Recommendations for the Future of the U.S. Geoscience Information Network

Introduction

The U.S. Geoscience Information Network (USGIN) is a partnership of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS), who formally agreed in 2007 to develop a national geoscience information framework that is distributed, interoperable, uses open source standards and common protocols, respects and acknowledges data ownership, fosters communities of practice to grow, and develops new Web services and clients. The National Science Foundation (NSF), the Department of Energy through a connection with the National Geothermal Data System, and the USGS through a partnership with the ScienceBase project jointly funds USGIN.

This document provides a recommended road map and objectives for USGIN activities over the next 5 years, based on discussions and input from a working group consisting of USGIN stakeholders (Table 1.) The Working Group held several teleconferences over an 8 week period and focused on two interrelated topics: (1) how community-driven programs approach the issue of sustainability and (2) development of a strategic direction for USGIN to advance its goals and objectives.

Table 1. Working Group Participants

Name	Affiliation
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)
Viv Hutchison	US Geological Survey (USGS)
Carol Meyer	Earth Science Information Partners (ESIP)
Anna Milan	National Oceanic and Atmospheric Administration (NOAA)
Satish Sankaran	ESRI
Stephen Richard	Arizona Geological Survey, USGIN
Erin Robinson	Earth Science Information Partners (ESIP)
Jerry Weisenfluh	Kentucky Geological Survey
Ilya Zaslavski	Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI)

Background

Representatives of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS) met in Denver February 21-22, 2007, to discuss opportunities for making their data more accessible and interoperable across agencies. They recommended that the USGS and State Geological Surveys work together to create a distributed, national "Geological Information Network" (GIN) of digital Earth Science data using common standards and protocols, preserving ownership, credit, and control of data, and building on existing data systems (AZGS Open-file Report 2008-01, 2008).

The intention of the US Geoscience Information Network is to benefit the geological surveys by reducing the cost of online data publication and access provision, and to benefit society through easier (lower cost) access to public domain geoscience data. This information supports environmental planning, resource-development, hazard mitigation design, and decision-making. GIN supposes that sharing resources for system development and maintenance, standardizing data discovery and creating better access mechanisms, causes cost of data access and maintenance to be reduced (see Shapiro, 2000). A study by the German Institute for Standardization concluded that the economic benefits of standardization range between 0.2 and 0.9% of the gross national product (DIN, 2000; Blind et al., 2011). These studies focused on standardization in a wide variety of business domains, we suggest that they also apply in the informatics domain. Although anecdotal, consider how the music industry landscape has changed with standardized file formats and metadata schemes for recordings, or the seamless connection of most printers to computers using standard interfaces and interchange formats. Standardized access to rich data resources will create collaborative opportunities in science and business. Development and use of shared protocols and interchange formats for data publication will create a market for user applications, facilitating geoscience data discovery and utility for the benefit of society.

Since the 2007 meeting, the geoinformatics community has continued to evolve, with the emergence of new activities and partnerships that may impact USGIN. Following are some examples:

- The National Geoinformatics Community (NGC) is an informal collection of academic institutions and projects that advance geoinformatics at all levels via outreach, advocacy, and fostering communities of practice. The NGC concept has evolved from several workshops, town hall meetings, and the work of an exploratory committee that met with existing and successful community efforts such as UNAVCO[1], Incorporated Research Institutions for Seismology (IRIS)[2], Joint Oceanographic Institutions (JOI)[3], and Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUASHI)[4]. Identified objectives for NGC include cataloging and communicating community efforts, identifying and facilitating development of community standards/specifications, and gaining a better understanding of the geoinformatics-related requirements of geoscientists and educators. No formal organizational structure has been established at this point for NGC.
- The USGS Community for Data Integration (CDI) was established in 2009 to provide a forum within the USGS to facilitate data discovery, exchange, and interoperability of data and information for scientific computing, to create awareness of relevant data products, and to promote interdisciplinary science through the interaction and integration of scientific information. The CDI also provides a mechanism to deploy

[1] http://www.unavco.org/unavco.html
[2] http://www.iris.edu/hq/
[3] http://www.oceanleadership.org/2004/joint-oceanographic-institutions-to-lead-us-efforts-in-iodp/
[4] http://www.cuahsi.org/

consistent processes, protocols, and data management to implement the USGS Data Integration Strategy[5]. The CDI is supported by the Core Science Systems (CSS) Mission Area of the USGS in partnership with the USGS Office of Science Quality and Integrity. Community-led working groups meet monthly, and annual workshops encourage collaboration, showcase and exchange ideas, and create networking opportunities across the USGS partner network. Additionally, CSS supports one staff member to host a virtual monthly meeting, keeping the community informed of innovative data integration contributions occurring in the USGS and by USGS partners.

