***American Mineralogist*—"Outlooks"—Submitted 20Jan2014**

Revision #1: Submitted 13May2014

Data-Driven Abductive Discovery in Mineralogy

Robert M. Hazen[1*]

[1]Geophysical Laboratory, Carnegie Institution of Washington,

5251 Broad Branch Road NW, Washington, D. C. 20015, USA.

ABSTRACT

Traditional pathways to discovery in Earth sciences rely in large measure on deductive and inductive approaches, by which measurements and observations are made in the context of established principles or testable, predictive hypotheses about the natural world. Vast, but largely untapped, Earth and life science data resources offer a potentially revolutionary alternative "abductive" approach to investigate Earth's co-evolving geosphere and biosphere. We therefore advocate a strategic, data-driven program for accelerated scientific discovery.

Keywords: data mining, statistical mineralogy, mineral evolution, cyberinfrastructure, philosophy of mineralogy

---

* E-mail address of corresponding author: rhazen@ciw.edu

INTRODUCTION

Discovery in Earth sciences relies to a significant degree on induction and deduction—classic approaches to reasoning that focus on the observation, modeling, and ultimately (one hopes) predictive explanations of known patterns and phenomena in nature. These powerful methods have proven successful in documenting and comprehending many aspects of the natural world, but they are inherently inefficient at discovering new complex patterns that require multivariate analysis of large datasets or synthesis of diverse types of data. Consequently, recognition of such gradual global processes as biological evolution by natural selection (Darwin 1859; Beddall 1968), continental evolution by plate tectonics (Wood 1985; Hazen 2012), atmospheric and ocean oxygenation by photosynthesis (Holland 1984; Canfield 2014), and climate change (World Meteorological Organization 1989; Weart 2008) required decades of integrated data synthesis preceding discovery and acceptance of critical Earth phenomena.

Today, the Earth and life sciences benefit from vast and ever-expanding data resources in numerous disciplines—data that for the most part serve the needs of focused communities of researchers. However, the potential now exists for a revolutionary integration and synthesis of these diverse data resources, leading to an alternative "abductive" approach to investigate Earth's co-evolving geosphere and biosphere. A growing number of Earth scientists thus advocate a systematic, data-driven quest for the accelerated discovery of hidden patterns in data resources from varied, interconnected disciplines (Fayyad et al. 1996; Hazen et al. 2011; Keller and Schoene 2012; National Science Foundation 2012; Bolukbasi et al. 2013; earthcube.org). Today's scientific enterprises generate terabytes per day of new data, yet these vast resources are woefully underutilized because they are not linked into a single platform (see, however, National Science Foundation 2012). This "Outlooks" contribution examines strategies for linking existing and new data resources, as well as methods for coupling integrated data platforms with existing statistical analysis and visualization capabilities. Thus, we envision a new kind of open-access "scientific instrument" that could transform the Earth sciences.

Deduction, Induction, and Abduction

Most philosophers of science recognize two complementary modes of logical reasoning by which many discoveries arise. By most definitions, deductive reasoning begins with a general premise that is asserted to be true, and then draws specific inferences from that generalization that must also be true. Thus:

- Earth's atmospheric oxygenation influenced the partitioning of redox-sensitive elements.
- Molybdenum, rhenium, nickel, and cobalt are redox-sensitive elements.
- Therefore, we conclude by deduction that atmospheric oxygenation must have influenced the partitioning of Mo. Re, Ni, and Co.

In deduction, specific conclusions represent a subset of the initial general premise. Studies of the partitioning of redox-sensitive elements are thus conducted in the context of well-established physical and chemical principles, and are not expected to yield surprising or anomalous results that contradict the original premise. Such efforts are critical to providing a solid foundation for scientific progress by filling in gaps in what we know we don't know, but they do not usually represent the most efficient path to discovering fundamentally new phenomena.

The complementary inductive mode of reasoning begins with observations of particular instances of a generalization, which then lead to predictions of further instances of the generalization (or to the generalization, itself). Thus:

- Each of the last 5 supercontinent cycles led to episodes of enhanced mineralization during intervals of continental convergence.
- B, Be, Hg, and Mo are mineral-forming elements.
- Therefore, we predict by induction that B, Be, Hg, and Mo minerals will display enhanced mineralization during intervals of continent convergence.

Unlike deduction, the specific predictions of induction are not necessarily contained within the initial premise and thus they cannot follow with certainty. Because one starts with instances of a generalization, and not an established premise, opportunities for discovering unexpected or anomalous patterns may be enhanced. Thus, for example, Hazen et al. (2012) found an anomalous absence of Hg mineralization during the assembly of the Mesoproterozoic Rodinian supercontinent—an anomaly that parallels emerging data from other studies (e.g., Huston et al. 2010; see below). The tradition in Earth sciences (and crime novels) of collecting data to discriminate amongst multiple working hypotheses (Chamberlin 1890) is inherently inductive in nature, and remains a powerful strategy for discovery.

