Perspectives on EarthCube from a National Research Center

National Center for Atmospheric Research

Computation and Information Systems Laboratory

Tim Hoar, Rich Loft, Seth McGinnis, Don Middleton, Eric Nienhouse,

Doug Schuster, Henry Tufo, Matthew Woitaszek, and Steve Worley

with Mike Daniels, NCAR Earth Observing Laboratory, and Matthew Mayernik and Mary Marlino, NCAR Library

NSF's EarthCube thrust is a timely and challenging initiative. Over the past decade, NCAR, its collaborators, and other organizations have made remarkable progress in advancing our community data management capabilities for a broad range of geoscience initiatives. In the climate community, we have seen the WCRP/CMIP archives emerge via the Earth System Grid Federation (ESGF) as coherent multi-model ensembles that enable critical global studies such as the Intergovernmental Panel on Climate Change (IPCC) to take place. NCAR's Research Data Archive (RDA) provides numerous reanalyses of climate data, which are critical resources for many climate-related research endeavors. In weather forecasting, the THORPEX Interactive Grand Global Ensemble (TIGGE) provides a globally networked archive of near-realtime weather forecasts with archive centers on three continents. These efforts, and others like them, have to some degree addressed the foundational needs for federated systems, managed data, common formats, and consistent metadata conventions. Research studies that involve analysis across such collections can generally proceed fairly smoothly. However, once one needs to include the analysis of heterogeneous observational data and/or is engaged in interdisciplinary research, the situation becomes dramatically more complex. It can be difficult to discover data resources and, when you do, there is often a serious lack of common data formats, metadata standards, and/or tools that can work across the disparate data sets.

Our overall vision for EarthCube is a program that dissolves data barriers, supporting people with cyberinfrastructure that can provide the increasing automation necessary to tackle emerging and cross-disciplinary problems. We highlight three science areas that could be advanced via future EarthCube innovation collaboration:

Enabling integration of cross-disciplinary data with increasing automation: We envision an EarthCube infrastructure that spans organizational and disciplinary bounds, supporting cross-disciplinary research. We have submitted several use case whitepapers to the EarthCube Science Requirements repository that highlight this challenge and bring the requirements to life. In the first, we describe a fictional researcher "Ursula" in her effort to work across disparate and unfamiliar data resources [McGinnis]. The framework allows the user to find methodologies,

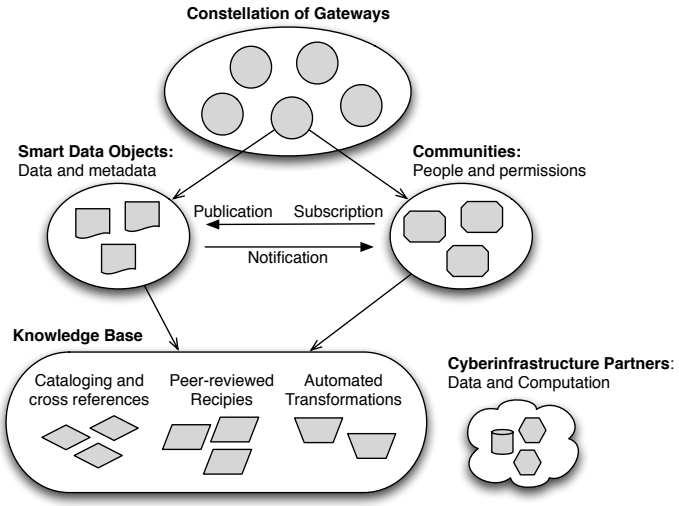
software, and data sets – all of which are aware of their interoperability capabilities and integrated with an automated experiment management system – and then combines data and computation in a unique way to perform novel research that is reproducible. The second is a view from an ecosystem expert, "Hector", working in Latin America [Worley]. Here there is a need to integrate disparate data sources from local to global observations, model data, and information from multiple service agencies. Again, the envisioned interworkability environment is extremely helpful in that it is rich with ancillary metadata, data gathering between science realms uses ontological relationships, analysis is performed with familiar tools, visualization and data integration is browser based, and data curation, software ID tagging, and restart capability leverage cloud-based resources.

Advancing the Conduct of Field Programs: NSF called out field programs as one of the areas that EarthCube should serve and advance. Field programs, like sensor instrumentation, play an important role in providing many communities with critical, fresh observational data about our world, and millions of dollars are spent on the deployment in hopes of obtaining the desired dataset. These data stream in real-time from many sources, are of varying temporal and spatial domains and need to be shared globally with researchers in an interactive environment, often with limited available bandwidth such as onboard an aircraft or over other remote satellite links. Today, emphasis is being placed on data quality as observations are increasingly assimilated into forecast models to improve guidance during missions. As data volumes increase, automation and widespread visualization is playing a critical role in improving data quality to the level required by models. We envision EarthCube providing new capabilities that fundamentally elevate how these complex scientific endeavors are conducted. A vision for an interactive "GeoHub" is described as a compelling science and technology story in which the real time integration of data guides field campaigns [Daniels, et al].

