USGS
science for a changing world

Geoinformatics 2007—Data to Knowledge

Proceedings

May 17-18
San Diego, California

Scientific Investigations Report 2007-5199

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia: 2007

For sale by U.S. Geological Survey, Information Services
Box 25286, Denver Federal Center
Denver, CO 80225

For more information about the USGS and its products:
Telephone: 1-888-ASK-USGS
World Wide Web: http://www.usgs.gov

Suggested citation:
Brady, S.R., Sinha, A.K., and Gundersen, L.C., editors, 2007, Geoinformatics 2007—Data to Knowledge, Proceedings: U.S. Geological Survey Scientific Investigations Report 2007-5199, 104 p.

Manuscript approved for publication September 10, 2007.
Prepared by Reston Publishing Service Center.
Editing by Marilyn A. Billone.
Photocomposition and design by Cathy Y. Knutson.
Cover design by Jenifer Bracewell.
For more information concerning this report, please contact
Shailaja R. Brady, U.S. Geological Survey, 911 National Center, Reston, VA 20192, srbrady@usgs.gov.

Preface

Conference Summary:
The Geoinformatics 2007 Conference presented the effort by the geosciences and information technology communities to respond to the growing need for utilizing multidisciplinary geoscience datasets and tools to understand the complex dynamics of earth and planetary systems. The vision of the community of a fully integrated geosciences information network that provides free access to high-quality earth and planetary science-related data, tools, and services was highlighted at the meeting. The fusion of spatial and nonspatial data from earth and planetary sciences needed to develop a coherent scientific understanding of the Earth's four-dimensional evolution and architecture is emphasized in the proceedings volume. Information technology (IT) research related to knowledge-based mediation and information integration techniques for four-dimensional data models; visualization of multiscale, four-dimensional information spaces; data-level interoperability; metadata modeling and interchange; and metadata-based access to data and services were leading topics of discussion and presentations. More significantly, issues of data management, moderated by A. Krishna Sinha (Virginia Polytechnic Institute and State University) and discussed by panelists Chris Greer (National Science Foundation), Linda Gundersen (U.S. Geological Survey), Ian Jackson (British Geological Survey), Lesley Wyborn (Geoscience Australia), Peter Fox (National Center for Atmospheric Research), Deborah McGuinness (Stanford University), and Shinji Takarada (Japanese Geological Survey), leading to solutions of global challenges such as natural resources and hazards, provided the focus for the conference and will continue to drive geoinformatics research in the foreseeable future.

Conference Sponsorship:
Geoinformatics Division of the Geological Society of America, U.S. Geological Survey, British Geological Survey, National Science Foundation, American Geophysical Union, California Institute for Telecommunications and Technology, and San Diego Supercomputer Center

Organizing Committee:
A. Krishna Sinha (Chair), Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.; Linda Gundersen, U.S. Geological Survey, Reston, Va.; Peter Fox, National Center for Atmospheric Research, Boulder, Colo.; Dogan Seber, San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

Sally R. Brady (Co-Chair), U.S. Geological Survey, Reston, Va.; A. Krishna Sinha (Co-Chair), Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.; Linda Gundersen, U.S. Geological Survey, Reston, Va.; Bruce Johnson, U.S. Geological Survey, Reston, Va.; Robert Raskin, Jet Propulsion Laboratory, Pasadena, Calif.; Ilkay Altintas, San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.; Rahul Ramachandran, Information Technology and Systems Center, University of Alabama, Huntsville, Ala.; Peter Cornillon, Graduate School of Oceanography, University of Rhode Island, Narraganset, R.I.; and Hassan Babaie, Department of Geology, Georgia State University, Atlanta, Ga.

Financial support was provided by the U.S. Geological Survey, National Science Foundation Division of Earth Science (EAR), San Diego Supercomputer Center, California Institute for Telecommunications and Information Technology, and Virginia Polytechnic Institute and State University Foundation.

Conference support was provided by the Geological Society of America (Erin Pitner, Nancy Carlson), California Institute for Telecommunications and Information Technology (Jennifer Zimmerman); and Department of Geosciences, Virginia Polytechnic Institute and State University (Mary McMurray).

Contents

Geoinformatics 2007—Data to Knowledge

Edited by Shailaja R. Brady, A. Krishna Sinha, and Linda C. Gundersen

Envisioning a Geoinformatics Infrastructure for the Earth Sciences: Technological and Cultural Challenges

By Linda C. Gundersen[1]

[1]U.S. Geological Survey, Reston, Va.

Since its inception in 2000, geoinformatics has been envisioned as serving all the earth sciences and enabling the understanding of complex systems. For the past 6 years the earth and computer science community has struggled with clearly defining the path to that vision, obtaining sufficient resources, maintaining interest across the broad spectrum of scientists that geoinformatics intends to serve, and achieving coordination and interaction among scientists nationally and internationally. There is hope, however, that we are approaching a true turning point in geoinformatics that has as its hallmark a general convergence on a suite of technologies and concepts that could begin the process of growing an integrated infrastructure.

Several major technology experiments and community database efforts are beginning to bear fruit, and numerous workshops and town halls have garnered unprecedented interest and agreement in the community. The development of Geoscience Markup Language (GeoSciML) by the Interoperability Working Group of the Commission for the Management and Application of Geoscience Information (CGI), a commission of the International Union of Geological Sciences (IUGS), is one such breakthrough. GeoSciML is a geoscience-specific XML-based GML (Geography Markup Language) application that supports interchange of geoscience information. In recent testing, it demonstrated that it could be used to provide the interoperability needed to bring disparate geologic maps together. A number of efforts, such as the EarthChem Portal, GEON (The Geosciences Network), and the North American Geologic Map Data Model, showed that community databases, standards, and ontologies can be created and implemented cooperatively. The well-attended Geoinformatics Town Hall at the American Geophysical Union Meeting in December 2006 was among the first to articulate a national and potentially international collaboration across the community (Fox and others, 2006).

In early 2007, several workshops occurred among geological surveys nationally and internationally, and a National Science Foundation (NSF) workshop among geoinformatics practitioners was held. Collectively, these workshops have resulted in a more definitive vision and convergence around a few key concepts and technologies. Common elements defined in these workshops describe an infrastructure that is distributed, interoperable, uses open-source standards and common protocols, respects and acknowledges data ownership, enables science, fosters communication and collaboration, shares resources and expertise, and develops community databases, new Web services, and clients that are sustained over time.

Each of these concepts and technological choices agreed upon in these workshops offers a myriad of challenges that will now have to be analyzed with agreed-upon resolutions. To do this will take a sustained effort, supported by government, industry, and academia. It will involve carefully and thoughtfully implementing a real "community of practice" (Wenger, 1998) in geoinformatics.

In a recent NSF workshop report (Edwards and others, 2007), the authors state that "infrastructure is fixed in modular increments, not all at once or globally. Because infrastructure is big, layered, and complex, and because it means different things locally, it is never changed from above. Changes take time and negotiation, and adjustment with other aspects of the systems involved." The authors also describe the evolution of systems to true infrastructure. They define an evolving spectrum that starts with a single system that is centrally organized and controlled, to networks that link systems with control partially or wholly distributed among the system nodes, to internetworks (or Webs) that are networks of networks based on coordination rather than control. These internetworks represent true infrastructure and require communities of practice to coordinate and implement. Currently, many systems are being built and ad hoc gateways constructed to link systems without much coordination.

Communities of practice are not limited to a community with a common interest or "domain," but can encompass practitioners who share experiences and learn from each other across domains. They develop a shared repertoire of resources: experiences, stories, tools, vocabularies, and ways of address-

ing recurring problems. Coordination, as well as standards of practice and reference materials, grow out of this experience. The critical benefits of communities of practice include creating and sustaining knowledge, leveraging of resources, and rapid learning and innovation. Not only will communities of practice help grow the cyberinfrastructure needed to understand and analyze complex systems, but they will promote the other essential ingredient—communication among scientists. The primary barrier to integrated science is the difficulty in translating the vocabulary of one scientific field to another. A geomorphologist who understands the control geology has on the health of vegetation in a riparian habitat has significant difficulty translating this information to the ecologist specializing in bird mortality. Simply making data interoperable will not create an infrastructure for the earth sciences to conduct analyses of complex systems. It will take a sustained, long-term, focused effort across the earth sciences and computer sciences to build the tools (databases, gateways, Web services, and applications), and to change the culture (common vocabularies, data sharing, and open-source standards).

References Cited

Edwards, P.N., Jackson, S.J., Bowker, G.C., and Knobel, C.P., 2007, Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, NSF Grant 0630263, 55 p.

Fox, P., Gundersen, L., Lehnert, K., McGuinness, D., Sinha, K., and Snyder, W., 2006, Toward broad community collaboration in geoinformatics: Eos, v. 87, no. 46, p. 513.

Wenger, E., 1998, Communities of practice: learning, meaning, and identity: Cambridge University Press, 336 p.

OneGeology—The Birth of a Global Geoscience Spatial Data Infrastructure, or Just Another Noble Aspiration?

By Ian Jackson[1]

[1]Information Directorate, British Geological Survey, Keyworth, Nottingham, United Kingdom

Contrary to the thinking of many in positions of power and influence in the political and environmental domain, the environment is not just restricted to that bit of our world above the ground; the subsurface is pretty important too. Think earthquakes and landslides; minerals and mining; and groundwater and pollution. Like most things environmental, few of these issues respect national frontiers and if, as seems wise, we want to assess and address global environmental problems at a global scale, then we need access to global environmental data. In the topographic domain that often tends to be dominant in the global Internet (GI), and in one or two of the environmental domains above ground, (for example, meteorology) we have, comparatively speaking at least, extensive and relatively consistent data. This is not so in the geoscience domain. Unfortunately, for geoscientists (and the rest of society), digital geoscience spatial data across the globe, even at small scales, are either unavailable, out of date, of variable quality, or inconsistent.

It is true that in several countries, extensive geophysical data exists and there is basic digital geological map data of reasonable quality and coverage. But in many more countries, it is incomplete or is not present at all—even within Europe (which poses an interesting resource and priority challenge for a post-INSPIRE European Union and its member states). Even where there is good national data, the chances of it being interoperable, let alone harmonizable, are small. This is not news to those within the geoscience community; many of us are well aware that we need to accelerate the development and promulgation of simple, basic, and essential digital geological map standards and specifications to improve the interoperability and sharing of data. In Europe, North America, and Australasia, people are working hard on trying to move structural interoperability forward and some of the readers of this abstract will be aware of the development of a high-level geoscience data model and the interchange format GeoSciML. Semantic interoperability, needed to deliver some form of scientific homogeneity (in other words, harmonized geoscience data) is another story and will take some very serious concentration and effort by the global geoscience community in the area of terminology and classification. While geoscientists may thrive on independence and diversity, digital systems generally do not.

Despite the good work and progress with GeoSciML, moving the development and promulgation of standards forward is a slow and unglamorous process, and, outside the domain of informatics, it is regarded as a pretty esoteric business whose purpose and value is rarely fully understood. Developing a standard tends to be a very abstract occupation, somewhat detached from reality; at least that is what many geoscientists think.

Enter the United Nations International Year of Planet Earth (IYPE2008) and the germ of an idea to create a 1:1,000,000-scale global geological map dataset—a concept now known as OneGeology and the subject of this abstract. At the beginning of 2006, the 1:1,000,000 idea was presented somewhat immaturely and at short notice to the General Assembly of the Commission for the Geological Map of the World (CGMW) in Paris. What if we used IYPE2008 as a stimulus to begin the creation of a digital geological map of the planet at 1:1,000,000 scale? Could we design and initiate a project that uniquely mobilizes geological surveys around the world as part of an ongoing IYPE2008 contribution, to act as the drivers and sustainable data providers of this global dataset? Further, could we synergistically use this geoscien-

tist-friendly vehicle of creating a tangible geological map to accelerate progress of an emerging global geoscience data model and interchange standard? Finally, could we use the project to transfer know-how to developing countries and reduce the length and expense of their learning curve, while at the same time producing geoscience maps and data that could attract interest and investment? These aspirations, plus the chance to generate a global digital geological dataset to assist in the understanding of global environmental problems and the opportunity to raise the profile of geoscience as part of IYPE2008, seemed more than enough reasons to take the proposal to the next stage.

Since that CGMW meeting in February 2006, the concept has been disseminated to organizations and individuals around the globe and has matured considerably. In addition to the support of CGMW, the project has attracted the support of the International Union of Geological Sciences (IUGS), United Nations Education, Scientific and Cultural Organization (UNESCO), and the International Steering Committee for Global Mapping (ISCGM), not to mention the IYPE2008 Management Team. But crucial to the success of the concept is getting geological surveys to sign up and commit their data and resources to the cause. At the time of this writing, more than 50 surveys from around the world have agreed to participate. My own organization, the British Geological Survey, has decided to make the project its prime contribution to IYPE2008, as well as offer to continue to play a full role in the project's coordination, and agreed to support the initial kick-off meeting, which took place in Brighton, United Kingdom, in March 2007. Several countries, including the United Kingdom, plan to pilot the methodology and make data available during 2007-08.

The proposed methodology differs from the usual method of making available geological data for a continent or the globe—usually one editor or editorial unit compiling information from a variety of sources and in recent years using a GIS to produce the cartographic result. The OneGeology proposition is a completely modern paradigm. It is planned as a distributed model, and at the most technically sophisticated end, will see a Web feature service, a dynamic set of geological map data served mostly on a national basis by individual geological surveys and other bodies (for example, the polar and marine surveys and research bodies) to a Web portal or portals, and as such will be frequently updated and improved by the data providers and reflect the most up-to-date data they possess. To achieve its goals, the project team will have combined state-of-the-art skills in geoscience data modelling and information management with worldwide expertise and experience in geoscience. The project will obviously be closely interlinked with the IUGS Commission Working Group, developing the global geoscience data model and exchange language—GeoSciML.

The key international players attended the kick-off meeting in the United Kingdom. At this meeting, they endorsed and initiated the project and discussed the project plan and the specifications for the geological and information systems. During 2007, the first test datasets are anticipated to become available and we will progressively add data through 2008 so that we can present the first results at the International Geological Congress in Oslo, Sweden, in 2008. What then? We geologists work in geological time, so what we are talking about here will take no time at all. We will continue to add and upgrade the data through time to progressively provide the most complete coverage of the planet at our target scale of 1:1,000,000, and also (more excitingly) add data at a higher and more useful resolution.

Even at the outset there was a realization that the project would not be able to obtain 1:1,000,000-scale geological map (lithology/chronostratigraphy) data everywhere, and, perhaps, in some parts of the world, any data at all. Yes, there will be major faults running along many national boundaries (the semantic interoperability problem) but what a great way to get a long overdue problem tackled! It is here the Google Earth philosophy comes into play—Be pragmatic, make available what you can now, and aspire to improve it in the future; after all, there is no such thing as a geological map that is complete.

There are, undoubtedly, many more bridges to cross, but what was a mere germ at the beginning of 2006 evolved into a proposal and that proposal is now a project with an initial budget and extensive international support. The hope is that as momentum gathers and more become aware of the project and its practical and altruistic benefits, they will come on board, contributing data and expertise. This presentation will present the progress of OneGeology to date and the technical and cultural issues it has encountered.

Building the AuScope Australian Earth Science Grid

By Lesley A.I. Wyborn[1] and Robert M. Woodcock[2]

[1]Information Services and Technology Branch, Geoscience Australia, Canberra, Australia

[2]Division of Exploration and Mining, CSIRO, Kensington, Australia

In 2006, the Australian Government announced a new funding initiative, the National Collaborative Research Infrastructure Strategy (NCRIS). NCRIS aims to provide Australian researchers with access to major research facilities, supporting infrastructure and networks necessary for world-class research. As an element of this strategy, $42.8 million was allocated to the Australian Earth Science Research community to build an integrated national geoscience infrastructure system called AuScope. The NCRIS AuScope funding has several parts. One part of the funding is to develop an advanced national infrastructure for acquiring and analyzing geophysical and geochemical data. Components in this infrastructure include a geotransect facility for imaging large-scale cross sections of the Earth's crust, an ion probe for advanced analysis of Earth samples, and the development of a virtual library of drill core

samples from across Australia. Another part of the AuScope infrastructure is the development of a high precision positioning system, which will build an enhanced national geospatial reference system.

To draw together information from this new national infrastructure and from other existing sources, the AuScope NCRIS funding will also be used to develop a world-leading geoscience geoinformatics network. This network will be called the AuScope Grid and will comprise a Data Grid and a Compute Grid. Combined, they will provide a distributed infrastructure that will enable the dynamic construction of an open-access, four-dimensional model of the Australian continent.

The Earth Science Data Grid is a national geoscience data network that will enable online access to information from new NCRIS geoscience infrastructure and from other sources in academia, industry, and government. The Data Grid will use open geospatial standards to allow real-time access to data, information and knowledge stored in distributed repositories. A key objective for the Data Grid is that it will be built on "end-to-end" science principles (in other words, open-access principles) whereby there will be access to the highly processed information and knowledge, as well as to the original raw data and the processing programs used to generate the results.

The goal of the Compute Grid is to facilitate quantitative geoscience analysis by providing an infrastructure and tools for advanced data mining and online computational modeling and simulation. Computationally demanding geoscience programs will be made available as Web services, and distributed across computing and storage resources in a manner that requires limited knowledge of the physical infrastructure.

The key to linking components and resources on the AuScope Compute Grid with the Data Grid will be service-based access to the geoscience information holdings using a common service interface and information models, including GeoSciML. Further development, maturing, and formalization of GeoSciML are essential to the success of the AuScope Grid. GeoSciML is being developed through the Interoperability Working Group of the Commission for the Management and Application of Geoscience Information (CGI), a commission of the International Union of Geological Sciences (IUGS). Similar information models are also required for geophysical and geochemical data. As with GeoSciML, it is desirable that these additional information models be developed as open standards and under the auspices of the relevant international scientific organizations.

There are no obvious technological barriers to what has been proposed in building the AuScope Grid. Nearly all the required technical elements have been tested in recent years in a series of interoperability testbed projects. Potential limitations are now seen as "social engineering" issues. One social limitation to developing the AuScope Grid to its full potential is the requirement to bring the relevant communities together at an international level, and to work collaboratively to develop open standards for information exchange of all of the required scientific content. Developing and ratifying these standards at an international level is only the first step; they also have to be widely accepted and adopted by the community.

More fundamentally, the proposed grid requires a transition to fully distributed systems whereby all components (tools, applications, compute resources, and data) are available online as globally distributed resources. This paradigm shift contrasts the current culture of "monolithic silos" whereby all the required components for a data-mining exercise or for a modeling and simulation experiment are hosted by a single organization. In some areas, this social transition is proving the most difficult challenge of all as distributed systems are sometimes perceived as reducing control of assets and resources. Distributed systems also rely heavily on mutual trust and open collaboration, and require that once resources are exposed as Web services that these are maintained and made available 24/7.

A Very Practical Geoinformatics Project: The Reality of Delivering a Harmonized Pan-European Spatial Geoscience Database

By Kristine E. Ch. Asch[1]

[1]Federal Institute for Geosciences and Natural Resources, Hannover, Germany

This paper will introduce and critically consider a project whose objective was to produce a harmonized geological spatial database for the continent of Europe. As with many informatics projects in other scientific domains, the technical challenges, while substantial, were significantly less than the organizational and cultural challenges. Solutions to all challenges were necessarily pragmatic and the lessons learned may prove relevant to other geoinformatics initiatives.

The IGME5000 (The 1:5,000,000 International Geological Map of Europe and Adjacent Areas) is a project within the German Federal Institute for Geosciences and Natural Resources (BGR), which is being undertaken under the auspices of the Commission for the Geological Map of the World (CGMW). Its aim is to create a harmonized geological dataset for the whole of Europe and adjacent areas at a scale (or resolution) of 1:5,000,000. The project's initial focus is on the pre-Quaternary geology and in addition to coverage of the terrestrial areas, it also integrates the much less well-known offshore domain. In total, there are approximately 2,570 area descriptions and 32,000 polygons, of which 27,500 are onshore and 4,500 are offshore.

The project was initiated in December 2004 and its scope extends not only to the whole of Europe, but also to areas of the Middle East, North Africa, and North America. Forty-

eight national geological surveys and more than 20 academic experts from across Europe have made substantial contributions to its goal.

In line with traditional CGMW projects, the IGME5000 deliverables included a printed map, which was published in February 2006; however, in contrast to previous CGMW projects, this printed map was the product of a digital spatial database. The IGME5000 is very much a geoinformatics exercise and since June 2006, it has been possible to access the dataset via a Web mapping application.

The IGME5000 faced some significant challenges. Foremost amongst these was the recognition that few, if any, true geoscience standards existed (or indeed exist) and that new standards needed to be developed. Another significant challenge was the integration of the onshore and offshore—geological domains, which necessitate very different approaches to their survey and depiction. Add to these the challenge of the logical classification of tectonic and genetic parameters, plus the complexities of metamorphic terrains and the "special" areas, such as the Alps or the Mediterranean Sea, and the technical and scientific difficulties become evident.

The organizational and cultural issues proved equally challenging. The project team had to deal with "national boundary faults" and mismatches; differences in the classification of age and lithology, and colors and symbols; idiosyncratic abbreviations; a plethora of input media; the diversity resulting from 48 different countries and geological survey organizations across 4 continents; the "variety" of opinion from more than 20 academic advisors; and the limitations of a small core team in BGR and linked to this a small budget (hence, the need to work on a voluntary basis). In addition, the complexity of multiple languages and the need to work in only one (English), the well-established independent (idiosyncratic?) behavior of geologists coupled with usually inadequate communication between them and informatics experts, and, finally, the seemingly strange concept of deadlines, made the scale of the challenges apparent.

The solution to these was pragmatic and as diverse as the challenges themselves. The IGME5000 was a marathon, not a sprint; the project would take time and would need persistence with both contributors and their superiors. The international and diverse nature of the participants meant the project would need to be inclusive and ensure adequate consultation. Lack of budget meant much time would be needed to spread the message, and in turn the BGR team would need to inspire and be enthusiastic with contributors. There would be a need for constant communication and patience. Standards would have to be based on the lowest common denominator. Expecting perfection was not realistic and accepting less than perfect contributions would allow progress. There was a need to maintain interest and, thus, to make sure the project focused on the scientific goals of certain experts. Substantial advantage was gained in choosing a suitable board of advisors and from being part of a global umbrella organization (the CGMW). Keeping things simple was essential (for example, a simple data table in addition to a sophisticated—and thus unused!—MS Access data input template). Last but not least, the language problem was, paradoxically, helped by the author being a nonnative, "nonperfect" English speaker.

The IGME 5000 has now delivered a digital spatial database of Europe's bedrock geology, but work on the project continues. The Web mapping application will be further developed and optimized. The project and its team are playing a role in the new EU Directive—Infrastructure for Spatial Information in Europe (INSPIRE). The IGME5000 is providing a basis for new digital standards for the CGMW. The availability of harmonized digital geological data for the whole of Europe has led to active involvement by BGR in a European Union project investigating geochemical fingerprints for determining the origin of food TRACE. The IGME5000 has also started to feed valuable practical experience into the new geoscience project, which is attempting to produce a spatial database of the whole planet—OneGeology.

In undertaking this project, many lessons have been learned. An extensive multinational project with many contributing nations and states requires common standards and these standards need to be developed on the basis of a lowest common denominator. Cooperation is a matter of communication and patience. Lack of budget does not necessarily mean no dedication or contribution, but it will mean that things take longer. Many nations mean many cultures. Indeed, adopting one language means a constant source of misunderstanding with which one has to cope. Set your target high but do not be disappointed if you do not reach all your scientific and technical objectives in full. Keep things simple and consider the end user at all times. Acknowledge and be positive about the support you have been given (even if it is not exactly what you need). Be aware that IT people and geologists are different species and that while they need to work closely together they will approach the task differently. Finally, recognize that in all of life's ventures, at some point discussions have to stop and hard decisions have to be made; otherwise a project like this will never be finished.

Integrated Geological Map Database (GeoMapDB) in Geological Survey of Japan, AIST

By Shinji Takarada[1], Daisaku Kawabata[1], Ryoichi Kouda[1], Jun-ichi Miyazaki[1], Yuichiro Fusejima[2], and Hisashi Asaue[1]

[1]Geological Survey oF Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

[2]Active Fault Research Center, Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Publications and Databases in GSJ

The Geological Survey of Japan (GSJ) published geological maps at scales of 1:50,000, 1:75,000, 1:200,000,

1:1,000,000, and 1:2,000,000 during the last 125 years. Other maps, such as active fault, active volcano, marine geology, hydrogeology, mineral resources, coal, oil and gas fields, intensity of aeromagnetic, and geothermal maps, are also published. GSJ published a total of 41 CD-ROM series, including 1:200,000-scale geological maps in vector and raster formats. GSJ established more than 10 databases in the Research Information Data Base (RIO-DB; http://www.aist.go.jp/RIODB/riohomee.html) of the AIST. The databases include active fault, active volcano, seamless digital geological map, geothermal drill core, marine mineral, geophysical exploration activity, basement rocks, geochemical map, crustal stress, and geological literature databases. These maps and databases have been published on printed sheets, CD-ROMs, and normal html-based Web sites. GSJ decided to integrate most of these maps and databases on a Web-based geographic information systems (GIS) to facilitate the accessibility of the geological data of the organization in 2005.

