Research Program
Phase II Proposal Instructions

Cover Page

Please submit this information as the cover to your proposal

Please choose the area of emphasis:

Medical; or
x **Science and Engineering**

Applicant Institution: Carnegie Institution of Washington

Project Title: The Co-Evolution of the Geo- and Biospheres: An Integrated Program for Data-Driven Abductive Discovery in the Earth Sciences

Project Time Period: (from 1/1/2015 to 12/31/2017)

Project Leader

Name/Title: Robert M. Hazen, Senior Staff Scientist, Geophysical Laboratory, and
Executive Director, Deep Carbon Observatory
Mailing Address: Carnegie Institution of Washington, Geophysical Laboratory, 5251 Broad Branch Road, Washington, DC 20015
Phone Number: 202-478-8962
Fax Number: 202-478-8901
Email: rhazen@ciw.edu

Research Program
Project Overview

Organization: Geophysical Laboratory, Carnegie Institution of Washington

Project Objectives/Aims	Implementation Timeline
Examples: *1. Use high-resolution confocal fluorescence microscopy for the measurement of molecular and bulk fluid velocities in nanoscale fluidic channels* *2. Investigate the use of externally applied "gate voltages" for controlling nanoscale fluid transport* *3. Apply the above in developing new, highly efficient technologies for biomolecular separations*	*Examples:* *1. Install requested microscopy equipment, hire technician and recruit research personnel in first four months* *2. Aim 1: Develop two-photon confocal microscopy techniques and image analysis methodologies in months four through twelve; continue fine-tuning in years 2 and 3* *3. Aim 2: Demonstrate use of external voltages to control thickness of electrical double layers in months eight through twelve; exploit this control to create electronic fluid control switches in year 2* *4. Evaluation: Final assessment by external project advisory committee at end of year 3*
Task 1: Deep-Time Data Resource Development A. Add geochronological data to deep-time mineral databases. B. Construct a Precambrian paleobiology database. C. Incorporate qualitative age information for proteins, pdb.org D. Add information on paleotectonic setting to mineralogy and petrology data resources.	A. Add ages to 5000 mineral localities per year; years 1, 2, & 3. B. Add a comprehensive tabulation of Precambrian fossils into paleobiodb.org; years 2 & 3. C. Determine relative ages for 5 metabolic enzyme groups per year; years 2 & 3. D. Develop numerical key for varied tectonic settings in year 1; enter data for 1000 localities in year 1; 2000 localities in years 2 and 3.
Task 2: Deep-Time Data Infrastructure Development A. Link petrology, mineralogy, & geochemistry data resources. B. Link data infrastructure to protein structure database. C. Link data infrastructure to paleobio.org database. D. Link data infrastructure to thermochemical modeling. E. Provide open access and disseminate protocols for the Deep-Time Data Infrastructure.	A. Develop and test linked Earth materials data infrastructure; year 1. B. Develop and test above linked to protein data resources; year 2. C. Develop and test above linked to paleobio.org; year 3. D. Develop and test above linked to thermochemical modeling; year 3. E. Implement open access to the Deep-Time Data Infrastructure and promote its use through lectures and publications; year 3.
Task 3: Data-Driven Discovery A. Apply statistical methods to analyze deep-time geochemical data for minerals and rocks. B. Apply Klee diagrams to analyze redox-sensitive element associations in minerals vs. time.	A. Identify positive and negative temporal correlations among redox-sensitive cations and anions in minerals, rocks, and proteins using pair correlations and principal component analyses; years 1-3. B. Record mineral diversity, disparity, and distribution for different geological eras; compare Klee diagrams for major element associations through time; years 1-3.

C. Apply statistical and graphical methods to analyze redox-sensitive trace elements in target minerals. D. Apply statistical methods to compare transition elements in rocks and minerals versus protein structures. E. Apply skyline diagrams to compare and contrast temporal patterns for different redox-sensitive elements F. Integrate and synthesize deep-time data. G. Disseminate results of data-driven discovery on Earth's evolving oxidation state and the rise of oxygen.	C. Graph temporal variations of redox-sensitive trace elements in sulfides and carbonates; use peak-fitting methods to identify maxima and minima; use linear regression analysis to establish trends; years 2-3. D. Develop timelines for the dominant oxidation states of redox-sensitive cations and anions for minerals and enzymes through deep time; years 2-3. E. Integrate efforts in 3-D (above) to produce a series of skyline diagrams; compare and contrast the timing and intensity of episodic mineralization. F. Integrate and synthesize all tasks above to develop a comprehensive model of Earth's changing near-surface (i.e., atmosphere, ocean, and shallow sub-surface) oxidation state over the past 4 billion years. G. The team, led by the PI and Project Manager, will submit 4 research papers and present 20 lectures/seminars in year 1. The team will submit 8 research papers and present 30 lectures/seminars in year 2. The team will submit 10 research papers, present 40 lectures/seminars, and organize thematic sessions at 2 international conferences in year 3. We also anticipate numerous research papers, lectures, and seminars in subsequent years.

Project Abstract: The Co-Evolution of the Geo- and Biospheres:
An Integrated Program for Data-Driven, Abductive Discovery in the Earth Sciences

Earth's living and non-living components have co-evolved for 4 billion years through numerous positive and negative feedbacks. Earth and life scientists have amassed vast amounts of data in diverse fields related to planetary evolution through deep time—mineralogy and petrology, paleobiology and paleontology, paleotectonics and paleomagnetism, geochemistry and geochrononology, genomics and proteomics, and more. Yet our ability to document, model, and explore these complex, intertwined changes has been hampered by a lack of data integration from these complementary disciplines. We propose a new program of data-driven discovery in the Earth and life sciences. We want to develop, curate, and integrate diverse data resources to focus on our planet's changing near-surface oxidation state and the rise of oxygen through deep time—a critical problem that exemplifies this co-evolution and underscores the opportunities and challenges of deciphering transient characteristics of Earth's history. Using abductive reasoning applied to our newly developed "Deep-Time Data Infrastructure" to discover patterns in the evolution of our planet's environment, we will create and merge the integrated data sets, statistical methods, and visualization tools that inspire and test hypotheses applicable to modeling Earth's past and today's changing environment.

Executive Summary: The Co-Evolution of the Geo- and Biospheres:
An Integrated Program for Data-Driven, Abductive Discovery in the Earth Sciences

Methodology/Implementation: Earth's living and non-living components have co-evolved for 4 billion years through numerous positive and negative feedbacks. Yet our ability to document, model, and explore these complex intertwined changes has been hampered by a lack of data synthesis and integration from many complementary disciplines—mineralogy, petrology, paleobiology, geochronology, proteomics, geochemistry, and more. Accordingly, our team from 6 academic institutions, coordinated by a Project Manager, propose to develop, curate, and integrate deep-time data resources to focus on our planet's dramatically changing near-surface oxidation state—a critical problem that touches on transformative geological and biological events through Earth history. The rise of oxygen exemplifies the co-evolution of rocks and life, and underscores both the tantalizing opportunities and technical challenges of deciphering transient characteristics of Earth's storied past.

Three aspects of this proposal are unique. First, by adding the dimension of geological time to existing data resources in mineralogy, petrology, Precambrian paleobiology, thermochemistry, and proteomics, we will gain important new insights regarding Earth's evolving oxidation states of the atmosphere, oceans, and near-surface environments. We will accelerate discovery by employing a variety of statistical methods and visualization tools to mine our new and growing data infrastructure. These discoveries will exemplify aspects of the co-evolution of the geosphere and biosphere that are applicable to modeling today's changing environment.

Second, we will take the critical first steps in building a comprehensive, integrated "Deep-Time Data Infrastructure"—what we envision as a 10-year, $10 million program, synthesizing many different data resources that share the dimension of geological time. This resource will be designed and implemented to tackle a host of problems related to Earth's complex evolution.

Third, we will exemplify the potential of data-driven scientific discovery. By exploiting the Deep-Time Data Infrastructure and employing powerful statistical and visualization methods, we will initiate a research program for abductive discovery in Earth sciences, where observations based on diverse and extensive data resources generate and test hypotheses.

3-Year Timeline: Our proposed program features 3 types of objectives. First is development of essential deep-time data resources. We will add age data for >5000 localities per year, as well as tectonic setting information, focusing on minerals of redox-sensitive elements. In years 2 and 3 we will add Precambrian (>542 million years) paleobiology data, including fossils and molecular/isotopic biosignatures, to paleobiodb.org. Also in years 2 and 3 we will incorporate a qualitative time dimension to the protein structure database (pdb.org).

Second, we will synthesize and integrate these data resources into the Deep-Time Data Infrastructure. In year 1 we will link data resources related to redox-sensitive elements in mineralogy, petrology, and geochemistry into an Earth Materials Data Infrastructure. In year 2 we will link this infrastructure to data on relative ages of enzymes that incorporate redox-sensitive transition elements. In year 3 we will integrate paleobiology and thermochemical data resources with the Deep-Time Data Infrastructure. We will also create an open-access interface and promote this new resource through a website, lectures, and publications.

Third, we will use the Deep-Time Data Infrastructure to probe Earth's oxidation state through Earth history. In years 1 to 3 we will search for deep-time correlations among trace and minor elements in rocks/minerals, enlist Klee diagrams and cluster analysis to identify correlations among numerous mineral or element pairings, and use quantitative geochemical models to document Earth's changing near-surface redox state. In years 2 and 3 we will graph temporal

variations of redox-sensitive trace elements in sulfides and carbonates, use peak-fitting methods to identify temporal maxima and minima, use linear regression analysis to establish trends, and develop timelines for the dominant oxidation states of redox-sensitive elements in minerals and enzymes through deep time. In year 3 we will integrate these efforts to produce a series of skyline diagrams to compare and contrast the timing and intensity of episodic mineralization. Ultimately, we will integrate and synthesize all tasks to develop a comprehensive model of the rise of oxygen—Earth's changing near-surface oxidation state over the past 4 billion years.

Personnel: To succeed, this effort requires an extraordinary range of expertise. Carnegie mineralogist Robert Hazen focuses on origins of life, mineral-molecule interactions, and the co-evolution of the geo- and biospheres. He introduced "mineral evolution" in 2008, leads the Deep Carbon Observatory, and is known for integrated scientific research and education efforts.

Mineralogist and crystallographer Robert Downs (University of Arizona) is the world's leader in mineralogical data resources. He maintains official mineral species and crystal structure databases and has close ties to all major Earth materials data resources.

Geobiologist Paul Falkowski leads Rutgers' Environmental Biophysics and Molecular Ecology Program. He has a unique combination of geological and biological expertise, a deep knowledge of deep time, and close ties to Rutgers' protein structure database—a key resource for this effort.

Peter Fox at RPI is a world leader in developing integrated data infrastructure, as well as exploring data with advanced statistical and visualization methods. He is an expert in the development of virtual observatories and semantic data frameworks in the geosciences.

