An Architectural Approach for Sharing, Discovery and Knowledge Dissemination on EarthCube

Christopher Mader[1], Benjamin Kirtman[1,2], Felimon Gayanilo[1], Joel Zysman[1], Nicholas F. Tsinoremas[1]

1. University of Miami Center for Computational Science
2. Rosenstiel School of Marine and Atmospheric Science, University of Miami

EarthCube aims to essentially create the ultimate virtual environment for "Team Science". Inherent to this goal is building bridges between data, resources and disciplines. To do this requires an "out of the box" approach that incorporates many of the "inside the box" technologies and methodologies that have been established by the different science communities. Here we present a "blue sky" concept that could serve as starting point for a design approach for this system. To put the concept in context, we first provide a brief use case scenario followed by a description of some of the key system characteristics.

Over the next decades, marine ecosystems throughout the U. S. will be put at especially great risk from climate change in part due to their inherent (and historical vulnerabilities). For example, marine ecosystems will be impacted by at least some of the following: sea level rise, ocean acidification, alternately drought in some areas and increased rainfall-runoff and sediment delivery in others, increasing ultraviolet exposure, increasing temperatures, tropical storm intensification, shifts in ocean productivity, increasing pollution, loss of habitat, and loss of biodiversity. Risks associated with other anthropogenic impacts will compound climate change stresses and will vary markedly, locally and regionally. To assess and synthesize impacts and risks, and provide people, businesses, and governments the information and interpretation needed to adapt to and manage environmental change, there is a need for improved understanding and predictions of the earth system from days to decades.

To better define and refine the requirements of the EarthCube one must determine real life use case scenarios. One such scenario is described in the white paper: *EarthCube Science Requirements: Disseminating Multi-Model Seasonal Forecast Data*, Kirtman and Huang. In that use case the grand challenge is to provide actionable environmental information based on the best available science. Meeting this challenge will necessarily involve heterogeneous observational data collected from satellites, ships, planes, buoys, subsurface ocean platforms and land based stations to understand, "initialize" and validate models that predict a component of the earth system or that integrates multiple components of the earth system. Meeting this challenge will also necessarily involve complex heterogeneous computational models covering a variety of space and time scales, medium (e.g., ocean, atmosphere, land) and physical, chemical and biological processes, and may even explicitly include interactive human influences. Users of such forecast information (e.g., water resource managers, energy providers) at all levels of sophistication require predictions with minimal uncertainty accompanied by reliable estimates of that uncertainty.

In the NSF EarthCube vision, the user should be able to gain access to all of these data and computational resources transparently through the environment provided by the EarthCube knowledge management system. By simply opening a web browser (or other search client), the user should be able to leverage resources distributed across the US, and the world, by entering a set of search terms or keywords related to the datasets needed to initialize and validate models, for example. The results of this search should return references to the datasets in a contextually relevant way, so that it is immediately clear which of the datasets are appropriate for use in initializing the models, which are appropriate for validation, and which are suitable for other purposes. Likewise, the user should be able to easily locate all of computational model

components needed to create an ensemble model (if desired). Ideally, the user could then provision computational (e.g., through XSEDE) and data capacity (e.g., SDSC Cloud) and deploy all of the necessary model and data components to run the ensemble model and analyze and visualize the output.

Characteristics of EarthCube

To meet the vision of simple resource discovery and easy integration, the EarthCube infrastructure will need to provide a minimum set of key system characteristics. These requirements are distinct from specific needs for capacity, bandwidth, and other technological requirements, all of which will be continuously changing over the life of the system. The constant features for the life of the system will be those that facilitate discovery, communication, coordination and governance of system resources. At a minimum the system should probably provide the following key characteristics.

A. EarthCube should make it possible to easily leverage existing resources

The GEO community currently has many cyberinfrastructure resources and standards available to it, such as THREDDS, OPeNDAP, and SensorML. These resources are valuable assets that are widely used and have a broad base of associated human expertise. Any realization of EarthCube must be able to easily assimilate these resources, and make it simple for them to be incorporated as part of the EarthCube infrastructure.