GeoMapDB

GSJ introduced a new Integrated Geological Map Database (GeoMapDB) in September 2006 (see fig. 1; http://iggis1.muse.aist.go.jp/en/top.htm). The GeoMapDB is based on a WebGIS (ArcIMS) technology, which makes it possible to browse, overlay, and search geological maps online. The purpose of this database is to make many kinds of geological maps produced by GSJ accessible to the general public. The database contains geological maps with scales ranging from 1:2,000,000 to 1:25,000. The database includes the 1:1,000,000-scale geological map of Japan (3rd edition), 1:200,000-scale seamless digital geological map of Japan (http://www.aist.go.jp/RIODB/db084/index_e.html), 1:200,000-scale geological map of Japan (raster at 150 dpi and vector formats), 1:50,000-scale quadrangle series (raster format), and the 1:25,000-scale environmental geologic map of the Tsukuba Science City (raster and vector formats). It is possible to search information using the attribute tables of maps in vector format. Legends and cross sections of the 1:50,000-scale quadrangle series and environmental map of Tsukuba city are available. Links to Quaternary volcanoes are also available. Links to other databases, such as geological literature, outcrop information, dating, and geological sample databases will also be made available soon. Three-dimensional (3D) display of the viewing area is also possible. Downloading the viewing image at 150 dots per inch (dpi) and original files

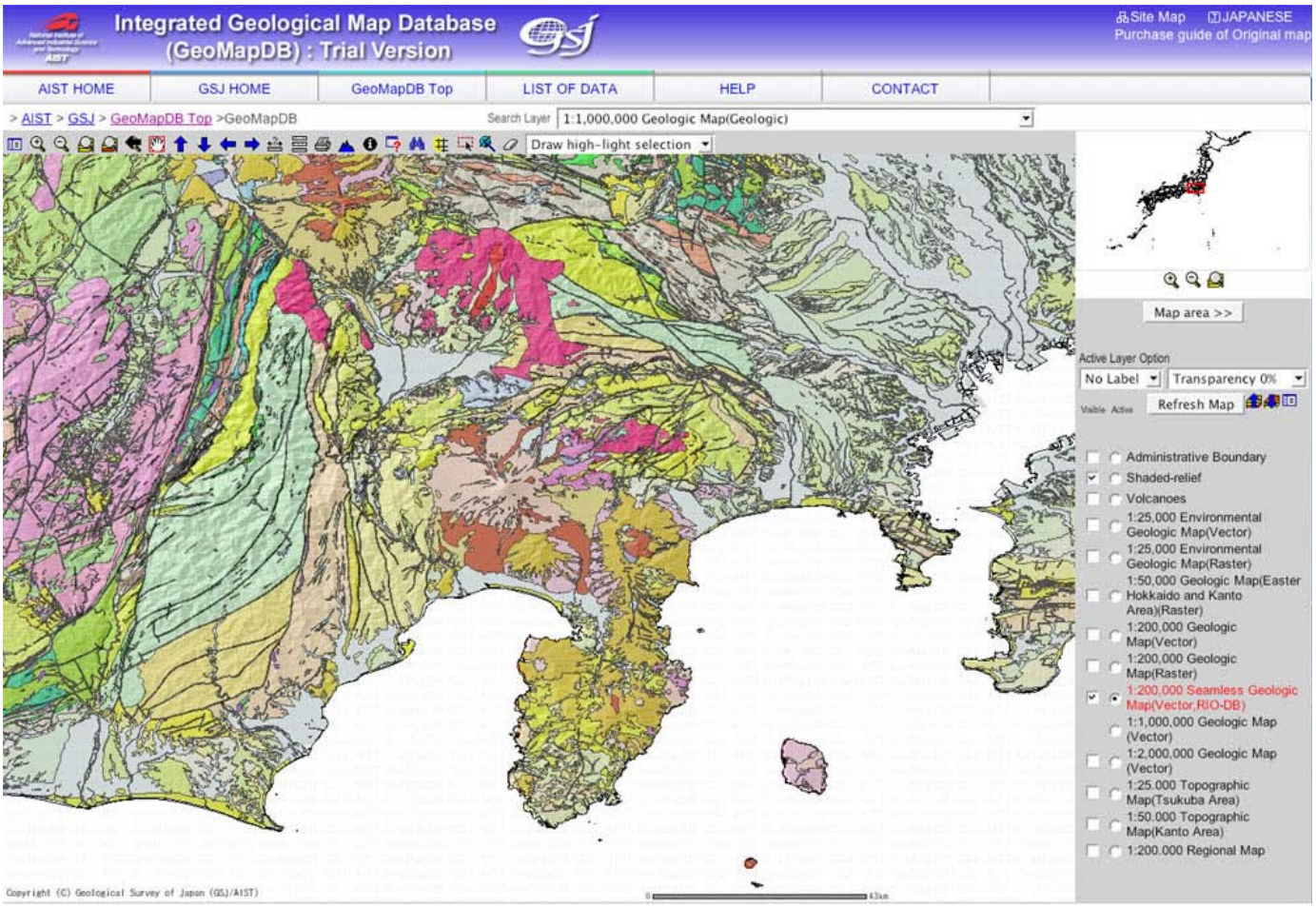

Figure 1. 1:200,000-scale seamless digital geologic map from the Integrated Geological Map Database.

in raster and vector formats is possible. Web Mapping Service (WMS) for the 1:1,000,000-scale geological map of Japan and 1:200,000-scale seamless digital map of Japan is available; thus, overlapping borehole data and landslide data from other agencies and overlaying on the Google Earth map is possible. GSJ decided to contribute the data from GeoMapDB to the OneGeology project (http://www.onegeology.org/), which aims to make a 1:1,000,000-scale digital geological map of the world using WFS. Collaboration with the GEO Grid project (http://www.geogrid.org/) is another major activity of the GSJ for the next several years.

Semantic Web Services in a Virtual Observatory (VO)

By Peter Fox[1], Luca Cinquini[2], Deborah L. McGuinness[3], and Patrick West[1]

[1]High Altitude Observatory (HAO), National Center for Atmospheric Research (NCAR), Boulder, Colo.

[2]Scientific Computing Division (SCD), National Center for Atmospheric Research (NCAR), Boulder, Colo.

[3]Knowledge Systems, McGuinness Associates and Stanford University, Stanford, Calif.

We present a set of four Web services resulting from our work on a semantic data framework in the setting of virtual observatories. These services allow a client service to search for data using three primary selections—parameter, date-time range, and instrument—and also to return appropriate service links to the actual data (the fourth service). These services use a shared understanding of the inputs, outputs, and preconditions as defined by a formal ontology, encoded in the World Wide Web Consortium's Ontology Web Language Recommendation and running in an Internet-accessible environment with Web Service Description Language (WSDL) bindings. The services can utilize reasoning services just as a user of the Web portal is able to. The service client can optionally utilize the ontology when it consumes the service for additional knowledge or may be used purely syntactically (like most existing Web services). We present these services within a specific domain context for the Virtual Solar-Terrestrial Observatory (VSTO).

Introduction

We are exploring ways of technologically enabling scientific virtual observatories—distributed resources that may contain vast amounts of scientific observational data, theoretical models, and analysis programs, and results from a broad range of disciplines. Simply, the main aim of a VO is to make all resources appear to be local and appear to be integrated. This is challenging because the information is collected by many research groups, using a multitude of instruments with varying instrument settings in multiple experiments with different goals, and captured in a wide range of formats. We must provide a means for an incoming user to discover, locate, retrieve, and use heterogeneous, and perhaps diverse, or interdisciplinary data of interest. We also must provide interfaces for requests from user applications and machine-generated requests for services.

We present the Web services aspect of our work on semantic integration of scientific data (Fox and others, 2006) in the context of the VSTO project. VSTO presently covers the fields of solar atmospheric physics and terrestrial middle and upper atmospheric physics. We used semantic Web technologies to create declarative, machine operational encodings of the semantics of the data to facilitate interoperability and semantic integration of data. We then semantically enabled Web services to find, manipulate, and present scientific data over the Internet. We describe our implementation of the Web service as part of our Virtual Observatory project (fig. 1).

Service-Oriented Architecture (SOA)

There are notable, successful examples of enabling e-science using an Internet-based service-oriented architecture using Web services. The International Virtual Observatory Alliance (IVOA; http://www.ivoa.net) developed a number of "simple access protocols" for application interoperability that are widely used in the astronomy community, including the Simple Spectrum Access Protocol and the Simple Time Access Protocol. Also, the Open Geospatial Consortium (OGC; http://www.opengeospatial.org) has, via their standards process, developed protocol standards, such as the Web Coverage Service (WCS), the Web Feature Service (WFS), and the Web Map Service (WMS)—together known as WxS. These services are also in wide use in applications needing integration via the geospatial coordinate system. We note these two examples in this context since they provide light-weight semantics via Web services. This means that while not containing formal semantic encodings, their names and basis in either particular data types or coordinate referencing and data-product types provide hard-coded semantic meaning to those clients accessing them. In essence, terms like image, spectrum, time, coverage, feature, and map have a well-defined meaning in those communities to provide great utility.

Our need to provide service in an SOA environment arises from collaborations we have with the Virtual Ionosphere-Thermosphere-Mesosphere Observatory (VITMO; http://vitmo.jhuapl.edu) and Madrigal (http://tmadrigal.haystack.mit.edu/madrigal) projects.

Semantic Data Framework

An unexpected outcome of the additional knowledge representation and reasoning was that the same data query workflow is used across the two disciplines. We are finding that it seems to generalize to a variety of other datasets as well

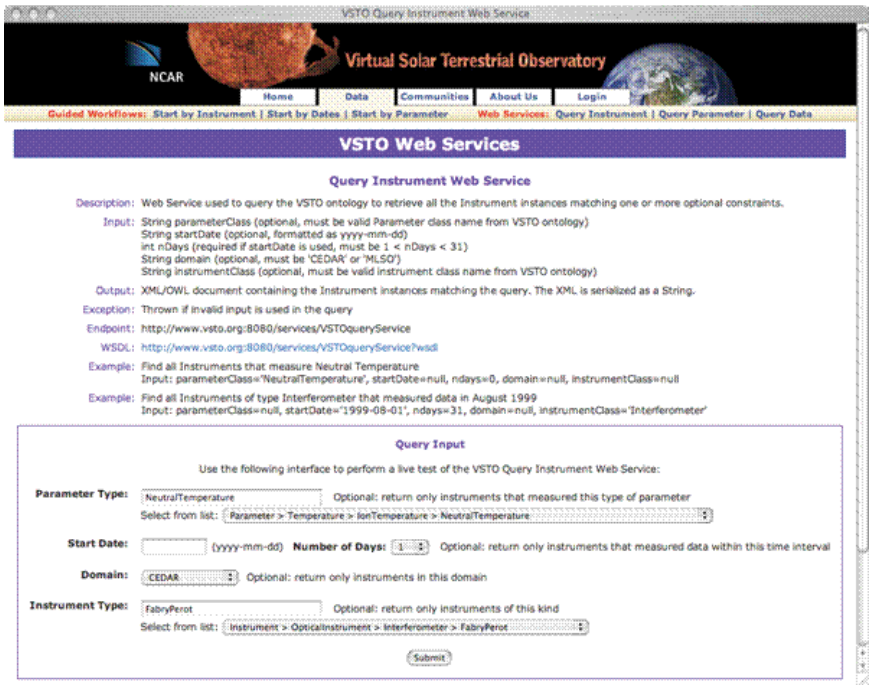

Figure 1. VSTO Web services end-point and input example for the query interface initiating an instrument search. The development of Web interfaces naturally followed from the Web portal interface functionality.

and we have seen evidence supporting this expectation in our work on other semantically enabled data-integration efforts in domains including volcanology, plate tectonics, and climate change (Fox and others, 2006). Given the value added by this basic knowledge representation and reasoning, we extended the method of access to support computer-to-computer interfaces, particularly via the commonly adopted service-oriented architecture implemented as Web services.

An example of the query instrument service follows in figure 2. A consumer of such a service, either another service, or client application may parse the OWL as XML without semantic meaning and use their own reasoning engine (or ours) to further work with the returned information.

Discussion and Conclusions

We currently have two clients using VSTO Web services: VITMO and the Madrigal Virtual Observatory. Now that our Web services are deployed at www.vsto.org, we are in a position to augment the search and query we provide in the VSTO