- The Federation of Earth Science Information Partners (ESIP) is a broad-based, distributed community of data and information technology practitioners that voluntarily come together to leverage collaborative synergies that create coordinated interoperability efforts across domain-specific communities. Participation in the ESIP Federation allows members to expose, gather, and enhance their own in-house capabilities in support of their organization's own mandates. The ESIP Federation has a 13-year track record of working on Earth science interoperability solutions that connect distributed and heterogeneous communities. Ultimately, these collaborations allow data interoperability a greater chance of success. The ESIP Federation's operations are managed by the Foundation for Earth Science and governed by a slate of officers and several standing committees, and two staff members. Development efforts are community driven, through the formation of working groups and clusters.

These activities and others may provide further opportunity for USGIN to develop a sustainability model that includes key partners by broadening the focus and leveraging activities under way.

USGIN Today

The basic hardware and software components of the GIN system are diagrammed in Figure 1.

In the past few years, USGIN has progressed from a conceptual idea to deployment of tools and capabilities critical to efficient access and dissemination of geosciences information. Examples include:

- A metadata catalog and search tool using the ESRI Geoportal, a free, open-source software product[6].
- An ISO metadata profile[7] that uses the OGC Catalog Service for the Web (CSW 2.0.2) as the service protocol.
- Formal collaboration agreements with GEON and the USGS National Digital Catalog to integrate their various catalogs resources. Catalog service specifications have been tested on the ESRI Geoportal tool and implemented with Deegree and Geonetwork.
- A USGIN file repository and Web tool to create metadata, register resources, and upload files. This is deployed as a Drupal v6 application[8]. The application pushes metadata to a Web-accessible directory for harvest into the Geoportal-based catalog when resources are published.
- A collection of Geography Markup Language (GML) simple-feature content models for use by State Geological Surveys to publish data to the National Geothermal Data System (NGDS)[9]. These models are built as Microsoft Excel spreadsheets to facilitate data loading; deployed services use an XML encoding of the content model as an interchange format.

[5] http://www.columbia.edu/~rb2568/rdlm/Faundeen_USGS_RDLM2011.pdf
[6] http://catalog.usgin.org/geoportal
[7] http://lab.usgin.org/usgin-iso-metadata-v1-1
[8] http://repository.usgin.org/
[9] http://www.geothermaldata.org/

- Open Geospatial Consortium (OGC)[10] Web Feature Services (WFS), deployed for the NGDS by the AZGS, the Illinois Geological Survey, and Kentucky Geological Survey. GML simple feature models are used because ArcGIS and various open-source software packages can consume WFS serving simple features. Thus, a widely deployed client platform is already available for service utilization.
- A Uniform Resource Identifier (URI) redirection application to enable linked data architecture by using http URIs[11]. The USGIN URI Redirection Engine is considered the backbone for USGIN's evolving linked data system. URI's directed at the URL http://resources.usgin.org/uri-gin/ can be rewritten using rules based on regular expressions. This application is used to resolve feature identifiers to documents that describe the feature.

