Role of government data agencies, virtual observatories, and communities of practice in sustaining EarthCube

R. Sky Bristol, Roland Viger, Rich Signell, Nate Booth, Mike Frame, Linda Gundersen, Cheryl Morris – U.S. Geological Survey

Summary

In addition to academic partnerships, an efficient, effective, and sustainable EarthCube will require collaboration with government entities which have legislative mandates to maintain and provide access to scientific information and information systems that complement those created by NSF grantees. This white paper describes three bases for interaction between government data agencies with other EarthCube participants. The first is the connection between NSF-sponsored research into cyberinfrastructure and that can be turned into operational infrastructure by government organizations in support of mandates to produce, maintain, and provide scientific data and information. The second basis deals specifically with the creation of a more comprehensive and integrated virtual observatory for earth science. The third basis involves the deliberate facilitation of communities of practice to foster collaboration and manage competition toward shared goals. Although the discussion focuses on the U.S. Geological Survey (USGS) and several specific partner organizations, the principles discussed are applicable to many government and nongovernment EarthCube partners. We conclude the discussion with some specific examples of USGS activities and existing partnerships that serve as examples for interaction between the USGS and EarthCube.

Federal agencies like the USGS have a strong role as a fundamental part of their missions to deploy the kind of cyberinfrastructure imagined in EarthCube. Additionally they have a strong role in providing support to the community through funding, partnering, providing sustainable archives for community use, sharing infrastructure, and actively engaging in education and research that provides solutions. By creating a shared vision of the priority problems to resolve and the goals to be reached within the next 10 years regarding cyberinfrastructure, the community will leverage its resources, find solutions and establish a cyberinfrastructure quicker and more efficiently and possibly move science and innovation forward at a faster pace.

Data Mandate

As discussed here, cyberinfrastructure is considered to include data acquisition, management, storage and accessibility, data mining and integration, and visualization. Cyberinfrastructure evolves rapidly over time and any organization must maintain a degree of currency with changing technology in order to be competitive in whatever market it exists. In the case of government, competitiveness is focused on relevance to society represented by taxpayers and funding representatives.

While the USGS and other earth science organizations have a mandate to provide publicly available and scientifically defensible data and information, we do not have a recognized and sanctioned mandate to execute a corresponding computer and information science program to research and develop the technologies and methods that would keep us

competitive in the larger marketplace. In its relatively long history, the USGS has been responsible for a number of important technological and engineering innovations and patents, many of which have contributed to profitable developments in the commercial sector. The USGS continues to provide leadership and innovation in many areas, but the increasingly rapid pace of advancement in data and information technologies is beginning to impact our ability to fully keep pace in the application of those technologies to our earth science domain. A further confounding factor is the ongoing impetus to consolidate and centralize the business-level information technology expertise and capacity into higher-level governmental entities. While understandable and even laudable from a business efficiency standpoint, this activity has at times negatively impacted the ability of the USGS to apply technology to science.

The National Geological and Geophysical Data Preservation Program (NGGDPP), a collaboration between the U.S. Department of the Interior (DOI) and State Geological Surveys, has served to coordinate the preservation and improve the accessibility of the thousands of collections and millions of artifacts collected by federal and state governments. This program has offered many opportunities to examine the mandates that drive Government agency data management more thoroughly. This effort has engendered conversation about the roles of these agencies as long-term data managers, as information agencies, and as scientific knowledge agencies. The argument has been made by NGGDPP members that the preservation and accessibility of data is fundamental to the combined missions of the Federal and State agencies, much more so than is reflected in the modest funding of the program.

This argument and the reason for the NGGDPP as a funded program are indicative of the continuous evolution in thinking about data management, preservation, and archiving that are also mirrored in the NSF Data Sharing Policy (*http://www.nsf.gov/bfa/dias/policy/dmp.jsp*). What were at one time acceptable practices of data management with individual principal investigators making choices about appropriate disposition of project data have evolved with technology to a desire on the part of funding agencies to see the data assets produced by projects made part of a larger system and available for increasing levels of data intensive science.

