects. However, the benefits gained from data curation and review far outweigh this additional cost as researchers have greater control over what is preserved and how, and data are more readily accessible not just to researchers in the same field but also to the government and private sectors as well as the general public. Nevertheless, access to data and associated metadata is often provided in a specific format, such as a formatted spreadsheet with custom column headings and annotations. While these formats are usually easy to understand for trained geochemists, they are often not machine-accessible or readable, which limits their potential for future reuse. Access can further be hindered by complex user account requirements. There are many examples of services being restricted to certain user groups, for example, within a certain institution or community. Finally, with the Open Science movement there are increasing demands not only for the data, but also for the underlying data structure, to be free and open for easier reuse and interoperability with other systems. Many research data systems today are free to access, especially since they are often built out of community or governmental initiatives as a service to the community. Nonetheless, several of the most common and well-maintained database management systems are based on proprietary software.

The logical structure of geochemical data

The concept of the life cycle can be used to design the technical infrastructure and principles for storing geochemical data in a database. At the heart of the technical infrastructure is a logical structure or "data model" that encapsulates the data themselves as well as a wide range of ancillary information that is required to document the procedures used to acquire and process these data. The data model standardizes how data and metadata relate to one another. It enables machine-readability and data access across the variety of sample or data types and analytical procedures.

Best suited for geochemical data is the "Observations, Measurements, and Samples" (OMS) technical standard approved by the Open Geospatial Consortium, an international alliance of businesses, government agencies, research organizations, and universities that sets standards for geospatial data. OMS (previously known as Observations and Measurements; ISO 19156:2023; Cox, 2010; Schleidt and Rinne, 2023) formally describes any data workflow as observations made using a specific measurement on a "feature of interest." The concept of a "feature of interest" encompasses both the target site and a sample taken at that site. For example, when describing a 50 mL water sample of the Pacific Ocean, both the water sample and the Pacific Ocean are features of interest during a specific sampling event. The water sample is the result of a documented sampling procedure carried out in a specific location within the ocean at a specific time (Fig. 3; Gordon et al., 2015)).

OMS is discipline and data-type agnostic, that is, the standard can be applied to any type of sample (e.g., solid, liquid, gaseous, terrestrial, planetary, synthetic), site, or observing station (e.g., deployment of sensor), using any type of measurement (e.g., laboratory, in situ, time series, remote sensing, (hyper)spectral, satellite imagery) to obtain any type of observation (e.g., chemical composition, physical properties, crystallographic structure). Due to its generic nature, OMS supports interoperability between different disciplinary implementations, such as between databases hosting observational data from different types of samples in disciplines as diverse as hydrology, rock geochemistry, soil geochemistry, and biogeochemistry (Hsu et al., 2017).

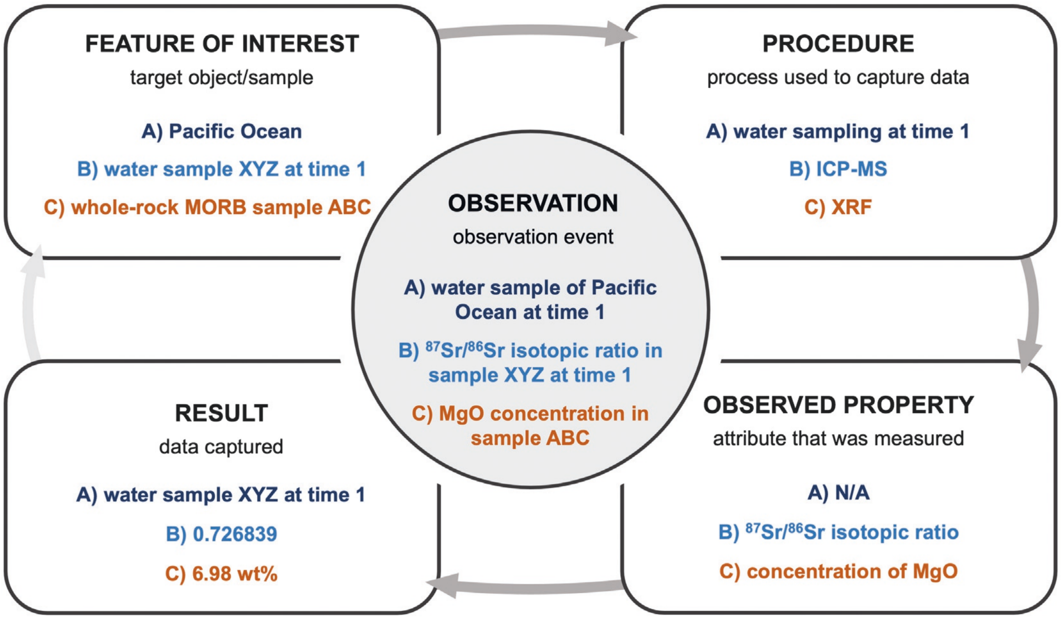

FEATURE OF INTEREST
target object/sample
A) Pacific Ocean
B) water sample XYZ at time 1
C) whole-rock MORB sample ABC
PROCEDURE
process used to capture data
A) water sampling at time 1
B) ICP-MS
C) XRF
OBSERVATION
observation event
A) water sample of Pacific Ocean at time 1
B) $^{87}Sr/^{86}Sr$ isotopic ratio in sample XYZ at time 1
C) MgO concentration in sample ABC
RESULT
data captured
A) water sample XYZ at time 1
B) 0.726839
C) 6.98 wt%
OBSERVED PROPERTY
attribute that was measured
A) N/A
B) $^{87}Sr/^{86}Sr$ isotopic ratio
C) concentration of MgO

Fig. 3 An observation as defined by the Observation, Measurements, and Samples model (Schleidt and Rinne, 2023). Feature of interest, procedure, observed property, and result are all components of an observation. *Arrows* signify symbolic workflow from a researcher's perspective. Three examples are provided to illustrate how this data model might be applied to a geochemical workflow. Redrawn after Gordon JM, Chkhenkeli N, Govoni DL, Lightsom FL, Ostroff AC, Schweitzer PN, Thongsavanh P, Varanka DE, Zednik S (2015) A Case Study of Data Integration for Aquatic Resources Using Semantic Web Technologies. U.S. Geological Survey. https://doi.org/10.3133/ofr20151004.

For example, the Observation Data Model (ODM2; Horsburgh et al., 2016) uses the OMS precursor "Observations and Measurements" to integrate sample data and time series data for applications in environmental research. Fig. 3 shows examples of how OMS can be applied to geochemical data. Thus, the OMS standard translates the data life cycle from sample and data acquisition to data reuse into technical language for data management. OMS continues to be updated to align with evolving technology and has been abstracted as the Semantic Sensor Network Ontology (Haller et al., 2018), for example, in order to further machine-readability.

Metadata for reusable, trustworthy geochemical data

In order to make geochemical data shareable and reusable, it is imperative to record and publish information about the data, that is, its metadata. Without information about the origin of the data, the analyzed sample, the analytical procedure, instrument calibration, and the way raw data (e.g., detector counts) have been processed into derived and eventually normalized compositions (e.g., weight percent or isotope ratios), it is impossible for others to evaluate and reuse these data, and to properly compare them to or integrate them with other data. For example, Gale et al. (2013) used such metadata to calculate interlaboratory correction factors and eliminate low-quality data for a global compilation of mid-ocean ridge basalts. In addition, without providing information about the content of a dataset, its authors, and where to access the data, online searches may not be able to find the data or allow the citation and attribution of those who generated and curated them.

As described in the 2007 Geochemical Society policy statement, metadata should "*1. allow reproduction of the experiment; 2. describe quality of the analytical data; and 3. allow comparison with other laboratories (standardization)*" (Geochemical Society, 2007). The 2009 Editors' Roundtable in geochemistry identified a similar set of critical metadata and recommended that these should be published alongside the primary analytical data in an accessible and ideally standardized format (Goldstein et al., 2014). At a minimum, these metadata should describe:

1. The sample itself, where and how it was collected/synthesized, including geographic location and its local context, geological age if applicable, and permissions;
2. The material that was analyzed and the method used to prepare the sample for each specific type of analysis undertaken;
3. The instrument and the analytical method, including instrument configuration and calibration, uncertainty quantification, and quality documentation; and
4. Data reduction techniques applied to the raw analytical data.

The scientific value of data increases with the quality and richness of the associated metadata. Good metadata facilitate more reuse and sustained impact of the data. One of the reasons why the synthesis databases of GEOROC and PetDB are widely used is that they allow fully flexible searches through decades of geochemical data based on common metadata, including geographic location, geological setting, age, and lithology. As a consequence, data included in these synthesis databases are more likely to be discovered and reused than if they were only available through their respective research article. The FAIR principles, explained in detail in Section "Principles, standards, and best practices for modern research data management," emphasize the relevance of "rich metadata," whereby metadata that are recommended by a disciplinary community are especially relevant for the reusability of data, for both humans and machines. Geochemistry is a markedly diverse discipline with respect to methodologies and analyzed materials, and therefore a wide range of metadata recommendations exist (e.g., Deines et al., 2003; Walker et al., 2008; Horstwood et al., 2016; Dutton et al., 2017; Courtney Mustaphi et al., 2019; Khider et al., 2019; Demetriades et al., 2020; Schaen et al., 2020; Brantley et al., 2021; Damerow et al., 2021; Abbott et al., 2022; Boone et al., 2022; Demetriades et al., 2022; Flowers et al., 2022, 2023; Peng et al., 2022; Wallace et al., 2022; Mahan et al., 2022; Klein and Eddy, 2023; Kohn et al., 2024). In general, metadata for any geochemical dataset should include:

- *Sample metadata* that provide information about geographic location of the sampling site, geological context (e.g., setting, stratigraphy, geological units) and age, sampling cruise/field program, sampling technique, sample classification, sample repository (for natural samples); experimental setup and conditions, composition of the starting mixture (for synthetic samples);
- *Analyzed material* (e.g., whole rock, mineral, inclusion), its classification (e.g., rock class, mineral species), and relationship to the sample;
- *Analytical or method metadata* that describe the analytical method: technique, instrument, laboratory, sample preparation method, errors, precision, standard values, data reduction, and correction procedures;
- *Bibliographic metadata* that record bibliographical information for the associated research publication; and
- *Administrative metadata* that record administrative aspects such as intellectual property rights, licensing, funder, grant, project(s), and acquisition permissions (including Indigenous rights).

Persistent identifiers for findable and citable data and samples

Foundational to modern digital data management and to enabling compliance with many of the standards for research data management such as the FAIR Principles (see Section "Principles, standards, and best practices for modern research data management") is the need to assign a persistent identifier (PID) to data, metadata, and artifacts. A PID is defined by the National Science and Technology Council (NSTC) as "*a globally unique, machine resolvable and processable identifier that has an associated metadata schema*" (National Science and Technology Council (NSTC), 2022). Applying "globally unique" PIDs means that the identifier unambiguously refers to exactly one single resource (physical or digital object) in the world (see McMurry et al., 2017; Jacobsen et al., 2020; Juty et al., 2020, for more advantages in using PIDs). PIDs provide a standard mechanism for registering metadata in a central, harvestable catalog and for retrieval of those metadata. The use of PIDs in modern data infrastructure is so critical for discovery, access, citability, and interoperability of digital resources that not only digital datasets are registered with PIDs, but also related artifacts such as analytical instruments and material samples as well as persons and organizations. For example, DOIs uniquely identify research articles or datasets as data sources. Samples collected for geochemical analysis can be registered with the International Generic Sample Number (IGSN). ORCIDs for authors and researchers are now required by many funding agencies and publishers, and laboratories and organizations obtain PIDs from the "Research Organization Registry" (ROR). PIDs can be linked to each other and thus help to build networks of information that vastly improve our ability to track data and sample citations and thus usage and impact.

Data structures

Data may be stored in a variety of structures that are used to integrate different data elements in a logical way to facilitate their effective use, persistence, and sharing. The choice of which data structure best suits the representation of data is usually determined by considerations of data volumes and complexity as well as the overall purpose and target audience. The different structures relevant to geochemical databases are described below and illustrated in Fig. 4.

Spreadsheets

The simplest form of organizing data in a structured manner is a table in a single data file (Fig. 4A). In geochemistry, these files are often Microsoft EXCEL spreadsheets, although other, nonproprietary tabular formats, such as CSV, are also commonly used. There are clear limits to the volume of data and complexity of metadata that can be reasonably or efficiently stored and retrieved via a single-file spreadsheet. Spreadsheet-based data compilations are a convenient solution for specific, well-defined research questions. Yet in many cases, as explained further in Section "Synthesis (compilation) databases," only minimal metadata are easily accommodated in these spreadsheets, which makes them difficult to reuse or translate/incorporate into broader or different research projects at a later stage. Yet, due to their simplicity, spreadsheets are very commonly used by individual researchers and/or laboratory scientists in geochemistry.

Relational databases

Instead of using the two-dimensional, flat file format of spreadsheets, relational database management systems (RDBMS) offer a powerful method of organizing and querying multidimensional, related data (e.g., Codd, 1970; Lehnert et al., 2000). RDBMS are

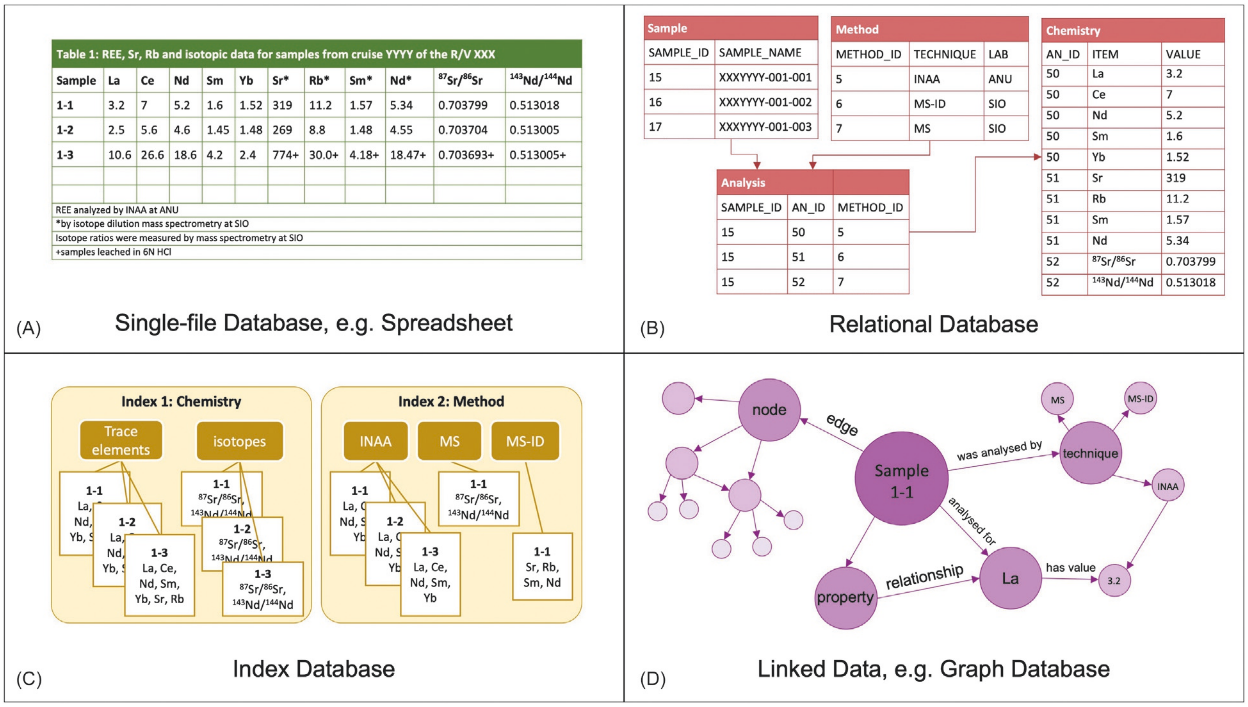


Table 1: REE, Sr, Rb and isotopic data for samples from cruise YYYY of the R/V XXX

Sample	La	Ce	Nd	Sm	Yb	Sr*	Rb*	Sm*	Nd*	$^{87}Sr/^{86}Sr$	$^{143}Nd/^{144}Nd$
1-1	3.2	7	5.2	1.6	1.52	319	11.2	1.57	5.34	0.703799	0.513018
1-2	2.5	5.6	4.6	1.45	1.48	269	8.8	1.48	4.55	0.703704	0.513005
1-3	10.6	26.6	18.6	4.2	2.4	774+	30.0+	4.18+	18.47+	0.703693+	0.513005+

REE analyzed by INAA at ANU
*by isotope dilution mass spectrometry at SIO
Isotope ratios were measured by mass spectrometry at SIO
+samples leached in 6N HCl

(A) Single-file Database, e.g. Spreadsheet

Sample

SAMPLE_ID	SAMPLE_NAME
15	XXXYYYY-001-001
16	XXXYYYY-001-002
17	XXXYYYY-001-003

Method

METHOD_ID	TECHNIQUE	LAB
5	INAA	ANU
6	MS-ID	SIO
7	MS	SIO

Analysis

SAMPLE_ID	AN_ID	METHOD_ID
15	50	5
15	51	6
15	52	7

Chemistry

AN_ID	ITEM	VALUE
50	La	3.2
50	Ce	7
50	Nd	5.2
50	Sm	1.6
50	Yb	1.52
51	Sr	319
51	Rb	11.2
51	Sm	1.57
51	Nd	5.34
52	$^{87}Sr/^{86}Sr$	0.703799
52	$^{143}Nd/^{144}Nd$	0.513018

(B) Relational Database

Fig. 4 Overview of different data structures exemplified using a subset of typical geochemical data: (A) single-file database; (B) relational database; (C) index database; (D) linked data. Example dataset is modified from Lehnert K, Su Y, Langmuir CH, Sarbas B, Nohl U (2000) A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1 (5). https://doi.org/10.1029/1999gc000026.

usually based on structured query language (SQL) and common examples include MySQL, Microsoft SQL Server (Microsoft Access), or PostgreSQL. Relational databases store data in a collection of tables linked by relational operators or "keys" (Fig. 4B). Two great advantages of this format compared to single-table spreadsheets are the minimization of duplicate data entries/records, referred to as normalization, and the speed at which data can be queried and retrieved (Codd, 1970). This format does, however, still require data to be recorded in a highly structured and rigid way. As an example, the Observations Data Model Version 2 (ODM2; Horsburgh et al., 2016) is a geoscience implementation, in a relational database structure, of the domain-agnostic Observations and Measurements standard (ISO 19156:2011; Cox, 2010, the precursor to OMS, see Section "The logical structure of geochemical data").

RDBMS require specialist skills to develop, operate, and administer relational databases; create workflows and interfaces for data ingestion and validation of data integrity; optimize performance of queries; and accommodate new requirements. Inexperienced users of RDBMS can accidentally delete or modify data and poor database design can lead to slow performance or cumbersome maintenance. This need for specialist expertise has limited the uptake of RDBMS in geochemical data management and for data compilations, as few research groups have the required skills, resources, or access to technical infrastructure. Therefore, relational databases are usually provided by dedicated data infrastructure projects as a community service to support general geochemistry research.

RDBMS may encounter performance issues when data volumes become very large and the need to accommodate a growing range of metadata makes the database schemas complex with a large number of data tables. Database "indexing" is used to enhance the speed of data retrieval by copying selected columns of data into a single, flat data structure. Indexes are used to quickly locate data, and each index entry is linked via relational operators (keys) to the relevant database tables for efficient access of the full dataset. This structure means that the index needs to be maintained simultaneously with a relational database, requiring additional storage space but optimizing compute time for data lookup and retrieval.

An index database, in contrast, introduces another layer of abstraction. Specifically designed to facilitate fast data search without the need to impose a rigid schema, an index database or "search index" groups data by certain common search categories ("index") and assigns subsets of data as entities or "documents" to each index (Quach and Ambalgekar, 2024). For example, "analytical method" and "chemical analyte" could each be assigned an index with the geochemical data for a particular sample linked as an entity to both indexes (Fig. 4C). One common, open-source index database management system is Elasticsearch.

Linked (open) data

Linked data represents a structure where relationships/connections between data are available to allow easy data access (CODATA RDM Terminology Working Group, 2024, Fig. 4D). Querying linked data enables the retrieval of both explicitly and implicitly derived information based on the data/properties themselves and their relationships. This connectivity enables the query to process the actual relationships between information and infer the answers from the network of data. Linked data are often associated with an ontology as a formal description of both properties and relationships of such data structures.

Graph databases are a common example of linked data that store information as a collection of nodes with properties and edges that represent relationships between nodes. In contrast to relational databases and metadata indexes, graph databases are optimized for querying the relationships between data records, that is, they focus on the inter-connectivity of data. Graphs therefore are a type of data network.