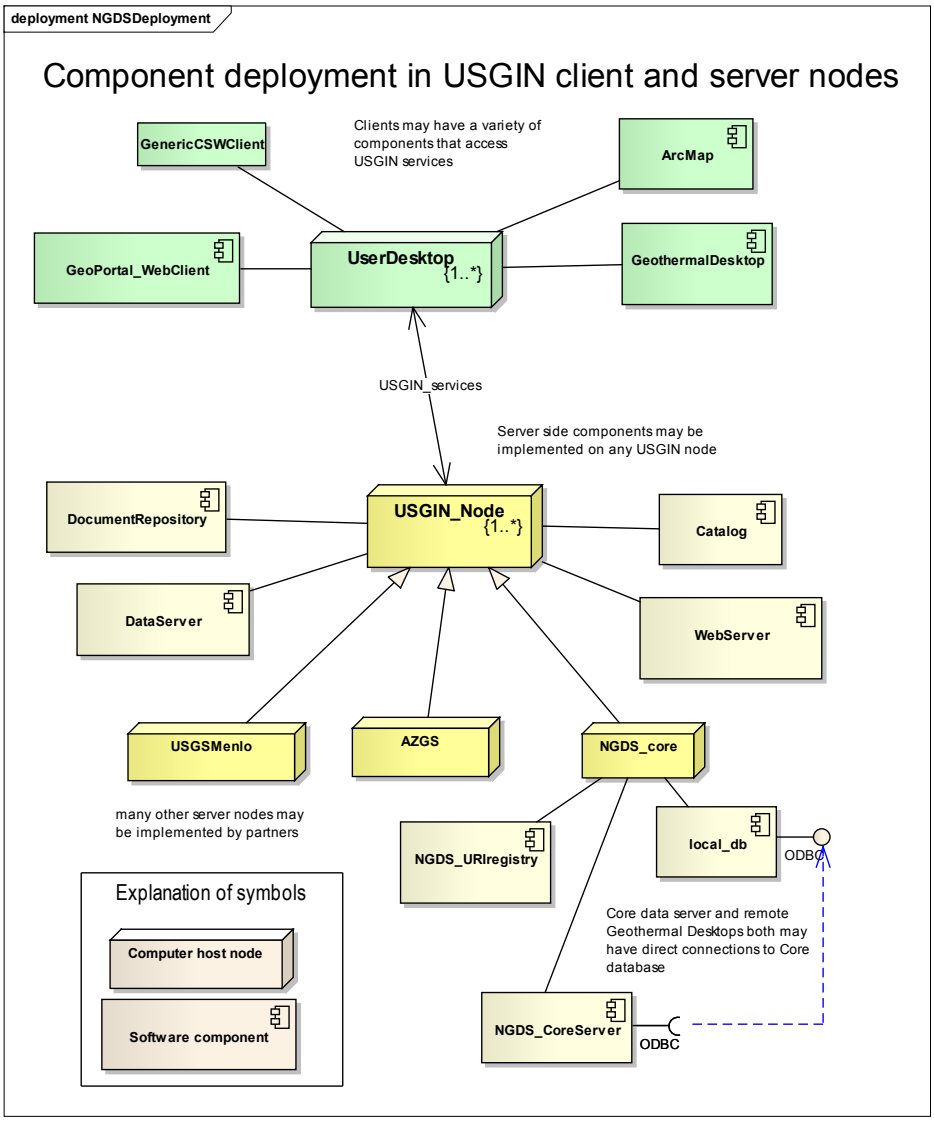

Figure 1. Example server nodes and components in USGIN network.

[10] http://www.opengeospatial.org/
[11] http://resources.usgin.org/uri-gin/uri-description/

Additionally, current USGIN team members from the AzGS have been participating in various standards projects. Recent activity includes:

- Contributions to the international GeoSciML markup language [12]
- Participation in the Energistics Metadata Work Group[13] that is contributing to revision of the ISO 19115 and 19119 metadata standards [14]
- Development of vocabularies for rock type and numerous geologic feature properties for use in populating GeoSciML instance documents [15]
- AASG-USGS development of a relational data base format for geologic map data [16]

USGIN community activities are documented on a Web site (http://lab.usgin.org) that includes reference information and links for services, specifications, and applications that are in use or are being considered for network use. Development activities are documented in blogs. USGIN invites anyone interested in contributing to the effort to register and contribute content to the Web site.

Achieving Sustainability: A Five-Year Vision

The USGIN Working Group envisions further development of tools and capabilities, in addition to extending the community of practice that currently involves 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[17].

