Towards a New Distributed Platform for Integrative Geoscience: An EarthCube Design Approach and Prototype Plan

Shanan E. Peters¹, Miron Livny², Kurt Squire³, Jack Williams⁴, Bill Hibbard⁵, Puneet Kishor⁶, Sam Batzli⁷, Patrick McLaughlin⁶, Susan Millar³, Tom Whittaker⁵, Nancy Wiegand⁵, Terry Millar⁸

¹Department of Geoscience, ²Morgridge Institute for Research Center for High Throughput Computing, ³Morgridge Institute, School of Education, ⁴Center for Climate Research, ⁵Space Science and Engineering Center, ⁶Wisconsin Natural History and Geological Survey, ⁷Cooperative Institute for Meteorological Satellite Studies, ⁸Provost’s Office, University of Wisconsin-Madison and UW Extension, Madison, WI 53706

Overview:
Understanding the current status and behavior of Earth systems, their long-term evolution, and the effect that human activity will have on their sustainability critically depends on the development of new approaches to integrating and interrogating disparate types of observed and modeled geoscience and biological data. NSF’s EarthCube’s initiative, whose stated mission is to “integrate data and information for knowledge management across the Geosciences,” is, therefore, timely and serves an important and very real national need.

To realize the ambitious objectives of EarthCube, we believe that a fundamental shift in how we, as a broadly-defined geoscience community, think about data distribution and modes of truly interdisciplinary collaboration and data integration is required. Our design approach is built upon a distributed, internet-like paradigm that loosely couples existing geoscience data resources using a fundamental organizing principle – space-time – combined with semantic web approaches that work from the data up.

We seek to test our design approach by establishing a prototype of a community-derived and -governed cyberinfrastructure (CI) that leverages both the face-to-face communication processes that are possible on a campus and our diverse geoscience community itself, which includes atmospheric and ocean sciences, geophysics and geodesy, sedimentary geology, paleobiology, structural geology, climatology, paleoclimatology, hydrology, hydrogeology, soil science, limnology, geochemistry, biological science, geological engineering, computer science, education, social science and the Wisconsin Geological and Natural History Survey.

Our general approach to the challenge of building a CI for such a diverse community rests fundamentally on experiences in many relevant, parallel domains showing that:

• Top-down centralization of data storage, data management, data standards protocols, and ontologies is often challenging, even when conducted within the boundaries of the individual domains that generate and use data and data products for scientific objectives.

• Portals that provide passive data access to human users, no matter how well developed and managed, are generally most useful to specialists who already know the data and to

2 http://www.nsf.gov/geo/earthcube/index.jsp
3 This Design Approach paper and four related Technical Solution papers have been prepared by a newly formed, collaborative working group at the University of Wisconsin-Madison that spans many colleges, centers, departments, and partners.
individuals who are able to train themselves and their communities in how to discover and use those data.

Therefore, **simply scaling-up or cementing in any of the current approaches in any of the geoscience domains—no matter how successful and well-designed—should not be the path forward.** Instead, we aim to adapt and design, and then test and implement, a prototype for building an innovative geoscience- and data-centric discovery and analysis platform that:

- accommodates and fully leverages, in an inclusive way, the diverse geoscience data, information, and knowledge systems that already exist and that will continue to be used and supported by individuals, communities, and organizations in geoscience;

- is based on a set of invariant data/information/knowledge structure and management principles that over time may be instantiated by different technologies;

- facilitates the development of creative, data-driven, third-party applications for discovering and exploring geoscience data that can provide added scientific and educational value to existing community data and knowledge resources;

- addresses head-on the significant challenges of “scaling-up” by including, right from the start, design elements that engage an international community;

- is based on a set of sound partnership and community-building principles and practices that involve trust, collaboration, citation, and community governance.

Below we describe how we believe our design approach for EarthCube will fundamentally improve geoscience research and teaching. We then provide a more concrete description of the CI we envision and outline an inclusive model of community-governance that focuses on developing an initial set of protocols and then deploying those protocols within the existing structure used by domain scientists. Finally, we provide a description for the initial phases of the operations and sustainability of our implementation, with special emphasis on how our inherently open and distributed design approach is sustainable for the long term.

