White Paper Submitted to the NSF EarthCube Charrette, November, 2011

Geological Survey Contributions to Building Cyberinfrastructure in the Geosciences

M. Lee Allison[1,3,4], Linda C. Gunderson[2], Kevin T. Gallagher[2,3], Stephen M. Richard[1], and Vivian Hutchison[2]

[1]Arizona Geological Survey, 416 W. Congress St., Ste. 100, Tucson, AZ 85701
[2]U.S. Geological Survey, Reston VA
[3]US Geoscience Information Network Steering Committee
[4]Association of American State Geologists

ABSTRACT

There are significant challenges to building the geoinformatics component of the cyberinfrastructure for the sciences as envisioned by the National Science Foundation (NSF) (Atkins, 2003). These challenges include: creating common standards and protocols; engaging the large community of resource providers; establishing practices for recognition of and respect for intellectual property; developing simple data access and resource discovery systems; building mechanisms to encourage development of web service tools for analyses; creating sustainable business models for continuing maintenance and evolution of information resources; and integrating the data management life-cycle into the professional and cultural practice of science. The Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS) agreed in 2007 to pursue design and implementation of the U.S. Geoscience Information Network (USGIN) as a facilitating step in realizing a geoinformatics component of a cyberinfrastructure for the sciences. The USGIN is building catalog and web services to access the large and diverse information holdings of the geological surveys in the United States as well as tools and applications to understand and analyze the information. Through adopting Open Geospatial Consortium (OGC) standards and coordinating the creation of the necessary components, we have been able to leverage not only the efforts of the geological surveys but of many initiatives in government, industry, and academic institutions to begin assembling a comprehensive and potentially global network of catalogs, services, and distributed data holdings.

We feel that the EarthCube vision to promote an integrated system for discovery and access to geoscience information can be realized with an approach emphasizing the use of open-source standards, connecting data provider and consumer software components using standardized interfaces and interchange formats, and committing resources to support community engagement and education. Community-based operating standards and protocols developed with a requirement-driven design methodology make incorporation of additional data and information to the USGIN is a simple process that reflects the evolution of the World Wide Web. However, meeting the challenges and realizing the potential of this system is as much an education and social engineering project as a technical project. Thus, the geoscience community must self-organize and make concrete decisions and commitments to technical specifications and conventions for resource registration, publication, citation, and preservation. EarthCube provides a golden opportunity to move this process forward and USGIN provides the opportunity to leverage an existing national network.

Necessity of a U.S. Geoscience Information Network among Geological Surveys

Geological surveys have unique resources and mission-specific requirements that include the gathering, archiving, and dissemination of long-term Earth science data. These resources constitute one of the largest, most extensive collections on the geology and natural resources of the United States. Historically, these data and information have only been available in paper format or in disparate digital systems, which require significant time and resources to explore, extract, and reformat. In early 2007, the Federal and State geological surveys in the United States agreed to the development of the U.S. Geoscience Information Network (http://usgin.org, http://lab.usgin.org) as a data integration framework that is distributed, interoperable, uses open-source standards and common protocols, respects and acknowledges data ownership, fosters communities of practice, and is based on web services and clients (Allison and others, 2008a). The "USGIN" as the network has come to be known, has attracted collaborators across government, industry, and academic institutes and working groups, including such organizations as the U.S. Department of Energy, Energistics, Microsoft Research, and the San Diego Supercomputer Center.

By using modern information technology and a loosely-coupled SOA design that provides standardized discovery tools for data access, the geological surveys and the general science community will benefit in multiple ways. First, information resources from each survey will be more readily available to the world audience. Second, interoperability will enable data and applications from external sources (databases, catalogs, and inventories) to be readily utilized with each participant's local data system. Third, a large, federated data network will create opportunities for the broader community, including academia and the private sector, to build applications utilizing this huge data resource, and to integrate it with other data. The breadth and depth of survey-based data constitute one of the largest data resources in the geosciences, in essence, a national data "backbone." By building upon existing community-based practices and buy-in, we help ensure that the network becomes self-sustaining.