Within this framework, the USGIN community can establish an identity for geological survey informatics practitioners, can assist in prioritizing technical development that is specific to the geological survey community, and can leverage development taking place in the larger community. Policies, protocols, and procedures for developing, reviewing, and distributing specifications can be adopted from established practices 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 with stakeholders by supporting participation in CDI, OGC, and ESIP meetings (immediate)

[12] http://geosciml.org
[13] http://www.energistics.org/metadata-work-group
[14] http://www.energistics.org/metadata-work-group
[15] https://www.seegrid.csiro.au/wiki/CGIModel/ConceptDefinitionsTG
[16] NCGMP09, http://ngmdb.usgs.gov/Info/standards/NCGMP09/
[17] The GSA and AGU national meetings are such large and diverse assemblies that it is difficult to achieve the small group personal interaction necessary to foster community development. AASG annual meetings provide an excellent forum for education and outreach to the state geologists, and obtaining feedback on priorities.

    - 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
    - Foster online collaboration in groups with particular objectives
- Deliver products
    - 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. It should be small enough to be agile, comprising 7 members, with 2 or 3 representatives from the states, USGS, and large geosciences communities.

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 should remain an ongoing priority for GIN. The Working Group proposes that growth of the community should be reinforced by collaborating on two test bed activities (suggestions outlined below), and engaging with more experienced communities at the Open Geospatial Consortium and ESIP. Depending on how the priorities are established, the test bed efforts will test and develop best 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 see tangible progress. Success in the OGC community is testament that test bed projects have fostered communication, alignment of activities, and exchange of expertise and capabilities in that community.

Historically, a small number of geoinformatics practitioners have been spread across a wide variety of discipline-focused organizations, effectively diluting the knowledge base. The number of practitioners is growing, and is reaching a point where a critical mass of individuals can be brought together to meld into a community. Identifying one or two annual meetings as gathering points for this community can foster this newly forming coalescence. This Working Group recommends the USGS CDI meetings, OGC quarterly meetings, and ESIP meetings as obvious candidates for consideration.

Implementation

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. Additionally, as a member of a community activity, the role of USGIN as a leader in the community must be organic and emergent. However, there are some implementation activities the Working Group identifies as essential for USGIN, listed as recommendations below.

Establish a Long-Term Governance Model

If USGIN is to represent the interests of the geological surveys in the larger geoinformatics community, a strong governance model is necessary to define USGIN as an entity. Such a model should contain information about the USGIN member base, designated speakers for the organization, the source of their authority, and how decisions for priorities and recommendations are made. A coordination group can be the starting point for a formal organization; one of its first major activities, however, will need to be a formal charter that defines the governance model - membership, leadership selection and terms, staffing, and decision making processes for the organization.

Develop a Business Model

Depending on decisions about the activities and responsibilities of a USGIN organization established in a governance charter, a defined level of financial support will be necessary and should be detailed in a business plan. Based on a cursory survey of existing organizations with intent similar to USGIN conducted by the Working Group, we recommend that a non-profit foundation model may be the most appropriate. This would define USGIN as a legal entity that could enter into contracts, receive funding, pay salary, and make grants of funding. A formal legal entity could support a small number of full time employees who look after day-to-day operations, planning, and logistics for community events. An important issue to be determined by the governance authority is whether there should be a stand-alone organization, or if it is more effective to form a group within the framework of ESIP, OGC, or some other existing organization. The implications for a separate business entity or affiliation with an existing organization need to be explored in the context of financial and intellectual property considerations that may result from the government agency nature of most geological surveys.

The business structure to support a community geoinformatics development process should promote grass root nucleation of interest groups that are the core of development efforts. Selection of groups to be nurtured by available funding should be based on community assessment of priorities and user demand. Resource allocation by government agencies in support of these activities must be tempered by programmatic priorities. Loosely knit organizations such as ESIP and the USGS-CDI provide organizational models, and the USGIN community will progress most efficiently by utilizing these existing groups to gather input and effort from the communities they attract.

Because many geological surveys have data archive and dissemination functions as part of their portfolio, some support for USGIN might be built into their operating expenses and overhead as users come to expect access to data resources through standardized network interfaces. Sharing of resources and reuse of components through USGIN will reduce the cost of these activities. Wide adoption of similar software, protocols and practices increases the number of stakeholders with an interest in supporting USGIN. Services and data streams deployed by the geological surveys must have sufficient value to the user community to support either a pay-for-use model, or to motivate continued public funding if the system it to be viable in the long term. USGIN can be seen as a metaphor—it is building an online version of the bricks and mortar libraries that historically have been the anchor of knowledge preservation and access for the geosciences.