Paleobiologist Andrew Knoll at Harvard is renowned for studies in paleobiology, geobiology, and life's evolution. An expert on the nature and evolution of early life, he has pioneered integrative paleontological research, using fossils to reveal Earth's environmental history.

Geochemist Dimitri Sverjensky of Johns Hopkins University is internationally recognized for work in geochemical thermodynamics—a field critical for inferring characteristics of Earth's ancient environments. He is expert in all aspects of water-rock-biomolecule interactions.

A Project Manager will play the critical role of coordination and integration, maintaining regular contact with all 6 teams, traveling at least twice a year to each node, and organizing Project annual meetings. He/she will ensure that data resources are integrated into a user-friendly, open-access interface. The Manager will actively engage in research, present lectures, and write publications related to the evolution of Earth's near-surface oxidation state. The Manager will have a PhD in Earth Sciences, with expertise in geobiology, geochemistry, and mineralogy/petrology; familiarity with concepts of Earth's changing near-surface environments in deep time; and familiarity with database development, management, and use.

PROJECT NARRATIVE—The Co-Evolution of the Geo- and Biospheres: An Integrated Program for Data-Driven, Abductive Discovery in the Earth Sciences

Introduction

Earth's living and non-living components are inextricably linked, as life and its near-surface environment have co-evolved for 4 billion years through sequential positive and negative feedbacks. Remarkable glimpses of this dynamic deep-time past are slowly coming into focus: Evidence is emerging for temporal changes in ocean and atmospheric chemistry; increases in mineral diversity; changing distributions of redox-sensitive trace elements and their isotopes; shifts in the character and rates of global tectonic mechanisms; the growth of continents; the nature of the supercontinent cycle; near-surface phenomena related to fluid-rock interactions; the origins and evolution of life; the rise of terrestrial ecosystems, notably their roots and soils; and myriad other consequences of the co-evolving geosphere and biosphere. We are beginning to tease out the nature and extent of these phenomena, placing them within the framework of geological space and time. Fascinating scientific questions and consequent opportunities abound.

In spite of this tantalizing scientific frontier, our ability to document, model, and explore implications of these complex, intertwined changes in the geosphere and biosphere has been hampered by a lack of deep-time data integration from multiple complementary disciplines—mineralogy, petrology, paleobiology, geochemistry, proteomics, geodynamics, geochrononology, paleotectonics, paleomagnetism, thermochemical modeling, and more. We therefore propose to develop, curate, and integrate deep-time data resources to investigate one key aspect of Earth history—the evolution of our planet's variable near-surface oxidation state through billions of years of complex feedbacks between life and its geological context. Numerous lines of evidence point to dramatic changes in composition and redox state of the atmosphere, oceans, and crust through deep time. The most notable "events" were associated with the episodic rise of O_2 through oxygenic photosynthesis and the concomitant burial of organic matter. However, the timing, rates, and drivers of this fascinating, and at times contentious, problem are still poorly understood and widely debated (Canfield 2014; Lyons et al. 2014). Furthermore, atmospheric and ocean compositions are also closely tied to other geochemical cycles, notably those of iron, carbon, nitrogen, sulfur, and hydrogen. Consequently, evidence for global change must be integrated from numerous sources of geological and biological data. The specific problem of understanding Earth's evolving oxidation state thus exemplifies some of the larger challenges associated with documenting Earth's complex history in deep time.

We propose a data-driven abductive strategy for discovery, focusing on the specific scientific mystery of Earth's changing oxidation state through deep time. Accordingly, we will (1) create and/or expand relevant open-access deep-time data resources; (2) integrate those resources into a Deep-Time Data Infrastructure; (3) exploit powerful statistical methods to discover illuminating correlations among these data; and (4) implement novel visualization tools that both accelerate discovery and facilitate the interpretation and dissemination of those discoveries. We will thus test existing hypotheses and promote discovery of previously hidden patterns related to 4 billion years of Earth's evolving near-surface oxidation state. These discoveries will inevitably lead to new hypotheses, including insights applicable to today's changing environment.

We anticipate three significant impacts of this effort. First, we will gain important insights regarding intertwined geological and biological causes and consequences of Earth's evolving near-surface oxidation state. Second, we will take the first key steps in establishing a Deep-Time Data Infrastructure, which ultimately will be employed to tackle a host of problems related to Earth history. Third, we will exemplify the powerful data-driven strategy for scientific discovery.

Part I: *Elaboration of the objectives/aims, including expected impacts:* The scientific motivation of this proposal is to understand Earth's changing near-surface oxidation state through deep time. To achieve this overarching 3-year goal, our proposed research will progress in three parallel but interrelated tracks, each with its own complementary set of aims and objectives.

Objective IA. Deep-Time Data Resource Development: Data-driven discovery requires reliable and comprehensive data resources. Studies of Earth's near-surface oxidation state rely on deep-time data on the nature and distribution of redox-sensitive elements, including their roles in rocks, minerals, and enzymes. Thus, we will take full advantage of existing open-access, deep-time data resources and associated infrastructure (see Table 1). However, critical gaps exist. Key mineral databases (i.e., mindat.org) do not incorporate geological time or tectonic setting. Protein structure databases (e.g., pdb.org) do not incorporate temporal information. The paleobiology database (paleobiodb.org) does not incorporate systematic information on Precambrian fossils (> 542 million years). Therefore, a major effort of this project will be to expand significantly geochemical data resources by adding age data for localities of minerals that incorporate redox-sensitive elements. We currently have added ages for minerals in ~2800 global localities. We propose to add ages for 5000 additional mineral localities in each of the project's 3 years.

In addition, in years 2 and 3 we will begin to integrate relative age information for proteins based on enzyme phylogenetics (see below), and we will design and construct a Precambrian paleobiology database for fossils and molecular/isotopic biomarkers. This new resource will be linked to The Paleobiology Database (paleobiodb.org; Alroy et al. 2008; Alroy 2010).

Table 1. Selected open-access data resources relevant to proposed deep-time research.

Web address	Content
rruff.info	Mineral species and properties
mindat.org	Mineral localities/associations/properties
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
ngdc.noaa.gov/mgg/geology/petros.html	Igneous rock geochemistry
http://volcano.si.edu	Volcanoes and their eruptions
melts.ofm-research.org	Thermodynamic modeling
lepr.ofm-research.org	High-T thermochemical data
metamorph.geo.uni-mainz.de/thermocalc/	Thermochemical modeling
vamps.mbl.edu/portals/deep_carbon/cdl.php	Subsurface microbial ecosystems
http://www.pdb.org/pdb/home/home.do	Protein structures
http://paleobiodb.org; see also fossilworks.org	The paleobiology database

Objective IB. Deep-Time Data Resource Integration: A vital step in achieving our objective of understanding Earth's changing near-surface oxidation state through deep time is to integrate existing and new deep-time data resources into one inter-operable infrastructure. In anticipation of this significant effort, we will convene a Deep-Time Data Workshop on Sunday, December 14, 2014 in San Francisco, immediately before the Annual Meeting of the American Geophysical Union. This gathering of world experts, which will be sponsored by the Deep Carbon Observatory, will address the needs and opportunities of the community and begin the coordination required to integrate diverse deep-time data resources. Should this grant be funded, we will immediately move in year 1 of our program to design and implement data infrastructure to link databases in mineralogy, petrology, and geochemistry—all data for Earth materials in which age, location, and composition (both elemental and isotopic) are key parameters.

In year 2, we will expand this infrastructure by implementing links to the protein structure database (pdb.org) at Rutgers University. An immediate challenge is that protein structures are

not directly associated with geological age. However, phylogenetic analyses of structurally related proteins can reveal relative ages that might be correlated with Earth's geochemical evolution (David & Alm 2011; Kim et al. 2013). For example, many deeply rooted enzymes that drive metabolism incorporate ferrous iron (Fe^{2+}), which was available abundantly in Earth's earliest oceans (Kim et al. 2013; Harel et al. 2014). Following the Great Oxidation Event (GOE) ~2.4 billion years ago, Fe^{2+} was removed from the upper oceans by reaction with oxygen, and then precipitated in sediments as iron sulfides by reaction with the H_2S generated by an expanded presence of sulfate-reducing bacteria (Canfield 1998). Enhanced fluid flow from the mantle transiently renewed the precipitation of iron formation ca. 1880 Ma (Chakrabarti et al. 2012, Rasmussen et al. 2012) and Fe^{2+} remained available in anoxic subsurface waters through most of the Proterozoic Eon (Johnston et al. 2010; Planavsky et al. 2009) but, with partial oxidation of the biosphere, new metals became available to evolving microbial populations. Consequently, more recently evolved enzymes incorporate molybdenum and copper (Williams 1981; Schwartz et al. 2001). This qualitative correlation between metal speciation in enzymes and the evolution of ocean chemistry suggests a plausible chronology that can be explored more rigorously with an integrated deep-time data infrastructure. This effort is central to our aim of teasing out previously unrecognized correlations between the evolving geosphere and biosphere.

In year 3 we will link the Deep-Time Data Infrastructure to paleobiology and thermochemical databases. We anticipate that linking data on Precambrian fossils (as well as those from the Phanerozoic Eon; http://paleobiodb.org) to those for rocks and minerals will be straightforward, because locality and age parameters are consistent. The integration of thermochemical data infrastructure is different in character, because thermochemical resources represent both compilations of data (e.g., calorimetric measurements of individual mineral species) and modeling tools to analyze the implications of coexisting mineral assemblages. Thus, for example, data on the distribution of copper minerals before and after the GOE will point directly to the changing oxidation state of Earth's near-surface environment. The linked thermochemical resources in our Deep-Time Data Infrastructure will be designed to facilitate that analysis.

Objective 1C. Data-Driven Discovery: A key motivation of this proposal is development of strategies to facilitate abductive discovery of patterns in the deluge of deep-time data (Hazen 2014). We focus on discovery of patterns in Earth's near-surface oxidation, but many other opportunities for exploiting a deep-time data infrastructure exist. In year 1 we will integrate large and growing amounts of rock, mineral, and other geochemical elemental and isotopic data as a function of geological time. We will employ standard statistical methods, including correlation matrices and principal component analysis, to quantify deep-time correlations among numerous redox-sensitive trace/minor elements in rocks and minerals. We will interpret and amplify these insights using quantitative geochemical modeling, including reaction path modeling, to document Earth's changing near-surface redox state over 3.5 billion years (e.g., Sverjensky & Lee 2010). In year 1 we will also develop Klee diagrams to enhance visualization of correlations among numerous mineral or element pairings as a function of geological time (see IIIC below).