A Community-Based Governance Model and CI Architecture: the Vision for USGIN

The design and evolution of USGIN is also based in a community of practice approach (Wegner, 1998) meaning that participants in USGIN learn, develop, evolve, and coordinate the building of the network with each other. The vision for USGIN, bulleted below, is still based upon the original principals that were articulated at the 2007 workshop (Allison and others, 2008b) and agreed upon by the AASG and USGS.

- Develop a coordinated, national geoscience framework to access and integrate state survey and USGS-information resources.
- Function as a "community of practice" in developing the geoscience network.
- Develop prototypes (pilots, test beds) to show proof of concept, to determine realistic levels of effort, and to compare costs and benefits while providing immediate benefits in the form of user services.
- Build the network through an iterative and evolutionary process.
- The basic architecture of the network should be distributed and leverage existing systems, map services, and data with local autonomy, by using standards to enable interoperability, portability, and reusability.
- Review, test, and adopt standards and protocols for developing the system including metadata and Open Geospatial Consortium (OGC) protocols and standards (http://www.opengeospatial.org/).
- Help develop and adopt GeoSciML (geoscience mark-up language) as a protocol and consider proposing it as a standard to the Federal Geographic Data Committee.
- Recognize that there are priority resources for which the geological surveys have mission requirements and inherent partnerships, including data and information on bedrock and surficial geology, geochemistry, geophysics, mineral and energy resources, geologic hazards, water resources, and subsurface information such as borehole and well data.
- Encourage Web clients and services to be developed and facilitate participation and

implementation by others in a manner that meets their own business model and needs.

- Reduce philosophical and cultural barriers that impede system development.
- Adhere to a code of conduct that respects and acknowledges data ownership and the work of others. Respect intellectual property and data provenance, use "branding" in data services to acknowledge data sources. Develop usage measurements and utilize them with Web clients and services.
- Develop a database-citation format.
- Acknowledge that geological surveys need to recognize interoperable, web-enabled information resources as part their mission. The surveys also must seek partnerships to leverage resources, develop, and implement the vision.

The Design Process: Conceptual Elements for Data Integration and System Design

When more fully implemented, we envision a scenario where any user can search all USGIN catalogs through a simple web interface, which might be provided by any geological survey or other network participant. Applications being developed and tested now will enable browsing of available data geospatially for a specific area, and then accessing selected data through web services. Data that are provided in common interchange formats can be utilized by any number of applications, including in-house, freeware, and proprietary commercial products that implement OGC standards. It is intended that the original data source would be credited with the download. This type of "decoupled" system where the data providers need not know details about the clients or user applications and vice versa provides ease of use and contrasts sharply with centralized systems where data can only be accessed by a dedicated client that is custom built for that application. This latter design restricts or prohibits interoperability and hinders open integration of data and services.

The most critical system components of USGIN include standardized catalog services to register and discover resources, web map services to display georeferenced images, and feature services to transport data (Richard et al., 2009). The USGIN project is currently implementing discovery using OGC Catalog Service for the Web (CSW), georeferenced map-image delivery using OGC Web Map Service (WMS), and geologic data using Web Feature Service (WFS) (Richard & Grunberg, 2010). Wherever possible, we are leveraging the results of open source projects to avoid duplicating development effort, and to keep the cost of implementation as low as possible. We are also developing or working with collaborators on CSW services including the USGS on the ScienceBase catalog, and the GEON portal (http://www.geongrid.org/). We have tested the open source client application CatalogConnector (http://sourceforge.net/projects/catalogconnector/), and the ArcGIS Geoportal client, as well as a simple web client built using open layers to provide access to catalog services. Most GIS software packages already function well as WMS clients. An ArcGIS client for GeoSciML WFS being developed for USGIN will load data into the standard format for the publication of geologic maps (NCGMP09) for client-side utilization. Another critical aspect of USGIN will be the development of tutorials and workshops to assist others to bring new data and services into the network.