Most mineralogical research is firmly grounded in deduction and/or induction. Most investigators, most of the time, start with an established deductive premise or an inductive generalization consistent with observations about known phenomenon and then collect new data to test the validity of one or more explanatory hypotheses, or to develop new hypotheses.

These deductive and inductive efforts stand in contrast to “abduction”, which is a form of logical inference that begins with the accumulation of reliable data independently of a premise or generalization. Analysis of these data, including statistical “data mining” approaches, then point to previously unrecognized patterns and correlations, and ultimately to the development of potentially new hypotheses to explain those patterns. Discoveries that lead to “paradigm shifts”—for example, James Hutton’s recognition of gradual geological change and deep time (Hutton 1795), Charles Darwin’s elucidation of evolution by natural selection (Darwin 1859), and the collective development of the concept of plate tectonics (Wood 1985)—tend to be intrinsically abductive in character, even if the initial collection of data was motivated in a deductive/inductive context. Each of these transformative discoveries required synthesis and integration of vast amounts of diverse data resources accumulated over decades to articulate a new framing of the natural world. Abduction thus provides a pathway to discovering what “we don’t know we don’t know.”

Data-Driven Discovery

For most of the history of science abduction has proven a difficult and time-consuming path to discovery. A lifetime of meticulous data collection and thoughtful synthesis, at times amplified by creative intuition or blind luck, may be required to recognize previously hidden patterns in diverse data. Only through decades of intimacy with observations, and recognition of subtle quirks and idiosyncrasies in data, will some significant patterns emerge from the noise. Such abductive discoveries do not easily come to the impatient or distracted researcher, which should serve as an important justification for the support of dedicated specialists who devote their lifetimes to a focused scientific pursuit.

The development of large and expanding data resources, coupled with powerful computation methods, has the potential to change the nature of abductive scientific discovery. Advances in cyberinfrastructure are poised to integrate data from numerous sources into semantically cohesive data platforms (Berner-Lee et al. 2001; Hey and Trefethen 2005; Fox and Hendler 2009; Hey et al. 2009; McGuinness et al. 2009; Hazen et al. 2011; Narock and Fox 2011). Furthermore, new and widely available statistical methods and visualization procedures are providing the means to interrogate these data resources in new ways and thus to tease out subtle correlations that are otherwise inherently invisible to the human brain (Card et al. 1999; Hammer et al. 2001; Peter and Shneiderman 2008; Fox and Hendler 2011; Kim et al. 2013).

Such data mining and discovery efforts that exploit large databases and enhanced data interrogation techniques to seek patterns are much in the news, particularly with respect to investment (Kovalerchuk and Vityaev 2000) and national security (Gellman and Poitras 2013) applications. In science and medicine new data resources also have the potential to reveal previously unrecognized phenomena. For example, seismological data (from nuclear test ban verification efforts), coupled with ocean floor topography, geochronology of ocean basalts, and paleomagnetism data, were critical in the discovery of patterns that elucidated mechanisms of plate tectonics (e.g., Wood 1985). Today, analyses of hidden patterns in genome databases to map viral evolution (Holmes 2007; Lam et al. 2010) and statistical exploration of medical records to find potential causal factors in pervasive diseases (Clos 2001; Berka et al. 2009) represent growing applications of abductive strategies.

Similar opportunities await the mineralogist and petrologist. Earth materials scientists have accumulated vast amounts of data on rocks, minerals, and geofluids, including their major element, minor element, and isotopic compositions; optical, electrical, magnetic, elastic, and other physical properties; atomic structures, as well as the variations of those structures with pressure, temperature, and composition; petrologic context and associated minerals; their ages; thermochemical parameters and phase relations; tectonic settings and geologic context; and even their mineral-hosted microbial ecosystems (Table 1). These data are complemented by resources on the evolution of paleoatmospheres and paleooceans, geomicrobiology, paleontology, genomics and proteomics, paleotectonics, paleomagnetism, and observations of other terrestrial planets and moons. The ultimate goal of data-driven discovery is to create a single interoperable platform that offers access to multiple, heterogeneous dimensions in a new “mineral data space.” We can envision a time when integrated data resources provide the key for discovering and understanding numerous complementary aspects of Earth’s evolution in space and time.