In year 2 we will adapt statistical methods for discovering correlations among varied data. For example, we will search for temporal correlations for transition metals in proteomic versus mineral databases (see Objective IB above). We will also apply principal component analysis to trace element data in related groups of redox-sensitive minerals (i.e., carbonates and sulfides) from different geological eras. Also in year 2, as our data resources expand, we will begin to apply more sophisticated statistical approaches to detect peaks and episodicity in data represented by numbers of localities or number of specimens versus time (see IIIC below).

In year 3 we will implement visual exploration via 3D graphing and skyline diagrams to visualize complex multi-dimensional relations (see IIIC below). Another major effort in year 3 will be to link data from non-equilibrium thermodynamic models of ancient redox environments to all other databases. Models of weathering processes on the continents, and hydrothermal processes in the oceans, will enable fragmentary evidence and clues from other databases as to the changing redox environment to be integrated to develop an internally consistent picture of the near-surface environment on Earth through time. Finally, an ultimate objective of year 3 is to initiate open access to the Deep-Time Data Infrastructure, building from the key elements of successful virtual observatories, where access and integration are the driving tenants. Such infrastructure provides internet-enabled services for machine and human interfaces, allowing multi-dimensional "data spaces" to be assembled, making true exploratory discovery possible.

Part I—Expected Impacts: We predict three significant novel impacts. First, we are confident that our strategy of data-driven discovery, coupled with more traditional hypothesis-driven inquiries, will result in recognition of surprising new patterns and phenomena related to Earth's evolving near-surface oxidation state. We anticipate discoveries related to the extent and mechanisms of redox change in the oceans, the atmosphere, and the crust—processes that arise from complex feedbacks between the geosphere and biosphere. These patterns will inevitably emerge from our integration of deep-time data related to redox-sensitive elements in rocks, minerals, and enzymes, linked to complementary information from paleobiology, proteomics, and thermochemical modeling. Ultimately, we will gain a richer understanding of Earth's changing near-surface redox environment through 4 billion years of Earth history.

Second, we will take the first key steps in establishing the Deep-Time Data Infrastructure, which ultimately could be employed to tackle a host of problems related to Earth history. Here we have focused on data from mineralogy, petrology, geochemistry, paleobiology, paleotectonics, proteomics, and thermochemistry—all data that inform studies of Earth's changing near-surface oxidation state. Ultimately, if this program continues to its logical climax, we will also integrate data related to paleomagnetism, hydrology, sedimentation, genomics, volcanism, and the vast amount of data related to geological mapping in three dimensions. That vision can only be achieved by taking these first essential steps.

Third, we will exemplify a powerful new strategy for scientific discovery. Data-driven discovery provides a model for 21st-century science that explicitly recognizes the power of abduction and exploits as yet untapped opportunities represented by the accelerating deluge of Earth and life science data (Hazen 2014). Such a strategy in no way subsumes deduction and induction as effective pathways to scientific discovery; indeed, the abductive approach explicitly relies upon the accumulation of traditional measurements and observations. Abduction amplifies the vast and growing flood of data inspired by deductive and inductive discovery. Ultimately, when data from numerous geological and biological fields 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 and life sciences.

Part IIa. *Relationship of the objectives/aims to the present state of knowledge in the field:*

The Problem of Earth's Changing Near-Surface Oxidation State: The central scientific focus of this proposal is to understand the changing near-surface oxidation state of Earth's oceans, atmosphere, and crustal rocks and minerals through deep time. This problem has been the subject of extensive previous study, much of it data-driven (Canfield 2014; Lyons et al. 2014). Current thinking was strongly influenced by the seminal synthesis of Heinrich Holland (1984, 2006),

who recognized significant compositional changes in Earth's oceans and atmosphere, including a major transition at ~2.4 billion years ago—changes he ascribed to the "Great Oxidation Event" (GOE), made possible by the rise of oxygenic photosynthesis. Holland's analysis and subsequent research emphasize the complex integrated nature of this question, with critical impacts from physical (e.g., gravitational escape of hydrogen; photo-oxidation), chemical (e.g., continental and submarine weathering reactions), geological (e.g., volcanism; subduction), and biological (e.g., photosynthesis; biomineralization) processes (e.g., Holland 2006; Kasting 2013).

Atmospheric chemistry was quickly recognized to be a consequence of feedbacks principally among O, C, H, N, and S species, mediated initially by volcanism and weathering reactions, notably with Mg, Ca, and Fe minerals (Walker 1977; Kump et al. 2001; Catling & Claire 2005; Kump & Barley 2007; Falkowski & Isozaki 2008), with subsequent dominant biological inputs, notably through methanogenesis (Rye & Holland 1998; Catling et al. 2001; Pavlov et al. 2003; Zahnle et al. 2013), and ultimately by oxygenic photosynthesis and creation of significant near-surface redox gradients (Canfield et al. 2000; Kump et al. 2001; Kasting 2001; Kasting & Siefert 2002; Towe 2002; Knoll 2003; Bekker et al. 2004; Barley et al. 2005; Catling & Claire 2005; Kump & Barley 2007). Numerous lines of evidence contribute to an understanding of the timing and extent of the GOE and subsequent oxygenation associated with or following Neoproterozoic glaciation: weathering rates of redox-sensitive minerals (Rasmussen & Buick 1999; England et al. 2002); Ce and Fe anomalies in Precambrian paleosols (Rye & Holland 1998; Murakami et al. 2001); evidence (often controversial) from Precambrian fossils and molecular and isotopic biomarkers (Brocks et al. 1999; Brasier et al. 2005; Dutkiewicz et al. 2006; Buick 2008; Rasmussen et al. 2008; Allwood et al. 2009; Bosak et al. 2009; Flannery & Walter 2012; Noffke et al. 2013); sulfur isotopic data (Farquhar et al. 2001, 2007; Mojzsis et al. 2003; Ono et al. 2003; Bekker et al. 2004; Fike et al. 2006; Domagal-Goldman et al. 2009; Guo et al. 2009; Hofmann et al. 2009; Halevy et al. 2010; Halevy 2013); and insights on the carbon, nitrogen, and phosphorus cycles (Donnelly et al. 1990; DesMarais et al. 1992; Halverson et al. 2005; Elie et al. 2007; Falkowski & Godfrey 2008; Laakso & Schrag 2014). In addition, detailed studies of redox-sensitive trace elements in Archean black shales point to possible "oxygen oases" prior to the GOE (Anbar et al. 2007; Frei et al. 2009; Olson et al. 2013; Partin et al. 2013).

Changes in ocean chemistry, including their intertwined salinity (Hardie 1996; 2003; Lowenstein et al. 2001; Dickson 2002); pH (Walker et al. 1981; Caldeira and Kasting 1992; Kah et al. 2004; Kah and Riding 2007); and concentrations and speciation of redox-sensitive elements and their isotopes (e.g., Canfield 1998; Anbar & Knoll 2002; Habicht et a. 2002; Ohmoto et al. 2006; Canfield et al. 2007; Ono et al. 2007; Kaufman et al. 2007; Papineau et al. 2007; Scott et al. 2008; Partin et al. 2013), also attract significant attention. Mineralogical data also point to gradual oxidation of the subsurface crustal environment (Golden et al. 2013; Large et al 2014). In the Phanerozoic Eon, redox gradients created by interactions among buried reduced organic carbon, circulating oxidized meteoric water, and hydrothermal solutions led to significant ore-forming events, many of which may have been microbially mediated (Laberge 1973; Lovley et al. 1991; Anbar & Holland 1992; Suzuki & Banfield 1999; Fayek et al. 2005; Reith et al. 2006; Konhauser et al. 2007). Thermochemical reaction path calculations are critical in understanding these effects (Sverjensky & Lee 2010; Markl et al. 2014). Proposed studies of redox-sensitive trace elements will enhance our understanding of the timing and extent of subsurface changes.

Finally, an important outstanding problem—one only touched on in this proposed study—is the nature of possible changes in the oxidation state of the upper mantle, and the extent to which the redox states of the crust and mantle have been coupled over time (Kump et al. 2001;

Rohrbach et al. 2007; Burgisser & Scaillet 2007). Here, again, variations in redox-sensitive trace elements may provide keys to deciphering changes in deep time (e.g., Canil 1997, 2002).

Stepping back from the specific question of Earth's oxidation, two related types of prior discovery inspire this proposal. First, varied recent findings reveal that the geo- and biospheres, long treated as virtually unrelated domains, have co-evolved in many ways. Co-investigators of this proposal have contributed significantly to this emerging paradigm shift in the geosciences. Falkowski and colleagues have demonstrated significant feedbacks between ocean/atmospheric chemistries and microbial metabolism that appear to make it extremely difficult for the redox state of the planet to change (Fennel et al. 2005; Falkowski & Godfrey 2008; see 1B above). Hazen, Downs, Sverjensky et al. discovered that 2/3rds of mineral species occur as a consequence of biologically-mediated changes, notably redox changes, in Earth's near-surface environment (Hazen et al. 2008, 2011, 2014). Knoll and colleagues articulated significant changes through time in skeletal biomineralization, with consequences for facies development in precipitated sediments (Knoll 2003b; Pruss et al. 2010) and clear physiological linkages between environmental changes and evolutionary events (Knoll et al. 2007; Knoll 2013). Geobiological thinking extends to speculations about subsurface origins of life, the global carbon cycle, rates and modes of rock weathering, and even possible roles of microbes in the initiation of subduction and modern-style plate tectonics. Bio-geo feedbacks thus dominate key aspects of Earth history.

Coupled with these findings are recent advances in "mineral evolution" by Hazen and colleagues, who study Earth's changing near-surface mineralogical characteristics (including diversity and distribution of species, elemental/isotopic compositions, and morphologies) through deep time. Though initially focused on qualitative changes in mineralogy (Hazen et al. 2008), recent studies incorporate a quantitative, data-driven approach. For example, Hazen et al. (2012) examined mercury (Hg) minerals in what has been called a "brute force use case" of data mining. Their analysis consumed ~1.5 person-years of effort, mostly involving locating and evaluating hundreds of scattered references in a dozen languages. Three unexpected results emerged: (1) Almost all Hg mineral localities correlate with 3 episodes of supercontinent assembly at ~2.7, 1.8, and 0.4 billion years. (2) An as yet unexplained billion-year gap in Hg mineralization occurred at 1.8 to 0.8 billion years. (3) The largest Hg deposits are coeval with Carboniferous coal measures, suggesting links between the C and Hg cycles. These abductive discoveries, though focused on a single rare element with relatively few localities and no essential biological roles, demonstrate the untapped potential of data-driven discovery. Accordingly, each of the three parallel aspects of our proposal builds on existing knowledge bases in Earth, life, and data sciences.