Since these existing resources and standards represent the entire set of research areas covered by GEO, from the ocean to edge of the atmosphere and beyond, a major focus of the EarthCube infrastructure should be to establish technology that makes it possible to define and expose the relationships between these existing resources, in terms of their functional capabilities and compatibilities, as well as in terms of the meanings of these resources with respect to the earth system. For example, an investigator studying climate change should be able to find information about relevant sea surface conditions from any pertinent data services by using concepts and terms relevant to his area of research. This should be possible even if the original (or main) purpose for acquiring the data had nothing to do with climate change. However, in addition to directing the investigator to the relevant datasets, EarthCube should also be able to provide information about how these data can be used with other resources (e.g., computational resources) available through the EarthCube.

B. The deployment of new EarthCube resources should be simple

Just as it should possible to easily leverage existing resources, it should also be easy to add new resources to the EarthCube.

Smooth integration of new resources can be greatly facilitated by the existence of standards for the interoperability of resources and services, as well as standards for the description of data. Many of these kinds of standards already exist, but the organizational and governance structure of the EarthCube will need to have the capability to identify the need for new standards, and to create and implement these standards when necessary. In addition to various standards the EarthCube will also need to have a robust, well tested, set of infrastructure components to enable the deployment of new resources. The specification, implementation and maintenance of these standard components will also need to be managed by the EarthCube governing community.

C. EarthCube should support distributed administration and governance of resources

In order to create flexibility, nurture scientific creativity, and take advantage of local expertise, the EarthCube should implement a governance model that supports distributed administration and governance of the resources available on the infrastructure. In reality, the sheer number of

distributed, word-wide, resources envisioned for EarthCube would make central a administration model for the system completely unworkable.

While there should be some kind of central governance structure for the development of standards and the specification of the behavior of the key infrastructure components, the deployment, governance and maintenance of specific resources (e.g., computational capacity) should be done by individual autonomous organizations. Enablement of this kind of model is largely made possible by the establishment of standards as mentioned above, which ensure that even though each individual site is independently administered and developed, this is being done following the same set of basic rules.

D. EarthCube should support Open participation
Any individual or organization that wants to participate as part of the EarthCube infrastructure should able to do so. Participation in EarthCube would mean making data, computational, educational or other resources available through the infrastructure. The technological infrastructure and governance model should allow any size organization, from an individual up to a government agency (for example), to contribute resources to the system.

Key to meeting this requirement is the development of infrastructure components that are easy to deploy and use. And also possibly the establishment of support organizations, analogous to something like sourceforge for hosting open source projects or wordpress.com for hosting blogs, where entities that lack either the computational resources or expertise to directly participate as part of the EarthCube can publish and make their resources available.

E. Discovery of relevant resources is critical
If the EarthCube infrastructure is to be a world-wide distributed network of resources, then it is absolutely essential that investigators can easily discover and access those specific resources that are most relevant to their areas of interest.

Possibly the single most important feature of the EarthCube will be its ability to easily provide users with all of the relevant resources they need (for example) to answer a particular question, conduct specific research, or develop engaging educational curricula. The development of standards (as discussed above) is an essential aspect of enabling easy discovery. Equally important is the development of technologies that can exploit these standards in order to connect users with the most relevant EarthCube resources for their needs.

F. As is orchestration and coordination of resources
Much as the discovery of relevant resources is key to the infrastructure being useful, so is the ability to coordinate the use of these resources. One of the key features that a knowledge infrastructure for discovery should provide is the ability to use resources that have been contributed by various organizations in novel ways to discover new knowledge.

As with enabling discovery, the establishment of standards for the description and exchange of data, as well as interoperability standards for computational resources is a condition for enabling orchestration and coordination. And again, as with discovery, the application and development of technologies that exploit these standards is also essential. We believe that semantic technologies, such as Web Ontology Language (OWL), can play a significant role in meeting both of these requirements.

G. EarthCube should be implementable in an iterative fashion
Finally, the strategies, governance mechanisms, and technologies used to build and deploy EarthCube must support an iterative and progressive implementation of the system. This is perhaps obvious, but worth stating and defining explicitly, since the full vision of EarthCube

will take years to implement and should include currently existing resources (which can be include in the earliest phases of system), as well as yet to be defined, or even imagined, resources.

