

CyberGeo

Thoughts on cyberinfrastructure for the geosciences, with particular emphasis on aspects complementary to high-performance computing and networking.

Motivation

As the National Science Foundation (NSF) converts cyberinfrastructure from concept to action, understanding is needed about how this general notion [ref Blue Ribbon report] should be made specific to advance research and education in the geosciences.

Though the needed specificity is taking shape in relation to high-performance computing and networking [ref Vision report?], equivalent progress has yet to be achieved for other aspects of the cyberinfrastructure vision. This paper is intended to help fill that gap.

Definitions

Cyberinfrastructure (sometimes abbreviated CI) is a recently coined term¹, and there is no widely agreed, precise definition. It may be said to encompass computer-related information technologies that serve as infrastructure for some domain of activity. As used by NSF, this domain is the entirety of science and engineering, but in the context of this document, the notion of cyberinfrastructure is limited to capabilities that support research and education within or proximal to the geosciences.

<p>Infrastructure</p> <p>1: the underlying foundation or basic framework (as of a system or organization)</p> <p>3: the system of public works of a country, state, or region; also: the resources (as personnel, buildings, or equipment) required for an activity</p>	
<p>Fig. 1: Definitions selected from the <i>Merriam-Webster Online Dictionary</i>.</p>	<p>Fig. 2: <i>Thinkmap Visual Thesaurus</i> depiction of words related to a concept dubbed “the manner of construction of something and the arrangement of its parts.”</p>

Two (of three) dictionary definitions for “infrastructure,” and a related thesaurus entry (figures 1 and 2), are germane to defining cyberinfrastructure. By augmenting the two definitions—infrastructure as foundation and as public works—with a third concept about a participatory design environment, the following three subsections form a tripod on which to build a strategy for effective cyberinfrastructure in the geosciences.

¹ The prefix “cyber” stems from “cybernetics,” but common usage no longer refers only to that field, instead spanning nearly everything related to computers, networks or robots.

Cyberinfrastructure as Underlying Foundation

This definition emphasizes functionality. Cyberinfrastructure (for the geosciences) must comprise systems, services, frameworks and capital equipment that enable and support (i.e., that leverage) other, more specialized systems, services and activities. To be effective in the sense of a foundation, such cyberinfrastructure must be well tested, dependable, and adaptable, i.e., general-purpose or based on useful abstractions. Experimental cyberinfrastructure must not be ruled out, but investigations should include assessments of (potential) reliability and adaptability.

Cyberinfrastructure as a System of Public Works

This definition emphasizes the *common good*, which of course is not independent of functionality. However, the common good also depends upon policies and decision-making processes (analogous, e.g., to those surrounding public services and regulated utilities) that reflect widely held values such as transparency, interface standardization and universal, low-cost or free access to tools and services.

Cyberinfrastructure as a Socio-Technical Environment

This notion emphasizes *ongoing, participatory design* as an essential characteristic, one that distinguishes cyberinfrastructure from other types of infrastructure. This reflects a pair of closely related observations:

- “The deep and enduring changes of our ages are *not technological but are social and cultural, in their core substance*” [Fischer, 2006].
- Computers are “apart from other communication and information technologies (e.g., television) that are passive and cannot conform to the users' own tastes and tasks. Passive technologies offer some selective power, but they cannot be extended in ways that the designer of those systems did not directly foresee” [Fischer, 2002].

Taken together, the above concepts of cyberinfrastructure demand a combination of technical robustness, sound policies, and frameworks for evolution, in which users may creatively extend the artifacts they are given, overcoming barriers that often exist between consumers and designers [Brown & Duguid, 2000].

Direction and Ideals

We assume NSF/GEO intends to foster and support cyberinfrastructure projects and programs—collectively dubbed *CyberGeo* herein—that benefit the geosciences. From the preceding section, one may conclude that the boundaries of CyberGeo cannot be discerned from prior definitions alone, so the subsections below propose an overarching purpose and a set of guiding principles to help set directions and inform decisions about what is or is not central to CyberGeo.

Purpose of CyberGeo

The overarching CyberGeo purpose is:

To establish a reliable, socio-technical environment that leverages creativity and learning in the geosciences.

[*Alternate 1*: establish reliable and adaptable information services & technologies that facilitate the exploration & communication of concepts across the geosciences community.]

[*Alternate 2*: establish reliable and adaptable information services & technologies that support or accelerate a rich set of research & education advances in the geosciences.]

Principles of CyberGeo

The principles listed below are intended to help set priorities and shape projects in ways that maximize the leverage gained via cyberinfrastructure for the geosciences.