**Vision for EarthCube:**

Our vision for EarthCube is science- and data-focused. The icon for our vision consists of a tetrahedron linking the solid, fluid, living Earths and the rock record. This icon illustrates the basic principle that motivates us: **Earth systems are interconnected, and there is no logical “center” that should dominate this effort.** As one very simple example of Earth systems connections, every year, pCO$_2$ measurements from Mauna Loa,

---

Hawaii go up and down by more than 5 ppm, a rate of change that far exceeds the steady ~1 ppm annual increase that can be attributed to human activity. This strong annual variability in pCO₂ is the result of two things: 1) the annual growth and decay of CO₂, H₂O, and sunlight-derived biomass, and 2) a present-day northern hemisphere bias in the distribution of continental land area. Thus, today our planet literally breaths, and it does so with a period of one year because plate tectonics has resulted in an unequal hemispherical distribution of landmass, because Earth’s rotational axis is tilted, and because large amounts of terrestrial (and marine) photosynthetic biomass forms and decays in the northern hemisphere as winter comes and goes. These types of interconnections between Earth systems occur on many different spatial and temporal scales.

The unifying structure we propose for advertising, discovering, and connecting locally-managed and highly diverse data spanning all of Earth systems science is space-time. A space-time data model (consisting of points, lines, polygons, and volumes referenced to discrete or continuous times or time intervals) is simple in concept, but complicated in detail because the data span an enormous range of scales, conventions, degrees of specification, and precision and uncertainty. Nevertheless, all data in geoscience (and many other sciences) can be fundamentally reduced to a common space-time structure.

Figuratively speaking, the use of a data structure Δ₁ for research entails integrating and analyzing subsets of Δ₁ in order to produce information and new knowledge. There may be many different specific data elements in Δ₁, but here we specifically emphasize:

\[ \text{space (} x, y, z \text{) and time (} t \text{)}. \]

Coordinates of records in Δ₁ can be written \( \Delta_{i}(\tau, d^i) \), where \( \tau = \{x, y, z, t\} \), and \( d^i \) is a vector of other domain-specific data types, each with its own semantics and structures. Research using only \( \Delta_1 \) can be thought of as accessing subsets of the subspace \( \tau \times d^i \). Typically, and to oversimplify, Earth systems research is supported by individual data structures:

\[ \Delta_{i}(\tau, d^i), \Delta_{i}(\tau, d^j), \Delta_{i}(\tau, d^k), ..., \Delta_{i}(\tau, d^n), \]

But, most research tends to be restricted to a small number of the subspaces \( \tau \times d^i \). **We propose that a short term vision for EarthCube should be to creatively facilitate new data discovery using arbitrary subsets of the full space:**

\[ \tau \times \overrightarrow{d_1} \times \overrightarrow{d_2} \times ... \overrightarrow{d_n}. \]

The space-time coordinates, \( \tau = \{x, y, z, t\} \), common to geoscience data provide a deep organizational principle for the facilitation of EarthCube that is consistent with current data structures, the advent of new data structures, and the ability of the facilitation to adjust to all such changes. This simple organizational principle is extremely powerful because it can meaningfully connect data from a weather station to a stratigraphic rock section to a fossil to an oak tree to an earthquake to a gold deposit to a coupled ocean-atmosphere climate model run,
and even to human sociopolitical and economic data and information. A specific existing example that combines diverse data under a version of this paradigm is VisAD, discussed in detail in an accompanying white paper.\textsuperscript{5} Establishing connections in the many other dimensions that characterize geoscience data can be achieved through semantic web approaches, which are built from the data-up.\textsuperscript{6}

In addition to being technically focused, our vision for EarthCube is closely linked to our desire to rapidly enable the creation of new data-driven applications for teaching and learning. As an example, imagine a student out in the field with her mobile internet device. This student connects to an EarthCube (EC) center instantiation via an HTML5 interface, which is standards-based and therefore accessible to all mobile platforms. The server (with the student’s permission) automatically knows where she is by virtue of the location services provided by that mobile device. Because EC knows location, it knows, based on its distributed network of geospatially accessible data sources, what rocks the student is standing on, what rocks are beneath her, what fossils are in those rocks, the soil type she is on, the economic uses of the geological resources around her, the trees that occur natively there, her current plate tectonic velocity, the current weather conditions and average climate conditions, the underlying mantle thermal structure, the local gravity field, recent seismic activity and other geohazards, etc. In short, because location is specified, EC will be able to draw simultaneously on many different and independent data and information resources that are maintained and vetted by the experts in their respective communities but that are published to EC in a unified way using the simplest possible space-time-based protocols.

