Title

The future low-temperature geochemical data-scape as envisioned by the U.S. geochemical community

Authors
Brantley, Susan L
Wen, Tao
Agarwal, Deborah A
et al.

Publication Date
2021-12-01

DOI
10.1016/j.cageo.2021.104933

The Future Low-Temperature Geochemical Data-scape as Envisioned by the U.S. Geochemical Community

Susan L. Brantley[1,15], Tao Wen[2], Deborah Agarwal[3], Jeffrey G. Catalano[4], Paul A. Schroeder[5], Kerstin Lehnert[6], Charuleka Varadharajan[7], Julie Pett-Ridge[8], Mark Engle[9], Anthony M. Castronova[10], Richard P. Hooper[11], Xiaogang Ma[12], Lixin Jin[9], Kenton McHenry[13], Emma Aronson[14], Andrew R. Shaughnessy[15], Louis A. Derry[16], Justin Richardson[17], Jerad Bales[10], Eric M. Pierce[18]

1. Earth and Environmental Systems Institute and Department of Geosciences, The Pennsylvania State University, University Park, PA, USA
2. Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA
3. Advanced Computing for Science Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
4. Department of Earth and Planetary Sciences, Washington University, St. Louis, MO, USA
5. Department of Geology, University of Georgia, Athens, GA, USA
6. Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
7. Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley CA, USA
8. Department of Crop and Soil Science, Oregon State University, Corvallis, OR, USA
9. Department of Geological Sciences, The University of Texas at El Paso, El Paso, TX, USA
10. Consortium of Universities for the Advancement of Hydrological Science, Inc, Cambridge, MA, USA
11. Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA
12. Department of Computer Science, University of Idaho, Moscow, ID, USA
13. National Center for Supercomputing Applications, University of Illinois, Urbana, IL, USA
14. Department of Microbiology and Plant Pathology, University of California, Riverside, USA
15. Department of Geosciences, The Pennsylvania State University, University Park, PA, USA
16. Department of Earth and Atmospheric Sciences, Cornell University, Ithaca NY, USA
17. Department of Geosciences, University of Massachusetts Amherst, Amherst, MA, USA
18. Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN USA

Corresponding author:
Susan L. Brantley, Earth and Environmental Systems Institute and Department of Geosciences, The Pennsylvania State University, University Park, PA, USA
Email: sxb7@psu.edu

CRediT authorship contribution statement

Susan L. Brantley: Funding acquisition, Supervision, Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Tao Wen:** Supervision, Conceptualization, Investigation, Writing - review & editing. **Deborah Agarwal:** Supervision, Conceptualization, Investigation, Writing - review & editing. **Jeffrey G. Catalano:** Supervision, Conceptualization, Investigation, Writing - review & editing. **All other authors:** Writing - review & editing.

Abstract

Data sharing benefits the researcher, the scientific community, and the public by allowing the impact of data to be generalized beyond one project and by making science more transparent. However, many scientific communities have not developed protocols or standards for publishing, citing, and versioning datasets. One community that lags in data management is that of low-temperature geochemistry (LTG). This paper resulted from an initiative from 2018 through 2020 to convene LTG and data scientists in the U.S. to strategize future management of LTG data. Through webinars, a workshop, a preprint, a townhall, and a community survey, the group of U.S. scientists discussed the landscape of data management for LTG – the data-scape. Currently this data-scape includes a "street bazaar" of data repositories. This was deemed appropriate in the same way that LTG scientists publish articles in many journals. The variety of data repositories and journals reflect that LTG scientists target many different scientific questions, produce data with extremely different structures and volumes, and utilize copious and complex metadata. Nonetheless, the group agreed that publication of LTG science must be accompanied by sharing of data in publicly accessible repositories, and, for sample-based data, registration of samples with globally unique persistent identifiers. LTG scientists should use certified data repositories that are either highly structured databases designed for specialized types of data, or unstructured generalized data systems. Recognizing the need for tools to enable search and cross-referencing across the proliferating data repositories, the group proposed that the overall data informatics paradigm in LTG should shift from "build data repository, data will come" to "publish data online, cybertools will find". Funding agencies could also provide portals for LTG scientists to register funded projects and datasets, and forge approaches that cross national boundaries. The needed transformation of the LTG data culture requires emphasis in student education on science and management of data.

Keywords

Data management, data repositories, geochemistry, metadata, data sharing, open science

Highlights

1. Scientists use a wide variety of data repositories for heterogeneous LTG datasets
2. Both structured and unstructured databases are needed to store LTG data online
3. Powerful search tools and data portals are needed to enable LTG data discovery

1. Introduction

Scientific communities and publishers within geosciences are publishing their data online and promoting new ways to analyze these data (e.g. Asch and Jackson, 2006; Christensen et al., 2009; Horsburgh et al., 2011; Aspen Institute, 2017; Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI), 2018; Cousijn et al., 2018; Bergen et al., 2019; ESIP Data Preservation and Stewardship Committee, 2019; Gil et al., 2019; Stall et al., 2019; Liu et al., 2020; U.S.G.S., 2020a). Some publishers have promoted and agreed to the so-called Findability, Accessibility, Interoperability, and Reusability of digital assets (FAIR Data Principles). A few geoscience communities (e.g., climate, oceanography, cryosphere, ecology, genetics, atmospherics, and agricultural science) have progressed toward these goals in terms of managing their data online. The growth of the Open Science and Open Data movement has led publishers and data repositories in the Earth Sciences to collaborate as part of Coalition for Publishing Data in the Earth & Space Sciences (COPDESS, http://www.copdess.org), a group that is promoting best practices for data in publications in geosciences (COPDESS, 2020). Now, journals managed by the American Geophysical Union have opted into the 'Enabling FAIR Data' project to increasingly require data to be submitted to trusted, certified data repositories where they can be cited with a digital object identifier (DOI). The explosion in the use of sensors, remote sensing, automatic instrumentation, data analytics, and the increasing storage of data online in a globally connected information system is driving an increasingly efficient and accessible data management system or "data-scape" in the Earth Sciences.

However, as this movement has progressed, improvements remain slow in many subfields of geoscience, including low-temperature geochemistry, referred to here in this paper as LTG. For example, the transition in late 2018 to requiring basic data sharing for submissions to the journal of *Geochimica et Cosmochimica Acta* resulted in initial resistance by many authors. Today, a majority of authors choose to attach their data to the published manuscript as supporting material, which often remains behind a paywall. This approach is generally preferred by many authors as this does not require time-consuming data formatting or input protocols for a separate repository. As enforcement of new data management policies has intensified by journals and funding agencies, submissions to geochemical data repositories have increased for rock chemistry (Albarede and Lehnert, 2019). In addition, papers are beginning to appear that describe meta-analyses for topics as wide-ranging as arsenic and methane in groundwater (Podgorski and Berg, 2020; Wen et al., 2021), soil organic carbon (Gomes et al., 2019), and nutrients in rain and groundwater (Amos et al., 2018), and these papers highlight the utility of more extensive data sharing. Nonetheless, resistance to data management in repositories remains in the LTG community, as it does for other communities.

To understand this situation and to chart an appropriate roadmap for forward movement for management of LTG data within one country (U.S.), a two-year initiative was pursued to discuss the LTG data-scape (funded by the U.S. National Science Foundation, NSF). Four webinars were run (see Acknowledgements) and a 2.5-day workshop was held in February 2020 in Atlanta (Georgia, U.S.) with participants from data science and geochemistry communities from within the NSF-funded LTG community. Workshop participants posted this paper in a preprint form at EarthArXiv (BRANTLEY et al., 2020), soliciting reader comments (none were posted). The posted paper was also sent to 350 geochemists funded by the NSF with i) a survey soliciting feedback and ii) an invitation for an online discussion. The survey and discussion included 27 and 24 participants respectively. This paper summarizes the outcome of all these discussions, noting that the participants were biased toward practicing geochemists with only a small number of data scientists. Thus, this paper is unusual compared to many other papers about data management in that it is mostly from the perspective of bench and field scientists within one country (U.S.). The intent was to consider the problem of data management with respect to the specific characteristics of LTG data and to propose a forward trajectory as new data systems are developed in the future. This paper is necessarily informed from that perspective because of the funding, but it is offered also as an invitation for other scientists worldwide to contemplate the LTG data-scape into the future.

For this paper, "LTG" describes any geoscience that investigates earth processes pertaining to the chemistry of surficial Earth materials including water and biota. This field includes, but is not limited to, chemical and biogeochemical cycling of elements, aqueous processes, mineralogy and chemistry of earth materials, the role of life in the evolution of Earth's geochemical cycles, biomineralization, medical mineralogy and geochemistry, and the geochemical aspects of critical zone science and geomicrobiology. In addition to these topics, LTG also includes tools, methods, and models pertaining to the fields listed above. This LTG definition is drawn from the definition currently used by the NSF for the U.S. LTG community.

At the workshop, we recognized that some sub-sets of the LTG community have already self-organized their approaches to data management, sometimes initiating their own best practices for data management systems (e.g., Table 1). To enable conversation at the workshop among more sub-sets of the LTG and data informatics communities, a short lexicon of terms was compiled (Table 2). We discovered that words were often used differently by domain scientists (geochemists) and data scientists, and even sometimes by different individuals within each community. The lexicon was also helpful for participants from communities that had yet to develop data management systems (e.g., Table 3).

The main questions at the workshop addressed data management and sharing from different perspectives. We focused on three areas. First, who are the different stakeholders interested in coordinated management of LTG data, and what does each of them want to achieve? To answer this question, we discussed what we perceive to be the characteristics of the optimal management system from the perspective of different stakeholders (e.g., data producers, data users, modelers, funders, journal editors, government agencies, the public). Second, we asked, how can we best secure the longevity of data for the future given that a typical research project in LTG in the U.S. is only three years without possibility of renewal? In this regard we noted that data archived in older papers can still be read, while data in "aging" electronic peripheral devices such as floppy disks can only be read by specialty workers, emphasizing the importance of the type of media for storage and the resources available for data storage (e.g. CHRISTENSEN et al., 2009). Similarly, data stored within proprietary software may not be accessible in the future if the software changes or is not maintained. Finally, we looked at the question, what does the data life cycle look like today for LTG? We noted that many LTG practitioners only collect small volumes of data and publish it in papers, while others pursue meta-analysis of multiple datasets. Although the original intent of the effort was to provide a definitive roadmap, it may not be surprising that we did not develop an "answer" here, but rather we describe a broad trajectory for a future data-scape for LTG data in the U.S. as a step forward.

2. Characteristics of LTG data

Geochemical data are highly heterogeneous in usage, type, volume, structure, dimensionality, quality, and character. The one trait that these data tend to share is that they often summarize chemical analysis or features related to chemical makeup along with estimates of sensitivity, reproducibility, accuracy, and type of analysis. An important characteristic of geochemical data is also that they are used not only by other chemists and geochemists, but also by scientists from other fields (e.g., environmental science, geophysics, agronomy, public health) as well as sometimes by the public (e.g., water quality, air quality).

Given these many types of and uses for LTG data, the structure of the data varies from one dataset to another. Analyses can focus on the 100+ elements, the 200+ stable and radiogenic isotopes, 5000+ minerals, or the thousands of inorganic and organic species that have been identified. A schematic example showing chemical analyses that might be made for one soil sample is shown in Figure 1. A few data characteristics are emphasized below.