Arctic Science and Data Integration Challenges: For the past several years, NSF has supported the development of the Cooperative Arctic Data and Information Service (CADIS), which was initially aimed at providing comprehensive data management support for the Arctic Observing Network (AON). CADIS has been quite successful in providing a system where each funded AON project can "publish" their data collections along with metadata using a simple web interface. CADIS data is quite diverse, however, and it has generally proven to be quite a challenge to provide higher level services and data integration across the holdings. A new Advanced CADIS (ACADIS) effort will expand support for the broader Arctic science community; its data integration and other challenges are described in another EarthCube white paper [Parsons]. ACADIS represents an effort with some qualities and challenges in common with the much broader EarthCube initiative, including data integration and the diverse teams needed to make progress.

To more clearly identify the key components and technologies that may be involved in EarthCube, we consider the underlying assumptions of McGinnis's "Ursula" cross-disciplinary research use case. In the scenario, EarthCube first makes it simple for the user to find both data and a process description maintained by specialists at a climate research center. The gateway provides tools that can perform basic data manipulations, including subsetting, interpolations, and reformatting, using custom inputs as well as community-provided scripts. Intermediate data is stored and processing is performed on remote resources via EarthCube infrastructure. At

every step of the way, community-provided data sets and transformations are citable, and the final product can be published via EarthCube as well. The simple conceptual diagram below was developed during discussion of the challenges inherent in cross-disciplinary endeavors.



From the use case workflow, several elements of a possible future system become apparent. First, EarthCube gains access to existing data and metadata, as well as people in communities, through federation. More than just files in a repository, data collections must be "smart"; that is, imbued with metadata, knowledge, and methods that provide for interworkability via automated means. People can subscribe to data sets and genres, and data systems notify users of changes. EarthCube then can assemble a greater base of knowledge through cataloging, libraries of published data sets and procedures, and automated operations supported by the underlying infrastructure. The science stories presented here imply the need for community knowledge management and semantic web capabilities.

Building on Success

From the breadth of disciplines and organizations involved in geoscience research, we believe that EarthCube will exist as a constellation of data sources, providers, and gateways that will allow communities to continue taking advantage of specialized tools but participate in the broader EarthCube infrastructure. By federating existing gateways, the communities, data, and knowledge contained in each can be leveraged to enable cross-domain discovery and will lead to further possibilities for automatic data interchange and experiment management. As noted earlier, the geosciences have an impressive range of community CI to build upon. These also provide a wealth of hard-won experience in trying to tackle the challenging problems associated with heterogeneous, distributed, federated environments and the interdisciplinary research that needs support. What follows are a few considerations related to the design of an EarthCube system that arise from our experience building some of the large data systems mentioned earlier, mostly over the past decade.

Systems of Systems

EarthCube will need to provide services for a range of data providers, including individual PI data publication (as part of an NSF data management plan), specific project and departmental collections, and the many large archives that serve as community reference collections. Some

of this may be accomplished via relatively simple web services and interfaces, and some with managed software stacks. To be successful, though, EarthCube will need to provide federation across a broad range of large archives and existing/emergent data networks. This is the System of Systems paradigm found in the Global Earth Observing System of Systems (GEOSS), the emerging WMO Information System (WIS), the Earth System Grid Federation (ESGF), and others described in EarthCube whitepapers, such as OGC and DataONE [DataONE].

When federating broadly – even globally – especially with the large data archives it will be very important for EarthCube to focus on defining interfaces/services such that any organization can easily "play" in the federation. While software stacks and reference implementations can be extremely valuable, they do not always mesh well with existing infrastructure or facilitate the development of innovative new capabilities. This has been one of the challenging areas in some of the national/international federated distributed data systems mentioned earlier.

There are other opportunities to note. The ESGF has recently integrated Globus Online (GO) [Foster] into its architecture as another mechanism for providing data delivery to customers. This experience suggests that a general-purpose set of services such as GO can be creatively reused in a range of specific community contexts where data delivery and upload are handled as a flexible service.

Transformative science will occur as a result of innovation, but innovation can be hard to predict or generate. If EarthCube can define and deliver standards and simple, well-understood interfaces to its federated collections of data, knowledge, and services, and the system is sufficiently open, innovation has a better chance of happening.

Standards are Key

In addition to perhaps sharing architectural strategy with some of these efforts, EarthCube will want to interface with global programs such as these and provide interoperability as well. Well-defined standards – formal, commercial, community, and defacto ones – are a primary key enabler for success and should be a primary focus of EarthCube governance. Across some of the projects in this space, several standards/technologies have been prominent including SAML, PKI/X.509, OpenID, OPeNDAP, THREDDS, netCDF, the Climate and Forecast (CF) metadata conventions, OGC, RDF, OWL, the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH), Globus, Content Syndication and related standards (e.g. ATOM, RSS, OpenSearch), and others.