Explore Testbed Opportunities

A testbed is a platform for technological experimentation in a framework that allows for rigorous, transparent, and replicable testing of new technology (http://en.wikipedia.org/wiki/Testbed; Percivall, 2006). The Working Group

agreed that one or two testbed activities should be proposed to engage geological surveys in development, adoption, and deployment of standard interchange protocols and document formats. This series of activities represents a good starting point to nucleate USGIN. Several test bed activities were suggested:

1. OneGeology – US testbed. The USGS and AASG collaborate to implement WMS and WFS services according to the OneGeology profile (GeoSciML portrayal for WMS, GeoSciML WFS for supporting WFS). USGS handles national maps (surficial deposits, glacial deposits, bedrock geology of US, etc.) and State Geological Surveys deploy services for state-scale geologic maps. The community must agree on an integrated portrayal scheme for services to attain some interoperability between map services.
2. Observation service test bed— the state geological surveys and USGS deploy observation services for geochronologic data, geochemical data, gravity stations, water quality data, or some similar commonly available site or sample-based data. The community would need to agree on service profiles and interchange formats, as well as procedures to avoid data duplication. This testbed activity could build on and enhance both the National Geothermal Data System and EarthChem.
3. Integrated catalog capability – the achievement of interoperable metadata – demonstrate an ability to harvest between catalog nodes hosted by state surveys and USGS, execute search against multiple nodes, edit tools that work against multiple nodes (with user authentication and access control).

Each of these proposed testbeds would utilize existing service protocols and interchange formats to the maximum extent possible, as well as off-the-shelf-open source software or widely deployed commercial software (e.g. ArcGIS) for service deployment. Service profiles would also consider existing clients or client development frameworks (OpenLayers, Flash, etc.) in their design.

The protocols and interchange formats used for testbed activities will be developed with consideration of existing technology and standards (following the founding principles of USGIN). The test bed activities in the geological survey community will define a scope and provide a foundation to promote the use of specifications developed in our community by the larger geoinformatics community. Adoption of some of these specifications as ‘standards’ by USGS and AASG for use by those organizations will lend authority and motivate wider adoption.

The transition from use case and test bed activities to production deployments and agreement on ‘standard’ specifications for data discovery and access must be propelled by active interest from the user communities who have a stake in the outcome. Part of the testbed planning should include identifying and contacting target communities, and exploring possibilities that they might contribute to the costs of system development and maintenance.

Develop Marketing Strategy

Development of a USGIN education and outreach strategy to inform and engage data providers of GIN direction, infrastructure, and activities is recommended for USGIN’s development. Stakeholders will need to know how to realign their existing approaches to data delivery to best interact with USGIN, what new resources are available, how to use them, and any new opportunities for information utilization. A marketing plan should also include monitoring of network resource usage, and collection of input from the user community to identify what is working well, what needs fixing, what is not being used, and what new capabilities would be useful.

Development of educational material, giving talks and running workshops requires significant time and effort, and is generally difficult to support with project-based funding alone. This is one area in which personnel with funding specifically assigned to these tasks is vital. A designated outreach person would interact with system developers to understand the function of the system well enough to explain it to stakeholders. In turn, they would interact with the user community not only to offer education about USGIN network resources, but also to get valuable feedback about USGIN products.

Costs

In order to assess costs, the Working Group assembled an inventory of what components are likely to be necessary to sustain USGIN.