These discussions, previous descriptions of the USGS as a "Natural Science and Information Agency" (National Research Council 2001), and ongoing evolution of data management and archiving policy within the Federal Government, and USGS in particular, prompted review of the original Organic Act of 1879 and more funding authorizations for data and information elements of Programs in the USGS. The NGGDPP analysis of these legislated mandates concluded that the USGS has a significant role in building and sustaining cyberinfrastructure in support of its earth science mission.

This mandate for the USGS, coupled with the increasing pace of technological development and demands of data intensive science, point toward a necessary interdependence with the EarthCube research and cyberinfrastructure initiative. The USGS can contribute research and development in some cases where the organization has programmatic funding for cyberinfrastructure. We can also turn research into operational elements within USGS data and information platforms, which in turn will contribute new capabilities to the broader earth science community that uses USGS data. The USGS can also contribute use cases from across our scientific research programs and technical assistance to decision making entities to help inform and guide EarthCube research and development.

Virtual Observatory

EarthCube expresses the need for NSF to build a comprehensive virtual observatory for geosciences in the United States and beyond. While NSF geosciences cyberinfrastructure projects have provided significant benefit in the research and education communities, there is general recognition that this task demands a well coordinated community approach and a clearer pathway developed to sustain the investments made. USGS also has a need for a comprehensive virtual observatory, for organizing the environmental monitoring it conducts, to enhance mission science, and to drive computational models that simulate and integrate the complex earth systems processes necessary for sound natural resource and hazard planning and management. Our partnerships with emergency management and response networks, with global monitoring systems, and with multiple federal agencies and academic institutes engaged in the geosciences requires a robust cyberinfrastructure capable of real-time integrated science.

The USGS can play an important role in the development of EarthCube by providing a research and operations component that: (1) drives research via specific use cases (Operations-to-Research), and: (2) sustains support for research developed infrastructure (Research-to-Operations). This partnership would benefit USGS by strengthening its mission responsibility to provide information and science in societally relevant ways. USGS science and information products would combine with partner observations across the landscape and become federated through a common data backbone. It would demonstrate the science partnerships between academia and federal research which provides an ideal framework for incubating new techniques and technologies, critical for strengthening American competitiveness in the world, while anchoring Earthcube in real natural resource and hazard management needs.

Use Cases - USGS can provide salient applied science and research use cases for what is needed (and what can be gained) from such an integrated system and provide partnership and support to realize these. Use cases derived from current USGS needs such as the National Groundwater Monitoring Network or the National Network of Reference Watersheds provide real world distributed network integration examples with direct links into USGS Programs.

Sustained Support - Once stages of EarthCube have worked through the incubation stage to published research products, the USGS can work with partners to operationalize these capabilities into robust and scalable cyberinfrastructure. While new resources are certainly needed for USGS to fulfill this, USGS programs recognize the importance of this role, described in the 2007 USGS Science Strategy as well as many of the emerging Strategic Science Planning Team reports underway. Once operational, USGS can provide consistency, long-term stability, and integrity to EarthCube components.

Communities of Practice

The competitive proposal process used in science is an excellent and proven way to bring about the best ideas. It works well for the scientific principle of “standing on the shoulders” of scientists and projects that have gone before. The competitive proposal process promotes innovation and novel approaches by forcing proposers to do their homework on what has already been done within and across scientific disciplines. Competition does not, however, generally promote collaboration and does not by itself form the best method for building and sustaining shared infrastructure.

Deliberately facilitated communities of practice can provide a bridge between competition and collaboration to promote meaningful knowledge sharing and cooperation toward the EarthCube architecture. The USGS has experimented with the community of practice concept following a 2006 workshop on scientific information management (USGS SIR 2007). The workshop featured a keynote address from Etienne Wenger, and much of our development of the community of practice is based on his 1998 book on the topic (Wenger 1998). The USGS Community for Data Integration (CDI 2011) along with the Earth Science Information Partners (ESIP; http://esipfed.org/) should be considered as both models for and partners with an EarthCube Community of Practice ("EarthCube CoP"). The remainder of this section discusses lessons learned from the history of the CDI and possible application of those concepts to an EarthCube CoP.