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://dataportal.ucar.edu/schemas/vsto_all.owl#"
         xmlns:vsto="http://dataportal.ucar.edu/schemas/vsto.owl"
         xmlns:cedar="http://dataportal.ucar.edu/schemas/cedar.owl#"
         xmlns:mlso="http://dataportal.ucar.edu/schemas/mlso.owl#"
         xmlns:owl="http://www.w3.org/2002/07/owl#" xml:base="http://dataportal.ucar.edu/schemas/vsto_all.owl">
  <vsto:FabryPerot rdf:ID="cedar_instrument_5000">
    <vsto:hasDescription>South Pole Fabry-Perot Interfer Spectr</vsto:hasDescription>
    <vsto:hasName>SPF</vsto:hasName>
    <vsto:hasIdentifier>5000</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5015">
    <vsto:hasDescription>Arrival Heights Fabry-Perot Interf Sp</vsto:hasDescription>
    <vsto:hasName>AHF</vsto:hasName>
    <vsto:hasIdentifier>5015</vsto:hasIdentifier>
  </vsto:FabryPerot>
```

Figure 2. VSTO instrument query output excerpt returning OWL documents with semantic information on the available instruments according to the input selections in figure 2. Figure 2 shows the example end point for the Query Instrument semantic Web service. The Web service inputs (all optional) and their types are described, and the end-point service address is given along with a link to the Web Services Description Language (WSDL) document content for the service. Two examples are given and below that there is a Query Input form that allows a potential user of the service to scope the types of queries that they may wish to make (including the semantic filters discussed earlier). Figure 2 contains an excerpt from an example query response beginning with information about 2 of the 13 valid instruments.

portal by installing VSTOWeb services at remote locations. These services would then be accessed when a user navigates the query workflow, resulting in a distributed set of queries using Web services displayed to the user.

We are also beginning work on capturing provenance. We plan to leverage the Proof Markup Language (PML-P) (Pinheiro da Silva and others, 2006)—an interlingua for representing provenance, justification, and trust information. Our initial work will just use the PML-P portion of the ontology—just focusing on where information came from and which services were called, later focusing on exposing the actual reasoning performed. Once captured in PML, the Inference Web toolkit (McGuinness and others, 2004) may be used to display information about why an answer was generated, where it came from, and how much the information might be believed and why.

We have also reviewed the semantic Web services with respect to needs for the NSF-funded Geosciences Network (GEON) project, the NASA-funded Semantically-Enabled Scientific Data Integration (SESDI) project, and the NASA-funded SKIF project, and plan to add these as motivating use cases to provide the most robust and smart Web service that we can.

The VSTO project is funded by the National Science Foundation, Office of Cyber Infrastructure under the SEI+II program, grant number 0431153. The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research with substantial sponsorship from the National Science Foundation.

References Cited

Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., and Solomon, S., 2006, Semantically-enabled large-scale science data repositories, *in* Cruz, I., and others, eds., The Semantic Web—ISWC 2006, Proceedings of the Fifth International Semantic Web Conference, Athens, Ga., November 5-9, 2006: Lecture Notes in Computer Science, v. 4273, p. 792-805.

McGuinness, D., and Pinheiro da Silva, P., 2004, Explaining answers from the Semantic Web: The inference Web approach: Web Semantics: Science, services and agents on the World Wide Web, v. 1, no. 4, p. 397-413.

Pinheiro da Silva, P., McGuinness, D., and Fikes, R., 2006, A proof markup language for semantic Web services: Information Systems, v. 31 nos. 4-5, June-July 2006, p. 381-395. [previous version, KSL Tech Report KSL-04-01.]

Semantic Mediator Architecture for Environmental Science

By Luis E. Bermudez[1] and John Graybeal[1]

[1]Research and Development, Monterey Bay Aquarium Research Institute, Moss Landing, Calif.

Adopting metadata specifications is often insufficient to achieve interoperability among geospatial information communities due to the heterogeneity of the values in metadata annotations. In geological sample databases, semantic heterogeneities could occur in the rock types, sample technique type, sampling platform, and analysis procedures, to name just a few. For example, when a sample is collected by hand by a diver, the Petrological database (PetDB) calls it "dive," while the SamplesDB database at Monterey Bay Aquarium Research Institute (MBARI) uses the term "hand collected."

The Marine Metadata Interoperability (MMI) project is working to address semantic conflicts. The work is guided by community collaborations and supported via the MMI site (http://marinemetadata.org). MMI focuses on several activities to achieve semantic interoperability: (1) encouraging reuse of existing vocabularies; (2) providing best practices for publishing controlled vocabularies so that they are interoperable within the Semantic Web; (3) hosting workshops to create and map controlled vocabularies; and (4) providing tools and guidance to solve semantic heterogeneities.

Along these lines, MMI has developed an architectural concept and a prototype implementation of a semantic mediation service. In this paper, we present this architecture and implementation, and discuss its potential application to the geosciences.

The two basic solutions to solve semantic heterogeneities are the mediator-wrapper approach, and the enforced standard approach. The wrapper is a piece of software build on top of the data sources, which serves the metadata and data via a common model. Traditionally, queries in the central system are translated to queries in local systems. The enforced-standards approach requires that every data source provide the data and metadata according to a single standard, including using the same exact terms to specify semantic meaning. Clearinghouses that harvest metadata in a particular format are an example of this approach.

The MMI proposition to solve semantic heterogeneities is a mixture of the two previous approaches. It requires metadata be made available in a standard format, yet allows the controlled vocabularies in the system to be heterogeneous. We propose that the semantic mediator is a reusable, sharable component of a service-oriented architecture (see fig. 1). A centralized mediator facilitates lookup services, registry of vocabularies, mappings, and queries that other components of the system could use.

This mediator, which is available at http://marinemetadata.org/semor, is being used at Integrated Ocean Observing System (IOOS) Tethys (OOS Tethys), which is the Southeast-

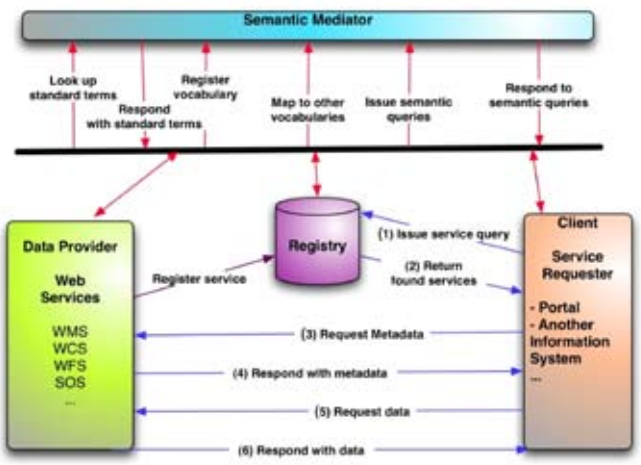

Figure 1. Semantic mediator.

ern University Research Association/Marine Metadata Interoperability (SURA/MMI) demonstration. The semantic mediation service is based on Semantic Web technologies, such as OWL and resource description framework (RDF) to store and retrieve controlled vocabularies represented in ontologies. It has a Web user interface and a Simple Object Access Protocol (SOAP) Web service to interact with it programmatically. It is currently based on a combination of Sesame and Jena, and support for Simple Protocol RDF Query Language (SPARQL) Protocol and RDF Query Language SPARQL/SOAP is planned in the near future. Mapping of vocabularies is performing via the Vocabulary Integration Environmental (VINE) tool, a stand-alone application specialized in creating custom mappings. Currently, the mappings and rules are loaded to the semantic mediator service manually, and the semantic mediator regenerates all the relationships (including inferred ones) in the knowledge base every time a new ontology is added.

These semantic capabilities will prove increasingly useful in a wide range of environmental science applications, as more data systems are directly and indirectly linked to provide interoperable services. The MMI-developed semantic architectures and tools, along with many others presented on the MMI Web site, are targeted to develop interoperable environmental data systems on a national and international scale.

Semantic Integration of Heterogeneous Volcanic and Atmospheric Data

By Deborah L. McGuinness[1], Peter Fox[2], A. Krishna Sinha[3], and Robert Raskin[4]

[1]Knowledge Systems, McGuinness Associates and Stanford University, Stanford, Calif.

[2]High Altitude Observatory (HAO), National Center for Atmospheric Research (NCAR), Boulder, Colo.

[3]Department of Geology, Virginia Polytechnic Institute and State University, Blacksburg, Va.

[4]Jet Propulsion Laboratory, Pasadena, Calif.

The vast majority of explorations of the Earth system are limited in their ability to effectively explore the most important (often most difficult) problems because they are forced to interconnect at the data-element, or syntactic, level rather than at a higher scientific, or semantic level. In many cases, syntax-only interoperability is the state-of-the-art. Currently, in order for scientists and non-scientists to discover, access, and use data from unfamiliar sources, they are forced to learn details of the data schema, and other people's naming schemes and syntax decisions. These constraints are limiting even when researchers are looking for information in their own discipline, but they present even greater challenges when researchers are looking for information spanning multiple disciplines, including some in which they are not extensively trained. Our project, the Semantically-Enabled Scientific Data Integration (SESDI), aims to demonstrate how ontologies implemented within existing distributed technology frameworks will provide essential, reusable, and robust support necessary for interdisciplinary scientific research activities.

Introduction

Our project is aimed at enabling the next generation of interdisciplinary and discipline-specific data and information systems. Our initial focus is the integration of volcanology and atmospheric data sources in support of investigations into relationships between volcanic activity and global climate.

This work is aimed at providing scientists with the option of describing what they are looking for in terms that are meaningful and natural to them, instead of in a syntax that is not. The goal is not simply to facilitate search and retrieval, but also to provide an underlying framework that contains information about the semantics of the scientific terms used. Our system is expected to be used by scientists who want to do processing on the results of the integrated data, thus the system must provide access to how integration is done and what definitions it is using. The missing element in previous systems in enabling the higher level semantic interconnections is the technology of ontologies, ontology-equipped tools, and semantically aware interfaces between science components. We present the initial results of using semantic technologies to integrate data between these two discipline areas to assist in establishing causal connections, as well as exploring as yet unknown relationships.

Semantic Data Integration Methodology

Our effort depends on machine-operational specifications of the science terms that are used in the disciplines of interest. We are following a methodology that we believe is yielding candidate reference ontologies in our chosen domains. We have identified specific ontology modules that need construc-

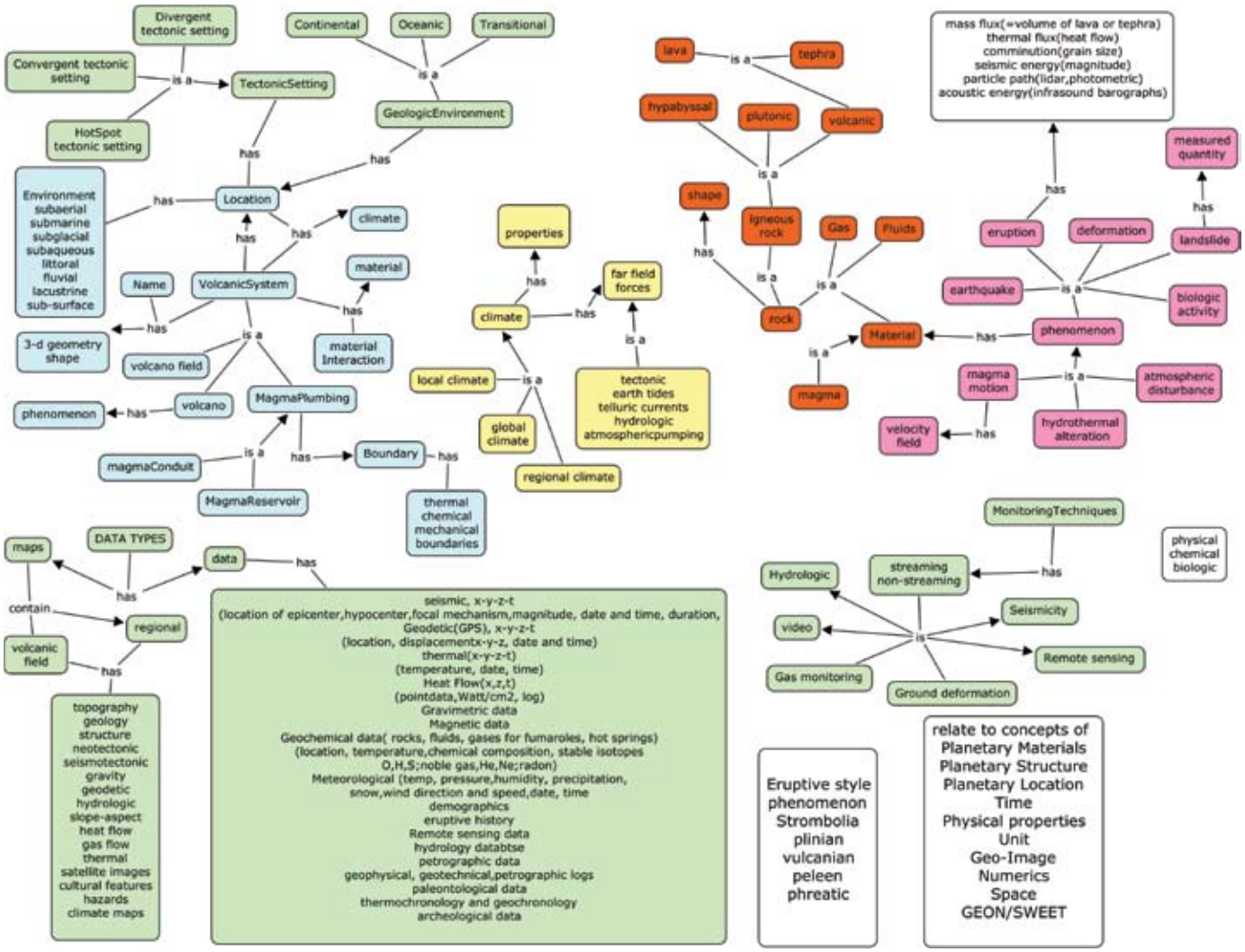

Figure 1. Volcano ontology CMAP fragment. Volcanoes can be classified by composition, tectonic setting, environmental setting, eruption type, activity, geologic setting, and landform. The ontology currently contains upper level terms in these areas and is being expanded according to the needs of the project and is being reviewed by additional domain experts. The initial focus is on gathering terms, putting them into a generalization hierarchy (using "isa" links in the diagram), and connecting the terms through properties (using "has" links in the diagram), as well as identifying equivalence relationships (using "sameas" links) and partonomic information (using "ispartof" links).

tion in the areas of volcanoes, plate tectonics, atmosphere, and climate. We have begun construction of two of the modules along with the help of a set of selected experts in the areas. Prior to a workshop, we identify a small set of subject matter experts. We also provide some background material for reading about ontology basics. Additionally, prior to our face-to-face meetings with experts, we identify foundational terms in the discipline and provide a simple starting point for organizing the basic terminology. While we do not want to influence the domain experts on their terminology, we find that we make more progress if we provide simple starting points using agreed-upon terminology. We then bring together a small group of the chosen domain experts and science ontology experts with a goal of generating an initial ontology containing the terms and phrases typically used by these experts. We use our task of researching the impact of volcanoes and global climate to focus the discussions to help determine scope and level of granularity.

We held a meeting with volcano experts and generated an initial ontology containing terms and phrases used to classify volcanoes, volcanic activities, and eruption phenomena. We use a relatively simple graphical tool (CMAP) for capturing the terms and their relationships. A portion of the initial volcano ontology is shown in figure 1 (from Sinha and others, 2006; McGuinness and others, 2006).

We held a second workshop to create a plate tectonics ontology. We identified domain experts and used the same science ontology experts as used in our volcano ontology meeting. The resulting terminology description is shown in figure 2. In this meeting, we also focused on gathering the primary

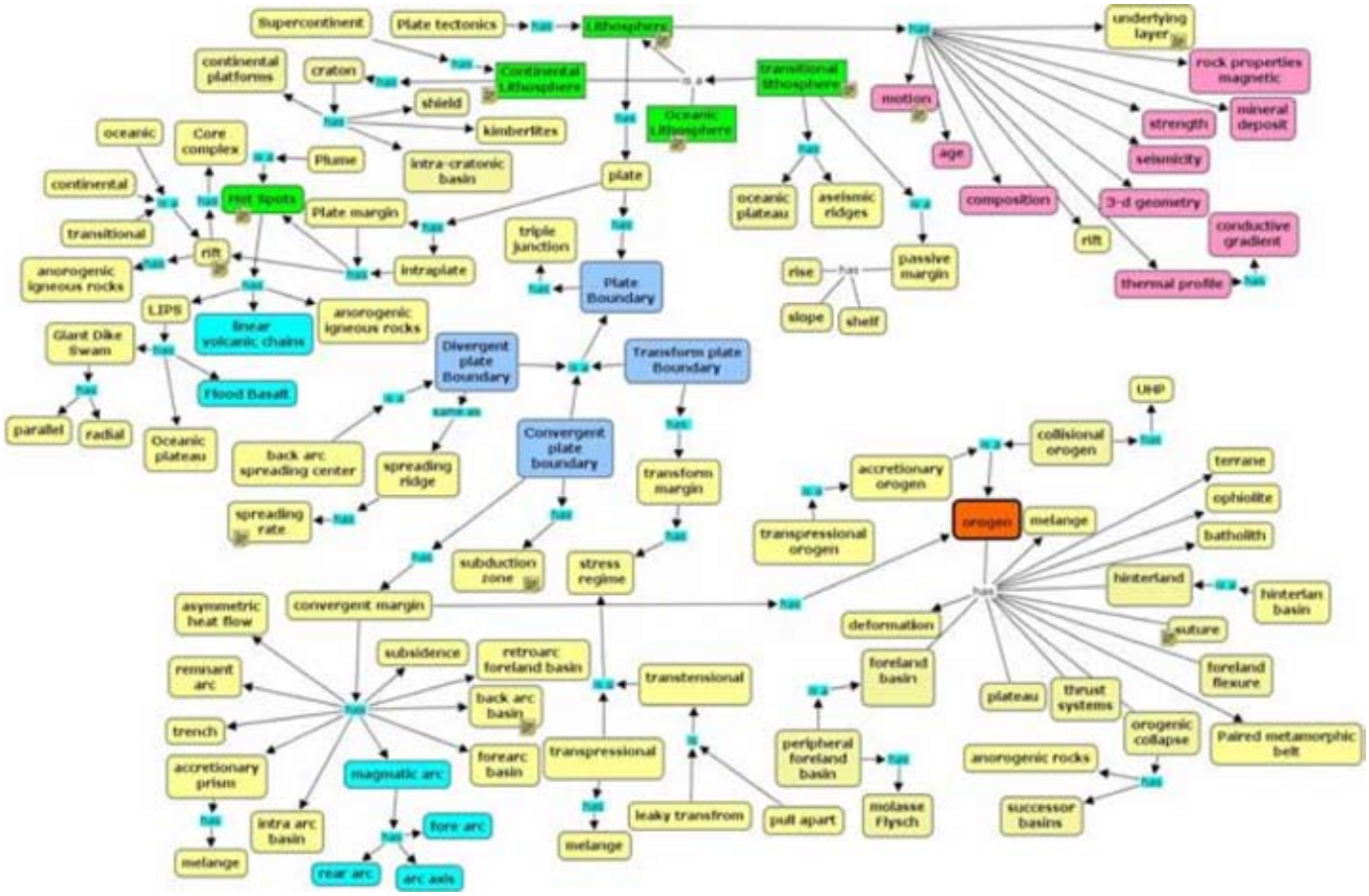

Figure 2. Plate tectonics ontology CMAP fragment. Both in preparation for and in followup from the domain workshops, we are reviewing relevant existing vocabularies and ontologies. We have reused terminology from SWEET (http://sweet.jpl.nasa.gov/ontology/), GEON ontologies (http://kbis.sdsc.edu/GEON/resources/resources.html), and the Virtual Solar Terrestrial Observatory (McGuinness and others, 2006) instrument and observatory ontologies. We are also gathering some of the starting points for the atmosphere and climate ontologies from SWEET and related ontologies.

class terms (for example, plate boundary, lithosphere, and so on), putting them into a generalization hierarchy, and identifying important properties relating the terms.

Our candidate datasets for registration with the developed ontologies and for use in our data integration use case include the WOrld VOlcano DATabase (WOVODAT) in collaboration with Yellowstone researchers and the U.S. Geological Survey, and the Nevada Test site database in collaboration with Los Alamos researchers. We are presently identifying the corresponding atmospheric and climate record databases. Our approach will be to establish initially a Web portal based on one of the current best-of-breed semantic data frameworks (for example, VSTO, GEON) to prototype the access to the data from the volcano and atmosphere disciplines separately. Then we will provide Web service access to both sources so that the statistical application we will utilize for the data integration can query and retrieve data.

Discussion and Conclusion

We have begun an effort that utilized ontologies to provide the capture term meanings in distinct but related science domains with a goal of facilitating research into relationships between the domains. We currently have starting points for reference ontologies in volcanoes and plate tectonics. We have also begun the homework on atmosphere and climate and will be holding workshops to generate reference ontologies with domain experts. We also will hold workshops to vet the ontologies among the multiple communities. So far, our findings are that our methodology for creating starting points for reference ontologies is working well in terms of gathering terms and relationships, and reaching agreement among the initial domain and science ontology experts.

Based on the successful use of semantics in data integration for the VSTO and GEON projects, our next step for SESDI is to articulate a use case that drives the way and type of data integration needed to solve a specific scientific problem. Our candidate is to examine the statistical relation

between the height of the tropopause and related forcings. This height is very sensitive to forcing so that the fingerprint of volcanic and solar forcings are very distinct.

References Cited

Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., and Solomon, S., 2006, Semantically-enabled large-scale science data repositories, *in* Cruz, I., and others, eds., The Semantic Web—ISWC 2006, Proceedings of the Fifth International Semantic Web Conference, Athens, Ga., November 5-9, 2006: Lecture Notes in Computer Science, v. 4273, p. 792-805.

McGuinness, D.L., Sinha, A.K., Fox, P., Raskin, R., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., and Lin, K., 2006, Towards a reference volcano ontology for Semantic scientific data integration: Eos, v. 87, no. 36, Joint Assembly Supplement Abstract IN42A–03.

Sinha, A.K., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., Fox, P., McGuinness, D.L., Raskin, R., and Lin, K., 2006, Towards an ontology for volcanoes, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Geoinformatics 2006—Abstracts: U.S. Geological Survey Scientific Investigations Report 2006-5201, p. 52.

Semantics and Science: Facilitating a Community Centric View

By Danielle Forsyth[1]

[1]Thetus Corporation, Portland, Oreg.

Scientific information is pouring in from satellites, sensors, instruments, cameras, documents, and devices. This raw information is processed into an ever-increasing number of data products for use by the scientific, policymaking, academic, and corporate communities. Often, these communities need to share information for collaboration, yet the languages of these communities differ. Communities use different terms and relationships to describe and define their information, they trust different sources, they have different experts, and they rely on different and ever-changing models.

Understanding complex systems requires cross-community collaboration and historical perspective. Communities must be connected so that complex systems can be understood, boundary conditions can be examined, and sensitivity to potentially impacted or dependent systems can be understood. Forcing a common language for these interconnected communities will not happen, and if it did, it would result in a lack of needed information fidelity and context.

Semantic approaches to cross-community collaboration can facilitate the necessary knowledge sharing and reusability to abstract meaning in support of an overall understanding of complex systems.

This presentation demonstrates how different business, policymaking, and scientific communities can collaborate using multiple domain or problem representations (knowledge models). It illustrates the interconnections and information/knowledge sharing between seemingly unrelated communities where serendipitous discovery matters. In addition, the presentation demonstrates how these rich semantic models can be utilized to find similar and related information and to discover unexpected connections between seemingly arbitrary pieces of information.

The presentation centers on a map-based interface that allows different communities to visualize changing information and relationships (in a spatial environment) based on their domain or problem of interest. Within this semantically rich and spatial environment, different groups of users can review metadata, make annotations, create new relationships, view information history, and filter their views to focus on the appropriate level of knowledge and information detail to address their questions. New knowledge can be captured and community members can be automatically notified of information of interest.

While semantic models are touched on, the focus of the presentation is on their use in filtering information to allow users to quickly focus on relevant and needed information in their own domains.

CyberIntegrator: A Highly Interactive Scientific Process Management Environment to Support Earth Observatories

By Rob Kooper[1], Luigi Marini[1], Jim Myers[1], Barbara Minsker[1], and Peter Bajcsy[1]

[1]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Ill.

The concept of Earth observatories has been evolving over the past decade and it is surrounded with the multitude of "informatics" needs for a successful deployment. In the context of Earth observatories, informatics refers to problems related to expected large amounts of often highly complex data that have to be analyzed; information has to be interactively extracted from raw data and then understood by domain scientists. In all application areas where Earth observatories are being designed and built (for example, Watershed Assessment, Tracking and Environmental Results (WATERS), Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), National Ecological Observatory Network (NEON), Geosciences Network (GEON), and Optical Remote

Imaging Observatory of New Mexico (ORION)), scientists desire to learn from their data about a spectrum of complex phenomena surrounding us; however, the informatics challenges for a domain expert can significantly hinder any learning progress. Regardless of a domain—geologic, hydrologic, ecologic, environmental, or sensor—general informatics challenges exist related to (a) data volume and computational requirements, (b) data, analysis and resource complexity management, and (c) the heterogeneity of information technologies supporting scientists. Our goal is to support scientists building Earth observatories to overcome these general challenges.

This paper presents a novel process management environment called CyberIntegrator to support diverse analyses in Earth observatories. These analyses are very time-consuming and hard to reproduce because of the lack of in-silico scientific process management and because of the diversity of data, software, and computational requirements. The motivation for our work comes from the need to build the next generation of in-silico scientific discovery processes that require (a) access to heterogeneous and distributed data and computational resources; (b) integration of heterogeneous software packages, tools, and services; (c) formation and execution of complex analytical processing sequences; (d) preservation of traces about scientific analyses; and (e) design of secure collaborative Web-based frameworks for sharing information and resources.

The goal of the presented work is to describe a modular architecture and key features of a workflow that provides a process management environment for automating science processes, reducing the human time involved and enabling scientific discoveries that would not be possible without supporting software and hardware infrastructure. Our approach to solving the above problem is based on adopting object-oriented software-engineering principles, designing a modular software prototype, and focusing on user interfaces that simplify complex analyses including heterogeneous software integration.

Figure 1 shows the overall architecture of the developed process management environment by bringing together the top level features with the low-level object-oriented software-engineering principles. From the functionality perspective, the CyberIntegrator software could be viewed as a system for (a) browsing and searching available data, tools, and computational resources; (b) accessing available datasets, tools, and computational resources; (c) bringing them together; (d) executing one tool at the time or a sequence of tools, (e) monitoring and controlling executions, (f) efficiently utilizing available data, tools, and computational resources; (g) collecting information, or provenance, about the process flow to help later reconstruct the thought process of the scientist; and (h) assisting the scientist using the provenance gathered by the community. The key architectural components are the editor, engine, executors, applications, collections of registries, and optional metadata repository and event broker. They are all written in Java. The registries could be viewed as repositories of high-level descriptions of available data, tools, and resources. The metadata store provides a repository for gathered information about workflow execution and it is based on a resource description framework (RDF) format. Finally, the event broker is a component for handling a stream of data or events, and it is based on Java Message Service (JMS) application programming interface (API) for sending messages between two or more clients.