Some geochemical data are sample-based. A "sample" is a physical object that can be archived (Table 2). Samples refer to both laboratory- and field-derived objects and can include any medium from liquids to solids to gases. They can derive from any of the 5000+ minerals known to form naturally (FLEISCHER, 2018) or from the large number of possible mixtures of these minerals (e.g. rocks, rock aggregate, sediments, soils). In addition, geochemists also study non- and nano-crystalline materials (HOCHELLA et al., 2019). Of great importance among the non-crystalline materials are all the different types of organic matter (e.g. HEMINGWAY et al., 2019) as well as living and non-living organisms and biotic waste materials. Finally, geochemists are not just interested in analyses of natural samples: they also investigate the human-made (i.e., engineered) materials and -associated wastes (i.e., incidental materials).

With each sample, geochemists can complete bulk analyses but they also can separate a single sample into multiple daughter sub-samples or they can extract the materials for different species or different associations or affinities (e.g. PICKERING, 1981) as exemplified in Figure 1. Thus, Earth materials (e.g., rocks, soils) are ground for bulk analysis while, in addition, individual fragments are separated and analyzed or targeted for analysis in a thin section using a variety of spectroscopic or microscopic tools. Similarly, when organisms are analyzed, the analysis can be for the bulk or for a specific part such as the leaves, trunk, xylem, brain, otolith, etc., and for each body part, the analysis can target the bulk or a sub-part such as the entrained water (e.g. ORLOWSKI et al., 2016). And of course, each of these sample-based analyses can target concentrations of different species: for example, elements, molecules, isotopes, isotopically-labelled molecules, etc. In addition, geochemical analyses do not just consist of tabulated analytical data; rather, they consist of spectra, diffractograms, photographs, spectrograms, and other types of images or pixelated data that are often not reported as tables. The volume of data associated with these datasets can be much, much larger than sample-based analytical data. Thus, whereas early datasets could be accommodated in a notebook, these newer and larger data volumes can only be accommodated in online data systems (Figure 2).

In contrast to sample-based data, LTG geochemists also collect time-series ("longitudinal") or field-based measurements (taken without collecting a sample) of liquids, gases, biota, and solids. Some of these time-series measurements are made by field workers, but increasingly, measurements are made with sensors (e.g. KIM et al., 2017) or remote sensing (e.g. BERATAN et al., 1997). Temporal variations are measured in real-time or intermittently over long durations (e.g. BENSON et al., 2010). Advances occurring in the technology of sensors and sensor networks are rapidly driving new types of data collection for water quality, soil and rock characteristics, gas composition, and biological properties.

Regardless of whether their measurements are sample-based, field measurement-based, or time-series, LTG scientists place great stock in new types of analyses. The upshot of this is that many LTG papers summarize data that are purely research grade. As shown schematically in Figure 3, these measurements are highly non-routine (one-of-a-kind or first-of-a-kind), in contrast to more established, routine measurements with accepted standards. Figure 3 emphasizes that, as innovation in the measurement protocol decreases from left to right, the ease of data management increases.

Finally, in addition to these sample-, field- and sensor-based measurements, many geochemical "data" now increasingly consist of model set-up (including input parameters), outputs, and/or calculations. One type of model output that is often thought of as data include measurements reported from instruments where manufacturers keep data processing protocols proprietary, leaving open access to raw data limited and sequestered behind a paywall limited to licensed users. Other types of model output are also stored and used by geochemists. For example, global oceanic chemistry models used by oceanographers and geochemists can yield very large datasets of salinity or trace element content versus location. These models can include predicted data, so-called "re-analysis" data, model workflows, and model programs, and often the community wants to have access to all of these "data" sets (KALNAY et al., 1996). In addition to the output "data", the tabulated input values are also of importance for each model run.

Given all of this heterogeneity in data types and model outputs, some LTG datasets are large in volume while others are very small. For example, model-related output "data" are commonly associated with very large "data" volumes, as are sensor or remote sensing data, both of which can provide high-spatiotemporal resolution. In contrast, many sample-based datasets may be relatively small in volume, at least partly because of the expense and time necessary to collect, prepare, sub-sample, and analyze (Figure 1). However, almost all geochemical data are large in terms of types of metadata that are needed. 'Metadata' refers to the information related to "who, what, when, where, how" for the data values (e.g. MICHENER, 2006; PALMER et al., 2017; WEN, 2020).

3. Lack of best practices, standards, and harmonization

The design of effective data repositories – whether for LTG or other disciplines – depends not only on characteristics of the data as described above, but also upon the goal of the investigator and the overall workflow for data generation and processing (RUEGG et al., 2014). As a result, even where many examples of a certain type of data have been collected, and even when they may be organized into online libraries, it is rare in LTG that there is a generally accepted standard for the data. For example, quantitative phase analysis of Earth materials, whether they are rocks, soils, sediments, or something else, is fundamental to LTG, and there are several libraries for such data (Table 1), but formats for sample preparation for X-ray diffraction, data collection, and meta-analysis have not been established within the community. In another example, the team behind one NSF-supported geochemical data repository (EarthChem Library) emphasized the most common methods and sample types into templates for petrologists to submit rock chemical data. When the team used the same template for communities beyond petrology, they were met with resistance because non-petrologists preferred templates tailored to their own workflows. As a consequence of the many workflows, practicing LTG scientists consistently reported that data and metadata protocols from highly standardized data repositories were difficult to implement for their own datasets. For example, sometimes metadata that is important to one discipline might not asked for in a specialized template (e.g., a soil scientist might want to indicate the soil order in a template for chemical composition but have no place to include that information), or metadata is required that was not collected (e.g., a soil scientist might not know the geologic age of a given formation).

The variety of workflows that characterize LTG is not just a consequence of competing egos or laboratories. Rather, the different workflows result from groups asking different questions about different processes in different types of environments that require different approaches. For example, soil scientists and geologists collect and analyze soils to pursue questions within LTG. But the former analyzes only the <2 mm fraction (because it impacts soil fertility the most) while the latter use the entire sample for analysis (because they calculate mass balance compared to parent rock). Thus, for routine analyses of different types of soils, the National Cooperative Soil Survey (NCSS) database (N.R.C.S., 2020) is useful because all the soils have been sieved in the same way before an analysis, but this database is not necessarily useful for mass balance calculated by geologists (BRIMHALL AND DIETRICH, 1987). In another example, many in-vitro analytical methods have been developed to assess the health impact and bioaccessibility of contaminants in dust particles in the human lungs (WISEMAN, 2015) but these protocols differ significantly from analyses aimed to understand leachability in environmental systems (PICKERING, 1981).

Another reason for the lack of agreement on standards and protocols of measurement and reporting data results from LTG practitioners' strong emphasis on development of new and/or non-standardized technique – for example in sampling methodology, chemical extraction, analytical technique, and laboratory protocol. This emphasis results not only in innovative new methodologies, but also in a lack of data standards, difficulty in creating templates for data or metadata input, and ultimately, difficulty in comparing datasets within the LTG community. Here, data standards are defined as policies or protocols that determine how geochemical data and metadata should be formatted, reported, and documented. Many LTG scientists have not heard of nor used standards such as the Observations and Measurements Protocol of the International Organization for Standardization (ISO) (COX, 2011). Likewise, few LTG scientists are aware of the so-called 'Requirements for the Publication of Geochemical Data' which were agreed upon in 2014 by an editors' roundtable (a roundtable that included geochemists). These requirements explain how to report data and metadata in structured, standardized manners (GOLDSTEIN et al., 2014).

Even where geochemical data are already compiled and accessible in one place such as the Water Quality Portal [co-sponsored by the U.S. Geological Survey (USGS), the Environmental Protection Agency (EPA), and the National Water Quality Monitoring Council (NWQMC)], the data are not harmonized, i.e., units, formats, analytical methods, detection limits, and other parameters are not presented consistently (e.g. SPRAGUE et al., 2016; SHAUGHNESSY et al., 2019). Apparently, data standards for agreed-upon units and measurement protocols have never emerged because i) communities have never felt enough need for or placed enough value on such standardization or ii) variations in protocols were simply necessary to answer the proposed research questions. Neither have LTG scientists addressed, as a community, how to cite and reward or incentivize scientists who collate, curate, synthesize, and share published data for LTG or for other communities (data interoperability). The lack of standards, formats, and norms has in turn hampered the development of automated flows of geochemical data into databases. For these and other reasons, geochemical data compilations have grown slowly (LEHNERT AND ALBAREDE, 2019).

4. Current data management systems

To date, a variety of data management systems have been used by LTG scientists, including storage in notebooks, offline data infrastructures (e.g., individual computers), published works (e.g., theses, preprints, and journal publications and supplemental material), and online data infrastructures (e.g., personal webpages, dedicated data repositories). A schematic showing the trend of data management is shown in Figure 2. As emphasized by the red-shaded arrow, the number of data values diminish from left to right as data are culled after quality control checks or data are not deemed important enough to save. The most structured form of data management system indicated on Figure 2 is a shared online relational database (upper right). Only a few of these are available for LTG data (see, for example, Supplementary Material). Such databases represent the most structured and demanding management systems, but they also promote the easiest data discovery, re-use for meta-analysis, and collaboration.

Some of the data repositories that have a track record of success for data types of interest to LTG (time-series water data, rock chemistry, atmospheric radiation measurements, CO_2 flux, etc.) are summarized in Table 1. Some of these are maintained and used as libraries (e.g., for spectra, electron micrographs, or diffraction patterns) and not data repositories. Such libraries do not generate DOIs for the data provider and may only retain a limited number of examples for each entity. An instructive example for mineralogy is the International Centre for Diffraction Data (ICDD) that offers a detailed (behind the paywall) library of experimental and theoretical mineral structure data that serves as a reference for identification and quantification of minerals. Other open-source databases for mineral structures are also available (e.g., Mineralogical Society of America Crystal Structure database).

Given that only a few highly structured targeted databases for LTG data are available, and that libraries are not true data repositories, many other LTG data types lack appropriate repositories (a few examples are listed in Table 3). For these "orphaned" data types, scientists either publish their data in a journal article or its supplement, leave it unpublished on their computer or in a thesis, publish it online on their personal website, or use generalized and unstructured data repositories that can accommodate any type of data file and can assign a DOI to the dataset. These generalized data repositories provide little curation of metadata and do not police data quality. On the other hand, they generally provide long-term storage and require that the data provider record a modicum of metadata to allow indexing and to enable search features.

Some of these general-purpose repositories operate behind a firewall or paywall, while some are open and free. Some can be used by anyone while others are limited to specific clientele (e.g., from a specific university, country, or funded program) or types of data. For example, geochemists in the USGS use ScienceBase (U.S.G.S., 2020c), geoscientists funded by the U.S. Department of Energy (DOE) use ESS-DIVE (see Supplemental Material) for ecosystem and watershed data (VARADHARAJAN et al., 2019) and the ARM data center for cloud and aerosol properties, and EDX for data related to fossil fuel energy (N.E.T.L., 2020). Other such generalized data repositories are also becoming available through publishers, universities, federal agencies, and private entities. Examples that are used by some NSF-funded geochemists are EarthChem Library and CUAHSI's HydroShare (see Supplemental Material). No portal links to all the many data repositories used by LTG scientists.

Despite the examples in Table 1, most LTG scientists are not using data repositories. Thus, even for those parts of LTG science for which data management systems have been developed, many practitioners of LTG do not understand the repositories, how to use them, how to manage their data efficiently to prepare to ingest data into the repository, nor what kind of science they could enable. The problem is somewhat circular in nature because some of the difficulties in data management could be reduced by 'best practices' in data management throughout the data life cycle, but often the data repository itself is simply not well suited to the scientists' data needs, leaving it less likely to be used (Figure 4). The bottleneck where LTG scientists are not uploading data into online repositories (Figure 2) is likely impacting the kind of LTG science that is completed (Figure 4).