In spite of the promise of data-driven discovery, pitfalls abound. Data resources must be approached with a firm grounding in chemical and physical principles; an awareness of the meaning, quality, and sources of the data employed; and a keen sense of intuition. Synthesis of unreliable or biased data from varied sources may lead to false or misleading trends. For example, geochemical data employing different analytical instruments or standardization procedures may display subtle systematic differences (Pyle et al. 2002; Donovan et al. 2003). Other sources of bias reflect logistical factors: Recent studies of mineral distributions in space and time (e.g., Hazen et al. 2012; Grew and Hazen 2014) are invariably biased by the proximity of the most scrutinized deposits to major academic institutions. Therefore, any interrogation of integrated data resources must be undertaken within a framework of established deductive and inductive discovery.

Brute-Force Use Cases

Integrated data resources for abductive discovery in mineralogy do not yet exist. However, based on recent “brute-force use cases,” we can be confident that previously unrecognized patterns and correlations will emerge from the thoughtful integration and evaluation of reliable data.

Brute-force use cases involve time-consuming, manual accumulation of relevant data, either through literature searches or acquisition of new measurements. Such efforts have been undertaken in many facets of the Earth sciences. Consider two recent examples from the field of “mineral evolution”—studies of variations in the diversity and distribution of the minerals of beryllium and mercury through deep time, which demonstrate the potential of this concept as a means to recognize tectonic patterns; search for critical resources; generate insights regarding the evolution of ocean and atmospheric chemistry; and document subtle ongoing feedbacks among terrestrial life, weathering, soils, and climate.

Grew and Hazen (2014), for example, accumulated age information for 122 Be mineral localities, including the earliest known occurrences of all but 2 of the 112 known minerals in which Be is an essential element. They collated these data from examination of 300 references in 10 languages. This large dataset, assembled with minimal preconceptions about what trends might emerge, revealed 5 significant episodes of Be mineralization at approximately at 2600-2700, 1850-1750, 950-1050, 550-600, and 300 Ma—times in part associated with intervals of supercontinent assembly. Similar trends are now emerging from data-intensive studies of granite pegmatites (Tkachev 2011), as well as the ages of many thousands of individual detrital zircon crystals—large datasets that add robustness to the interpretation of episodic mineralization over at least the past 3 billion years (Valley et al. 2005; Campbell and Allen 2008; Rino et al. 2008; Hawksworth et al. 2010; Condie and Aster 2010; Condie et al. 2011; Voice et al. 2011). Additional details in the Be dataset reveal other intriguing pulses of Be mineralization, for example at ~1.3 Ga associated with extensional environments—a time not well represented in the episodic zircon record.

In a similar effort, Hazen et al. (2012) surveyed 128 mercury mineral localities, including the earliest known occurrences for 89 of the 90 known Hg species—a study that required examination and synthesis of data in more than 400 references in a dozen languages. Once again, this brute-force effort led to the discovery of previously hidden patterns. In particular, 3 unexpected results emerged:

(1) The ages of almost all Hg mineral localities correlate with 4 episodes of supercontinent assembly. Data fit to Gaussian curves at 2.69 ± 0.04, 1.81 ± 0.05, 0.53 ± 0.05, and 0.32 ± 0.07 Ga correlate with the assemblies of Kenorland, Nuna, Pannotia, and Pangaea, respectively—patterns consistent with those observed for zircon, molybdenite, and the minerals of Be and B.

(2) An as yet unexplained billion-year gap in Hg mineralization occurred between 1.8 and 0.8 billion years, an interval that included assembly of the supercontinent of

Rodinia. This interval may correlate with the innovation of microbial Hg methylation, perhaps coupled with changes in ocean chemistry (Canfield 1998). Alternatively, these data on Hg minerals may contribute to evidence that the tectonic setting of Rodinian assembly differed from that of other supercontinents (Cawood et al. 2009; Huston et al. 2010).

(3) The largest known Hg deposits from ~0.3 Ga are coeval with Carboniferous coal measures, suggesting co-burial of organic carbon and Hg, with subsequent hydrothermal mobilization and re-deposition. Thus, the mineralization of Hg—a rare element with no biological function—is now coupled to the evolving terrestrial biosphere.

This study was based entirely on published and web resources, yet it consumed more than 1 person-year of effort, mostly devoted to locating and evaluating previously published data in sometimes obscure sources, as well as integrating those data with lists of minerals approved by the International Mineralogical Association (Downs, 2006; rruff.info/ima) and Hg locality information (principally in mindat.org).

These abductive discoveries, though focused on single rare elements with relatively few localities, demonstrate the untapped potential for a new strategy of discovery based on development and mining of enhanced data resources. Such brute-force studies of Earth's near-surface rocks and minerals are by no means limited to the temporal occurrence and aerial distribution of mineral species. For example, Faquahar et al. (2000, 2001, 2007) collated extensive data on sulfur isotopic fractionation in varied lithologies versus age. They found remarkable mass-independent effects, a presumed consequence of upper atmosphere photolysis of sulfur compounds, that are largely constrained to formations > 2.25 Ga. They ascribed this finding to enhanced ozone shielding—a conclusion heralded as the “smoking gun” for the timing of the Great Oxidation Event and its irreversible transformation of atmospheric chemistry (Canfield 2014).