IIa1. Deep-Time Data Resource Development: We will rely on, and promote continued development of, extensive existing open-access data resources (Table 1; Mutschler et al. 1981; Ghiorso & Sack 1995; Holland & Powell 1998; Ghiorso et al. 2002; Stixrude & Lithgow-Bertelloni 2005, 2011; Lehnert et al. 2007; Hazen 2014). We will significantly enhance existing mineralogical (mindat.org) and paleobiology (paleobiodb.org) data resources.

IIa2. Deep-Time Data Resource Integration: We will build on developments of integrated Earth science data infrastructures, including the Virtual Solar-Terrestrial Observatory (Fox et al. 2009), the Biological and Chemical Oceanography Data Management Office portal (Rozell et al. 2013), and the Global Change Information System (Tilmes et al. 2013; Ma et al. 2014). However, a systematic abductive discovery strategy based on a data infrastructure that integrates and interrogates diverse deep-time data resources has not previously been attempted.

IIa3. Data-Driven Discovery: We will exploit developments in data-driven discovery, including established statistical methods for data analysis and visualization techniques (Peter & Shneiderman 2008). Data mining efforts that exploit large databases and enhanced data interrogation techniques to seek patterns (Tan et al. 2005; Hey et al. 2009; Shoshani & Rotem 2010; Rodriguez and Laio 2014) are much in the news lately, particularly with respect to investment and national security applications. In science and medicine new data resources also have the potential to reveal previously unrecognized phenomena (e.g., DeepDive; Niu et al. 2012). For example, explorations of genome databases or medical records represent growing applications of abductive strategies. A number of scientists (including Hazen, Fox, and Downs) advocate abductive research as a motivation for EarthCube, a major data infrastructure project spearheaded by NSF (see IIc3 below). However, no grant support presently exists, nor are other groups pursuing a data-driven, deep-time abductive discovery strategy as outlined here.

Part IIb. *Relationship of the objectives/aims to work in progress by the project personnel:* The six key personnel in this proposal are all deeply involved in studies that are synergistic and complementary to this proposal. This proposed project will thus leverage their internationally recognized ongoing research programs, expertise, colleagues, and infrastructures.

Robert Hazen (Carnegie Institution), who introduced "mineral evolution" (Hazen et al. 2008, 2009, 2011, 2012, 2013a, 2013b, 2014; Golden et al. 2013; Hazen 2013, 2014; Grew & Hazen 2014), is also known for integrated scientific research and education efforts (Hazen 2005; Hazen & Trefil 2009; Trefil & Hazen 2012). He has spent the past 20 years exploring aspects of the co-evolving geosphere and biosphere through more than 4 billion years of Earth history.

Robert Downs (University of Arizona) maintains the official IMA mineral and crystal structure databases (both at rruff.info). He has developed critical open-access data resources in mineralogy (Downs 2006; Downs & Wallace 2003). As member of IMA's Executive Board he leads the international effort to integrate data resources into mainstream mineralogical culture.

Paul Falkowski (Rutgers University) heads the Environmental Biophysics and Molecular Ecology Program in the Institute of Marine Sciences and Dept. of Earth & Planetary Sciences. He has pioneered studies of the coevolution of protein structure and transition metal selection in early metabolism of microbial life. (Falkowski et al. 2008; Kim et al. 2013, Harel et al. 2014).

Peter Fox (Rensselaer Polytechnic Institute) pioneered development of virtual observatories (Fox et al. 2006, Fox & McGuinness 2011), semantic data frameworks (Fox et al. 2009; Narock & Fox 2012), and data science infrastructures for exploratory science (Fox & Hendler 2009, 2011) in the geosciences and environmental science (Tilmes et al. 2013; Ma et al. 2014).

Andrew Knoll (Harvard University), an expert on the nature and evolution of early life, has pioneered integrative paleontological research in which fossil evidence of evolution is interpreted in the context of environmental history (Knoll 2003a, 2003b, 2013). Knoll has nearly four decades of field experience in Proterozoic terrains.

Dimitri Sverjensky (Johns Hopkins University) is expert in all aspects of the evolution of water-rock-biomolecule interactions from molecular (Jonsson et al. 2009; Parikh et al. 2011) to planetary (Sverjensky & Lee 2010) scales. He has developed new thermodynamic databases for aqueous species and minerals from surficial conditions on Earth (Sverjensky et al. 1991) up to the elevated pressures and temperatures of the upper mantle (Sverjensky et al. 2014).

In addition to these investigators, a Project Manager will play the critical role of coordination and integration, maintaining regular contact with the 6 teams, traveling at least twice a year to each node, and organizing annual meetings. He/she will ensure that data resources are integrated into a user-friendly, open-access interface. The Manager will actively engage in research, give

lectures, and write publications related to the evolution of Earth's near-surface oxidation state. The Manager will hold a PhD in Earth Sciences with expertise in mineralogy/ petrology, geobiology, and geochemistry; knowledge of concepts related to Earth's changing near-surface environments in deep time; and familiarity with database development, management, and use.

Part IIc. *Relationship of the objectives/aims to work in progress at other institutions:* This proposal for a Deep-Time Data Infrastructure is complementary to 3 types of ongoing data-driven research at a number of institutions around the world (See also IIa above).

IIc1. Development of Deep-Time Data Resources. More than a dozen efforts around the globe focus on development of open-access data resources that incorporate the dimension of geological time (Table 1). Most significant are extensive, expanding databases in petrology and mineralogy. Other projects, not yet open access, focus on specific minerals [pyrite (Large et al. 2014); zircon (e.g., Valley et al. 2005; Condie & Aster 2010; Voice et al. 2011); carbonates (Xiaoming Liu, pers. comm.)] or isotopes [especially C, S, and Sr (Farquahar et al. 2001, 2007; Prokoph et al. (2008), but also Si (Chakrabarti et al. 2012)]. We will rely on all of these resources, and continue to be in close touch with the scientists who developed and sustain them.

IIc2. Application of Deep-Time Data Resources to Earth Sciences. We are coordinating with many Earth scientists who use data resources to probe Earth's changing near-surface environment (e.g., Valley et al. 2005; Farquhar et al. 2007; Keller & Schoene 2012; Large et al. 2014). We are sponsoring the "Deep Carbon in Deep Time" Short Course in October in Vancouver prior to the Geological Society of America's annual meeting, and the "Deep-Time Data Workshop" in San Francisco prior to the Annual Meeting of the American Geophysical Union. These workshops will be sponsored by the Deep Carbon Observatory and will gather leaders in deep-time data development and analysis, as well as many early-career scientists.

IIc3. Development of an Integrated Earth Science Data Infrastructure. Development of integrated data resources, which requires expertise in multiple scientific disciplines as well as data science, is complementary to the discipline-driven efforts noted above. Of note, EarthCube is a major NSF-sponsored program to develop integrated cyber-infrastructure for the Earth sciences (NSF 2012). Downs, Fox, and Hazen participate extensively in meetings related to the organization and objectives of EarthCube, and have played a significant role in refocusing that effort to consider the type of applications that hundreds of thousands of Earth scientists worldwide might want to make of this new resource. We were thus instrumental in initiating the first of more than 30 stakeholder workshops both for young geoscientists and for specific sub-discipline areas.

Our proposed effort is complementary, not competitive, to EarthCube's ambitions for at least 3 reasons. First, EarthCube focuses primarily on building research coordination networks and cyberinfrastructure building blocks—tasks that remain distant and distinct from our research objectives. EarthCube is primarily concerned with finding ways to interconnect the vast and growing data resources related to environmental concerns of the modern world. Applications to deep time appear to lie many years in the future, and what does emerge may not be relevant to studies of the co-evolution of the geosphere and biosphere. Second, our project is distinct from NSF priorities for EarthCube, which explicitly exclude support for database development—a key aspect of our proposal. Mineral, protein, and paleobiology databases are at present inadequate for our purposes because they do not include the dimension of deep time. Third, our model of an integrated data infrastructure is motivated by a set of scientific questions related to Earth's evolving oxidation state that, unlike EarthCube, links Earth sciences with biological sciences. Our ambitions thus extend beyond the limits presently envisioned for EarthCube. Ultimately, we

hope to make significant progress several years before EarthCube is available to users, and thus we will provide a compelling example of a successful integrated data infrastructure.

Part III. *Description of proposed methods and procedures related to each aim:* Many of our proposed methods and procedures have been introduced in Sections I and II above. Here we highlight examples of specific methods related to our proposed activities.

IIIA. Deep-Time Data Resource Development: Objectives outlined in 1A (above) principally involve enhancing existing mineralogical (mindat.org), protein structure (pdb.org), and paleobiology (paleobiodb.org) data resources to explore Earth's near-surface oxidation in deep time. Our efforts will proceed on three fronts. (1) We will "inhale" existing databases, some of which are not presently available through open-access sources. (2) We will seek out "dark data" resources, especially archives of geochronology data that are not publicly available. (3) We will interrogate published literature and add deep-time mineral and paleobiology data into our developing data resources. Our efforts will focus on minerals that incorporate redox-sensitive elements, including Fe, Mn, Cu, Ni, Co, Mo, U, C, S, As, and P.

We will coordinate with the open-access Paleobiology Database (paleobiodb.org) in our development of a Precambrian paleobiology database (e.g., Alroy et al., 2008; Kiessling et al., 2010; Peters & Heim 2010; Alroy 2010). A significant novelty associated with our paleobiology efforts will be incorporation of the growing body of robust molecular biomarker fossil data, which have proven especially important in interpreting Precambrian microbial ecosystems.

IIIB. Deep-Time Data Resource Integration: Building an integrated Deep-Time Data Infrastructure involves structuring of often sparse (i.e., we will not have continuous deep-time coverage), but annotated, multi-dimensional data in a form suitable for visual and algorithmic exploration and analyses. As such, compact graph-based representations are desirable (cf., array-based or lower-dimensionality spreadsheets; Brisaboa et al. 2014). In addition, there are standardized vocabularies for these structures (RDF Data Cube Vocabulary 2014) that accommodate annotations (especially provenance) as well as links to related concepts and datasets. Bridging into virtual observatory settings, time representations and the all-important temporal reasoning have also advanced to a viable state (West et al. 2009). Thus, a key activity will be to maintain the key backbone deep-time data while applying/developing the necessary extract-transform-load (ETL) to instantiate the time-based graphs (e.g., Lebo et al. 2011). The data infrastructure must fully accommodate quality assessment and data integrity verification.

Incorporation of thermochemical data and modeling capabilities into the Deep-Time Data Infrastructure involves additional attention to representation of physical and chemical processes in the structural representation of the data space. In other words, explicit representations or annotations of model/equation relationships among elements in the data cube are possible and desirable [data in dimension A are related to dimension B by chemical rate equation C representing chemical process D at time T and rate R; (Khandewal & Fox 2014)].