5. Lessons learned

Several important lessons were learned (Table 4) by inspecting the history of a few U.S.-centric LTG data management systems (see, Supplemental Materials). Figure 2 shows a conceptual schematic for the evolution of these management systems. From bottom to top on Figure 2, systems increasingly allow efficient and easy data discovery outside of the data producers' home group, improving the ease of collaboration across groups and disciplines. At the same time, however, increasing the utility and efficiency for the data user from top to bottom on Figure 2 entails more formalized and rigid rules for formatting and uploading data (i.e., from left to right on the graph), limiting flexibility for the data provider. Progress along the large arrow from left to right and bottom to top on the diagram also requires increasing effort by the community to prioritize data standards. With data standards, data harmonization is more likely, and data access therefore becomes easier for the data user, but formatting demands increase for the data provider. Six lessons with respect to LTG gleaned from the initiative are summarized below and in Figures 3-4 and Table 4. The order of subsections below roughly moves from lessons about the more general aspects of workflows to lessons that are more specific to data management systems in LTG.

5.1. The data enterprise from measurement to meta-analysis is complex and provides multiple opportunities for error, but systematic management of data and metadata leads both to improvements in the quality of the dataset and identification of large-scale trends within the data.

Few individuals in LTG understand the entire trajectory of data from sample collection / sensor deployment to publication. Errors can creep in at all steps and only a very few people within this enterprise can assure the quality of the data. These personnel tend to be those who made or supervised the measurements or who were responsible for reference standards, methodologies, instrumentation upkeep, and quality assurance measures. These personnel need to be involved in organization of metadata and assurance of data quality. Even when the data volume is small, metadata often becomes highly complex, especially if the information is to be of lasting usefulness [a point also made for ecological data (MICHENER, 2006)]. LTG metadata is complex partly because interpretation of chemical analyses requires understanding details of sub-sampling, extractions, or density separations before analysis (Figure 1).

As data are moved from the laboratory notebook to compiled datasets to shared data repositories along the trajectory in Figure 2, many opportunities for errors arise and data systems necessarily accrue errors. While most data management systems have very limited capacity to check for data quality, systematic data management promotes discovery of issues related to data quality or organization or metadata, and large-scale trends and patterns in the data can become apparent. Thus, even though compilation of data can be accompanied by error, systematic data and metadata management generally improves the overall quality of data sets and makes them more valuable. It is even possible that development of data management systems would lead to better tools for finding data quality issues.

5.2. As determined by their specific goals, LTG scientists participate in many different workflows, produce data with different structures and metadata, and make different choices with respect to how and where they publish their data, contributing to a proliferation of data management systems.

Some sampling and analytical strategies in LTG are routine. “Routine” data are relatively easy to standardize and manage in structured repositories (Figure 3). Example of “routine” data are measurements of solute concentrations, pH, alkalinity, and other parameters completed on water samples by the National Water Quality Laboratory (USGS) or completed based on standard methods (APHA, 1998).

In contrast, data developed from non-standardized analytical techniques or after refinements of specific issues with respect to collection or analysis of novel types of samples are inherently non-routine. These data generally are more difficult to archive in standardized data management frameworks and may also require extensive metadata, including discussions of analytical technique and clear disclosure of underlying assumptions.

Even with samples undergoing mostly routine analyses, some samples are treated differently and can be difficult to formally enter into standardized data management systems. This is because a geochemist may have to use one workflow of separation / extraction / analysis for one rock sample and another for a second sample of different composition. For example, a low-sulfur red shale generally requires one type of analytical workflow while a high-sulfur black shale requires another because bulk elemental analysis is affected by sulfur content. Overall, LTG scientists generally do not use the same method of sample collection, preparation, nor analysis.

The result of such variability is that the many combinations of sample preparations and chemical / mineralogical / isotopic analyses makes data compilation in a structured repository a complex process (NIU et al., 2014). Data management systems for LTG are thus like so-called "quality management systems" developed by large institutions to manage their data (RIEDL AND DUNN, 2013; U.S. NATIONAL ACADEMY OF SCIENCES ENGINEERING AND MEDICINE, 2019) in that they must facilitate different levels and types of reporting protocols (Figure 3). The result of all this complexity is proliferating approaches to data management driven by competition and different preferences among individuals, teams, projects, networks, universities, agencies, and even countries. As of October 2020, 63 data repositories were listed within the Enabling FAIR Data Project Repository Finder (https://repositoryfinder.datacite.org/) where the search term "geochemistry" was utilized.

5.3. LTG scientists often resist sharing data in data management systems.

Geochemists at the workshop stated that they want sustainable, long-term repositories for their data so that they can have accountability with funding agencies, so they can brand their data as their own, and so that they can promote use and citation of their data by other scientists and the public. But we learned that most LTG scientists do not publish their data in online data repositories, nor do they train their students in those activities. The few workshop scientists who had used repositories did it generally because they were required by journal editors or mandated by a funder. The result has been generally slow growth of geochemical databases (LEHNERT AND ALBAREDE, 2019).

Even some of the LTG scientists who had used repositories expressed resistance to the process. The reasons for such resistance within LTG in some cases is similar to resistance observed in other scientists (TENOPIR et al., 2015; BRASIER et al., 2016). For example, sometimes the resistance in LTG scientists stems from the natural tension between data providers and those who pursue meta-analysis. LTG scientists also sometimes expressed fear about loss of control of the data or possible misuse of their data by others (see, also, TENOPIR et al., 2015). Such fears were even expressed when embargoes were offered to limit the use of data for various periods of time, although embargoes can address the above concerns to some extent.

But the most commonly cited reasons for resistance to the use of data repositories were the time-consuming nature of inputting data and metadata and the related lack of a reward structure for data management. This driver of resistance is directly related to the complexity of LTG data and metadata, a complexity that is sometimes but not always shared by other data types (see also, TENOPIR et al., 2015). In most cases, data management falls on the geochemists who are completing the analyses because most geochemists do not have data managers. This may explain why, as pointed out (for ecological data) (MICHENER, 2006), "Obtaining metadata may be the most challenging aspect of data management. The investigators who collect, manipulate, perform QA [quality assurance] on, and initially analyze their particular part of the project's information … have little intrinsic incentive to take the time to formalize and structure this knowledge, except for what is needed for reports and publications."

5.4. Scientists generally have not developed standards for data and metadata in LTG, and the resulting lack of data harmonization makes use of shared datasets cumbersome.

An important result of the lack of systematic data sharing within LTG is the lack of agreement on data standards and lack of data harmonization. For example, in the USGS National Water Information System, one of the best maintained online data repositories for LTG data in the U.S., 32 different name-unit conventions are used for dissolved nitrate alone (SHAUGHNESSY et al., 2019). Only rarely within LTG have monitoring networks and government agencies imposed common standards across specific projects. Of course, the multiplicity of questions, samples and analyses, lack of agreement on data and metadata standards, and general lack of data harmonization makes data management more difficult and may contribute to selection of research with a micro-scale or local focus rather than a focus on regional or global problems where many datasets must be collated together (Figure 4). The large number of important questions that can be answered within the current framework has served the LTG community well. But the circle shown schematically in Figure 4 emphasizes that the LTG community neither prioritizes nor rewards systematic data publication in repositories and this slows the pace of research on regional or global problems.

In contrast, other communities have successfully brokered data sharing agreements (e.g., climate, biological oceanography, seismology) and best practices have been endorsed for data publication and data citation that apply across multiple domains (e.g., LEHNERT AND HSU, 2015; ESIP 2019; DATA CITATION

SYNTHESIS GROUP, 2014; STALL et al., 2019; COPDESS, 2020). Scientists within our LTG initiative hypothesized that the community does not (yet) value data standards nor harmonization enough to reward the time required for agreement and implementation of standards. If more LTG data were intended for integration with other groups' or other disciplines' datasets, or if this integration were highly valued and rewarded, then the hard work of data standardization would occur. But the development of Earth system models now demands interoperability of datasets, and LTG practitioners increasingly want to standardize and share more data.

5.5. The activities of development and maintenance of shared relational databases are highly time- and resource-consuming.

Building cyberinfrastructure that facilitates access to geochemical data along the trend shown in Figure 2 is expensive, skill-requiring, and time-consuming. The exact cost of building and maintaining datasets or data repositories depends upon the type of database. For example, although relational databases are more powerful than flat files, they are also more difficult to maintain over time. They are also less intuitive for subject-matter experts, and require more planning and documentation (CHRISTENSEN et al., 2009). In actual U.S. dollars, the annual cost of maintaining EarthChem's PetDB (Table 2) is $250,000/year, including institutional overhead at the level of 54%. This does not include resources for new developments to keep up with changing technology demands. For large, multi-investigator projects, data management can cost 20-25% of the cost of the measurements themselves (BALL et al., 2004). The costs of maintenance are at least partly related to the need to maintain utility in the face of ongoing evolution of computer hardware and software and web applications. A part of the problem is that research datasets are ever-changing, but very little money is typically available for changing data management structures or new metadata fields, etc. It is of course always possible to write code to migrate data from one system to the next. However, this also costs time and money. The costs of such activities along with the utility of some data may explain why in some cases, datasets are being prepared by commercial entities rather than through free data sharing among scientists.

All these issues are amplified because of the large number of skillsets needed in a data management team – skillsets that are generally not found in a small set of individuals. For example, information technology researchers with the skill sets to develop new cyberinfrastructure are generally less interested in maintaining old infrastructure. Furthermore, personnel managing data cyberinfrastructures must not only support the software and hardware but must also provide help to the community of users. This latter requires people with geochemical skills and very few people currently have both data management and geochemical skillsets.

5.6. Where geochemical databases have been successful, they have been focused on specific data types and have either been funded over long periods of time or organized by small groups of dedicated scientists.

A few entities have built very focused databases for geochemical data. For example, PetDB and Geochemistry of Rocks of the Oceans and Continents (GEOROC) are successful synthesis databases for petrologic data, as is the CUAHSI Hydrologic Information System (HIS) for time-series water quality data (see Supplementary Material). The first two databases exclude large sectors of materials of interest to LTG while the second database is built for time series but is not as easy to use for depth profiles of soil porewater, for example. Another successful data repository used in LTG is the USGS Produced Water Database (Table 1).

These databases and other long-term repositories (Table 1) share some attributes. First, they target only a subset of data as defined by their mission or funding: PetDB, for example, was funded by NSF's RIDGE Program to collate the geochemistry of igneous and metamorphic rocks of the ocean floor. These databases do not include the geochemistry of all rock types even though they have accepted similar geochemical data for other materials. Second, successful databases tend to receive consistent funding over many years from government agencies, private foundations, libraries, or universities, or are led by a small group of dedicated scientists (<12) who attract data from other contributing scientists.

6. What is needed for the future LTG data-scape

Publicly accessible geochemical databases accelerate collaboration among scientists and across disciplines and promote dialogue with the public (CHRISTENSEN et al., 2009; BRANTLEY et al., 2018). Without compiled datasets, very little coordinated design of data gathering strategies occurs, leaving gaps in geochemical understanding (Figure 4). Without publication of data in accessible venues, the information is not usable by communities outside of the original audience. Furthermore, the value of scientific data increases to other scientists and to the public when data can be accessed even after a given program or project is terminated and such longevity of data can be enhanced by systematic data sharing (BALL et al., 2004; CHRISTENSEN et al., 2009). As an example, background soil chemistry data from decades in the past can be used to assess pollution impacts or health risks for activities that are ongoing today (e.g. BRECKENRIDGE AND CROCKETT, 1998; U.S. NATIONAL ACADEMY OF SCIENCES ENGINEERING AND MEDICINE, 2017). On the other hand, if a decision-maker or scientist or member of the public must peruse multiple publications and web pages to pull together a dataset, or must laboriously adjust the units of a dataset because the data are not harmonized (SHAUGHNESSY et al., 2019), the time needed for such activity can limit deep analysis (LIU et al., 2020).