Large-scale community data resources, such as EarthChem/PetDB (Lehnert et al. 2000, 2007), are especially relevant for discoveries in statistical petrology, geochemistry, and mineralogy (see http://www.earthchem.org/citations/petdb for examples). In one such effort, Keller and Schoene (2012) employed a database of 70,000 analyses of continental igneous rocks to discover evidence for significant lithospheric disruption at ~2.5 Ga—a time just prior to the Great Oxidation Event. Their comprehensive overview of secular variations in major and incompatible elements in basalt reveals a significant decrease in mantle melt fraction at that time—a trend not obvious without a large and relatively unbiased data set. Keller and Schoene (2012) concluded that atmospheric oxidation may be linked in part to redox changes associated with crustal evolution. The availability of large-scale, community supported, persistent, and quality-controlled data resources is critical to the success of such endeavors.

Accumulations of mineral data also point to Earth’s gradual subsurface oxidation. Golden et al. (2013) gathered new and published trace element analyses of the rhenium content of 422 molybdenite (MoS_2) specimens from 135 localities with known ages from 2.91 billion years to 6.3 million years. Rhenium is a redox-sensitive element that is mobilized in its Re^{7+} form only under relatively oxidized subsurface conditions. This brute-force data effort revealed two statistically significant trends: (1) Systematic increases in average and maximum trace concentrations of Re in molybdenite since 3.0 Ga point to enhanced oxidative weathering by subsurface fluids, and (2) episodic molybdenum mineralization correlates with five intervals of supercontinent assembly from ~2.7 Ga (Kenorland) to 300 Ma (Pangaea).

These and other examples demonstrate that brute-force methods have the potential to reveal hidden correlations; numerous other trends in published mineralogical data are undoubtedly awaiting discovery. However, brute-force data recovery methods are inherently time-consuming and correspondingly inefficient. A far better strategy is to develop and further enhance community supported data resources.

A Strategy for Data-Driven Discovery in Mineralogy

A strategy for data-driven discovery in mineralogy is emerging (Hazen et al. 2013). The first steps relate to sustainable maintenance, expansion, and quality control of existing databases for rocks, minerals, and other Earth materials (Table 1). The most basic need and responsibility of any natural history discipline is the accurate, timely, comprehensive, and accessible archiving of data on species. No task is more fundamental to the long-term stability and integrity of a field, nor should such data management be left to the unfunded good will of individuals, no matter how skilled and well intentioned they may be. Some data resources such as EarthChem and Volcanoes (Table 1) enjoy significant, if not guaranteed long-term, Federal support. However, it is astonishing that the official web-based list of approved mineral species (rruff.info/ima), which is freely available and widely used by the international community, has no long-term institutional home or financial support. And, until recently, the server for LEPR and MELTS—widely used open-access resources for thermochemistry—resided in the bedroom of its founder, Mark Ghiorso. The international Earth materials community, therefore, needs to initiate action on several fronts:

(1) Encourage professional journals to adopt policies and collaborations by which all newly published data will be deposited in approved, sustainable, quality-controlled, open-access sources.

(2) Endorse the International Geo Sample Number (IGSN) procedures of the System for Earth Sample Registration (Lehnert and Klump 2008; www.geosamples.org) and require use of this system before publication.

(3) Continue to encourage and support efforts to transfer previously published data into open-access repositories.

(4) Identify and encourage recovery of "dark data" resources, including unpublished hard copy and electronic format data accumulated by individuals. Encourage publication of these data resources through electronic data journals with digital object identifier (doi) information.

(5) Create active and engaged user communities to ensure quality control of data resources, which must be properly vetted prior to incorporation into open-access sources.

(6) Establish data publication procedures and data citation policies that ensure proper credit and motivation for data producers.

(7) Identify and exploit sources of funding for work being done now and for long-term institutional support of critical data resources.

Note that several of these steps will require both institutional and cultural changes within the Earth sciences community. Effective change will take time and effort, but we can anticipate a time when the larger Earth materials community recognizes the critical importance of shared, high-quality data resources and underscores the responsibility of all researchers to contribute to this infrastructure.

The second facet of fostering data-driven discovery in Earth materials research involves integrating existing databases into a larger Earth Materials Data Infrastructure (Hazen et al. 2013). Ultimately, we can envision linking separately maintained data resources into a federated, centrally-governed data framework in which diverse data resources are semantically compatible and linked (see, for example, National Science Foundation 2012; earthcube.org). Eventually, a more expansive opportunity lies in the integration of Earth materials data with other complementary disciplines, including paleobiology, proteomics, paleotectonics, and planetary sciences.