IIIC. Data-Driven Discovery: Methods and procedures to achieve our primary goal of understanding details of Earth's changing near-surface redox state through 4 billion years rest primarily on mining our Deep-Time Data Infrastructure with statistical and visualization techniques to seek previously unrecognized relations among diverse types of data. We will employ Klee diagrams to explore connections among redox-sensitive elements. For example, preliminary studies of coexisting elements in all known minerals (Figure 1) reveal an as yet unexplained affinity of cobalt for arsenic (surprisingly, 60% of known Co minerals incorporate As). Studies of such redox-sensitive element couplings through deep time will reveal changes in

near-surface oxidation state; Klee diagrams are the ideal visualization tools. We will also adopt statistical methods to characterize temporal variations Hammer et al. 2001). With small data sets (<1000 age data) we have resorted to standard Gaussian curve fitting procedures (e.g., Hazen et al. 2012). However, a much more nuanced analysis of peak significance versus time is possible with larger data sets by applying Monte Carlo mean kernel density analysis (Aster et al. 2004; Condie & Aster 2010), by which true peaks are more easily resolved from noise.

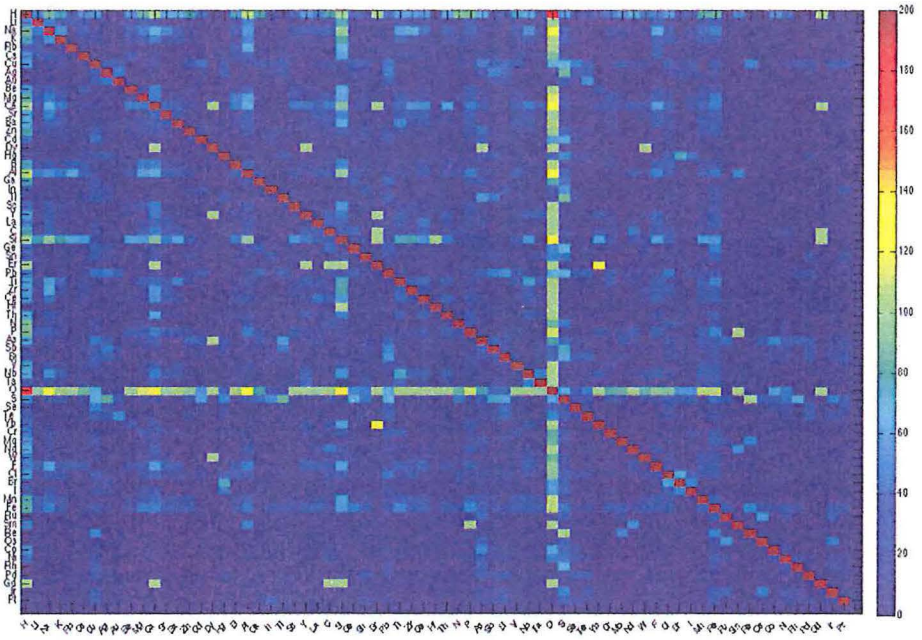

Figure 1. A Klee diagram with 72x72 matrix elements, each representing the percentage of mineral species containing element X that also contain element Y. "Hot spots" indicate elements that are geochemically linked. Many of these hot spot features reveal previously undocumented relationships between redox-sensitive elements.

We also require statistical tests to evaluate the significance of changing trace element or isotopic compositions versus time. We will apply significance tests based on linear regression analysis of the log of a compositional parameter X as a function of time, using indicator variables as described in Montgomery et. al. (2006). In this method the regression equation is:

$$\log(X) = \beta_0 + \beta_1 t + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_i x_i,$$

where t is time and β_i are regression parameters, and x_i are indicator variables. Indicator variables are set to 0 or 1 depending on which era the data point is in. For example, x_6 might represent samples from the most recent Cenozoic Era, so a data point from this era has $x_6 = 1$ while $x_2 = x_3 = x_4 = x_5 = 0$. Subsequent regression analysis of the values for β_i can reveal significant changes in average $\log(X)$ concentration with respect to time (Golden et al. 2013).

Skyline diagrams (Figure 2) are particularly useful in comparing subtle temporal patterns of minerals or elements. For example, we have observed striking episodicity in the mineralization of several elements, including B, Be, Hg, Mo, U, and Cu (Hazen et al. 2014). These patterns are similar to those for zircon ages (Condie & Aster 2010; Voice et al. 2011), granite pegmatites (Tkachev 2011), and pyrite (Large et al. 2014), and they may correlate with episodes of supercontinent assembly (e.g., Nance et al. 2014). However, subtle differences exist and skyline diagrams will enhance our ability to discern secondary patterns and trends.

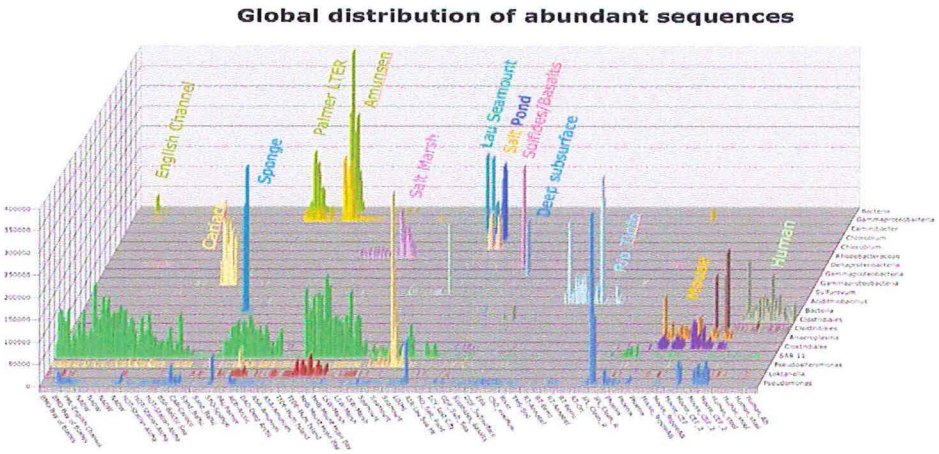

Figure 2. A skyline diagram (here showing genomic sequences of marine fauna) allows nuanced comparison and pattern recognition for frequency distribution data.

Part IV. *Technical problems that may be encountered and how they will be addressed:* The complexities of integrating heterogeneous data resources are well documented, and constitute important challenges to be overcome. Deep-time data in the Earth and life sciences must be integrated with consistent temporal and spatial parameters, yet underlying conventions vary among different disciplines. Recent progress in linking extant vocabularies and conventions with open-world reasoning (versus trying to harmonize or articulate specific mappings; Bechhofer et al. 2008) allows semantic mediation and probabilistic approaches to ameliorate heterogeneities noted above and to accommodate gaps or errors in parameters (e.g. Laskey 2006).

- Geochronological techniques have provided reliable ages for many rocks and minerals. However, diagenesis, metamorphism, hydrothermal alteration, and weathering can cause successive mineralogical alterations and thus impose multiple relevant ages for a specific location. Therefore, we must develop methods for incorporating metadata.
- Most repositories of Earth materials data record geographical locations; however, different coordinate systems must be recognized and converted for consistency. Deep-time data must also include depth information, because vertical sections (for example from mines) typically transect a range of ages. However, this information is rarely provided in data resources.
- Vast amounts of valuable and relevant "dark data" reside in individual and corporate archives. Creating incentives to provide data access continues to be a challenge, though recent calls from governments worldwide for open, accessible data are having an effect. In response, various means of establishing credit and attribution for such openness with data are also appearing (e.g., Thompson Reuters Data Citation Index). Our proposed deep-time data project will thus contribute to much needed changes in the sociology of our science.

Part V. *Description of facilities, equipment, and resources:* Virtually all hardware and software required for cyber-infrastructure developments are in place and operational. Resources include the DCO collaboration and data portal (deepcarbon.net); a multi-processor, large memory and storage server hosted in the main (24x7) computer facility at RPI (operated by RPI/TWC); and the DCO computer cluster (http://deepcarbon.net/DCOcomputer) with 640 processor cores, 544GB system memory, and 154TB system storage. This project will take advantage of software licenses in the installed systems, as well as free and open source software.

This proposal builds on significant, if largely uncoordinated and therefore underutilized, data resources. The most obvious resources are deep-time databases in Earth and life sciences (Table 1). Our proposed effort relies on cooperation and coordination with scientists who manage these

large and growing resources. Other key resources are to be found in uncollated data of literally tens of thousands of relevant publications. Our 6 institutions individually and collectively enjoy world-class library and information infrastructure, upon which we will continue to rely.

EarthCube, once organized and launched, constitutes an important potential resource for our project. An ultimate goal will be to integrate the Deep-Time Data Infrastructure with the much more extensive proposed scope of EarthCube, thus creating much more expansive opportunities for data-driven discovery. We will continue to participate and promote this long-term possibility.

Part VI. *Plans for this project beyond the proposed time period, including financial support:* We view this proposal as the critical first steps in a 10-year, ~$10 million international effort to coordinate all accessible resources related to Earth and life deep-time data. Data infrastructure developed during the initial 3-year effort will be critical to creating a committed international data community, to secure additional funding from other sources, and to establish a new culture among our scientific peers to support open-access collection, curation, and integration of our fields' immense but untapped data resources. Three facets of these efforts should be noted.

- We have support from the Alfred P. Sloan Foundations for mineral database development , plus Sloan support for Earth materials (but not deep-time) data infrastructure development . In addition, discipline-bound research of the Co-Investigators is currently supported by from agencies and foundations. Furthermore, the key data resources upon which we will rely are supported by as much as $50 million per year.
- Hazen has proposals pending for deep-time data acquisition from NSF and the Simons Foundation . He is also a team member on two pending NASA Astrobiology Institute proposals with mineral evolution components.
- The Robert and Margaret Hazen Foundation has pledged for the duration of this project to support database development and mineral evolution research. These funds will be distributed primarily to support early-career scientists at the University of Arizona, Johns Hopkins University, and the Carnegie Institution.

Part VII. *Advisory Board:* We do not plan to convene a formal advisory board during the initial stages of this program. We will, however, solicit advice, cooperation, and, where appropriate, collaborations among the numerous scientific stakeholders in deep-time data. We will hold two deep-time data workshops in Fall 2014 (both workshops sponsored in part by the Deep Carbon Observatory) to begin this important coordination of international efforts.

Efforts of the Project Manager will be critical, not only in integrated and synthesizing the research of the 6 nodes, but also in conducting numerous face-to-face meetings and other exchanges with international leaders of the data resources noted in Table 1. The Project Manager will coordinate annual meetings of the Deep-Time Data Infrastructure team, at which we will seek advice and cooperation from these leaders—steps critical in building an effective open-access resource. The Project Manager, along with the lead scientists, will conduct numerous seminars and lectures to meet prospective users, describe the project, and disseminate its results.