Diversity of geochemical data systems

Geochemical data infrastructure comprises a variety of data systems that fulfill different purposes in the life cycle of geochemical data. Terms used to describe these different components of data infrastructure are variably applied in the literature and in common usage with different meanings. For example, the term "database" is frequently used for any type of data systems, although databases are only one component in the much wider ecosystem of geochemical data infrastructure. Here, we primarily use the definitions from the Committee on Data of the International Science Council (CODATA) Research Data Management Terminology (RDM) as summarized in Table 1.

Components of geochemical data infrastructure can be broadly divided into three categories that align with the main phases in the life cycle of geochemical data: (1) data systems that support data management in the laboratory, (2) data repositories, and (3) data compilations (synthesis databases). Each of these categories is described in more detail below and their relationships are summarized in Fig. 5. In addition, sample registries play a critical role as they support persistent access to sample metadata and the use of persistent identifiers for samples. As described above, these PIDs are critical for unambiguously linking geochemical data across digital data systems to the physical sample they were generated from. The sections below serve as a general description with specific examples relevant to geochemistry following in Section "Status of today's geochemistry data ecosystem" and Table 2.

Sample registries

Sample registries are a type of repository which catalog, archive, and provide access to sample metadata that characterize and describe sample provenance, compositional or textural features, and the physical sample archive. Sample registries may register samples with globally unique and persistent identifier systems, for example, with the IGSN. Such identifier systems ensure that the physical sample and its citation in a digital data system or an electronic publication are unambiguously and persistently linked to the sample's metadata stored in the sample registry catalog. Persistent identifiers such as the IGSN allow linking subsamples to their parent samples and tracking the usage of a sample through multiple analytical procedures and laboratories (Klump et al., 2021a).

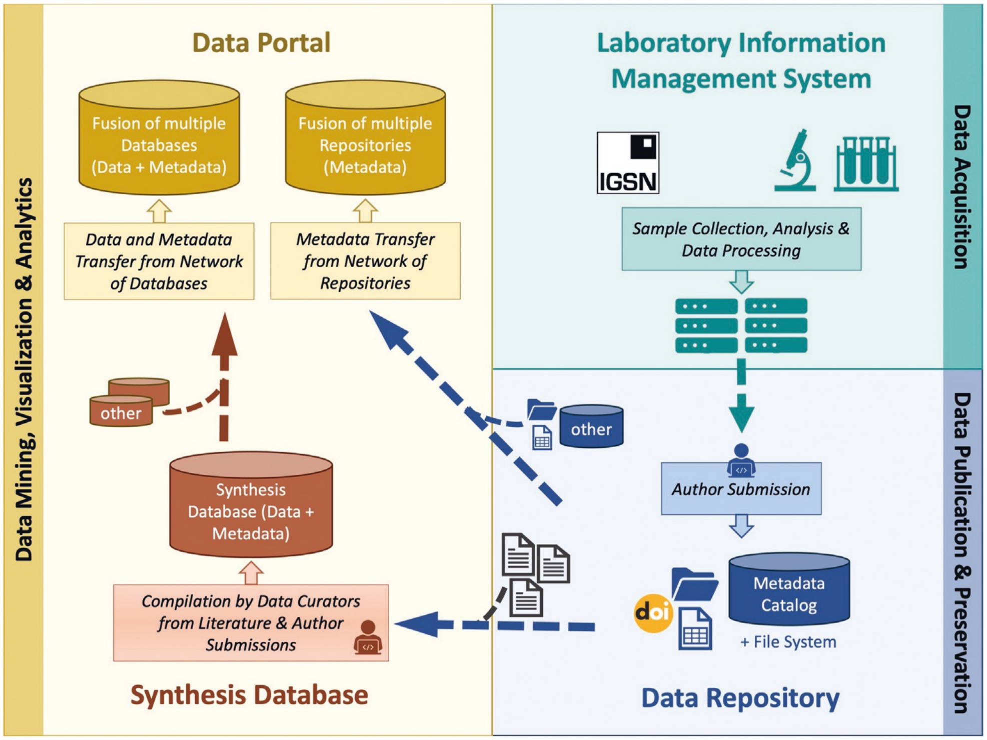

Data Portal
Fusion of multiple Databases (Data + Metadata)
Fusion of multiple Repositories (Metadata)
Data and Metadata Transfer from Network of Databases
Metadata Transfer from Network of Repositories
other
Synthesis Database (Data + Metadata)
Compilation by Data Curators from Literature & Author Submissions
Synthesis Database
Data Mining, Visualization & Analytics
Laboratory Information Management System
IGSN
Sample Collection, Analysis & Data Processing
Data Acquisition
other
Author Submission
doi
Metadata Catalog
+ File System
Data Repository
Data Publication & Preservation

Fig. 5 Schematic example of different data systems within the database ecosystem, following the data life cycle from data acquisition to data publication and preservation to data compilation and reuse.

Currently, the only openly accessible, independent sample registry for geological samples is SESAR (System for Earth and Extraterrestrial Sample Registration). However, IGSN registration is also built into several data systems and software tools for field and laboratory sample management (e.g., AusGeochem; Boone et al., 2022). An IGSN should be generated by the owner of the sample (collector or curator) as early as possible in the life of the sample to take full advantage of its features. IGSNs can be embedded in QR codes used to label the sample and its various preparations from field to lab to publication to archive.

Data management systems in the laboratory

Digital data infrastructure in the laboratory aids researchers in automatically acquiring and processing their data, as well as documenting their methods and storing the metadata (e.g., instrument parameters) required to reproduce any of the measuring or processing steps. Two main types of data systems support data management in the laboratory:

1. *Laboratory Information Management System*, referred to as LIMS, is a software to support researchers with the storage and management of laboratory information and workflows (Skobelev et al., 2011). LIMS is widely used in the commercial sector, in medical and pharmaceutical laboratories, laboratories of regulatory agencies, and other large-scale analytical facilities to increase efficiency and consistency of laboratory operations, ensure quality of measurements, support data processing, and minimize errors in reporting. LIMS' main purpose is to track samples through all steps of the analytical process in the laboratory and record any deviations from the routine procedure (Skobelev et al., 2011). Many LIMS were developed in-house, especially in the early days of digital data in the 1980s, but today there is a wide range of commercial LIMS software available that can be configured and customized to support data management needs of a laboratory. The implementation and maintenance of LIMS often requires a level of financial resources and information technology skills that many geochemical laboratories cannot afford.
2. *Electronic Lab Notebooks (ELNs)* are software solutions that support the digital recording and documentation of laboratory procedures and events including calibration, analysis, and experiment, replacing hand-written notes in paper notebooks (Dolan and Whipple, 2023). ELNs improve reproducibility of analytical results and procedures and lessen the burden and risks of manually recording and transcribing data and procedures. They also allow the direct incorporation of data and metadata from instruments. However, they merely record this information and do not automatically generate standardized data and metadata. While most ELNs are proprietary cloud-hosted solutions, there are community-developed open-source ELNs and a small number of commercial ELNs with open-source codebases as well as free (to nonprofit organizations) ELNs with proprietary codebases (Higgins et al., 2022).

In recent years, various laboratory data management systems for geochemical and geochronological laboratories have been developed or are still under development as part of cyberinfrastructure programs, large research infrastructure networks, research projects, and institutional initiatives (e.g., EPOS, AuScope). Among them are Sparrow (Quinn et al., 2021), AusGeochem (Boone et al., 2022), the Sample Analysis MicroInformation System (SAMIS; Bennett et al., 2022), and Selenocene (Kenah et al., 2020). These softwares are designed to fulfill the needs and requirements of specific laboratories and research communities. Use beyond their immediate purpose is often hampered by lack of flexibility, scalability, and sustainability, and in some cases, limited knowledge of their existence. They sometimes also act as a "black box" with limited transparency on detailed processes within the system, which may create unnecessary dependence of laboratories on particular systems.

Where databases and repositories usually only cater to the most processed data level, laboratory data management systems often cover multiple or all data levels from raw data to derived products and model outputs (see NASA Data Levels, n.d.). The great advantage of such laboratory data management systems is that raw data, including all analytical metadata, are captured without additional effort from the researcher (and the human errors inherent in retrospective data management practices). However, since these systems largely deal with unpublished data, conformance with current ethical best practices, for example, by not exposing data to third parties, is key to maintain trust and to protect and support researchers and overcome the reluctance of many geochemists to share their raw data and processing methods (see Section "Principles, standards, and best practices for modern research data management"). This trust is critical because exposure could threaten the competitive edge for research grants and many geochemists see little support and reward in the academic credit system for the time spent on diligent data management.

Data repositories and archives

Data repositories archive, curate, and disseminate data to enable current and future (re)use of these data in science to drive new discoveries. Different types of data repositories can be distinguished based on their overall scope: "*generic*" repositories provide a general service to the community, such as DOI registration, regardless of the provenance or content of the data; "*institutional*" repositories serve the members of a particular institution; "*disciplinary or domain*" repositories serve a particular scientific discipline and, therefore, are often far more specialized (i.e., more detailed, rigorous curation) than generic or institutional systems by offering templates, workflows, and user interfaces that are tailored to the data types and practices of the specific science community; "*programmatic*" repositories are those specific to a particular research program resulting in highly specialized, bespoke data infrastructures, often managing multidisciplinary data types. Zenodo, Dryad, or Figshare are examples of generic repositories. A range of programmatic and domain repositories are listed in Table 2.

Ideally, repositories offer comprehensive technical, organizational, and social solutions in order to deliver trustworthy data publication and preservation services: they need to ensure that the data they curate have rich documentation that enables discovery, access, and reuse in research and for public use. This entails instruction, guidance, and assistance to data producers. Repositories need to enable preparation of standards-compliant data packages for sustainable and trustworthy long-term data management and dissemination. One of the principles defined by the research data community to provide guidance for the stewardship of data, and to help repositories ensure the quality of data for their continued reuse as valuable research assets, is the Open Archival Information System (OAIS) reference model. The OAIS defines terminology and functional requirements for a trustworthy data repository, and its most recent version is published by the International Standards Organization (ISO) as ISO 14721-2012 (Consultative Committee for Space Data Systems, 2012). A description of other standards and principles guiding data management in repositories is provided in Section "Principles, standards, and best practices for modern research data management." Use of persistent identifiers such as Digital Object Identifiers (DOIs), rich and machine-readable metadata and data, and intuitively usable tools that support the deposition, access, analysis, and visualization of data are crucial responsibilities of repositories. Broad and inclusive community engagement should entail user support, instruction, and outreach to assist and respond to the needs of diverse users and their changes over time.

Data are submitted to repositories in the form of files, either as a single file, multiple files forming a dataset or as multiple datasets combined into a data package or bundle. Linked to these files is a set of metadata that describe the content and provenance of the dataset (e.g., title, abstract, authors, contributors, keywords), provide legal and administrative information (e.g., license, access constraints, release date), and cite related resources (e.g., DOI of the scholarly publication using the data, IGSNs of samples in the dataset). A repository is likely to define the range of data types, data structures, and file types that it accepts, based on its scope, mission, and capacity. Proprietary formats are often discouraged by research data repositories as they may become unsupported and hinder reuse. Disciplinary repositories may develop and promote templates for structuring and documenting data that facilitate compliance with community-specific expectations for data documentation. For example, the EarthChem Library provides several data submission templates for bulk and in situ analyses that also include recommendations for specific controlled vocabularies to be used (e.g., Team EarthChem, 2022).

Most datasets deposited in a repository represent original publications of a specific dataset that can be cited in a research paper (the research interpretation derived from the data) and/or in subsequent data compilations that reuse all or parts of these data. Using the example of the OSIRIS-REx Sample Analysis mission, for each of the >60 methods used to analyze the sample collected from asteroid Bennu, a distinct data package is generated that contains one or multiple data and metadata files and that is archived in a consistent format in the Astromaterials Data Archive as the domain repository for laboratory data of astromaterials (Lehnert et al., 2022). While these separate data publications may not immediately aid the scientific interpretations drawn from synthesizing multiple of these data products, individual archiving with the appropriate metadata for each method ensures the long-term reusability of the data for purposes not yet thought of today. Storage in a community-agreed standard format ensures easier interpretation, quality assessment, and reuse of the data files, and data users have greater trust in individual datasets.

Only disciplinary, or domain, repositories are in a position to enforce standard data formats specific to the discipline and the respective data type. In return, they also provide the required curation and support. As outlined above, curation of data and metadata provides valuable quality assessment and control prior to data publication, as well as ensuring that community standards are followed. This review of data and metadata will ensure that such information is provided in accordance with legal, disciplinary, and community ethical norms, in accordance with CoreTrustSeal requirements for appraisal, data quality, and confidentiality/ethics that govern trusted digital repositories (CoreTrustSeal Standards and Certification Board, 2022). CoreTrustSeal defines four levels of curation from acceptance-as-is to basic curation through to enhanced and data-level curation, where both metadata and data are reviewed and enhanced. It has been shown that the datasets published in curated domain repositories often accrue more citations than datasets in generalist repositories, since they are more discoverable to their community and more reusable due to the quality of the metadata and discipline-specific quality assessment and control (QA/QC) on the data themselves (e.g., Piwowar et al., 2007; Colavizza et al., 2020; Digital Science et al., 2021).

Another important service that repositories increasingly provide is to track the use of a given dataset and make metrics such as number of downloads or citations in the scholarly literature available to data providers and users. For as long as data publications are not yet indexed in scholarly citation trackers such as Google Scholar and Scopus, these metrics are a valuable incentive for researchers to share their data, making it possible to demonstrate the impact of their data and give credit for good data management practices, as well as to enable reporting to funders.

Synthesis (compilation) databases

The most common meaning of the term "database," and indeed the one described in the introductory examples above, is that of a data compilation that aggregates similar data from multiple sources into a single, common storage structure where the data can be readily searched, sorted, summarized, and otherwise manipulated. We here use the term "compilation database" to describe a data system in which previously unconnected data are harmonized and integrated. We use "data compilation' to refer to single-file, spreadsheet database files and "synthesis database" for more complex, often relational databases. These compilation databases are critical in modern data analytics as they strive to make data "analysis-ready" for the application of statistical and other computational methods (e.g., machine learning, neural network analysis) and for modeling and simulations. Synthesis databases enable search and retrieval of individual data points as well as entire data tables or data files, as long as the databases are structured properly and sufficient metadata are included that describe the primary analytical data and the samples from which they were acquired. As a consequence, they are extremely versatile and can be used for a variety of different applications.

The main contents of any compilation database are the "*Primary Data*," the results of chemical analyses of natural or synthetic samples, including major oxide and trace element concentrations, radiogenic and stable isotope ratios, noble gas contents, and derived data such as analytical ages and models. These chemical properties may have been measured on different types of samples and on different components of the sample (e.g., minerals, melt inclusions, presolar grains in meteorites). "*Secondary Data*" are the metadata included in the compilation. The value of a compilation database and its long-term utility are to a large extent, dependent on the quality and comprehensiveness of the metadata included. Two distinct types of compilation databases can be distinguished based on their size and purpose.

1. "*Spreadsheet compilations*" are often created by individual researchers or research groups for a particular project or topic, a specific region, material, or chemical property of interest. These spreadsheets may contain any number of analytical measurements compiled from publications and/or unpublished datasets. They are usually organized by samples and chemical properties into rows and columns, with or without metadata such as sample locations, methods, uncertainties, and references for the source of the data. This type of data compilation has in the past formed the basis of influential publications such as the ones mentioned in the policy statement of the Geochemical Society to highlight the impact of geochemical databases (see Section "Introduction"). However, spreadsheet-based compilations have significant limitations (see also Section "Spreadsheets"):
   - Many spreadsheet compilations are generated to support specific projects and will not be maintained beyond the duration of the project, thus becoming quickly outdated.
   - Different compilations even of the same or similar data types have custom formats and structures and cannot easily be merged.
   - They often only include a minimum set of metadata that is selected for a specific purpose, which means they are difficult to reuse. They also often lack essential metadata about data quality and data provenance since the creator of the compilation prescreens the suitability of the data before adding them to the compilation.
   - It is difficult to include expansive metadata in spreadsheets due to the flat file structure. In order to describe the method by which the data points for the different variables in the compilation were acquired, an additional column would need to be added for each variable as each value in the column may have a different provenance.
   - They are not a citable resource as long as they are not deposited in a repository and registered with a central metadata index. Instead, they are often stored on PCs, on other digital media, or in personal accounts in the cloud where they are not accessible to others and may at some point in the future be lost. Other researchers who are interested in the same data may not know about them and may create a new compilation, thus duplicating the work.

Nevertheless, spreadsheet-based data compilations are the most commonly used type of database in geochemistry since they are so easy to compile and use. Increasingly, they are now submitted to repositories to become citable community data resources and reference datasets (e.g., Class and Lehnert, 2012; Weiss et al., 2016; Wörner, 2021; Pilger, 2022; Rasmussen et al., 2022; Stracke et al., 2022).

2. "*Synthesis Databases*." A quite different form of compilation databases are the large-scale and long-term curated data systems, here referred to as "synthesis databases," that may store millions of data points in complex relational or linked data structures. These synthesis databases may be restricted to members of a research group, laboratory, or an organization or they can be public and open to the global research community, serving thousands of users, including students and educators, the commercial sector, and researchers in domains beyond geochemistry. Such synthesis databases may feature complex architectures with different functional layers for data storage, data search, and data visualization and analysis. Examples for such synthesis databases include the community database of the Sedimentary Geochemistry and Paleoenvironments Project, the USGS National Geochemical Database, GEOROC, and PetDB.

A key characteristic of synthesis databases are interactive web interfaces with which users can obtain exactly the dataset they need for their particular analysis, regardless of whether they are a skilled geochemist, a scientist from another earth science discipline, a teacher, a student, or a member of the public. Graphical User Interfaces (GUIs) should present to the user intuitive tools to rapidly and simply create queries to the database that use the metadata in the database as search criteria. Geographical location, sample type, geological context, or analytical method can be used as single search criteria or in combination to find the data of interest. Some databases also allow users to find samples with specific chemical properties (e.g., "find all samples with Nb <0.8 ppm"). Users can then select which chemical properties they would like to add to their search result. The search result is displayed as an HTML table in the search application or can be downloaded as a data compilation for further analysis. These interfaces are a requirement to be able to access and make use of the full versatility of synthesis databases. As synthesis databases grow to tens of thousands of analyses and millions of individual measurements, the ability to evaluate and manipulate data through traditional approaches such as visual inspection and/or plotting from spreadsheets becomes increasingly difficult. A user interface to a large database should, therefore, allow the user to evaluate data on-line to avoid wasting time and resources downloading massive datasets and then manually extracting the often small subset of data necessary for their analysis. These interfaces are an important point for future growth of databases, as advanced web services and computational interfaces (APIs) could provide direct links to modeling software and thus support machine learning applications.

The value of synthesis databases is without question. However, development and maintenance of these synthesis databases are labor-intensive and expensive. To this day, data entry and curation is mostly a manual process, where data and metadata are typed directly into spreadsheets or copied from electronic data tables and are compiled and ingested into relational databases through different workflows and software solutions. Metadata about the analytical method, data quality, or samples need to be transcribed from inside the text or from figure captions into the databases. In addition, fast-changing technical environments and the growing demands for computational applications require continuous IT support and (web) development. What is ultimately required for an efficient search across thousands of data entries is to receive data in a standardized way and then present that data to the user in a manner best suited for analysis.

Data portals

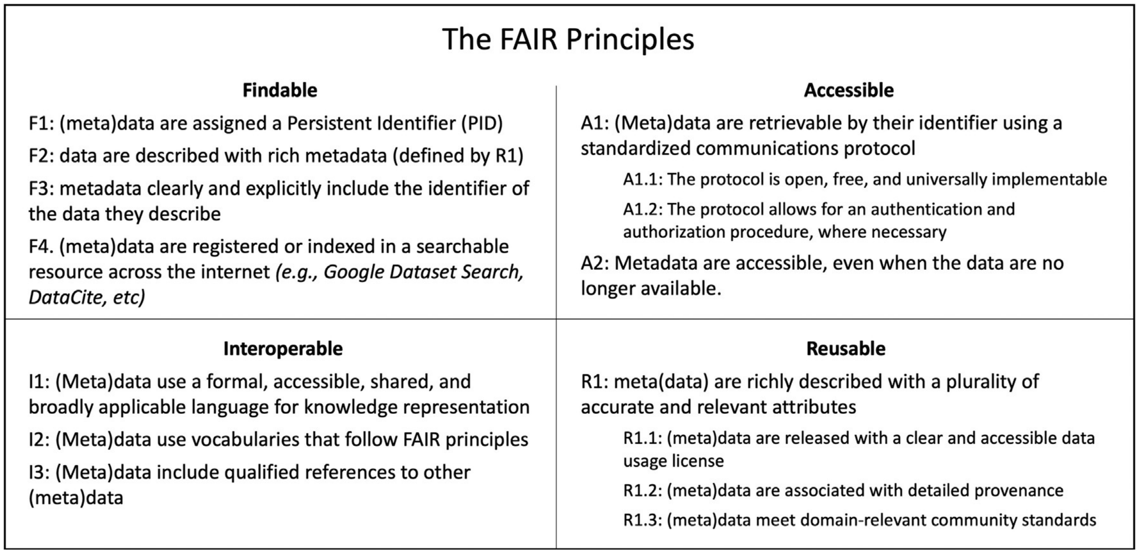
Finally, data portals even further advance data discovery and access as they provide a technical infrastructure that allows researchers to find, extract, and integrate data from multiple independent data providers through a single user interface. Often, data portals only harvest the metadata of contributing databases/repositories into a single metadata index. They therefore allow discovery of data across different providers, but not the advanced filtering and integration of the data themselves. Unfortunately, often insufficient metadata are provided that do not enable acknowledgment/citation of the original researcher and their funding sources. Few data portals have achieved full interoperability of the primary data held within different data systems/infrastructures as it is much more complex and requires participating data systems to make their data and metadata available in a consistent structure and with agreed-upon vocabularies. One successful example is the XML schema developed by EarthChem for encoding data and metadata of igneous and metamorphic rock geochemistry that has enabled the EarthChem Portal to serve as a single point of access to fully interoperable data in a globally distributed network of more than five geochemical databases (Hsu et al., 2011).