Our aim has been to design a workflow that works with descriptions (metadata) of data, software tools, and computational resources for easy integration and hierarchical organization. This distributed metadata (also called "registries") about data, tools, or resources can be modified with a text or Extensible Markup Language (XML) editor. The workflow system will pull in all the information from the registries, load a requested workflow, and execute the workflow accordingly. The benefits of such workflows for domain scientists are (a) the simplicity of integrating existing software within the workflow system; (b) the benefits of running, reusing, repurposing, and sharing workflows with other scientists; and (c) receiving feedback from the system during workflow creation based on the provenance gathered.

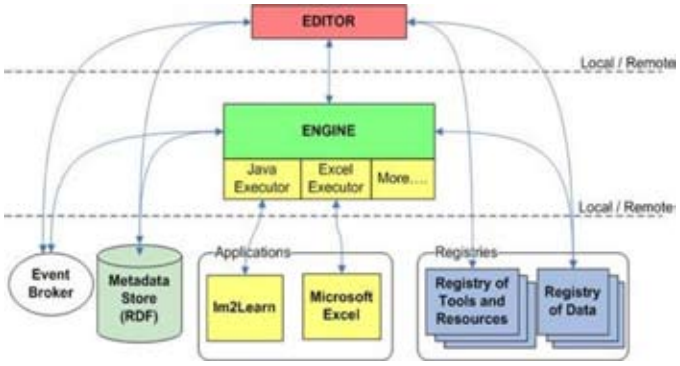

Figure 1. The overall architecture of CyberIntegrator.

CyberIntegrator editor provides a user-friendly interface for browsing registries of data, tools, and computational resources; creating workflows in a step-by-step exploration mode; reusing and repurposing workflows; executing process flows, locally or remotely; aiding research explorations using a provenance-to-recommendation pipeline; and incorporating heterogeneous tools and linking them transparently. Figure 2 shows the graphical user interface of CyberIntegrator editor. There are three top panes (left to right: Data pane, Tools pane, and Resources pane) and one pane on the bottom. The Data pane lists all datasets loaded or generated so far for processing. The Tools pane lists the tools currently loaded from local and remote registries. The Resources pane lists computational resources (executors) available for a particular tool. The bottom pane contains several tabs with information about the CyberIntegrator execution. The Help tab contains help text about the tool currently selected. The Steps tab contains a

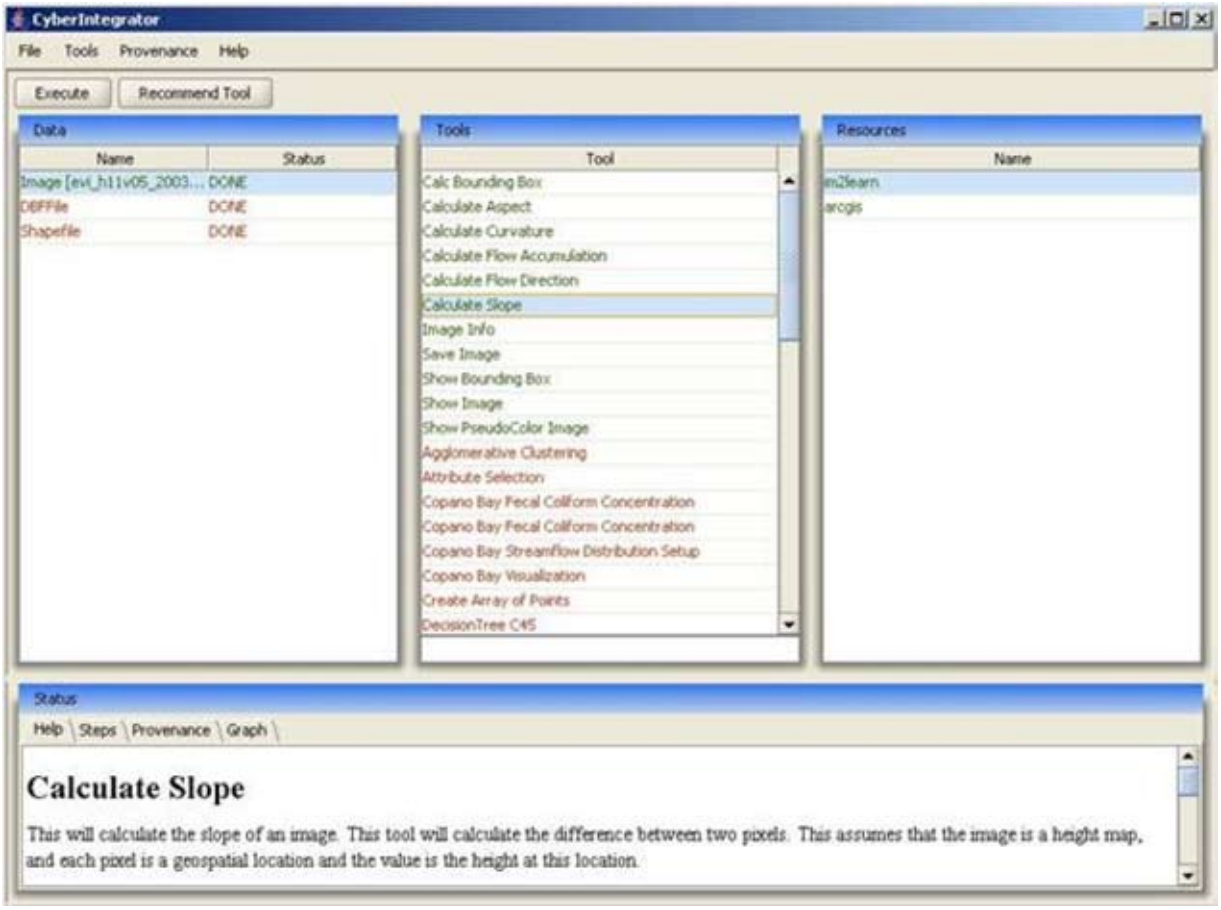

Figure 2. CyberIntegrator.

list of the tools run so far with information about each execution. The Graph tab includes a graphical representation of the sequence of steps. The current workflow can also be saved and either reloaded or shared with others later. The scientist can ask the system to recommend a tool based on the current data by selecting the button above the Data pane.

The key contributions of our work on CyberIntegrator can be summarized as follows: The main computer science novelty of our work lies in (1) formalizing the software integration framework using object-oriented software-engineering principles; (2) designing a browser-based modeling paradigm for step-by-step composition of workflows; (3) gathering provenance during workflow creation and execution; (4) using the provenance for tool recommendation feedback for workflow autocompletion; and (5) providing capabilities to publish, run, monitor, retrieve, and reuse and repurpose workflows from local and remote computational resources for long-running workflows (for example, referring to large simulation and streaming data analyses). The main technical contributions are also in testing and demonstrating the prototype process management system with several application scenarios from environmental and hydrologic engineering sciences. Our process management prototype is also supporting Waters/CLEANER and Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) communities in the context of building Earth observatories. The software is available for download at http://isda.ncsa.uiuc.edu/download.

DIA Engine: Semantic Discovery, Integration, and Analysis of Earth Science Data

By Abdelmounaam Rezgui[1], Zaki Malik[1], and A. Krishna Sinha[2]

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Va.

[2]Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.

Introduction

Geoscientists have generated massive volumes of earth science data for decades. Most of the produced data, how-

ever, remain isolated "knowledge islands"; the ability to find, access, and properly interpret these large data repositories has been very limited. This is due to two main reasons: the absence of data-sharing infrastructures that scientists may use to advertise their data, and lack of a "common language" that scientists can use to properly interpret other providers' data. As a result, the discovery, integration, and analysis of earth-science data have remained difficult. This, in turn, has precluded the meaningful use of the available data in answering complex questions that require information from several data sources. To address this problem, we have developed Discovery, Integration and Analysis (DIA)—a service-oriented, Web-based computational infrastructure that enables scientists to utilize semantically enabled technologies to discover, integrate, and analyze earth-science data. It also promotes tool sharing through Web services. It provides a collaborative environment where scientists can share their resources for discovery and integration by registering them through well-defined ontologies (Sinha and others, 2006). DIA is developed using a variety of technologies, including the following: Enviromental Systems Research Institute's ArcGIS Server 9.1, Web services, .NET, Java, and JNBridge 3.

Architecture of the DIA Engine

The DIA engine is a Web-accessible system that provides three classes of functionalities: discovery, integration, and analysis. Data discovery enables users to retrieve datasets, while data integration enables users to query multiple resources along some common attributes to generate previously unknown information called data products. Data analysis may be used to verify certain hypotheses or to refine the data product.

To describe DIA's architecture (fig. 1), we will use the following query as our type example: (Q) Find A-type plutons in Virginia and identify the correlation between these plutons and their geophysical (for example, gravity) properties. The core of DIA's engine consists of five components:

User Interface—This is an ArcGIS Server.NET map viewer Web application. DIA provides a menu-based interface that enables users to specify a large number of complex queries. Map-based queries can be refined by specifying a bounding box that identifies a pair of latitude-longitude points which delimits the query's spatial scope. After the query's spatial scope is specified, the user uses DIA's drop-down menu to indicate the filters (A-type igneous rock filter, in our running example) and (or) tools to be applied to the data samples discovered in the query's spatial scope.

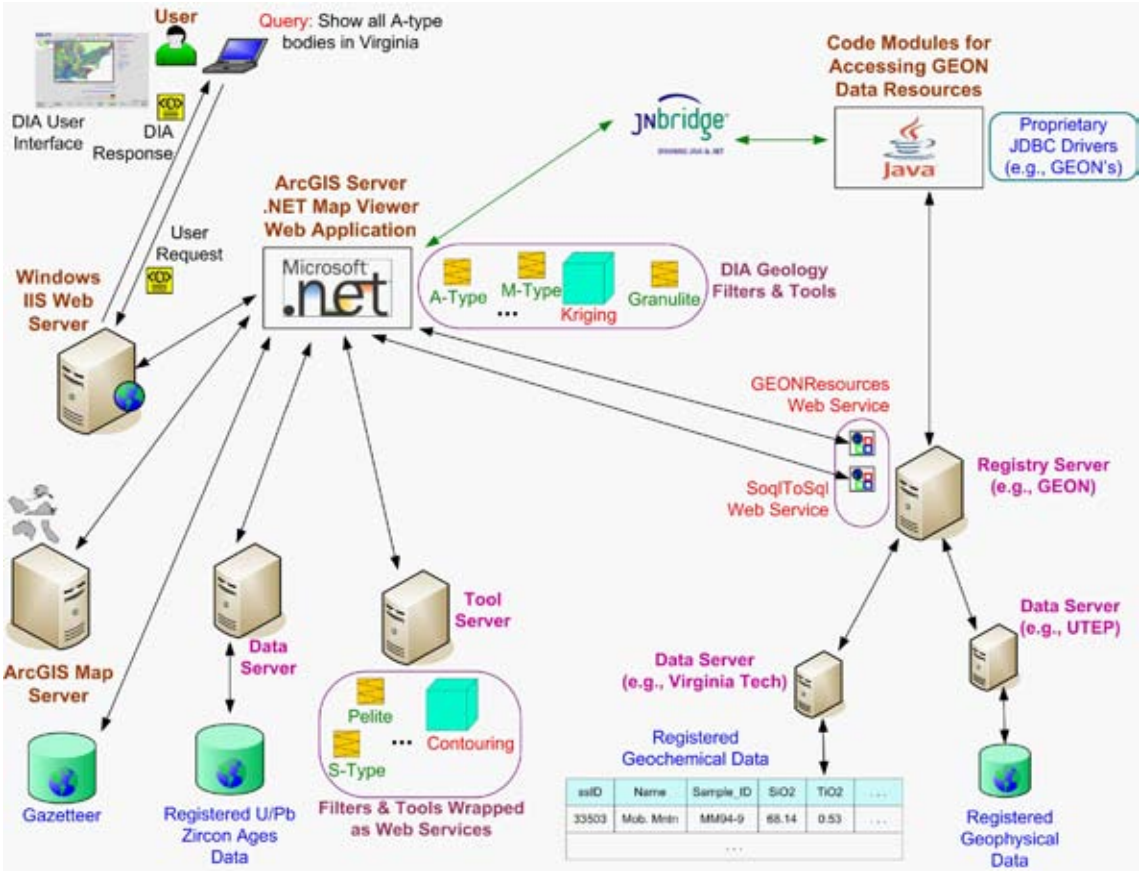

Figure 1. DIA's software architecture.

Web Servers—Typically, DIA uses two Web servers. The first Web server is responsible for routing users' queries to DIA's query processor and the second ensures communication between DIA's query processor and its own map server. In a minimal deployment, a single Web server may be used for both purposes. In the current DIA's implementation, we use a single instance of Windows IIS Web server as DIA's Web server.

Map Server—This component is an ArcGIS map server that provides maps to DIA's query processor.

Registry Servers—These servers could be distributed worldwide, and provide directory functionalities (registration of data and tools, indexing, search, and so on) The providers of resources advertise their resources on registry servers.

Query Processor (QP)—This is the core component of the DIA engine. It is responsible for producing the results for users' queries and delivering them to the Web server. Essentially, the QP consists of two subcomponents: (a) the query interpreter and (b) the geology and mapping filters and tools. The former is a .NET module that interprets queries and identifies the appropriate filters and (or) tools to be invoked to answer each query. The latter is a large set of .NET modules that perform DIA's core functionalities, including filters (for example, A-Type igneous rock filter), tools (for example, kriging) and map management routines (for example, coloring of geological bodies and sample points). Query processing consists of two phases: (a) data and tool discovery and (b) filtering and integration.

Data and Tool Discovery—During this operation, the DIA engine identifies and retrieves the resources (data and tools) required to answer the user's query. To illustrate, consider the previous A-Type query. When the QP receives the query from the Web server, it determines the type of data required to answer the query. In this case, the QP determines that data associated with the keyword "Geochemistry" is the query's target. The QP then interacts with one or several registry servers to retrieve the needed data. An example of registry server is available at http://www.geongrid.org. To interact with Geosciences Network (GEON) server, DIA invokes a GEON Web service called GEONResources that provides functions for searching and getting the metadata information for resources registered through GEON portal. When invoking GEONResources, DIA's QP indicates that it is searching datasets registered with the keyword "Geochemistry" and that contain data samples in the query's spatial bounding box. For each returned database, the DIA system executes a two-step process. First, it builds a virtual query (expressed in Simple Ontology-Based Query Language (SOQL)—a language developed by GEON's researchers at the San Diego Supercomputer Center (SDSC)) that requests all the data (in other words, columns) that are necessary to apply the filter specified by the user. The DIA system then invokes a GEON Web service called SoqlToSql that translates this SOQL query into an SQL query. In the second step, DIA submits the SQL query to the GEON server that interacts with the actual database server, gets a record set containing the relevant data samples, and returns the data to the DIA engine.

Filtering and Integration—Data filtering is a process in which the DIA engine transforms raw data into a data product. After DIA retrieves the datasets relevant to the user's query, it determines whether the filter(s) to apply or tool(s) to use is locally available. If so, the filter or tool is applied to the datasets and the query result is displayed to the user. If not, DIA searches for the needed filter or tool in registry servers. DIA is able to invoke any external tool that is wrapped as a Web service. In the case of the given A-type query, the A-Type filter is already available in DIA and also made available as a Web service for external users.

Integration in DIA is a process in which the results of several subqueries are produced and then overlaid in the user interface. In the case of our A-Type query, DIA first follows the same workflow as for determining A-Type bodies to produce the result of kriging gravity data. DIA looks up registry servers for gravity data in the selected area of interest and then retrieves the raw gravity data from its provider(s) (for example, http://paces.geo.utep.edu). DIA then determines whether a kriging tool is locally available. Since such a tool is already included in DIA's implementation, it is invoked and no external registry servers are searched. When the output of the kriging tool is generated, DIA overlays it on the previously generated results (in other words, A-Type plutons) making it possible for the user to have a natural and easily interpretable view of the integration's result (fig. 2).

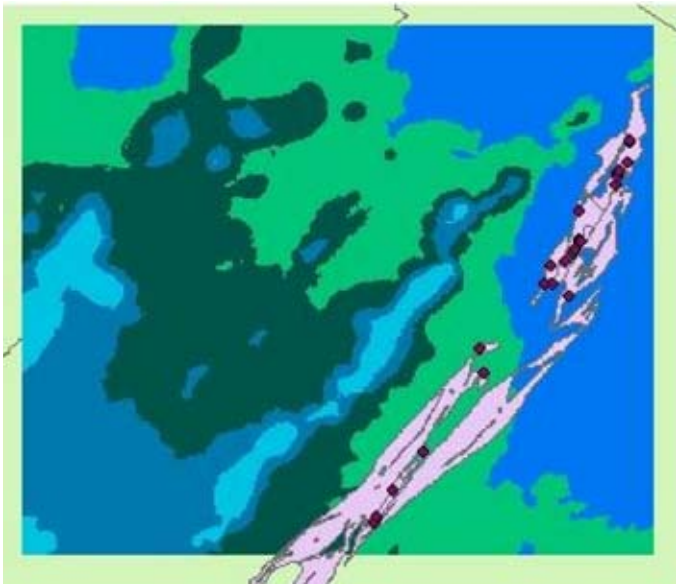

Figure 2. Semantically enabled integration of data products where A-type plutons and gravity fields have been merged through the DIA engine.

Conclusion

We suggest that the semantic integration of data and tools can be implemented through the DIA engine, a system that enables geoscientists to discover, integrate, and analyze earth-science data. DIA also demonstrates the potential of the service-oriented design paradigm to enable scientists to share tools in addition to data. The DIA engine is now in its final pre-release phase. Its beta version is currently accessible at http://mapserver.geos.vt.edu/DIA.

We expect that as the Semantic Web matures, more geoscientists will adopt ontologies as a means for data and service sharing and integration. The DIA engine is designed for Web-based geoinformatics systems. These systems would provide an infrastructure where scientists worldwide would be able to discover, integrate, and analyze data.

Acknowledgment

This research is supported by the National Science Foundation, award EAR 0225588 to A.K. Sinha.

Reference Cited

Sinha, A.K., Malik, Z., Rezgui, A., and Dalton, A., 2006, Developing the ontologic framework and tools for the discovery and integration of earth science data; Annual report, June 2006, Blacksburg, Va., Virginia Polytechnic and State University, Department of Geosciences, available online at http://geon.geol.vt.edu/pubreps/Virginia Tech Annual Report 2006.doc/. (Accessed June 12, 2007.)

Customizing a Semantic Search Engine and Resource Aggregator for Different Science Domains

By Sunil Movva[1], Rahul Ramachandran[1], Xiang Li[1], Phani Cherukuri[1], and Sara Graves[1]

[1]Information Technology and Systems Center, University of Alabama at Huntsville, Huntsville, Ala.

The goal for search engines is to return results that are both accurate and complete. The search engines should find only what the users really want and everything they really want. Search engines (even metasearch engines) lack semantics. The basis for search is simply string matching between the user's query term and the resource database; thus, the semantics associated with the search string is not captured. For example, if an atmospheric scientist is searching for "pressure" related Web resources, most search engines return inaccurate results, such as Web resources related to blood pressure. Noesis is a metasearch engine and a resource aggregator that has been designed to utilize domain-specific ontologies to provide specialized scoped search capabilities. Noesis uses domain ontologies to help the user scope the search query to ensure that the search results are both accurate and complete. Semantics are captured in the domain ontologies by defining terms along with their synonyms, specializations, generalizations, and related concepts. These domain ontologies are used by Noesis to guide the user to refine their search query and thereby reduce the user's burden of experimenting with different search strings. Noesis also serves as a resource aggregator as it categorizes the search results from different online resources, such as education materials, publications, datasets, and Web search engines that might be of interest to the user. Noesis can be customized for use in different domains by configuring it to access different ontologies or to search different online resources. Currently, we have a general purpose atmospheric science version of Noesis available, and we are creating a coastal ecology version specialized for the Gulf of Mexico Research Collaborative, a National Aeronautics and Space Administration (NASA) applications project. In addition, Noesis portlets are planned for the Earth Science Information Partners (ESIP) Federation Environmental Information Exchange and Geospatial One Stop.

Towards Debugging Maps Generated by GEON Applications Through Provenance

By Nicholas Del Rio[1] and Paulo Pinheiro da Silva[1]

[1]Department of Computer Science, University of Texas at El Paso, El Paso, Tex.

Tool Overview

Geoscientists need the capability to understand and debug maps generated from the highly distributed Geosciences Network (GEON) applications and workflows in order to accept them, particularly when a resulting map exhibits unexpected or anomalous properties. On one hand, visualization techniques can help a scientist to understand intermediate and final results of a complex GEON application but not the underlying processes that derived these results. On the other hand, provenance provides information about sources and methods used to derive results, which can also increase the understanding and acceptance of GEON-generated maps by scientists. Although rarely used in combination, visualization and provenance techniques together may further increase geoscientists' understanding of GEON maps by providing a complete picture of their generation. Scientists would be able to evaluate final results, derivation processes, and any intermediate result derived during the GEON processes. Probe-It! is a single tool that provides geoscientists with the capability to

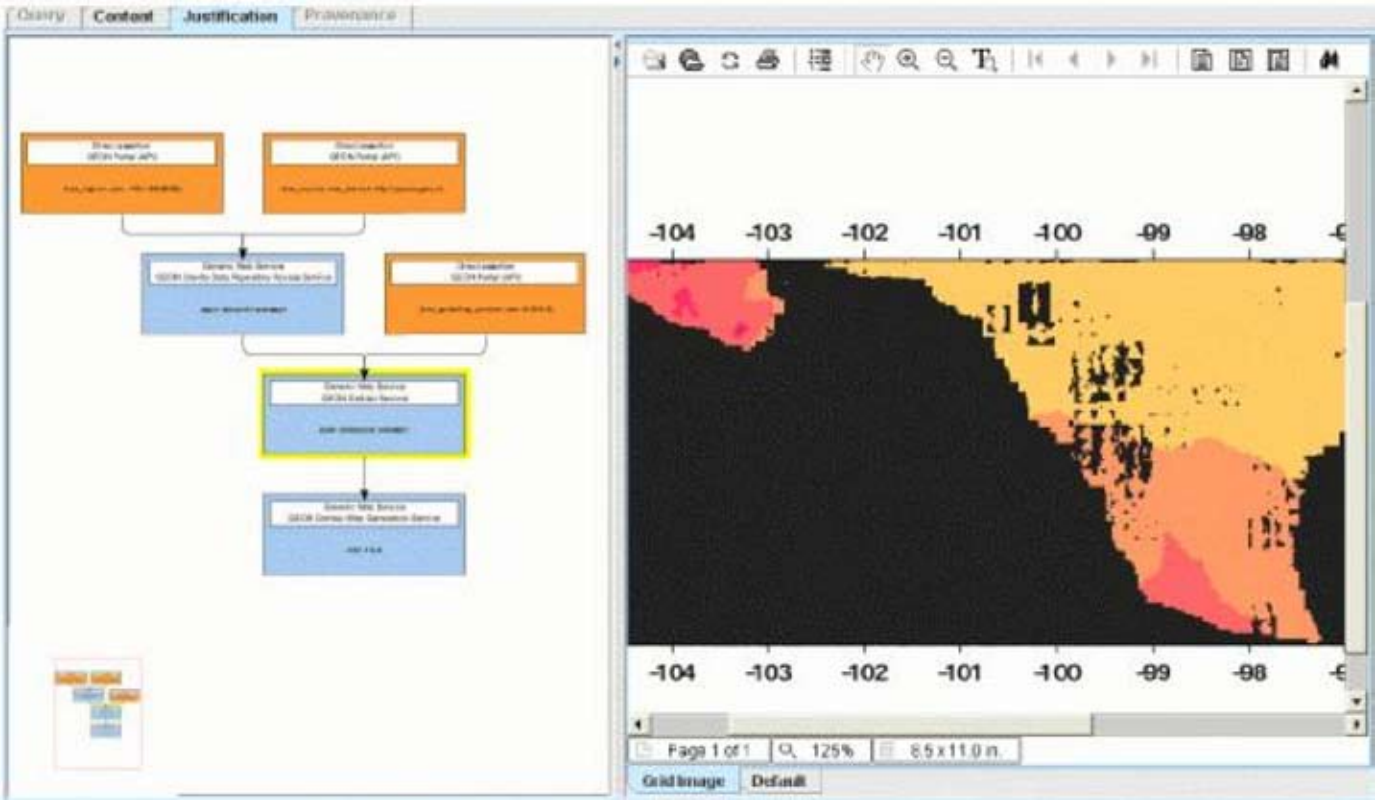

Figure 1. ProbeIt! Snapshot.

visualize provenance associated with GEON map generation in order to aid the scientist in the understanding and debugging of an unexpected map.

Because GEON maps can be generated using remote resources, such as sensory data, remote databases, and services, assessing the quality and correctness of the resultant maps is difficult because of the following: (1) scientists may not know the history associated with some data source; and (2) scientists may not know details about the underlying services applied to their data. Probe-It! addresses those issues by providing visual access to both the sources and processes used to derive a map. For example, through the use of Probe-It!, scientists can move the visualization focus from intermediate and final results of a contour map workflow to the associated provenance trace back and forth. We believe that providing geoscientists with both a description of the map generation process and the means to visualize the associated intermediate and final results will allow scientists to determine whether their resultant maps are correct. Because the most effective visualization varies between scientists, Probe-It! also provides a framework for associating particular visualizations to a particular information type. For example, a GEON scientist may prefer to view spatial datasets as a two-dimensional plot on a map, while another scientist might prefer to view the same data as a graph. Probe-It! is flexible enough to facilitate a multitude of views.

Figure 1 highlights ProbeIt!'s justification view, which outlines the provenance trail of a contour map workflow that consists of three Web services: dataset retrieval service, a smoothing service, and a contouring service; the resultant map is a contour of gravity data. The provenance associated with this workflow execution includes everything from the specified map region, provided by a geoscientist, to the final contour map of the specified region. The arrows indicate dataflow between services, while each node of the graph represents an invoked Web service and its associated output.

Evaluation Plan

In order to verify that our tool Probe-It! aids scientists in debugging anomalous results, a moderately sized study comprising of gravity experts around the globe is being initiated. Each participant will be asked to identify a correct map. In this task scenario, we will present four gravity contour maps of a region specified by each subject, three of which have had their workflow altered in such a way as to corrupt the final result. The scientists are asked to identify the correct map using only Probe-It! as a resource. If the subjects can both identify the correct map and indicate why the other candidate maps are unsatisfactory, then we can claim that our tool provides both a comprehensive and digestible trace of a workflow execution. If the subjects fail to identify the correct map or error source, we can at least get an insight as to what additional functions might

have facilitated success and integrate those missing features into our tool.

Visualization Tool for Oracle Spatial

By Yinghui Li[1]

[1]Department of Computer Science, San Diego State University, San Diego, Calif.

This is a teaching tool that is developed for the users who are willing to learn Oracle Spatial Database to understand the concepts and the formats of Oracle Spatial Database. The user interface of the system is written in Java, including Java2D. The back-end of the system is using the Oracle 9i Spatial Database server. Oracle Java Database Connectivity (JDBC) Thin driver is used to establish the database connections.

There are four modes in this tool: Spatial Data, Spatial Query, User Data, and User Query. This tool has online help files that introduce how to use these four modes. Users can practice using the tutorial modes (Spatial Data and Spatial Query modes) and then create their own spatial tables, insert their own spatial objects, and do some spatial analysis in User Data and User Query modes. Advanced users could use this visualization tool to show their graphics. In this tool, users can delete tables, modify data, load data from text files, zoom in and zoom out the graphics, print preview and print the graphics, and select an object from a list for identification.

Support Collaboration Geosciences Research Through the GEON Portal

By Kai Lin[1] and Chaitan Baru[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

Successful collaborations are crucial to geosciences research. In this demo we will demonstrate all the aspects of the MyProjects system in the Geosciences Network (GEON) ortal for supporting geosciences collaborations. The goal is motivated by the observation that most geosciences projects have multiple researchers, often at multiple sites with different schedules. It is therefore difficult to share information, coordinate tasks, and maintain consistency. Many successful experiences have proved that modern information technologies, such as forum, wiki, and many other tools, offer opportunities to reduce this difficulty. We seek to organize these cooperative tools in a better way with other cyberinfrastructure technologies to support distributed cooperative work in a geosciences environment.

MyProjects is a system in the GEON portal which allows users to create new projects and add other people into the project teams. The notion of project is a very flexible concept. It supports not only simple personal tasks but also large organizational tasks. Four predefined roles in the system are listed as follows, based on decreased privileges: leader, contributor, observer, and guest; that is, a leader can do whatever a contributor can do in a project, and so on. A specific role is assigned to each team member.

By default, each project consists of nine main components: configuration management, team, a news board, wiki, discussion forums, to-do lists, project planning, problem reporting, and some space for sharing resources. Each part has several operations that can be applied to it, and each operation is associated with a lowest role. A team member can perform an operation only if the member's role is not lower than the lowest role of the operation. For instance, four operations can be applied to a project team: view team members, view team member e-mail addresses, modify team, and modify team member role. By default, the lowest role associated with viewing team member e-mail addresses is observer; therefore, a guest will not be able to see e-mail addressed to other team members.

The configuration management component provides the capability of customizing project components and changing the lowest role of each operation for the selected components. For example, a project may decide not to have any forum and wiki and allows only leaders and contributors to view team member e-mail addresses.

Adding new members to a project team can be done by sending an e-mail invitation or by inputting people from other project teams. Once a person accepts the invitation, that person is assigned an initial role as a contributor. But this can be changed by a person who can modify the team.

Project news, including news drafts, unreleased news, and current news, can be posted on the news board. By default, only project leaders can delete news; contributors can add news to the board, edit existing news, and view unreleased news; observers can check archived news; and guests can only read the current news. Users also can choose whether to send released news via e-mail to all project members when the news is posted.

The project wiki is an effective place for mass-collaborative authoring that allows contributors to add, remove, and edit, content with links to other resources. Project forums build some online discussion groups for contributors to exchange open messages on any interesting topics.

Project task management is supported by three components: to-do lists; project planning; and problem reporting. A to-do list usually contains several tasks that need to be done. The tasks can be closed once they are completed. Important dates in a project's timeline can be set up as milestones by using the project-planning component, so that every contributor knows what needs to be done by what time. Troubles with data quality, hardware, and software in a project can be reported through the problem report component. When a problem is reported, a team member can be assigned to solve the problem. A problem can be closed after it is solved by reporting what has been completed.