Each sub-section below describes a piece of what the LTG scientists who participated from the U.S. in our initiative concluded as to what is needed to move forward on this vision.

6.1. Globally unique sample identifiers

Once more LTG data are shared, the problem of ambiguity in sample identification could remain. Recognizing this, the participants in our initiative concluded that the community, funders, and journals all should require that LTG scientists use globally unique identifiers such as International Geo Sample Numbers (IGSN) (IMPLEMENTATION ORGANIZATION OF THE IGSN, 2020) or Archival Resource Keys (ARK) (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2020). By providing information about provenance, sampling time, depth and other metadata, these identifiers perform analogously to a birth certificate for a sample. Use of identifiers does not imply that the sample is archived but such identifiers might allow sample discovery if they are archived. Apps could be developed to create identifiers prior to or concurrent with sample collection, even in the field. Funding agencies could reward investigators for use of identifiers in reporting.

6.2. Publication of all data

Workshop participants concluded that all primary LTG data should be shared publicly with appropriate metadata at the time of journal publication so that data can be used by other scientific communities, other LTG scientists, and the public. This will maintain the relevance of the discipline within the context of all of Earth science as more and more Earth system models are developed. LTG journals and government publications should consider mandating this, and should similarly consider mandating that computer code be made available and linked to journal articles, reports, and data in repositories (LIU et al., 2020). This could improve documentation and error checking for both data and codes, many of which currently have little external vetting.

The workshop participants concluded that most of this LTG data should be published in online data repositories with DOIs (instead of in journal paper supplements). In that way, researchers can be evaluated efficiently for published data by peers (in peer review), by managers (in assessing salaries, promotion, tenure), and by agencies (in determining funding). Some LTG practitioners pointed out, however, that measurements produced in some process-oriented sciences are so small in volume that they do not even warrant summary in a table in a paper, let alone in a repository. Likewise, there are types of data (diffractograms, spectra, photomicrographs, wellbore logs, development-grade data such as on the left of Figure 3) for which specialized repositories do not yet exist. Publishing these small-volume or unusual data side-by-side with all explanations, interpretations, and metadata – within a journal paper or its supplement – in some cases might be better than in a repository if these data are highly likely to be mis-interpreted. The problem with this is that such data are difficult to find, let alone meta-analyze. Recognizing this, some publishers no longer accept data in supplements as part of the 'Enabling FAIR Data' movement (COPDESS, 2020).

To accomplish their goals, LTG scientists need both archived (unchanging) and versioned (modifiable and updatable) datasets. Some LTG datasets must be maintained as stationary entities (long-term archives) while others are continuously updated or corrected over time (self-described longitudinal or versioned datasets). For example, water chemistry data have been used to investigate the impact of hydraulic fracturing on groundwater (Shale Network, Table 1). When meta-analyses are published (WEN et al., 2019), the data are referenced both as a growing dataset site hosted by the CUAHSI HIS (doi:10.4211/his-data-shalenetwork), but also as a separately archived version of the dataset sampled at the time of analysis (doi:10.26208/8ag3-b743). To archive the data as a versioned dataset was not possible in the CUAHSI HIS, and so the scientists published it in their university data repository. That repository allowed archiving of a long-term copy of the data, whereas the other site showed only the entire, growing dataset. From the perspective of data producers, it is particularly important to archive the dataset analyzed in publications to ensure the reproducibility of the relevant research or modeling. On the other hand, scientists also need to update datasets and attach version numbers to evolving data. Thus, data management systems should provide curation that tracks provenance, provides versioning capabilities, and allows citations (e.g., DOIs). Such utilities could be provided in different data management systems or within one system.

6.3. Data management must be streamlined and incentivized

To break out of the circular problem shown in Figure 4, data management should be streamlined and rewarded. To streamline the management will require that LTG scientists implement best practices of data handling throughout each project. Some researchers have begun to propose such practices (THOMER et al., 2018) and some point out that efficient data and metadata management ultimately makes presentation and publication easier. Researchers should plan for data management in advance of their research. At the same time, however, funders should recognize that this requires additional funding for personnel time, hardware, or software. For larger projects, data management team members could be embedded into science teams. To enable improved data management, LTG scientists want agencies to fund the additional time and infrastructure, while protecting resources for the science itself.

Data scientists at the workshop pointed out that the use of consistent data templates pulled from existing resources or standardized analytical laboratory reports could be a cost-effective way to streamline the collection of consistent metadata. These formats could use community-defined, non-propriety data formats. The utility of creating such formats is that it can help standardize data within and outside of investigator groups and can lead toward data harmonization. Some pointed out that geochemical workflows could be supported and automatically recorded by intelligent software such as Laboratory Information Management Systems. At the same time, however, such systems can be expensive and time intensive to implement and are usually only implemented in large laboratories or for very large datasets, both of which tend to plot to the right on Figure 3.

6.4. A "bazaar" of data management systems

The participants of our initiative considered which of two realizations would be preferred for the ecosystem of data repositories for LTG. The first that was discussed was the development of one large repository, a data "superstore", for most LTG data, regardless of the country of origin, funding agency, university, sub-discipline, or investigator. For example, the LTG program at NSF could fund a data management system that was required for NSF-funded LTG science but was open to non-NSF scientists. The second scenario, a "street bazaar" for data systems, would consist of many repositories for LTG data, all differing in data volume, data type (generalized or specific), access characteristics, etc., much as shown in Table 1. Such repositories would be managed by many different entities.

In general, the first scenario was not considered to be feasible nor desirable. First, LTG datasets are already distributed among repositories across the world and within the U.S. and many data are stored in sites managed by non-US and non-NSF scientists (for example, see Table 1). Likewise, some already-functioning specialized data management systems (Table 1) could be better places for LTG data publication than a generalized NSF-branded or LTG-branded repository. Furthermore, some datasets might be well-managed in different ways in different data management systems with different data measurement protocols, promoting different types of science. For example, a critical zone observatory or a national park might host its own data repository as an example of a site-based data curation system (PALMER et al., 2017) or might be best spread across multiple repositories. Hence, multiple data repositories must be expected and should be encouraged, and a street bazaar of data management systems, scenario two, is not only inevitable but could be desirable because competition would drive improvements. Perhaps data providers will eventually choose data repositories the same way they choose journals for their publications (in consultation with the scientific community, editors, managers, and funders), establishing a hierarchy of valued repositories.

6.5. Both structured and unstructured data management systems

Within the bazaar, LTG scientists need both flexible management systems for datasets where measurement methods are less routine or still under development, and highly structured and managed data systems for datasets with established standards for measurement. Structured data systems should only be built for very large and important datasets where the measurements are more or less routine and the community agrees upon the need for and utility of the database. Two examples discussed previously manifest this finding: namely the development of a highly structured database for rock chemistry (PetDB) and the development of a highly structured database for water chemistry and other hydrological data (CUAHSI HIS). These communities had rough measurement standards and protocols already, and agreed on the utility of the data, and so they self-organized with funding from NSF and USGS respectively and developed standardized data management systems. At the LTG workshop, it was unanimously agreed that the specialized, targeted, and highly structured data repositories that are currently successful in managing data for specific communities (upper right on Figure 2) should be maintained as preferred repositories for their respective sub-disciplines (as long as their community finds them useful).

Without such agreed-upon formats and goals, other communities instead need data management systems that allow data to be stored in less structured systems that are more intuitive to subject-matter experts, generally easier for data archival, and easy to re-structure (CHRISTENSEN et al., 2009). This is largely because it can be difficult and time-consuming to format and input large volumes of metadata into structured data management systems even when they are designed specifically for an individual dataset; likewise, such data input often does not make sense for less routine data (Figure 3). Thus, funding agencies should promote development of less-structured, generalized long-term data repositories for other data types (e.g., Table 3). These repositories can host almost any kind of dataset, without any requirements about data structure. Generalized data repositories are not organized around a research question and thus can adapt as the science changes. They are instead organized by an entity (a library or university or country or funding agency, for example) or are associated with a broad scientific target topic (water, climate, etc.). Good examples that have been funded by U.S. federal agencies are CUAHSI HydroShare, EarthChem Library (described in Supplementary Material), the NASA-funded EOSDIS Distributed Active Archive Centers (DAACs, https://earthdata.nasa.gov/eosdis/daacs), the USGS Sciencebase (https://www.sciencebase.gov/catalog/), and the DOE ESS-DIVE (VARADHARAJAN et al., 2019). These generalized data repositories are not as rigid in their metadata requirements, do not provide rigorous data curation, and are simpler and more intuitive to use: these characteristics are important because of shifting reporting requirements and evolving science targets.

Of course, by definition, this second type of unstructured data storage is not as useful to some data users (Figure 2) because datasets are compiled with different characteristics. But the need for less structured data systems emerged from both the rock and water communities (see Supplementary Material) largely because of the time commitment needed for uploading of data and metadata into more structured databases. Therefore, even after the highly structured databases became successful (e.g., PetDB and CUAHSI HIS), less structured data systems that allow easier collations of data without the time-consuming input and metadata format requirements were needed. The two highly disparate communities – petrologists and water scientists – both separately discovered the need for i) structured data management systems and ii) less structured systems.

6.6. Pathways for prioritized growth of databases

Workshop participants agreed that a path must be made available to nucleate and grow specialized, targeted, and highly structured databases for specific data (e.g., PetDB, CUAHSI HIS). For example, some of these might nucleate within the generalized and unstructured data repositories (e.g., EarthChem Library, HydroShare, ESS-DIVE). Such a transition might organically occur when the volume of data reaches a critical or threshold value, when the need for the data becomes critical, or when the user base becomes large (BALL et al., 2004). Not every dataset or data type will follow this trajectory, but for a small number of datasets, funding could be made available on a competitive basis within the standard proposal format. The data systems that move all the way to the upper right on Figure 2 will likely answer specific, important, and compelling questions that enable meta-analysis for broad, enduring problems.

One intriguing mechanism for developing a specialized database is the so-called team-science or research-consortium model. In this mechanism, a group of scientists self-nucleate to compile their data into a structured database with the enticement of at least one co-authored publication. The scientific question and the publication are the focus of the effort rather than the production of a database. Thus, the benefits of data compilation are not restricted to the data user. An excellent example of such team science that is developing a structured and specialized database is the Sedimentary Geochemistry and Paleoenvironments Project (https://sgp.stanford.edu; SGP). Such efforts may be particularly successful when a limited type of data is targeted (for SGP, shale geochemistry) and when a highly dedicated group manages the effort. For such an effort to be successful, the data must answer more than one scientific question, and funding agencies must spur such groups forward. Some groups using the EarthChem Library for specialized datasets have also self-nucleated with help from the EarthChem Library team.