To accomplish this vision, we need to identify and integrate key data resources, while providing computational tools that can be used to select, analyze, and visualize data. Ultimately, a comprehensive Earth materials data infrastructure could be linked to artificial intelligence and machine learning capabilities to accelerate data-driven discovery. Cyberinfrastructure programs such as EarthCube, which enjoy significant community support as well as Federal funding, are moving scientific research in these bold new directions. However, the mineralogy-petrology research community needs to increase its commitment to these efforts if we are to take maximum advantage of emerging opportunities.

Conclusions

Two significant and novel impacts are likely to result from a program of abductive discovery in mineralogy. First, unanticipated mineralogical discoveries will be made. New phenomena related to mineral crystal chemistry, trace element and isotope distributions, mineral associations and geologic context, fluid-rock interactions, and interactions with the biosphere, all in the framework of geological space and time, are certain to emerge. The abductive approach, coupled with more traditional hypothesis-driven inquiries, will inevitably lead to discovery of new patterns in nature.

Second, and more significant for the future of mineralogy and petrology, data-driven discovery provides a model for 21^{st}-century science that explicitly recognizes the power of abduction and exploits opportunities represented by the explosion of Earth science data. Such a strategy in no way subsumes deduction and induction; indeed, the abductive approach explicitly relies upon the accumulation of traditional measurements and observations. Abduction amplifies the vast amounts of data inspired by deductive and inductive discovery. Ultimately, when data from petrology, mineralogy, geochemistry, paleontology, geodynamics, proteomics, irreversible thermodynamics, and geochronology are integrated with newly adapted statistical analysis and visualization capabilities, we will enjoy a wholly new kind of “scientific instrument”—an open-access engine of discovery that could transform the Earth sciences.

Acknowledgements

Robert Downs and Peter Fox provided invaluable feedback on an earlier draft of this contribution, portions of which were adapted and modified from a report (12 April 2013) submitted to MSA Council by the Ad Hoc Committee on Earth Materials Data

(Members: Kerstin Lehnert, Chair; Elizabeth Cottrell, Robert Downs, Peter Fox, Mark Ghiorso, Robert Hazen, Surendra Saxena, Frank Spear). I thank John Armstrong, Robert Downs, Paul Falkowski, Andrew Knoll, and Dimitri Sverjensky for constructive advice during this article's preparation, and Mark Ghiorso, Kerstin Lehnert, Keith Putrika, and two anonymous reviewers for detailed and thoughtful reviews of the manuscript. Financial support for data infrastructure development has been provided by the Deep Carbon Observatory, the Alfred P. Sloan Foundation, the National Science Foundation, and the Carnegie Institution of Washington.

References

Alroy, J. (2010) The shifting balance of diversity among major animal groups. Science, 329, 1191-1194.

Alroy, J., Aberhan, M., Bottjer, D.J., Foote, M., Fürsich, F.T., Harries, P.J., Hendy, A.J.W., Holland, S.M., Ivany, L.C., Kiessling, W., Kosnik, M.A., Marshall, C.R., McGowan, A.J., Miller, A.I., Olszewski, T.D., Patzkowsky, M.E., Peters, S.E., Villier, L., Wagner, P.J., Bonuso, N., Borkow, P.S., Brenneis, B., Clapham, M.E., Fall, L.M., Ferguson, C.A., Hanson, V.L., Krug, A.Z., Layou, K.M. Leckey, E.H., Nürnberg, S., Powers, C.M., Sessa, J.A., Simpson, C., Tomasovych, A., and Visaggi, C.C. (2008) Phanerozoic trends in the global diversity of marine invertebrates. Science, 321, 97-100.

Beddall, B. G. (1968). Wallace, Darwin, and the theory of natural selection. Journal of the History of Biology, **1,** 261–323.

Berka, P., Rauch, J., and Djamel, A.Z. [Editors] (2009) Data Mining and Medical Knowledge Management: Cases and Applications. Hershey, Pennsylvania: IGI Global.

Berner-Lee, T., Hendler, J., and Lassila, O. (2001) The semantic web. Scientific American, 284(5), 34-43.

Bolukbasi, B., Berente, N., Cutcher-Gershenfeld, J., Dechurch, L., Flint, C., Haberman, M., King, J.L., Knight, E., Lawrence, B., Masella, E., McElroy, C., Mittleman, B., Nolan, M., Radik, M., Shin, N., Thompson, C.A., Winter, S., Zaslavsky, Allison, M.L., Arctur, D., Arrigo, J., Aufdenkampe, A.K., Bass, J., Crowell, J., Daniels, M., Diggs, S., Duffy, C., Gil, Y., Gomez, B., Graves, S., Hazen, R., Hsu, L., Kinkade, D., Lehnert, K., Marone, C., Middleton, D., Noren, A., Paerthree, G., Ramamurthy, M., Robinson, E., Percivall, G., Richard, S., Suarez, C., and Walker, D. (2013) Open data: Crediting a culture of cooperation. Science, 342, 1041-4042. DOI:10.1126/science.342.6162.1041-b

Campbell, I.H., and Allen, C.M. (2008) Formation of supercontinents linked to increases in atmospheric oxygen. Nature Geoscience, 1, 554-558.