The GIN approach to data integration involves adopting existing components and leveraging work from other projects and by other developers. Multiple projects underway at both USGS and AASG will deliver the key components to enable and deploy USGIN. The USGIN approach is to contribute to a data integration framework that is adopted and promulgated voluntarily because it works and meets the needs of both data providers and data users. Both the USGS and AASG are developing components, specifications, and services collaboratively and semi-independently within this organic framework. This adaptability is a core attribute that is fostering implementation not only across both USGS and AASG, but to a rapidly growing broader community (Keller et al, 2007). Numerous partnering efforts are in negotiation but significant ones are established. The following describes some of those partnerships.

Data Integration at the USGS
The USGS Science Strategy (USGS Circular 1309, 2007) released in 2007 identified data integration as one of its cross-cutting strategic science directions and states: *"The USGS will use its information resources to create a more integrated and accessible environment for its vast resources of past and future data. It will invest in cyberinfrastructure, nurture and cultivate programs in Earth-system science informatics, and participate in efforts to build a global integrated science and computing platform."*

USGS is using SOA design principles in constructing a new architecture for all USGS data and science applications; a complex challenge for a 131-year-old institution that has been collecting earth science data since its inception (Gallagher et al, 2007). This effort requires operating on many aspects of architecture creation simultaneously, while dealing with extensive legacy analog and digital data. The approach is to create tools and services that assist with the scientists' work flow process while addressing all aspects of the data management life-cycle. The USGS supports the use of Open Geospatial Consortium standards and is working with Unidata (https://www.unidata.ucar.edu/) to implement their data access and management tools: THREDDS (Thematic Realtime Environmental Distributed Data Services) and NetCDF (Network Common Data Form), as well as their data access protocol OPeNDAP (Open-source Project for a Network Data Access Protocol).

Projects are underway that include building a federated database network, a master metadata catalog called ScienceBase, creating new web services for discovery of data, creating community specific data models and vocabularies, creating easy to use registry and data upload applications, providing tools for modelers to integrate modeling outputs, and building integrated earth system science applications (Gundersen, 2008). These efforts are driven by the USGS Community for Data Integration (CDI). The development of ScienceBase is leveraging the technology being used in USGIN to employ an open standards cataloging method (OGC-CSW). This specification will provide, among other things, a way for ArcGIS users to query directly for all available map-type services that can be incorporated directly into ArcGIS projects. In addition, other USGS catalogs such as the National Digital Catalog of Data and Materials (http://datapreservation.usgs.gov/index.shtml) can be readily accessed from a single search. Visualization tools such as the National Map (http://nationalmap.gov/), the Mineral Resources On-Line Spatial Data service (http://mrdata.usgs.gov/), and the National Water Information System (http://waterdata.usgs.gov/nwis) are also searchable using the ScienceBase catalog.

National Geothermal Data System: A coalition of state geological surveys (via AASG) is expanding and enhancing the National Geothermal Data System (NGDS - www.geothermaldata.org) by creating a national, sustainable, distributed, interoperable network of data providers representing all 50 states that will develop, collect, serve, and maintain geothermal-relevant data that operates as an integral compliant component of NGDS (www.stategeothermaldata.org). The data exchange mechanism is built on the USGIN protocols and standards.

Data are exposed from the state geological surveys through the NGDS, by digitizing at-risk legacy, geothermal-relevant data (paper records, samples, etc.), publishing existing digital data using standard web and data services, and through limited collection of new data in areas lacking critical information.

Goals are to enhance States' abilities to preserve and disseminate geothermal data; facilitate geothermal resource characterization and development efforts; expand the scope of data available to the geothermal community; foster new services and applications built by third-parties to take advantage of the system's capabilities and content; contribute materially to creation of a national geoinformatics system through implementation and deployment of NGDS; and increase operational support for geoinformatics infrastructure through broader user base.