Where datasets are crucial enough, agencies could begin to require and reward data harmonization. Alternately, an agency could fund groups to help communities begin to broker agreed-upon reporting formats, along the lines of the community-driven strategy followed by ESS-DIVE, which involved domain experts and data scientists (http://ess-dive.lbl.gov/community-projects/). Some funders have also promoted the development of "translators" or thesauruses for controlled vocabularies used. For example, Skomos/OZCAR (https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/en/) provides lists of closely related controlled vocabulary terms and their sources with links to the source of each one. As pointed out for a related problem by SCHROEDER (2018), however, computers can help impose some harmonization but if algorithms to relate datasets are not agreed upon, then cybertools cannot solve the problem.

6.7. Certification of data repositories

The appropriate repositories in the LTG data-scape of the future could include certified sites run by a scientific organizations, publishers, government agencies, or universities. These repositories should be well supported and secure and should use file formats that ensure long-term preservation. Storing the data in a specific spreadsheet format rather than a comma-separated values (CSV) file might limit users' ability to use the data in the future if proprietary format conventions are changed. Thus, the use of non-proprietary data formats is preferred. Upon deposition in the repository, the dataset should be given a DOI for use in journal publications. In some cases, repositories will be hosted on a single server while others might be distributed data management systems (e.g., CUAHSI HIS or the NASA DAACs). These latter are also sometimes referred to as portals because they point to data that are housed on servers distributed among participants. If a data repository is available for a specific type of data, then the editor or program manager or funder should encourage (or enforce) publication in that repository.

Currently, only a few government agencies, funders, publishers, universities, or community organizations have articulated guidelines for certification of repositories (RE3DATA.ORG, 2020; THE FAIRSHARING TEAM, 2020) but participants in our initiative felt such certification is useful. For example, the USGS defines a trusted digital repository as "one whose mission is to provide reliable, long-term access to managed digital resources to its customers, now and in the future." The USGS also stipulates four criteria for a "trusted digital repository" and provides an internal certification for such repositories (https://www.usgs.gov/about/organization/science-support/office-science-quality-and-integrity/trusted-digital-repository). Specifically, the repository must 1) accept responsibility for the long-term maintenance of the material that is archived on the site; 2) be able to support not only the repository but also the digital information within the repository; 3) show "fiscal responsibility and sustainability"; 4) follow commonly accepted conventions and standards; and 5) participate in system evaluations defined by the community. Some of the repositories certified on the USGS site are run by the USGS while others are run by other entities (e.g., the Incorporated Research Institutions for Seismology or IRIS). Other data repository certification protocols are being developed, including one that currently has 16 requirements (CORETRUSTSEAL.ORG, 2020).

6.8. Better data search tools and portals

Without a superstore or designated repository for all LTG data, better tools to navigate the bazaar of data are needed. In effect, the LTG participants advocated that we change the paradigm from "build data repository, data will come" to "publish data online, cybertools will find": less money for building data repositories and more for improving the capabilities of tag and search. With this new paradigm, every data provider would put their data into a certified data repository with appropriate metadata that are tagged during upload or after (voluntarily or mandated), enabling future data discovery. Some researchers might go into datasets posted by others and tag them, just as internet users tag online photographs for Google Search, and funding agencies could reward this activity if specific data types were deemed especially important. While this shift would mean that reusability and interoperability of data would not be possible until tagging and search tools became available, the data publication process would be less onerous for the data providers, and would likely result in more data uploads with metadata. Of course, greater adoption of data standards would enable more efficient data search and discovery.

Another idea that emerged during this initiative and that would enable data discovery was that funders of LTG science should build portals to register their LTG projects, similar to the BCO-DMO portal built for oceanographic and polar projects funded by the NSF (NATIONAL SCIENCE FOUNDATION BIOLOGICAL AND CHEMICAL OCEANOGRAPHY DATA MANAGEMENT OFFICE, 2020). All projects funded through a given program would be required to register within the site and each project would be required to either upload project data to the portal site itself, or provide a link to project data in another online data management system. The portal could thus provide data management and navigation services at no cost to the program-funded projects and would promote discovery of data funded by the agency.

Funding should be prioritized for cybertools to find the data that have been placed online in trusted secure data repositories and to cross-reference samples with unique identifiers. Examples of these types of search tools are beginning to appear. In recognition of the difficulty of harvesting data from papers and supplements, for example, the NSF has funded tools to find such data (xDD, 2020). The Enabling FAIR Data Project (Repository Finder) also provides a search tool for data repositories (https://repositoryfinder.datacite.org/). (However, not all the data systems summarized in Table 1 are returned by the finder.) The Data Observation Network for the Earth (DataONE), a community project that links data repositories and provides data search functionality (https://www.dataone.org/), currently enables cross-search amongst registered member nodes using indexed metadata.

Another example is Google Dataset Search, which is built around a metadata vocabulary and codes created and maintained by Schema.org. Schema.org, only recently adapted to Earth science data through the NSF-funded EarthCube 418 (https://www.earthcube.org/p418) and 419 projects (https://www.earthcube.org/p419), provides structured vocabulary that can be used to encode metadata, keywords, and web URLs into a machine-readable format. Google Dataset Search crawls these encoded datasets, extracts metadata attributes, and catalogs them for search. The result is a catalog of datasets from many different sources, including data repositories, that can easily be searched via datasetsearch.google.com or from a more community-specific portal such as GeoCodes (e.g, https://geocodes.earthcube.org/geocodes/textSearch.html). End users in different disciplines can query and discover data across scientific domains and disciplines from a single access point. Such capabilities for dataset search would drive growth of controlled vocabularies that can be indexed.

6.9. Education in geochemical data science

All of the lessons learned and community needs suggest that the LTG community must educate students and early career researchers to promote a culture shift toward systematic data management. For example, the lack of data harmonization will only be resolved when LTG practitioners themselves develop and accept standardized formats and controlled vocabularies across their discipline. This will likely only happen if the community begins to prioritize and reward integrated databases and meta-analyses. Some educational resources are already available including training modules for data management by the USGS (U.S.G.S., 2020b) and massive open online courses on the basics of data science. In addition, one team has developed a course to educate geoscience students about the basics and advanced knowledge of data science using genuine research data and peer-reviewed research (WEN et al., 2020). Students can also attend workshops for data science at geoscience conferences offered by agencies, scientific societies, and many of the data initiatives already mentioned throughout this paper. These workshops often enable participants to gain first-hand experience in using data science for addressing geoscience questions.

7. Conclusions

The LTG community increasingly recognizes the value of data sharing but more guidance and education of the community is needed to push this recognition forward toward systematic data management. A group of LTG and data scientists from the U.S. participated in a multi-year initiative that led to advocacy for a change in paradigm from “build data repository, data will come” to “publish data online, cybertools will find”. This powerful and tractable paradigm shift will require funding agencies to work together to cross between the domains of basic science and information science. The group supported the notion that both highly structured (specialized) and less-structured (more generalized) data repositories are needed for LTG data. All of these data transformations within LTG require a new emphasis on data science for training the next generation of LTG scientists. As this data-scape emerges along with powerful cybertools for search, increasingly powerful answers to societal questions will arise.

8. Computer Code Availability

No code or software has been developed for this research.

Acknowledgements

This paper was initiated during a workshop held February 18 - 20 (2020) that was funded by the U.S. National Science Foundation (EAR 19-39257 to SLB). The workshop, where twenty-four institutions were represented, was organized by S. L. Brantley, T. Wen, D. Agarwal, and J. G. Catalano. All participants engaged in the ideas in the paper but some, not listed as coauthors, did not work on this document. Those latter participants included E. Barrera, P. Bennett, O. Harvey, R. Hazen, N. Kabengi, M. Leon, S. Morrison, C. Reinhard, J. Tang, and K. Williams. Some authors could not participate in the workshop. Four webinars were run prior to the workshop: Deb Agarwal (11/25/2019); X. Ma (12/16/2019); K. Lehnert (1/17/2020); J. Bales (1/31/2020). An earlier version of this manuscript was posted at EarthArXiv (https://eartharxiv.org/repository/view/1839/), and was sent out to 350 LTG scientists funded in the U.S. with an online survey. Twenty-seven scientists responded to the survey and 24 scientists participated in an online townhall discussion. Support for P.A.S. is acknowledged from EAR-GEO-1331846. D.A. and C.V. were funded as part of the ESS-DIVE project by the U.S. Department of Energy's Office of Science under Contract No. DE-AC02-05CH11231. Helpful reviews from two anonymous reviewers and Francis Albarede and associate editor Pierre Lanari are acknowledged.

Tables

Table 1. Subset of datasets, data portals, and libraries for low-temperature geochemists

Title	Description	Website or Citation
Alberta Geological Survey (AGS) Open Data Portal	Data related to the geology of Alberta Canada that are published by the Alberta Geological Survey.	https://geology-ags-aer.opendata.arcgis.com/
American Mineralogist Crystal Structure Database	A crystal structure database that includes every structure published in the American Mineralogist, The Canadian Mineralogist, European Journal of Mineralogy and Physics and Chemistry of Minerals, as well as selected datasets from other journals.	http://rruff.geo.arizona.edu/AMS/amcsd.php
Ameriflux	Ecosystem carbon, water, and energy fluxes.	https://ameriflux.lbl.gov/
Aqua-Mer	A database and toolkit for researchers working on environmental mercury geochemistry	https://aquamer.ornl.gov/
Atmospheric Radiation Measurement (ARM) Data Center	Data center stores data and observations of cloud and aerosol properties and their impacts on Earth's energy balance.	https://adc.arm.gov/discovery/#/
BCO-DMO (Biological and Chemical Oceanography Data Management Office)	A portal to find data and related information from research projects funded by the Biological and Chemical Oceanography Sections and the Office of Polar Programs at the U.S. National Science Foundation	https://www.bco-dmo.org/
Critical Zone Data sets	Sensor, field, and sample data for the critical zone (highly interdisciplinary).	http://criticalzone.org/national/data/datasets/
Crystallo-graphy Open Database	Crystal structures of compounds and minerals (not biopolymers).	http://www.crystallography.net/cod/
CUAHSI Hydrologic Information Systems (HIS)	Portals providing hydrologic information of different types.	https://www.cuahsi.org/data-models/portals/
CUAHSI HydroShare	Repository for hydrologic data and models that enables users to share, access, visualize, and manipulate hydrologic data types and models.	https://www.hydroshare.org
DOE ESS-DIVE	Repository for environmental data related to US DOE's Office of Science Environmental Systems Science program.	http://ess-dive.lbl.gov/
DRP (Digital Rocks Portal)	A portal to data describing porous micro-structures, especially for the fields of hydrocarbon resources, environmental engineering, and geology.	https://www.digitalrocksportal.org/
EarthChem Library	Repository for geochemical datasets (analytical data, experimental data, synthesis databases).	http://earthchem.org/library
ECOSTRESS Spectral Library	The ECOSTRESS spectral library is a compilation of over 3400 spectra of natural and human-made materials.	https://speclib.jpl.nasa.gov/
EDI (Environment-al Data Initiative)	NSF funded data portal for data from the Long-Term Ecological Research network.	https://portal.edirepository.org/nis/home.jsp
US EPA WQX	U.S. Environmental Protection Agency's water quality monitoring data from lakes,	https://www.epa.gov/waterdata/water-quality-data-wqx