- Leveraged Activities: There are no inherent limits on the types of activities that benefit from CyberGeo; some may be highly specialized and lie on the leading edge, and others may be multi-disciplinary and nearly universal.
- Dual Priorities: Given limited resources, CyberGeo priorities emphasize two classes of technology, those serving large segments of the GEO community and those enabling advances that otherwise would be unachievable. CyberGeo will not duplicate or replace general-purpose commodities that are widely and economically available to the public (on the Internet, e.g.).
- Evolving Boundaries: As the geoscience community gains experience in using cyberinfrastructure, there will be natural extensions of the CyberGeo boundaries to encompass additional technology. Concepts that are experimental or discipline-specific may—through abstraction, engineering and testing—gain the attributes of true infrastructure.
- Central versus Distributed Activity: CyberGeo must embody a continuously evolving balance of centralized and decentralized capabilities. This reflects the reality that most functions—even those which, at some stage, may be performed best in a large, central facility—eventually become better matched to the systems owned and operated by individual community members and institutions.
- Elevating Semantics: The general trend of cyberinfrastructure is toward ever-higher levels of semantics—i.e., meaning—embedded in the tools and the data flows of the community.
- Transcending the Disciplines: **[To be written**, with emphasis on how common abstractions (IDV/LDM/NetCDF/CDM/GALEON, e.g.) enable and support interdisciplinary advances.]
- Standards and Transparency or Openness: **[To be written**, perhaps with material from NSF's CI Vision doc on "Resource Collections" and "Reference Collections," on international standards/certifications, and on policies for accessibility and use.]
- [other principles?]:

Five Primary Classes of CI-Enhanced Activity

[This entire section is largely unwritten, though a few ideas/sources are noted.]

Earth-Systems Observation

[To be written, perhaps with material NSF uses to flesh out the "Strategic Plan for Collaboratories, Observatories and Virtual Organizations" section in its CI Vision. This section should stress the type of mutually beneficial relationship that has developed between Unidata & NOAA, as well as the model of participatory observation manifest in

SuomiNet, all made possible by LDM and IDD. To complete this well, I must learn more about observational leveraging in CUAHSI, GEON, IRIS, etc.]

Earth-Systems Simulation

[To be written, with much referencing of NSF material on high-performance computing, as well as GEO-specific literature such as *Cyberinfrastructure for Environmental Research and Education* (2002) and *Cyberinfrastructure for the Atmospheric Sciences in the 21st Century* (2004). This section should highlight the community-model concept, noting the importance of NetCDF as underpinning for standardized data representation and interfaces (among models, visualization tools & GIS, e.g.).

We might reference the recent NY Times article by Markoff, titled Software Out There: “The Internet is entering its Lego Era.”

To complete this well, I should learn more about simulation in OCE & EAR, but the section should be fairly brief because the focus of the paper is not HPC.]

Data Analysis and Synthesis

[To be written, with reference to other GEO material as above, plus progress made in IPCC assessment. This section should highlight the value of abstract data typing (i.e., data models), encapsulation (behind APIs and Web Services), third-party metadata, and polymorphism, as manifest in IDV, NetCDF, OpenDAP, THREDDS, CDM, GALEON, GIS and so forth.

To complete this, I must learn more about progress in LEAD and beyond Unidata.]

Scholarly Communication

[To be written, with reference to work by Lagoze and van de Sompel on the relationship between digital-library advances (e.g., institutional repositories, arXiv & Fedora) and scholarly communication. Such communication increasingly entails (or should entail) transport or referencing of data sets, computer programs, maps, images, and so forth, sometimes with embedded interactivity, as manifest in IDV bundles and (to a lesser degree) in THREDDS catalogs.

This section should utilize Fischer’s paper (titled Beyond “Couch Potatoes”: from Consumers to Designers and Active Contributors), especially these points:

- “Cultures are substantially defined by their media and their tools for thinking, working, learning, and collaborating.”
- “Skilled domain workers are not novices or naïve users, particularly with respect to domain concepts. They are people who have computational needs, who want and need to be designers in personally meaningful tasks. They make serious use of computers in their work, but they are not interested in becoming professional computer scientists.”
- “Extensible systems, together with power users who can perform modifications, enable a process of *co-adaptivity* between users and system (Mackay, 1990). Users learn to operate a system and adapt to its functionality, and systems are modified to adapt to the practices of its users.”

As I see it, the implications of Fischer's points are that sharing the artifacts of computer use and of design (i.e., data sets, metadata, programs, etc.) is increasingly critical to the flourishing of geoscience communities of practice and communities of interest.]

Learning and Decision-Making (with Scaffolding)

[To be written, with emphasis on the positive educational impacts of Unidata, DLESE and related projects. So far the new information-technology landscape has influenced the Nation's education system only modestly ... but I might put something in here about the Gates Foundation's support for high-tech alternative schools...

