Flexibility in an EarthCube Design Approach: Ideas from IRIS and its Partners

Tim Ahern (IRIS, Director of Data Services),
Keith Koper (University of Utah, Chair of DMS Standing Committee),

Vision for Earthcube

EarthCube's goal of easing access to geoscience data across the fields supported by the NSF GEO Directorate is simply stated but will be challenging to realize. EarthCube aspires to define and meet a range of fundamental and complex cyberinfrastructure needs of the Geoscience community for the decades ahead. To ensure that the geoscience research community is fully engaged in the definition, design and implementation of EarthCube, an initial strategy should be to build on existing strengths within the Geosciences and deliver early prototypes that deliver on the often-stated goal of a system that supports integration and delivery of diverse types of geoscience and other needed data (e.g. social sciences data).

The Geosciences are inherently multi-disciplinary. Embedded in the EarthCube concept is commitment that the geoscience community is ready and well-positioned, perhaps more than other scientific disciplines, to engage in building a system that will meet the ambitious goal of cross-disciplinary data integration and knowledge management. Research problems in the earth sciences are complex, requiring massive amounts of diverse observational data, the development of sophisticated models, leveraging of significant computational resources, and the best scientific minds to address the challenges. Earth science problems also have a direct and important societal impact.

The system that must be developed will be challenging both technologically as well as sociologically. Understanding the needs of the geoscience community will require close and effective connections between the geoscientists that will define the system and the cyber technologists that will build the system. It is important that domain specialists from the Geosciences take the lead in setting the goals for EarthCube and work closely with cyber specialists to begin to design and implement a new and exciting framework that places emphasis on solving practical, but fundamental, Earth science problems, rather than exploring new cyber technologies that may be remote from current geoscience needs.

Our vision for a pathway for EarthScope is captured in the following concepts:

- The goals for EarthCube should be identified by Geo-domain scientists.
- The design and implementation should begin in close collaboration with cyber–specialists in a way that brings current cyber-technologies to bear on geoscience applications.
- Coordination **within** a geoscience domain should take place first as vertical integration within existing domains and include integration between national and international centers in the same domain (e.g. seismology)

- Horizontal integration across domains should build upon coordination within the domains
- Support for standardized formats will be a requirement for horizontal integration. Additionally ready-to-use data, corrected as required, should be available in simple, easy to use formats
- The system will evolve with time and therefore must be a loosely coupled and flexible system throughout its lifetime
- Technology will change and the EarthCube solution will also need to evolve
- There are existing strengths throughout the geo-data community (low hanging fruit) that can take the lead in developing short-term successes and enable community buy-in

Leveraging geospatial and temporal characteristics. The Earth science community is diverse in the types of data needed for its research. However, an attribute most geoscience (and especially geophysical) data have in common is they can be geospatially referenced, providing unique opportunities for data integration and discovery. In addition to the geospatial aspect, another unifying characteristic is often that of time. These factors provide immediate opportunities for data interactions in that the discovery and integration of many types of geoscience data can be made using space-time queries. To exploit this characteristic of much of the data, EarthCube could begin by developing standard naming conventions, consistent with international standards where possible, such as those established by the Open Geospatial Consortium (OGC).

A system to support geoscience research for the next decade must be built in a manner that allows the system to evolve to meet newly identified needs as well as to take advantage of new technologies as they emerge. Decentralized, flexible, loosely-coupled systems are the key to the long term success of EarthCube.

Community Based Governance Model

An important factor in establishing EarthCube and ensuring its long-term success will be the governance and management structures that are adopted to guide its creation and evolution. The governance structure should emphasize vision, breadth, balance, and commitment. The focus of the management structure should be on stability, efficiency, quality, competence, and usability.

It will be critical in the early stages of development to provide clarity on the purpose and long-term goals of EarthCube. If the primary purpose of EarthCube is to serve the research needs of the academic community, the governance and management structures should be firmly established with this purpose in mind. Structures should be embedded for appropriate linkages with collaborating entities, such as other data and IT organizations, but the governance structure should be clearly charged with advancing the interests of the intended users.