Title	Description	Website or Citation
	streams, rivers, and other types of water bodies.	
GDR (Geothermal Data Repository)	Data collected from researchers funded by US Dept. of Energy Geothermal Technologies Office.	https://gdr.openei.org/
GeoReM (Geological and Environmental Reference Materials)	Max Planck Institute database for reference materials (rocks, glasses, minerals, isotopes, biological, river water, seawater).	http://georem.mpch-mainz.gwdg.de/
GEOROC (Geochemistry of Rocks of the Oceans and Continents)	Max Planck Institute database with published analyses of rocks (volcanic rocks, plutonic rocks, and mantle xenoliths).	http://georoc.mpch-mainz.gwdg.de/georoc/
Geosciences Data Repository for Geophysical Data	Collection of geoscience databases (including geochemistry) accessed by GDRIS.	http://gdr.agg.nrcan.gc.ca/gdrdap/dap/search-eng.php
GLiM (Global Lithology Map)	Database with spatial data on global lithology at a resolution of 1:3,750,000.	https://www.geo.uni-hamburg.de/en/geologie/forschung/geochemie/glim.html
Global spectral library to characterize the world's soil	Library of vis-NIR spectra for predicting soil attributes.	https://www.sciencedirect.com/science/article/pii/S0012825216300113#s2105
Global whole-rock geochemical database compilation	Compilation of >1,000,000 whole rock geochemical measurements compiled from ~13 other databases and >1,900 other sources.	https://zenodo.org/record/3359791#.X6wKb2dKjq0
GLORICH (Global River Chemistry Database)	Database with river chemistry and basin characteristics for global watersheds.	https://www.geo.uni-hamburg.de/en/geologie/forschung/geochemie/glorich.html
Handbook of the thermo-gravimetric system of minerals and its use in geological practice	Dataset of thermal properties of minerals from the Hungarian Institute of Geology.	https://mek.oszk.hu/18000/18031/18031.pdf
International Centre for Diffraction Data	Mineral and inorganic materials powder diffraction database. (behind paywall).	http://www.icdd.com
Images of Clay	A library of SEM images of clay, mostly for teaching purposes.	https://www.minersoc.org/images-of-clay.html?id=2
Karlsruhe Crystal Structure Depot (Das Kristallstrukturdepot)	A repository for crystal structures linked to publications in German journals that is run by FIZ Karlsruhe.	https://www.fiz-karlsruhe.de/en/produkte-und-dienstleistungen/das-kristallstrukturdepot
LEPR (Library of Experimental Phase Relations)	Published experimental studies of liquid-solid phase equilibria relevant to magmatic systems.	http://lepr.ofm-research.org/YUI/access_user/login.php
mindat.org	Database of mineral occurrence and general mineral properties.	https://www.mindat.org
MetPetDB	Database for metamorphic petrology.	https://tw.rpi.edu/web/project/MetPetDB
MG-RAST	DOE resource for microbial community datasets, many of which are annotated with environmental data.	https://www.mg-rast.org/
Mineral Spectroscopy Server	Data on mineral absorption spectra in the visible and infrared regions of the spectrum and Raman spectra of minerals.	http://minerals.gps.caltech.edu/FILES/Index.html

Title	Description	Website or Citation
Mössbauer spectral library	Further development of the database of the Mössbauer Effect Data Center.	http://mosstool.com/
NADP National Atmospheric Deposition Program	U.S. precipitation chemistry database, including nutrients, acids, base cations, and mercury.	http://nadp.slh.wisc.edu/
National Cooperative Soil Survey Soil Characterization Data	Includes soil chemical, physical, and mineralogical data for soil profiles across the U.S.	https://ncsslabdatamart.sc.egov.usda.gov/
National Water Quality Portal	Water quality monitoring data collected by over 400 state, federal, tribal, and local agencies.	https://www.waterqualitydata.us/
NAVDAT (North American Volcanic rock Data)	Web-accessible repository for age, chemical and isotopic data from Mesozoic and younger igneous rocks in western North America.	https://www.navdat.org/
ORNL DAAC for Biogeochem. Dynamics	Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (NASA's archive of record for Terrestrial Ecology)	https://daac.ornl.gov
PetDB	Database of geochemical data for igneous & metamorphic rocks.	https://search.earthchem.org
RRUFF Project	Database of Raman spectra, X-ray diffraction and chemistry data for minerals.	https://rruff.info/
SGP (Sedimentary Geochemistry and Paleoenviron-ments Project)	Database of shale geochemistry to answer questions about early environments on Earth	https://sgp.stanford.edu/about
Shale Network database	Water quality data in regions of shale gas development in northeastern USA.	Shale Network, 2015. doi:10.4211/his-data-shalenetwork
Skomos	Skomos manages the hierarchical vocabulary for OZCAR/Theia and has links to other thesaurus including GCMD (NASA), EnvThes (EU, eLTER), Eionet, FAO/GACS (incuding Agrovoc, Agrisemantic), ANAEE (Fr/EU), LusTRE (EU), SKOS (UNESCO).	https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/en/
SPECTRa Project (Submission, Preservation and Exposure of Chemistry Teaching and Research Data)	This project aims to disseminate primary data for chemistry from academic research laboratories.	http://www.ukoln.ac.uk/repositories/digirep/index/Deliverables#SPECTRa.html
StabisoDB	StabisoDB currently comprises $\delta^{18}O$ and $\delta^{13}C$ data of more than 67.000 macro- and microfossil samples including benthic and planktonic foraminifers, benthic and nektonic mollusks, brachiopods, and fish teeth and conodonts.	https://cnidaria.nat.uni-erlangen.de/stabisodb/

Title	Description	Website or Citation
Supplemental data for clay mineral journals	Material deposited as supplemental material from *Clays and Clay Minerals*.	http://www.clays.org/Journal/JournalDeposits.html
Tethys RDR	Open access data repository run by the Geological Survey of Austria (GBA) to publish data generated in cooperation with GBA.	https://www.tethys.at/
Theia	Array of Earth Surface datasets, including atmosphere, biosphere, cryosphere, land surface and terrestrial hydrosphere.	https://in-situ.theia-land.fr
TraceDs	Experimental studies of trace element distribution between phases.	http://traceds.ofm-research.org/access_user/login.php
USGS high resolution spectral library	The spectral library was assembled to facilitate laboratory and field spectroscopy and remote sensing for identifying and mapping minerals, vegetation, and manmade materials.	https://www.usgs.gov/labs/spec-lab/capabilities/spectral-library
USGS NWIS	Chemical and physical data for surface and groundwater in the USA.	https://waterdata.usgs.gov/nwis
USGS Produced Water Database	Chemistry of produced waters from oil and gas fields.	https://www.sciencebase.gov/catalog/item/59d25d63e4b05fe04cc235f9
VentDB	Geochemical Database for Seafloor Hydrothermal Springs funded by US NSF for data management for seafloor hydrothermal spring geochemistry.	http://www.earthchem.org/ventdb
Allard Economic Geology Collection	Collection of data and samples from >750 mines worldwide. Data includes locations, rocks, minerals, photographs, and deposit type information.	http://128.192.226.15/

Table 2. A lexicon for a few data science terms

Term	Definition as used by geochemists
Controlled vocabulary	A set of terms that are used to describe measurables so that different data providers do not identify the same observable with different nomenclature
Data curation	Inspection of data for quality, inclusion of metadata, etc. after or before it is uploaded to a repository
Data discovery	The process by which data users search, discover, collect, and evaluate the data from various sources in order to extract patterns in the data
Data harmonization	The process by which a compilation of data of the same type of measurement are re-calculated or re-normalized into the same units or species or reporting protocol so that meta-analysis of the large dataset can proceed directly from the data
Data quality	The characteristics that determine if data are fit for the purpose intended, including accuracy, relevance, accountability, reliability, and completeness[1]
Data repository	A site where multiple datasets are archived together. Data repositories can be of many types, which include general purpose repositories that accept any types of data (e.g., Figshare, Dryad), funder or institutional or national cross-domain repositories (e.g., ESS-DIVE, CUAHSI HIS), and domain-specific repositories that are theme-based (e.g., NCBI, PetDB). Repositories in the first two categories and sometimes the third typically issue DOIs. Importantly, a data repository may or may not require specific preparation, analytical methods, and/or data reporting styles.
Data set or database	A group of data values for a given project, with some metadata.
Data standards	Documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data
DOI	A unique digital object identifier that allows a researcher to find a published paper or dataset.
Distributed data system	A system where one can access data from multiple users but the data sets themselves reside on the providers' server.
FAIR principles	Findable, accessible, interoperable, reusable principles.[2]
Identifier	An alphanumeric tag for a sample that is findable online.
Interoperable	Data can be used straightforwardly with other data and in multiple workflows.
Library	A repository of examples of a specific type of data (differs from a repository in that it generally has examples of each category but not all data in one place for all categories). Depositing data into a library allows others to find the data because of its location but DOIs are generally not assigned as data are deposited.
Meta-analysis	Analyzing data collected by different investigators perhaps at different times, or in different places, and sometimes with different techniques.
Metadata	Descriptors about data that answer the questions of who? what? how? when? where?, etc.
Portal	An online site that allows a user to find many datasets.
Quality assurance of data	A management approach that focuses on implementing and improving procedures so that problems do not occur in the data.
Quality control of data	An approach that seeks to identify and correct problems in the data product before the product is published.[1]
Query	A request to find data with certain metadata characteristics (e.g., find groundwater data from Idaho).
Registration	Getting an unique identifier for a sample.
Relational database	A database that allows the user to find data related to one another by various metadata (e.g., are there data for porewater and mineralogy and organic matter for this soil horizon in this location?).
Sample	A physical entity that could be archived.
Template	Form with pre-set structure for data input.

[1] NATIONAL ACADEMY OF SCIENCES ENGINEERING AND MEDICINE (2019)
[2] WILKINSON et al. (2016)

Table 3. Examples of LTG data currently without a dedicated public database

Data type	Notes
X-ray diffractograms for specimens and reference materials	International Centre for Diffraction Data maintains a database behind a paywall
Data from LTG laboratory experiments	
Synchrotron data	
2D images (spectra, SEM photomicrographs, aerial photographs)	Some photographic, thin section, SEM, and other type libraries are available for teaching purposes (not for depositing research data)
3D datasets (computer-enhanced tomographic images, etc.)	

Table 4. Lessons learned and what LTG needs for the future data-scape

Six Lessons Learned

1. The data enterprise from measurement to meta-analysis is complex and provides multiple opportunities for error, but systematic management of data and metadata leads both to improvements in the quality of the dataset and identification of large-scale trends within the data.
2. As determined by their specific goals, LTG scientists participate in many different workflows, produce data with different structures and metadata, and make different choices with respect to how and where they publish their data, contributing to a proliferation of data management systems.
3. LTG scientists often resist sharing data in data management systems.
4. Scientists generally have not developed standards for data and metadata in LTG, and the resulting lack of data harmonization makes use of shared datasets cumbersome.
5. The activities of development and maintenance of shared relational databases are highly time- and resource-consuming.
6. Where geochemical databases have been successful, they have been focused on specific data types and have either been funded over long periods of time or organized by small groups of dedicated scientists.

Nine Needs of the LTG Community with Respect to Data Management

1. LTG scientists should use globally unique sample identifiers.
2. LTG scientists should publish all their primary data with appropriate metadata at the time of journal publication.
3. LTG scientists should streamline data management and appropriate data management should be rewarded.
4. LTG scientists need a dynamic “bazaar” of data management systems.
5. The LTG “bazaar” should include both structured and unstructured data management systems.
6. The LTG community should develop pathways to identify and develop highly structured databases that contain important data for priority questions.
7. Data management systems chosen by LTG scientists should be certified for reliable long-term access.
8. The LTG community needs to develop better data-search tools and portals that enable data discovery.
9. The LTG community must prioritize educational activities to promote geochemical data science.