Our work is further leveraged by Downs, Fox, and Hazen, who are founding members (and Hazen was the first Chairperson) of the Mineralogical Society of America's Committee on Data Science, whose formal responsibilities include oversight of Earth materials data infrastructure and policy, and all of whose members are important contributors to the data resources we seek to integrate. We will continue to participate and to coordinate with that international body.

Finally, should this effort evolve as we envision into a 10-year, ~$10 million program that integrates all aspects of deep-time data in the Earth and life sciences, then we will constitute an international advisory board that represents the full constellation of stakeholders.

Part VIII. Bibliography

Allwood, A.C., Grotzinger, J.P., Knoll, A.H., Burch, I.W., Anderson, M.S., Coleman, M.L., and Kanik, I. (2009) Controls on development and diversity of Early Archean stromatolites. Proceedings of the National Academy of Sciences USA 106:9548-9555.

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.

Anbar, A.D., and Holland, H.D. (1992) The photochemistry of manganese and the origin of banded iron formations. Geochimica et Cosmochimica Acta 56:2595-2603.

Anbar, A.D., and Knoll, A.H. (2002) Proterozoic ocean chemistry and evolution: A bioinorganic bridge? Science 297:1137-1142.

Anbar, A.D., Duan, Y., Lyons, T.W., Arnold, G.L., Kendall, B., Creaser, R.A., Kaufman, A.J., Gordon, G.W., Scott, C., Garvin, J., and Buick, R. (2007) A whiff of oxygen before the great oxidation event? Science 317:1903-1906.

Aster, R., Borchers, B., and Thurber, C. (2004) Parameter Estimation and Inverse Problems. NY: Elsevier Academic Press.

Barley, M.E., Bekker, A., and Krapez, B. (2005) Late Archean to early Paleoproterozoic global tectonics, environmental change and the rise of atmospheric oxygen. Earth and Planetary Science Letters 238:156-171.

Bechhofer, S., Yesilada, Y., Stevens, R., Jupp, S., and Horan, B. (2008) Using ontologies and vocabularies for dynamic linking, Internet computing. IEEE 12:32-39.

Bekker, A., Holland, H.D., Wang, P.-L., Rumble, D. III, Stein, H.J., Hannah, J.L., Coetzee, L.L., and Beukes, N.L. (2004) Dating the rise of atmospheric oxygen. Nature 427:117-120.

Bosak, T., Liang, B., Sim, M.S., and Petroff, A.P. (2009) Morphological record of oxygenic photosynthesis in conical stromatolites. Proceedings of the National Academy of Sciences USA 106:10939–10943.

Brasier, M.D., Green, O.R., Lindsay, J.F., McLoughlin, N., Steele, A., and Stoakes, C. (2005) Critical testing of Earth's oldest putative fossil assemblage from the 3.5 Ga Apex chert, Chinaman Creek, Western Australia. Precambrian Research 140:55–102.

Brisaboa, N.R., Ladra, S., and Navarro, G. (2014) Compact representation of Web graphs with extended functionality. Information Systems 39:152-174.

Brocks, J.J., Logan, G.A, Buick, R., and Summons, R.E. (1999) Archean molecular fossils and the early rise of eukaryotes. Science 285:1033-1036.

Buick, R. (2008) When did oxygenic photosynthesis evolve? Philosophical Transactions of the Royal Society London B363:2731-2743.

Burgisser, A. and Scaillet, B. (2007) Redox evolution of a degassing magma rising to the surface. Nature 445:194-196.

Caldeira, K. and Kasting, J.F. (1992) Susceptibility of the early Earth to irreversible glaciation caused by carbon dioxide clouds. Nature 359:226-228.

Canfield, D.E. (1998) A new model for Proterozoic ocean chemistry. Nature 396:450-453.
Canfield, D.E., Habicht, K.S., and Thamdrup, B. (2000) The Archean sulfur cycle and the early history of atmospheric oxygen. Science 288:658-661.
Canfield, D.E., Poulton, S.W., and Narbonne, G.M. (2007) Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. Science 315:92-95.
Canfield, D.E. (2014) Oxygen: A Four-Billion Year History. Princeton, New Jersey: Princeton University Press.
Canil, D. (1997) Vanadium partitioning and the oxidation state of Archean komatiite magmas, Nature 389:842-845.
Canil, D. (2002) Vanadium in peridotites, mantle redox and tectonic environments: Archean to present. Earth and Planetary Science Letters 195:75-90.
Catling, D.C. and Claire, M.W. (2005) How Earth's atmosphere evolved to an oxic state: A status report. Earth and Planetary Science Letters 237:1-20.
Catling, D., Zahnle, K., and McKay, C. (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. Science 293:839-843.
Chakrabarti, R., A.H. Knoll, S.B. Jacobsen, and W.W. Fischer (2012) Silicon isotopic variability of Proterozoic cherts. Geochimica et Cosmochimica Acta 91:187-201.
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.
David, L.A., and Alm, E.J. (2011) Rapid evolutionary innovation during an Archean genetic expansion. Nature 469:93-96.
DesMarais, D.J., Strauss, H., Summons, R.E., and Hayes, J.M. (1992) Carbon isotope evidence for the stepwise oxidation of the Proterozoic environment. Nature 359:605-609.
Dickson, J.A.D. (2002) Fossil echinoderms as monitor of the Mg/Ca ratio of Phanerozoic oceans. Science 298:1222-1224.
Domagal-Goldman, S.D., Kasting, J.F., Johnston, D.T., and Farquhar, J. (2008) Organic haze, glaciations and multiple sulfur isotopes in the Mid-Archean Era. Earth and Planetary Science Letters 269:29-40.
Donnelly, T.H., Shergold, J.H., Southgate, P.N., and Barnes, C.J. (1990) Events leading to global phosphogenesis around the Proterozoic/Cambrian boundary. In Phosphorite Research and Development, Geological Society Special Publication 52:273-287.
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.
Dutkiewicz, A., Volk, H., George, S.C., Ridley, J., and Buick, R. (2006) Biomarkers from Huronian oil-bearing fluid inclusions: an uncontaminated record of life before the Great Oxidation Event. Geology 34:437.
Elie, M., Noueira, A.C.R., Nédélec, A. Trindade, R.I.F., and Kenig, F. (2007) A red algal bloom in the aftermath of the Marinoan Snowball Earth. Terra Nova 19:303-308.
England, G.L., Rasmussen, B., Krapez, B., and Groves, D.L. (2002) Paleoenvironmental significance of rounded pyrite in siliciclastic sequences of the Late Archean Witwatersrand Basin: Oxygen-deficient atmosphere or hydrothermal alteration. Sedimentology 49:1133-1136.
Falkowski, P.G., and Godfrey, L.V. (2008) Electrons, life and the evolution of Earth's oxygen

cycle. Philosophical Transactions of the Royal Society London B27:2705-2716.
Falkowski, P.G., and Isozaki, Y. (2008) The story of O_2. Science 322:540-542.
Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034-1039.
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.
Fayek, M.J., Utsunomiya, S., Pfiffner, S.M., Anovitz, L.M., White, D.C., Riciputi, L.R., Ewing, R.C., and Stadermann, F.J. (2005) Nanoscale chemical and isotopic characterization of Geobacter Sulfurreducens surfaces and bio-precipitated uranium minerals. Canadian Mineralogist 43:1631-1641.
Fennell, K., Follows, M., and Falkowski, P.G. (2005) The co-evolution of the nitrogen, carbon and oxygen cycles in the Proterozoic ocean. American Journal of Science 305:526-545.
Fike, D.A., Grotzinger, J.P., Pratt, L.M., and Summons, R.E. (2006) Oxidation of the Ediacaran ocean. Nature 444:744-747.
Flannery, D.T., and Walter, R.M. (2012) Archean tufted microbial mats and the Great Oxidation Event. Australian Journal of earth Sciences 59:1-11.
Fox, P., and Hendler, J. (2009) Semantic eScience: Encoding meaning in next-generation digitally enhanced science. In: The Fourth Paradigm: Data Intensive Scientific Discovery, Eds. T. Hey, S. Tansley, and K. Tolle, Microsoft External Research, pp. 145-150.
Fox, P., and Hendler, J. (2011) Changing the equation on scientific data visualization. Science 331:705-708.
Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., and Solomon, S. (2006) Semantically-enabled large-scale science data repositories. 5th International Semantic Web Conference (ISWC06), ed. Cruz et al., Springer-Verlag, Berlin, LNCS 4273:792-805.
Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J., and Middleton, D. (2009) Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. Computers and Geosciences, special issue on Geoscience Knowledge Representation for Cyberinfrastructure 35:724-738.
Fox, P., McGuinness, D.L., and the VSTO team (2011) Semantic cyberInfrastructure: The virtual solar-terrestrial observatory. Cambridge University Press monograph on Geoinformatics: Cyberinfrastructure for Solid Earth Sciences, Ed. C. Baru and R. Keller. pp. 21-36.
Frei, R., Gaucher, C., Poulton, S.W., and Canfield, D.E. (2009) Fluctuations in Precambrian atmospheric oxygenation recorded by chromium isotopes. Nature 461:250–253.
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, doi:10.1029/2001GC000217.

Golden, J., McMillan, M., Downs, R.T., Hystad, G., Stein, H.J., Zimmerman, A., Sverjensky, D.A., Armstrong, J., and Hazen, R.M. (2013) The Great Subsurface Oxidation "Event": Evidence from Re Variations in Molybdenite (MoS_2). Earth and Planetary Science Letters 366:1-5.

Grew, E.S., and Hazen, R.M. (2014) Beryllium mineral evolution. American Mineralogist 99:in press.

Guo, Q.J., Strauss, H., Kaufman, A.J., Schroder, S., Gutzmer, J., Wing, B., Baker, M.A., Bekker, A., Jin, Q.S., Kim, S.T., and Farquhar, J. (2009) Reconstructing Earth's surface oxidation across the Archean-Proterozoic transition. Geology 37:399-402.

Habicht, K.S., Gade, M., Thandrup, B., Berg, P., and Canfield, D.E. (2002) Calibration of sulfate levels in the Archean ocean. Science 298:2372-2374.

Halevy, I., Johnston, D.T., and Schrag, D.P. (2010) Explaining the structure of the Archean mass-independent sulfur isotope record. Science 329:204-207.

Halevy, I. (2013) Production, preservation, and biological processing of mass-independent sulfur isotope fractionation in the Archean surface environment. Proceedings of the National Academy of Sciences USA 110:17644-17649.

Halverson, G.P., Hoffman, P.F., Schrag, D.P., Maloof, A.C., and Rice, A.H.N. (2005) Toward a Neoproterozoic composite carbon-isotope record. Geological Society of America Bulletin 117:1-27.

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.

Hardie, L.A. (1996) Secular variation in seawater chemistry: An explanation for the coupled secular variation in the mineralogies of marine limestones and potash evaporites over the past 600 m.y. Geology 24:279-283.