Leadership at the governance level by the academic community will be essential to ensure that the goals of EarthCube are focused on advancing fundamental geoscience research and that the structures and implementation evolve to meet changing needs and new opportunities. Sustaining committed and enlightened leadership will require that the governance structure have the authority to lead and maintain stable implementation, coupled with the flexibility and vision to innovate and evolve. The leadership should be composed of individuals who are committed to the long-term stability and multi-disciplinary foundation of EarthCube.

Various models exist for how EarthCube programs and services could be implemented. Within NSF, it could be established in an existing or new Directorate level program, with guidance under the proscribed constraints of a FACA Advisory Committee. It could be a community-organized, non-corporate amalgamation of separately funded projects like GeoPrisms, or a facility program operated under a consortium-governed, non-profit corporation like IRIS, UNAVCO or UCAR. A formally organized federation could draw on the capabilities of existing or new domain-specific facility programs. Other models exist, but “form should follow function” - the forms of governance and management that are developed should be tightly linked to the functions that are to be undertaken to implement the goals of EarthCube.

IRIS has had 25 years of experience in providing data collection and distribution services to seismology. How IRIS works for and with its community could bring an important perspective to the definition of the EarthCube governance system. One of the reasons for IRIS’ success is that through the consortium structure, the seismological community governs and guides what IRIS does and works closely with the facility to define the cyberinfrastructure technologies and resources that are needed for the domain scientists to do their work. The challenge for EarthCube is to build a system, across the geoscience domains, to lead the EarthCube developments in much the same way that IRIS has interacted with the seismological community.

Some elements of the IRIS model that might be considered as EarthCube is created include:

- Membership in EarthCube should be formalized and include members from the geoscience community in universities, not-for-profit consortia, and related research organizations.
- Leadership of the governance for EarthCube should rotate on a regular basis and come from within the member organizations .
- Working groups should be formed from the EarthCube organizations and associated organizations and include both domain and cyberinfrastructure representatives.
- Associated membership should include representatives from for-profit corporations, other government agencies, and similar organizations.

- EarthCube should be staffed at a sufficient and sustained level to interact with the governance committees and working groups to implement the components of EarthCube as defined by the governance committees.

Conceptual Cyberinfrastructure Architecture

We believe it is essential for the system to be service oriented and flexible. Rather than a top-down approach, it should be a system of flexible, loosely-coupled components that can be accessed as needed for a given scientific problem. The term mashup is now commonly used to refer to a web application that combines data from two or more sources to create a new service. Traction can be gained by implementing a data mashup (dashup?) that brings together relevant data and products across domains as needed for a scientific problem. By making data and products available through web accessible techniques (e.g. URLs, URIs) data integration can take place leveraging a system of loosely-coupled and flexible web based services.

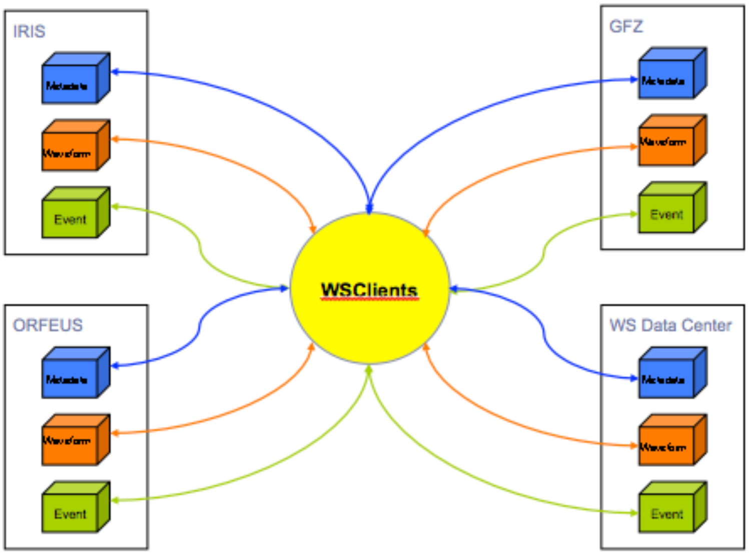
Flexibility and Coordination within a Specific Geoscience Domains. IRIS has more than two decades of experience in the coordination of data exchange within the seismological domain. IRIS has worked successfully with both domestic partners (e.g. USGS, Regional Seismic Networks) and international partners in the International Federation of Digital Seismographic Networks (FDSN) to define the data exchange formats, exchange protocols, and, more recently, standardized services that have resulted in a successful international system that also works effectively domestically.