Principles, standards, and best practices for modern research data management

In recent years, as Open Science policies have led to the proliferation of research data infrastructure, three key sets of discipline- and technology-agnostic high-level principles and guidelines have been published to guide the description, management, curation, and sharing of research and scientific data: FAIR, TRUST, and CARE, which are described further in Sections "The FAIR principles: Enhancing discovery and reuse of geochemical data," "TRUST principles for data repositories that store geochemical data," and "The CARE principles." Albeit at a high level of abstraction, these three sets of principles are designed to enable multidisciplinary sharing of data and are now commonly used across all research disciplines. Adherence to these principles will ultimately enable geochemical data to be integrated with trusted ethical data from many other disciplines, a critical step toward addressing grand societal challenges such as the 17 United Nations Sustainability Goals (Wyborn and Lehnert, 2021; Klöcking et al., 2023). These principles will also lay the foundation for an ecosystem that will nurture global geochemical data networks and enhance trust in sharing and publishing of geochemical data. Although most relevant to large data systems, these principles are becoming increasingly important to researchers as a growing number of funders and publishers expect compliance for any new data publications.

The FAIR principles: Enhancing discovery and reuse of geochemical data

The FAIR principles, first defined by Wilkinson et al. (2016), provide guidelines to enhance the reusability of scientific data under the groupings of Findable, Accessible, Interoperable, and Reusable (Fig. 6). The principles recognize that as the data deluge grows, humans are no longer able to handle the volumes of research data at the scope, scale, and timeliness now required. The FAIR Principles of Wilkinson et al. (2016) are not just for humans, but put specific emphasis on providing guidelines on how machines can be enabled to automatically find and reuse data. Implementation of these guidelines in turn requires new technical



The FAIR Principles

Findable

F1: (meta)data are assigned a Persistent Identifier (PID)

F2: data are described with rich metadata (defined by R1)

F3: metadata clearly and explicitly include the identifier of the data they describe

F4. (meta)data are registered or indexed in a searchable resource across the internet *(e.g., Google Dataset Search, DataCite, etc)*

Accessible

A1: (Meta)data are retrievable by their identifier using a standardized communications protocol

A1.1: The protocol is open, free, and universally implementable

A1.2: The protocol allows for an authentication and authorization procedure, where necessary

A2: Metadata are accessible, even when the data are no longer available.

Interoperable

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2: (Meta)data use vocabularies that follow FAIR principles

I3: (Meta)data include qualified references to other (meta)data

Reusable

R1: meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1: (meta)data are released with a clear and accessible data usage license

R1.2: (meta)data are associated with detailed provenance

R1.3: (meta)data meet domain-relevant community standards

Fig. 6 The FAIR principles as defined by Wilkinson et al. (2016).

developments to enable machine-actionable data access and reuse, as well as convergence of data from multiple online sources into coherent analysis-ready datasets. Many funding agencies now make compliance with FAIR a requirement (e.g., NASA Science Mission Directorate, 2022; ARDC, 2020; NSF, n.d.; Horizon Europe Funding, n.d.). Since the FAIR principles are very abstract, several guidelines are available to help researchers and repositories implement and/or comply with them (e.g., Mons et al., 2017; Jacobsen et al., 2020).

Under the FAIR principles, each digital dataset must be accompanied by a structured metadata file that provides machine-readable and searchable explicit descriptions of the context of the data. Domain-specific standards are essential to define what context is crucial for the reuse of their specific data types. For geochemical data, such context would include descriptions of provenance of the sample, analytical method, analytical data quality, researcher, funder, and organization.

Principles F1, F3, and F4 require generic technical implementations that apply to all data types. For example, to enable discovery, citation, and accreditation and to track usage, data should be stored in a sustainable, trusted repository that is accessible online, and a globally unique, resolvable PID should be assigned to the dataset. Metadata should be discoverable globally via, DataCite or Google Dataset Search. Most geochemistry journals now require a PID (usually a DOI) for any dataset cited in a publication that links to the repository where the dataset is stored. F2, in contrast, can vary from community to community and will be influenced by the (geochemical) data type and the intended purpose for finding and reusing data. By definition, "rich" metadata allow users to discover, understand, and process the data by machines (Brümmer et al., 2014; GO FAIR, n.d.).

To optimize findability of a dataset by multiple methods and search engines, it is essential for geochemical communities to define standardized metadata descriptors, preferably using standardized controlled vocabularies (Jacobsen et al., 2020). For geochemistry, these metadata should include sample classification, analytical method, data type, and laboratory (see Section "The ecosystem of geochemical data management"; Goldstein et al., 2014; Chamberlain et al., 2021). Information on the geospatial location of the site a natural sample is collected from is desirable, but in some sensitive cases may need to be restricted, while for synthetic samples this attribute is usually irrelevant.

Ideally, data should be available online with minimum access constraints and no requirement to register or login (A1.1). Providing data in multiple file formats for download, such as SQL, CSV, or even NetCDF and HDF5 for large data volumes, may further increase accessibility. Where there are authentication and authorization requirements, these need to be specified and easily understood (A1.2). Increasingly journals ask for Data Availability Statements (sometimes called Data Accessibility Statements) for each article, which tell readers where any data associated with a paper are available from and, if there are restrictions, under which conditions the data can be accessed. Principle A2 ensures that the knowledge that a geochemistry dataset was once accessible is preserved through the metadata record, even if a later, updated version has since been published.

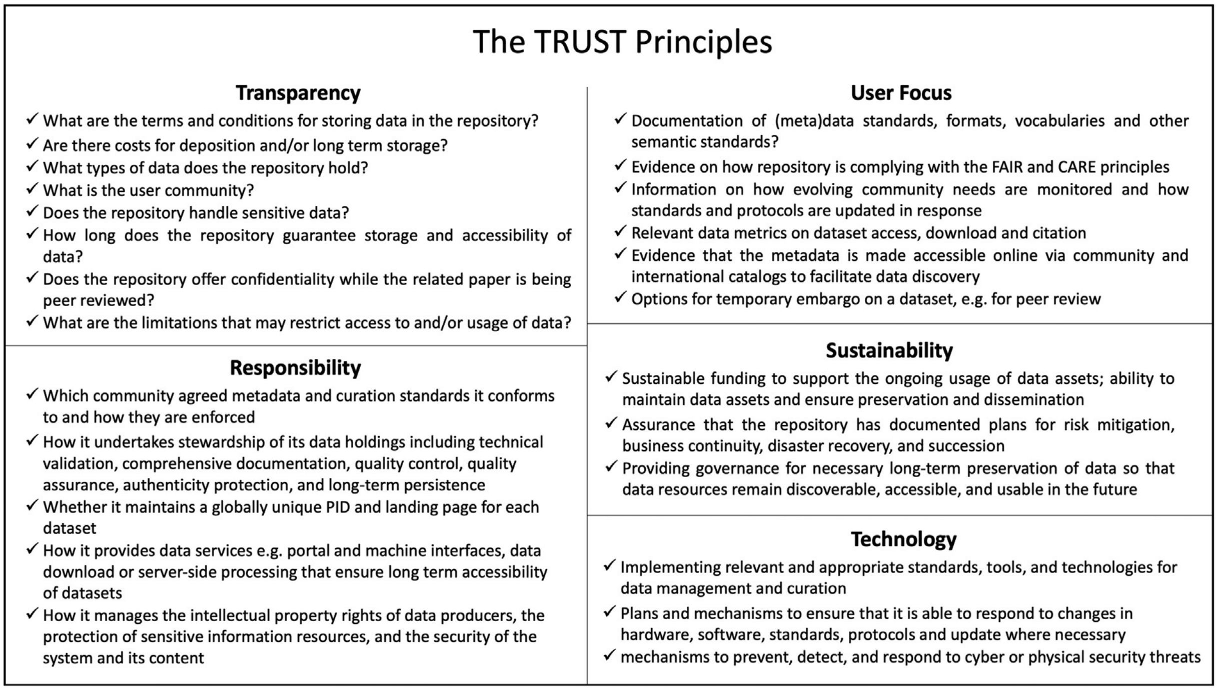
Wilkinson et al. (2016) define interoperability as "*the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort*." The interoperability principles are essential to enabling seamless integration of equivalent datasets from multiple organizations and repositories. They are difficult to achieve for geochemical data, due to the lack of agreed community standards and controlled vocabularies for describing the metadata and data elements of a dataset. Anecdotal evidence suggests that geochemical data comply with the 80:20 Rule of Data Science, in that researchers spend 80% of their time finding, cleaning, and reorganizing their data to be ready for use, and a mere 20% of their time actually analyzing and reporting on their aggregated datasets (Ruiz, A., 2017; Kim and Hardin, 2020).

Like the interoperability principle, the reusability principles are hard to achieve for geochemical data. For R1.1, machine-readable licenses and copyright information need to be permanently linked to the data that indicate how they can be reused and by whom. The license used should preferably have minimal restrictions (e.g., CC By 4.0; Australian Research Data Commons, 2019). For Principle R1.2, it is important that (meta)data provide detailed provenance information that cites any source data and algorithms that generated the data, including documentation of peer review and any QA/QC procedures that have been applied to the dataset. Above all, for Principle R1.3, (meta)data need to meet discipline-relevant community standards that must be developed at the international level and preferably endorsed by an authoritative source such as a scientific union or society.

How can geochemical data comply with the FAIR Principles? Compliance of geochemical data with the FAIR principles requires community agreement on implementation choices for each of the FAIR principles. The challenge is that geochemical data are markedly diverse, making it nearly impossible to define a single one-size-fits-all formula to make every geochemical dataset comply with the FAIR guiding principles. Although some of the FAIR principles are shared across all datasets (e.g., assigning a globally unique and persistent identifier to metadata, F1, and making (meta)data retrievable by their identifier using a standardized communications protocol, A2), several of them will need implementations that are very specific to individual geochemical data types (e.g., I2 (meta)data use controlled vocabularies that follow FAIR principles; R1.3 (meta)data meet domain-relevant community standards). Hence, what is appropriate, feasible, and achievable, particularly with respect to choices of metadata and data standards, vocabularies and data formats, ways of recording provenance, and services used to access the data, largely depends on the individual geochemical communities and their objectives, drivers, funding requirements, and capabilities.

TRUST principles for data repositories that store geochemical data

Digital data repositories are a crucial element for the implementation of data policies by governments, agencies, and funders aiming to ensure open access and preservation of FAIR data. Repositories need to demonstrate that they operate in a reliable manner, that they ensure quality and longevity of the data under their stewardship, and that they align their services with the needs of their users. The TRUST Principles (Transparency, Responsibility, User Focus, Sustainability, and Technology), introduced by Lin et al. (2020),



The TRUST Principles

Transparency

- ✓ What are the terms and conditions for storing data in the repository?
- ✓ Are there costs for deposition and/or long term storage?
- ✓ What types of data does the repository hold?
- ✓ What is the user community?
- ✓ Does the repository handle sensitive data?
- ✓ How long does the repository guarantee storage and accessibility of data?
- ✓ Does the repository offer confidentiality while the related paper is being peer reviewed?
- ✓ What are the limitations that may restrict access to and/or usage of data?

Responsibility

- ✓ Which community agreed metadata and curation standards it conforms to and how they are enforced
- ✓ How it undertakes stewardship of its data holdings including technical validation, comprehensive documentation, quality control, quality assurance, authenticity protection, and long-term persistence
- ✓ Whether it maintains a globally unique PID and landing page for each dataset
- ✓ How it provides data services e.g. portal and machine interfaces, data download or server-side processing that ensure long term accessibility of datasets
- ✓ How it manages the intellectual property rights of data producers, the protection of sensitive information resources, and the security of the system and its content

User Focus

- ✓ Documentation of (meta)data standards, formats, vocabularies and other semantic standards?
- ✓ Evidence on how repository is complying with the FAIR and CARE principles
- ✓ Information on how evolving community needs are monitored and how standards and protocols are updated in response
- ✓ Relevant data metrics on dataset access, download and citation
- ✓ Evidence that the metadata is made accessible online via community and international catalogs to facilitate data discovery
- ✓ Options for temporary embargo on a dataset, e.g. for peer review

Sustainability

- ✓ Sustainable funding to support the ongoing usage of data assets; ability to maintain data assets and ensure preservation and dissemination
- ✓ Assurance that the repository has documented plans for risk mitigation, business continuity, disaster recovery, and succession
- ✓ Providing governance for necessary long-term preservation of data so that data resources remain discoverable, accessible, and usable in the future

Technology

- ✓ Implementing relevant and appropriate standards, tools, and technologies for data management and curation
- ✓ Plans and mechanisms to ensure that it is able to respond to changes in hardware, software, standards, protocols and update where necessary
- ✓ mechanisms to prevent, detect, and respond to cyber or physical security threats

Fig. 7 The TRUST principles as defined by Lin et al. (2020).

were collaboratively developed and endorsed by members from across the various sectors of the digital repository community, and are designed as a set of guiding principles to demonstrate TRUSTworthiness of a digital data repository and confirm that they are reliable and capable of appropriately managing any data that they hold. A repository depends on the interaction of people, processes, and technologies to support secure, persistent, and reliable services. Its activities and functions are supported by software, hardware, and technical services. Together, these provide the tools to enable the delivery of the TRUST Principles (see Fig. 7).

The Transparency principle requires repositories to be "*transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence*" (Lin et al., 2020). Transparency is about how the operational infrastructure and procedures allow users (researchers, institutions, agencies) to ascertain which repositories are sustainable and able to provide a long-term home for their data. The Responsibility principle requires repositories "*to be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service*" (Lin et al., 2020). Responsible TRUSTworthy repositories should adhere to metadata and data standards for geochemical data as this will enhance discovery and reusability of a dataset and assure contributors and users that datasets in the repository will be interoperable with other equivalent datasets from anywhere around the globe. A responsible repository will also maintain a consistent PID and landing page for the dataset to ensure that it will remain accessible over time and can be consistently cited and referenced in scholarly publications.

The User Focus principle aims "*to ensure that the data management norms and expectations of target user communities are met*" (Lin et al., 2020). For generic repositories, the user community is very broad and expectations are not very high: it is unlikely that any of them will support the specific needs of the geochemistry community, in particular for (meta)data standards and controlled vocabularies. The Sustainability principle requires a repository to "*sustain services and preserve data holdings for the long-term*" (Lin et al., 2020). Sustainability is critical to any researcher as the repository needs to be able to provide uninterrupted access to its data assets for current and future use cases and communities. Ideally, the repository can provide evidence that it has sustainable funding over a period of decades.

The Technology principle requires TRUSTworthy repositories "*to provide infrastructure and capabilities to support secure, persistent, and reliable services*" (Lin et al., 2020). The consumers of the data should not need to know that changes have been made in the backend and the data presentation should not be affected.

How can geochemical data repositories comply with the TRUST Principles? Modern digital repositories for geochemical data are taking the place of the traditional analog libraries that stored and preserved the wealth of geochemical data in hundreds if not thousands of paper journals. Hence to assure long-term access to geochemical datasets, geochemical data repositories should follow the TRUST principles and at a minimum provide: Evidence that the repository has sustainable funding over extended time frames; Evidence of a long-term data preservation policy; Documentation of basic geochemical data curation and review; Evidence on how the relevant technologies are used for making data accessible online and are updated where required. In practice, these principles are hard to achieve, especially for smaller, community-led domain repositories, and at present few geochemical repositories are fully TRUST-compliant. The main hurdle is financial sustainability as many domain repositories struggle for long-term funding. In order to achieve broadly TRUSTworthy repositories, the geochemistry community globally needs to engage with and support geochemical data repositories by actively contributing to the development and maintenance of geochemical data standards, by adopting these standards throughout the geochemical data life cycle, and by advocating for the sustained support—financial and otherwise—of geochemical data infrastructure. It is also essential that researchers correctly cite the DOIs of datasets from repositories as proposed by Stall et al. (2023) so that the data citations in scholarly publications can be picked up by the indexers, thus enabling quantification by researchers, repositories, and funders of uptake and impact of datasets in repositories and calculation on Return of Investment (ROI).

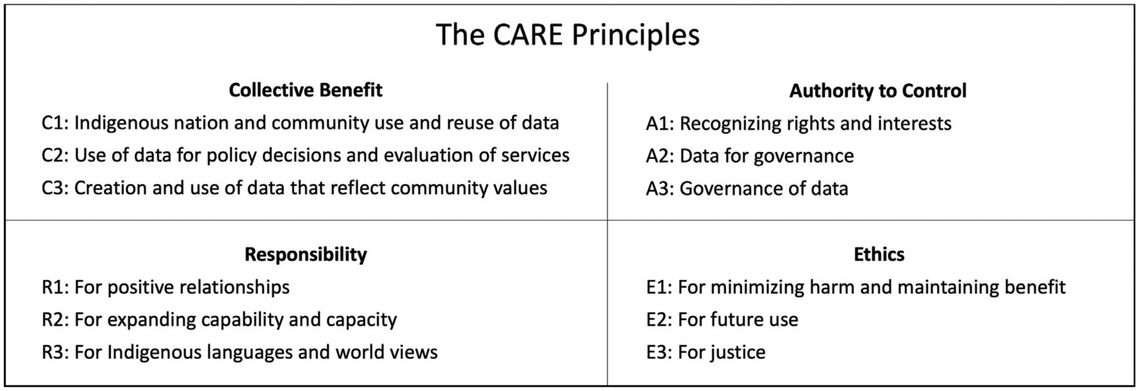
The CARE principles

The CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) were developed in consultation with Indigenous peoples, scholars, nonprofit organizations, and governments to address concerns about secondary use of data and samples (Carroll et al., 2020). The principles apply only to Indigenous data and samples and seek to (1) respect Indigenous data sovereignty and protect Indigenous rights and interests in Indigenous data, including traditional knowledges and local contexts; and (2) support open data, including secondary use in machine learning, broad data sharing, and big data initiatives (Carroll et al., 2020; Williamson et al., 2022). As with the FAIR principles, the CARE principles provide guidelines only and are subject to interpretation. Increasingly, papers are emerging that provide more details on how researchers and institutions can operationalize the CARE principles as well as emphasizing that the CARE principles should be applied in parallel with the FAIR and TRUST principles (e.g., Carroll et al., 2021; Taitingfong et al., 2023).

There are four groups of CARE principles defined by Carroll et al. (2020) and summarized in Fig. 8. Data ecosystems shall be designed and function in ways that enable Indigenous peoples to derive benefit from the data (Collective Benefit). Indigenous peoples' rights and interests in Indigenous data must be recognized, and their authority to control such data must be respected (Authority to Control). Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous peoples' self-determination and collective benefit (Responsibility). Indigenous peoples' rights and well-being should be the primary concern at all stages of the data life cycle and across the data ecosystem (Ethics).

How can geochemistry adopt the CARE Principles? In the past, Indigenous rights and knowledge have not always been respected, and there are instances of geochemical samples being taken from Indigenous lands without permission as forms of helicopter or parachute research (e.g., Adame, 2021; Stefanoudis et al., 2021). Worse still, as examples of parasitic research, in some cases local knowledge has been exchanged with the Indigenous community, but this is not acknowledged in the scientific literature and/or there is no attempt to enable the Indigenous community to have access to these data and have a say in how they can be used/reused (The Lancet Global Health, 2018). Indigenous Data Sovereignty refers to the right of Indigenous peoples to exercise ownership over Indigenous data, which, by definition, includes the control of data about their peoples, lands, waters, samples, and resources, from collection to use and reuse (Carroll et al., 2020; Williamson et al., 2022). That is, the term "Indigenous data" does not just refer to Indigenous artifacts or specific data on Indigenous peoples.