Figures

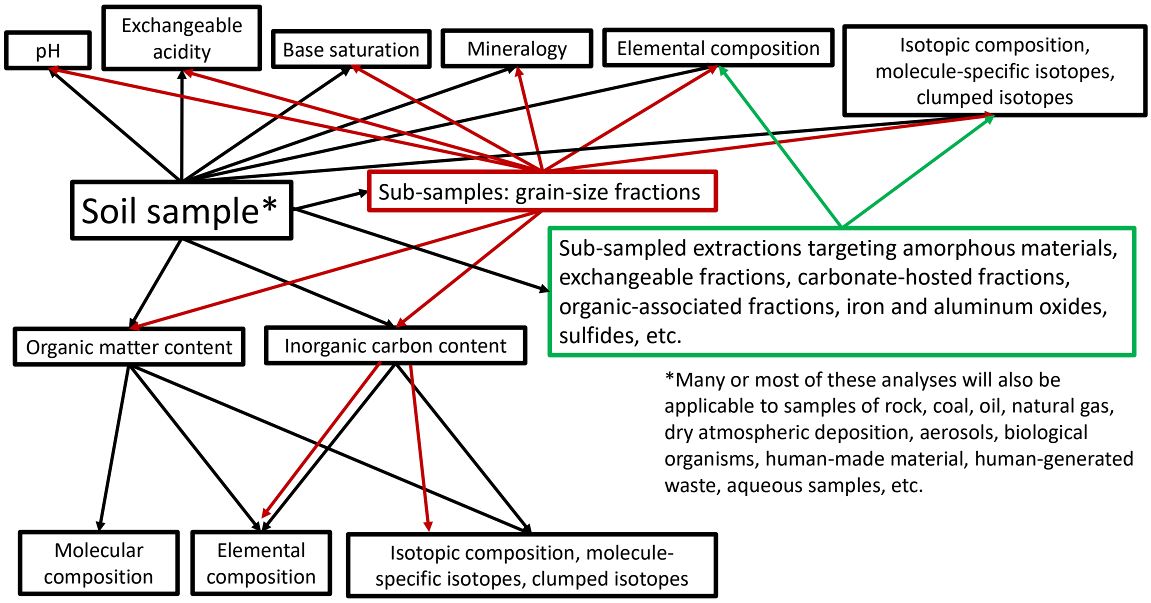
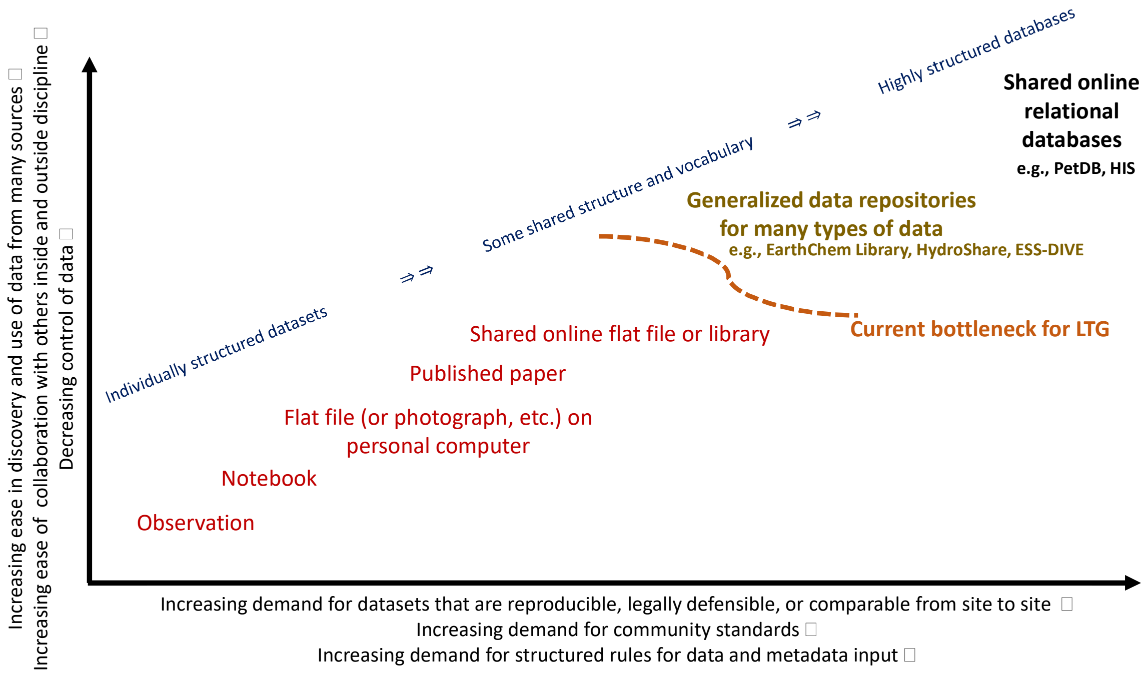


Figure 1. A schematic of different analyses and types of sub-samples or extractions that are sometimes completed on a given soil sample. Many of these would be applicable to other types of LTG samples. The schematic is shown to provide a sense of the number of analyses and sub-samples and extractions that are often completed in creating a LTG dataset, even from a single sample. The format of the data for each box could take the form of tabular data, photographs, spectra, diffractograms, etc. and the metadata associated with each box could include information about sample collection, field notes, geological and environmental details, filtration/separation/extraction/etc. details, instrumentation details, analytical details, and data processing details.



863

864 Figure 2. A schematic showing relationships among different types of management of LTG data. Data are
865 shown schematically as the pink-colored shaded area. Currently, LTG scientists need to store more data in
866 online data repositories. Only datasets that are prioritized by the community or funding agencies will be
867 stored in the most structured (and costly) repositories. Other LTG data should be deposited in generalized
868 data repositories that provide flexibility in management of data and metadata.

869

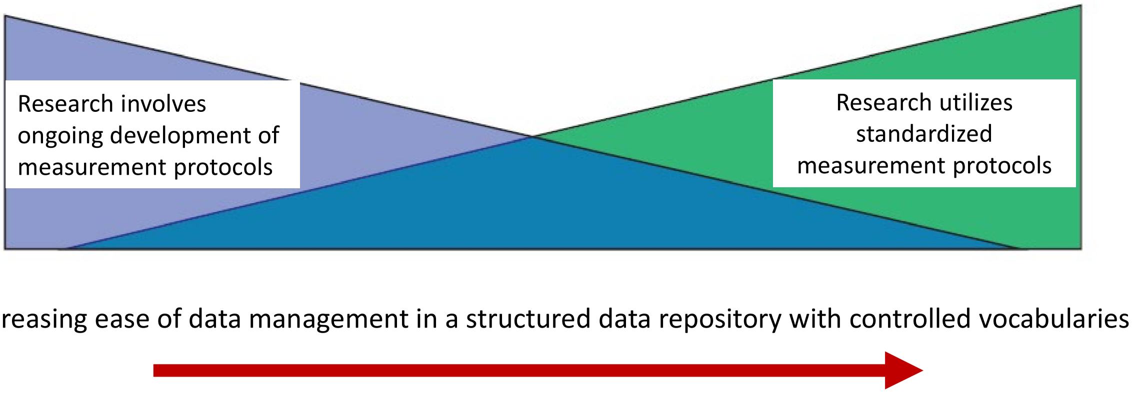


Figure 3. Schematic emphasizing how the ease of development of standardized data management protocols increases across the range from data that are highly non-routine (on the left in purple) to those that are highly routine (on the right in green). Figure adapted from a similar figure for management of data quality (Riedl and Dunn, 2013; National Academy of Sciences Engineering and Medicine, 2019).

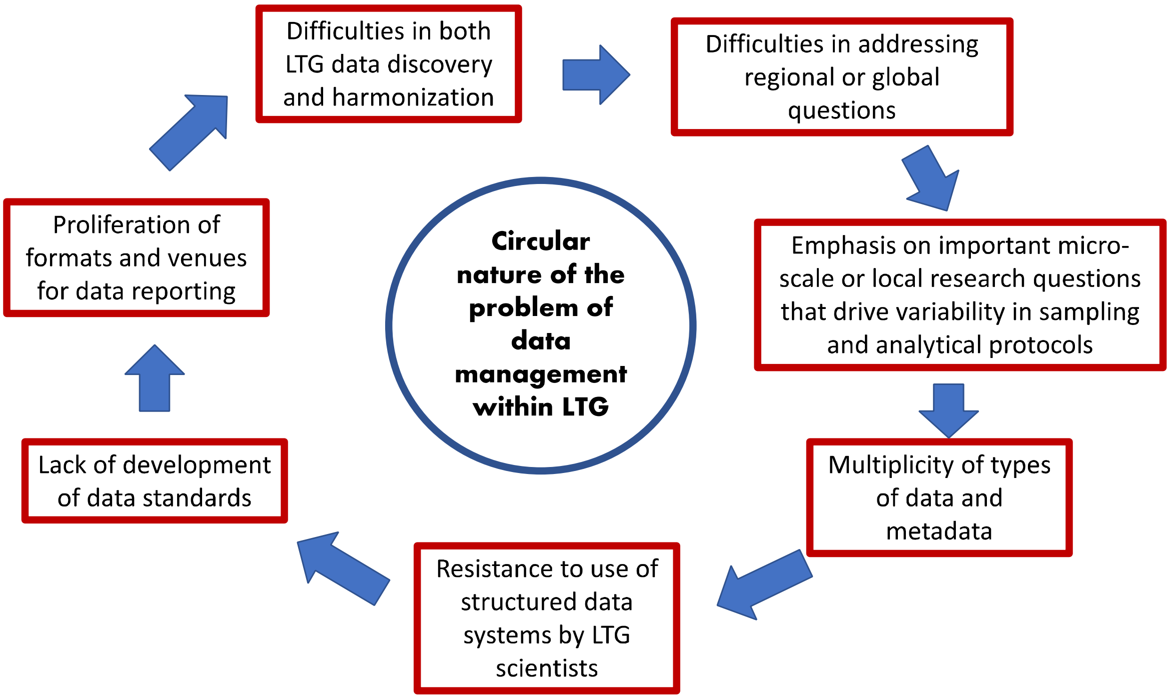


Figure 4. Summary of the circular nature of choices driving data management by LTG scientists. The culture of LTG has not established a need for data standards, data harmonization, nor data reporting, and this may impact the type of science that is completed.

References Cited

Albarede, F. and Lehnert, K., 2019. The Scientific Impact of Large Geochemical Data Sets*American Geophysical Union Fall Meeting 2019*, San Francisco, CA.

Amos, H. M., Miniat, C. F., Lynch, J., Compton, J., Templer, P. H., Sprague, L. A., Shaw, D., Burns, D., Rea, A., Whitall, D., Myles, L., Gay, D., Nilles, M., Walker, J., Rose, A. K., Bales, J., Deacon, J., and Pouyat, R., 2018. What Goes Up Must Come Down: Integrating Air and Water Quality Monitoring for Nutrients. *Environmental Science & Technology* **52**, 11441-11448.

APHA, 1998. *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association, Washington D.C.

Asch, K. and Jackson, I., 2006. Commission for the Management and Application of Geoscience Information (CGI). *Episodes* **29**, doi.org/10.18814/epiiugs/2006/v29i3/009.

Aspen Institute, 2017. Internet of Water: Sharing and Integrating Water Data for Sustainability.

Ball, C. A., Sherlock, G., and Brazma, A., 2004. Funding high-throughput data sharing. *Nature Biotechnology* **22**, 1179–1183.