Getting Security "Right" is Very Important

There is generally a strong desire for scientific data to be completely open; an open policy simplifies matters enormously. The reality, however, is that some collections are associated with policies and processes that demand security in terms of authentication and authorization. Even absent policy, researchers often prefer their work-in-progress to remain private until some time after publication. Thus, data access and security requirements quickly ripple throughout the network and impact systems, services, applications, and tools, requiring that EarthCube must be architected from the outset to support robust and flexible access controls.

As a foundation for data access controls, EarthCube must integrate (or extend) existing community-based identity management from its communities. Technologies such as

InCommon, Shibboleth, OpenID, PKI and so forth have been successfully used but more convergence is needed in this area. EarthCube is going to need to pay careful attention to this important area, and make sure that security development is closely tied in with service providers, and that the user community is delivered a good experience. Also, many existing archives and federations already have various security systems that will need to be integrated. Assuming that EarthCube has a strong social CI infrastructure, there are opportunities here to address identity in very powerful ways. Good decisions in this area are critical and need to be aligned well with community needs and existing CI with long-term support (e.g. XSEDE).

Community Publishing and Citation

Large projects and centers will often have data managers who serve as a bridge between scientific communities and the data resources they need for their research. This has proven to be an effective model for a certain scale of activities. Looking across data-producing projects ranging from small to medium – particularly in light of new NSF data management plan rules – we need better, sustainable strategies for broad community data publishing. In the CADIS work mentioned above, a lot of effort was put into delivering self-publishing tools for data collections and related metadata, and these have worked very well for the researchers. Small university groups often have limited resources available to them which makes transitioning to more advanced workflows extremely difficult. EarthCube will need to focus additional energy on community publication along with data citations, which are key to scientific transparency, reproducibility, and assessment of impact. In the related white paper [Mayernik] the authors discuss the data citation landscape, related issues, and approaches.

Computational and Data Scalability Remains a Growing Concern

Even simple use cases become challenging when the data sets are hundreds of terabytes in size, turning data movement and data transformations into resource-intensive problems. This is a difficult problem now, and getting more difficult rapidly. For example, the current CMIP5 climate simulation effort is expected to generate in excess of 10 petabytes of data distributed and replicated around the world. As EarthCube becomes a fundamental component of scientific investigations, issues regarding the consumption of storage or computational resources, such as estimating the feasibility of executing certain workflows, must be considered. The need for an ability to automatically decide to move a computation to a dataset, or vice versa, is growing commensurately with data volumes and the complexity of analysis workflows.

Community Knowledge Management and Access

In EarthCube, existing standards in each community must be represented, and cross-discipline understanding codified. Areas of standardization, or at least translation, include data structure (format, metadata, physical units, grid projections, and geographical and time references) and data discovery (semantics of descriptive language, differences between disciplines, and data provenance and trust). One promising approach to this problem is through the use of taxonomies and ontologies to connect and organize communities. One consideration is that ontologies might be more useful if layered, so that expert knowledge can be presented at the appropriate level for the intended use. As noted in the Ursula exercise above, the use of semantic knowledge is the first step in enabling automated data transformation, such as making

the system capable of identifying possible operation sequences necessary to prepare an input data set for a particular processing step, for example.

NCAR as an EarthCube Collaborator

NCAR's mission is to support the academic and research community in developing a better understanding of our planet. For EarthCube, we bring upwards of two petabytes of managed scientific data collections, a number of existing standalone and federated data management systems, an aggressive field project program, connectivity to many inter-agency and international data projects, and many science stories that we hope EarthCube can address. We are also providers of a wealth of community cyberinfrastructure including data management software and systems, supercomputers, petascale data storage, digital preservation, and analysis and visualization tools that well-support the broad geosciences and data integration operations. We are thus keen to partner with EarthCube participants across any or all of these areas.

References

Other EarthCube White Papers

[DataONE] DataONE-Enabling Cyberinfrastructure for the Biological, Environmental, and Earth Sciences. http://earthcube.ning.com/group/earthcube-design-approaches/forum/topics/white-paper-dataone-enabling-cyberinfrastructure-for-the

[Daniels] GeoHub: An Interactive eScience Facility for Global Field Observations http://earthcube.ning.com/group/technology-solutions/forum/topics/new-whitepaper-geohub-an-interactive-escience-facility-for-global

[Mayernik] Use Case: Citing Integrated Data Sets. http://earthcube.ning.com/group/user-requirements/forum/topics/use-case-citing-integrated-data-sets

[McGinnis] Example Use Case in Climate Impacts Research. http://earthcube.ning.com/group/ user-requirements/forum/topics/example-use-case-in-climate-impacts-research

[Parsons] Interdisciplinary data discovery and use—The Arctic experience. http://earthcube.ning.com/group/user-requirements/forum/topics/interdisciplinary-data-discovery-and-use-the-arctic-experience

[Foster] Research Data Lifecycle Management as a Service / Globus Online. http://earthcube.ning.com/group/technology-solutions/forum/topics/research-data-lifecycle-management-as-a-service

[Worley] Testing Relationships by Integrating Disparate Data Sources. Upload to EarthCube website pending.