MyProjects provides some storage space for each project to share resources with other project members. Contributors can create folders and upload files into the folders. The system‘s built-in version-control system lets contributors upload new versions to existing files saved in the system. Resources found by GEON search also can be saved in the project space.

A search engine is available in MyProjects to find news, discussion messages, and resources in project space. A user can choose to search in all the participated projects or just within a single project. Some MyProjects functions, like computational job monitoring and resource integration tools, are still under design or development.

Managing a Parameter Sweep for Earthquake Simulation Research

By Choonhan Youn[1], Tim Kaiser[1], and Dogan Seber[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

A Grid middleware consists of a set of tools and technologies that allows users to access Grid resources and applications using a common set of protocols and services. It also provides seamless computing and information environments for science and engineering communities. Within such a Grid computing environment, a Grid portal provides a user interface to a multitier Grid application development stack, which is the Grid middleware. The Geosciences Network (GEON) infrastructure that is naturally distributed with users and resources spread across 16 different partner sites in the U.S., provides portal, middleware, and data resources to facilitate scientific discovery using applications, tools, and services for domain scientists through a cyberinfrastructure. It consists of both service-oriented Web/Grid framework and application toolkits, using a Web service model and a portlet programming model to represent applications. Based on those grid environments, we have developed the Synthetic Seismogram (SYNSEIS) toolkit to be a computational platform that facilitates synthetic seismogram calculations in two-dimensional and three-dimensional media for research and educational purposes (fig. 1). It is a Web-based simulation and analysis system and is one of the services provided within the GEON network. SYNSEIS

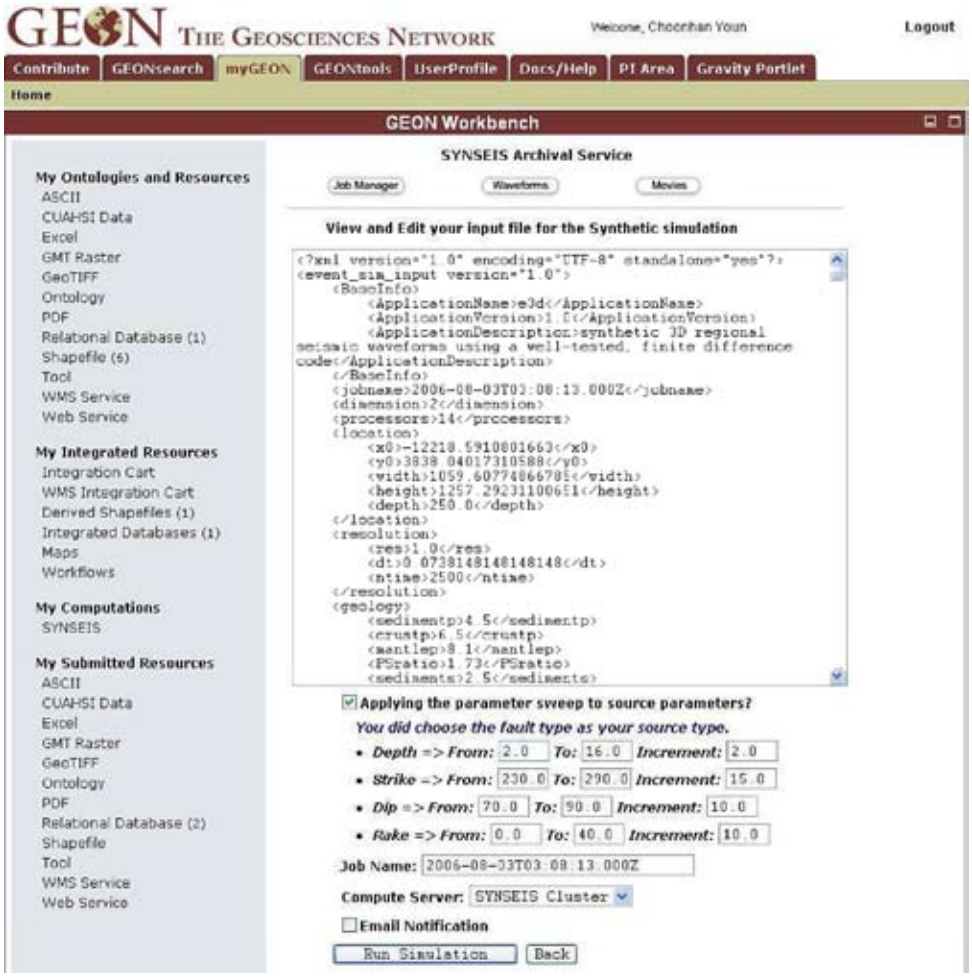

Figure 1. GEON—The Geosciences Network Workbench Web page.

provides information management, computational analyses, and simulations all wrapped into Web services and Grid computing. The E3D simulation software used by SYNSEIS is designed to simulate seismic wave propagation through the Earth's interior. We grid-enabled E3D using our own dialect Extensible Markup Language (XML) inputs, running crustal models through Web services. The XML inputs for this application include: structural information, which contains cell dimension; number of time steps; source parameters; geology; and number of stations. We demonstrate how one can use a simple management scheme to perform a parameter sweep and spread the work across a collection of computational resources, using an application that was not specifically designed to perform parameter sweeps. In particular, we identify the earthquake simulations in SYNSEIS as an example application that can benefit from running on multiple computational resources, and subsequently promote the sharing of computational resources among partner sites involved in the GEON project.

GeoSciML Testbed 2: Demonstrating the Exchange of Geologic Map Information Using a Common Data Transfer Schema and Open Geospatial Consortium Technologies

By Boyan Brodaric[1], Bruce R. Johnson[2], Francois Robida[3], and International Union of Geological Sciences–CGI Interoperability Working Group[4]

[1]Geological Survey of Canada, Ottawa, Ontario, Canada

[2]U.S. Geological Survey, Reston, Va.

[3]Bureau de recherches géologiques et minières (BRGM), 3-Avenue Guillemin, Orleans, France

[4]International Union of Geological Sciences (IUGS)–Interoperability Working Group of the Commission for the Management and Application of Geoscience Information (IUGS-CGI), Ottawa, Canada

Geoscience Markup Language (GeoSciML) is a standard data schema for exchanging geological features using Open Geospatial Consortium (OGC) Web service technologies and standards. It is being developed via open participation in the international community by the Interoperability Working Group of the Commission for the Management and Application of Geoscience Information (CGI), a commission of the International Union of Geological Sciences (IUGS). GeoSciML 1.0 is an application of GML (Geography Markup Language) and is derived primarily from the North American Digital Geologic Map Data Model (NADM, http://www.namd-geo.org) and the eXploration and Mining Markup Language (XMML, http://xmml.arrc.csiro.au/). It is being considered as the exchange format for OneGeology, a collaborative project to deliver, via the Web, geologic maps for the World at scales of about 1:1,000,000.

GeoSciML 1.0 has recently been evaluated in a second international testbed. Data providers from eight agencies in six countries (Canada, United States, United Kingdom, France, Sweden, and Australia) participated in this testbed. Geologic map information was served by each agency in GeoSciML 1.0 format, using OGC Web Mapping Service (WMS) and Web Feature Service (WFS) standards. For the testbed, multiple Web clients carried out three predefined use cases: (1) viewing geologic maps from several data providers and downloading geologic data as GeoSciML for a user-selected feature; (2) querying multiple maps and downloading geologic data as GeoSciML for multiple features; and (3) reclassifying multiple geologic maps using a common classification for each of the GeoSciML attributes of geologic age and rock type, and displaying the result as a derivative map.

A live demonstration of the testbed will use both Web-based and desktop clients and present the three use cases from multiple data sources. In addition to illustrating the benefits of a standard data transfer schema, the demonstration will also highlight the need to develop semantic approaches to reconcile heterogeneous data content.

Interactive Immersive Visualization of Geoscience Data

By Oliver Kreylos[1], Gerald W. Bawden[2], Magali I. Billen[3], Eric Cowgill[3], Bernd Hamann[4], Margarete A. Jadamec[3], Louise H. Kellogg[3], Oliver G. Staadt[5], and Dawn Y. Sumner[5]

[1]Department of Computational Science and Engineering, University of California—San Diego, La Jolla, Calif.

[2]U.S. Geological Survey, Sacramento, Calif.

[3]Department of Geology, University of California—Davis, Davis, Calif.

[4]Department of Computer Science, University of California—Davis, Davis, Calif.

[5]Department of Geology, University of California—Davis, Davis, Calif.

The geosciences are increasingly challenged to manage, process, visualize, and interpret the large quantities of data generated by high-accuracy, high-resolution imaging and sensing technologies, or large-scale computer simulations of complex phenomena. We are developing and using interactive visualization software to view, interact with, and manipulate observed and (or) simulated geophysical, geodynamical, and geologic data. The innovation of our approach is the highly effective use of human interaction in immersive three-dimensional virtual reality (VR) environments. While immersive (head-tracked stereoscopic) visualization allows us to detect features in large and complex data more effectively, interactive tools substantially simplify the construction or manipulation of three-dimensional shapes to isolate and identify those features, and to perform quantitative measurements of structures emerg-

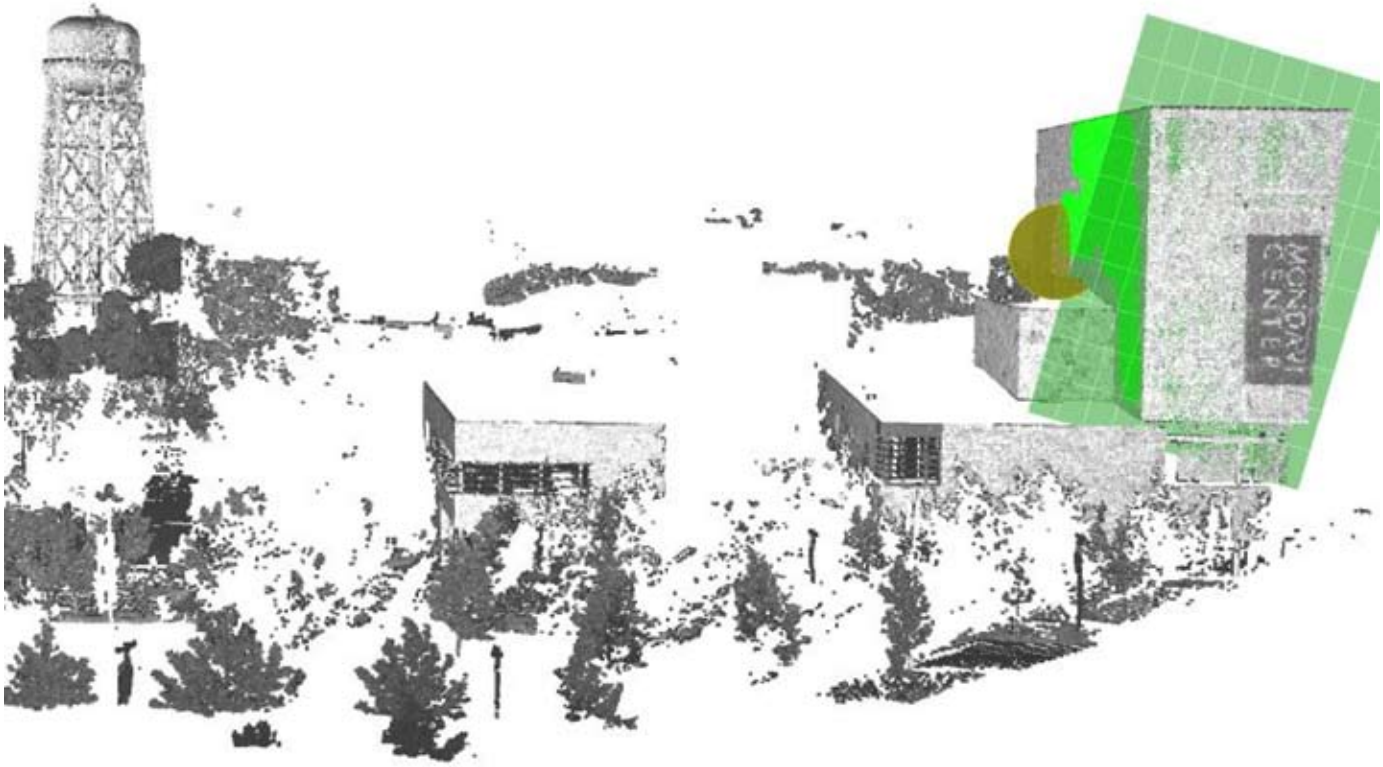

Figure 1. Screen shot from the LiDAR viewer, showing a part of University of California—Davis campus (including the water tower and the Mondavi Center for the Performing Arts). The image shows how a user can select features by using a "three-dimensional paint brush" (selected points are highlighted in green), and can quantify the position/orientation of features by, for example, extracting equations of best-fit planes (planes are visualized as green transparent rectangles).

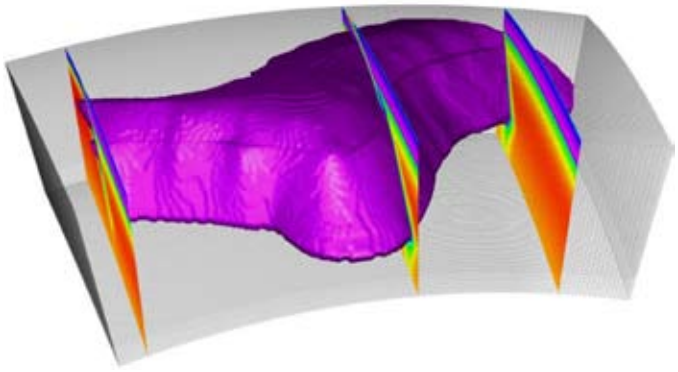

Figure 2. Screen shot from the gridded volumetric visualizer, showing an early model of a subduction zone underneath Alaska. The subducting slab was isolated by an isosurface of viscosity, and structures in the surrounding mantle are visualized using color-mapped slices at user-defined positions and orientations.

ing from the data. We tailor our visualization approach to the specific scientific problems to take full advantage of each visualization method's strengths, using both three-dimensional perception and interaction with data and ongoing simulations, to fully utilize the skills and training of geoscientists exploring their data in VR environments. In the process, we are developing a suite of tools that are adaptable to a broad range of scientific and engineering problems. We demonstrate our approach on several geophysical and geological datasets, including tripod-based and airborne light detection and ranging (LiDAR) data, seismic tomography, geodynamics computational models, volumetric data from thin serial sections of Archean rock samples, and earthquake hypocenters. A LiDAR viewer (see fig. 1) allows a user to be fully immersed in point cloud data to assess data quality and to analyze complex targets that are hard to identify in standard, nonimmersive visualizations. Software mapping tools allow "virtual field studies" in regions that are otherwise inaccessible to human geologists. A visualization software for gridded volumetric data (see fig. 2) focuses on exploratory data analysis by supporting real-time extraction of derived shapes, such as color-mapped slices, isosurfaces, or particle traces. Interactive measurement tools then allow scientists to quantify their observations. Our software is based on a "VR operating system" that can be used with

standard computers and a range of immersive three-dimensional environments (for example, GeoWall, ImmersaDesk, and CAVE). We have demonstrated that our approaches have clear advantages over other commonly used data analysis methods. Additional information about our work can be found at http://www.keckcaves.org.

Earth Science Community Implementation Through Iteration and Testing (ITIT) Resources Through a Unified Data and Analysis Portal

By Yehuda Bock[1], Paul Jamason[1], Reuy-juin Chang[1], Feng Pang[1], Sharon Kedar[2], Danan Dong[2], and Brian Newport[2]

[1]Scripps Institute of Oceanography, University of California—San Diego, La Jolla, Calif.

[2]Jet Propulsion Laboratory, Pasadena, Calif.

We are in the process of merging the capabilities of three NASA-funded projects under the umbrella of the National Aeronautics and Space Administration (NASA) Access Project, "Modeling and On-the-fly Solutions for Solid Earth Sciences (MOSES)," to facilitate data mining and modeling of rapidly expanding multidisciplinary geoscience datasets. (1) The SCIGN (Southern California Integrated GPS Network)-REASoN project is focused on the combination, validation, archive, and delivery of high-level data products and data-mining capabilities from space geodetic measurements, in particular from over 600 Canadian Galactic Plane Survey (CGPS) stations in Western North America. (2) The QuakeSim project is developing linked Web service environments for supporting high-performance models of crustal deformation from a variety of geophysical sensors, including global positioning system (GPS) and seismic instruments. (3) The Solid Earth Natural Hazards (SENH) Research and Applications Development Program's GPS and seismic integration project has developed a prototype real-time GPS/seismic displacement meter for seismic hazard mitigation and monitoring of critical infrastructure. The focus of the MOSES project is to enable direct interaction between modelers and data or data-product providers using Web services within a unified portal architecture. Modeling applications include, for example, time series analysis of continuous and real-time data (for example, RDAHMM and st_filter programs) and fault dislocation modeling (for example, Simplex program). Community resources include access to extensive infrastructure and distributed data archive holdings, an online map server/client linked to a GIS database, a "GPS Explorer" data portal that is extensible to heterogeneous datasets, and "Geophysical Resource Web Services." We present an interactive display of the current capabilities of the unified data portal and solicit feedback from community members.

A Deployable GEON LiDAR Processing and Analysis System

By Efrat Jaeger-Frank[1], Sandeep Chandra[1], Christopher J. Crosby[2], Viswanath Nandigam[1], Ashraf Memon[1], J. Ramon Arrowsmith[2], Ilkay Altintas[1], and Chaitan Baru[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[2]School of Earth and Space Exploration, Arizona State University, Tempe, Ariz.

Distribution, processing, and analysis of large light distance and ranging (LiDAR, also known as airborne laser swath mapping (ALSM)) datasets push the computational limits of typical data distribution and processing systems. The high point density of LiDAR datasets makes processing difficult for most geoscience users who lack the computing and software resources necessary to handle these massive data volumes. Over the past two years, as part of the Geosciences Network (GEON) project, we have developed a three-tier architecture—the GEON LiDAR Workflow (GLW)—to facilitate community access to LiDAR datasets. The GLW uses the GEON portal, a workflow system based on the Kepler scientific workflow environment, and a set of services, for coordinating distributed resources using emerging Grid technologies and the GEON-Grid clusters. The GLW is available to the community via the GEON portal and has proven itself as an efficient and reliable LiDAR data distribution and processing tool.

The increasing popularity of the GLW has led to several requests for deployment of the system to additional sites and projects. We are currently in the process of creating an automatic deployment of the system that requires a minimal amount of user intervention (known as a "roll"). As an initial phase, we have replicated the processing services originally deployed on a GEON cluster at Arizona State University, to additional compute clusters at the San Diego Supercomputer Center and at the University NAVSTAR (Navigation Signal Timing and Ranging) Consortium (UNAVCO) in Boulder, Colo. With the GLW deployed on multiple processing clusters, we can improve the system load balancing and provide a failover site. We further plan to enhance the system by utilizing a Grid scheduler to map jobs onto the Grid clusters by taking into account the availability of the corresponding compute, storage, and networking resources. Deploying the GLW on distributed sites also imposes additional requirements on the system for increased robustness, and more system monitoring information. For example, users are interested in tracking the execution state of their LiDAR processing job in real time. A number of new features were added to the GLW to address these requirements. We use the Kepler "provenance" capability, which collects job provenance data, to enhance the GLW's job-monitoring interface to provide users with live job status monitoring. The data provenance is also useful when publishing results and sharing GLW products among scientists. With

these enhancements, we expect to make the GLW more robust and useful to a wide range of earth science users. The GLW is currently deployed on 3 sites, and has 126 users who have submitted a total of 1,250 LiDAR processing requests for a total of over 500 gigabytes of data.

LiDAR-in-the-Box: Serving LiDAR Datasets Via Commodity Clusters

By Viswanath Nandigam[1], Chaitan Baru[1], Sandeep Chandra[1], and Efrat Frank[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

The Geosciences Network (GEON, http://www.geongrid.org) is an National Science Foundation-funded project to create an information technology infrastructure that facilitates a collaborative, interdisciplinary science effort in the field of earth sciences. GEON facilitates data registration, ingestion, and integration of a range of geoscience data types, including LiDAR (light distance and ranging) data. LiDAR datasets can be used to create high-quality digital earth-surface models, which are useful in a variety of geoscience and geospatial applications. The recent, rapid increase in the rate of acquisition and popularity of these datasets far outpaces the resources available to most geoscientists for processing and using these data.

GEON provides a novel approach for processing and distributing LiDAR datasets and derived products using a high-performance backend database machine, a portal as the front-end user interface, and the Kepler scientific workflow system for managing the computations. Currently, the LiDAR datasets are stored in an IBM DB2 database running on one of the nodes of DataStar, an IBM supercomputer system that is one of the computational resources in the TeraGrid. This node is linked to a large disk subsystem via a high-end fibre channel link. This machine configuration is well suited to handle the massive amounts of LiDAR data, which frequently exceed several millions of data points per dataset.

We propose a new approach to hosting LiDAR data based on commodity clusters, which can provide a better price/performance solution. This is achieved by taking advantage of DB2's "partitioned database" feature. In this approach, the LiDAR database tables would be partitioned across multiple machines or "nodes." Each partition is managed by an independent database manager, each with its own data, configuration files, indexes, and transaction logs. This architecture provides better scalability; new machines can be added to the complex and the database can be expanded across them. In this paper, we describe this new parallel database architecture for hosting LiDAR data. We refer to this system as the "LiDAR-in-the-Box" because one of the benefits of this approach is that individual researchers will be able to deploy such a system at their sites. We will describe the approach that will be used to make the LiDAR in the Box easily deployable.

A Petascale Cyberfacility for Physics-Based Seismic Hazard Analysis

By Philip Maechling[1], Thomas Jordan[2], J. Bernard Minster[3], Reagan Moore[3], Carl Kesselman[4], and The CME Collaboration[1]

[1]Southern California Earthquake Center, Los Angeles, Calif.

[2]Department of Earth Sciences, University of Southern California, Los Angeles, Calif.

[3]University of California—San Diego, La Jolla, Calif.

[4]Information Sciences Institute, University of Southern California, Los Angeles, Calif.

Current applications of probabilistic seismic hazard analysis (PSHA) employ empirical attenuation relationships to model the propagation and attenuation of seismic waves between the source and receiver; however, these relationships cannot easily account for three-dimensional structural variations (for example, basin effects) and source complexities (for example, directivity effects). A goal of the Southern California Earthquake Center (SCEC) is to use earthquake simulations to improve PSHA. For this purpose, SCEC is deploying as part of its Community Modeling Environment (CME) a new cyberfacility (PetaSHA) that can execute PSHA computational pathways and manage data volumes using the Nation's high-performance computing resources. The objectives are to extend deterministic simulations of strong ground motions above 1 hertz for investigating the upper frequency limit of deterministic ground-motion prediction; improve the resolution of dynamic rupture simulations by an order of magnitude for investigating the effects of realistic friction laws, geologic heterogeneity, and near-fault stress states on seismic radiation; and compute physics-based PSHA maps and validate them using seismic and paleoseismic data. The cyberfacility comprises several computational platforms that vertically integrate hardware, software, and wetware (technical expertise). One of these platforms, CyberShake, employs advanced workflow management tools to compute and store the large suites of ground-motion simulations needed for physics-based PSHA mapping. We are also developing a science gateway for the broader community to access the CME simulation capabilities and data products.

An Integration Scheme for Geophysical Studies of the Continental Lithosphere: An Update

By G. Randy Keller[1], Eva-Maria Rumpfhuber[1], and Aaron Velasco[2]

[1]School of Geology and Geophysics, University of Oklahoma, Norman, Okla.

[2]Department of Geological Sciences, University of Texas at El Paso, El Paso, Tex.

Modern studies of the processes that create and deform the continental lithosphere have both fundamental scientific and societal implications. With the development of ambitious projects such as EarthScope, the data emerging make it possible, in fact essential, to build three-dimensional models that have the highest spatial resolution possible, are tied to geologic constraints, and provide information on the composition and physical state of the materials that constitute the lithosphere. This ambition can obviously be best accomplished by measuring a broad range of measurements of P-wave velocity (Vp), S-wave velocity (Vs), density, magnetic properties, electrical properties, thermal properties, seismic anisotropy, attenuation (Q), temperature, and so on, for volume elements. In addition, interfaces that represent features such as stratigraphic boundaries, the Moho, faults, and boundaries of magmatic bodies and other discrete masses must also be mapped in order to properly characterize a region of the lithosphere. This goal can only be achieved through a highly integrated approach that takes advantage of all of the geological and geophysical constraints available. In most cases, controlled source and natural source seismology have the potential to provide the greatest spatial resolution of discontinuities and regions with characteristic seismic velocity, anisotropy, or Q. Because each of these types of seismic data are measured and analyzed by a variety of techniques, developing an integration scheme for seismic results is an important first step in building integrated models of lithospheric structure.

The diverse types of seismic data and analysis techniques each have their own sensitivities and spatial resolution, and when used alone, can constrain some aspects of the lithospheric structure; however, when used together with other types of geophysical and geological data, the combined approach has the potential to produce a better constrained model that also reflects multiple physical parameters. For example, controlled source experiments yield the Vp structure, and sometimes Vs structure, of the crust and uppermost mantle with the analysis of refraction and wide-angle reflection data. In particular, analysis of the PmP phase (Moho reflection) yields a good estimate of the average Vp of the crust (Vpave) for the crust. In addition to providing an independent measure of crustal thickness that complements the wide-angle data, receiver function analysis can constrain the Vp/Vs ratio utilizing full-crustal reverberation(s) from teleseismic earthquakes. Thus, a simple form of integration involves using the Vp/Vs ratio from receiver functions and Vpave from refraction measurements, to solve for the average Vs (Vsave) of the crust. When refraction and wide-angle reflection data and several receiver functions nearby are available, we have devised schemes whereby three-dimensional voxel-based models and two-dimensional models with interfaces can be derived using tomographic inversion in the first case, and ray-based techniques in the second case. In either case, gravity, magnetic, and electromagnetic data can easily add extra constraints. The ultimate goal is to add geologic and geodynamic results to make the result four-dimensional in nature.

Atlas of the Cryosphere: A Web Map Service for the Earth's Frozen Regions

By John Maurer[1]

[1]National Snow and Ice Data Center (NSIDC), University of Colorado at Boulder, Boulder, Colo.

Introduction