While this student is exploring the current EarthCube status of her location, imagine her wondering "what did the world look like at the time that this rock formed"? She might then select a lithostratigraphic rock unit, which identifies a semantic geoscience data element (e.g., “Dakota Formation”). EarthCube then discovers the space-time relationship of “Dakota Formation” using that semantic and serves out paleogeographic context (space is not constant in time), general circulation model results for that time interval under a range of model parameterizations, historical estimates of biodiversity and regional ecology, paleontological and geochemical records etc. and then compares these historical conditions to the conditions that are around her today, quantitatively in some cases, qualitatively in others.

Next, imagine her wondering, “what will this world look like in 250 years, if CO\textsubscript{2} rises 300 ppm?” She drags a time-slider interface element forward into the future. EC then serves out to the user general circulation model outputs for that condition, modeled responses in vegetation, climate, etc. The comparisons once again get made directly to the EarthBase current status of the location. Metadata associated with all of the models and primary data sources would be accessible, should she need to pursue more detailed scientific questions that require retrieving the lower-level or raw data from its original source or understand data integrity and uncertainty.

Dr. Phoebe Cohen, Education and Outreach Specialist for MIT’s NASA Astrobiology Team, described the educational possibilities of our vision in the following way:

\textsuperscript{5} See complementary Technical Solution paper by Hibbard et al., An EarthCube Technical Solution: the VisAD Data Model.

\textsuperscript{6} See complementary Technical Solution paper by Wiegand and Revercomb.
Imagine an 8th grade Earth science teacher – let’s call her Jodie. Jodie is new to teaching, and new to the rural area she teaches in. She wants to give her students an authentic field experience, but she doesn’t know how – her degree is in education and she’s only taken a few Earth science courses. Jodie visits EC and quickly is fed a wealth of information about her area. In just minutes, Jodie is able to discover what types of rock and soil their school is built on. She can look up climate data for their region going back a century, and access climate models that can tell her what the climate might be like in another 100 years. She notices that there is a fossil locality nearby and makes a note to find out how to take her students there on a future field trip. She discovers that there was an earthquake in the area 10 years ago and that the entire region was a shallow inland sea 85 million years ago. She then accesses curriculum guides and the Science Common Core State standards to help her incorporate all of this information into her classroom. A field trip template linked to a smart phone app makes it easy for Jodie to lead her class on trips around their school, creating meaningful and rich experiences for her students.”

We believe that this is an achievable outcome for EarthCube in a relatively short time frame.

**Conceptual CI Architecture:**

As noted in the ACCI Task Force on Data and Visualization Final Report, data sets and their associated metadata are critical records of scientific output and information about our natural world, and many important electronic data sets must be preserved in ways that allow their use into the foreseeable future. We present a parallel Technical Solution paper where we give an example of a relevant response to data-intensive science of high-energy physics - the Open Science Grid.

Here we take as a definition of ‘cyberinfrastructure’ the one adopted by the NSF Advisory Committee for Cyberinfrastructure:

> “Cyberinfrastructure consists of computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible.”

Our design approach attempts to adhere to this definition by building on many lessons learned (see accompanying Technical Solution papers) and by *interfering as little as possible with the geoscience communities that have already made important and successful decisions about how to manage and distribute their data*. We will do this by:

1. developing a guiding set of CI abstractions that can serve the evolution of data/information/knowledge generation and needs in geoscience;
2. developing an initial set of geoscience-specific protocols for advertising (in a machine-driven internet sense) locally stored and managed data using these abstractions;
3. supporting computer scientists to work with geoscience communities to facilitate the deployment and implementation of these protocols in their own local data facilities;

---

4) enabling collaboration among domain geoscientists, computer scientists, and teaching with technology specialists to develop the first of many possible instantiations of a data integration center that is built upon the distributed foundational web services of EarthCube, but whose sole purpose is to interrogate, explicitly link, and then dynamically cache and expose geoscience data resources;

An analogy for the EarthCube CI center we aim to build is Google, a centralized organization that generates and manages no content of its own, but that has developed algorithms for crawling locally-managed content exposed by www protocols and has hardware facilities for caching results and making prioritized lists of content based on semantics and site rankings.