Canfield, D.E. (1998) A new model for Proterozoic ocean chemistry. Nature, 396, 450-453.

Canfield, D.E. (2014) Oxygen: A Four Billion Year History. Princeton, NJ: Princeton University Press.

Card, S.K., Mackinlay, J.D., and Shneiderman, B. (1999) Reading in Information Visualization: Using Vision to Think. San Francisco, California: Morgan Kaufmann.

Cawood, P.A., Kroner, A., Collins, W.J., Kusky, T.W., Mooney, W.D., and Windley, B.F. (2009) Accretionary orogens through Earth history. Geological Society [London] Special Publication, 318, 1−36.

Chamberlin, T.C. (1890) The method of multiple working hypotheses. Science, 15, 92-96.

Clos, K.J. [Editor] (2001) Medical Data Mining and Knowledge Discovery. Dortrecht, Netherlands: Springer.

Condie, K.C., and Aster, R.C. (2010) Episodic zircon age spectra of orogenic granitoids: The supercontinent connection and continental growth. Precambrian Research, 180, 227-236.

Condie, K.C., Bickford, M.E., Aster, R.C., Belousova, E., and Scholl, D.W. (2011)

Episodic zircon ages, Hf isotopic composition, and the preservation rate of continental crust. Geological Society of America Bulletin, 123, 951-957.

Darwin, C. (1859) On the Origin of Species. London: John Murray.

Donovan, J.J., Hanchar, J.M., Picollo, P.M., Schrier, M.D., Boatner, L.A., and Jarosewich, E. (2003) A Re-examination of the rare-earth-element orthophosphate standards in use for elecrtron-microprobe analysis. Canadian Mineralogist. 41, 221-232.

Downs, R.T. (2006) The RRUFF Project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. Program and Abstracts of the 19th General Meeting of the International Mineralogical Association in Kobe, Japan. O03-13.

Downs, R.T., and Hall-Wallace, M. (2003) The American Mineralogist Crystal Structure Database. American Mineralogist, 88, 247-250.

Fox, P., and Hendler, J. (2009) Semantic eScience: Encoding meaning in next-generation digitally enhanced science. In Hey, T., Tansley, S., Tolle, K. [Editors] The Fourth Paradigm: Data-Intensive Scientific Discovery. Redland, WA: Microsoft External Research, pp. 145-150.

Farquhar, J., Bao, H., and Thiemens, M.H. (2000) Atmospheric Influence of Earth's Earliest Sulfur Cycle. Science, 289, 756-758.

Farquhar, J., Savarino, I., Airieau, S., and Thiemens, M.H. (2001) Observations of wavelength- sensitive, mass-independent sulfur isotope effects during SO_2 photolysis: Implications for the early atmosphere. Journal of Geophysical Research, 106, 1-11.

Farquhar, J., Peters, M., Johnston, D.T., Strauss, H., Masterson, A., Wiechert, U., and Kaufman, A.J. (2007) Isotopic evidence for mesoarchean anoxia and changing atmospheric sulphur chemistry. Nature, 449, 706-709.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) From data mining to knowledge discovery in databases. AI Magazine, Fall 1996, 37-54.

Fox, P., and Hendler, J. (2011) Changing the equation on scientific data visualization. Science, 331, 705-708.

Gellman, B. and Poitras, L. (2013). U. S. intelligence mining data from nine U.S. Internet companies in broad secret program. The Washington Post, June 7, 2013, A1.

Ghiorso, M.S., and Sack, R.O. (1995) Chemical Mass Transfer in Magmatic Processes IV. A revised and internally consistent thermodynamic model for the interpolation and extrapolation of liquid-solid equilibria in magmatic systems at elevated temperatures and pressures. Contributions to Mineralogy and Petrology, 119, 197-212.

Ghiorso, M.S., Hirschmann, M.M., Reiners, P.W., and Kress, V.C. III (2002) The pMELTS: A revision of MELTS for improved calculation of phase relations and major element partitioning related to partial melting of the mantle to 3 GPa. Geochemistry Geophysics Geosystems, 3, 10.1029/2001GC000217.