The National Snow and Ice Data Center (NSIDC) "Atlas of the Cryosphere" Web site (http://nsidc.org/data/atlas) allows visitors to explore and dynamically map the Earth's frozen regions (figs. 1, 2). Viewed from a polar perspective, the available data sources include snow cover, sea ice extent and concentration, glaciers, permafrost, ice sheets, and other critical components of the Earth's cryosphere. Users can zoom in to a specific region on the Earth as well as overlay country borders, major cities, and other geographic information. This site should act as a useful tool in science and education efforts surrounding the International Polar Year (IPY) and beyond by providing a geographic tool for viewing snow and ice on the planet. In addition to providing an interactive Web interface, maps and data sources contained in the Atlas of the Cryosphere are also accessible via the Open Geospatial Consortium (OGC) Web Map Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS). These international specifications provide a framework for sharing maps and geospatial data over the Internet.

This paper will provide an overview of the Atlas of the Cryosphere, describe its interoperability with other OGC-compatible software applications, as well as outline some of the "lessons learned" in developing an OGC-enabled map server from disparate geospatial data sources.

The development of this application was supported by National Aeronautics and Space Administration's Earth Observing System (EOS) program and was developed using MapServer, an open-source development environment for building spatially enabled Internet applications.

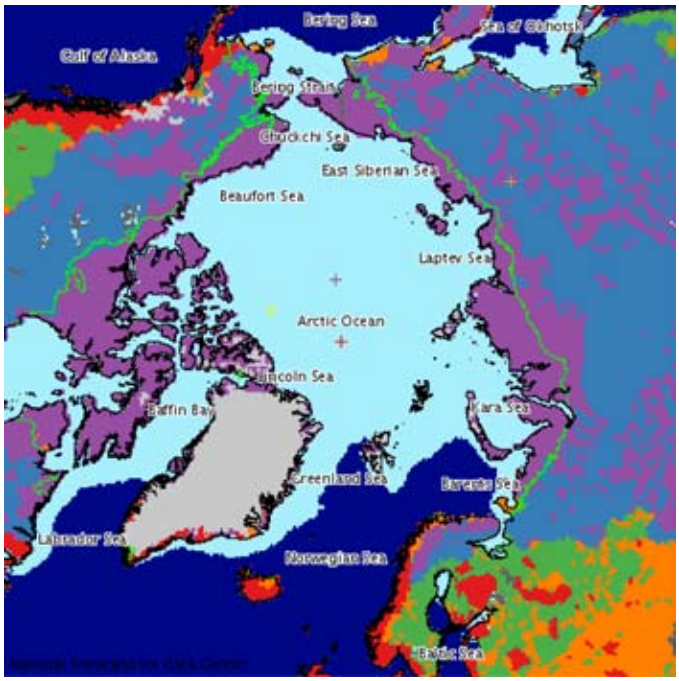

Figure 1. Example of map available on Atlas of the Cryosphere Web site (Arctic view).

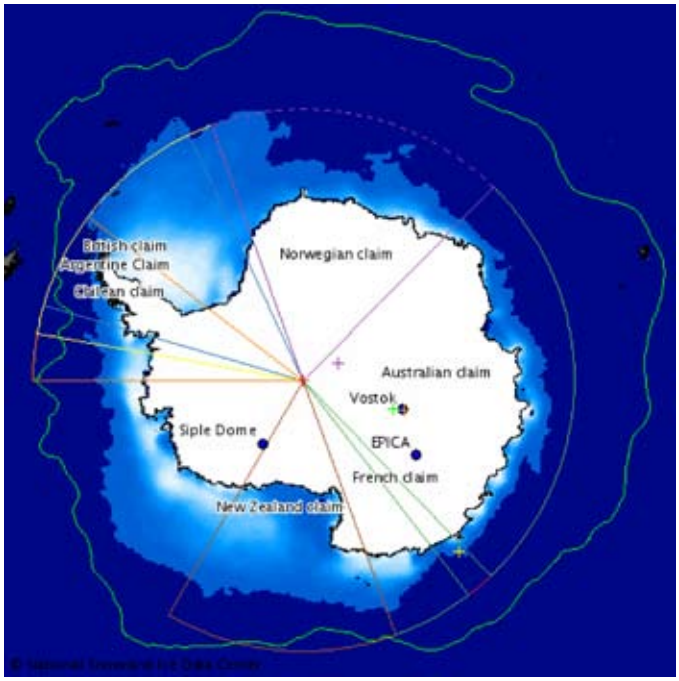

Figure 2. Example of map available on Atlas of the Cryosphere Web site (Antarctic view).

Atlas Features

The atlas allows for the following activities: (1) dynamically visualizes Earth's snow and ice; (2) explores the planet from a polar perspective for both the northern and southern hemispheres; (3) customizes maps by zooming in and out and by selecting from a variety of basemaps and overlays; (4) views monthly climatologies of snow and sea ice to see how and where the cryosphere shrinks and grows over the course of a year; (5) looks up definitions for unfamiliar cryospheric terms; (6) accesses maps and source data through WMS, WFS, and WCS; and (7) provides Really Simple Syndication (RSS) feed for keeping informed about atlas additions and modifications.

Selectable Parameters

The following parameters may be selected:

1. Cryosphere-related parameters such as glacier locations; glacier outlines; ice core locations: ice sheet accumulation; ice sheet elevation; permafrost classification; permafrost extent; sea ice concentration; sea ice extent; seasonal snow classification; snow extent; snow water equivalent; treeline (northern limit of forests); and more.
2. Other parameters such as Antarctic Circle; Arctic Circle; cities; countries; Equator; geographic features (land, sea, and ice); International Date Line; latitude and longitude; North Pole; South Pole; Tropic of Cancer; Tropic of Capricorn; U.S. states; and more.

Open-Source GIS

NSIDC has leveraged use of the following open-source software packages for the development of the atlas and for manipulation of its diverse data sources, such as reformatting and reprojecting:

- MapServer—A development environment for building spatially enabled Internet applications.
- Geospatial Data Abstraction Library (GDAL)—A data translation and processing library for raster geospatial data formats.
- OGR Simple Features Library—A data translation and processing library for vector geospatial data formats.
- PROJ.4—A cartographic projections library for enabling reprojections.
- libgeotiff—A GeoTIFF library for reading and writing GeoTIFF information tags.

Interoperability

Maps can be generated through the provided Atlas of the Cryosphere Web interface. Alternatively, interoperable and customizable data access to Atlas maps and source data are enabled via the following Open Geospatial Consortium, Inc. (OGC) specifications:

- Web Map Service (WMS): Provides map and legend images of selected data layers and base maps.
- Web Feature Service (WFS): Provides vector source data in Geographic Markup Language (GML) format.
- Web Coverage Service (WCS): Provides raster source data in GeoTIFF format.

Using these services, maps and data can be accessed for your spatial region of interest at your desired resolution or output size. These services are accessible through the construction of a URL string that contains certain required and optional parameters for customizing output. Several OGC-compatible clients are also available for handling these interactions through point-and-click graphical user interfaces (GUIs), such as MapServer, ArcGIS 9 Service Pack 2, ENVI Zoom 4.3.1, Google Earth 4, uDig, QGIS, and GRASS, to name just a few. Client support for OGC services is constantly improving and expanding as popularity for these specifications increases.

Remote access through a client is normally obtained via an OGC "GetCapabilities" URL, as in the following examples for the northern and southern hemispheres of the atlas, where the "service" parameter can be WMS, WFS, or WCS:

http://nsidc.org/cgi-bin/atlas_north?service=WMS&request=GetCapabilities

http://nsidc.org/cgi-bin/atlas_south?service=WMS&request=GetCapabilities

Other possible OGC requests include

- GetMap—Get an image of selected data layers.
- GetLegendGraphic—Get a map legend for selected data layers.
- GetFeature—Get vector source data in Geography Markup Language (GML) format.
- GetCoverage—Get raster source data in GeoTIFF format.

These requests can include various standardized options for customizing output, including the ability to limit results to a particular spatial region. For more information, visit http://opengeospatial.org. A summary for applying these requests to the Atlas of the Cryosphere is also available at http://nsidc.org/data/atlas/ogc_services.html.

Lessons Learned

European Petroleum Survey Group (EPSG) Codes

In order for users to access maps and source data remotely via OGC Web Services (OWS), the application must list the available map projections for these services. These map projections must be specified as EPSG Codes. These codes are developed and maintained by the EPSG—now known as the International Association of Oil and Gas Producers (OGP)—and are a shorthand way of specifying a map projection and other coordinate parameters. Unfortunately, requiring the use of EPSG codes limits your OWS to the map projections that are already defined in the EPSG Geodetic Parameter Dataset. This can be especially limiting for polar projections, very few of which are currently available. As a result of this limitation, NSIDC submitted a request to have several new polar projections added to the EPSG Geodetic Parameter Dataset. This request was granted and eventually incorporated into the latest release (version 6.12; see EPSG Codes 3408-3413); however, these new codes can take several months to be incorporated into the various open-source applications and into the OWS clients that use them.

Custom Ellipsoids and Datums

Several of NSIDC's datasets are projected on two somewhat unconventional datums, neither of which is available in popular image processing or GIS packages such as ENVI and ArcGIS, often leading the georeferencing information to be ignored when opening GeoTIFFs that are in coordinate systems using these datums. This is something to consider when attempting to distribute geospatial data that will be easily readable for users without extra configuration to be handled on their end. If there is not a strong justification for using an unconventional datum, one should consider reprojecting onto something like WGS84 prior to distribution, which is especially popular and appropriate for global- or large-scale earth- science applications.

Wraparound and Overlapping the Pole

There are particular issues that need to be addressed when attempting to display global data in polar projections. If you have a dataset that is in a latitude and longitude (Plate Carrée) projection or other such global projections, there can be problems when attempting to display these data in a polar projection on the fly. Namely, if you have data that overlap the ±180 degree latitude line, they will likely be distorted in a polar projection because of wraparound at this boundary. In addition, data that overlap the pole will also likely be distorted, and an annulus will likely prevent data from being viewed directly at or near the pole itself. A work-around is to reproject and store your global source data in a common polar projection prior to displaying them in a polar-projected MapServer application so that the reprojection is not done on the fly.

Optimizing MapServer Performance

For improved speed and to avoid unexpected artifacts such as those mentioned above, reproject all data sources into identical or very similar projections ahead of time rather than relying on MapServer to do this on the fly. When accessing large raster data files, use a tiling scheme to improve access speed, as was done for NSIDC's 250-meter resolution Moderate Resolution Imaging Spectroradiometer (MODIS) Mosaic of Antarctica (MOA) product. This enables the map server to access smaller subsets of the larger file at any given time rather than always needing to access the entire file. Lastly, while incorporating external OWS services into your application promotes interoperability, it can noticeably slow perfor-

mance, depending on the remote server and how greatly their projections differ from that of your own application.

Conclusions

Between the public release of the Atlas of the Cryosphere on February 2, 2007, and the end of March 2007, there have been 1,548 unique visitors to the site, already demonstrating its popularity. This project has been an attempt to bring together the important large-scale, climatological features of the cryosphere into a single, user-friendly, and Web-accessible interface that not only provides dynamic visualization but also a flexible and interoperable means for obtaining the source data for these features as well, thereby going beyond the criticism that many scientists have about applications that merely provide "pretty pictures." By providing a simple Web interface, this and other map server applications are also more broadly accessible to the general public, compared to other geospatial applications that may require special software to download and (or) broadband Internet access such as Google Earth. Additional cryospheric datasets, supported polar EPSG Codes, and interface features will be developed according to time and demand. If you have questions, comments, or suggestions, please contact NSIDC User Services at nsidc@nsidc.org.

Building the Interface Facility for Centimeter-Scale, 3D Digital Field Geology

By John S. Oldow[1], Charles M. Meertens[2], Carlos L.V. Aiken[3], J. Douglas Walker[4], and J. Ramon Arrowsmith[5]

[1]Department of Geological Sciences, University of Idaho, Moscow, ID

[2]University NAVSTAR (Navigation Signal Timing and Ranging) Consortium (UNAVCO), University of Colorado, Boulder, Colo.

[3]Department of Geosciences, University of Texas, Dallas, Richardson, Tex.

[4]Department of Geology, University of Kansas, Lawrence, Kans.

[5]School of Earth and Space Exploration, Arizona State University, Tempe, Ariz.

The acquisition and analysis of field data is the cornerstone of the solid earth sciences and these data provide the context and content for most inquiries in geological sciences. Geoscientists have traditionally collected these data by analog methods (for example, pencil, pen, and paper). A fully digital approach is critical, however, so that information is presented in a geospatially referenced frame for data transfer, scale and projection manipulation, and registration between datasets; this approach requires a cyberinfrastructure for the earth sciences. Geoscientists of the future will collect field data in a digital environment, which will consist of the following: equipment for acquiring information in a digital and georeferenced format; software that allows seamless transfer, integration, and exploration of data, and; online processing for manipulation of large datasets generated or to be used in the field. To implement the transition from the analog-dominated system of today to the digital future, we propose to build a facility consisting of a pool of shared hardware and software supported with instruction by expert users. This facility will allow users to build and work in a digital environment that represents the Earth's surface or landscape in three-dimensional with accuracy approaching the scale of a centimeter.

Lithospheric Structure of Northern Africa and Western Eurasia

By Minoo Kosarian[1] and Charles Ammon[1]

[1]Geosciences Department, The Pennsylvania State University, State College, Pa.

Although much progress has been made over the last few decades towards understanding the structure of the Earth, many questions regarding the details of Earth's lithospheric structure remain unanswered. The primary goal of this study was to gain a better understanding of upper and lower continental crustal composition and structure to improve our knowledge of the tectonic evolution of the Earth. To accomplish this goal, we focused on the estimation of first-order seismic structure using receiver functions, and the construction of a library of shear-velocity structures in the vicinity of seismic stations across western Eurasia and northern Africa using receiver functions and tomography-based surface-wave dispersion estimates.

We used 171 stations recording a total of about 6,000 teleseismic events producing more than 100,000 seismograms. The distribution includes 78 stations in the Middle East and Asia, 57 stations in Europe, and 36 stations in central and northern Africa. We have examined receiver functions for 119 stations with the best data for the period of 1990-2004, and applied the receiver function stacking procedure of Zhu and Kanamori (2000, JGR) to estimate Poisson's ratio and crustal thickness. The structures are classified into five tectonic environments—explicitly shields, platform, Paleozoic orogenic belts, Mesozoic to Cenozoic orogenic belts, and rifts based on Condie's (1989) global classifications. The results show a slightly lower value of Poisson's ratio $\sigma = 0.25$ for shields compared to the orogenic-belts with $\sigma = 0.26$. Crustal thickness ranges from 32 to 47 kilometers (km) with an average of 38 km and a standard deviation of 3 km for the shields. The less well sampled platforms show a wider distribution of crustal thickness, ranging from 30 to 58 km with an average 42 km and a standard deviation of 9 km. Orogenic regions show the largest variation in crustal thickness with values from 20 to 55 km and standard deviations in the range of 8 to 10 km. We combined observations obtained in this study with receiver functions results from other published analysis. In total, we

have integrated observations from 606 stations located in different geologic settings. The compiled results show a value of $\sigma = 0.26$ for Poisson's ratio and crustal thickness (H) = 39 km for crustal thickness in shields and platforms, and $\sigma = 0.26$-0.27 with H = 35-37 km for the orogenic belts.

Reference Cited

Zhu, H., and Kanamori, H., 2000, Moho depth variation in southern California from teleseismic receiver functions: Journal of Geophysical Research, v. 105, p. 2969-2980.

Deploying Data Portals

By Sandeep Chandra[1], Kai Lin[1], and Choonhan Youn[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

A fundamental objective of the Geosciences Network (GEON) project (http://www.geongrid.org) is to develop data-sharing frameworks, and in the process, to identify best practices and develop capabilities and tools to enable advances in how geoscience research is done. The GEON portal framework, which plays a key role in facilitating this objective, is implemented using GridSphere and a portlet-based framework that provides a uniform authentication environment with access to a rich set of functionality, including the ability to register data and ontologies into the system; smart search capability; access to a variety of geoscience-related tools; access to Grid-enabled geoscience applications; and a customizable private work space from which users have access to the scientific workflow environment, which allows them to easily author and execute processing "pipelines" for data analysis, modeling, and visualization. In this paper, we describe the modular hardware and software components that we have developed, which make it possible to easily deploy a data-sharing portal environment for any application domain.

In practice, deploying a portal framework like GEON portal requires an understanding of the various portal software components and their dependencies that go into engineering such a system. In addition to sites within the GEON network, the GEON software infrastructure is increasingly being adopted in other projects, such as the Chesapeake Bay Environmental Observatory (CBEO), the Network for Earthquake Engineering Simulations (NEES), an Archeoinformatics project, and the National Ecological Observatory Network (NEON). Based on this experience, we have developed a modular packaging of the various components of the system to allow easy installation and configuration of the hardware, middleware, and other software components.

The data portal infrastructure consists of the following components:

1. A Portal Server, which runs the portal software. The nominal system is a "rack-mounted," server-class machine with 750 gigabyte (GB) raw disk (5 x 146 GB hot-swappable Serial Attached SCSI (SAS) drives), dual core 3.0 GHz Intel Xeon processors, 4 GB of random-access memory (RAM), dual gigabit network interfaces, and redundant power supplies. The portal server runs the GEON software stack, including the portal software, and provides connectivity to and interoperability among the other GEON systems.

2. A Data Server, which provides storage and other data management services. This system is also a "rack-mounted," server-class machine with dual core 3.0 gigahertz Intel Xeon processors. It includes 5 x 300 GB Redundant Arrays of Inexpensive Disks (RAID) SAS drives for a total of 1.5 terabytes of raw drive space. The data nodes are configured with RAID disks in order to deal with unforeseen disk failures.

3. A Certificate Authority (CA) Server, which manages user accounts. The CA server system has the same basic configuration as the Portal Server, but with 2 GB RAM and the disk size reduced to about 36 GB, since the CA Server tasks are not input-output intensive.

The Portal Server runs the Rocks cluster management software and a standardized "GEON software stack," which includes the GEON Portal and its dependent libraries, including software tools developed in the GEON project. The so-called "core" portal functionality is generic (for example, search and data ingestion capabilities) and can mostly be leveraged "out of the box" by other projects. The Data Server provides the capability to host data registered through the portal and also provides other data management services. The Data Server runs the SDSC Storage Resource Broker (SRB) software, which provides a number of built-in data management services. The Data Server could also host additional data management services (for instance, geographic information systems (GIS) software). The Portal Server can communicate with Web services hosted at remote locations using the standard Web service protocols.

The CA Server runs the Grid Account Management Architecture (GAMA) software, which manages user accounts through the portal. The CA Server is installed using the Rocks software and the GAMA roll. Once installed, the system is fully configured as a CA. The portal software is also preconfigured to communicate with this GAMA server for managing user accounts.

For more information, please see the following:

1. Ilya Zaslavsky, Chesapeake Bay Environmental Observatory (CBEO) Project (http://geon16.sdsc.edu:8080/gridsphere/gridsphere) (Accessed October 11, 2007.)

2. NEESit – Enabling Earthquake Engineering Research, Education and Practice (http://neesphere.sdsc.edu:8080/gridsphere/gridsphere) (Accessed October 11, 2007.)

3. National Ecological Observatory Network (NEON) (http://neon.sdsc.edu:8080/gridsphere/gridsphere) (Accessed October 11, 2007.)

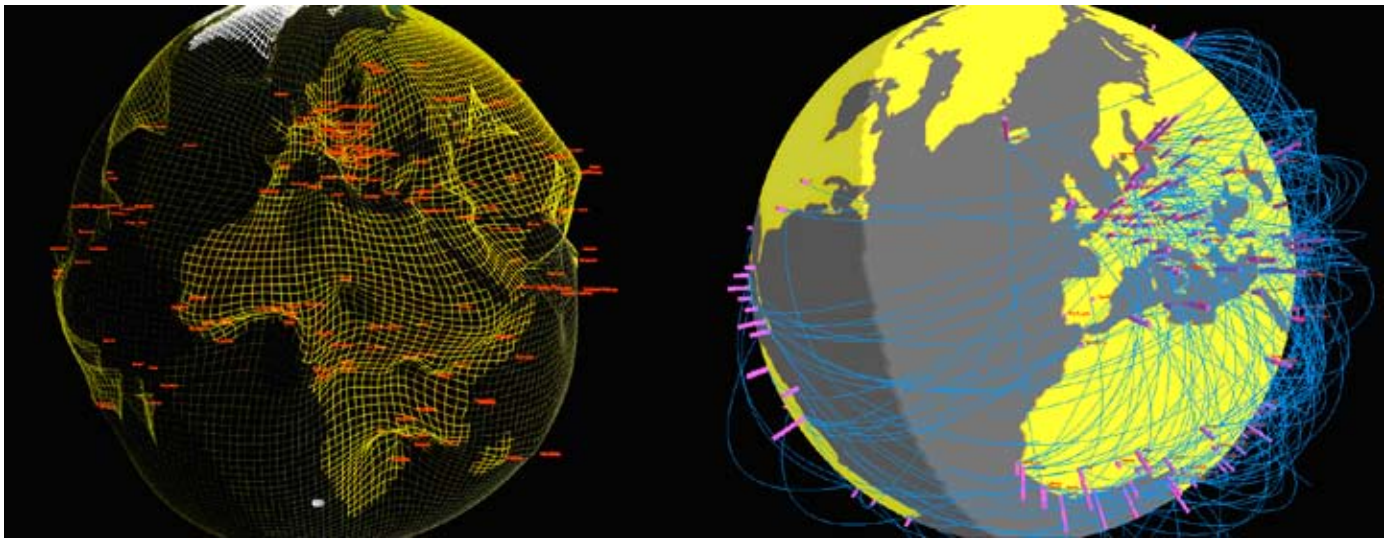

Figure 1. The diplomatic exchange data visualized with analytical tools on the right, and with global context visualization on the left.

4. Karan Bhatia, Kurt Mueller, and Sandeep Chandra, Grid Account Management Architecture (GAMA) (http://grid-devel.sdsc.edu/gridsphere/gridsphere?cid=gama) (Accessed October 11, 2007.)

Dynamic Visualization of Geographic Networks Using Surface Deformations with Constraints

By Basak E. Alper[1], Selcuk Sumengen[2], and Selim Balcisoy[2]

[1]Department of Media Arts and Technologies, University of California—Santa Barbara, Goleta, Calif.

[2]Department of Electrical Engineering and Computer Science, Sabanci University, Orhanli Tuzla, Istanbul, Turkey

This paper proposes a visualization system for geographic network datasets, which aims to convey both low-level details of the data and high-level contextual information with two different visualization modalities. The first modality, global context visualization, represents time-series spatial network data within geographic context through a real-time animation of three-dimensional map deformations. It maintains spatial framework while providing a qualitative feel of the data by only exhibiting dominant and (or) interesting features. The second modality provides a set of interactive analytical tools based on conventional node and link displays, which reveal accurate statistical details of the data on demand.

The global context visualization technique employs a modified graph drawing algorithm based on spring embedders, which position network nodes according to the time-series data being fed to it. Our contribution lies in projecting complex network datasets into a real-time virtual environment where the geographic framework is preserved. Applied constraints limit variation of network nodes by favoring inherent geographic distribution of nodes. The graph optimization solution is reached using an implicit integration scheme and allows the system to visualize data in real time. As the position of network nodes change, the surface is redrawn to fit on the new positions of the nodes. The geographic map projected on the surface deforms and enables viewers to read the data variation as a map deformation (fig. 1). This representation gives a strong qualitative impression and enables viewers to summarize the nature of the data.

The first phase of the global context visualization employs spring embedders for drawing a graph in which each geographic location corresponds to a single node and nonspatial data components correspond to relations between these nodes. The output graph visualization reflects input data by positioning related nodes closer. In this sense, our technique can be compared to force-directed placement methods; however, the proposed technique does not follow force-directed placement in any precise sense, but instead exploits its key features. The single most important distinction lies in the geographic constraints applied on the system. These constraints enable the spring-embedder system to reach a configuration that will lead to a deformed map where geographic layout is preserved to some extent for assuring intuitive recognition.

In the second phase, the surface covering the nodes is adjusted to fit on the modified positions. As a result, the geographic map projected on the surface deforms and highlights variations in the data. This approach exploits a priori knowledge of the viewer about the physically accurate version of the map. In other words, map deformation facilitates comprehension of nonspatial variables with respect to the geographic framework. Once users have a sense of the overview of the data, they can dig into the details by using analytical tools provided. Interactively responding to users, these tools expand

informative quality of the visualization through direct manipulation of visualization parameters. They can explore network data as height bar animations over three-dimensional maps or as arcs showing connections. They are able to filter data through selecting nodes or selecting the range of the displayed data.

We examined the proposed method using two different datasets. The first dataset comprises the domestic U.S. air flights among 231 airports between 1991 and 2004. The second dataset is the diplomatic exchange of data among 128 nations through years 1815 to 1966 (see fig. 1).

Geoinformatics for Geochemistry (GfG): Integrated Digital Data Collections for the Earth and Ocean Sciences

By Kerstin Annette Lehnert[1] and Sri Vinayagamoorthy[2]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, N.Y.

[2]Center for International Earth Science Information Network, Columbia University, Palisades, N.Y.