Also, emphasize workforce development, per the human-resources emphasis in the "Cyberinfrastructure for Atmospheric Science" paper and per Fischer's writing about "meta-designers" and about communities of practice/interest.]

Use Cases

The value of cyberinfrastructure and—more germane here—its potential utility in the geosciences, may be illuminated by examples. Each "use case" below sketches a scenario where appropriate cyberinfrastructure will enable investigation and learning that would be impeded absent such capabilities. Due to the specific emphasis of this paper, the selected examples do not focus on the value of high-performance computing and networking, though such examples are abundant. [Note: I've given little thought to whether or not these use cases are sufficient or optimal.]

Extreme Events

Extreme Earth-system events are of increasing concern to Americans, as reflected in the media and in heightened priorities at local, state and federal agencies, such as the NSF Geosciences Directorate (NSF/GEO)². Notable—re cyberinfrastructure—are possibilities for "just-in-time" investigations of extreme and *unexpected* events, regardless of where they occur or which Earth-system components pertain (e.g., the atmosphere, hydrosphere or lithosphere).

CyberGEO can support unprecedented surveillance of planet Earth by:

- Supporting synthesis of multiple and novel types of information, such as might be required to address plume dispersion events, say of pollen, volcano ash, smoke from forest fires, human-produced toxins and the like.
- Enhancing interoperability and generally enabling integration of heterogeneous observing systems, including sensor networks just now being envisaged;

² From the summary of *NSF GEOSCIENCES BEYOND 2000 – Understanding and Predicting Earth's Environment and Habitability* at <http://www.nsf.gov/pubs/2000/nsf0028/nsf0028.htm>:

"The provision of reliable information on geophysical phenomena, both natural and human-influenced, that is well-targeted to meet societal needs is a significant product resulting from geoscience research. Losses in the United States from geophysical disasters have risen rapidly. Single extreme events, such as hurricanes, tornadoes, earthquakes, volcanic eruptions, solar storms, and floods can cause losses of several billion dollars and severely disrupt commerce and daily human activity. ... Earthquakes, severe storms, solar storms, and biological invasions represent threats, but we have the opportunity to mitigate these threats for society. Predictions of extreme planetary events can help save lives and/or lessen property damage."

- Offering universal, low-cost access to near-real-time flows of data in all spheres of interest to GEO, including means to monitor extreme, high-impact events;
- Enabling wider use application of dynamically adaptive or steered observing systems, such as are employed in LEAD and at JPL: “NASA’s Jet Propulsion Laboratory (JPL) has implemented a Volcano Sensor Web (VSW) in which data from ground-based and space-based sensors that detect current volcanic activity are used to automatically trigger the NASA Earth Observing 1 (EO-1) spacecraft to make high-spatial-resolution observations of these volcanoes.” [Davies, 2006]
- Linking Earth-system observation to science education at many levels, capitalizing on students’ natural curiosity and promoting scientific literacy.

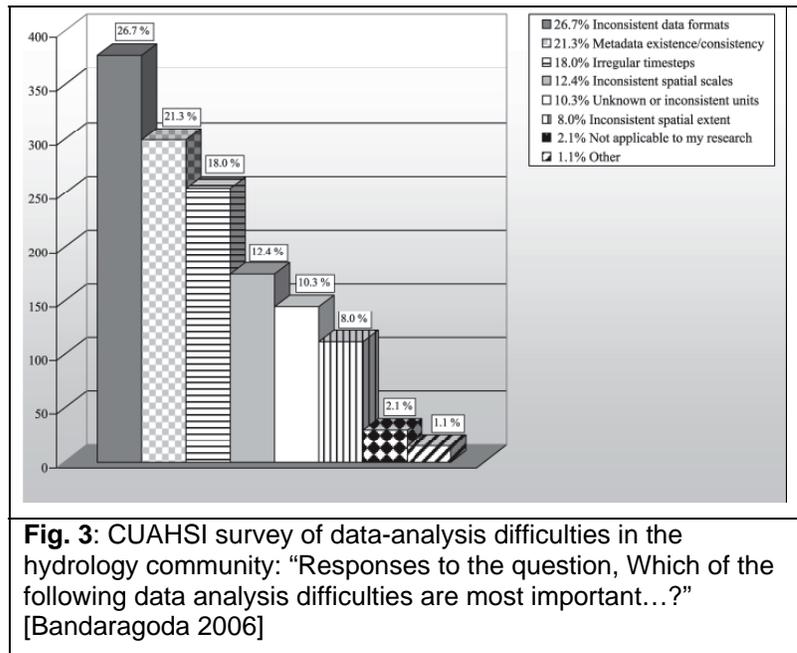
Data Repurposing

A major cost of science is the human effort required to utilize unfamiliar data, a cost that explodes in cross-disciplinary studies. An indication of this impedance may be found in a recent survey concerning the “Cyberfrontier” in hydrology (see Figure 3):