There are many advantages of this design approach to EarthCube. For example, we do not need to reinvent the wheel. There is a very large and active community that is dedicated to developing web-focused geospatial standards, notably the Open Geospatial Consortium\(^8\). Recent advances in web service architectures also hold promise. Other communities have also been active in this general area, including GEON, GBIF, DataOne, GCMD, NSDIC, Reverb, Unidata, EOSDIS, and others. Thus, we believe that some of the tools required to build and implement the distributed components of EarthCube already exist, at least in concept.

What has been missing or inadequate up to this point is: 1) the engagement of a large number of domain scientists who best know their data and its present and future uses, 2) a science-driven paradigm for how existing and diverse data elements should be openly exposed to a network or CI and then related to one another on a fundamental and scientific level, and 3) a shared vision of, and sufficient resources to create, a geoscience-focused CI for data exploration and discovery that is built first upon a robust scientific paradigm and second upon a solid IT/CS architecture (as opposed to focusing on building databases and portals for hypothetical data that could be added to them at some point).

**Web Service Architecture:** A web service-based architecture, built on RESTful principles, forms the foundation of our initial distributed design approach for EarthCube. Web services provide many specific advantages. We note that many of these advantages are enumerated in the technical paper submitted by IRIS, which has implemented RESTful web services to great effect in their own center-level geophysical data and data product archive facility. Our conceptual CI approach is collaborative in spirit and in practice with IRIS's efforts.

IRIS's highly successful web services have focused on vertical distribution of data and data products within the geophysics community and beyond. Our effort is focused on achieving a high-level of horizontal integration of many very different data types, from satellite-collected ocean and atmospheric data to climate model simulation results to hydrologic data as well as fossil, geochemical, biological, and rock record data. Thus, our focus is on the development of protocols that can be deployed within a RESTful architecture and that expose data in a consistent and fundamental space-time dimension. Other dimensions of data integration will be accomplished using the semantics and attributes of the data and metadata that become exposed by these space-time-focused web services. **We believe that the combination of**

---

\(^8\) [http://opengeospatial.org](http://opengeospatial.org).
IRIS’s experience and success in developing robust vertical web services and our horizontally-focused but highly congruent CI approach comprise a powerful starting point for rapidly exposing, algorithmically exploring, and then dynamically relating all of the geoscience data types that must ultimately become part of EarthCube.

Web services also have several disadvantages, but most of these characterize the internet generally. For example, if a third party uses a web service and the owners of that service make a change to the data or the service itself, the third party’s ability to use the service will be interrupted. Web services are, therefore, not highly reliable. However, the internet is unreliable in exactly the same way (e.g., broken hyperlinks, somebody powers down a server accidentally, etc.) but these limitations are outweighed by the advantages of such a high-degree of flexibility and interoperability. Part of our challenge is to improve reliability by providing communities with guidance for how to build EarthCube-ready web services that are robust and dependable.

**Design Process:**

Not every step in the data chain may produce data suitable for publication and/or dissemination on the web, and not every level of data may be usable without specialized training, software, or understanding. Discipline specific scientific data formats have existed for as long as science, and for good reason. Although specialized data formats do solve discipline-specific problems, they also pose a hurdle for data sharing and reuse, particularly in an inter-disciplinary context. Yet, somewhere in the data chain, data do become suitable for publishing on the web. It is that step that is the target of EarthCube. Thus, **an important component of our design process involves working with a diverse geoscience community to develop basic principles and guidelines for publishing geoscience data and data products to EarthCube using web services.** Much as journals have principles and guidelines that must be adhered to prior to publishing scientific papers, our design process for EarthCube must address two important aspects of data publication, discussed below.