Hardie, L.A. (2003) Secular variations in Precambrian seawtare chemistry and the timing of Precambrian aragonite seas and calcite seas. Geology 31:785-788.

Harel, A., Bromberg, Y., Falkowski, P.G., and Bhattacharya, D. (2014) The evolutionary history of redox metal-binding domains across the tree of life. Proceedings of the National Academy of Sciences USA, in press.

Hazen, R. M. (2005) Genesis: The Scientific Quest for Life's Origin. Washington, DC: Joseph Henry Press, 339 p.

Hazen, R.M. (2013) Paleomineralogy of the Hadean Eon: A preliminary list. American Journal of Science 313:807-843.

Hazen, R.M. (2014) Data-driven abductive discovery in mineralogy. American Mineralogist, in press.

Hazen, R.M., and J.S. Trefil (2009) Science Matters, 2nd Edition. New York: Doubleday, 360 p.

Hazen, R.M., Papineau, D., Bleeker, W., Downs, R.T., Ferry, J.M., McCoy, T.J., Sverjensky, D.A., and Yang, H. (2008) Mineral evolution. American Mineralogist 93:1693-1720.

Hazen, R.M., Ewing, R.J and Sverjensky, D.A. (2009) Evolution of uranium and thorium minerals. American Mineralogist 94:1293-1311.

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., Hysted, 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., Jones, A.P., Kah, L., and Sverjensky, D.A. (2013a) Carbon mineral evolution. Reviews of Mineralogy and Geochemistry 75:79-107.
Hazen, R.M., Sverjensky, D.A., Azzolini, D., Bish, D.L., Elmore, S., Hinnov, L., and Milliken, R.E. (2013b) Clay mineral evolution. American Mineralogist 98:2007-2029.
Hazen, R.M., Liu, X., Downs, R.T., Golden, J., Grew, E.S., Hystad, G., Estrada, C., and Sverjensky, D.A. (2014) Mineral evolution: Episodic metallogenesis, the supercontinent cycle, and the co-evolving geosphere and biosphere. Economic Geology, *in press*.
Hey, T., Tansley, S., and Tolle, K. [Editors] (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Redland, WA: Microsoft External Research.
Hofmann, A., Bekker, A., Rouxel, O., Rumble, D., and Master, S. (2009) Multiple sulfur and iron isotope composition of detrital pyrite in Archaean sedimentary rocks: A new tool for provenance analysis. Earth and Planetary Science Letters 286:436-445.
Holland, H.D. (1984) The Chemical Evolution of the Atmosphere and Ocean. Princeton Series in Geochemistry. Princeton University Press, Princeton, New Jersey.
Holland, H.D. (2006) The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society London 361:903-915.
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.
Johnston, D.T., Poulton, S.W., Dehler, C., Porter, S., Husson, J., Canfield, D.E., and Knoll, A.H. (2010) An emerging picture of Neoproterozoic ocean chemistry: Insights from the Chuar Group, Grand Canyon, USA. Earth and Planetary Science Letters 290:64-73.
Jonsson, C. M., Jonsson, C. L., Sverjensky, D. A., Cleaves II, H. J., and Hazen, R. M., 2009. Attachment of l-Glutamate to rutile (α-TiO2): A potentiometric, adsorption, and surface complexation study. Langmuir 25:12127-12135.
Kah, L.C., and Riding, R. (2007) Mesoproterozoic carbon dioxide levels inferred from calcified cyanobacteria. Geology 35:799-802.
Kah, L.C., Lyons, T.W., and Frank, T.D. (2004) Low marine sulphate and protracted oxygenation of the Proterozoic biosphere. Nature 431:834-837.
Kasting, J.F. (2001) The rise of atmospheric oxygen. Science 293:819-820.
Kasting, J.F. (2013) What caused the rise of atmospheric O_2? Chemical Geology 362:13-25.
Kasting, J.F. and Siefert, J.L. (2002) Life and the evolution of Earth's atmosphere. Science 296:1066-1068.
Kaufman, A.J., Johnston, D.T., Farquhar, J., Masterson, A.L., Lyons, T.W., Bates, S., Anbar, A.D., Garvin, J., and Buick, R. (2007) Late Archean biospheric oxygenation and atmospheric evolution. Science 317:1900-1903.
Keller, B., and Schoene (2012) Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gya ago. Nature 485:490-493.
Khandewal, A., and Fox, P. (2014) Towards a web ontology language for descriptions of continuous processes. Journal of Web Semantics, under review.
Kiessling, W., Simpson, C., and Foote, M. (2010) Reefs as cradles of evolution and sources of biodiversity in the Phanerozoic. Science 327:196-198.
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 London B368:20120257.

Klein, C. (2005) Some Precambrian banded iron-formations (BIFs) from around the world: Their age, geologic setting, mineralogy, metamorphism, geochemistry, and origin. American Mineralogist 90:1473-1499.

Knoll, A.H. (2003a) Life on a Young Planet. Princeton University Press, Princeton, New Jersey.

Knoll, A.H. (2003b) Biomineralization and evolutionary history. Reviews in Mineralogy and Geochemistry 54:329-356.

Knoll, A.H., Bambach, R.K., Payne, J., Pruss, S., and Fischer, W. (2007) A paleophysiological perspective on the end-Permian mass extinction and its aftermath. Earth and Planetary Science Letters 256:295-313.

Knoll, A.H. (2013) Systems paleobiology. Geological Society of America Bulletin 125:3-13.

Konhauser, K.O., Amskold, L., Lalonde, S.V., Posth, N.R., Kappler, A., and Anbar, A. (2007) Decoupling photochemical Fe(II) oxidation from shallow-water BIF deposition. Earth and Planetary Science Letters 258:87-100.

Kump L.R., and Barley, M.E. (2007) Increased subaerial volcanism and the rise of atmospheric oxygen 2.5 billion years ago. Nature 448:1033-1036.

Kump, L.R., Kasting, J.F., and Barley, M.E. (2001) Rise of atmospheric oxygen and the "upside down" Archean mantle. Geochemistry, Geophysics, Geosystems, 2, Paper #2000GC000114.

Laakso, T.A., and Schrag, D.P. (2014) Regulation of atmospheric oxygen during the Proterozoic. Earth and Planetary Science Letters 388:81–91.

LaBerge, G.L. (1973) Possible biological origin of Precambrian iron-formations. Economic Geology 68:1098-1109.

Large, R.R., Halpin, J.A., Danyushevsky, L.V., Maslennikov, V.V., Bull, S.W., Long, J.A., Gregory, D.D., Lounejeva, E., Lyons, T.W., Sack, P.J., McGoldrick, P.J., and Claver, C.R. (2014) Trace element content of sedimentary pyrite as a new proxy for deep-time ocean-atmosphere evolution. Earth and Planetary Science Latters 389:209-220.

Laskey, K.B. (2006) MEBN: A Logic for Open-World Probabilistic Reasoning, MARS online http://hdl.handle.net/1920/461.

Lebo, T., Erickson, J., Ding, L., Graves, A.,Williams, G., DiFranzo, D., Li, X., Michaelis, J,, Zheng, J., Flores, J., Shangguan, Z., McGuinness, D.L., and Hendler, J. (2011) Producing and Using Linked Open Government Data in the TWC LOGD Portal, in Linking Government Data, ed. Wood, David, pp. 51-72, Springer, New York.

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

Lovley, D.R., Phillips, E.J.P., Gorby, Y.A., and Landa, E.R. (1991) Microbial reduction of uranium. Nature 350:413-416.

Lowenstein, T.K., Timofeeff, M.N., Brennan, S.T., Hardie, L.A., and Demicco, R.V. (2001) Oscillations in Phanerozoic seawater chemistry: Evidence from fluid inclusions. Science 294:1086-1088.

Lyons, T.W., Peinhard, C.T., and Planavsky, N.J. (2014) The rise of oxygen in Earth's early ocean and atmosphere. Nature 506:307-314.

Ma, X., Fox, P., Tilmes, C., Jacobs, K., and Waple, A. (2014) Providing global change information for decision making: capturing and presenting provenance. Nature Geosciences, in press.

Markl, G., Marks, M.A.W., Derrey, I., and Gühring, J.-E. (2014) Weathering of cobalt arsenides: Natural assemblages and calculated stability relations among secondary Ca-Mg-Co arsenates and carbonates. American Mineralogist 99:44-56.

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.

Mojzsis, S.J., Coath, C.D., Greenwood, J.P., McKeegan, K.D., and Harrison, T.M. (2003) Mass-independent isotope effects in Archean (2.5 to 3.8 Ga) sedimentary sulfides determined by ion microprobe analysis. Geochimica et Cosmochimica Acta 67:1635-1658.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2006) Introduction to Linear Regression Analysis, Fourth edition. Hoboken, New Jersey: John Wiley & Sons, 612 pp.

Murakami, T., Utsinomiya, S., Imazu, Y., and Prasadi, N. (2001) Direct evidence of late Archean to early Proterozoic anoxic atmosphere from a product of 2.5 Ga old weathering. Earth and Planetary Sciences Letters 184:523-528.

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

Nance, R.D., Murphy, J.B., and Santosh, M. (2014) The supercontinent cycle: A retrospective essay. Gondwana Research 25:4-29.

Narock, T., and Fox, P.A. (2012) From science to e-science to semantic e-science: A heliophysics case study. Computers & Geosciences 46:248-254.

Niu, F., Zhang, C., R\'e, C., and Shavlik, J. (2012) DeepDive: Web-scale knowledge-base construction using statistical learning and inference. Second International Workshop on Searching and Integrating New Web Data Sources. *VLDS*, pp. 25-28.

Noffke, N., Christian, D., Wacey, D., and Hazen, R.M. (2013) Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old Dresser Formation, Pilbara, Western Australia. Astrobiology 13:1103-1124.

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

Ohmoto, H., Watanabe, Y., Ikemi, H., Poulson, S.R., and Taylor, B.E. (2006) Sulphur isotope evidence for an oxic Archean atmosphere. Nature 442:908-911.

Olson, S.L., Kump, L.R., and Kasting, J.F. (2013) Quantifying the areal extent and dissolved oxygen concentrations of Archean oxygen oases. Chemical Geology 362:35-43.

Ono, S., Eigenbrode, J.L., Pavlov, A.A., Kharecha, P., Rumble, D. III, Kasting, J.F., and Freeman, K.H. (2003) New insights into Archean sulfur cycle from mass-independent sulfur isotope records from the Hamersley Basin, Australia. Earth and Planetary Science Letters 213:15-30.

Ono S., Shanks, W.C., Rouxel, O., and Rumble, D. (2007) S-33 constraints on the seawater sulfate contribution in modern seafloor hydrothermal vent sulfides. Geochimica et Cosmochimica Acta 71:1170-1182.