The Geoinformatics for Geochemistry (GfG) program, a collaborative enterprise by the Lamont-Doherty Earth Observatory (LDEO) and the Center for International Earth Science Information Network (CIESIN), integrates and consolidates the development, maintenance, and operation of four closely related Geoinformatics projects, comprising the digital data collections for geochemistry (EarthChem, PetDB, and SedDB), and the System for Earth Sample Registration (SESAR) that administers global unique identifiers for samples. Systems within the GfG program represent core databases for geochemistry and the broader Geosciences, enabling data stored in these databases to be discovered and reused by a diverse community now and in the future. The systems dynamically evolve in response to community needs and technical innovation, and contribute proactively to the construction of a digital information infrastructure that supports the next generation of Geoscience research and education, by establishing links to other Geoinformatics activities, and by pursuing developments in interoperability.

The GfG program provides the technical infrastructure (hardware, software, and services), the required range of expertise (a team of scientists, data managers, database administrators, Web application programmers, Web designers, and project managers) and the organizational structure for the execution of the individual project components. It is managed in a professional and sustainable environment that ensures reliable services, a high level of data quality, and the long-term availability of the datasets. All systems within the program are operated in a dynamic modus operandi, continuously responding to the needs and demands of the community and to changes in technologies, metadata and interface standards, data types, policies and procedures for data publication and data access, and organizational structures to retain their value to the science community. GfG is dedicated to educate and train the science community, as well as students and teachers, in the use of the data collections through short courses, internships, and lectures, and to advancing the establishment of a new workforce for Geoinformatics through training and education of project staff.

The GfG program includes the following elements:

- System engineering and development—Data modeling and database development; development of data submission and ingestion procedures; development of Web applications; and interoperability interfaces;
- System operation—Database administration; system maintenance, security, and backups; risk management;
- Data management—Data compilation/solicitation; data and metadata entry and quality control; user support; and long-term archiving;
- Education, outreach, and community liaison—Short courses, workshops, lectures, and exhibits; and collaboration with other Geoinformatics efforts, nationally and internationally; and
- Program and project management—Integrated management of program and its individual projects.

GfG System Infrastructure and Architecture

The core infrastructure on which the GfG systems are developed and operated is illustrated in figure 1. It consists of a set of Web, application, mapping, and database servers; all are Sun Microsystems servers. WebLogic application server

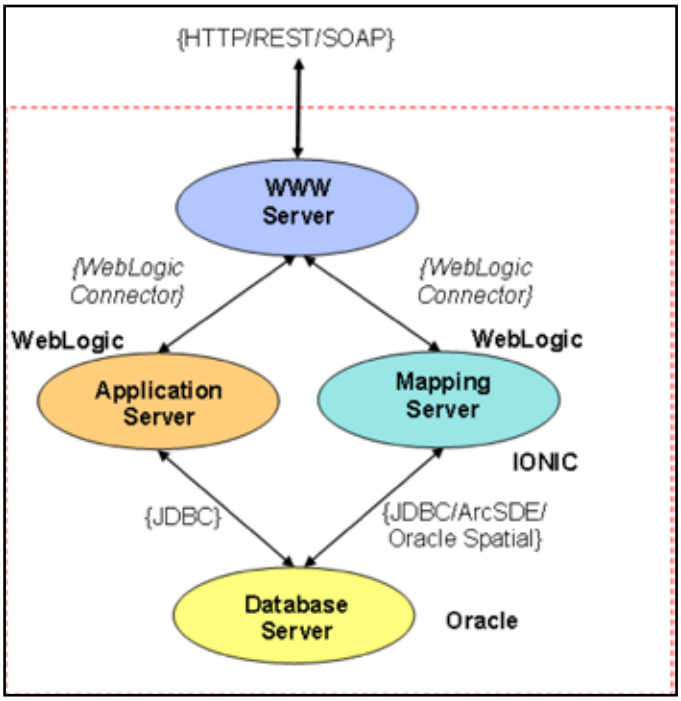

Figure 1. Infrastructure of GfG.

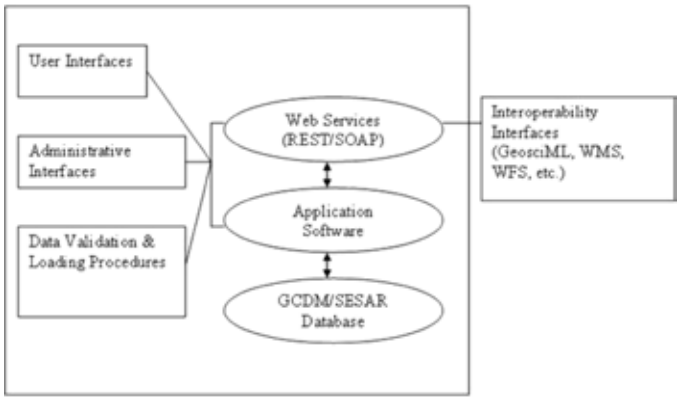

Figure 2. Service-oriented architecture (SOA) of GfG.

software is used for developing Web applications. IONIC software is used for serving geospatial data and for developing mapping and visualization tools.

Using this infrastructure, we are developing a service-oriented architecture (SOA) to implement an application framework consisting of query models and data cache that support Web and interoperability interfaces. This framework is used across each individual project, thereby making the operation and maintenance of GfG systems more efficient and sustainable.

At the core of this architecture are the following databases: the new Geo-Chemistry Data Model (GCDM) and the SESAR data model. The application software layer consisting of query model and data cache is being developed using object-oriented design (OOD) methodologies on a Java/J2EE platform supported by a WebLogic application server. Web services for querying and serving data are being developed on top of the application software layer. The user interfaces are built based on Web services, as well as on application software modules. The Web services will include interoperability interfaces to serve analytical and geospatial data related to samples to external client systems. A conceptual diagram of this service-oriented architecture is illustrated below in figure 2.

Geochemistry Data Model (GCDM)

Geochemical data served by the GfG includes a substantially broader range of measurements and materials, such as sediment cores, hydrothermal spring fluids and plumes, and xenoliths, and requires the following features to be supported by the underlying data model:

- Description of spatial and temporal components of samples and measurements (for example, depth in core, time-series and sensor measurements, and point analyses on a microprobe slide);
- Capability to store “derived” (model) types of observed values, such as age models for cores or end-member compositions for seafloor hydrothermal springs;
- Capability to track relationships between samples and subsamples,
- Ability to integrate data at any level of sample granularity; and
- Capability to accommodate analytical metadata at the level of individual measurements.

Based on these requirements, we have developed a more generic, integrative, and flexible model for geochemical data, the Geo-Chemistry Data Model (GCDM), to serve as the core data structure for our entire suite of geochemical databases (Djapic and others, 2006; Lehnert and others, 2006). This data model is compliant with standards defined in GeoSciML, a markup language developed by the IUGS Commission for Geoscience Information to represent Geoscience information associated with geologic maps and observations (Cox, 2006). Attributes in GCDM, such as method, sample, and item measured, can be mapped to corresponding types within GeoSciML; others, such as observation point or observed value, can be incorporated into the GeoSciML concepts of method, event, and measured value. We will use GeoSciML to serve geochemical data via interoperable Web services. We have presented and discussed the model with the community at various occasions (Geoinformatics, 2006, 2006 AGU Fall Meeting, workshop with the IODP Applications Development team at TAMU), and received valuable feedback and validation of the model. Updates to the model are in progress and will make it even more generic and widely applicable.

Along with the new data model, we will implement the International Geo Sample Number (IGSN), the emerging global unique identifier for samples that will allow building enhanced interoperability with other data systems at the sample level. SESAR data model is at the core of the implementation of IGSN for broad and diverse earth samples ranging from holes, cores, and dredges, to individual samples and subsamples. SESAR enables the unique identification of samples and integration of sample data from various sources and systems.

Web Interfaces

Each GfG system includes the following Web-based interfaces:

- Query and browse;
- Visualization and analysis;
- Administrative; and
- Data validation and loading.

Interoperability Interfaces

To maximize the use of the GfG data collections, we are implementing interoperability interfaces using the service-oriented architecture described above, to allow open access to the following client systems:

- Analytical Data Services: Geochemical data access via Web services that are based on the XML schema developed by EarthChem to serve complete sample data and

metadata. This schema will continue to evolve towards compliance with GeoSciML as a community standard, and will enable other systems and tools to access GfG data in real time.

- Geospatial Data Services: OGC compliant WMS and WFS services for serving sample locations. Selected data, metadata, and a link to a sample profile will be included. The service will enable any OGC compliant client to overlay, visualize, and analyze relevant geospatial data layers from multiple sources in conjunction with GfG layers.

References Cited

Cox, S.J.D., ed., 2006, Observations and measurements: Open Geospatial Consortium, Inc. document OGC 05-087r4, version 0.14.7, 168 pages, available online at http://portal.opengeospatial.org/files/?artifact_id=17038. (Accessed June 18, 2007.)

Djapic, B., Vinayagamoorthy, S., and Lehnert, K.A., 2006, Serving geochemical data using GeoSciML compliant Web service: Next step in developing a generic geochemical database model: Eos, v. 87, no. 52, Fall Meeting Supplement, Abstract IN51B–0813.

Lehnert, K.A., Vinayagamoorthy, S., Djapic, B., and Klump, J., 2006, The digital sample: Metadata, unique identification, and links to data and publications: Eos. v. 87, no. 52, Fall Meeting Supplement, Abstract IN53C–07.

OpenGIS® Web Feature Service (WFS) Implementation Specification, Version 1.1.0, Open Geospatial Consortium Inc., Document: OGC 06-027r1, Date: 2006-02-12, available online at http://www.opengeospatial.org/standards/wfs/. (Accessed June 18, 2007.)

GIS Interpolation of Stratigraphic Contact Data to Reconstruct Paleogeography: Deriving a Paleolandscape Model of the Base Zuni Sequence and Subsequent Cretaceous Islands in Central Texas

By Shane J. Prochnow[1], Melissa Mullins[2], Stephanie V. Capello[3], Anna F. Perry[3], Steven W. Ahr[3], Isaac T. Westfield[3], Kari L. Fallert[4], Song Gao[4], Kirstin T. Hartzell[4], and Kenna R. Lang[4]

[1]Center for Applied Geographic and Spatial Research, Baylor University, Waco, Tex.

[2]Center for Reservoir and Aquatic Systems Research, Baylor University, Waco, Tex.

[3]Department of Geology, Baylor University, Waco, Tex.

[4]Department of Environmental Studies, Baylor University, Waco, Tex.

Geographic information systems (GIS) raster modeling technology might constitute a quantum leap in visualization and data interpretation for paleogeographic studies. Our research uses raster modeling to interpolate paleotopography and stratigraphic thickness maps within a roughly 4,200-km^2 study area in south-central Texas (fig. 1*A*). Only readily available geospatial data was used. High-resolution geological maps (1:24,000 scale) published by the Bureau of Economic Geology (BEG) at the University of Texas at Austin were digitized and georeferenced using ArcGIS™ software products. Paleosurface modeling focused on the following four stratigraphic features: (1) basal Zuni sequence boundary set on Precambrian granite and Cambrian to Ordovician marine sedimentary rocks; (2) fluvial to deltaic Hensel Sand (Lower Cretaceous); (3) marine to tidal Glen Rose Formation (Lower Cretaceous); and (4) marine Walnut Clay (Lower Cretaceous). The basal Zuni sequence boundary is a continental-scale, angular unconformity, while the younger surfaces involved in this study are relatively conformable. Stratigraphic contact traces between these surfaces were converted to three-dimensional (3D) data by extracting elevation values from 28-m (meter) resolution digital elevation models (DEM) developed by the U.S. Geological Survey (USGS) National Elevation Dataset (NED) using ArcGIS™ 3D Analyst. The NED has a published vertical accuracy of 7 m to 14 m. ArcGIS™ Geostatistical Analyst was then used to interpolate paleosurface rasters using the simple kriging method based upon 12 nearest points along contact traces and their respective elevation values. The elevation of the interpolated raster was corrected for structural deformation since the Cretaceous using the ArcGIS™ Spatial Analyst Raster Calculator by setting the base of the Walnut Clay (mostly coincident with the upper surface of the Glen Rose) as a Z-axis datum and adjusting the upper surface of the Hensel Sand and the base of the Zuni sequence elevation, respectfully. The interpolated paleosurface rasters have estimated vertical root mean square and standard error within 5 m, roughly the same as the published NED accuracy used for the derivation of elevation. The thickness of the Hensel Sand and Glen Rose was also estimated by using the Raster Calculator to subtract their corrected bounding surface interpolations. The thickness maps were verified by using published measured section data that accompanied the original geologic maps. Three-dimensional modeling of the paleosurfaces was visualized in ArcGIS™ ArcScene. The paleorelief for the base of the Zuni sequence in the study area was about 351 m. The interpolated paleosurface grid for the base of the Zuni sequence has an estimated vertical root mean square error of 1.8 m and an average vertical standard error of 4.9 m. The total paleorelief for the upper boundary of the Hensel Sand is estimated at 356.5 m, but is locally lower in relative elevation than the base of the Zuni sequence uncon-

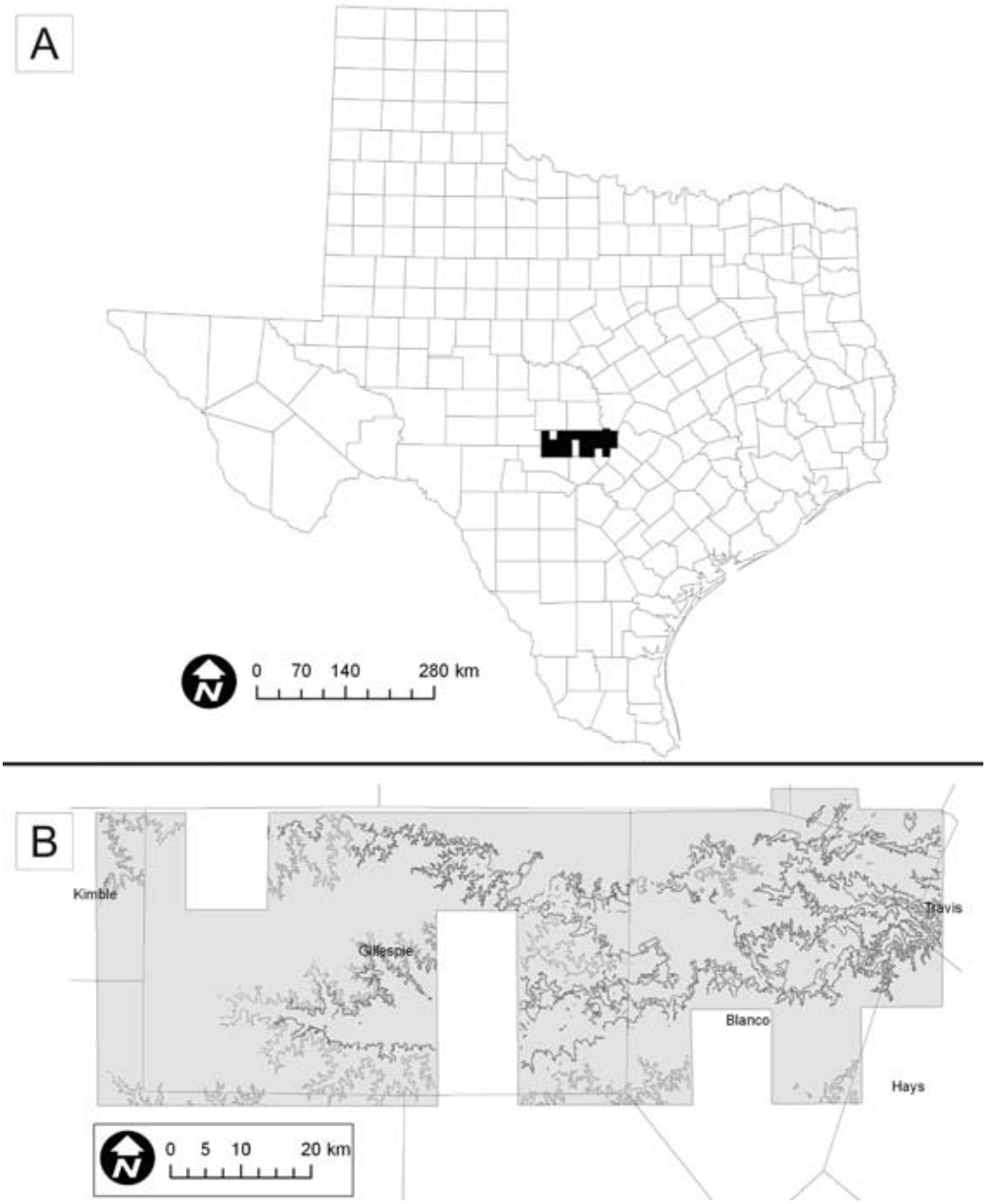

Figure 1. *A,* The study area is located in central Texas where Paleozoic and Cretaceous bedrock crops out. *B,* The coverage of published 1:24,000-scale geologic maps and the respective contact traces used for this study.

formity. The interpolated upper surface grid for the Hensel Sand has an estimated vertical root mean square error of 3.7 m and an average vertical standard error of 4.9 m. The total paleorelief for the lower boundary of the Walnut Clay (upper boundary of the Glen Rose) is estimated to have been 197.0 m, but is locally lower in relative elevation than both the base of the Zuni sequence unconformity and the upper surface of the Hensel Sand. The interpolated lower boundary surface grid for the Walnut Clay has an estimated vertical root mean square error of 4.4 m and 1.4 m of average vertical standard error. The total thickness of the Hensel Sand within the study area ranged from 0 to 96.6 m thick. The Hensel Sand is locally absent on paleohighlands (interfluves) of Paleozoic rock that constitute about 7.5 percent of the study area (fig. 2*A*). The Glen Rose ranges from 0 to 183.8 m thick in the study area. The area of zero Z-axis values (Glen Rose thickness) indicates

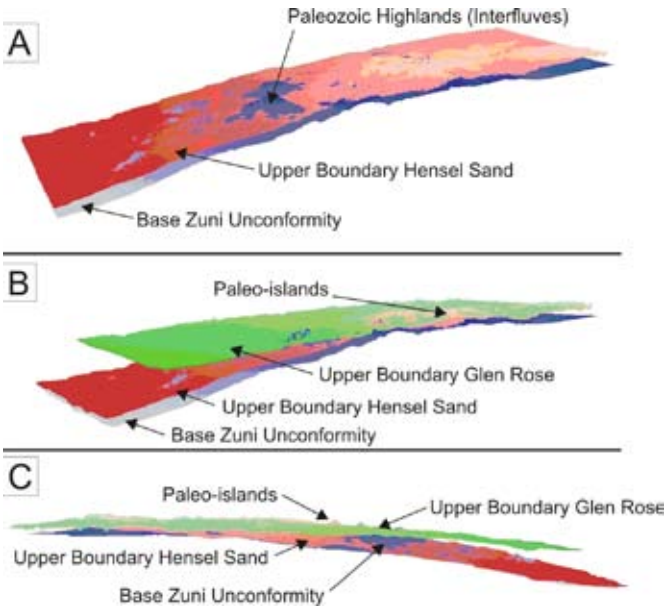

Figure 2. ArcGIS™ ArcScene three-dimensional model of interpolated paleosurfaces. *A*, The upper boundary of the Hensel Sand (Cretaceous fluvial) over the base Zuni sequence unconformity showing Paleozoic highland interfluves. View is from the northeast. *B*, Same as *A*, but including the upper boundary of the Glen Rose and showing Cretaceous paleoislands. *C*, Same as *B*, but shown looking from the south.

terrestrial exposure above where the marine Glen Rose was deposited, probably reflecting islands that persisted throughout the deposition of this first marine unit of the Zuni transgression (figs. 2*B*, *C*). These islands shifted updip (northwest) relative to the Paleozoic interfluves, but still account for about 7.5 percent of the total interpolated grid cells (fig. 2*B*). The paleotopography at the base of the Zuni sequence may have been largely controlled by faulting, but was comparable in total relief and slope as the modern topography of the study area. The interfluves during the Hensel deposition may preserve well-developed paleosols because these features were never scoured by Cretaceous fluvial systems, but were exposed for long periods of time to the atmosphere, and were eventually buried by subsequent low-energy marine sediments. The Glen Rose eventually buried most of the Paleozoic highlands (interfluves) in the center of the study area (figs. 2*A, B*), but laterally terminates on islands of Hensel Sand in the northwestern portion of the study area (figs. 2*B*, *C*). This suggests that island systems may have existed in the study area throughout the deposition of the Glen Rose. The paleoislands in the study area were subsequently buried by the locally thin (< 2 m thick) Walnut Clay. Again, these paleoislands may better preserve well-developed paleosols since they were buried by low-energy deposits and exposed to the atmosphere for a longer interval than other areas where the upper boundary of the Hensel Sand is preserved. Raster GIS modeling also allows for the rapid and consistent calculation of unit thickness, surface slopes, and boundary relationships across large areas to better analyze geologic data and serve as a predictive tool. This study demonstrates how raster GIS modeling can be used, in particular with stratigraphic and paleogeographic studies, by spatially interpolating known data into buried, obscured, or missing areas. The ability for GIS technology to characterize paleogeographic features, including paleoislands, has potentially profound implications for geological study. These implications include, but are not limited to, identifying paleotopographic features for paleontological studies of terrestrial habitats, refining depositional models and interpolating geologic surfaces for petroleum exploration, and studying the connection between paleolandscape position and paleosol formation and preservation.

Global Earth Observation Grid (GEO Grid): Development of an IT Infrastructure for the Earth Science Using Satellite Data

By Shinsuke Kodama[1], Ryosuke Nakamura[1], Naotaka Yamamoto[1], Hirokazu Yamamoto[1], Koki Iwao[1], Masashi Matsuoka[1], Satoshi Tsuchida[1], and Satoshi Sekiguchi[1]

[1]Grid Technology Research Center, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

Global Earth Observation Grid (GEO Grid) is aiming at providing an E-Infrastructure to understand our Earth with more insight and, more precisely, with greater speed and ease. Within the E-Infrastrucutre, we are able to share data, application programs, and scientific workflows without a deep knowledge of information technology (IT), which provides the grid technology. GEO Grid securely and rapidly provides large archives of earth observation satellite data and integrated service with various observation databases and geographic information systems (GIS) data, while making them easy to use.

The core contents of the system are the observation data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) developed by Ministry of Economy, Trade and Industry (METI), Japan and geoscientific information, such as geological and environment technology data, accumulated for a long period of time at the National Institute of Advanced Industrial Science and Technology (AIST). As a core technology, the GEO Grid provides international standard compliant grid technology and develops systems (fig. 1; see also http://www.geogrid.org/). The ASTER data contains multispectral images of visible-near infrared region (visible near-infrared (VNIR)—3 bands), short infrared region (shortwave infrared (SWIR)—6 bands), and thermal infrared region (thermal infrared (TIR)—5 bands) with a spatial resolution of

15 m (for VNIR). We constructed a hard disk-based archive system for the ASTER data that reaches more than 1 million scenes. The level 0 ASTER data (raw data) are stored in this system and higher products are generated by on-demand processing. These products contain orthorectified images and 15-m resolution Digital Elevation Models (DEMs), generated from Band 3N (nadir) and 3B (backward) images.

We will present several sample applications using ASTER data and the GEO Grid system; pyroclastic-flow simulation using an ASTER DEM; integration with other satellite images and ground truth data for the accurate global land-use change detection and carbon-cycle modeling (http://kushi.geogrid.org/); and integration with the geologic GIS data. The GEO Grid system supports Web Map Service (WMS) to provide the ASTER images and maps generated from ASTER data. This allows users to integrate ASTER image and GIS data, such as geologic maps, easily. We had several experiments in which the ASTER image is overlaid on a geologic map provided from the Geological Survey of Japan or the Geoscience Network in United States by using WMS. We also plan to support the Web Coverage Service (WCS) which allows users to use ASTER DEM/Ortho data for scientific analyses. The pyroclastic-flow simulation application provides a possible coverage map of pyroclastic-flow deposits caused by a volcanic dome collapse, and that can contribute to make an emergency volcanic hazard map. An energy line model is used for calculation of the maximum possible flow distance (Takarada and others, 1993). The user can choose collapse point and physical parameters of pyroclastic flow and submit a job using a Web browser. The resulting map is provided by using the WMS, and appears on a Web browser.

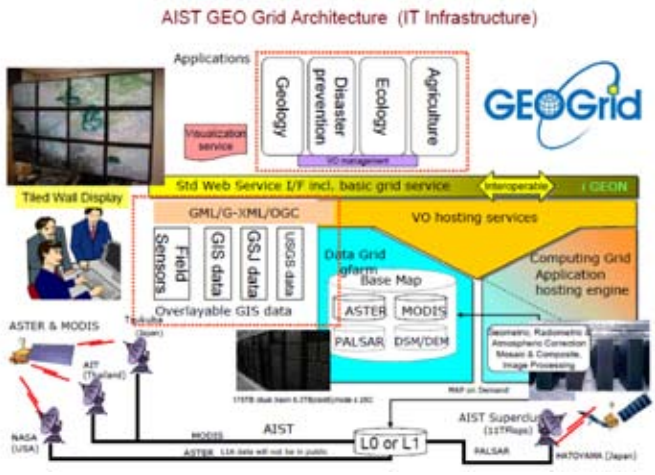

Figure 1. AIST GEO Grid architecture (IT infrastructure).

HydroXC: A Common Schema for Hydrologic Data Transfer and Object Descriptions

By Michael Piasecki[1], Bora Beran[1], Jon Roe[2], and Stephanie Liu-Barnes[3]

[1]Department of Civil, Architectural and Environmental Engineering, Drexel University, Philadelphia, Pa.

[2]National Weather Service, Silver Spring, Md.

[3]APEX Digital Systems, Silver Spring, Md.

The National Weather Service (NWS) Office of Hydrologic Development (OHD) has been sponsoring the development of the HydroXC initiative, which is now in Phase III. The overall goal of this initiative, which is also supported by a consortium comprised of members of the hydrologic community originating in government (National Weather Service, U.S. Geological Survey (USGS), Natural Resources Conservation Service, and U.S. Army Corps of Engineers), academia (Duke University, Drexel University, University of Pittsburg, and University of Virginia), and also private companies (APEX Digital Systems, Environmental Science Research Institute, Vieux and Associates, Wier and Associates), is to develop a general Extensible Markup Language (XML) schema that can be used for data transfer between entities interested in hydrologic data. While some efforts are already underway to "schematize" hydrologic information (for example, the HydroML effort lead by USGS), most efforts are very specific to a certain task prompting the need for an exchange vehicle that is generic enough such that any type of hydrologic information can be described and packaged in a language that is machine readable (in other words, XML).

The third phase of this project will address pragmatic requirements for using the XML schema, and for making it more specific to daily needs of hydrologic software users. Current research efforts are focused on further evolving the HydroXC XML schema that has been derived from the Standard Hydrologic Exchange Format (SHEF) developed at the NOAA National Weather Service. To do so, the current thrust is to focus on the derivation of several specific hydrologic object representations, including a few examples that highlight the new areas of the schema (in other words, a reservoir object, a flow rating curve object, a stream reach object, and a cross-section object). The focus is on compiling descriptions that are useful for data exchange by identifying key attributes that are also used by other standards (for example, GML or HydroML). Additional work is underway to use the HydroXC schema to develop data adapters that are capable of reading and writing messages between some proprietary format and HydroXC-compliant XML.

This work will demonstrate the general composition of the schema, which is aligned along elements used for defining components and example instantiations of the derived object descriptions. We will also outline the inclusion of internal

standards (like ISO 19115 and 8601, GML, and EPSG codes) into this schema and the importance they have in representing geospatial and temporal referencing, as well as some of the challenges that arrive from attempting to incorporate legacy systems.

Implementing the Collaboratory for the Study of Earthquake Predictability (CSEP)

By Maria Liukis[1], Danijel Schorlemmer[1], Philip Maechling[1], and Thomas Jordan[2]

[1]Southern California Earthquake Center, University of Southern California, Los Angeles, Calif.

[2]Department of Earth Sciences, University of Southern California, Los Angeles, Calif.

The Collaboratory for the Study of Earthquake Predictability (CSEP) is developing the infrastructure for facilities to conduct earthquake forecast experiments. It provides a controlled integration environment with standardized software stack for the development and installation of forecast experiments. The processing infrastructure has to allow for rapid computations using distributed computing facilities, but also needs to run on desktop computers for research activities. Module design of the CSEP software focuses on reproducibility of any forecast experiment. Furthermore, program codes need to be validated and distributed to other than Southern California Earthquake Center (SCEC) testing facilities. We will discuss the design challenges and present the software concept, development strategies, ways for participating in development, and the flexibility for customizing our open-source software.