More recently IRIS has pursued the development of web services. We have found that this model of development has resulted in a rapid adoption by the community and has also received the support of the FDSN in terms of a flexible architecture that will meet the needs of the international and domestic seismological communities.

Web Services provide a flexible loosely-coupled method of providing access to data from distributed centers and make it straightforward to develop a federated system of data centers within the seismological domain. There are two fundamental types of web services, those based upon the standards based, somewhat heavy-weight, SOAP model of web services and those that are built upon REST style web services that are the conceptually simpler method and focus on point-to-point communication using http. IRIS' experience (shared by others in the geosciences) has shown that while SOAP services are potentially more powerful, the required infrastructure that must be in place for both the service provider and the service consumer is much greater and results in limited adoption. REST web services are provided and consumed by simple URLs that specify the service and a variety of parameters that control the query or invoke processing options. REST services are recommended as a way of encouraging early development of data integration within and between EarthCube domains, IRIS and its partners in the FDSN (that includes both international and domestic partners) have identified five specific capabilities that are essential components for the first federated system of seismological data centers. We believe that these can also represent the key services for many other domains. While the specifics of the systems will differ, the basic services will be the same. These five primary services are:

1. **Data Availability.** A service that supports geospatial, temporal, and type-of-instrument queries. This service returns information that identifies what data with the requested attributes exist at specific data centers.
2. **Metadata.** A service that returns the relevant metadata describing the observational data in a manner in which the observational data can be understood.
3. **Data Requests.** This service will result in the return of the actual observational data. In IRIS's case this is the return of the sensor data as a time series. For other domains the specific type of data may differ. For some fields the volume of data returned will be relatively modest, for others the volume may be huge and this may affect the way in which data are returned. For instance IRIS's services that return time series data have been shown to return more than a terabyte per day for an individual user and could represent an end member in terms of data volumes returned.
4. **Other Related Domain Data.** Depending on the type of data there may be other products of direct interest. For example, for seismologists, and other geoscientists, catalogs of earthquake parameters are of high interest.
5. **Higher Level Products.** Many geoscience fields produce derived products that are of more interest than the raw observations. Services that can identify and return products that meet geospatial and time constraints are of general interest.

Each geoscience domain should coordinate the development of unified web services that support the above five fundamental types of services. Geoscientists within the domain are best suited to define the data types and how the implementation of services should be done within a domain. EarthCube might consider the formation of a SoftWare Assist Team (SWAT) that can provide support for the definition and perhaps implementation of web services within a domain. The SWAT team will minimize the level of effort required within a domain as well as facilitate eventual cross-domain development efforts.



Federated Web Services. By federating web services across geographically separated data centers, client applications can access data and products using identical web service based techniques from multiple centers.

The implementation of uniform web services at geographically distributed domain data centers will allow the abstraction of database and data repositories in a manner where the specific details of local data management do not need to be exposed to the outside user. By further coordination of the query parameters (e.g. parameter naming, query response, etc) this allows the inputs and the outputs to and from any data center to have the same expected behavior. This allows the centers to participate in a federated system of centers. Within the seismological domain, this coordination is beginning to take place through the FDSN. We believe this is an effective model that should be considered by other geoscience domains as well. IRIS has some initial documentation of the strawman FDSN web services and these can be found at http://www.iris.edu/ws.

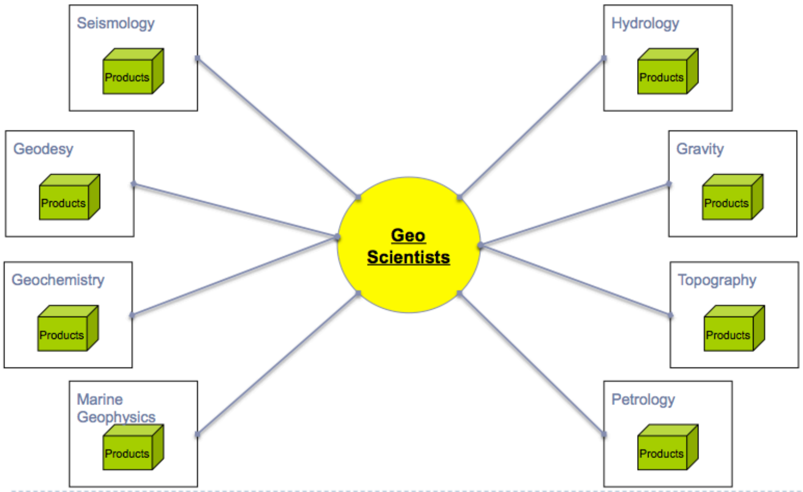
Flexibility and Coordination across Geoscience Domains