Papineau, D., Mojzsis, S.J., and Schmitt, A.K. (2007) Multiple sulfur isotopes from Paleoproterozoic Huronian interglacial sediments and the rise of atmospheric oxygen. Earth and Planetary Science Letters 255:188-212.

Parikh, S. J., Kubicki, J. D., Jonsson, C. M., Jonsson, C. L., Hazen, R. M., Sverjensky, D. A., and Sparks, D. L., 2011. Evaluating glutamate and aspartate binding mechanisms to rutile (α-TiO_2) via ATR-FTIR spectroscopy and quantum chemical calculations. Langmuir 27:1778-1787.

Partin, C.A., Bekker, A., Planavsky, N.J., Scott, C.T., Gill, B.C., Podkovyrov, V., Maslov, A., Knohauser, K.O., Lalonde, S.V., Love, G.D., Poulton, S.W., and Lyons, T.W. (2013) Large-

scale fluctuations in Precambrian atmospheric and oceanic oxygen levels from the record of U in shales. Earth and Planetary Science Letters 369-370:284-293.

Pavlov, A.A., Hurtgen, M.T., Kasting, J.F., and Arthur, M.A. (2003) Methane-rich proterozoic atmosphere. Geology 31:87-90.

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.

Peters, S.E., and Heim, N.A. (2010) The geological completeness of paleontological sampling in North America. Paleobiology 36:61-79.

Planavsky, N., Rouxel, O., Bekker, A., Shapiro, R., Fralick, P., and Knudsen, A. (2009) Iron-oxidizing microbial ecosystems thrived in late Paleoproterozoic redox-stratified oceans. Earth and Planetary Science Letters 286:230–242.

Prokoph, A., Shields, G.A., and Veizer, J. (2008) Compilation and time-series analysis of a marine carbonate database $\delta^{18}O$, $\delta^{13}C$, $^{87}Sr/^{86}Sr$, and $\delta^{34}S$ through Earth history. Earth Science Reviews 87:113-133.

Pruss, S., Finnegan, S., Fischer, W.W., and Knoll, A.H. (2010) Carbonates in skeleton-poor seas: New insights from Cambrian and Ordovician strata of Laurentia. Palaios 25:73-84.

Rasmussen, B., and Buick, R. (1999) Redox state of the Archean atmosphere: Evidence from detrital heavy minerals in ca.3250-2750 Ma sandstones from the Pilbara Craton, Australia. Geology 27:115-118.

Rasmussen, B., Fletcher, I.R., Brocks, J.J.; et al. (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455:1101-1109.

Rasmussen, B., Fletcher, I.R., Bekker, A., Muhling, J.R., Gregory, C.J., and Thorne, A.M. (2012) Deposition of 1.88-billion-year-old iron formations as a consequence of rapid crustal growth. Nature 484:498-501.

RDF Data Cube Vocabulary, World Wide Web Consortium, on line and http://www.w3.org/TR/vocab-data-cube/ (last accessed 7 February 2014).

Reith, F., Rogers, S.L., McPhail, D.C., and Webb, D. (2006) Biomineralization of gold: Biofilms on bacterioform gold. Science 313:233-236.

Rodriquez, A., and Laio, A. (2014) Clustering by fast search and find of density peaks. Science 344:1492-1496.

Rohrbach, A., Ballhaus, C., Golla-Schindler, U., Ulmer, P., Kamenetsky, V.S., and Kuzmin, D.V. (2007) Metal saturation in the upper mantle. Nature 449:456-458.

Rozell, E., Fox, P.A., Zheng, J., and Hendler, J. (2012) S2S architecture and faceted browsing applications. Proceedings of the 21st International Conference Companion on World Wide Web, pp. 413-416, ACM.

Rye, R. and Holland, H.D. (1998) Paleosols and the evolution of atmospheric oxygen: A critical review. American Journal of Science 298:621-672.

Schwartz, G., Mendel, R.R., and Ribbe, M.W. (2009) Molybdenum cofactors, enzymes and pathways. Nature 460:839-847.

Scott, C., Lyons, T.W., Bekker, A., Shen, Y., Poulton, S.W., and Anbar, A.D. (2008) Tracing the stepwise oxygenation of the Proterozoic ocean. Nature 452:456-459.

Shoshani, A., and Rotem, D. [Editors] (2010) Scientific Data Management: Challenges, Technology, and Deployment. Boca Raton, FL: Taylor & Francis, 534 p.

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.
Suzuki, Y., and Banfield, J.F. (1999) Geomicrobiology of uranium. Reviews in Mineralogy 38:393-432.
Sverjensky, D.A., and Lee, N. (2010) The Great Oxidation Event and mineral diversification. Elements 6:31-36.
Sverjensky, D. A., Harrison, B., and Azzolini, D. (2014) Water in the deep Earth: the dielectric constant and the solubilities of quartz and corundum to 60 kb and 1,200°C. Geochimica et Cosmochimica Acta, in press.
Sverjensky, D.A., Hemley, J.J., and D'Angelo, W.M. (1991) Thermodynamic assessment of hydrothermal alkali feldspar-mica-aluminosilicate equilibria. Geochimica et Cosmochimica Acta 55:989-1004.
Tan, P.-N., Steinbach, M., and Kumar, V. (2006) Introduction to Data Mining. Boston, MA: Addison-Wesley.
Tilmes, C., Fox, P.A., Ma, X., McGuinness, D.L., Privette, A.P., Smith, A., Waple, A., Zednik, S., and Zheng, J.G. (2013) Provenance representation for the National Climate Assessment in the Global Change Information System. IEEE Transactions on Geoscience and Remote Sensing 51:5160-5168.
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.
Towe, K.M. (2002) The problematic rise of Archean oxygen. Science 295:798-799.
Trefil, J.S., and R.M. Hazen (2012) The Sciences: An Integrated Approach, 7th edition. Hoboken, NJ: John Wiley & Sons, 555 p.
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.
Walker, J.C.G. (1977) Evolution of the Atmosphere. Macmillan, New York.
Walker, J.C.G., Hays, P.B., and Kasting, J.B. (1981) A negative feedback mechanism for the long-term stabilization of Earth's surface temperature. Journal of Geophysical Research 86:9776-9782.
West, P., Rozell, E., Zednik, S., Fox, P., and McGuinness, D.L. (2009) Semantically enabled temporal reasoning in a virtual observatory. In OWL Experiences and Direction (OWLED), R. Hoekstra and P.F. Patel-Schneider, editors.
Williams, R.J.P. (1981) Natural selection of the chemical elements. Proceedings of the Royal Society London B213:361-397.
Zahnle, K.J., Catling, D.C., and Claire, M.W. (2013) The rise of oxygen and the hydrogen hourglass. Chemical Geology 362:26-34.

Project Budget—Justification and Other Funding

Budget Justification: The total estimated cost for this study of Earth's changing oxidation state and initial development of the Deep-Time Data Infrastructure is $4,071,926. We request $1,390,676 from the W. M. Keck Foundation for 3 years to initiate this program, primarily for support of early-career investigators ($990,676), a Project Manager ($305,000), plus travel ($76,000), and supplies ($19,000). We will provide (from institutional and other grant support) an additional $330,000 for PI and Co-I salaries; $690,000 in matching funds for postdoctoral fellows and associates; $380,000 for graduate student researchers; $75,000 for supplies; and $150,000 for travel (including our annual meetings with invited participants). We also commit $1,056,250 in facilities/overhead for non-Keck expenses (65% rate). Please note that exact calculation of matching funds is difficult; here we report good-faith conservative estimates of contributions from the 6 partner institutions.

The Project Manager will be based at the Carnegie Institution, but will travel frequently to the other nodes. Researchers at the Carnegie Institution ($577,676 from Keck funds) and the University of Arizona ($168,000) will enhance existing mineralogy and petrology data resources. Researchers at Harvard University ($152,000) will develop the Precambrian paleobiology database and correlate changes in fossil microbial communities to Earth's evolving near-surface oxidation state. Scientists at Rutgers University ($152,000) will explore temporal changes of transition metals in enzymes and correlate those changes with observations of redox-sensitive elements. Researchers at Johns Hopkins University ($152,000) will support integration of geochemical modeling with the Deep-Time Data Infrastructure, and will apply reaction path modeling to evaluate redox conditions implied by temporal changes in mineralogy. Scientists at RPI ($189,000) will lead the effort to synthesize and integrate these varied data resources to create a Deep-Time Data Infrastructure.

This proposal does not request funds for equipment, construction, or remodeling.

Other Funding: We will leverage Keck Foundation funds with resources from the 6 lead investigators, each of whom receives significant institutional and grant support for traditional inductive and deductive discovery in their specialties (but not in Deep-Time Data Infrastructure or its applications). Discipline-bound research of the Co-Investigators is currently supported by > $5 million from agencies and foundations. Furthermore, key data resources upon which we will rely are supported by as much as $50 million per year. We have funding in hand from the Sloan Foundation ($400 K to Hazen for deep-time mineral data development; $500 K/year to Fox in part for data infrastructure development) and the Robert and Margaret Hazen Foundation ($100 K/year pledged for the duration of this project). In addition, each of the 6 lead institutions provides significant support in the form of salary, infrastructure, and grants to support graduate students and postdoctoral fellows.

Pending Funds: Hazen has proposals pending that incorporate deep-time data acquisition from the Simons Foundation ($560 K), NSF ($450 K), and NASA (proposed Astrobiology Institute nodes at Rutgers University and at the Carnegie Institution).

No part of this proposal has yet been considered (or declined) by any other funding agency.

Need for Keck Support: Connecting data resources across geological and biological disciplines in a formal way to search for new patterns is a high risk/high reward endeavor that is not likely to be funded by any government agency, nor by most other private entities. Funding of this project will require a visionary foundation, because the abductive, data-driven approach to discovery is both nontraditional and broadly interdisciplinary. Unlike most "hypothesis driven" research programs, we cannot predict what we will find, but we can be certain that a more vivid history of our complex planetary home will result.

Recognition Statement

The principal anticipated outcome of this effort will be a web-based "Deep Time Data Infrastructure"—an open-access resource for data-driven discovery. The home page of this web site will prominently feature the W. M. Keck Foundation logo and the statement "We gratefully acknowledge support from the W. M. Keck Foundation, without whom this resource would not have been possible."

We anticipate more than 20 related publications during the 3-year span of this proposal, as well as numerous subsequent publications, as we continue to exploit and disseminate this extraordinary new resource. We will acknowledge the generous support of the W. M. Keck Foundation in every related publication. We will also incorporate the Keck Foundation logo into every seminar, lecture, poster, and other public presentation. Hazen had more than 140 public lectures, including 21 named lectureships, in the past 4 years, and all of the co-investigators give many presentations every year. We therefore anticipate numerous opportunities to acknowledge the generosity of the W. M. Keck Foundation.