Hence, for future research in geochemistry, it is important that the relevant Indigenous communities are engaged throughout the whole geochemical data life cycle, even prior to any geochemical sampling being undertaken, that local knowledges are included, and more importantly that the Indigenous community has access to the data and a voice in how they can be used, including any commercialization of data and results. For future sample acquisition, it is essential that the relevant Indigenous communities are engaged prior to any samples being collected, and that wherever possible, local knowledge is included in the collection process to avoid incidents such as the unauthorized sampling of the Bishop Tuff in California and other cases elsewhere (Sahagún, 2021). Likewise, under certain protocols, any data derived from that sample can also be subject to conditions. For example, the Nagoya Protocol (Buck and Hamilton, 2011) addresses access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity and stipulates that where commercialization has taken place on samples from Indigenous lands, the local communities may need to be compensated (e.g., Sheridan, 2004).



The CARE Principles

Collective Benefit	Authority to Control
C1: Indigenous nation and community use and reuse of data C2: Use of data for policy decisions and evaluation of services C3: Creation and use of data that reflect community values	A1: Recognizing rights and interests A2: Data for governance A3: Governance of data
Responsibility	**Ethics**
R1: For positive relationships R2: For expanding capability and capacity R3: For Indigenous languages and world views	E1: For minimizing harm and maintaining benefit E2: For future use E3: For justice

Fig. 8 The CARE principles as defined by Carroll et al. (2020).

Operationalizing the CARE principles by both repositories and researchers is not as advanced as the FAIR and TRUST principles. To address this discrepancy, the Indigenous Metadata Bundle Communiqué was developed by the Collaboratory for Indigenous Data Governance (https://Indigenousdatalab.org/), ENRICH: Equity for Indigenous Research and Innovation Coordinating Hub (https://www.enrich-hub.org/) and Tikanga in Technology (https://www.waikato.ac.nz/rangahau/koi-te-mata-punenga-innovation/TinT) (Taitingfong et al., 2023). The communiqué provides a conceptual framework for use throughout the data ecosystem, including repositories, and provides specificity for Indigenous peoples' data. The communiqué provides guidance on the Collective Benefit and Authority to Control principles and identifies five categories for metadata elements: governance, provenance, lands and waters, protocols and local contexts notices, and labels (https://localcontexts.org/). Categories that support the Responsibility and Ethics principles include language, PIDs, classification systems, Indigenous names and taxonomies, data quality, and relationships to FAIR and TRUST.

Standards and best practices for geochemical data

"Standards" are the foundation of many products and processes in our lives. They allow us to use a credit card anywhere in the world; to fit different types of light bulbs into our lamps; to trust the reliability of household appliances; and to connect our phones to cellular networks in different countries. Standards play a critical role in information technology, providing specifications for the formats, structures, terminologies, and processes in information architectures to make them compatible and reliable. Standards are implemented in browsers, search engines, and other software that connect multiple components on the web. Standards are also critical in research data management. They ensure trust in the quality of data and services; enable software to interact with data; and allow exchange and integration of data between distributed data resources. The CODATA RDM Terminologies (2024) define a standard as a "*set of agreed-upon and documented guidelines, specifications, accepted practices, technical requirements, or terminologies that have been prepared by a standards developing organization or group, and published in accordance with established procedures*" (CODATA RDM Terminology Working Group, 2024, Table 1). Standards can be established at the international, national, or regional level, but they do require endorsement by a recognized authority and broad adoption. This is a major issue in the field of geochemistry since there is no internationally recognized standards body to coordinate and determine standards (see Section "A question of authority: Who needs to act?"), equivalent to those being set up in the International Union of Pure and Applied Chemistry (IUPAC), International Union of Crystallography (IUCr), or the International Federation of Digital Seismograph Networks (Klöcking et al., 2023). Presently, the closest entity to true standards in geochemistry are the globally accepted values for certified reference materials that are used for calibration, method validation, quality control, and quality assurance and are fundamental to comparing measurement results between different laboratories (Jochum and Enzweiler, 2014). Reference materials are continuously analyzed and reanalyzed in laboratories around the world and monitored and compiled in the GeoReM database (Jochum and Nohl, 2008). For most other aspects, at best geochemistry only has recommendations and best practice documents that do not qualify as formal standards because there is no endorsed international agreement. In addition, very few of the recommendations themselves are readily available and FAIR-compliant. However, FAIR, machine-actionable standards that are endorsed through broad community agreement are a fundamental requirement for the discipline to move forward and harness the future potentials of harmonized geochemical data. Internationally agreed standards and best practices are essential for compliance with the FAIR principles. In the original definition of the FAIR principles, Wilkinson et al. (2016) emphasized that the principles were designed to enhance machine-readability of data. Schultes et al. (2020) further developed the concept of FAIR Implementation Profiles (FIPs) in which communities that wish to interoperate data declare which standard, vocabulary, or format they use, ensuring that these FIPs are themselves also FAIR and accessible online.

We here distinguish between content standards, best practices, and technical standards. In order to achieve machine-to-machine interoperability of geochemical data and to ensure that the scientific content (including attributes in both the metadata and the data) is accurately documented, agreement is required on:

1. minimum variables to describe individual geochemical data types;
2. controlled vocabularies and their definitions for data description; and
3. documentation of standards for reporting data, including any formats used to structure the data.

Content/scientific standards

As argued by Goldstein et al. (2014): "*Access to the complete data, upon which new scientific discovery and knowledge is based, is a fundamental requirement for the reproducibility of scientific results.*" In geochemistry, the same data type is often collected by multiple laboratories and institutions using multiple analytical techniques. As a consequence, even if each technique follows an agreed protocol and the analytical objectives are comparable, it is often difficult to directly compare data acquired in different laboratories because of differences in terminologies used for metadata, column headings, and qualitative descriptive values within a dataset. These types of discipline-specific knowledge are controlled by "content" or "scientific" standards.

Use of shared terminology is crucial for accurate communication of scientific content and a critical enabler for integration of data (Cox et al., 2021), particularly in machine-to-machine interoperability of data as advocated by the FAIR principles (Wilkinson et al., 2016). Controlled vocabularies are used to create an understanding between the user community of the meaning, context, and definition of the words used to describe both real-world phenomena and abstract concepts. Further, since creating a global network of geochemical data will involve multilingual groups, precision on concepts, and terminology between multiple languages will become critical. Effective translations require formal definitions of each term of a vocabulary based on concepts, not just a simple word-for-word translation of consecutive words, which is a limitation of most on-line translation tools (David et al., 2022).

Controlled vocabularies can be ordered in increasing complexity Zeng (2008); Duerr et al. (2023); Vanderbilt et al. (2010) and Madin et al. (2007) and can range from:

1. Controlled list of terms with no definitions;
2. Glossaries that have agreed definitions;
3. Taxonomies where terms are ordered into hierarchical groups or categories based on particular characteristics;
4. Data models that are abstract models, which organize elements of data and standardize how they relate to one another and to the properties of real-world entities;
5. Thesauri, which are sets of terms representing concepts and the hierarchical, equivalence, and associative relationships connecting them; and
6. Ontologies that are quite abstract and provide formal representations of the knowledge (concepts) and the relationships within a scientific discipline.

Definitions for each of these terms can be found in Table 1. Controlled lists, glossaries, and taxonomies are currently sufficient for most geochemical use cases to understand the terms and names used in both data and metadata. The GeoSciML and EarthResourceML vocabularies of the International Union of Geological Sciences (IUGS) Commission for the Management and Application of Geoscience Information (CGI) are particularly valuable because they have been developed and maintained as an online reference since 2005 and 2020, respectively, with contributions from many nations: they are also endorsed by the IUGS and therefore qualify as formal standards.

In addition to vocabularies, there are also a growing number of publications that are documenting efforts to establish agreement on minimum variables for various geochemical data types (e.g., see list in Klöcking et al., 2023). However, most of these guidelines are listed in text-rich, unstructured documents and PDFs and still need to be modernized into technical implementations and documents that themselves comply with the FAIR principles. In addition, only very few are internationally agreed and followed by all members of a specific subdiscipline or for individual data types. As a consequence, there are very few true "standards" in geochemistry, although there is an increase in published best practices that could eventually lead to convergence on international standards.

Documentation of best practices

In contrast to formal standards, best practice documents describe workflows and QA/QC procedures recommended (but not yet widely adopted) for a particular data type by a specific community. Simpson et al. (2018) define a community best practice as "*a methodology that has repeatedly produced superior results relative to other methodologies with the same objective. To be fully elevated to a best practice, a promising method will have been adopted and employed by multiple organizations.*" In recent years, many community best practices have been developed and published. For example, Wallace et al. (2022) have produced a document outlining best practice for acquiring and publishing data from tephra samples. Similar best practices exist for some of the geochronological methods (Schaen et al., 2020; Flowers et al., 2022).

Informed decisions on whether or how to (re)use any digital analytical dataset are always dependent on a consideration of what practices have been used to obtain the data and the provision of information about the quality specifications (Peng et al., 2022). Thorough documentation of QA/QC procedures undertaken on a datasets could revolutionize this process and greatly enhance interoperability, quality assessment, and reuse of datasets. This has been shown by the success of subdisciplinary or regional groups publishing community best practices (e.g., Oceans Best Practice Site; Przeslawski et al., 2022).

Publication of QA/QC procedures and validation tests applied to any dataset is crucial to enabling a prospective user of a published dataset to evaluate its fitness for purpose. However, it takes considerable time to get international agreement on the precise QA/QC standards to be applied to a particular analytical instrument or dataset. As an alternative, Peng et al. (2022) developed a set of global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. They recommend that for any dataset there should be at least three documents: (1) a data specification that is a detailed description of the data (including quality attributes) so that it can be created, supplied to, and used by another party; (2) the determination of the tests that will be used to validate whether the dataset meets the documented specification; and (3) publication of the test results of the validation of a specific dataset against the agreed test procedures. Each one of these documents should be published with a DOI and linked to the dataset. Note that (1) and (2) could be developed by a community and could relate to a particular instrument or data type. Convergence and harmonization of specifications and tests across multiple communities will lead to greater levels of standardization across analytical techniques and ultimately could produce agreed international protocols and standards. The ever-increasing demands for machine-readability of data, particularly for AI and ML applications, are leading to a proliferation of online vocabularies many of which can only be read by specific APIs and/or apply to unique datasets. Many geochemical datasets have lists of terms that are specific to a single dataset only. Users are often uncertain as to which ones are the most authoritative vocabularies and convergence is becoming a high priority. The size of the community that creates the vocabulary can be local, regional, international or global, and the larger the size of a community that reliably and sustainably creates and maintains a vocabulary, the bigger the community is that can easily share and understand the data. As convergence on agreed specifications progressively takes place, and if vocabularies become adopted at national and international levels, they can become official standards that are preferably endorsed by a recognized authority such as a scientific union or society. The geochemistry community therefore needs to come together internationally and, for each analytical method, describe best practices for data reporting and data exchange. This will not be an easy task because of the diversity of instruments, analytical methods, and sample types within geochemistry. However, progress is being made and there already is a growing number of publications that are documenting efforts to establish agreement on minimum variables and controlled vocabularies for various geochemical data types (e.g., see list in Klöcking et al., 2023).

Technical standards

Alongside the standards and best practices that capture geochemistry knowledge and procedures, there is a requirement for technical standards that (1) improve geochemical data management, particularly in the laboratory; and (2) enable knowledge to be effectively shared between human and machine-actionable systems on the Internet.

Examples of technical standards that improve data sharing are

- File formats and languages such as CSV, HTML, XML (eXtensible Markup Language), and JSON (JavaScript Object Notation).
- Computational frameworks such as RDF (Resource Description Framework) which is a standardized way to interchange data and metadata, including complex hierarchical relationships. There are various extensions from RDF, including SKOS (Simple Knowledge Organization System) and OWL (Web Ontology Language).
- Communication protocols such as APIs.
- Persistent identifiers such as DOI, IGSN, ORCID, and ROR.
- Technical implementations of content standards, such as machine-actionable structured vocabularies.

Many of these technical standards are endorsed by major International standards bodies such as the International Organization for Standardization (ISO), Open Geospatial Consortium (OGC), World Wide Web Consortium (W3C), and Institute of Electrical and Electronics Engineers (IEEE). They are domain-agnostic and are foundational to being able to integrate geochemical data with data from other disciplines.

Technical standards do not need to be understood by all members of the geochemistry research community. Geochemists may need to partner with technical experts in order to ensure the best structuring and uptake of the scientific content standards that they develop, and to ensure their availability in online vocabulary repositories. Nonetheless, adoption of more of these technical standards within geochemical workflows will ultimately contribute to making data management and data sharing easier for geochemists, particularly in multidisciplinary projects. A potential future solution across scientific disciplines is the Cross-Domain Interoperability Framework (CDIF) proposed by Gregory and Hodson (2023).

Status of today's geochemistry data ecosystem

Present data ecosystem

Following the definitions and distinctions of Section "The ecosystem of geochemical data management," Table 2 provides an overview of data systems that are currently available for use by geochemists globally. Fig. 9 illustrates the roles of these different data systems and their services to support data management during a typical geochemical research workflow. Note that the list of data systems in Table 2 is only a snapshot of an ever-evolving ecosystem. SedDB and MetPetDB are just two examples of legacy databases that are no longer actively maintained, while the rise of Open Science and "big data" approaches has resulted in the development of a great number of new data systems in recent years. For up-to-date information, re3data (https://www.re3data.org/) provides a curated registry of research data repositories that also includes some synthesis databases and laboratory data management systems. The DataCite Search (https://commons.datacite.org/) provides a registry of repositories, although to date it is less comprehensive than re3data. FAIRsharing.org (https://fairsharing.org/) offers a registry of databases as well as repositories. Table 2 is not comprehensive of all data systems relevant to geochemistry. The list only includes data systems that have committed to implementing the FAIR, TRUST, and CARE principles where appropriate (see Section "Principles, standards, and best practices for modern research data management"). In particular, at least basic geochemical data curation and review must be provided. Generalist and institutional data repositories, such as Dryad, Figshare, or Zenodo, have been specifically avoided since they do not offer any geochemical data curation. Furthermore, with the exception of laboratory data management systems, those data systems that are defined by national boundaries are also excluded, for example, systems that only accept data generated with support from a specific

Table 2 Overview of current actively curated and maintained geochemical data systems, grouped by their purpose following Section 3.6. Note that only international disciplinary and programmatic systems that provide data review and curation are listed here. Sources include re3data and FAIRsharing.org.

	Sub-Discipline	*Scope of Data System*
Data Compilation: Synthesis Databases		
Astromaterials Data Synthesis	Cosmochemistry	Database of laboratory analytical data generated on astromaterials samples. Incorporates the MetBase legacy data. https://search.astromat.org/
PetDB	Igneous geochemistry	The Petrological Database of the Ocean Floor (PetDB) compiles geochemical data from igneous and metamorphic rocks, glasses and minerals. https://search.earthchem.org/
GEOROC	Igneous geochemistry	The GEOROC Database (Geochemistry of Rocks of the Oceans and Continents) compiles published analyses of igneous and metamorphic rocks, glasses, minerals and inclusions from all geological settings on Earth. https://georoc.eu/
GeoReM	Geological reference materials	GeoReM (Geological and Environmental Reference Materials) is a database for reference materials of geological and environmental interest, such as rock powders, synthetic and natural glasses as well as mineral, isotopic, biological, river water and seawater reference materials. https://georem.mpch-mainz.gwdg.de/
EarthChem Portal	General geochemistry	Single point of access to independently operated databases for geochemistry, petrology, mineralogy. Combined search across PetDB, USGS, GANSEKI and GEOROC data; also includes the dormant NAVDAT, SedDB, MetPetDB compilations http://portal.earthchem.org/
LEPR / TraceDs	Experimental petrology	The Library of Experimental Phase Relations and Trace element Distribution experimental database compiles results of published experimental studies involving phase equilibria relevant to natural systems, including major elemental partitioning in magmatic systems and trace element partitioning between silicate liquids, mineral phases and aqueous solutions. https://lepr.earthchem.org/
IsoArcH	archaeometry	Community-driven platform for isotope research in bioarchaeology and forensic sciences; isotope database for bioarchaeological samples that consists of georeferenced isotopic, archaeological, and anthropological information. https://isoarch.eu/
GlobaLID	Pb isotopes for archaeometry	GlobaLID is a Global Lead Isotope Database and aims to facilitate the reconstruction of raw material provenances with lead isotopes, especially in archaeology. https://globalid.dmt-lb.de/
IsoBank	Stable isotope geochemistry	Stable isotope measurement data originating from any context. https://isobank.tacc.utexas.edu/
DataONE	general	Network of interoperable data repositories facilitating data sharing and discovery. https://www.dataone.org/
World Ocean Atlas (WOA)	oceanography	Temperature, salinity, oxygen, phosphate, silicate, and nitrate means based on profile data from the World Ocean Database (WOD). https://www.ncei.noaa.gov/products/world-ocean-atlas
WaterIsotopes.org	Stable isotope hydrogeochemistry	The Waterisotopes Database compiles stable H- and O-isotopic data from a range of public domain and private sources, including the global network of precipitation-monitoring stations operated by the Global Network for Isotopes in Precipitation (GNIP). https://wateriso.utah.edu/waterisotopes/index.html
EBAS	Atmospheric chemistry	EBAS is a database hosting observation data of atmospheric chemical composition and physical properties. EBAS hosts data submitted by data originators in support of a number of national and international programs ranging from monitoring activities to research projects. EBAS is developed and operated by the Norwegian Institute for Air Research (NILU). https://ebas.nilu.no/
OnePetrology	Igneous geochemistry, geochronology	OnePetrology provides geochronology, isotope, and geochemistry data for igneous rock types from around the world, with especially high resolution in Asia. OnePetrology was developed and is maintained by the IUGS BIG Science Program Deep-time Digital Earth (DDE). https://petrology.deep-time.org/
IODP LIMS Database	Oceanography, sedimentary geochemistry	The International Ocean Discovery Program Laboratory Information Management System (IODP LIMS) database contains samples/data for IODP expeditions beginning in 2009 (Expeditions 317 and beyond) and any new samples/data generated from legacy expeditions (DSDP/ODP Legs 1-210 and IODP Phase 1 Expeditions 301-312). http://web.iodp.tamu.edu/LORE/
Sedimentary Geochemistry and Paleoenvironments Project (SGP)	Sedimentary geochemistry	The Sedimentary Geochemistry and Paleoenvironments Project (SGP) is a consortium that compiles analyses of the sedimentary geochemical record. https://sgp-search.io/

(*Continued*)

Table 2 (Continued)