Energy Industry Metadata Standards Working Group*:* The USGIN project is participating in the Energistics' consortium's Metadata Standards Working Group (http://www.energistics.org/metadata-work-group), to develop a petroleum industry metadata profile that is compatible with metadata services for other geoscience domains.

OneGeology: The OneGeology (1G, www.onegeology.org) initiative to make accessible online digital geologic map data for the world has 116 participating countries, providing more than 120 map services from 46 nations using OGC WMS and WFS through a dynamic web portal. OneGeology–Europe (1G-E, www.onegeology-europe.eu/) is a European Union project in which 26 national geological surveys and organizations are collaborating to build a continent-wide geoscience data network. Developers from 1G-E and USGIN continue to collaborate on common standards, protocols, procedures, specifications, and design with the goal of making the two systems fully compatible and interoperable. Emerging practices from the global project, 1G, and the regional initiatives 1G-E, and USGIN, provide a foundation to create a comprehensive global digital data network of geoscience (and geospatial) information. The next step is providing structured data for geoscience features using OGC WFS's utilizing GeoSciML as the data transport schema.

Sustainability

One of the challenges facing not only the field of data integration but all of geoinformatics is sustainability. Many worthwhile projects have disappeared at the end of the grant funding cycle because of the lack of long-term cyberinfrastructure to maintain them. A benefit of the geological surveys is that they are government entities that will likely continue with their core missions and thus, providing continuing development of USGIN. A USGIN sustainability path is emerging as additional groups and companies adopt the framework creating a broad user and contributor base, with growing demand for its services. A broadly deployed system means that the cost of maintenance can be spread among a larger community so that no one group or organization is burdened with it. Loose coupling between data providers and consumer applications reduces the number of critical components, and use of standardized services and interchange formats enables data and application portability to facilitate preservation. The initial validation of this approach by USGS and AASG set the stage for national deployment and continuity from the start. Subsequently, the use of USGIN in NGDS, and participation by a growing cadre of companies, State and Federal agencies, and data integration and networking projects in related sciences, augurs well for the evolving approach.

Roadmap

The USGIN Working Group envisions further development of tools and capabilities and extending the community of practice involving geoinformatics practitioners from the USGS and state geological surveys. Promoting engagement and participation of the state geological surveys, and increasing communication between the states, USGS, and other stakeholders are prerequisites for community development. A key element of community building is personal interaction; face to face meetings take time and money. We propose that maximum impact can be achieved by using the existing USGS CDI, Open Geospatial Consortium (OGC), and ESIP meetings to bring stakeholders together.

Within this framework, the USGIN community can establish an identity for geological survey informatics practitioners, work to prioritize technical development that is specific to the geological survey community, and leverage development taking place in the larger community. Policies, protocols, and procedures for developing, reviewing, and distributing specifications can be adopted from practice developed by existing organizations, for example the OGC. Documenting and promoting best practices through demonstrations, education, and outreach within the geological survey community is paramount for fostering deployment of interoperable services for data discovery and distribution.

These presuppositions and objectives predicate priorities for the next five years:

- Community building
    - Promote face to face engagement by supporting participation in CDI, OGC, and ESIP meetings (immediate)
    - Organize coordinating committee to shepherd community
- Prioritize effort
    - Nucleate efforts based on program and project requirements and personal interests
- Identify specific deliverable products (two test beds, 6-12 months; ongoing for duration)
- Improved communication
    - Online collaboration in groups with particular objectives
- Deliver product
    - Demonstrate capabilities and usefulness (18-24 months; 6-12 months after deliverables are identified)
- Develop and disseminate outreach and educational materials
    - Workshops, tutorials, online resources, publications (12-36 months, ongoing)

Although these objectives initially are sequential, as the community evolves, all of these will need to proceed in tandem. Approximate time horizons are indicated for key steps in the process for some initial high priority activities.