Conceptual Architecture

To meet the functional and governance objectives outlined above, at least in the near-term, requires an architecture that at minimum can: (1) Exploit existing resources, paradigms and standards; (2) Support and enforce the implementation of new standards; and (3) Excel at enabling easy *Discovery* of EarthCube resources. Meeting these minimum requirements will establish a foundation on which other features can be added, such as orchestration of resources and inference of new knowledge. It also important to view the technological infrastructure as one tier of the EarthCube system, the other tier being the governance and community support structures that must be established in order for the system to truly function well. Here we present some ideas regarding a conceptual architecture that could do well at meeting the above three minimum requirements, that we hope can be used a starting point for discussion.

Key characteristics of this conceptual architecture include:

- Semantic capabilities for discovery of resources and inference of new information
- A peer-to-peer network model
- An Android Intent-like mechanism to support coordination and orchestration

We picture the EarthCube infrastructure as a network of interconnected nodes, much like the internet, on which are deployed EarthCube *Resources* (ECResource). This network, the ECNetwork, will run on the internet as a kind of meta-network, meaning that the resources deployed on the ECNetwork are also available through the internet (like the World Wide Web), but the ECNetwork includes technologies that expose information about the resources, and also the functionalities of these resources, in an EarthCube specific context. Like the internet, the ECNetwork will be a globally interconnected network of computational capacity, data, and other resources. The primary goal of the ECNetwork is to support *Discovery* of *Resources*, as well as *Cooperative Computation*.

The nodes in the ECNetwork are institutional level entities (like internet domains), and are maintained and governed by the organizations and individuals that want to participate as part of the EarthCube. An organization or individual that participates as part of the ECNetwork would be referred to as an ECHub. Each individual ECHub maintains an autonomous, locally administered network of ECResources. These ECResources can include, but would not be limited to, data services (e.g., a THREDDS server), computational capacity, educational resources and human expertise.

An ECHub would be created on the ECNetwork (essentially declared) by deploying an ECDiscoveryService (ECDS), a software component, which enables hub administrators to catalog and describe the resources available from the hub. Each ECDS participates as part of a peer-to-peer (P2P) network with other ECDiscoveryServices deployed by other ECHubs.

Deployed ECDiscoveryServices broadcast information about resources hosted locally to other ECDiscoveryServices on the network. This resource information is added to a local catalog maintained by each receiving ECDS, creating a distributed catalog of resources available within the ECNetwork. This kind of network topology has been used reliably in other network service discovery applications (e.g., BitTorrent). What distinguishes the ECNetwork from a simple multicast domain name system (mDNS), for example, is that the catalogs maintained by the ECNetwork will contain rich, semantic, descriptions of the services available through the network.

The ECNetwork design would follow the same paradigm described in most P2P networks, but modified to allow for an authoritative source for data lookups and retrievals. In this scenario, each ECDS will carry its own catalog as well as basic information about other hubs. As clients browse the network, immediate requests will go to the closest (geographic or logical) hub for lookup. If the closest hub does not contain the actual catalog information, it would perform a lookup to the hub that does have the data. By distributing the catalog among many peers, resiliency and response could be increased. Yet by retaining the concept of an authoritative source each hub can ensure that the data being disseminated is accurate and correct. The P2P model also would make possible distributed download of data (if desired) to assist in making sure that bottlenecks are minimized for large data flight.

Resource descriptions could be created using standard ontologies and encoded in a knowledge representation language like Web Ontology Language (OWL). These ECResource description standards would include existing ontologies, be based on existing metadata or other standards, or be new ontologies that are developed to describe specific kinds of resources. OWL is a W3C specification, and in addition to providing a standard way to encode resource descriptions, also provides the ability to include formal semantics as part of these descriptions. The inclusion of context and meaning (semantics) within the resource descriptions is key to enabling the EarthCube vision. EarthCube users will be able to find and orchestrate ECResources using terminology that is relevant to their areas of interest or research, by entering simple queries that return information about available resources categorized by how these resources can be used and coordinated within the ECNetwork.

The fact that the ECDiscoveryServices are semantically enabled is what creates the ability of the ECNetwork to be a true infrastructure for knowledge management. The ECDiscoveryServices not only help locate resources, but also provide a tool to manage the meaning of these resources based on context and application. For example, an ECDS can describe all of the potential applications of a particular resource, the dependencies of these applications on other resources, where to find these other resources, as well as documentation describing their application. The semantic capabilities of the ECDS also create the potential to build inference components (using reasoners like Pellet[1] that can use this semantic information to create new knowledge or new relationships between resources.