	Sub-Discipline	*Scope of Data System*
Data Publication, Curation and Archiving: Data Repositories with Domain Expertise		
EarthChem Library (ECL)	General geochemistry	An open-access repository for geochemical datasets (analytical data, experimental data, synthesis databases) and other digital resources relevant to the field of geochemistry. The EarthChem Library offers data preservation and access, including long-term archiving and registration of data with Digital Object Identifiers (DOIs). https://www.earthchem.org/
Astromaterials Data Archive	General cosmochemistry	The Astromaterials Data System is NASAâ€™s designated archive for laboratory analytical data acquired on samples returned from space by NASA missions as well as meteorites in NASAâ€™s collections. Astromat provides trusted repository services for researchers to publish and archive astromaterials sample data in compliance with Open and FAIR Data policies of funding agencies and publishers. Astromat welcomes contributions of a broad range of data for extraterrestrial materials, including but not limited to, compositional data for samples of lunar rocks, meteorites, and other astromaterials and the minerals, melt and fluid inclusions, chondrules, and presolar grains that they contain. Astromat also accepts geochemical synthesis datasets; geochronological data; petrographic descriptions of samples; kinetic data from geochemical and petrological experiments. AstroMat is developed and operated at the Lamont- Doherty Earth Observatory of Columbia University and funded by NASA. https://repo.astromat.org/
GFZ Data Services / DIGIS Geochemical Data Repository for GEOROC	General Earth Sciences / igneous geochemistry	The GFZ German Research Centre for Geosciences assigns Digital Object Identifiers (DOI) to datasets. These datasets are archived by and published through GFZ Data Services and cover all geoscientific disciplines. They range from large dynamic datasets deriving from data intensive global monitoring networks with real-time data acquisition to the full suite of highly variable datasets collected by individual researchers or small teams. These highly variable data (â€˜long-tail dataâ€™) are small in size, but represent an important part of the total scientific output. The GEOROC data repository, hosted by GFZ Data Services as a dedicated data center, hosts research data within the scope of the GEOROC database: geochemical compositions of rocks, glasses, minerals and inclusions from all geological settings on Earth. The repository is curated by the Digital Geochemical Data Infrastructure (DIGIS) project at GÃ¶ttingen University. https://dataservices.gfz-potsdam.de/
PANGAEA	General Earth Sciences	Open-access library for archiving, publishing, and disseminating georeferenced data from the Earth, environmental, and biodiversity sciences. Originally evolving from a database for sediment cores, it is operated as a joint facility of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM) at the University of Bremen. PANGAEA holds a mandate from the World Meteorological Organization (WMO) and is accredited as a World Radiation Monitoring Center (WRMC). It was further accredited as a World Data Center by the International Council for Science (ICS) in 2001 and has been certified with the Core Trust Seal since 2019. The successful cooperation between PANGAEA and the publishing industry along with the correspondent technical implementation enables the cross-referencing of scientific publications and datasets archived as supplements to these publications. https://www.pangaea.de/
Multiscale TROPIcal CatchmentS (M-TROPICS)	Biogeochemistry, critical zone	The CZO Multiscale TROPIcal CatchmentS (M-TROPICS) provides the international scientific community with unique decennial time series of meteorological, hydrological, geochemical, and ecological variables in tropical environments. The CZO M-TROPICS involves academic and governmental partners in tropical countries (Cameroun, India, Lao PDR, and Vietnam) and is included in the Research Infrastructure OZCAR, the French contribution to the international CZO initiative. https://mtropics.obs-mip.fr/
DDE Data Publisher & Repository	General Earth Sciences	Data Publisher & Repository is a repository of geoscience data established with the support of Deep-time Digital Earth international big science program (DDE). https://repository.deep-time.org/
NOAA (National Oceanic and Atmospheric Administration, USA)	Atmospheric, environmental, oceanic chemistry	NOAA's National Centers for Environmental Information (NCEI) are responsible for hosting and providing public access to one of the most significant archives for environmental data on Earth with over 20 petabytes of comprehensive atmospheric, coastal, oceanic, and geophysical data. NCEI provides archive services for much of the data collected by NOAA scientists, observing systems, and research initiatives. We manage a comprehensive collection of environmental information from a broad range of time periods, observing systems, scientific disciplines, and geographic locations. Our world class stewardship services preserve data for future use on behalf of our science-user communities and the general public. https://www.ncei.noaa.gov/

Table 2 (Continued)

Sub-Discipline	Scope of Data System
Data Management in the Laboratory	
Sparrow	Data system for geochronology and geochemistry labs. Sparrow allows laboratories to organize analytical data and track project- and sample-level metadata. Its metadata management interface streamlines tasks such as controlling embargo, identifying and linking geologic and publication metadata, and generating aggregate summaries. https://sparrow-data.org/
Data Management in the Laboratory	
Sparrow	Data system for geochronology and geochemistry labs. Sparrow allows laboratories to organize analytical data and track project- and sample-level metadata. Its metadata management interface streamlines tasks such as controlling embargo, identifying and linking geologic and publication metadata, and generating aggregate summaries. https://sparrow-data.org/
AusGeochem	Cloud-hosted open geochemistry data platform designed to simultaneously act as a geosample registry, a geochemistry data repository and an active research tool. The primary objective of the platform is to provide sample and data management services to Australian geochemistry laboratories and their users. https://www.auscope.org.au/ausgeochem
Sample Analysis Micro-Information System (SAMIS)	Comprehensive data management system specifically designed for planetary science sample return missions (OSIRIS-REx).
Sample Management	
SESAR	Sample registration service and preservation of sample metadata allocating agent of the IGSN to allow unambiguous referencing of samples to the data and publications generated by their study, and to the people and institutions, who collect, curate, fund and study them; sample curation support, catalog search. https://www.geosamples.org/
AusGeochem	See above.

funding agency or from a specific, national research project. Note that development of many of these data systems is project- or researcher-driven, which means that coverage of different subdisciplines within geochemistry and their data types is neither systematic nor equal. Finally, this selection is in a large part based on the re3data search which, although it is the most complete index for research repositories to date, has a noticeable regional bias toward Europe and North America (i.e., the Global North; https://www.re3data.org/browse/by-country/).

Limitations of today's geochemistry data ecosystem

Geochemistry has come a long way since its birth in 1838, yet compared to other disciplines, there is still much ground to cover in order to fully achieve FAIR and digital geochemical data. Modern geochemical databases enable innovative science built on computational analysis of curated synthesis datasets; sample and data management systems provide researchers with the necessary support to consistently generate high-quality datasets; and domain repositories offer data and metadata curation as well as publication to ensure that datasets are FAIR and due credit can be given to researchers, laboratories/institutions, and funders. These capabilities have significantly improved the overall reproducibility of research results and the reusability of individual datasets, yet there is still ample room for improvement.

Culture of ownership and attribution

Academic databases are commonly built from personal or community projects. In contrast, most of the more comprehensive, long-lasting infrastructures are maintained by government agencies, such as national geological surveys, which can make them less accessible to the wider community. There are also examples of commercial systems, especially for laboratory data management systems, where software is often developed and sold by instrument manufacturers or affiliated companies. It is important to remember that any such data systems do not have ownership over the data, but only over the infrastructures they develop and maintain. Data ownership usually remains with the authors who first created or published them, for example, through research articles, in accordance with the licenses of publication. However, this fact is not always reflected in the way data from these systems are cited. To give proper attribution, authors reusing data compiled from any database should reference both the data system and all of the original sources that have contributed to the compilation dataset. In practice, this could mean tens to hundreds of additional citations and requires data systems to make bibliographic or provenance metadata easily accessible. Unfortunately, the digital age makes it very easy to copy and web scrape data from multiple sources online. In research publications, care is taken to cite previous ideas, but frequently metadata attributes that enable identification of ownership, facilitate attribution to funders, researchers, and institutions, and enable quantification of impact are not included in the copied data. Even though under permissive licenses this

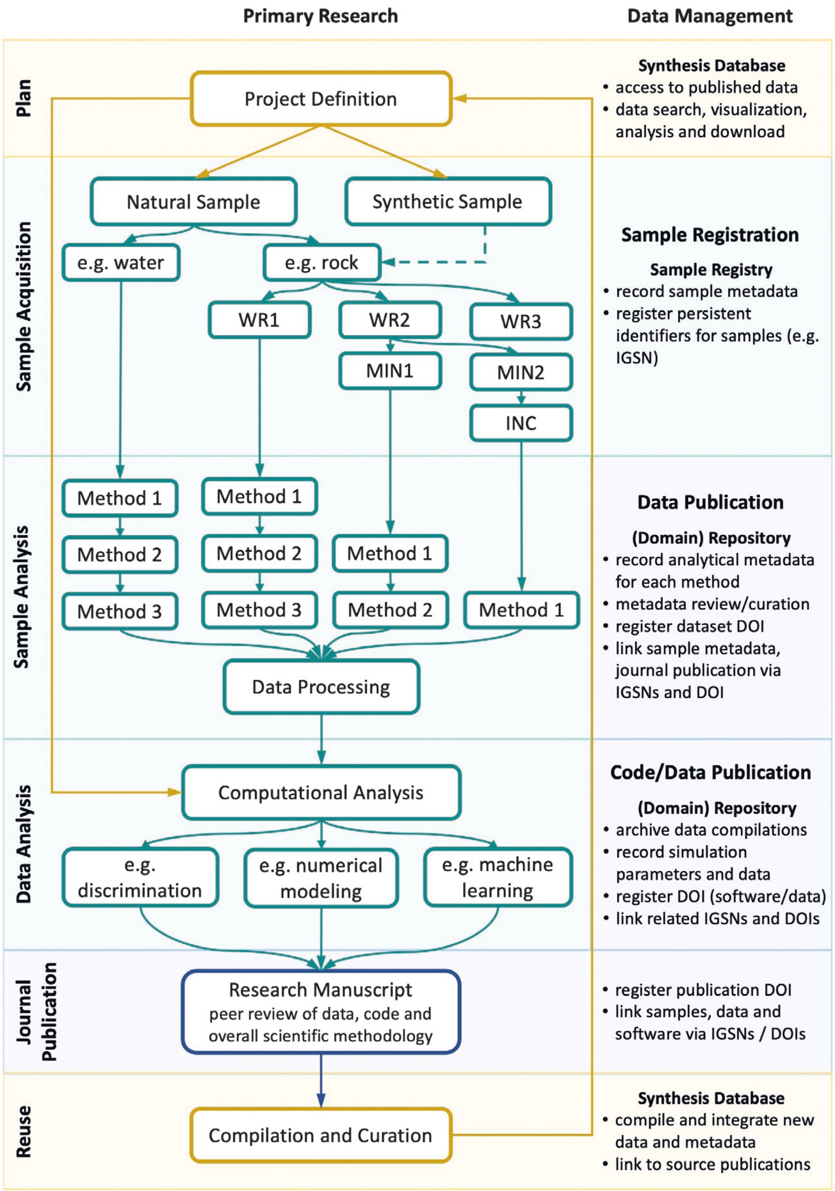

Primary Research
Data Management
Plan
Project Definition
Synthesis Database
• access to published data
• data search, visualization, analysis and download
Sample Acquisition
Natural Sample
Synthetic Sample
e.g. water
e.g. rock
WR1
WR2
WR3
MIN1
MIN2
INC
Sample Registration
Sample Registry
• record sample metadata
• register persistent identifiers for samples (e.g. IGSN)
Sample Analysis
Method 1
Method 2
Method 3
Method 1
Method 2
Method 3
Method 1
Method 2
Method 1
Data Processing
Data Publication
(Domain) Repository
• record analytical metadata for each method
• metadata review/curation
• register dataset DOI
• link sample metadata, journal publication via IGSNs and DOI
Data Analysis
Computational Analysis
e.g. discrimination
e.g. numerical modeling
e.g. machine learning
Code/Data Publication
(Domain) Repository
• archive data compilations
• record simulation parameters and data
• register DOI (software/data)
• link related IGSNs and DOIs
Journal Publication
Research Manuscript
peer review of data, code and overall scientific methodology
• register publication DOI
• link samples, data and software via IGSNs / DOIs
Reuse
Compilation and Curation
Synthesis Database
• compile and integrate new data and metadata
• link to source publications

Fig. 9 Schematic research project and data management workflow that summarizes data systems and services available during different research stages and highlights the complementarity of primary research and sample and data management. *INC*, inclusion; *MIN*, mineral; *WR*, whole rock. Sample registries, data repositories and synthesis databases have well-defined roles within this workflow. Laboratory data management systems may support sample analysis only or span the full range from sample registration to data publication.

may be allowable, this type of behavior raises issues about the ethics of data duplication and republication, and highlights the need for documentation of best practices for the identification of data aggregations, data republication, and mirroring of data to multiple sites to ensure that researchers, funders, repositories, and any person/institute involved are appropriately credited (Klump et al., 2021b).

Yet even if data sources are referenced and provided with every data download—as is the case with all data systems described in this chapter and indeed is a requirement of the FAIR principles—many journals are still unwilling to include overly long reference lists in the main article, forcing authors instead to include these data citations in a supplementary document that usually is not indexed by Scopus or GoogleScholar and is not automatically linked to the manuscript. In addition, authors are not always aware of these citation requirements, sometimes only opting to name the data system in their manuscript without any further references or indeed any documentation of how they subselected data from a database. This practice means that not only is there no due credit given to the authors who originally generated and published the data, but it is also not possible to reproduce these datasets to verify the scientific outcomes. There are many examples of this behavior in the literature citing the GEOROC and PetDB databases.

Another related example are citations to data processing software such as Glitter, which again usually only mention the software name but do not cite the original authors of the reference material values used within the software application.

To overcome this complexity of data citation, a working group of the Research Data Alliance (RDA) has proposed so-called reliquaries, through which a collection of aggregated individual datasets could be referenced with a single citation (Agarwal et al., 2021). All source datasets within a compilation would receive a citation, yet only this "data reliquary" would need to be included in the article reference list. Unfortunately, work is still ongoing to develop a scalable solution that enables credit for each individual element of this "reliquary," and a period of testing will be required before journals might implement this new technology.

In the meantime, the system of data citation is largely built on trust, and many researchers are only willing to openly share their data provided the roles that they play in the generation of the geochemical data are guaranteed to be recognized and attributed (i.e., formal acknowledgment as authors/contributors, for example, sample collection, generation of the analytical data in the laboratory, publication of the data). The lack of proper attribution currently still causes widespread skepticism toward the value of data publication. As highlighted in Chamberlain et al. (2021) and Klöcking et al. (2023), geochemistry is still broadly lacking a culture of data sharing, in stark contrast to other scientific disciplines where such practices are well established. Instead, fear and distrust are still encouraging a culture of "data my'ning" and the relatively high cost of appropriate data management compared to the small data volumes further discourages adoption of FAIR data practices (Stuart et al., 2018; Tedersoo et al., 2021; Digital Science et al., 2021).

Lack of standards and community-endorsed best practices

Geochemical data make up the "long tail" of Earth Science data, since due to the intricacy and wide variety of analytical procedures they are low-volume but highly variable (Heidorn, 2008). As a consequence, even if data are shared, often they are not machine-readable and of variable quality. They are also very fragmented, with each subdiscipline or community currently determining their own data practices and customs.

Many controlled vocabularies are available as appendices in papers or shared as PDFs but few are FAIR and able to be interrogated online and more importantly, reused by others for their own datasets, thus enhancing interoperability. As the size and diversity of the communities wanting to discover and reuse geochemical data grow, there is increasing demand for more sophisticated data models and ontologies to increase the understanding of the terms used and to enhance automated machine-to-machine access and analysis of multiple geochemical datasets. There is now an urgency for the geochemistry community to work together to modernize much of the existing wealth of knowledge on geochemical terms and vocabularies into globally shared online resources as well as better structure these into machine readable data models and ontologies suitable for access on the web (e.g., GeoSciML, EarthResourceML, CGI Vocabularies). Cox et al. (2021) provide guidelines on how to convert legacy vocabularies that are currently in text-rich, unstructured documents and PDFs into machine-readable semantic resources, which can then be used for unambiguous data annotation, which in turn increases data interoperability and enables online real-time data integration. Any vocabularies that support widely used data systems should be made available in online vocabulary services and registers such as the International Union of Geological Sciences (IUGS) Commission for the Management and Application of Geoscience Information (CGI) Vocabulary Register, Research Vocabularies Australia, and the National Environmental Research Council Vocabulary Server.

While willingness to share data or developing and adhering to community standards require a culture change, limitations of the technical infrastructure within the data ecosystem are also holding the community back. Most of the current data systems lack tools for automation, making the process of preparing, contributing, or accessing data laborious and highly system-dependent. Since many of these data systems are tied to specific projects or subdisciplinary and/or national interests, the data landscape itself is also fragmented and characterized by a general lack of global awareness, communication, and cooperation. Data themselves are often not published in a machine-readable format, which together with a dearth of consistent technical standards hinders the integration and interoperability of data between distributed data systems and further contributes to the fragmentation of the geochemical data landscape. This fragmentation also has measurable financial repercussions: the European Commission has estimated the annual direct and indirect costs of managing nonstandardized research data at EUR 10.2bn and EUR 16bn per year, respectively (European Commission, 2018).

Legacy data vs. new data types

The fundamental power of geochemical synthesis databases lies in the aggregation of data acquired over a large span of years by different laboratories. However, this aggregation also poses the great challenge of how to meaningfully compare data collected using very different methodologies and standards. Many of the legacy datasets contained in geochemical databases have relatively sparse metadata that make them easy to dismiss by more modern standards and QA/QC criteria. When analyzing large compilation datasets, in particular, even a single missing analytical value could lead to the removal of an entire legacy dataset. In addition, even canonical values used for data processing and normalization may change, and if these corrections are nonlinear or the reference values used are not preserved, it may be impossible to recalculate previous data for comparison with more modern data (e.g., Giuliani et al., 2024). If there is sufficient sample left, it is often easier to reanalyze the sample, but this is not always an option.

At the same time, geochemical methods are constantly evolving and demands for new data types can arise. For example, the recent advances in analytical instrumentation, that can quantitatively measure elemental concentrations and isotopic compositions of samples in situ across large areas or at increasingly higher spatial resolution, have seen a proliferation of 2D and 3D imagery to be stored as primary analytical data. Many current data systems are not equipped to host large volumes of imagery or to provide users with tools to work with these images. The combination of spatial imaging with precise point analysis or textural data requires new infrastructure developments as well as the development of community standards to ensure that these new data types and associated software tools are preserved in FAIR formats (Einsle et al., 2023). Beyond data storage, geochemistry might also need to adopt more advanced computational techniques such as parallel computing on high-performance computing (HPC) infrastructures in order to cope with the analysis and visualization of data volumes that are already in the region of terabytes and approaching the "big data" realm (e.g., Prodanović et al., 2023).

Finally, digital storage is also constantly evolving and digital media and software solutions can quickly become unstable. Standards and vocabularies are constantly being extended to cover more diverse use cases. Databases and repositories need to invest resources into maintaining and upgrading their systems to ensure continuous data access and availability of other services, which they often have little to no funding for. Many funding agencies have yet to recognize this growing dilemma and several valuable digital collections have been lost simply due to the deterioration of the media that they are stored on and/or the deprecation of the tools required to read the data. Data rescue efforts are becoming increasingly common to upgrade and preserve data not only from deteriorating storage formats or media, including paper records, but also from electronic data tables trapped behind a paywall without perspective of maintenance as the respective journal is moving toward adopting open science practices.

Lack of funding and sustainability

Many data systems, especially disciplinary ones that only serve the geochemical community, are developed and maintained out of fixed-term grants or contracts. Often, these grants are provided from national science funders and are therefore actively competing with research funding. Such lack of sustainability is one of the largest threats to "big data"-research in geochemistry today, and there are many examples of programmatic databases appearing and disappearing over the years (e.g., MetPetDB, NAVDAT). Currently, the most stable data services with long-term funding are those operated by scientific agencies or commercial enterprises. The willingness for scientific innovation and implementation of new technologies is then dependent on the specific policies of these agencies or companies and could be decoupled from cutting-edge research needs.

Insufficient funding and support, coupled with a multiplication of external requirements following the adoption of the FAIR, TRUST, and CARE principles, jeopardize the survival of domain-specific data systems. Yet many of these systems are too valuable for the geochemical community to lose. As discussed in Section "Data curation, preservation, accessibility, and reuse," the review of data and metadata provided by domain repositories significantly increases the quality and completeness of published geochemical datasets. The value of global synthesis databases is that they preserve and harmonize decades of data ensuring continued ROI for millions of dollars in analytical and personnel costs. A simplistic back-of-the-envelope calculation for the 36,973,060 data values currently compiled in the GEOROC database, assuming a present-day average analytical cost in the range of US$5–30 per data point, returns a total value of US$180 million to US$1.1 billion. Similar calculations can be made for IEDA, the data facility that hosts PetDB, EarthChem and SESAR, whose total costs are estimated to represent less than 5% of the annual national research budget for the science programs IEDA serves. These numbers are likely to be a significant underestimation of the actual cost of acquiring these data, as analytical costs would have been far more expensive in the past and the data also represent thousands of work hours for research and technical staff over the decades. When taking into account that the GEOROC data holdings are just a small subsample of all geochemical data, the true value of the data compiled in geochemical data systems globally is staggering and underlines the overall contribution these systems make to the geochemical research community. In contrast, the maintenance costs for these data systems are only a fraction of a percent of their accumulated value. The current precarious funding situation of many geochemical data systems, therefore, puts the whole research community at risk of a devastating material loss of knowledge.

Global north vs. global south

Global North and Global South are terms that are increasingly being used to group countries based on characteristics around socioeconomics and politics. The United Nations Conference on Trade and Development (UNCTAD) groups the United States, Canada, Europe, Israel, Japan, South Korea, Australia, and New Zealand into the Global North, while the Global South comprises Africa, Latin America and the Caribbean, Asia (excluding Israel, Japan, and South Korea), and Oceania [excluding Australia and New Zealand; UNCTAD, Statistical Portal, Classifications (https://unctadstat.unctad.org/EN/Classifications.html].

Like many other disciplines, geochemistry, and especially the geochemical data landscape, suffers from a dominance of the Global North. Although the CARE principles exclusively apply to Indigenous data, they raise awareness of related issues for researchers in the Global South. There are many examples of parachute science by research groups from the Global North without a single local coauthor or even the acknowledgment of support by researchers from the Global South. Further, many datasets on samples from the Global South reside only in data systems in the Global North and can be behind paywalls, inaccessible to those countries that the original samples were taken from (e.g., Goodenough and Mills, 2021). A simple search for the keyword "geochemistry" reveals that all of the datasets currently indexed in Google's Dataset Search are hosted by data providers in the Global North. A similar picture appears in the provenance statistics of re3data (see Section "The ecosystem of geochemical data management"). The most common approach to addressing this disparity focuses on increasing access to data hosted in the Global North, and indeed to increase the (re)publication of Global South data in data systems of the North. Largely driven by a lack of funding, the struggle to train and retain experts in the Global South creates a continued reliance on the Global North Reidpath and Allotey (2019). Language is often an additional barrier to knowledge transfer and dissemination between North and South (Arenas-Castro et al., 2024).