When asked how much of their ... time is spent on data preparation, more than 80% said they spend at least one-tenth of their research time on data preparation, and 12% ... said they spend more than half their research time on this task. [Bandaragoda 2006]

CyberGEO can foster unprecedented ease of data use, enabling uses that go beyond the intentions of the data-set authors, by:

- Developing and supporting wider use of advanced data-access systems (e.g., NetCDF, OpenDAP and GIS/WCS) that encapsulate abstract data models within Web services and other programmable interfaces.
- Advancing and supporting wider use of metadata with high-level semantics and post-facto (i.e., third-party) creation, thereby enabling forms of polymorphism among data sets and data streams. Relevant progress is manifest in THREDDS, CDM, GALEON and IDV.



Multidisciplinary Studies of Water

[To be written, Draw on Ben’s ideas...]

Multidisciplinary Studies of Volcanoes

[To be written, Draw on Ben’s ideas...]

Other Use Cases?

CyberGeo as a Set of Services and Supported Technologies

The foregoing offers a high-level view of CyberGeo and—via use cases—examples of how appropriate cyberinfrastructure would facilitate exploration and communication of concepts across the geosciences community. The following subsections describe a set of proposed CyberGeo services and (supported) technologies. These were selected to cover the five primary classes of CI-enhanced activity (see Table 1) with practical units of work that might be performed with a combination of NSF/OCI and NSF/GEO funding.

These potential services are partitioned into two categories, to distinguish those that require significant experimentation before full implementation (with user support, etc.).

Proposed Services and Technologies	Classes of CI-Enhanced Activity				
	Earth-Systems Observation	Earth-Systems Simulation	Data Analysis & Synthesis	Scholarly Communication	Learning/Decisions w/ Scaffolding
Well-Proven					
1. Distributed (High-Performance) Computation					
2. Distributed (High-Volume) Data Curation					
3. Near-Real-Time Data Flows					
4. Geographic Information Systems (GIS)					
5. Generalized Data Analysis & Visualization					
6. Data Encapsulation (Web Services & APIs)					
7. Multi-Media & Interactive Documents					
8. Digital Libraries & Institutional Repositories					
9. Settings for User Engagement & Reseeding					
Others?					
Experimental					
10. Data Mining					
11. Data Assimilation/Fusion					
12. Workflow Choreography					
13. Ontology-Based Reasoning					
14. Meaningful Artifacts of iScience					
Others?					

Table 1: Coverage of cyberinfrastructure-enhanced activities by proposed CyberGeo service offerings.

Well-Proven Services

The capabilities and technologies sketched in this section have exhibited significant practicality and utility in one or more of the geoscience disciplines, so their provision as part of CyberGeo would entail more emphasis on user-support, reliability and socio-technical aspects than on experimentation or proof of concept.

1. Distributed (High-Performance) Computation

[To be written, ...]

2. Distributed (High-Volume) Data Curation

[To be written, ...]

3. Near-Real-Time Data Flows

[To be written, ...]

4. Geographic Information Systems (GIS)

[To be written, perhaps within rather than separate from the following item...]

5. Generalized Data Analysis & Visualization

[To be written, ...]

6. Data Encapsulation (Web Services & APIs)

[To be written, ...]

7. Multi-Media & Interactive Documents

[To be written, ...]

8. Digital Libraries & Institutional Repositories

[To be written, ...]

9. Settings for User Engagement & Reseeding

[To be written, highlighting some “reseeding” examples, per Fischer’s model of phases in evolving complex environments (e.g., Harry Edmon’s impact on NEXRAD use in Unidata & NOAA; Charlie Zender’s impact on NetCDF use/development).]

Others?

Experimental Services

The capabilities and technologies sketched in this section show promise of utility in the geoscience, but their provision as part of CyberGeo would entail significant emphasis on experimentation and proof of concept.

10. Data Mining

[To be written, ...]

11. Data Assimilation/Fusion

[To be written, ...]

12. Workflow Choreography

[To be written, ...]

13. Ontology-Based Reasoning

[To be written, ...]

14. Meaningful Artifacts of iScience

[To be written, ...]

Others?

Strategic Priorities

[To be written: Define a clear path from experimental cyberinfrastructure to well-established & supported cyberinfrastructure (that is not heavily dependent upon commercialization).]

Accountability and End-User Engagement

[To be written: Describe user-centered means for gaining advice & setting policy.]