A critical component to help achieve the vision for a Geoscience Information Network is to reinforce the development of a community of practitioners. To foster a sense of identity and organization for the community, we recommend formation of a coordination group with representatives from the scientific and IT communities. This group will consist of representatives from the USGS, state geological surveys, and the broader community; while the group should be broad based it should still be small enough to be agile.

Community development is beginning to occur through collaborations within the CDI at the USGS, and through the AASG Geothermal Data project managed by the Arizona Geological Survey. Recruitment and training to bring in individuals interested in the nexus of information engineering and geoscience is an ongoing priority. We propose that growth of the community should be reinforced by collaborating on two test bed activities engaging with more experienced communities at the Open Geospatial Consortium and ESIP. Depending on priorities established, these efforts will test and develop practices, data publication specifications, and interoperability formats using map, feature, and observation services. Data registration, catalog, and discovery specifications should be enhanced to promote accessibility. Activity organized around specific priorities and objectives is essential so that participants receive a return on their investment in time and effort and have the sense that something is getting done. Tests beds have fostered communication, alignment of activities, and exchange of expertise and capabilities in the OGC community. Engagement of students in the test-bed deployments will be key to training the workforce necessary to build and maintain the system.

Outcomes

Evolution of the current Balkanized geoinformatics practice into a more cohesive and effective community has been and will continue to be an incremental process. The role of USGIN as an entity in this larger community requires organization, planning, promotion, and funding. An advisory committee cannot plan for all aspects of this process – As a community activity, it must be organic and emergent process, but there are some strategies that can be identified as essential in providing valuable outcomes.

- Provide a **Catalogue Service** that is adaptable to several existing standards.
- Provide an **Access Portal** into the USGIN.
- Provide **Documentation** so that the public can provide additional access portals, either through the Web or through desktop software applications.
- Establish a **Long-Term Governance Model** in order to represent the geological surveys in the larger geoinformatics community and provide a sense of leadership. This will also provide a source of authority on how decisions for priorities and recommendations are made.
- Develop a **Business Model** in order to build financial support, whether through an independent non-profit foundation or through the incorporation with an existing organization (such as ESIP or OGC), having such a legal entity would allow the ability to enter contracts, receive funding, pay salary, and make grants of funding. Since many geological surveys have data archive and dissemination functions as part of their portfolio, some support for the system might be built into their operating expenses and overhead.
- Explore additional **Test-bed Opportunities** to utilize existing service protocols and interchange formats, as well as off the shelf open-source software or widely deployed commercial software for service deployment. These efforts must include identifying and contacting target communities and exploring possible contributions to the costs of the system development and maintenance.
- Develop a **Marketing Strategy** as an education and outreach program to inform and engage data providers who will need to realign their existing approaches to data delivery, to make users aware of new resources and how to use them, and to interest students in geoinformatics as a career. Monitoring of network resource usage and collection of input from the user community on what is working and what is not should be the basis of this strategy.

Summary

The US Geoscience Information Network is developing and deploying a framework for distributed, loosely coupled, interoperable data publication and access utilizing standardized service interfaces and interchange formats. This network is being developed by the state and federal geological surveys, but the approach has applicability across the geoscience domain, and reflects the evolution of the World-Wide Web into a linked-data information system. Meeting the challenges and realizing the potential of this system is as much an education and social engineering project as a technical project. The geoscience community must self-organize and make concrete decisions on technical specifications and conventions for data registration, publication, citation, and preservation. EarthCube provides a golden opportunity to move this process forward.

Acknowledgements

Parts of this paper are drawn from a draft report on the future of USGIN, prepared by a working group under the leadership of Viv Hutchison of the USGS and Steve Richard of AZGS. Members of the Working Group include: David Arctur, Open Geospatial Consortium (OGC); Bob Cook, Oak Ridge National Laboratory (ORNL) & DataOne; Rob Fatland, Microsoft; David Ferderer, US Geological Survey (USGS); Ted Habermann, National Oceanic and Atmospheric Administration (NOAA); Carol

Meyer, Earth Science Information Partners (ESIP); Anna Milan, National Oceanic and Atmospheric Administration (NOAA); Satish Sankaran, ESRI; Erin Robinson, Earth Science Information Partners (ESIP); Jerry Weisenfluh, Kentucky Geological Survey; Ilya Zaslavski, Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI).