This paper is inevitably biased toward data systems and data management workflows in the Global North, although an effort has been made to reference this bias and to include relevant developments from the Global South. Importantly, the IUGS is increasingly including researchers from the Global South, for example, through the Deep-time Digital Earth (DDE) program and the Global

Geochemical Baselines commission. CODATA are also fostering active participation of and collaboration with the Global South. Nonetheless, the geochemical data landscape will likely remain biased for some time.

Future opportunities for digital geochemical data

The future of geochemical data is digital. With increased capacity of electronic data storage and data sharing infrastructure, it is now technically possible to create global networks of geochemical data from multiple data sources and institutions. Such global networks are a requirement for harnessing the wealth of geochemical data through modern computational tools, which would revolutionize the impact of geochemical data on societal issues of global significance. To name just one example, geochemical data are relevant to many of the United Nations Sustainable Development Goals such as ′6: Clean Water and Sanitation′ or ′13: Climate Action′ (e.g., Bundschuh et al., 2017; Gill, 2017; Alexakis, 2021; Wyborn and Lehnert, 2021). However, the ability to combine data from multiple sources is limited by lack of standardization of geochemical data, including vocabularies, ontologies, and agreed minimum variables for each type of data and each analytical method. To automate data sharing in the future, and close any gaps in coverage, while also maintaining the high quality and detail required for individual geochemical datasets, there is an urgent need for a culture change within the geochemical community, further technical advances, and the development of community-agreed digital standards. With these additional developments, and if sustainability of data is guaranteed, global democratization of geochemical data will be possible.

Capabilities and challenges of computational geochemistry

The rise of machine learning and other advanced statistical and computational approaches in geochemistry presents enormous opportunities. The number of publications employing machine learning tools has increased dramatically since 2017 and large datasets derived from synthesis databases are a primary data source for these studies (Petrelli, 2024). This trend will only continue to grow and to fully harness this potential, it is now more important than ever that databases provide their data and metadata in fully machine-actionable, analysis-ready format.

Despite the numerous innovations and potential new discoveries afforded by the application of machine learning techniques and artificial intelligence to geochemical data, these new methods also come with both technical and ethical challenges. "Data cleaning" is an important step in any machine learning workflow on real datasets; however, it might cause a significant reduction of the input dataset as legacy or poorly documented data may be rejected due to missing data or metadata. In response, imputation of certain missing (meta)data is mathematically possible, but often not a viable approach in a rigorous scientific setup. For geochemistry in particular, such modified datasets might even be detrimental as it is often the anomalies that yield most insight. Nonetheless, there is also evidence that while high-resolution images are more time-consuming and expensive to produce, they do not necessarily yield better results than lower-resolution data (Nathwani et al., 2023). These results are encouraging as they could help to optimize both storage capacity and analytical expenses during the current explosion of microanalytical data. Yet the question remains of where to draw the line between real and "invented" data, and as a community, we will need to weigh up benefits such as the saved cost of lower-resolution data against the reliability and ethics of machine learning and artificial intelligence.

As numerical applications in geochemistry become more complex, the importance of FAIR geochemical software will only increase and it might be necessary to revisit the challenge of storing model or simulation data. While computational approaches such as machine learning are very powerful, they can also be expensive to run and might eventually require databases of simulation data and/or the storage of such model data alongside primary geochemical data.

Toward a culture of open and FAIR data sharing

A number of examples exist in other scientific disciplines of global networks to openly share data from multiple institutions. Klöcking et al. (2023) cite several of these efforts in seismology, climate science, crystallography, pure and applied chemistry, and biodiversity. In crystallography, for example, authors are expected to archive their data in a computer-readable, standard format within domain repositories, such as the Worldwide Protein Data Bank (wwPDB), and formal data validation is part of the peer review process for journals (Spek, 2020). In seismology, the International Federation of Digital Seismograph Networks (FDSN) was established as early as 1984 to develop and promulgate a universal standard for the distribution of broadband waveform data and related parametric information (Dziewonski, 1994).

The need for a similar shift in data culture in geochemistry has been highlighted by Chamberlain et al. (2021). Although the Open Science movement has reached geochemistry and first changes are being instigated, there are a number of factors that are holding the community back. Chief among those is the fact that proper data management requires significant commitment of time and effort, and unfortunately there is still little recognition of these investments in the classic academic track record (Piwowar et al., 2007; Kim and Stanton, 2012; Klöcking et al., 2023). Fear of being scooped as well as wanting to uphold tradition will continue to outweigh the benefits of open data sharing until such data (and software) publications are recognized and rewarded by academic hiring and grant review panels. The recent amendment of the German Research Foundation (DFG) policies to recognize article preprints, datasets, or software packages as research outcomes is an important step toward more balanced research assessment

(DFG, 2022). Being able to appropriately give credit through compound citations and data reliquaries will further contribute to the culture change (Agarwal et al., 2021).

However, many data management tools are difficult to access beyond the Global North, meaning that widespread adoption is currently restricted to the more affluent nations. An important and urgent goal for geochemistry as well as for other disciplines is to work toward global "democratization" of data. Initially, increasing access and participation for the Global South may involve the development of data management workflows that are user-friendly to nonexperts and do not require expensive infrastructure such as a high-speed Internet connection. In the long term, however, it is paramount to support and foster the establishment of local data systems and work toward interoperability of a globally distributed geochemical data network.

Technical advances that enhance data sharing

Increasingly, data are born digital and connected to community-agreed protocols. With the rising roll-out of smart field notebooks/ apps, electronic laboratory notebooks, and custom-made data processing software for a variety of widely used analytical instruments, data are increasingly being born digital, that is, captured directly from acquisition. Automating the process of data recording straight from the instrument means that all instrument settings and ancillary parameters are recorded in conjunction with the primary geochemical data and stored in standardized protocols. Many instrument manufacturers are actively contributing by following community-agreed protocols where these exist and making much of their software and workflows open. This development reduces human error in data and metadata transcription and effortlessly leads to greater standardization of geochemical data measured by the same analytical method. However, yet again most of these developments are currently only accessible and/or viable for researchers in the Global North.

Technical standards enhance machine accessibility. Many of the technical standards described in Section "Standards and best practices for geochemical data" have the potential to streamline and improve geochemical data management. As they mature and are adopted by geochemical data systems, these standards will ultimately contribute to making data management easier for geochemists. These technical standards will be mostly invisible from the research community but constitute a valuable contribution to FAIR geochemical data. FAIR implementation profiles (FIPs; Schultes et al., 2020) offer a comprehensive summary of all standards adopted by a particular data infrastructure or community and could further accelerate convergence and interoperability between distinct data systems (Prent et al., 2024).

The increase of data publications in either (domain) repositories or dedicated data journals, such as Scientific Data (Nature) or Earth System Sciences Data (Copernicus), means that it is now possible to publish all metadata and raw data associated with a dataset. In turn, this means that any analytical data will become increasingly well documented, for example, enabling better and more objective comparison of data from different laboratories or quantitative comparative studies using large data compilations from multiple sources. In addition, a new platform dedicated to method publications has recently been proposed (Profeta et al., 2022). Such a method directory would allow laboratories to publish their workflows and methodologies, with any changes documented through version control. Researchers would then only need to cite the appropriate methodologies in scientific articles, allowing authors to focus on the discussion of scientific results while also ensuring that all information required for reproducibility and reusability of analytical data are appropriately and consistently documented.

Community-agreed scientific standards for geochemical data

The concept of scientific standards, especially the importance of standardized metadata, is described in detail in Sections "The ecosystem of geochemical data management" and "Standards and best practices for geochemical data." Unfortunately, as noted above, most standards available for geochemistry to date are mainly presented as best practice descriptions in journal or technical articles, predominantly published as PDFs. There is hence considerable effort required to modernize these documents and make all geochemical standards available as online machine-actionable resources. The recently formed CODATA OneGeochemistry Working Group is seeking to facilitate this process by harnessing and harmonizing existing groups working toward global data sharing and promulgating best practices and standards. OneGeochemistry is following initiatives such as the Digital Standards Initiative of the International Union of Pure and Applied Chemistry (IUPAC) or the GeoSciML initiative of the International Union of Geological Sciences (IUGS). As well as translating existing standards and best practices into machine-actionable resources, OneGeochemistry's mission is to facilitate community participation, discussion, and consensus to enhance best practices for all geochemical data types and converge on global standards (Prent et al., 2023). Scientific standards across all geochemical data types would allow databases to become more coordinated, by implementing common data schemas, formats, and vocabularies to make data easier to find, access, and reuse. Yet it is not just standardization within the geochemical community that matters today; it is also important that any standards developed can be interoperable with those from many other scientific datasets and beyond to social science data and other datasets. The EU-Horizon-funded, CODATA-led project *'WorldFAIR: Global cooperation on FAIR data policy and practice'* is currently working with a set of 11 disciplinary and cross-disciplinary case studies to advance implementation of the FAIR principles and, in particular, to improve the interoperability and reusability of digital research objects, including data. One of these case studies is Geochemistry (Hodson, 2024).

Ensuring longevity: Sustainability of geochemical data

Many of the databases described above are project-based and do not have sustained funding. Yet, persistent resources are required to support consistent and long-term data management, and the maintenance of databases, repositories, and higher level research infrastructures. AuScope, EPOS, and NFDI are three examples of national or transregional data infrastructure projects that are receiving long-term government funding on decadal timescales. Yet even these large-scale initiatives are not guaranteed to be continued in perpetuity. Databases such as PetDB and GEOROC have so far survived for 25 years; however, their funding remains precarious and reliant on national funding agencies and the goodwill of their host institutions. For domain repositories, in particular, long-term staffing solutions are vital to provide consistent data curation and assist researchers with data submissions. System maintenance and keeping up to date with technical developments further requires consistent support by software developers and engineers.

Although it has long been shown that the benefits of open data infrastructure far outweigh the costs of building and maintaining this infrastructure (e.g., Ball et al., 2004), data infrastructure remains undervalued compared to cutting-edge research and thus most data systems struggle for long-term survival. This leads to an environment where data systems are competing with geochemical researchers for funding through short-term projects. Funding is usually restricted to the development of new technologies or services, while system maintenance and curation are neglected. As a consequence, many of the temporary research grants granted to develop project-based databases or other systems that are independent of the main, established infrastructure are discontinued after only a few years and quickly orphaned. Thus, the splitting of the limited funds available to national research foundations inevitably leads to an enormous waste of resources. Furthermore, many of the most successful data infrastructures transcend national borders, catering to the global scientific demand. Unfortunately, such internationality further limits the funding options of such systems, requiring consortial structures and multilateral agreements between national funding agencies or institutions. Competition is a waste of resources and what is required instead is a global network jointly managed and supported through diverse national contributions. Successful examples of such networks already exist in other disciplines, such as the Global Biodiversity Information Facility (GBIF), the International Federation of Digital Seismograph Networks (FDSN), and the IUGG Global Heat Flow Database.

The advantages of a well-funded, sustainable data infrastructure are extensive. As shown in Section "Lack of funding and sustainability," the current value of analytical data compiled global data systems amounts to billions of USD. The additional cost to maintain and increase this value through active curation of the database is negligible in comparison; however, it does require long-term stability of funding. Not only are there measurable financial benefits as well-maintained, centralized infrastructures save resources (Ball et al., 2004). Infrastructure development would also be able to focus on those areas that are of most assistance to the community: maintaining and improving well-used services; developing new tools for data processing, analysis, integration, etc.; as well as expanding user support, investing in training materials and other educational resources. Development of interoperability with other infrastructures could be prioritized, for example, by building stable, guided workflows to connect repositories with journal publication platforms and synthesis databases with modeling software.

A question of authority: Who needs to act?

Funders and publishers are increasingly mandating appropriate data management, including data management plans in proposals and the requirement for publication of data in repositories rather than as appendices of journal articles. However, these mandates are not always consistently enforced as, for example, few funding agencies retrospectively verify if data management plans have been followed. There is also no formal (financial) support for data management, with institutions and data systems having to step up to support researchers using their own resources. The open research data (ORD) pilot program by the European Commission, however, is one example of enforceable open access policies being extended to research data (EU Open Access Strategy, n.d.). Since 2017, all projects receiving Horizon 2020 funding have been required to not only publish open access scientific publications but also to make any research data open by default, with a limited set of exceptions, for example, to protect privacy and ethical considerations.

What is more, different funding/publishing bodies often give conflicting advice and there are as yet no clear universal guidelines for best practice geochemical data management. As a consequence, the AGU has observed a marked increase in submissions to generalist repositories, rather than domain repositories, as authors seek out the easiest and fastest route to "ticking the box" (Vrouwenvelder and Stall, 2023). Clearly, more dedicated guidance and discipline-specific support are required to encourage data deposition in domain repositories, and support workflows of coordinated submission of manuscript and data.

Data citation also remains a challenge. Citation limits in the main article, originating from printed journal issues but often still enforced for online articles, mean that data citations are often only included in supplementary documents. This is especially the case for studies relying on large compilation datasets. The data reliquaries under development by the RDA "Complex Citations" Working Group will be one possible solution to this problem (Agarwal et al., 2021). Another, and simpler one, is for journals to ensure that any citations included in supplementary documents are also indexed so that due credit is given (Stall et al., 2023). In the meantime, better education and training of reviewers and editors would ensure greater consistency across journals and better support of authors.

Seismology and crystallography have both benefited from the dominance/authority of a single scientific union (the International Union of Geodesy and Geophysics, IUGG, and International Union of Crystallography, IUCr, respectively; see examples above). In stark contrast, and as shown by Fig. 10, geochemistry and geochemical data are relevant to 5 distinct science unions and

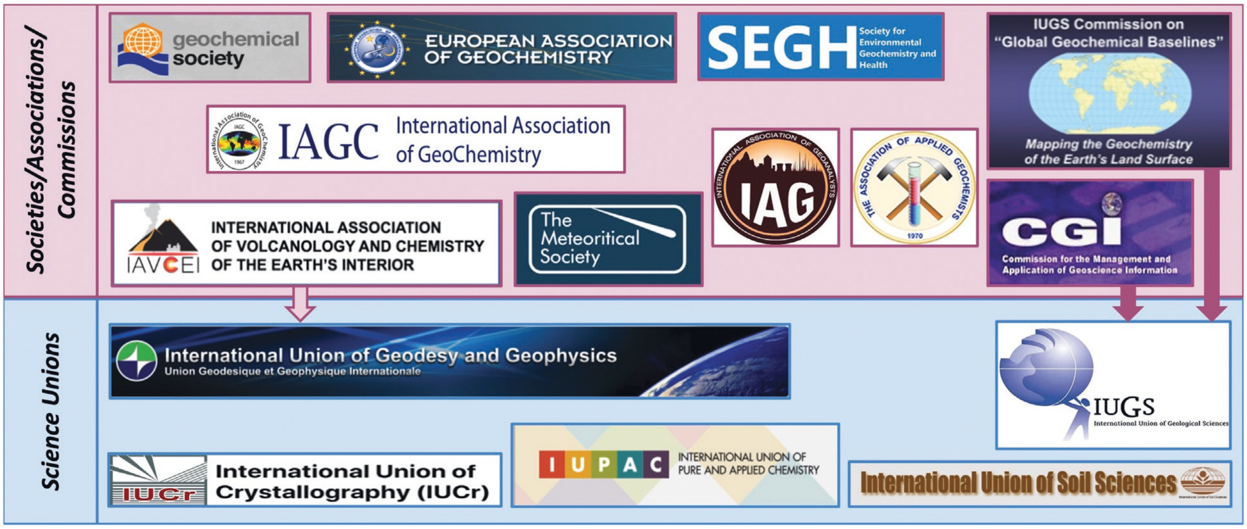

Societies/Associations/ Commissions
geochemical society
EUROPEAN ASSOCIATION OF GEOCHEMISTRY
SEGH Society for Environmental Geochemistry and Health
IUGS Commission on "Global Geochemical Baselines"
Mapping the Geochemistry of the Earth's Land Surface
IAGC International Association of GeoChemistry
IAG International Association of Geoanalysts
The Association of Applied Geochemists 1970
INTERNATIONAL ASSOCIATION OF VOLCANOLOGY AND CHEMISTRY OF THE EARTH'S INTERIOR
IAVCEI
The Meteoritical Society
CGI Commission for the Management and Application of Geoscience Information
Science Unions
International Union of Geodesy and Geophysics
Union Geodesique et Geophysique Internationale
IUGS International Union of Geological Sciences
IUCr International Union of Crystallography (IUCr)
IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY
International Union of Soil Sciences

Fig. 10 The 15 international societies/associations/commissions (*pink*) and science unions (*blue*) that include geochemical data within their remit. *Dark arrows* indicate formal Commissions of the International Union of Geological Sciences (IUGS); the *light arrow* indicates an association of the International Union of Geodesy and Geophysics (IUGG). The other societies and associations might be affiliated with a union but not controlled by that union. The OneGeochemistry initiative is currently endorsed by the Geochemical Society, the European Association of Geochemistry, the Association of Applied Geochemists, the International Association of Geoanalysts, the Meteoritical Society, and the IUGS Commission on Global Geochemical Baselines.

at least 10 societies, associations, or commissions. This abundance of different and independent official bodies constitutes a considerable obstacle to endorsing, promoting, and enforcing global standards for geochemical data. There is no single guideline that can be applied to all geochemists equally. The OneGeochemistry initiative is seeking to overcome this dilemma by operating across, and thus providing a platform to unite, all geochemical societies, associations, and unions (Prent et al., 2023).

Finally, academic institutions also have an important part to play. As more resources on Open Science and FAIR data are being made available, these materials should be included in the curriculum. Data management is becoming an essential part of the academic publishing culture, and therefore teaching these skills is a crucial part of career development for graduate students, academic staff, and faculty. Equally, engagement in FAIR data practices should be rewarded and considered in assessing candidates during hiring procedures.

Conclusion: Toward global democratization of geochemical data

From its early beginnings in the 19th century to the modern digital age of open science and artificial intelligence, geochemistry has come a long way and the approach to geochemical data has evolved accordingly. We now find ourselves at a crossroads where new, digital techniques are beginning to take hold—and have, indeed, already revolutionized other disciplines—while common data management practices are still largely governed by historical limitations. In the age of the microanalytical revolution, geochemical data volumes are increasing exponentially and both the research and the data communities need to adapt to accommodate this growing volume of information as well as new data types. Journals and funders are now requiring the formal publication of datasets in repositories. Increasing applications of machine learning techniques to large geochemical data compilations highlight the enormous value added by the curation and harmonization efforts undertaken by domain data systems. The FAIR, TRUST, and CARE principles are starting to become operationalized for geochemical data. Although there is a lack of machine-actionable, community-agreed, and internationally endorsed standards for geochemical data, many best practice guidelines and vocabulary recommendations have been published in recent years. There is now an urgent need for the international geochemistry community to come together and agree, formalize, and maintain those standards that are fundamental to the interoperability and machine-readability of FAIR geochemical data.

From the first data table to relational and graph databases, the technology now exists to make a Global Network of Geochemical Data a reality. All that is required is sustainable funding and a community choice to embrace the technical advances already made, continue efforts of consensus and standardization, and keep pushing the boundaries for a more globally connected science. Nevertheless, a commitment to global management of data also requires efforts of well-funded communities to support those of low- to middle-income countries. There is no place for parasitic science in this modern world of open data, and all integration efforts should be guided not only by the FAIR but also by the TRUST and CARE principles.

Acknowledgments

M.K. was supported by the German Research Foundation (DFG; grant numbers 437919684 and 503863705). K.L. acknowledges funding from US NSF Award Number 2148939 and NASA Grant Number 80NSSC19K1102. Detailed discussions and informal review by A. Lanati significantly improved this article. We are grateful to Brenhin Keller for a very constructive review and to A. Sturm and Z. Eriksen for additional feedback. We thank A. Shahar for expert editorial handling.

References

Abbott P, Bonadonna C, Bursik M, Cashman K, Davies S, Jensen B, Kuehn S, Kurbatov A, Lane C, Plunkett G, Smith V, Thomlinson E, Thordarsson T, Walker JD, and Wallace K (2022) *Community Established Best Practice Recommendations for Tephra Studies—-From Collection through Analysis*. Zenodo. https://doi.org/10.5281/ZENODO.3866266.

Adame F (2021) Meaningful collaborations can end 'helicopter research'. *Nature*. https://doi.org/10.1038/d41586-021-01795-1.

