A National Geoinformatics Community (NGC)

J. Douglas Walker (University of Kansas)
Linda Gundersen (USGS)
Lee Allison (Arizona Geological Survey)

Hassan Babaie (Georgia State University)
Doug Fils (Ocean Leadership)
Steve Richard (Arizona Geological Survey)
Ramon Arrowsmith (Arizona State University)
Cinzia Cervato (Iowa State University)
Steve Whitmeyer (James Madison University)

Summary
The motivation for further developing and involving the National Geoinformatics Community (NGC) concept in EarthCube is.:

I can't integrate what I can't find
I can't use something I don't understand
I don't want to use something I don't trust
I can't use something that isn't there any more

The scientific, societal, and educational problems the geoscientific community is being asked to address are increasingly complex and thus require the application of multiple datasets and advanced computational capabilities for integrated data analysis. However, we cannot integrate data we cannot find, understand/evaluate, or trust. It is also critical that data, tools, and services are persistent and do not disappear at the end of projects or with time. We cannot jointly analyze multiple datasets without the required tools. The National Geoinformatics Community effort is aimed at addressing these critical issues and developing and offering community-based solutions.

Framing Statement
For the past 10 years, Geoinformatics has evolved into a critical and growing activity in the Earth Sciences. It has potential transformative impacts in all areas of research, outreach, and education. Aspects of Geoinformatics include not only data discovery, access, and delivery, but also data, model, and system interoperability, data management, and archiving of past, current, and future research and education studies. This is emphasized by the new and emerging data reporting requirements being enforced by the US National Science Foundation and other funding agencies. It also appears to be the motivation behind the EarthCube effort.

Activities on data and service interoperability amongst the federal and state geological surveys are already underway with the Geosciences Information Network (GIN) project (http://usgin.org/). This activity provides a common data catalog and specifications, as well as tools and web services for discovery, integration, and analysis within a virtual network. Although GIN is open to all data providers and users, no such networking effort currently exists for the

academic communities, whether funded by the NSF or others. The academic community has struggled with how to self organize across the major earth science domains, though there are some significant successes that are providing portals and data management for certain sub-domains.

National Geoinformatics Community - Evaluating and addressing community needs and opportunities

The following sections present aspects of community needs as well as opportunities. The strength at present, of the National Geoinformatics Community, is in understanding general geoscience with special and unique emphasis in semantics and E&O (education and outreach) activities. In addition, the community is especially aware of needs of manpower with appropriate expertise.

Impediments to progress

The impediments to progress in the development of a Geoinformatics infrastructure:

- Absence of technical capabilities
- Lack of funding
- Lack of trained people to do work
- Lack of involvement/buy in
- Need for facilitators to bring people, managers, and tools together
- Lack of opportunity for innovation--thinking in a rut
- Domain provinciality
- Diversity of interests in the community

Cross-cutting issues

The most concise summary of the motivation for the NGC concept are the issues of integration, usability, trust, and persistence, listed above, and elaborated below.

Finding and combining data, tools, etc.

Problem: I can't integrate what I can't find.

Data discovery is a critical activity. The most important resources required to make Geoinformatics useful are 1. Find a simple way for data or application producers to make their products known, and 2. Enable data and application consumers to locate, evaluate, and access the resources they need, and 3. Integrate the results in order answer the science question(s) at hand. A National Geoinformatics Community should first and foremost foster the development and adoption of conventions and specifications to make a domain-wide distributed catalog service or system a reality, along with the training necessary for its use, and professional development opportunities and recognition necessary to motivate participation by scientists in the growth and maintenance of this resource. In addition, the NGC could help foster development of tools to mediate semantics and content (integration models). Without these the data cannot be integrated.

The community organization should support planning and implementation of specific interoperability experiments designed to test and demonstrate key system components and

concepts. Documentation for such experiments would provide the basis for specifications of community interoperability practices. The organization can sponsor topical sessions or workshops at professional meetings. Basing such sessions on interoperability experiments suggested above, or on specifications that are currently in development, would provide a way to bring in new input, and to communicate the results of such experiments.

Understanding what we have found
Problem: I can't use something I don't understand.

Just as the data accessed can be heterogeneous, the user base for Geoinformatics data is likewise varied. The user level can vary from someone interested in viewing a certain data set to an expert domain scientist wanting to model across several sets of data. At any point along the spectrum, the user needs to easily understand what information is present. A critical issue is to make data and tools easily understood by different audiences. It is clear, however, that establishing explanations and metadata that accommodate different levels of users is a time-consuming and difficult process.

This is an area where a the NGC could greatly facilitate use by the larger Geoscience community. Because the group spans the spectrum from data managers to E&O practitioners to domain sciences, a priority effort could be the more complete documentation and explanation of various datasets, data analysis tools, and end-user products and models. This is especially important given the drive for interoperability described above.

Determining the accuracy and quality of data, tools, etc.
Problem: I don't want to use something I don't trust

This is perhaps the most difficult aspect to tackle. To ensure that quality assurance and quality control is documented and maintained requires understanding the best practices at all levels. This includes the need for specifications, procedures, and standards not only for the data, but also software engineering, analysis tools, and data management for users to assess and evaluate the provenance, quality and accuracy of data and tools. These must be practices that the community endorses and are clearly informed by ongoing research. Again, this is an area where the breadth of the NGC community could be invaluable in the evaluation of QA/QC procedures and make recommendation to users, other communities, and publishers.

A somewhat related issue is that the desired trust in data requires an excellently trained, technically savvy, and science aware Geoinformatics workforce. Such a group is critical to all of the aspects mentioned above. Training and/or certification of such workers could be a core activity of NGC.

Keeping data and tools available for future use.
Problem: I can't use something that isn't there any more

Longevity of data is critical in several respects. One aspect is ensuring that information and analysis tools are available in the long term even though a project may end. This is a frequent occurrence and results in the common problem of missing or defunct websites. This is the typical

"Link not found" problem. Many of these missing links were pointing to valuable datasets or tools that have possibly been funded by federal agencies and could provide an important resource.

Data longevity and archiving are complex issues. As noted above, a minimum requirement for any information, especially that being archived, is that the proper metadata and QA/QC (quality assurance/quality control) procedures have been followed and are documented. This is an especially critical issue and attests to the need for qualified Geoinformatics practitioners. The scope of data archiving is a component of a comprehensive NGC. The NGC is not the only group that is concerned with this topic but it is one where the group can have impact. There are internationally recognized standards and protocols for the preservation of data and software resources. The NGC can work with those groups and organizations with primary responsibilities, sufficient resources, and rigorous policies to keep data safe in the long term.

Summary

We need to make data and tools accessible to a wide range of clients via registries and/or catalogs that are easily understood by a variety of clients. In building products, services, and tools, the broad view of data and analysis requires investigators to step outside their areas of expertise. Thus, it is crucial to have access to documents that make it possible to understand the best practices of other domain experts.

Connection with EarthCube

The National Geoinformatics Community is potentially a critical component necessary for the success of the EarthCube concept. The community is made of individuals and groups that deal with the research and especially educational aspects of Geoinformatics on a constant basis. The idea of creating an overarching Cyberinfrastructure for the Geosciences and beyond meshes well with the goals of the community. In turn, the National Geoinformatics Community can offer vast experience and expertise in dealing with a host of diverse issues involved with research, education, and outreach.