**Licensing and Citation:** When all is said and done, after all the advancements in the technology and architecture and protocols, **the only way scientists and other data creators will participate is if they see some benefit for themselves.** The obvious benefit is the implicit one — every data creator is also a user, and when the set of participants is large, the chance of finding a data creator with the data that one wants is also higher. That said, the two following mechanisms, when implemented successfully, will make participating in EarthCube more obviously attractive: 1) being able to convey to potential users what they can do with one’s data, and 2) receiving accreditation for data publication by knowing who is using one’s data and for what purpose.

The first involves two parts: 1) establishing a way to express one’s desired use-conditions, in other words, a license, and 2) implementing a way to convey that license. Creative Commons (CC) licenses, while designed for creative content, have characteristics that make them potentially useful for data.

Recognizing that CC licenses, as structured, were unsuitable for scientific data, CC established CC-Zero (CC0), a “no rights reserved” waiver that enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive interests in their...
works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.” We are testing ways of implementing CC licenses as both RDFa and JSON packets. Every data request, carried out by requesting a resource from a unique Uniform Resource Identifier (URI), the foundation of web services, will be accompanied by its licensing information.

Tracking the usage of data published to EarthCube and providing meaningful citation summaries for data providers is more challenging because usage can be classified into many different types. We are working on ways to track basic usage (such as number of requests to a specific web service data element) as well as more complex usage statistics that will provide all EarthCube data providers with rich, real-time data citation summaries.

Community-Based Governance Model:
Although the technological challenges of implementing a inclusively horizontal, scalable EarthCube are considerable, an equal or greater challenge is strengthening and interlinking the many different geosciences communities of practice. The US geosciences community of geoscientists numbers tens of thousands, and is divided into many domain-specific communities, often called disciplines, sub-disciplines, and sub-sub disciplines. Each of these has its own highly specialized ontologies, conceptualizations, data models, and governing research questions. The linkages among these many communities of practice vary widely in number, strength, and kind, with some communities well connected to others and others less so. Even scientists who are very good at the technical aspects of data management and distribution face translational challenges when working with scientists from other disciplines.

Cross-disciplinary communication problems are as old as science itself, and we do not claim to have “the answer” to these problems. However, we pose and seek to find answers to a promising question: Given limited resources and a vast and heterogeneous US geosciences community, what is the most effective way of creating strong communities of practice that can share best practices and move the entire geoscience community towards an open and frictionless transfer of data and knowledge?

The US geosciences community can be very roughly categorized into three broad communities of practice: 1) Informaticists, or scientists and computer scientists with expertise in information management and knowledge systems; 2) Data Generators, or individual investigators/research groups that are actively collecting and using observational and experimental data; and 3) Synthesizers, a rapidly growing community of scientists who are exploring research questions that demand some level of data synthesis and analysis/meta-analysis.

Each community of practice has its real and urgent data infrastructure, and therefore EarthCube-related, needs, and all communities are highly interdependent. A true EarthCube effort must ultimately make advances on all fronts. However, based on many parallel experiences at UW-Madison (the accompanying technology solution papers provide examples), we believe that the Synthesizers are the most effective community on which to focus our initial effort. This community is already pre-adapted and motivated to learn good practice, and the wide heterogeneity of its members’ skills, tools, and backgrounds provides a rich opportunity to
improve the efficiencies of knowledge production and transmission. Additionally, the
Synthesizers are best positioned to act as intermediaries between the Generators and the
Informaticists, transmitting science-driven informatics needs in one direction and best
informatics practices and methods in the other. This group also contains a large proportion of
early and mid-career scientists, who are well-versed in social networking tools, open-source
software, and other key elements of modern technology and data analysis tools.

How to best engage and strengthen the community of practice represented by the Synthesizers,
who are numerous and scattered across many disciplines and knowledge domains? We
believe that multiple approaches are needed, and that we must learn and adapt as we go. Core
principles that initially will govern our approach include:

1) **The central goal is to strengthen existing communities of practice.** This means
improving the connectivity among and within the many communities of practice. This can
be done through technical means (e.g., by helping disseminate new advances in web
services and approaches, such as RESTful services), and equally or more critically by
improving the ties within and among key communities of practice.