Agarwal DA, Damerow J, Varadharajan C, Christianson DS, Pastorello GZ, Cheah Y-W, and Ramakrishnan L (2021) Balancing the needs of consumers and producers for scientific data collections. *Ecological Informatics* 62: 101251. https://doi.org/10.1016/j.ecoinf.2021.101251.

Alexakis DE (2021) Linking DPSIR model and water quality indices to achieve sustainable development goals in groundwater resources. *Hydrology* 8(2): 90. https://doi.org/10.3390/hydrology8020090.

ARDC. (2020) *FAIR Data Guidelines for Project Outputs*. Zenodo. https://doi.org/10.5281/ZENODO.6559007. https://zenodo.org/record/6559007.

Arenas-Castro H, Berdejo-Espinola V, Chowdhury S, Rodríguez-Contreras A, James ARM, Raja NB, Dunne EM, Bertolino S, Emidio NB, Derez CM, Drobniak SM, Fulton GR, Henao-Diaz LF, Kaur A, Kim CJS, Lagisz M, Medina I, Mikula P, Narayan VP, O'Bryan CJ, Oh RRY, Ovsyanikova E, Pérez-Hämmerle K-V, Pottier P, Powers JS, Rodriguez-Acevedo AJ, Rozak AH, Sena PHA, Sockhill NJ, Tedesco AM, Tiapa-Blanco F, Tsai J-S, Villarreal-Rosas J, Wadgymar SM, Yamamichi M, and Amano T (2024) Academic publishing requires linguistically inclusive policies. *Proceedings: Biological Sciences* 291(2018): 20232840.

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, and McGillivray B (2020) The citation advantage of linking publications to research data. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0230416.

Commons Australian Research Data (2019) *ARDC Research Data Rights Management Guide*. Zenodo. https://doi.org/10.5281/ZENODO.5091580. https://zenodo.org/record/5091580.

Badia A (2020) The Data Life Cycle. pp. 1–29. Springer International Publishing. https://doi.org/10.1007/978-3-030-57592-2_1.

Baird D (1993) Analytical chemistry and the 'big' scientific instrumentation revolution. *Annals of Science* 50(3): 267–290. https://doi.org/10.1080/00033799300200221.

Ball CA, Sherlock G, and Brazma A (2004) Funding high-throughput data sharing. *Nature Biotechnology* 22(9): 1179–1183. https://doi.org/10.1038/nbt0904-1179. PMID: 15340487.

Bennett C, Haenecour P, Crombie K, Fitzgibbon M, Ferro A, Hammond D, McDonough E, Westermann M, Barnes J, Connolly H, and Lauretta D (2022) SAMIS: The OSIRIS-REx sample analysis micro-information system. In: *Goldschmidt2022 abstracts*. European Association of Geochemistry.

Boone SC, Dalton H, Prent A, Kohlmann F, Theile M, Gréau Y, Florin G, Noble W, Hodgekiss S-A, Ware B, Phillips D, Kohn B, O'Reilly S, Gleadow A, McInnes B, and Rawling T (2022) AusGeochem: An open platform for geochemical data preservation, dissemination and synthesis. *Geostandards and Geoanalytical Research* 46(2): 245–259. https://doi.org/10.1111/ggr.12419.

Borg G (2020) On "the application of science to science itself:" Chemistry, instruments, and the scientific labor process. *Studies in History and Philosophy of Science Part A* 79: 41–56. https://doi.org/10.1016/j.shpsa.2019.05.008.

Brantley SL, Wen T, Agarwal DA, Catalano JG, Schroeder PA, Lehnert K, Varadharajan C, Pett-Ridge J, Engle M, Castronova AM, Hooper RP, Ma X, Jin L, McHenry K, Aronson E, Shaughnessy AR, Derry LA, Richardson J, Bales J, and Pierce EM (2021) The future low-temperature geochemical data-scape as envisioned by the U.S. Geochemical Community. *Computers & Geosciences* 157: 104933. https://doi.org/10.1016/j.cageo.2021.104933.

Brümmer M, Baron C, Ermilov I, Freudenberg M, Kontokostas D, and Hellmann S (2014) DataID: Towards semantically rich metadata for complex datasets. In: *Proceedings of the 10th International Conference on Semantic Systems, SEM '14*. ACM.

Buck M and Hamilton C (2011) The Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. *Review of European Community and International Environmental Law* 20(1): 47–61.

Bundschuh J, Maity JP, Mushtaq S, Vithanage M, Seneweera S, Schneider J, Bhattacharya P, Khan NI, Hamawand I, Guilherme LRG, Reardon-Smith K, Parvez F, Morales-Simfors N, Ghaze S, Pudmenzky C, Kouadio L, and Chen C-Y (2017) Medical geology in the framework of the sustainable development goals. *Science of The Total Environment* 581–582: 87–104. https://doi.org/10.1016/j.scitotenv.2016.11.208.

Candela L, Castelli D, Manghi P, and Tani A (2015) Data journals: A survey. *Journal of the Association for Information Science and Technology* 66(9): 1747–1762.

Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker JD, Anderson J, and Hudson M (2020) The CARE principles for indigenous data governance. *Data Science Journal* 19. https://doi.org/10.5334/dsj-2020-043.

Carroll SR, Herczog E, Hudson M, Russell K, and Stall S (2021) Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data* 8(1). https://doi.org/10.1038/s41597-021-00892-0.

Chamberlain KJ, Lehnert KA, McIntosh IM, Morgan DJ, and Wörner G (2021) Time to change the data culture in geochemistry. *Nature Reviews Earth & Environment* 2(11): 737–739. https://doi.org/10.1038/s43017-021-00237-w.

Class C and Lehnert K (2012) *PetDB Expert MORB (Mid-Ocean Ridge Basalt) Compilation*. https://doi.org/10.1594/IEDA/100060. https://ecl.earthchem.org/view.php?id=274.

CODATA RDM Terminology Working Group. (2024) *CODATA RDM Terminology (2023 version): Overview*.

Codd EF (1970) A relational model of data for large shared data banks. *Communications of the ACM* 13(6): 377–387. https://doi.org/10.1145/362384.362685.

Consultative Committee for Space Data Systems. (2012) *Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-M-2. Also published as ISO 14721:2012*. https://public.ccsds.org/Pubs/650x0m2.pdf.

CoreTrustSeal Standards and Certification Board. (2022) *CoreTrustSeal requirements 2023-2025*.

Courtney Mustaphi CJ, Brahney J, Aquino-López MA, Goring S, Orton K, Noronha A, Czaplewski J, Asena Q, Paton S, and Brushworth JP (2019) Guidelines for reporting and archiving 210Pb sediment chronologies to improve fidelity and extend data lifecycle. *Quaternary Geochronology* 52: 77–87. https://doi.org/10.1016/j.quageo.2019.04.003.

Cox SJD, Gonzalez-Beltran AN, Magagna B, and Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. *PLOS Computational Biology* 17(6): e1009041.

Cox S (2010) *Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, v2.0.0. OGC 10-004r3*. http://www.opengis.net/doc/as/om/2.0.

Damerow JE, Varadharajan C, Boye K, Brodie EL, Burrus M, Chadwick KD, Crystal-Ornelas R, Elbashandy H, Alves RJE, Ely KS, Goldman AE, Haberman T, Hendrix V, Kakalia Z, Kemner KM, Kersting AB, Merino N, O'Brien F, Perzan Z, Robles E, Sorensen P, Stegen JC, Walls RL, Weisenhorn P, Zavarin M, and Agarwal D (2021) Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. *Data Science Journal* 20(1): 11. https://doi.org/10.5334/dsj-2021-011.

David R, Specht A, O'Brien M, Wyborn L, Drummond C, Edmunds R, Filippone C, Machicao J, Miyairi N, Parton G, Pignatari Drucker D, Stall S, and Zimmer N (2022) *Multilingual Data Challenges in Professionalizing Data Stewardship Worldwide*. Zenodo. https://doi.org/10.5281/zenodo.6588167.

Deines P, Goldstein SL, Oelkers EH, Rudnick RL, and Walter LM (2003) Standards for publication of isotope ratio and chemical data in Chemical Geology. *Chemical Geology* 202(1–2): 1–4. https://doi.org/10.1016/j.chemgeo.2003.08.003.

Demetriades A, Huimin D, Kai L, Savin I, Birke M, Johnson CC, and Argyraki A (2020) *International Union of Geological Sciences Manual of Standard Geochemical Methods for the Global Black Soil Project*. International Union of Geological Sciences Commission on Global Geochemical Baselines. https://doi.org/10.5281/ZENODO.7267967.

Demetriades A, Johnson CC, Smith DB, Ladenberger A, Adánez Sanjuan PA, Argyraki A, Stouraiti C, de Caritat P, Knights KV, Prieto Rincón G, and Simubali GN (2022) *International Union of Geological Sciences Manual of Standard Methods for Establishing the Global Geochemical Reference Network*, p. 515. Athens, Hellenic Republic: IUGS Commission on Global Geochemical Baselines. https://doi.org/10.5281/ZENODO.7307696.

DFG. (2022) *Package of Measures to Support a Shift in the Culture of Research Assessment*. https://www.dfg.de/en/news/news-topics/announcements-proposals/2022/info-wissenschaft-22-61.

Digital Science, Simons N, Goodey G, Hardeman M, Clare C, Gonzales S, et al. (2021) *The State of Open Data 2021*. Digital Science. Report. https://doi.org/10.6084/m9.figshare.17061347.v1.

Dolan L and Whipple EC (2023) Electronic lab notebooks in practice. In: *Research Data Access & Preservation (RDAP) Annual Meeting*. https://hdl.handle.net/1805/32016.

Duerr R, Buttigieg P, Berg-Cross G, Blumberg K, Whitehead B, Wiegand N, and Rose K (2023) Harmonizing GCW cryosphere vocabularies with ENVO and SWEET: Towards a general model for semantic harmonization. *EarthArXiv*. https://doi.org/10.31223/x5c966.

Dutton A, Rubin K, McLean N, Bowring J, Bard E, Edwards RL, Henderson GM, Reid MR, Richards DA, Sims KWW, Walker JD, and Yokoyama Y (2017) Data reporting standards for publication of U-series data for geochronology and timescale assessment in the Earth sciences. *Quaternary Geochronology* 39: 142–149. https://doi.org/10.1016/j.quageo.2017.03.001.

Dziewonski AM (1994) The FDSN: History and objectives. *Annals of Geophysics* 37(5). https://doi.org/10.4401/ag-4191.

Dziewonski AM and Anderson DL (1981) Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors* 25(4): 297–356. https://doi.org/10.1016/0031-9201(81)90046-7.

Einsle JF, Lehnert K, Klöcking M, Edwards GH, Bermanec M, Anderson E, and Sylvester P (2023) Making Geochemical Microanalytical Imagery Accessible and Reusable. https://agu.confex.com/agu/fm23/meetingapp.cgi/Paper/1309965.

European Commission. (2018) *Cost-Benefit Analysis for FAIR Research Data: Cost of Not Having FAIR Research Data*. LU: European Commission, Directorate General for Research and Innovation and PwC EU Services. https://doi.org/10.2777/02999.

Flowers RM, Zeitler PK, Dani⊠ík M, Reiners PW, Gautheron C, Ketcham RA, Metcalf JR, Stockli DF, Enkelmann E, and Brown RW (2022) (U-Th)/He chronology: Part 1. Data, uncertainty, and reporting. *GSA Bulletin* 135(1–2): 104–136. https://doi.org/10.1130/b36266.1.

Flowers RM, Ketcham RA, Enkelmann E, Gautheron C, Reiners PW, Metcalf JR, Danišík M, Stockli DF, and Brown RW (2023) (U-Th)/He chronology: Part 2. Considerations for evaluating, integrating, and interpreting conventional individual aliquot data. *Geological Society of America Bulletin* 135(1–2): 137–161.

Gale A, Dalton CA, Langmuir CH, Su Y, and Schilling J-G (2013) The mean composition of ocean ridge basalts. *Geochemistry, Geophysics, Geosystems* 14(3): 489–518.

Gentemann C (2023) Why NASA and federal agencies are declaring this the Year of Open Science. *Nature* 613(7943): 217. https://doi.org/10.1038/d41586-023-00019-y.

Geochemical Society. (2007) *Geochemical Society Policy on Geochemical Databases*. https://www.geochemsoc.org/about/positionstatements/datapolicy.

Gill JC (2017) Geology and the sustainable development goals. *Episodes* 40(1): 70–76. https://doi.org/10.18814/epiiugs/2017/v40i1/017010.

Giuliani A, Oesch S, Guillong M, and Howarth GH (2024) Mica Rb-Sr dating by laser ablation ICP-MS/MS using an isochronous calibration material and application to West African kimberlites. *Chemical Geology* 649(121982): 121982.

Goldstein S, Lehnert K, and Hofmann A (2014) *Requirements for the Publication of Geochemical Data*. Interdisciplinary Earth Data Alliance (IEDA). https://doi.org/10.1594/IEDA/100426.

Goodenough K and Mills K (2021) Reflecting on the colonial legacy of geoscience in Africa. *Elements (Que.)* 17(5): 302. 302.

Gordon JM, Chkhenkeli N, Govoni DL, Lightsom FL, Ostroff AC, Schweitzer PN, Thongsavanh P, Varanka DE, and Zednik S (2015) *A Case Study of Data Integration for Aquatic Resources Using Semantic Web Technologies*. U.S. Geological Survey. https://doi.org/10.3133/ofr20151004.

Gregory A and Hodson S (2023) *Cross-Domain Interoperability Framework (CDIF) Working Documents*.

Haller A, Janowicz K, Cox SJD, Lefrançois M, Taylor K, Le Phuoc D, Lieberman J, García-Castro R, Atkinson R, and Stadler C (2018) The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web* 10(1): 9–32. https://doi.org/10.3233/sw-180320.

Hasterok D, Gard M, Bishop CMB, and Kelsey D (2019) Chemical identification of metamorphic protoliths using machine learning methods. *Computers & Geosciences* 132: 56–68. https://doi.org/10.1016/j.cageo.2019.07.004.

Hazen RM, Downs RT, Eleish A, Fox P, Gagné OC, Golden JJ, Grew ES, Hummer DR, Hystad G, Krivovichev SV, Li C, Liu C, Ma X, Morrison SM, Pan F, Pires AJ, Prabhu A, Ralph J, Runyon SE, and Zhong H (2019) Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization. *Engineering* 5(3): 397–405. https://doi.org/10.1016/j.eng.2019.03.006.

He Y, Zhou Y, Wen T, Zhang S, Huang F, Zou X, Ma X, and Zhu Y (2022) A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications. *Applied Geochemistry* 140: 105273. https://doi.org/10.1016/j.apgeochem.2022.105273.

Heidorn PB (2008) Shedding light on the dark data in the long tail of science. *Library Trends* 57(2): 280–299. https://doi.org/10.1353/lib.0.0036.

Hey T and Trefethen A (2003) The Data Deluge: An e-Science Perspective. pp. 809–824. Wiley. https://doi.org/10.1002/0470867167.ch36.

Higgins SG, Nogiwa-Valdez AA, and Stevens MM (2022) Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nature Protocols* 17(2): 179–189. https://doi.org/10.1038/s41596-021-00645-8.

Hinze W (2001) Transition to electronic publishing: Part 1. Electronic supplements. *Eos, Transactions American Geophysical Union* 82(22): 243. https://doi.org/10.1029/01eo00137.243.

Hodson S (2024) *WorldFAIR (D2.2) WorldFAIR's Experience with FIPs (second set of FAIR Implementation Profiles for each case study)*. Zenodo. https://doi.org/10.5281/zenodo.11236094.

Horsburgh JS, Aufdenkampe AK, Mayorga E, Lehnert KA, Hsu L, Song L, Jones AS, Damiano SG, Tarboton DG, Valentine D, Zaslavsky I, and Whitenack T (2016) Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software* 79: 55–74. https://doi.org/10.1016/j.envsoft.2016.01.010.

Horstwood MSA, Košler J, Gehrels G, Jackson SE, McLean NM, Paton C, Pearson NJ, Sircombe K, Sylvester P, Vermeesch P, Bowring JF, Condon DJ, and Schoene B (2016) Community-derived standards for LA-ICP-MS U-(Th-)Pb geochronology – Uncertainty propagation, age interpretation and data reporting. *Geostandards and Geoanalytical Research* 40(3): 311–332. https://doi.org/10.1111/j.1751-908x.2016.00379.x.

Hsu L, Lehnert KA, Walker JD, Chan C, Ash J, Johansson AK, and Rivera TA (2011) Maximizing data holdings and data documentation with a hierarchical system for sample-based geochemical data. In: *AGU Fall Meeting Abstracts*, vol. 2011, p. IN23C-1462.

Hsu L, Mayorga E, Horsburgh JS, Carter MR, Lehnert KA, and Brantley SL (2017) Enhancing interoperability and capabilities of earth science data using the observations data model 2 (ODM2). *Data Science Journal* 16. https://doi.org/10.5334/dsj-2017-004.

Hwang L, Fish A, Soito L, Smith M, and Kellogg LH (2017) Software and the scientist: Coding and citation practices in geodynamics. *Earth and Space Science* 4(11): 670–680. https://doi.org/10.1002/2016ea000225.

Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot MA, Crosas MA, Dumontier M, Evelo CT, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RWW, Imming M, Jeffery KG, Kaliyaperumal R, Kersloot MG, Kirkpatrick CR, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S-A, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson MD, Willighagen EL, Wittenburg P, Roos M, Mons B, and Schultes E (2020) FAIR principles: Interpretations and implementation considerations. *Data Intelligence* 2(1–2): 10–29. https://doi.org/10.1162/dint_r_00024.

Jochum KP and Enzweiler J (2014) Reference materials in geochemical and environmental research. In: *Treatise on Geochemistry*, pp. 43–70. Elsevier.

Jochum KP and Nohl U (2008) Reference materials in geochemistry and environmental research and the GeoReM database. *Chemical Geology* 253(1–2): 50–53.

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

Keller CB and Harrison TM (2020) Constraining crustal silica on ancient Earth. *Proceedings of the National Academy of Sciences* 117(35): 21101–21107. https://doi.org/10.1073/pnas.2009431117.

Keller CB and Schoene B (2012) Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. *Nature* 485(7399): 490–493. https://doi.org/10.1038/nature11024.

Keller B and Schoene B (2018) Plate tectonics and continental basaltic geochemistry throughout Earth history. *Earth and Planetary Science Letters* 481: 290–304. https://doi.org/10.1016/j.epsl.2017.10.031.

Kenah BI, Mercer CM, and Cohen BA (2020) Selenocene—A software tool to track the processing history of samples analyzed in the Mid-Atlantic Noble Gas Research Laboratory. In: *51st Annual Lunar and Planetary Science Conference (No. 2326)*, p. 1639.

Khider D, Emile-Geay J, McKay NP, Gil Y, Garijo D, Ratnakar V, Alonso-Garcia M, Bertrand S, Bothe O, Brewer P, Bunn A, Chevalier M, Comas-Bru L, Csank A, Dassié E, DeLong K, Felis T, Francus P, Frappier A, Gray W, Goring S, Jonkers L, Kahle M, Kaufman D, Kehrwald NM, Martrat B, McGregor H, Richey J, Schmittner A, Scroxton N, Sutherland E, Thirumalai K, Allen K, Arnaud F, Axford Y, Barrows T, Bazin L, Birch SEP, Bradley E, Bregy J, Capron E, Cartapanis O, Chiang H-W, Cobb KM, Debret M, Dommain R, Du J, Dyez K, Emerick S, Erb MP, Falster G, Finsinger W, Fortier D, Gauthier N, George S, Grimm E, Hertzberg J, Hibbert F, Hillman A, Hobbs W, Huber M, Hughes ALC, Jaccard S, Ruan J, Kienast M, Konecky B, Roux GL, Lyubchich V, Novello VF, Olaka L, Partin JW, Pearce C, Phipps SJ, Pignol C, Piotrowska N, Poli M-S, Prokopenko A, Schwanck F, Stepanek C, Swann GEA, Telford R, Thomas E, Thomas Z, Truebe S, Gunten L, Waite A, Weitzel N, Wilhelm B, Williams J, Williams JJ, Winstrup M, Zhao N, and Zhou Y (2019) PaCTS 1.0: A crowdsourced reporting standard for paleoclimate data. *Paleoceanography and Paleoclimatology* 34(10): 1570–1596. https://doi.org/10.1029/2019pa003632.

Kim AY and Hardin J (2020) "Playing the whole game": A data collection and analysis exercise with google calendar. *Journal of Statistics and Data Science Education* 29(sup1): S51–S60. https://doi.org/10.1080/10691898.2020.1799728.

Kim Y and Stanton J (2012) Institutional and individual influences on scientists' data sharing practices. *The Journal of Computational Science Education* 3(1): 47–56. https://doi.org/10.22369/issn.2153-4136/3/1/6.