Some EarthCube Resources will be exposed as services, such as data and computational services. Other EarthCube resources will include access to resources such as software source code, publications and documentation, and human expertise. Still other resources will include the ability to provision computational and storage capacity. Where applicable interoperability standards will need to be developed in order for computational service and data based resources to be orchestrated and coordinated.

A possible suitable paradigm around which to develop interoperability standards to support coordination of services could be modeled after (or be, if appropriate once mature) the currently being developed framework called *Web Intents*[2]. *Web Intents* is basically a lightweight framework for distributed inter-application communication modeled on the Android Intents messaging system (paradigm), which is a key feature of the Android operating system (for mobile devices).

Briefly, Android Intents allow applications that run under Android to intercommunicate in a loosely coupled fashion. The Android Intents model is loosely similar to the Apache River

[1] http://clarkparsia.com/pellet/
[2] http://blog.chromium.org/2011/08/connecting-web-apps-with-web-intents.html

(formerly JINI)[3] model for cooperation between distributed systems. Adroid Intents enable applications that need a particular service (e.g., display of an image) to be requested from, and sent to, another application capable of performing the service without needing to know anything about the target application that will ultimately perform the service. To do this the requesting application defines an Intent object that contains a description of the service needing to be performed along with information about the data needed to perform the service. The Intent object (request) is given to the operating system (Android) for resolution. Android acts as a broker and identifies one or more applications capable of rendering this service and hands the Intent to the appropriate application for completion. Applications and other software components register with the operating system the types of Intents they are capable of satisfying. The goal of the Web Intents project is to extend this model to handle inter-communication between applications over the web.