The development of ontologies, concept spaces, and other semantic web approaches is commonly used in many cyber approaches to enable cross-domain integration. Such concepts and constructs are likely to find important applications in parts of the geosciences, especially in ecology and bio-geology, and may eventually find broader application in integrating diverse geo-datasets. In the initial stages of EarthCube, however, more direct approaches, such as "linked-data" may be more promising for exploring geospatial and time series data and yield significant progress that will lead to more widespread adoption of EarthCube approaches as domain scientists will see results early in the process.

Key concepts that EarthCube should consider adopting include:

- Domain data should
    - be available in common formats used within that domain

- be available in widely used formats (e.g. NetCDF, HDF5, etc) when possible
- be easily converted to simple alternate formats (e.g. ASCII) that are readily understood.
- be available after preprocessing that makes corrections to the data to ease use by a non-domain specialist with unambiguous corrections already applied (e.g. instrument calibration, inter-laboratory corrections, etc).



Federated Cross Domain Services. By coordinating services within a domain and with fundamental space-time queries implemented in a similar manner within domains, the integration of data between geoscience domains becomes more tractable. Rather than integrating multi-domain data within one center or system, geoscientists can access information, as needed, through a system of distributed and flexible web services.

Once each domain has self-organized and made their data available via web services, the more difficult job of integrating geoscience data across domains becomes possible. A recent survey of the geoscience community initiated by IRIS showed that when scientists wanted to access data from other domains they preferred to receive the data in an easily understood ASCII representation (51.4%), the native format of that domain (47.2%) and much lower for CSV (11.1%), or XML/JSON representations (16.6%). The message here is that, at least initially, data across domains should be available in multiple, simple, preprocessed, easy-to-understand representations. As much domain specific expertise should be represented in the corrected data as possible.

Scientific Workflows

The development of REST style web services, especially when coupled with workflow engines, would be an important initial step in allowing integration of data across the geosciences. At IRIS we have a suite of approximately twelve operational web services and are aware that UNAVCO, IEDA, CUAHSI and EarthScope are developing similar tools. These services can be accessed directly by java, wget, scripts, browsers, etc. More importantly, we have demonstrated that workflow engines can also access these services remotely. This is an extremely powerful capability that could be leveraged as the heart of the flexible, decentralized, loosely coupled system for EarthCube.

The workflow system being investigated at IRIS is Trident, developed and made available by Microsoft Research, but these capabilities now exist within many other workflow engines as well (e.g. Kepler, Taverna, Wings). IRIS web services are capable of having the output of a service directed to a local cache that can be used as the input to the next step of a workflow. The IRIS Caching Artifact Buffer (ICAB) opens up the possibility of powerful workflow sequences that minimize data transfer over the Internet. A distinct advantage of the Trident system is that it captures provenance of the workflow consistent with the Open Provenance Model (http://openprovenance.org) that would allow potential reprocessing when key information (metadata) changes with time.

An implementation of the workflow model for EarthCube would be one where data from one EarthCube node are accessed, processed, and transformed (to a simple format or one of the EarthCube standard formats) by any or all services available at that EarthCube node. The resulting artifact could then be transferred to another EarthCube node or back to the client. A similar process could then be directed to take place at another EarthCube node, and so on. CUAHSI is promoting a concept very similar to this in their Paris Metro analogy.

Identification of a few specific data exchange formats.

There are several formats that serve large communities and have been shown to be capable containers for a wide variety of data sets. The EarthCube system should identify such formats and, in general, each domain should provide support for converting data into these formats. To make the data more easily understood across domains, the capability should exist to provide results in simple ASCII formats as well.