This report is based in part on work done at the Arizona Geological Survey supported by the Department of Energy under Award Numbers DE-EE0002850 and EE0001120, in conjunction with related work on development of the Geoscience Information Network (US GIN) supported by NSF grant EAR-0753154.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Allison, M. Lee, Dickinson, Tamara L. and. Gundersen, Linda C. (2008a). Final Report: A Workshop on the Role of State Geological Surveys and U.S. Geological Survey in a Geological Information System for the Nation, submitted to National Science Foundation for Award 0723437 to the Association of American State Geologists, March 1, 2007 to February 29, 2008. Arizona Geological Survey Open-file Report 08-01. 23p.

Allison, M. Lee, Gundersen, Linda C., Richard, Stephen M., and Dickinson, T.M. (2008b). Implementation Plan for the Geosciences Information Network (expanded abstract), In Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds. *Geoinformatics 2008—Data to Knowledge, Proceedings*: U.S. Geological Survey Scientific Investigations Report 2008-5172. pp. 9-11.

Allison, M. Lee, Gundersen, Linda C., Richard, Stephen M. (2011). Geoinformatics in the Public Service: Building a Cyberinfrastructure Across the Geological Surveys, In R. Keller & C. Baru, eds, *Geoinformatics*, Cambridge University Press,

Atkins, D. (2003). Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation Technical Report (http://www.nsf.gov/od/oci/reports/atkins.pdf).

Gallagher, Kevin T., Bristol, Sky R., and Gundersen, Linda C. (2007). A Data Integration and Interoperability Blueprint for the U.S. Geological Survey, (expanded abstract). In Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds. *Geoinformatics 2007-Data to Knowledge, Proceedings*: U.S. Geological Survey Scientific Investigations Report 2007-5199. pp. 58-60.

Gundersen, Linda, C. (2008). Answers to Earth Systems Science Questions: The evolution of Geoinformatics at the U.S. Geological Survey. In Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds. *Geoinformatics 2008—Data to Knowledge, Proceedings*: U.S. Geological Survey Scientific Investigations Report 2008-5172. pp. 2-3.

Hutchison, Vivian, and Richard, Stephen M., in preparation, Recommendations for the Future of the U.S. Geoscience Information Network, 11p.

Keller, G. Randy, Maidment, David, Walker, J. Douglas, Allison, M. Lee, Gundersen, Linda

C., and Dickinson, Tamara, M. (2007). A Community Workshop and Emerging Organization to Support a National Geoinformatics System in the United States, (expanded abstract). In Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds. *Geoinformatics 2007-Data to Knowledge," Proceedings*: U.S. Geological Survey Scientific Investigations Report 2007-5199. pp. 75-76.

Percival, George, Editor, 2006-03-20, Interoperability Testbed Policies and Procedures: Open Geospatial Consortium Inc., document OGC 05-129r1.

Richard, Stephen M., Allison, M. Lee, . Clark, Ryan J., and Grunberg, Wolfgang (2009). US GIN: Interoperable Geoscience Data Services on the Web – How Do We Get There? Geological Society of America Abstracts with Programs, **41**(7), 99.

Richard, Steven M. and Grunberg, Wolfgang, eds. (2010). Use of ISO19139 xml Schema to Describe Geoscience Information Resources v. 1.1. Arizona Geological Survey Open-file Report OFR-10-02, 134 p.

U.S. Geological Survey. (2007). Facing tomorrow's challenges—U.S. Geological Survey science in the decade 2007–2017. U.S. Geological Survey Circular 1309, 70 p.

Wenger, E. (1998). Communities of practice: learning, meaning, and identity. New York: Cambridge University Press.