Klein BZ and Eddy MP (2023) What's in an age? Calculation and interpretation of ages and durations from U-Pb zircon geochronology of igneous rocks. *Geological Society of America Bulletin* 136: 93–109.

Klöcking M, Wyborn L, Lehnert KA, Ware B, Prent AM, Profeta L, Kohlmann F, Noble W, Bruno I, Lambart S, Ananuer H, Barber ND, Becker H, Brodbeck M, Deng H, Deng K, Elger K, de Souza Franco G, Gao Y, Ghasera KM, Hezel DC, Huang J, Kerswell B, Koch H, Lanati AW, ter Maat G, Martínez Villegas N, Nana Yobo L, Redaa A, Schäfer W, Swing MR, Taylor RJM, Traun MK, Whelan J, and Zhou T (2023) Community recommendations for geochemical data, services and analytical capabilities in the 21st century. *Geochimica et Cosmochimica Acta* 351: 192–205. https://doi.org/10.1016/j.gca.2023.04.024.

Klump J, Lehnert K, Ulbricht D, Devaraju A, Elger K, Fleischer D, Ramdeen S, and Wyborn L (2021a) Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Science Journal* 20: 1–16.

Klump J, Wyborn L, Wu M, Martin J, Downs RR, and Asmi A (2021b) Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal* 20. https://doi.org/10.5334/dsj-2021-012.

Kohn BP, Ketcham RA, Vermeesch P, Boone SC, Hasebe N, Chew D, Bernet M, Chung L, Danišík M, Gleadow AJW, and Sobel ER (2024) Interpreting and reporting fission-track chronological data. *Geological Society of America Bulletin*. https://doi.org/10.1130/B37245.1.

Kragh H (2001) *From Geochemistry to Cosmochemistry: The Origin of a Scientific Discipline, 1915-1955*. Wiley-VCH. https://doi.org/10.1002/9783527612734.ch09.

Le Bas MJ and Durham J (1989) Scientific communication of geochemical data and the use of computer databases. *Journal of Documentation* 45(2): 124–138. https://doi.org/10.1108/eb026842.

Lehnert K, Su Y, Langmuir CH, Sarbas B, and Nohl U (2000) A global geochemical database structure for rocks. *Geochemistry, Geophysics, Geosystems* 1(5). https://doi.org/10.1029/1999gc000026.

Lehnert K, Profeta L, and Mays J (2022) Managing Analytical Data from Pristine Returned Samples in Compliance with NASA'S Data Strategy: The Astromaterials Data System. Authorea. https://doi.org/10.1002/essoar.10510838.1.

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, and Westbrook J (2020) The TRUST principles for digital repositories. *Scientific Data* 7(1). https://doi.org/10.1038/s41597-020-0486-7.

Liu H, Sun W-D, and Deng J (2019) Statistical analysis on secular records of igneous geochemistry: Implication for the early Archean plate tectonics. *Geological Journal* 55(1): 994–1002. https://doi.org/10.1002/gj.3484.

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, and Villa F (2007) An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* 2(3): 279–296. https://doi.org/10.1016/j.ecoinf.2007.05.004.

Mahan SA, Rittenour TM, Nelson MS, Ataee N, Brown N, DeWitt R, Durcan J, Evans M, Feathers J, Frouin M, Guérin G, Heydari M, Huot S, Jain M, Keen-Zebert A, Li B, López GI, Neudorf C, Porat N, Rodrigues K, Sawakuchi AO, Spencer JQG, and Thomsen K (2022) Guide for interpreting and reporting luminescence dating results. *Geological Society of America Bulletin* 135: 1480–1502.

Manten AA (1966) Historical foundations of chemical geology and geochemistry. *Chemical Geology* 1: 5–31. https://doi.org/10.1016/0009-2541(66)90003-9.

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot MA, Deck J, Dumontier M, Fellows DK, Gonzalez-Beltran A, Gormanns P, Grethe J, Hastings J, Hériché J-K, Hermjakob H, Ison JC, Jimenez RC, Jupp S, Kunze J, Laibe C, Le Novére N, Malone J, Martin MJ, McEntyre JR, Morris C, Muilu J, Müller W, Rocca-Serra P, Sansone S-A, Sariyar M, Snoep JL, Soiland-Reyes S, Stanford NJ, Swainston N, Washington N, Williams AR, Wimalaratne SM, Winfree LM, Wolstencroft K, Goble C, Mungall CJ, Haendel MA, and Parkinson H (2017) Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biology* 15(6): e2001414. https://doi.org/10.1371/journal.pbio.2001414.

Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, and Wilkinson MD (2017) Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. *Information Services & Use* 37(1): 49–56. https://doi.org/10.3233/isu-170824.

Morris PJ (2002) *From Classical to Modern Chemistry: The Instrumental Revolution*. Cambridge: Royal Society of Chemistry.

NASA Science Mission Directorate. (2022) *SPD-41a: Scientific Information Policy for the Science Mission Directorate*. https://science.nasa.gov/spd-41/.

Nathwani CL, Wilkinson JJ, Brownscombe W, and John CM (2023) Mineral texture classification using deep convolutional neural networks: An application to zircons from porphyry copper deposits. *Journal of Geophysical Research: Solid Earth* 128(2). e2022JB025933.

National Science and Technology Council (NSTC). (2022) *Guidance for Implementing National Security Presidential Memorandum 33 (NSPM-33) on National Security Strategy for United States Government-Supported Research and Development: A Report by the Subcommittee on Research Security*. https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf.

Niu X, Lehnert KA, Williams J, and Brantley SL (2011) CZChemDB and EarthChem: Advancing management and access of critical zone geochemical data. *Applied Geochemistry* 26: S108–S111. https://doi.org/10.1016/j.apgeochem.2011.03.042.

Parsons MA (2013) The research data alliance: Implementing the technology, practice and connections of a data infrastructure. *Bulletin of the American Society for Information Science and Technology* 39(6): 33–36. https://doi.org/10.1002/bult.2013.1720390611.

Peng G, Lacagnina C, Downs RR, Ganske A, Ramapriyan HK, Ivánová I, Wyborn L, Jones D, Bastin L, Shie C-L, and Moroni DF (2022) Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. *Data Science Journal* 21. https://doi.org/10.5334/dsj-2022-008.

Petrelli M (2024) Machine learning in petrology: State-of-the-art and future perspectives. *Journal of Petrology* 65(5): egae036.

Pignatelli A and Piochi M (2021) Machine learning applied to rock geochemistry for predictive outcomes: The Neapolitan volcanic history case. *Journal of Volcanology and Geothermal Research* 415: 107254. https://doi.org/10.1016/j.jvolgeores.2021.107254.

Pilger RH Jr (2022) *Radiometric Dates from the South American Andes and Adjacent Areas: A Compilation*. GFZ Data Services. https://doi.org/10.5880/digis.e.2023.007.

Piwowar HA, Day RS, and Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2(3): e308. https://doi.org/10.1371/journal.pone.0000308.

Prent AM, Hezel DC, Klöcking M, Wyborn L, Farrington R, Elger K, Profeta L, Nixon AL, and Lehnert K (2023) Innovating and networking global geochemical data resources through OneGeochemistry. *Elements (Que.)* 19(3): 136–137.

Prent A, Farrington R, Wyborn L, Nixon A, Elger K, Klöcking M, Hezel D, and Lehnert K (2024) *WorldFAIR (d5.3) Guidelines for Implementing Geochemistry FIPs*.

Prodanović M, Esteva M, McClure J, Chang BC, Santos JE, Radhakrishnan A, Singh A, and Khan H (2023) Digital Rocks Portal (Digital Porous Media): Connecting data, simulation and community. *E3S Web of Conferences* 367: 01010.

Profeta L, Lehnert K, Ramdeen S, Ji P, Nielsen RL, Ustunisik GK, Walker JD, Block KA, and Grossberg M (2022) The IEDA2 facility-harmonizing FAIR sample (meta) data for VGP research. In: *AGU Fall Meeting Abstracts (vol. 2022, pp. V42A-05).*

Przeslawski R, Pearlman J, and Karstensen J (2022) *PDataset Quality Information in Australia's Integrated Marine Observing System*. https://www.scidatacon.org/IDW-2022/sessions/431/paper/969/.

Quach S and Ambalgekar P (2024) *Elasticsearch: What It Is, How It Works, and It's Usage*. https://www.knowi.com/blog/what-is-elastic-search/.

Quinn D, Linzmeier B, Sundell K, Gehrels G, Goring S, Marcott S, Meyers S, Peters S, Ross J, Schmitz M, Singer B, and Williams J (2021) Implementing the Sparrow laboratory data system in multiple subdomains of geochronology and geochemistry. *EGU General Assembly*. https://doi.org/10.5194/egusphere-egu21-13832.

Rasmussen D, Plank T, Cottrell E, Johansson A, Lehnert K, and Hauri E (2022) *Dco-earthchem Melt Inclusion Expert Dataset*. https://doi.org/10.26022/IEDA/112364. https://ecl.earthchem.org/view.php?id=2364.

Reidpath DD and Allotey P (2019) The problem of 'trickle-down science' from the global north to the global south. *BMJ Global Health* 4(4): e001719.

Reinhardt C (2006) *Shifting and Rearranging: Physical Methods and the Transformation of Modern Chemistry*. Sagamore Beach: Science History Publications.

Reinhardt C (2019) IUPAC engagement in the instrumental revolution. *Chemistry International* 41(3): 35–38. https://doi.org/10.1515/ci-2019-0312.

Rudnick RL and Gao S (2003) Composition of the Continental Crust. pp. 1–64. Elsevier. https://doi.org/10.1016/b0-08-043751-6/03016-4.

Ruiz A (2017) *The 80/20 Data Science Dilemma*. https://science.nasa.gov/spd-41/.

Sahagún L (2021) *Caltech Says It Regrets Drilling Holes in Sacred Native American Petroglyph Site*. https://www.latimes.com/environment/story/2021-07-19/caltech-fined-for-damaging-native-american-cultural-site.

Schaen AJ, Jicha BR, Hodges KV, Vermeesch P, Stelten ME, Mercer CM, Phillips D, Rivera TA, Jourdan F, Matchan EL, Hemming SR, Morgan LE, Kelley SP, Cassata WS, Heizler MT, Vasconcelos PM, Benowitz JA, Koppers AAP, Mark DF, Niespolo EM, Sprain CJ, Hames WE, Kuiper KF, Turrin BD, Renne PR, Ross J, Nomade S, Guillou HA, Webb LE, Cohen BA, Calvert AT, Joyce N, Ganerød M, Wijbrans J, Ishizuka O, He H, Ramirez AA, Pfänder JA, Lopez-Martínez M, Qiu H, and Singer BS (2020) Interpreting and reporting 40Ar/39Ar geochronologic data. *GSA Bulletin* 133(3–4): 461–487. https://doi.org/10.1130/b35560.1.

Schleidt K and Rinne I (2023) *OGC Abstract Specification Topic 20: Observations, measurements and samples, v3.0.0. OGC 20-082r4*. http://www.opengis.net/doc/as/om/3.0.

Schönbein CF (1838) Ueber die Ursache der Farbenveränderung, welche manche Körper unter dem Einflusse der Wärme erleiden. *Annalen der Physik* 121(10): 263–281. https://doi.org/10.1002/andp.18381211007.

Schultes E, Magagna B, Hettne KM, Pergl R, Suchánek M, and Kuhn T (2020) Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. pp. 138–147. Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_13.

Shaw DM and Bankier JD (1954) Statistical methods applied to geochemistry. *Geochimica et Cosmochimica Acta* 5(3): 111–123. https://doi.org/10.1016/0016-7037(54)90011-3.

Sheridan C (2004) Kenyan dispute illuminates bioprospecting difficulties. *Nature Biotechnology* 22(11): 1337.

Simpson P, Pearlman F, and Pearlman J (2018) *Evolving and Sustaining Ocean Best Practices Workshop 15–17 November 2017 Intergovernmental Oceanographic Commission, Paris, France: Proceedings*. AtlantOS/ODIP/OORCN Ocean Best Practices Working Group. https://doi.org/10.25607/OBP-3. https://www.oceanbestpractices.net/handle/11329/410.

Skobelev DO, Zaytseva TM, Kozlov AD, Perepelitsa VL, and Makarova AS (2011) Laboratory information management systems in the work of the analytic laboratory. *Measurement Techniques* 53(10): 1182–1189.

Smith R (2001) Electronic publishing in science. *British Medical Journal* 322(7287): 627–629. https://doi.org/10.1136/bmj.322.7287.627.

Spek AL (2020) checkCIF validation ALERTS: What they mean and how to respond. *Acta Crystallographica Section E Crystallographic Communications* 76(1): 1–11. https://doi.org/10.1107/s2056989019016244.

Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, Parsons M, Robinson E, and Wyborn L (2019) Make scientific data FAIR. *Nature* 570(7759): 27–29. https://doi.org/10.1038/d41586-019-01720-7.

Stall S, Bilder G, Cannon M, Chue Hong N, Edmunds S, Erdmann CC, Evans M, Farmer R, Feeney P, Friedman M, Giampoala M, Hanson RB, Harrison M, Karaiskos D, Katz DS, Letizia V, Lizzi V, MacCallum C, Muench A, Perry K, Ratner H, Schindler U, Sedora B, Stockhause M, Townsend R, Yeston J, and Clark T (2023) Journal production guidance for software and data citations. *Scientific Data* 10(1). https://doi.org/10.1038/s41597-023-02491-7.

Stefanoudis PV, Licuanan WY, Morrison TH, Talma S, Veitayaki J, and Woodall LC (2021) Turning the tide of parachute science. *Current Biology* 31(4): R184–R185. https://doi.org/10.1016/j.cub.2021.01.029.

Stracke A, Willig M, Genske F, Béguelin P, and Todd E (2022) *Major and Trace Element Concentrations and Sr, Nd, Hf, Pb Isotope Ratios of Global Mid Ocean Ridge and Ocean Island Basalts*. GFZ Data Services. https://doi.org/10.5880/digis.e.2024.008.

Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft Mithu, and Astell M (2018) *Whitepaper: Practical Challenges for Researchers in Data Sharing*. https://doi.org/10.6084/M9.FIGSHARE.5975011.V1.

Taitingfong R, Martinez A, Carroll SR, Hudson M, and Anderson J (2023) *Indigenous Metadata Bundle Communique*. Collaboratory for Indigenous Data Governance, ENRICH: Equity for Indigenous Research and Innovation Coordinating Hub, and Tikanga in Technology. https://doi.org/10.6084/m9.figshare.24353743.

EarthChem Team (2022) *In Situ Analysis Template v3.1*.

Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen A, Pedaste M, Raju M, Astapova A, Lukner H, Kogermann K, and Sepp T (2021) Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data* 8(1). https://doi.org/10.1038/s41597-021-00981-0.

Health The Lancet Global (2018) Closing the door on parachutes and parasites. *The Lancet Global Health* 6(6): e593. https://doi.org/10.1016/s2214-109x(18)30239-0.

Ueki K, Hino H, and Kuwatani T (2018) Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. *Geochemistry, Geophysics, Geosystems* 19(4): 1327–1347. https://doi.org/10.1029/2017gc007401.

UNESCO. (2021) *UNESCO Recommendation on Open Science*.

Vanderbilt KL, Blankman D, Guo X, He H, Lin C-C, Lu S-S, Ogawa A, Ó Tuama E, Schentz H, and Su W (2010) A multilingual metadata catalog for the ILTER: Issues and approaches. *Ecological Informatics* 5(3): 187–193. https://doi.org/10.1016/j.ecoinf.2010.02.002.

Vicente-Saez R and Martinez-Fuentes C (2018) Open science now: A systematic literature review for an integrated definition. *Journal of Business Research* 88: 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043.

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-SA, Renaut S, and Rennison DJ (2014) The availability of research data declines rapidly with article age. *Current Biology* 24(1): 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

Vrouwenvelder K and Stall S (2023) *Where are AGU authors sharing their data & software?*. https://zenodo.org/record/8378089. 10.5281/ZENODO.8378089.

Walker JD, Bowers TD, Black RA, Glazner AF, Lang Farmer G, and Carlson RW (2006) A Geochemical Database for Western North American Volcanic and Intrusive Rocks (NAVDAT). Geological Society of America. https://doi.org/10.1130/2006.2397(05).

Walker DJ, Condon D, Thompson W, Renne P, Koppers A, Hodges K, Reiners P, Stockli D, Schmitz M, Bowring S, and Gehrels G (2008) *Geochron Workshop Reports Sponsored by EarthChem and EARTHTIME*. https://doi.org/10.5281/ZENODO.4313859.

Wallace KL, Bursik MI, Kuehn S, Kurbatov AV, Abbott P, Bonadonna C, Cashman K, Davies SM, Jensen B, Lane C, Plunkett G, Smith VC, Tomlinson E, Thordarsson T, and Walker JD (2022) Community established best practice recommendations for tephra studies—From collection through analysis. *Scientific Data* 9(1). https://doi.org/10.1038/s41597-022-01515-y.

Walters WH (2020) Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights* 33: 18.

Weiss Y, Hanyu T, Class C, and Goldstein S (2016) *PetDB Expert Dataset: Mantle Xenolith Data from Cratonic Settings (Sorted and Revised) from www.earthchem.org/petdb*. https://doi.org/10.1594/IEDA/100602. https://ecl.earthchem.org/view.php?id=961.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas MA, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, and Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). https://doi.org/10.1038/sdata.2016.18.

Williamson B, Provost S, and Price C (2022) Operationalising indigenous data sovereignty in environmental research and governance. *Environment and Planning F* 2(1–2): 281–304. https://doi.org/10.1177/26349825221125496.

Wing JM (2019) The data life cycle. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e26845b4.

Woelfle M, Olliaro P, and Todd MH (2011) Open science is a research accelerator. *Nature Chemistry* 3(10): 745–748. https://doi.org/10.1038/nchem.1149.

Wörner G (2021) *Geochemical Compositions of Igneous Rocks of the Central Andean Orocline*. GFZ Data Services. https://doi.org/10.5880/digis.e.2024.005.

Wyborn L and Lehnert K (2021) OneGeochemistry: Creating a global network of geochemical data to support the 17 United Nations sustainable development goals. In: *Goldschmidt2021 Abstracts*. European Association of Geochemistry.

Wyborn LAI and Ryburn RJ (1989) *PETCHEM Data Set: Australia and Antarctica - Documentation*. Geoscience Australia, Canberra. http://pid.geoscience.gov.au/dataset/ga/14256. Record 1989/019.

Zeng ML (2008) Knowledge organization systems (KOS). *Knowledge Organization* 35(2–3): 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160.

Zindler A and Hart S (1986) Chemical geodynamics. *Annual Review of Earth and Planetary Sciences* 14(1): 493–571. https://doi.org/10.1146/annurev.ea.14.050186.002425.

Relevant websites

AGU Data Sharing Agreement. https://www.agu.org/-/media/files/publications/g-cubed-data-sharing-guidance.pdf.

EU Open Access Strategy. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-access_en.

EU Open Science Policy. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.

Geochimica et Cosmochimica Acta Guide for Authors. https://www.sciencedirect.com/journal/geochimica-et-cosmochimica-acta/publish/guide-for-authors.

Global Open Science Cloud. https://goscloud.net.

GO FAIR. https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/.

Horizon Europe Funding. https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm.

Journal of Petrology Information for Authors. https://academic.oup.com/petrology/pages/general_instructions.

NASA Data Levels. https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-information-policy/data-levels.

Nelson Memo. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

NSF. https://libguides.sjsu.edu/datamanagement/nsfdmp.

OECD Recommendation of the Council Concerning Access to Research Data From Public Funding. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347.

Year of Open Science. https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research.

https://codata.org/initiatives/decadal-programme2/worldfair/onegeochemistry-wg/—CODATA OneGeochemistry Working Group.

https://copdess.org/—COPDESS.

https://www.coretrustseal.org/—CoreTrustSeal.

https://www.fdsn.org/—FDSN.

https://www.gbif.org/what-is-gbif—GBIF: The Global Biodiversity Information Facility.

https://georem.mpch-mainz.gwdg.de/—GeoReM.

https://georoc.eu/—GEOROC.

https://www.ieee.org/—Institute of Electrical and Electronics Engineers (IEEE).

https://www.iso.org—International Organisation for Standardisation (ISO).

http://ihfc-iugg.org/products/global-heat-flow-database/—IUGG Global Heat Flow Database.

http://geosciml.org/—IUGS GeoSciML.

https://iupac.org/what-we-do/digital-standards/—IUPAC Digital Standards.

https://www.ogc.org/—Open Geospatial Consortium (OGC).

https://www.ogc.org/—Open Geospatial Consortium (OGC).

https://search.earthchem.org/—PetDB.

https://www.geosamples.org/—SESAR.

https://www.w3.org/—World Wide Web Consortium (W3C).

https://worldfair-project.eu/—WorldFAIR project.