2) **All knowledge is local.** Scientific experts have the best available knowledge about any
given data type, and whenever possible, decisions about data-quality standards should
rest with them. It is to be expected that these standards and practices will vary among
communities. Hence our goal is not to prescribe standards but rather raise awareness of
good practices, and to offer examples of “success stories” – solutions that have been
developed well in one arena, and might be successfully adapted elsewhere.

3) **Scaling up requires empowering users.** The diversity of the geoscience community
precludes a strong centralized model, yet the size of the geoscience community requires
scalable solutions. We propose a networked approach, in which we strengthen the social
connections among individual scientists and scientific consortia.

4) **Success depends on continuous assessment and adaptation.** We do not know the
best approach a priori, and the technical solutions will continue to evolve as our hardware
and software capabilities continue to improve.

**Operations and Sustainability Model:**
We suggest a constellation of operation strategies, loosely organized by a “Center for
EarthCube Science Support.”

**Short courses and workshops.** These would be 1-week courses focused around case studies –
successful models of data sharing and scientific applications. These would have dual purposes
of providing participants with informatics training and building a strong community of practice.
Community-building exercises would take equal weight with training exercises. Because we are
all on the same technical learning curve together, these workshops would be open to graduate
students, postdocs, and faculty, rather than being limited to particular career levels. Selection for
participation in these workshops would be based primarily on a) an assessment of a candidate’s
willingness and ability to begin putting the lessons learned into immediate practice, and b) the
ability of a candidate to connect outwards to the broader community. All workshops will include
self-reflection by the participants and organizers and observation and assessment by formative evaluators.

**Social networking sites.** The distributed, data-driven platform we aim to develop is well adapted to creating dynamic new social networking sites that combine geoscience data and the people who use and interact with those data. We believe there are considerable opportunities to create dynamic social networking sites that reveal and facilitate science as a process by build upon many of the architectural elements outlined here, including data publication and licensing approaches. In essence, we believe that a geoscience person- and data-driven “Facebook-type” model for interacting socially with EarthCube is a real and exciting possibility that will both serve to build a sustainable community and promote open and participatory science.

**Informatics help desk.** A staff of dedicated technical experts, peopled by early-career scientists (M.Sc. and Ph.D.-level) with dual expertise in geoscience and informatics, should be established. This group would be actively engaged with the geoscience community, offering resources and suggesting solutions to common technical challenges. This group would pursue both passive interactions (e.g., monitoring and responding to queries posted in on-line chat rooms) and active interactions (e.g., attending scientific meetings and reaching out to individual scientists, organizing and hosting web forums) with the broader geoscience community.

**Pilot steering committee.** In all cases, our intent is to engage with the broad US and international geoscience community. However, as we develop our approaches, we propose to first develop and test them on the UW community of geoscientists. From the outside, UW may seem like a single monolithic institution (as perhaps the geosciences do to non-geoscientists). However, with over 2000 faculty spread out across three miles of campus, from every branch of the geoscience and related disciplines, and organized into a plethora of individual scientists, and research centers ranging from strongly centralized to decentralized models, in many ways, **UW is a microcosm of the larger US scientific community.** Hence, it is an ideal laboratory for testing and developing informatics solutions that are not tied to any single sub-discipline of geoscience. Our initial pilot steering committee will be members of our own campus who span much of the geosciences and who have national connections. We also anticipate partnering with IRIS, which has arrived at a web service-based solution to making seismic data accessible.

**Our Motivation:**
If, at the end of efforts, one of our undergraduate students can visit a general user interface that is built upon UW’s instantiation of EarthCube and then easily and intuitively explore the depths of data and knowledge that come from asking one simple question: “What do we know about Earth, right here,” wherever “here” may happen to be, then we will have succeeded in bringing to our campus an important new tool for teaching and research. We will have also overcome the fact that connections among members of sub-disciplinary communities who are geographically dispersed are often stronger than the connections among people from different disciplines who are located here on our own campus. EarthCube, therefore, provides us with a great opportunity to strengthen the "weak ties" that bind the internationally connected UW-Madison geoscience community. Thus, from the technical to the sociological, we believe that our UW-focused experiment will provide valuable guidance during the EarthCube development process.