The USGS CDI began in 2008 as the "Council for Data Integration," a steering committee for data integration activities arising out of the 2007 USGS Science Strategy (USGS CIRC 2007). USGS executive leadership and the council began advocating for a community of practice, and this collaborative entity soon eclipsed the council as the body most directly involved in working toward the data integration vision of the USGS Science Strategy.

A 2009 CDI workshop resulted in organizational funding for several "high value opportunities" identified by participants. This seed money was put together with existing projects to fund additional work from several Science Centers and led to ongoing collaboration between disparate and sometimes competing teams that persists today to apply the tools and infrastructure created to a major climate science initiative in the USGS and DOI. Leading up to and following a workshop in the summer of 2010, the CDI began spinning up new working groups, essentially communities in their own right, around specific topics.

Based on experiences from the 2010 projects and input from working group leaders, it became apparent that a change in funding model and community facilitation dynamics was necessary for 2011. While the development project sponsored in 2010 successfully produced good utility for science teams, the nature of that type of work began to edge toward violation of several principles espoused by Wenger:

- The need to show incremental and tangible progress on deliverables for funding began to verge on micro-management on the part of the facilitating organization.
- By funding project work, the USGS as an organization began to see the community as more of an organizational unit and less as a collaborative beyond organizational boundaries.
- The desire for unified products from the community was at odds with the cross-cutting interests of community members.

The CDI has grown from a few people testing the waters of the concept to a dynamic group of many interested scientists and technologists working together in smaller, focused communities. Community members are reporting many of the benefits discussed by Wenger in his quick-start guide (Wenger 2003) such as help with challenges and access to expertise. The USGS as a whole is beginning to see many of the organizational benefits described such as problem solving and innovation. The following discussion on the success factors laid out in Wenger's guide describe the current direction for the CDI community and the role of the USGS as an organizational sponsor.

A Community Needs:

- Domain that energizes a core group – The working groups on data management, technology stack, and metadata have arisen because of a common interest discovered during workshops. These communities are encouraged to stick with the topics until they are no longer energizing to a core group.
- Skillful and reputable coordinator – Working groups have generally been led by a "passionate practitioner," someone with excitement for the particular subject and an ability to lead the group with passion and energy. These leaders have indicated a need for help in the area of facilitation, helping with management of group logistics and communication for a smoothly running community.
- Involvement of experts – All of the working communities have reached out inside and outside the USGS for experts to speak with the group and to become actively involved in some cases. The USGS is committed to facilitating this activity through funding when necessary and other means such as executive-level request for participation.
- Address details of practice – For the most part, working communities have delved deep enough into the details of a given area to retain and encourage the involvement of their members. Communities have started to drift when they are not addressing the real meat of an issue at a level important to the daily work that brought the members together in the first place.
- Right rhythm and mix of activities – The smaller working communities (e.g., "Tech Stack Working Group") have generally found the right mix of regular online discussions and focused problem solving via a variety of means. The larger CDI group has drifted somewhat through longer than necessary meetings and trying to jam too much varied and high-level content into monthly conference calls. The USGS is committed to reevaluating the activities it is helping to sponsor on a regular basis to achieve the best mix.

The Organization Needs:

- Strategic relevance of domain – A team from across the USGS Mission Areas is in the process of developing a 10-year strategic plan for Core Science Systems, the part of the USGS working to directly sponsor and facilitate the CDI. The role of the CDI and its various communities that will evolve over time is recognized as a vital part of this strategy and is being discussed regularly.
- Visible management sponsorship, but without micro-management – As discussed above, the USGS is working toward the proper mix of high hopes and realistic expectations for the tactical and strategic benefits from the community. The CDI continues to have the support and sponsorship at the executive level through both the Core Science Systems Mission Area and the full Executive Leadership Team.
- Evolution of formal and informal structures (Governance) – The management of the CDI has been adapting to changes within the CoP so that it can continue contribute an important part of the long-term strategic trajectory of the USGS.