- Servers and software: (This list assumes that network infrastructure (internet connection, switches, firewalls, DNS, etc.) are in place)
    - Catalog and repository server: Linux/Tomcat/PostgresQL/ GeoNetwork or Geoportal (could run on Windows stack as well). Only one is essential, any number is possible.
    - To support a catalog system: a server for registries for identifiers and vocabularies, in addition to repositories for system specification documents and resources, like XML schemas that must be web accessible to support service operations.
    - Data server: Windows/ArcGIS Server (dbms optional if shapefiles are used as data source), or Linux/Tomcat/PostGresQL/Geoserver or Mapserver. One server with modest capabilities could serve perhaps 300 Mb of data, depending on load.
- Personnel:
    - Technical IT personnel: Need capabilities to deploy server software, load data, configure services, debug http traffic if there are problems. Data preparation (if standard interchange formats are being used) requires understanding of ETL using SQL queries and other techniques. Some understanding of XML and XML schema is required occasionally; GeoServer requires mapping from XML to database fields in an XML configuration. Someone with understanding of metadata content models and encoding is likely to be essential to get a catalog system working well.
    - Outreach and marketing: personnel dedicated to the production and maintenance of documentation and educational materials, as well as face-to-face and online training programs.
- Maintenance:
    - Individuals dedicated to system management, including arrangement of meetings, maintenance of hardware, user help lines, and network operation.
- Software development and testing:
    - Components for service deployment may be necessary if off the shelf solutions do not meet all requirements; applications for service and conformance testing, development of metadata, and network monitoring

Transition from current stovepipe data discovery and delivery systems to a loosely coupled, service based architecture can probably be done by large organizations like the USGS with the personnel and hardware they currently utilize, following whatever IT/hardware refresh cycles they currently use. Additions to hardware capacity in the form of servers, bandwidth, and online storage will be necessary to bring new data online, but this would be true no matter what approach to data delivery is adopted. The most significant additional investment will likely be in education of personnel who develop and deploy data and metadata services, in addition to human effort for data integration, documentation, and migration of existing data to new formats and delivery protocols. These investments will provide long-term return in staff capabilities if, as anticipated, these become widely accepted standard operating procedures. An additional return is the expected increase in utilization of the information resources and greater visibility for the agency providing data.

Summary

The future of USGIN depends on taking measured steps to build this initiative and see it succeed. A summary of recommended actions from this document include:

- Develop and deploy a marketing strategy to increase engagement of the State Surveys and stakeholders and cement a strong user base supporting USGIN.
- Establish a governance structure led by a technical working team with representatives from USGS, State Geological Surveys, and other geosciences-focused communities
- Decide if USGIN should be established as an independent non-profit 501(c)(3) foundation or a subdivision of an existing organization like OGC or ESIP
- Conduct a comparative analysis of preferred sustainability models and select one to implement that will provide a conduit for financial resources to support infrastructure, a 2-3 person staff, and servers for system infrastructure
- Inventory the State and Federal Geological Survey information architecture landscape to identify and prioritize use cases, and to identify capabilities and expertise to develop prototypes and testbed opportunities.
- Work to promote consistent best practices
- Engage with ESIP, NGC, DataOne[18] and other geosciences communities to promote interests of geological surveys in national cyber-infrastructure development.
- Select test bed opportunities to engage the geological survey community in the context of CDI, ESIP, NGC, OGC, DataOne and evolving NSF EarthCube activities.

References

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

Shapiro, Carl, 2000, Setting Compatibility Standards: Cooperation or Collusion?, *in* Dreyfuss, M. C., Zimmerman, D. L., and First, Henry, Expanding the boundaries of Intellectual Property: Oxford, UK, Oxford University Press, p. 81-102.

DIN, 2000, Economic benefits of standardization, Summary of results: DIN German Institute for Standardization, e.V., 39 pages, accessed at http://www.din.de/sixcms_upload/media/2896/economic_benefits_standardization.pdf 2011-10-03.

Blind, Knut, Jungmittag, Andre, and Mangelsdorf, Axel, 2011, The economic benefits of standardization, DIN German Institute for Standardization, e.V., 24 p., accessed at http://www.din.de/sixcms_upload/media/2896/GNN_2011_engl_FINAL.111681.pdf 2011-10-03.

Web sites:

[18] https://dataone.org

Overall Benefits of Financial Data Standardization:
http://www.gs1us.org/sectors/financial_services/overall_benefits_of_financial_data_standardization (accessed 2011-10-03)

Paper Submitted by:
Stephen Richard, Arizona Geological Survey, USGIN
Viv Hutchison, US Geological Survey