Grazulis, S., Daskevic, A., Merkys, A., Chateigner, D., Lutterotti, L., Quiros, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., and Le Bail, A. (2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platforms for world-wide collaboration. Nucleic Acids Research, 40, D420-D427. doi:10.1093/nar/gkr900

Grew, E.S., and Hazen, R.M. (2014) Beryllium mineral evolution. American Mineralogist, in press.

Hammer, Ø., Harper, D.A.T., and Ryan, P.D. (2001) PAST: Paleontological statistical software package for education and data analysis. Paleo-Electronica.org, issue 1.

Hawkesworth, C.J., Dhuime, B., Pietranik, A.B., Kemp, A.I.S., and Storey, C.D. (2010) The generation and evolution of continental crust: Journal of the Geological Society, 167, 229-248.

Hazen, R.M. (2012) The Story of Earth. New York: Viking-Penguin.

Hazen R.M., Bekker A., Bish D.L., Bleeker W., Downs R.T., Farquhar J., Ferry J.M., Grew E.S., Knoll A.H., Papineau D., Ralph J.P., Sverjensky D.A., and Valley J.W. (2011) Needs and opportunities in mineral evolution research. American Mineralogist, 96, 953-963.

Hazen, R.M., Golden, J., Downs, R.T., Hystad, G., Grew, E.S., Azzolini, D, and Sverjensky, D.A. (2012) Mercury (Hg) mineral evolution: A mineralogical record of supercontinent assembly, changing ocean geochemistry, and the emerging terrestrial biosphere. American Mineralogist, 97, 1013-1042.

Hazen, R.M., Cotrell, E., Downs, R.T., Fox, P., Ghiorso, M., Lehnert, K., Saxena, S., and Spear, F. (2013) Report of the Chair of the Ad Hoc Committee on Earth Materials Data, To the First 2013 Mineralogical Society of America Council Meeting, 23 April 2013, 6 pp.

Hey, T., and Trefethen, A.E. (2005) Cyberinfrastructure for e-Science. Science, 308, 817-821.

Hey, T., Tansley, S., and Tolle, K. [Editors] (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Redland, WA: Microsoft External Research.

Holland, H.D. (1984) The Chemical Evolution of the Atmosphere and Oceans. Princeton, NJ: Princeton University Press.

Holland, T.J.B., and Powell, R. (1998) An internally consistent thermodynamic data set for phases of petrological interest. Journal of Metamorphic Petrology, 16, 309–343.

Holmes, E.C. (2007) Viral evolution in the genomic age. PLOS Biology, DOI: 10.1371/journal.pbio.0050278

Huston, D.L., Pehrsson, S., Eglington, B.M., and Zaw, K. (2010) The geology and metallogeny of volcanic-hosted massive sulfide deposits: Variations through geologic time and with tectonic setting. Economic Geology, 106, 571-591.

James Hutton (1795) Theory of the Earth; with Proofs and Illustrations. Edinburgh: Creech. 2 volumes.

Keller, B., and Schoene (2012) Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gya ago. Nature, 485, 490-493.

Kim, J.D., Senn, S., Harel, A., Jelen, B.I., and Falkowski, P.G. (2013) Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases. Philosophical Transactions of the Royal Society, B 368, 20120257.

Kovalerchuk, B. and Vityaev, E. (2000) Data Mining in Finace: Advances in Relational and Hybrid Methods. New York: Kluwer Academic Publishers.

Lam, T.T., Hon, C.C., and Tang, J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. Critical Reviews in Clinical Laboratory Sciences, 47, 5–49.

Lehnert, K.A., and Klump, J. (2008) Facilitating research in mantle petrology with geoinformatics. 9th International Kimberlite Conference Extended Abstracts, 91KC-A-00250.

Lehnert, K.A., Su, Y., Langmuir, C.H., Sarbas, B., and Nohl, U. (2000) A global geochemical database structure for rocks. Geochemistry Geophysics Geosystems 1.

Lehnert, K.A., Walker, D., and Sarbas, B. (2007) EarthChem: A geochemistry data network. Geochimica et Cosmochimica Acta, 71, A559.

McGuinness, D.L., Fox, P.A., Brodaric, B., and Kendall, E. (2009) The emerging field of semantic scientific knowledge integration. IEEE Intelligent Systems, 24, 25-26.

Mutschler, F.E., Rougon, D.J., Lavin, O.P., and R.D. Hughes (1981) PETROS version 6.1 Worldwide Databank of Major Element Chemical Analyses of Igneous Rocks. National Geophysical Data Center, NOAA. doi:10.7289/V5QN64NM.

Narock, T., and Fox, P.A. (2011) From science to e-science to semantic e-science: A heliophysics case study. Computers & Geosciences. doi: 10.1016/j.cageo.2011.11.018

National Science Foundation (2012) A Community Roadmap for Earthcube Data: Discovery, Access, and Mining. Arlington, Virginia: National Science Foundation, 38 p.