The CDI started with less than a dozen committed people, mostly from one organizational unit inside the USGS, and has grown to over 200 members, many of whom are now participating from other government and nongovernment organizations. The CDI shares interests, topics, and some members and activities with a similar community, the Earth Science Information Partners, but retains a useful organizational identity in the USGS. More than any direct production of new data integration tools and capabilities, the CDI has

worked to create an intentional space for collaboration between disparate science and technology teams and individuals throughout the USGS where the free exchange of ideas has led to advancements across a broad spectrum.

We feel that the EarthCube endeavor and the "knowledge market" concept discussed in the EarchCube KM video (http://earthcube.ning.com/video/earthcube-km) will benefit greatly from an organizational investment in an EarthCube CoP. This investment will perforce come from both the NSF and partner organizations to the EarthCube endeavor. It will require energy and commitment from its individual participants, only some of which will be funded directly by EarthCube projects, research and development teams, and executive sponsors across multiple agencies and organizations. While retaining necessary identity as the "EarthCube CoP," the community will benefit from formal and informal relationships with other communities such as the CDI, ESIP, and others whose subject matter, membership, and cyberinfrastructure goals dovetail with the objectives of EarthCube.

Collaboration Examples

Technological and methodological innovations created through the EarthCube endeavor promise to both crystalize existing NSF/USGS cyberinfrastructure partnerships and develop new patterns of transitioning research into operations. The following examples offer some thoughts on direct involvement of the USGS community with EarthCube.

DataONE – Data Observation Network for Earth (DataONE) is an emerging organization seeking to ensure the preservation and access to environmental science data through a distributed framework and cyberinfrastructure. A variety of software tools are being written to provide user access to available data. The U.S Geological Survey is actively participating in DataONE cyberinfrastructure and community engagement activities. As DataONE continues to role out its cyberinfrastructure, USGS intends to continue standing up Member Node implementations to allow for consuming USGS data by various earth science community members, and facilitating access to community data by USGS scientists. USGS leadership in DataONE can also be leveraged by the EarthCube activity through both USGS participation in cyberinfrastructure and community engagement efforts. *http://dataone.org/*

Standards Leadership – The USGS also has considerable expertise and capabilities in the areas of Metadata standards and management (Federal Geographic Data Committee (FGDC), Biological Standards FGDC Biological Data Profile, National Vegetation Classification Standards) and supporting cyberinfrastructure. EarthCube should support a comprehensive metadata program in support of documenting its research activities and discovery of research results. Through leveraging USGS Standards, FGDC, ISO, and cross-walks to standards such as EML, Quality Control processes, open-source tools, and infrastructure for data hosting, EarthCube will rapidly gain access to several hundred of thousands metadata and datasets, reduce software development timelines, and help to insure interoperability within the community. *http://www.fgdc.gov/*

ITIS – Additional benefits to EarthCube in the areas of species and taxonomic information can be gained through USGS leadership and support for to the Integrated Taxonomic Information System (ITIS) partnership. The White House Subcommittee on

Biodiversity and Ecosystem Dynamics has identified systematics as a research priority that is fundamental to ecosystem management and biodiversity conservation. This primary requires improvements in the organization of, and access to, standardized nomenclature. ITIS was designed to fulfill these requirements through an accessible database with reliable information on species names and their hierarchical classification. *http://www.itis.gov/*