Benson, B. J., Bond, B. J., Hamilton, M. P., Monson, R. K., and Han, R., 2010. Perspectives on next-generation technology for environmental sensor networks. *Frontiers in Ecology and the Environment* **8**, 193-2010; doi:10.1890/080130.

Beratan, K. K., Peer, B., Dunbar, N. W., and Blom, R., 1997. A remote sensing approach to alteration mapping: AVIRIS data and extension-related potassium metasomatism, Socorro, New Mexico. *International Journal of Remote Sensing* **18**, 3595-3609.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **363**, 1299, doi: 10.1126/science.aau0323.

Brantley, S. L., Vidic, R. D., Brasier, K., Yoxtheimer, D., Pollak, J., Wilderman, C., and Wen, T., 2018. Engaging over data on fracking and water quality. *Science* **359**, 395-397.

Brasier, K. J., Jalbert, K., Kinchy, A. J., Brantley, S. L., and Unroe, C., 2016. Barriers to sharing water quality data: experiences from the Shale Network. *Journal of Environmental Planning and Management*, dx.doi.org/10.1080/09640568.2016.1276435.

Breckenridge, R. P. and Crockett, A. B., 1998. Determination of background concentrations of inorganics in soils and sediments at hazardous waste sites. *Environmental Monitoring and Assessment* **51**, 621-656.

Brimhall, G. H. and Dietrich, W. E., 1987. Constitutive mass balance relations between chemical composition, volume, density, porosity, and strain in metasomatic hydrochemical systems: results on weathering and pedogenesis. *Geochimica et Cosmochimica Acta* **51**, 567-587.

Christensen, S. W., Brandt, C. C., and McCracken, M. K., 2009. Importance of data management in a long-term biological monitoring program. *Environmental Management* **47**, 1112-1124, doi 10.1007/s0026-010-9576-1.

Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI), 2018. CUAHSI Strategic Plan, 2018 – 2023. In: https://www.cuahsi.org/uploads/pages/img/StrategicPlan_SinglePages.pdf (Ed.).

COPDESS, 2020. Commitment Statement in the Earth, Space, and Environmental Sciences. Coalition for Publishing Data in the Earth and Space Sciences, website (https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/).

CoreTrustSeal.org, 2020. CoreTrustSeal Certified Data Repositories,. World Data System of the International Science Council and the Data Seal of Approval, https://www.coretrustseal.org/, accessed 11/8/2020.

Cousijn, H., Kenall, A. G., E. , and al., e., 2018. A data citation roadmap for scientific publishers. *Sci Data* **5**, 180259; https://doi.org/10.1038/sdata.2018.259.

Cox, S. J. D., 2011. ISO 19156:2011 Geographic information – Observations and measurements. International Organization for Standardization.

Data Citation Synthesis Group, 2014. Joint Declaration of Data Citation Principles. In: Martone, M. (Ed.). FORCE11, https://doi.org/10.25490/a97f-egyk, San Diego, CA

ESIP Data Preservation and Stewardship Committee, 2019. Data Citation Guidelines for Earth Science Data, Ver. 2. *Earth Science Information Partners web page*, https://doi.org/10.6084/m9.figshare.8441816.

Fleischer, M., 2018. Glossary of mineral species. *Mineralogical Record*.

Gil, Y., Pierce, S. A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., Hsu, L., Kinter, J., Knoblock, C., Krum, D., Kumar, V., Lermusiaux, P., Liu, Y., North, C., Pankratius, V., Peters, S., Plale, B., Pope, A., Ravela, S., Restrepo, J., Ridley, A., Samet, H., Shekhar, S., Skinner, K., Smyth, P., Tikoff, B., Yarmey, L., and Zhang, J., 2019. Intelligent systems for geosciences: An essential research agenda. *Communications of the ACM* **62**, 76-84.

Goldstein, S., Hofmann, A., and Lehnert, K., 2014. Requirements for the Publication of Geochemical Data, Version 1.0. . Interdisciplinary Earth Data Alliance (IEDA), doi.org/10.1594/IEDA/100426.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., and Fernandes Filho, E. I., 2019. Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma* **340**, 337-350.

Hemingway, J. D., Rothman, D. H., Grant, K. E., Rosengard, S. Z., Eglinton, T. I., Derry, L. A., and Galy, V. V., 2019. Mineral protection regulates long-term global preservation of natural organic carbon. *Nature* **570**, 228-231.

Hochella, J., M. F., Mogk, D., Ranville, J., Allen, I., Luther, G., Marr, L., McGrail, E. P., Murayama, M., Qafoku, N., Rosso, K., Sahai, N., Schroeder, P. A., Vikesland, P., Westerhoff, P., and Yang, Y., 2019. Natural, incidental, and engineered nanomaterials and their impacts on the Earth system. *Science*, http://dx.doi.org/10.1126/science.aau8299.

Horsburgh, J. S., Tarboton, D. G., Maidment, D. R., and Zaslavsky, I., 2011. Components of an environmental observatory information system. *Computers & Geosciences* **37**, 207-218, http://dx.doi.org/10.1016/j.cageo.2010.07.003.

International Federation of Library Associations and Institutions, 2020. Archival Resource Key (ARK), https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8793, accessed 11/8/2020.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**, 35.

Kim, H., Dietrich, W. E., Thurnhoffer, B. M., Bishop, J. K. B., and Fung, I. Y., 2017. Controls on solute concentration-discharge relationships revealed by simultaneous hydrochemistry observations of hillslope runoff and stream flow: The importance of critical zone structure. *Water Resources Research* **53**, 1424-1443.

Lehnert, K. and Albarede, F., 2019. The Scientific Impact of Large Geochemical Data Sets*American Geophysical Union*, https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/492556.

Liu, Z., Mantas, V., Wei, J., Jin, M., and Meyer, D., 2020. Creating data tool kits that everyone can use. *EOS* **101**, 25-27, doi.org/10.1029/2020EO143953.

Michener, W. K., 2006. Meta-information concepts for ecological information management. *Ecological Informatics* **1**, 3-7.

N.E.T.L., 2020. Energy Data eXchange. National Energy Technology Laboratory, U.S. Department of Energy.

N.R.C.S., 2020. National Cooperative Soil Survey United States Department of Agriculture, Natural Resources Conservation Service, https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx, accessed aa/8/2020.

National Science Foundation Biological and Chemical Oceanography Data Management Office, 2020. BCO-DMO. https://www.bco-dmo.org/.

Niu, X., Williams, J. Z., Miller, D., Lehnert, K. A., Bills, B., and Brantley, S. L., 2014. An Ontology Driven Relational Geochemical Database for the Earth's Critical Zone: CZchemDB. *Journal of Environmental Informatics* **23**, 13.

Orlowski, N., Breuer, L., and McDonnell, J. J., 2016. Critical issues with cryogenic extraction of soil water for stable isotope analysis. *Ecohydrology* **9**, 1–5, doi:10.1002/eco.1722.

Palmer, C. L., Thomer, A. K., Baker, K. S., Wickett, K. M., Hendrix, C. L., Rodman, A., Sigler, S., and Fouke, B. W., 2017. Site-based data curation based on hot spring geobiology. *Plos One* **12**, 15.

Pickering, W. F., 1981. Selective Chemical Extraction of Soil Components and Bound Metal Species, *CRC Critical Reviews in Analytical Chemistry*. CRC Press, Boca Raton, FL.

Podgorski, J. and Berg, M., 2020. Global threat of arsenic in groundwater. *Science* **368**, 845-850.

re3data.org, 2020. Registry of Research Data Repositories, https://doi.org/10.17616/R3D.

Riedl, D. H. and Dunn, M. K., 2013. Quality assurance mechanisms for the unregulated research environment. *Trends in Biotechnology* **31**, 552-554; doi.org/10.1016/j.tibtech.2013.06.007.

Ruegg, J., Gries, C., Bond-Lamberty, B., Bowen, G. J., Felzer, B. S., McIntyre, N. E., Soranno, P. A., Vanderbilt, K. L., and Weathers, K. C., 2014. Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment* **12**, 24-30; doi:10.1890/120375.

Shaughnessy, A. R., Wen, T., Niu, X., and Brantley, S. L., 2019. Three principles to use in streamlining water waulity research through data uniformity. *Environmental Science and Technology*, doi:10.1021/acs.est.9b06406.

Sprague, L. A., Oelsner, G. P., and Argue, D. M., 2016. Challenges with secondary use of multi-source water-quality data in the United States. *Water Research* **110**, 252-261, dx.doi.org/10/1016/j.watres.2016.12.024.

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**, 27-29, doi.org/10.1038/d41586-019-01720-7.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., and Dorsett, K., 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *Plos One* **10**, 24.

The FAIRsharing team, 2020. FAIRsharing.org. University of Oxford, Oxford e-Research Centre.

Thomer, A. K., Wickett, K. M., Baker, K. S., Fouke, B. W., and Palmer, C. L., 2018. Documenting provenance in noncomputational workflows: Research process models based on geobiology fieldwork in Yellowstone National Park. *Journal of the Association for Information Science and Technology* **69**, 1234-1245.

U.S. National Academy of Sciences Engineering and Medicine, 2017. *Investigative Strategies for Lead-Source Attribution at Superfund Sites Associated with Mining Activities*. The National Academies Press, doi.org/10.17226/24898, Washington D.C.

U.S. National Academy of Sciences Engineering and Medicine, 2019. *Assuring Data Quality at U.S. Geological Survey Laboratories*. The National Academies Press, http://doi.org/10.17226/25524, Washington, D.C.

U.S.G.S., 2020a. Data Citation. The United States Geological Survey (USGS).

U.S.G.S., 2020b. Data Management. United States Geological Survey, https://www.usgs.gov/products/data-and-tools/data-management/training.

U.S.G.S., 2020c. ScienceBase A U.S. Geological Survey Trusted Digital Repository. United States Geological Survey, https://www.sciencebase.gov/catalog/, accessed 11/8/2020.

Varadharajan, C., Cholia, S., Snavely, C., Hendrix, V., Procopiou, C., Swantek, D., Riley, W. J., and Agarwal, D. A., 2019. Launching an accessible archive of environmental data. *EOS, Transactions of the American Geophysical Union* **100**, https://doi.org/10.1029/2019EO111263.

Wen, T., Agarwal, A., Xue, L., Chen, A., Herman, A., Li, Z., and Brantley, S. L., 2019. Assessing changes in groundwater chemistry in landscapes with more than 100 years of oil and gas development. *Environmental Science Processes and Impacts* **21**, 384-396, doi:10.1039/c8em00385h.

Wen, T., 2020. Data Sharing, in: Encyclopedia of Big Data. Springer International Publishing, Cham, pp. 1–3. https://doi.org/10.1007/978-3-319-32001-4_322-1

Wen, T., Bandaragoda, C., and Harris, L., 2020. Data Science in Earth and Environmental Sciences. HydroLearn, https://edx.hydrolearn.org/courses/course-v1:SyracuseUniversity+EAR601+2020_Fall/about.

Wen, T., Liu, M., Woda, J., Zheng, G., and Brantley, S.L., 2021. Detecting anomalous methane in groundwater within hydrocarbon production areas across the United States. *Water Research* **200**, 117236. https://doi.org/10.1016/j.watres.2021.117236.

Wiseman, C. L. S., 2015. Analytical methods for assessing metal bioaccessibility in airborne particulate matter: A scoping review. *Analytica Chimica Acta* **877**, 9–18.

xDD, 2020. Geodeepdive, A digital library and cyberinfrastructure facilitating the discovery and utilization of data and knowledge in published documents. Geodeepdive.org, https://geodeepdive.org/about.html, accessed 11/8/2020