Perer, A., and Shneiderman, B. (2008) Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. ACM Conference on Human Factors in Computing Systems, Florence, Italy.

Pyle, J.M., Spear, F.S., and Wark, D.A. (2002) Electron microprobe analysis of REE in apatite, monazite and xenotime: Protocols and pitfalls. Reviews in Mineralogy and Geochemistry, 48, 337-362.

Rino, S., Kon, Y., Sato, W., Maruyama, S., Santosh, M., and Zhao, D. (2008) The Grenvillian and Pan-African orogens: world's largest orogenies through geologic time, and their implications on the origin of superplumes. Gondwana Research, 14, 51-72.

Sepkowski, D. (2012) Towards "A natural history of data": Evolving practices and epistemologies of data in paleontology, 1800-2000. Journal of the History of Biology. DOI 10.1007/s10739-012-9336-6

Stixrude, L., and Lithgow-Bertelloni, C. (2005) Thermodynamics of mantle minerals – I. Physical properties. Geophysical Journal International, 162, 610–632.

Stixrude, L., and Lithgow-Bertelloni, C. (2011) Thermodynamics of mantle minerals – II. Phase equilibria, Geophysical Journal International, 184, 1180–1213

Tkachev, A.V. (2011) Evolution of metallogeny of granitic pegmatites associated with orogens throughout geological time. Geological Society of London, Special Publications 350, 7-23.

Valley, J.W., Lackey, J.S., Cavosie, A.J., Clechenko, C.C., Spicuzza, M.J., Basei, M.A.S., Bindeman, I.N., Ferreira, V.P., Sial, A.N., King, E.M., Peck, W.H., Sinha, A.K., and Wei, C.S. (2005) 4.4 billion years of crustal maturation: oxygen isotope ratios of magmatic zircon. Contributions to Mineralogy and Petrology, 150, 561-580.

Voice, P.J., Kowalewski, M., and Eriksson, K.A. (2011) Quantifying the timing and rate of crustal evolution: Global compilation of radiometrically dated detrital zircon grains. The Journal of Geology, 119, 109-126.

Weart, S.R. (2008) The Discovery of Global Warming: Revised and Expanded Edition. Cambridge, Massachusetts: Harvard University Press.

Wood, R.M. (1985) The Dark Side of the Earth. London: George Allen and Unwin.

World Meteorological Organization (1989) The Changing Atmosphere: Implications for Global Security, Toronto, Canada, 27-30 June 1988: Conference Proceedings. Geneva: Secretariat of the World Meteorological Organization.

Table 1. Selected open-access data resources for Earth materials research.

Web address	**Content**	**ref**
rruff.info	Mineral species and properties	1
rruff.info/ima	IMA official mineral list	
rruff.geo.arizona.edu/AMS/amcsd.php	Mineral crystal structures	2
smmp.net/IMA-CM/ctms.htm	IMA list of type minerals	
www.crystallography.net	Crystal structure data	3
http://cod.iutcaen.unicaen.fr	Powder diffraction data	
http://database.iem.ac.ru/mincryst/	Mineral crystal structures	
mindat.org	Mineral localities/associations/properties	
webmineral.com	Mineral species/properties	
athena.unige.ch/athena/	Mineral species/properties	
http://abulafia.mt.ic.ac.uk/shannon/ptable.php	Ionic radii table	
http://minerals.gps.caltech.edu/FILES/raman/Caltech_data/index.htm	Spectroscopic data	
earthref.org	geochemistry/geomagnetism data	
geokem.com	igneous rock chemistry	
georoc.mpch-mainz.gwdg.de	rock geochemistry	
metpetdb.rpi.edu	metamorphic petrology	
navdat.org	igneous rocks of North America	
earthchem.org	geochemistry, geochronology, petrology	
earthchem.org/petdb	petrology	

ngdc.noaa.gov/mgg/geology/petros.html	igneous rock geochemistry	4
http://volcano.si.edu	volcanoes and eruptions	
iza-structure.org/databases/	zeolite crystal structures	
melts.ofm-research.org	thermodynamic modeling	5,6
lepr.ofm-research.org	experimental data	
metamorph.geo.uni-mainz.de/thermocalc/	thermochemical modeling	7
phaseplot.com/Phase_Plot/Contents.html	mantle phase equilibria modeling	8
vamps.mbl.edu/portals/deep_carbon/cdl.php	subsurface microbial ecosystems	

1. Downs (2006); 2. Downs and Wallace (2003); 3. Grazulis et al. (2012); 4. Mutschler et al. (1981); 5. Ghiorso and Sack (1995); 6. Ghiorso et al. (2002); 7. Holland and Powell (1998); 8. Stixrude and Lithgow-Bertelloni (2005, 2011).