Geo Data Portal and "GeoProcessing Services" – The USGS Center for Integrated Data Analytics produced a completely standards-based engine for server-side processing of major climate data sets to provide data inputs to simulation models and other analytical tools. The project has contributed heavily to technology development and testing on a number of fronts from THREDDS server and the NetCDF data format with UNIDATA to the analytical use of Web Coverage Services provided by ESRI ArcGIS Server. The overall platform and architecture centered on the Open Geospatial Consortium (OGC) Web Processing Service standard shows tremendous promise as a pattern for additional "GeoProcessing" and other types of services in data intensive science. *http://pubs.usgs.gov/of/2011/1157/*

USGIN – The USGS is a partner in the NSF-sponsored effort for the U.S. Geoscience Information Network (USGIN). The partnership has involved participation in developing metadata standards and profiles, contributions to technical architecture, testing and evaluation of proposed technology solutions, and contribution of geoscience records from the NGGDPP. The fundamental distributed data network architecture of USGIN and its basis in the OGC Catalog Service for the Web (CSW) standard is foundational to the growing distributed network of USGS data providers. The architecture has also been adopted by the U.S. Department of Energy for the National Geothermal Data System currently under construction. Much of the data cataloging architecture within USGS has benefitted from work pioneered in the USGIN project and our connection to that effort. *http://usgin.org/*

ScienceBase – The USGS ScienceBase is a standards-based platform for scientific data management, documentation, and data exchange for team collaboration. While ScienceBase has a large-scale vision for enterprise-level data management and services, it is being constructed using smaller-scale projects to ground the architecture in real world use cases (e.g., NGGDPP, Landscape Conservation Cooperatives led by the U.S. Fish and Wildlife Service, etc.). The projects using ScienceBase as a "ground-level" data management platform are generating use-cases important to larger cyberinfrastructure development. With its open architectural approach, coupled with a committed programmatic support system, ScienceBase is a candidate to operationalize research products from EarthCube. *http://www.sciencebase.gov/*

The National Map – The flagship product of the USGS National Geospatial Program is The National Map. At its heart is a robust, quality-controlled "geospatial database" of the base data needed for the topographic map that is the legislative responsibility of the USGS. The evolving technological architecture used to manage and serve The National Map relies on standard web services. The National Map provides a framework for other earth science work, effectively forming a geospatial fabric on which to "anchor" nearly all scientific data assets produced and used by the USGS. Recent strides by the National Map program, in collaboration with several science teams, to enhance the OGC Web Map Context standard promise rapid advancements in the delivery of thematic map packages across a wide variety of platforms. *http://nationalmap.gov/*

National Geologic Map Database – The USGS National Cooperative Geologic Mapping Program (NCGMP) maintains a robust catalog and digital repository of geologic maps at multiple scales from federal, state, and academic partners. The National Geologic Map Database (NGMDB) provides a wealth of information and is working on new data delivery mechanisms to provide national geologic map web services for broader use and integration into many platforms. The NGMDB is a good candidate to work with the results of EarthCube cyberinfrastructure research in the advancements of its services and data delivery mechanism. *http://ngmdb.usgs.gov/*

References

Community for Data Integration (CDI). CDI Web Site. 2011. Accessed 10/15/2011. (*https://my.usgs.gov/confluence/display/cdi/Home*).

National Research Council. Basic Research Opportunities in Earth Science. 2001. National Academy Press, Washington, DC.

U.S. Geological Survey (USGS). Proceedings of the First U.S. Geological Survey Scientific Information Management Workshop, March 21–23, 2006. 2007. Scientific Investigations Report (SIR), 2007-5232, 94p.

U.S. Geological Survey (USGS). Facing Tomorrow's Challenges—U.S. Geological Survey Science in the Decade 2007–2017. 2007. U.S. Geological Survey Circular (CIRC), 2007-1309. 70p.

Wenger E. Cultivating communities of practice: a quick start-up guide. 2003. Retrieved 2011, from http://www.ewenger.com/theory/start-up_guide_PDF.pdf

Wenger E. Communities of Practice: learning, meaning, and identity. 1998. Cambridge University Press. ISBN: 0521663636, 9780521663632.