Higher Level Data Products

The CODMAC committee of the National Research Council identifies levels of data that most likely pertain to all domain data that are within the purview of EarthCube. While the specific definitions slightly vary across domains, one structure is that adopted by the EarthScope MREFC project supported by the Geoscience Directorate:

- Level 0 – Raw uncorrected observations
- Level 1 – Quality Assured observations
- Level 2 – Derived products using unambiguous techniques
- Level 3 – Derived products using higher level domain expertise
- Level 4 – Integrated Products using data from multiple domains

Much of the integration needed across domains will come through products generated within a domain that allow geoscientists from other domains to quickly get to derived information that they need for an integrated study.

Design Process

In the initial stages we believe the specific geoscience domains should work within a domain to identify the necessary services needing development. A requirements-driven design can best be done by the identification of actual workflows needed by the geoscience community to solve real problems. The EarthScope technical white paper prepared for EarthCube by Gurnis, et al. identifies this approach as one that captures characteristics of what a geoscientist needs from the EarthCube cyberinfrastructure. Domains should self organize but be cognizant of standardization that is important across EarthCube (e.g. space-time aspects, standards, etc.).

After specific domains have organized and have in place federated web-services within their domain, domain scientists from multiple domains should be brought together in a series of workshops that focus on a theme requiring data and information from several domains. The output of these workshops should be specific workflows showing the data required, the sequence of processing steps required, the toolset that is needed, and what output is developed.

It is likely that a series of working groups will be organized to address cross-domain integration between the geosciences activities, likely coordinated by the EarthCube governance organization.

Operations and Sustainability Model

The question of sustainability is a complex one that involves not only financial support, but technical issues related to data preservation and, especially in the geoscience, commitment to the collection and preservation of extended time series and synoptic global data sets. For many disciplines in the geoscience there is a compelling need to collect and sustain time series that span decades to centuries and even longer. Maintaining access to these data in uniform formats is a major challenge that can involve careful attention to changing metadata and instrumentation characteristics as well as preservation and refreshing of physical media.

Sustainability should be closely linked to the intrinsic value of investments in data collection and preservation. While it is possible to explore various commodity-based mechanisms for public and private support of geoscience data, a key challenge for EarthCube should be to demonstrate how sustained investments in data resources contribute to advances in knowledge and the public good. Only if the “value” is perceived as high enough will sustained investments be forthcoming. A model that is too strictly commodity-based may make it difficult to argue for support for data collections that are deeply rooted in fundamental research applications. This is especially important for extended time series and synoptic global data where the inherent value of the data may extend well beyond the lifetime or geographic extent of individual research projects. One role for EarthCube may be to identify these key synoptic data sets (e.g., topography/bathymetry, meteorological time series, earthquake records and catalogs, geomagnetic records) and explore mechanisms for inter-agency coordination and support for uniform archiving, access and long-term collection and preservation.

Long-term support (sustainability) for many geoscience data sets can be greatly enhanced through the design and support of observational systems that are multi-use and multi-parameter. In those fields (common in the geosciencea) where there is a close link between basic research and mission applications (e.g., observations of weather, climate, earthquakes, volcanoes, rivers, tides, etc), carefully coordinated commitments to interagency and interdisciplinary support can greatly enhance the prospects for sustainability. Facilitating the educational use, at all levels, of key geoscience data can also significantly increase their “value”. While this may not lead directly to significant financial support, the demonstrated importance of these data in educational applications can aid in making the case for public support of sustained observations that benefit both the research and educational communities. The most significant factors driving the cost of long-term observational networks are often not the sensors themselves, but those related to logistic support (site infrastructure, maintenance and data communication). To the extent that multiple sensors can be established at common sites (without degradation of data quality and integrity), the long-term investments can be minimized.

For data collections that are primarily intended to support advanced research applications, it is essential that federal agencies, including NSF, be prepared to accept the primary responsibility for the resources necessary for long-term collection and preservation. EarthCube can help broader engagement by fostering public-private partnerships that include continuing resources from NSF, resources from other government agencies (e.g., USGS, NOAA, EPA, DoE), as well as financial or in-kind contributions from cyber related corporations (e.g. Microsoft Research, Amazon, Google, IBM, Apple).