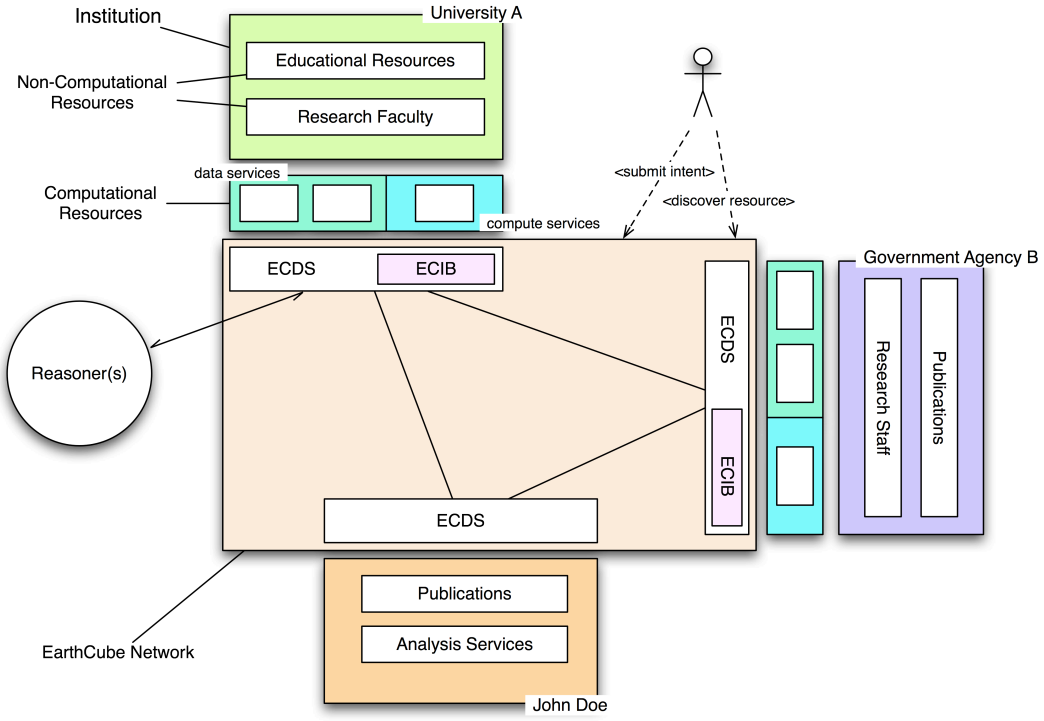

Figure 1: The ECNetwork and its relationship to the Institutions and Individuals who are participating as part of the network. The ECNetwork consists of the set of intercommunicating ECDiscoveryServices (ECDS) that have been deployed by the participating institutions. Independently of the ECNetwork itself, the institutions maintain the computational and non-computational resources that these institutions (ECHubs) advertise to the ECNetwork. ECHubs that wish to make some of their computational resources available for *Cooperative Computation* (orchestration or other automated access), deploy as part of their ECDS deployment an ECIntentBroker, which characterizes these resources and allows them to be part of the ECIntentSystem. ECHubs that do not have computational resources available for *Cooperative Computation* (e.g., John Doe) do not need to deploy the ECIB. The figure also shows a Reasoner (e.g., Pellet) that uses the semantic information contained in the ECNetwork to infer new relationships and/or categorizations, which can then be included as part of the knowledge contained in the ECNetwork. Users access the network by pointing a browser (web or specialized client) at any public ECDS.

An "Intents" system (i.e., the ECIntentSystem) could be developed for EarthCube and deployed on the ECNetwork. From a technological perspective, development of the ECIntentSystem could be done by developing an Intent broker (ECIntentBroker), which would then be deployed in conjunction with an ECDiscoveryService. The ECIB would describe which Resources are available for *Cooperative Computing* and the kinds of Intents they would be capable of fulfilling.

[3] http://river.apache.org/

For *Discovery* and resolution of Intents, the ECIB would participate through the ECDS as part of the P2P ECNetwork, following the paradigm described above.

We realize that the description of this conceptual architecture is incomplete, and that more components than described here would be required to make it a reality. However, we think that it is a reasonable approach to an architectural concept for EarthCube and look forward to discussing it within the community.

Process for design and implementation
The design and construction of the EarthCube ECNetwork could be done by technology project working groups, organized by the overall EarthCube governing committee. These working groups would develop specifications for specific aspects of the system, as well as direct the design and development of specific reference technologies. This is the basic model that has worked well for other open technology development organizations such as the Apache Foundation. However, the specifics of how these working groups are actually organized will need to be decided by the community as a whole.

As an example, the ECNetwork working group, would define the specifications for the ECNetwork in general, as well as the precise specifications for the ECDiscoveryService. The working group would then oversee the building of a reference implementation of the ECDiscoveryService, perhaps awarded as a contract within the community using an RFP process, or implemented by a project group formed by the members of the working group itself. This reference implementation would then be downloadable and installable at any site wishing to participate in EarthCube as an ECHub. Also, as is the case with widely used technologies, such as Java Servlet technology, other implementations of the specification could be made independently by other organizations. These other implementations should provide the same basic functionality as the reference implementation, but could in addition provide enhanced functionality that would be useful to specific communities or types of applications. Other working groups will be formed to implement the various other essential EarthCube infrastructure components.

Funding for the construction of reference implementations and other working group activities would need, at least initially, to be provided by NSF.

Governance
A number of governance models exist already for similar types of projects. These should be explored and evaluated for suitability. For example, two prominent, but different, models for open technology projects are the W3C model and the Apache Foundation model. Regardless of what governance model is ultimately used, the project should start by creating an overall EarthCube governing committee (or board) that will have the ability to create working groups or subcommittees to engage on specific projects during the organizing and early phases of the project. These subcommittees (or project groups) would be responsible for things like defining the ultimate governance model for EarthCube, definition of data and other standards, and the implementation of key project technologies. Members of the governing committee and project groups would be drawn from the EarthCube community and also include members from the NSF and possibly other interested government agencies.

Sustainability
Initially the entire project would need to be funded and sustained by the NSF. Private funding could be enticed if aspects of the EarthCube network proved to be valuable resources for private industry. For example, renewable energy companies (e.g., Wind Farm operators), could have a business interest in being tightly integrated with National Weather Service forecast systems available through the EarthCube. In which case, they could conceivably provide

funding for the development of advanced EarthCube system components or simply for the maintenance of specific ECResources. Ultimately, though, if EarthCube proves to be a valuable (indispensable) commodity for the GEO community, then it's reasonable to expect that the community together would have sufficient incentive to pool resources to improve and maintain the system. In this case, it may also be reasonable to establish a non-profit foundation to raise funds and maintain and improve the aspects of the system, as has been done for other projects (e.g., Wikimedia Foundation).