A Drastic Revision of Active Fault Database in Japan Based on the Redefined Relational Data Model

By Yuichiro Fusejima[1], Fujika Miyamoto[1], and Toshikazu Yoshoika[1]

[1]Active Fault Research Center, Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Outline of the Database

Active Fault Database of Japan (http://www.aist.go.jp/RIODB/activefault/cgi-bin/index_e.cgi) contains comprehensive information about active faults in Japan, sorted by the concept of "behavioral segments" (McCalpin, 1996). Each fault is subdivided into behavioral segments based on surface trace geometry and rupture history determined through paleoseismic studies. Faults shown on the index map are linked to a database of behavioral segments, which contains information about geologic and paleoseismic parameters, including slip rate, slip per event, recurrence interval, and calculated rupture probability in the future. Behavioral segments can also be searched by name or combination of fault parameters. All those data are compiled from journal articles, theses, and other documents.

Problems on the Data Input Method

The database was first formulated in 2002 by the Active Fault Research Center, Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology (AIST). The first edition of the database was launched online in March 2005, as part of the Research Information Database (RIO-DB), which is managed by AIST. Through the relational database management system (RDBMS) of ORACLE 9i, many users could easily read information similar to a catalog or a handbook, but using a Web browser; however, searching information using composite keywords was not possible. Furthermore, the order of the data could not be changed using composite attributes. These shortcomings were caused by the data input method. The data were entered using the MS-Excel spread-sheet software. Attempts were made to solve the problem. The inputted data were reconstructed into a pseudo data model, defined for a serialized data on two-dimensional spread-sheet schema; however, the pseudo data model could not solve the problem of data redundancy. Furthermore, standardization of the data specifications also was not possible. Because of the high level of data redundancy and the low level of standardization, flexible searching functions in the first edition of the database were not possible.

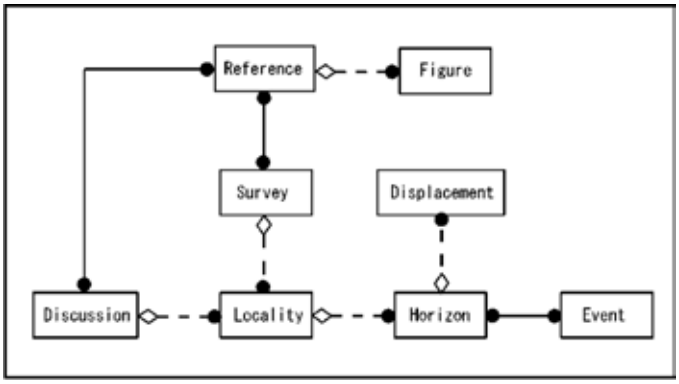

Figure 1. Top-level data model. Entity relationship diagram in IDEF1X.

Redefined Data Model of the Database

In general, all data of relational database must be inputted by RDBMS on a well-planned and well-designed data model. In this study, the data model was redefined to formulate a genuine relational database. The redefinition of the data model is based on the general contexture of published journal articles describing active faults; therefore, the titles of the constitutive entities/tables are Survey, Locality, Horizon, Displacement, Event, Discussion, Figure, and Reference. Sixty-five tables are defined. Many attributes and their data types, domains, and other specifications are also defined. Many more tables are constructed for appending legend code numbers of these attributes. Other attributes describing metadata are also defined. All these tables are normalized and connected by relationships. A top-level data model (entity relationship diagram in IDEF1X) is described in figure 1. The completed data model, which contains detailed definitions, will be presented in a poster at the conference.

Revision of the Database

Based on the redefined data model, a data input user interface system on an MS Access RDBMS was formulated. The new data were entered stepwise by the input system of the MS-Access and finally merged into the main RDBMS of ORACLE 10g. In August 2006, the database containing a large amount of data was revised based on the redefined data model. The flexible searching functions, which allow searching using composite keywords, other advanced searching methods, and changing the order of data using composite attributes, are successfully implemented using the revised database. Coincidentally, new field data collected from over 10,000 research localities are added into the new Japanese edition database. The database also shows metadata, bibliographies, figures, chronostratigraphies, and more detailed information. The new Japanese edition is presently translated into English. The new edition is developed to include analytical and GIS functions and more.

Reference Cited

McCalpin, J.P., ed., 1996, Application of paleoseismic data to seismic hazard assessment and neotectonic research: Paleoseismology, Academic Press, p. 439-493.

Community Science Ontology Development

By Robert Raskin[1], Peter Fox[2], Deborah L. McGuinness[3], and A. Krishna Sinha[4]

[1]Jet Propulsion Laboratory, Pasadena, Calif.

[2]High Altitude Observatory (HAO), National Center for Atmospheric Research (NCAR), Boulder, Colo.

[3]Knowledge Systems, McGuinness Associates and Stanford University, Stanford, Calif.

[4]Virginia Polytechnic Institute and State University, Blacksburg, Va.

Background of the Need for Formal Semantic Encodings

Scientists often spend a majority of their time locating and preparing a dataset before it even can be analyzed and put to use. Given that 21st century science will be blessed with massively large amounts of data, many of which can be used together synergistically, a great opportunity is not being realized. Knowing only the syntactic description of a dataset does not remedy this situation. Formal semantic encoding potentially enables the use of automated data integration, smart search, and interdisciplinary data fusion.

Best Practices and Modular Ontologies

Our experiences suggest that ontologies should be modular. Most aspects of science deal with hierarchical specializations of concepts, such as new classes of rock types and subtypes; hence, inheritance is very important. Ideally, concepts in an earth- science ontology should also inherit from more general ontologies in physics, chemistry, math, space, time, and so on.

Project and Agency Interest

National Aeronautics and Space Administration (NASA) is moving from an instrument-based to measurement-based strategy in its archival systems. This approach implies that full lineage of a dataset must be preserved to enable cross-platform integration. National Science Foundation has initiated calls for ontology development through its Office of Cyberinfrastructure.

Community needs are diverse, but share many common elements, such as the desire to read standard data formats and associate parameter names with meaningful scientific concepts. At these early stages of ontology development, funded work in one community should build upon the work of others, rather than be reinvented and not reused.

Current Ontology Efforts

Current ontology development work includes the following:

A. Semantic Web for Earth and Environmental Terminology (SWEET)—An upper-level ontology developed at NASA/JPL with coverage of the entire Earth system (Raskin, 2006; Raskin and Pan, 2005). (See fig. 1 for the ontology structure; (http://sweet.jpl.nasa.gov.)

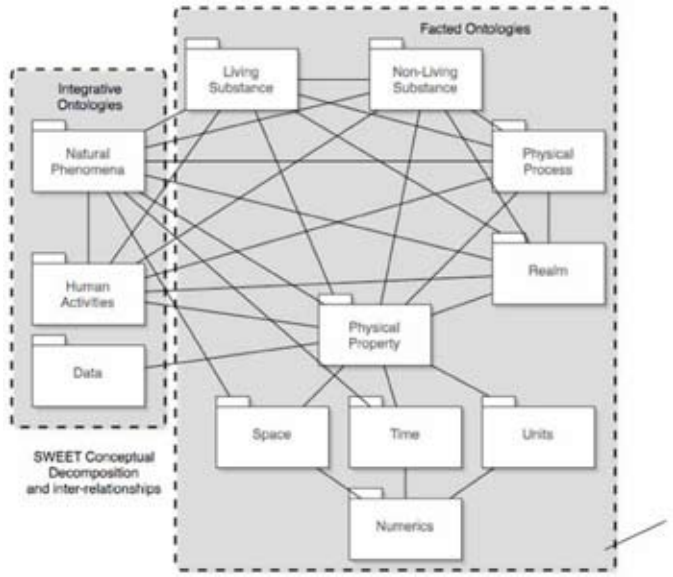

Figure 1. SWEET conceptual decomposition.

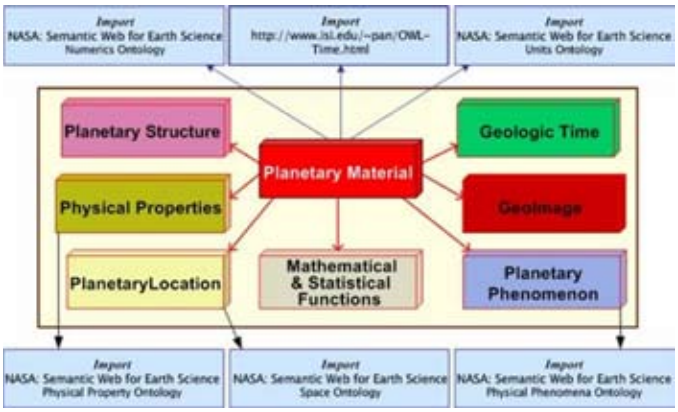

Figure 2. GEON planetary ontology framework.

B. Virtual Solar Terrestrial Observatory (VSTO)—Developed at NCAR with McGuinness Associates with coverage of solar atmospheric physics and terrestrial middle and upper atmospheric physics (McGuinness and others, in press; http://vsto.org).

C. Geosciences Network (GEON)—Developed at Virginia Tech with coverage of the solid Earth. (See fig. 2 for the ontology structure; http://geon.geol.vt.edu/geon/index.html.)

D. Marine Metadata Initiative (MMI)—Developed at Monterrey Bay Aquarium Research Institute (MBARI) with coverage of instrumentation and the marine world. (See fig. 3 for an excerpt; (http://marinemetadata.org.)

Ontology Tools for Collaboration of Communities

There has been a dire need for tools to support ontology evolution, validation, reasoning, comparison, merging, and so on. This is an emerging area of work for a number of organizations. MMI, for example, has stepped in and created a suite of tools for creating, comparing, and harmonizing ontologies with the goal of supporting ontologies for marine science, and science in general (MMI tools: http://marinemetadata.org/examples/mmihostedwork/ontologieswork). MMI convenes workshops where teams generate new ontologies, such as for instrumentation.

The http://www.PlanetOnt.org Web site is a collaborative community set up to share ontologies and infuse the experience of others. It provides services for: ontology version registration; comparison of ontologies; imported class dependencies; RSS feeds to notify dependent ontology owners of potential changes made; and discussion regarding an ontology or specific elements within an ontology. It provides a forum for identifying best practices, and for getting around specific limitations in OWL in a consistent manner. It is open to community involvement and welcomes submissions.

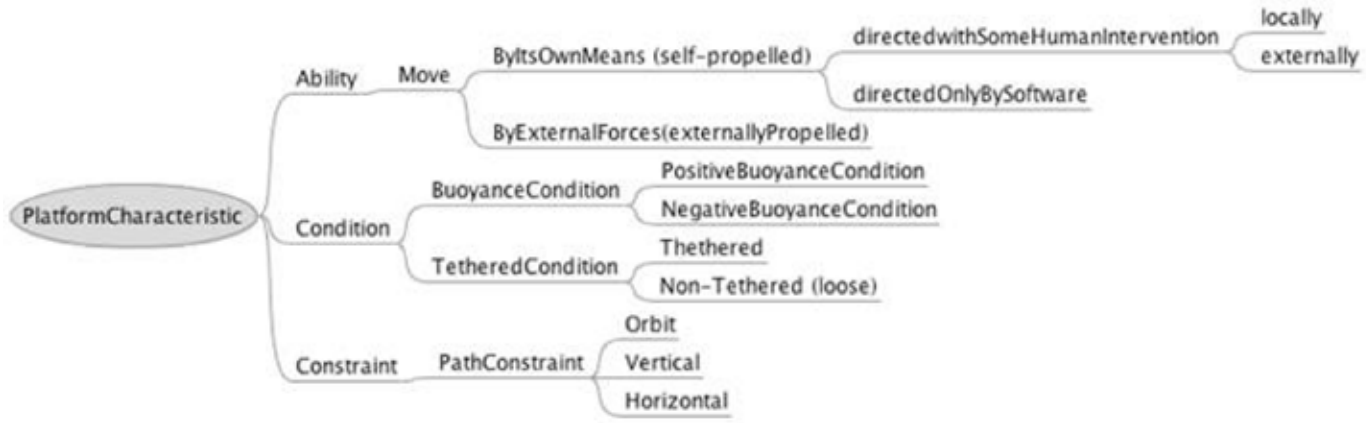

Figure 3. MMI ontology schematic.

Discussion and Conclusion

Ontologies should be developed collaboratively and incrementally from existing work. Ontology normalization shares much in common with database normalization; however, there is a further need to distinguish the general concept from the more specific. Specialized ontologies should import from the more general ones rather than repeat the general concepts. The http://www.PlanetOnt.org Web site supports this directional structure by identifying dependencies between any pair of ontologies.

References Cited

McGuinness, D., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J.L., and Middleton, D., in press, The Virtual Solar-Terrestrial Observatory: A deployed semantic Web application case study for scientific research, *in* Proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI-07), Vancouver, B.C., Canada, July 22-26, 2007.

Raskin, Rob, 2006, Ontologies for earth system science, *in* Sinha, A.K., ed., Geoinformatics: From data to knowledge: Geological Society of America Special Paper 397, p. 195-1990

Raskin, R.G., and Pan, M.J., 2005, Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET): Computers and Geosciences, v. 31, p. 1119-1125.

Ontologic Integration of Geoscience Data on the Semantic Web

By Zaki Malik[1], Abdelmounaam Rezgui[1], and A. Krishna Sinha[2]

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Va.

[2]Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.

Introduction

The World Wide Web was originally created for data sharing among scientists. Over the years, the Web has evolved from being merely a repository of data to a vibrant environment that allows users to make their data and applications Web accessible. As a result, a wealth of information and applications are now available, and their quantitative and qualitative management has become a primary issue. For instance, several research initiatives by geoscientists over the years have produced large amounts of data; however, the ability to find, access, and properly interpret these large data repositories has been very limited. Two main reasons for this lack of data sharing are the adoption of personal acronyms, notations, conventions, units, and so on, by each research group when producing data, and the current Web search methods that can be understood only by humans or custom-developed applications (Medjahed and others, 2003). Currently, machines merely display the data, and they are unable to process it any further due to lack of data and application semantics. This makes it difficult for other scientists to correctly understand the semantics of the data, and makes the automatic interpretation and integration of data simply infeasible. We suggest that for enabling the sharing, understanding, and integration of geosciences data and applications on a global scale, ontology-based registration and discovery is required.

Ontologies and the Semantic Web

The emerging "Semantic Web" is defined as an extension of the existing Web, in which information is given a well-defined meaning (Berners-Lee and others, 2001). The ultimate goal of the envisioned Semantic Web is to transform the Web into a medium through which data and applications can be automatically understood and processed. The concept of Web services (and other related technologies) is seen as a key enabler of the Semantic Web (Alonso and others, 2003; McIlraith and others, 2001). A Web service is a set of related functionalities that can be programmatically accessed through the Web. The convergence of business and government activities in developing Web service-related technologies (for example, SOAP; Universal Description, Discovery and Integration (UDDI); and WSDL)) is a sign for the large adoption of Web services in the near future (Curbera and others, 2002). Another key player in the envisioned Semantic Web is the concept of ontology. An ontology may be defined as a set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic. The Semantic Web is expected to offer data (organized through ontologies) and applications (exposed as Web services) enabling their understanding, sharing, and invocation by automated tools.

Data Ontologies

Recognizing the potential of the Semantic Web, we have defined a "planetary ontology" (through many workshops and scientific meetings) to provide the ontologic framework for earth-science data at many levels of semantic granularity. The planetary ontology includes concepts, concept taxonomies, relationships between concepts, and properties that describe concepts, as an initial step towards the development of ontologies for earth science. Figure 1 shows the high-level representation of the planetary ontology. High-level packages, such as Planetary Material, can be used to represent the nature (physi-

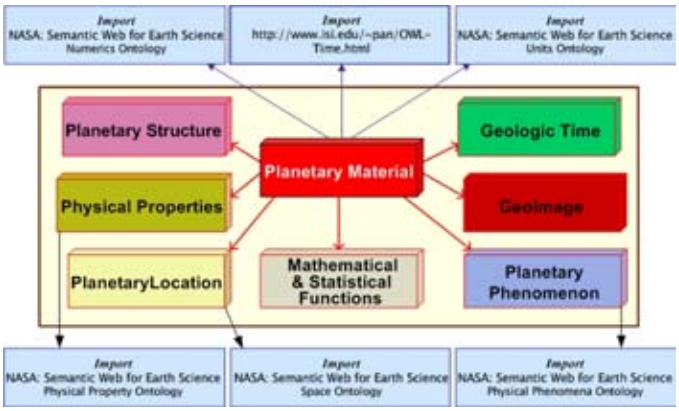

Figure 1. The planetary ontology framework.

cal/chemical) and state of substances and their properties. This figure also emphasizes the utilization of imported and inherited properties from additional packages (for example, Physical Properties, Location, and Planetary Structure) to fully define the concept of Planetary Materials. Ontologies will support the Semantic Web through (1) ease of registration to facilitate discovery, and (2) ability to query across multiple and diverse databases through interconnected disciplinary ontologies.

We have developed a prototype system as proof of concept, which uses the planetary ontology in discovering solutions to complex geoscience questions. Some of the geoscience tools that are required for integration within the Web environment are registered as Web services. For example, "data filtering tools" that distinguish between geologic bodies based on Magma Class (for example, A-Type, S-Type, M-Type, or I-Type) or metamorphic facies assemblages have been "wrapped" and registered as Web services. This enables users to utilize the tools without detailed knowledge of the operating system, development language environment, or the component model used to create these geoscience tools. Since only the input and output parameters need to be defined for Web service-based applications, it encourages reusability and reduces development time, effort, and cost.

Service Ontologies

As the Semantic Web matures, and more geoscientists adopt this paradigm, it is expected that a number of geoscience tools and services will be made accessible as Web services. This would require that, similar to data management practices, Web services also be ontologically registered. Annotating Web services with semantics would ensure that appropriate tools (in the form of Web services) are selected in an efficient and automatic manner for answering geoscience queries. Domain experts would provide formal specifications of geoscience concepts, enabling automated Web service usage. Moreover, since the Semantic Web is geared towards interactions involving minimal human intervention, a service ontology would

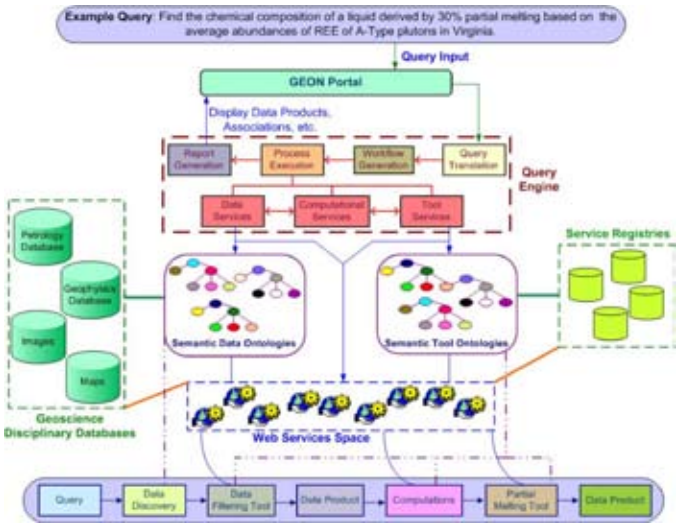

Figure 2. Semantic Web-enabled geoscience querying engine.

enable direct service-to-service communication and facilitate information transfer.

To fully understand the need for a service ontology, consider the geoscience query: "Find the chemical composition of a liquid derived by 30 percent partial melting (PM) based on the average abundances of Rare Earth Elements (REE) of A-Type plutons in Virginia." This query clearly requires access to a number of datasets and geoscience tools. Figure 2 provides a high-level overview of the four steps involved in answering the query. These are: finding the A-Type bodies in Va., computing the averages, using the REE definitions contained in the element ontology and exporting the data to a PM tool for computation, and displaying the results. The prototype query engine (Discovery, Integration and Analysis engine) developed by us is able to address the query.

The discovery of data pertaining to A-Type bodies (a class of igneous rocks) and which contains elemental data classified as REE requires that the geoscience data be registered to ontologies. The data ontologies now available to geoscientists (Sinha and others, 2006) allow access to multiple disciplines to fulfill this requirement (see fig. 2). Another major requirement for answering the query lies in the discovery of appropriate tools to carry out data filtering through both mathematical and domain-specific computations. It is expected that geoscientists will develop similar Web services (tools), using their own acronyms, and advertise them across multiple service registries (UDDI). Discovery of the required tools is only possible if the available tools and services have defined and precise semantics associated with them; thus, similar to data ontologies, "service ontologies" will also be required. Service ontologies are used for two purposes: to register services and to discover services. An ontology-based service description provides metadata information about the service provider, such as the service's categories and subcategories, the service's address, its parameters and their types, the service's output, the service's cost, etc. Service registries

expose ontology-based search interfaces that service clients use to discover services appropriate for a given task. Once the client selects a given service, the registry provides the service description that the client then uses to actually invoke the service; therefore, a service ontology will do for Web services what a data ontology has done for geoscience data.

Acknowledgment

This research is supported by the National Science Foundation award EAR 0225588 to A.K. Sinha.

References Cited

Alonso, G., Casati, F., Kuno, H., and Machiraju, V., 2003, Web Services: Concepts, Architecture, and Applications: Springer Verlag (ISBN: 3540440089).

Berners-Lee, T., Hendler, J., and Lassila, O., 2001, Semantic Web, Scientific American, v. 284, no. 5, p. 34-43.

Curbera, F., Duftler, M., Khalaf, R., and Nagy, W., 2002, Unraveling the Web Services Web: IEEE Internet Computing, v. 6, no. 2, p. 86-93.

McIlraith, S., Son, C., and Zeng, H., 2001, Semantic Web Services: IEEE Intelligent Systems, v. 16, no. 2, p. 46-53.

Medjahed, B., Bouguettaya, A., and Elmagarmid, A., 2003, Composing Web Services on the Semantic Web: Very Large Data Bases (VLDB) Journal, v. 12, no. 4, p. 333-351.

Sinha, A. K., Malik, Z., Rezgui, A., and Dalton, A., 2006, Developing the Ontologic Framework and Tools for the Discovery and Integration of Earth Science Data-Cyberinfrastructure Research at Virginia Tech, Annual Report., http://geon.geol.vt.edu/geon/index.html.

The Critical Zone Exploration Network: Your Growing Community

By Michael Hofmockel[1], Sue Brantley[1], Doug Miller[1], and Daniel deB. Richter[2]

[1]Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, Pa.

[2]Nicholas School of Environment and Earth Sciences, Duke University, Durham, N.C.

At Earth's surface, a complex suite of chemical, biological, and physical processes combines to produce soil from bedrock and sediments within the zone that extends from the outer limits of vegetation to the lower limits of ground water. This weathering engine transforms primary minerals, provides nutrients to nourish ecosystems and human society, mediates the transport of toxic components within the biosphere, creates water flow paths that shape and weaken bedrock, and contributes to the evolution of landscapes at all temporal and spatial scales. At the longest time scales, the weathering engine sequesters carbon dioxide (CO_2), thereby influencing the global carbon cycle, long-term climate change, and weathering rates. This Critical Zone supports all life on Earth (http://www.czen.org/node/254).

The Critical Zone Exploration Network (CZEN, http://www.czen.org) is a network of people, locations, tools, and ideas to investigate processes within the Critical Zone. Any group studying the Critical Zone is encouraged to participate. The CZEN Web site already provides many communication tools, like event scheduling, file upload, forums, list serves, and literature databases. A Fall 2007 data and information systems workshop is planned to discuss the ongoing collection and compilation of data for CZEN (http://www.czen.org/node/215).

Why join CZEN and create a group (http://www.czen.org/node/242)? Participation enables the following: (1) brings Critical Zone Observatories (CZOs) into close cyber-proximity; (2) provides economy of scale in technology overhead; (3) makes Web resources available to researchers from day one; and (4) allows CZOs the resources to build project-specific Web products, as the Web-based community-management system (CMS) is extendable and flexible.

Towards a Reference Plate Tectonics and Volcano Ontology for Semantic Scientific Data Integration

By A. Krishna Sinha[1], Deborah L. McGuinness[2], Peter Fox[3], Robert Raskin[4], Kent Condie[5], Robert Stern[6], Barry Hanan[7], and Dogan Seber[8]

[1]Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.

[2]Knowledge Systems, McGuinness Associates and Stanford University, Stanford, Calif.

[3]High Altitude Observatory (HAO), National Center for Atmospheric Research (NCAR), Boulder, Colo.

[4]Jet Propulsion Laboratory, Pasadena, Calif.

[5]Department of Earth and Environmental Science, New Mexico Institute of Mining and Technology, Socorro, N. Mex.

[6]Department of Geosciences, University of Texas at Dallas, Richardson, Tex.

[7]Department of Geological Sciences, San Diego State University, San Diego, Calif.

[8]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

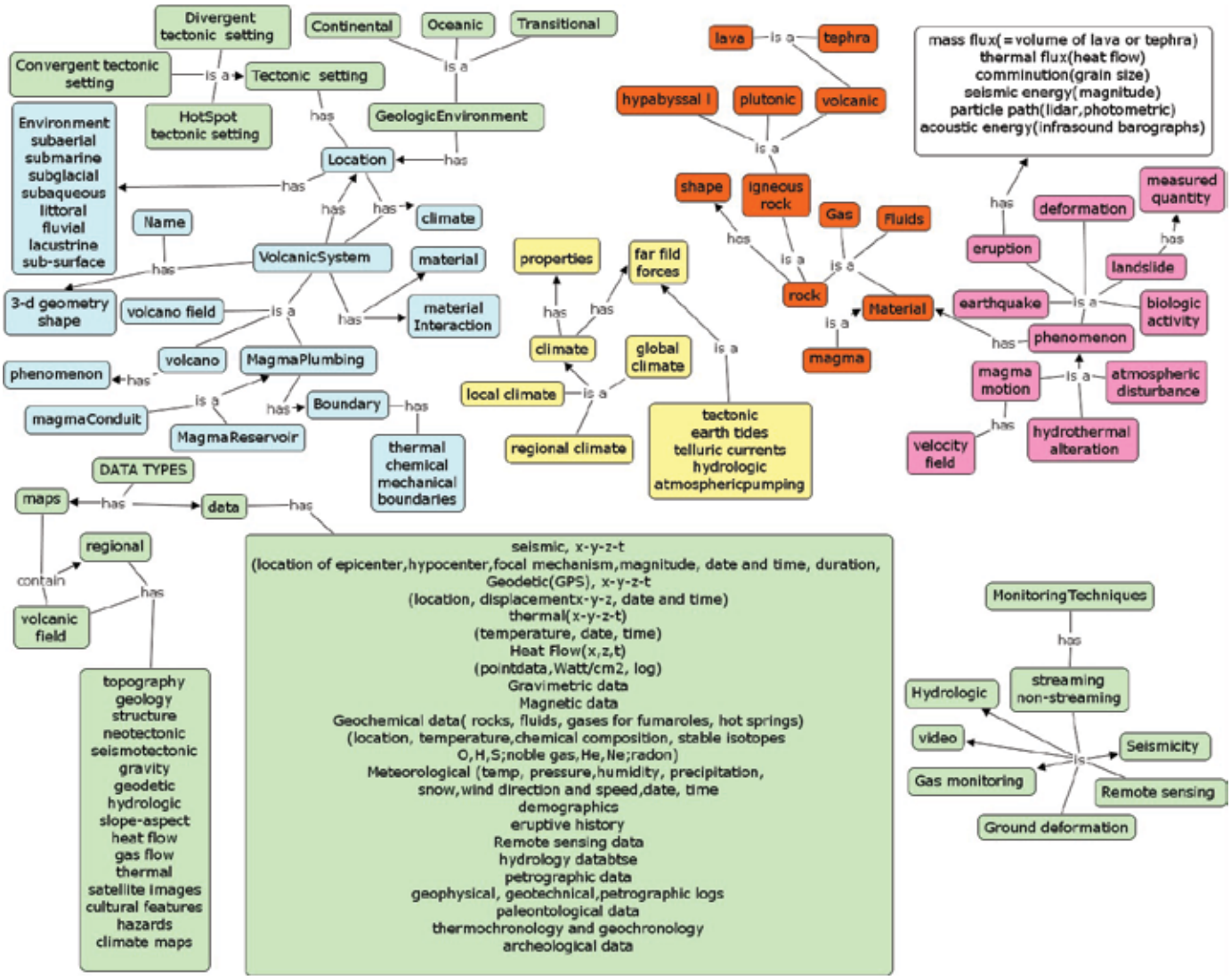

Figure 1. A class diagram representing the framework for linking data to volcanic features and processes.

Introduction

When scientific progress depends on integration of data across disciplines, it is critical for the users in the diverse disciplines to have access to the data in terms they can understand and use. Ontologies provide a method for encoding terms, term meaning, and term interrelationships in a machine interpretable format. In a geology setting, this means that ontologies provide a way of representing geologic terms and their associated properties. These encodings can enable interoperability and interdisciplinary data integration by allowing end users and agents to access precise, operational term definitions.

In support of a National Aeronautics and Space Administration (NASA)-funded scientific application (Semantically Enabled Science Data Integration Project (SESDI); http://sesdi.hao.ucar.edu/) that needs to share volcano and climate data to investigate relationships between volcanism and global climate, we have generated a volcano and plate tectonic ontologies. Our goal is to create reference ontologies—open terminology representations meant to be shared and reused by a broad community of users interested in the subject area, as well as to provide access to key volcanology and plate tectonic-related databases. We recognize that many of the features and products of volcanism are related to plate tectonic processes; thus, the availability of both ontologies allows us to map data associated with volcanic events, such as gas and particle ejecta, to plate tectonic settings and processes. Our goal is to support investigations into links between volcanism and climate change. This goal can be more rigorously addressed through semantic integration of data associated with the atmospheric response to volcanism and its plate tectonic setting.

Volcano Ontology

The near and far field effects of volcanic activity are common geologic phenomena observed in many parts of the world and may include possible catastrophic damage near eruptive

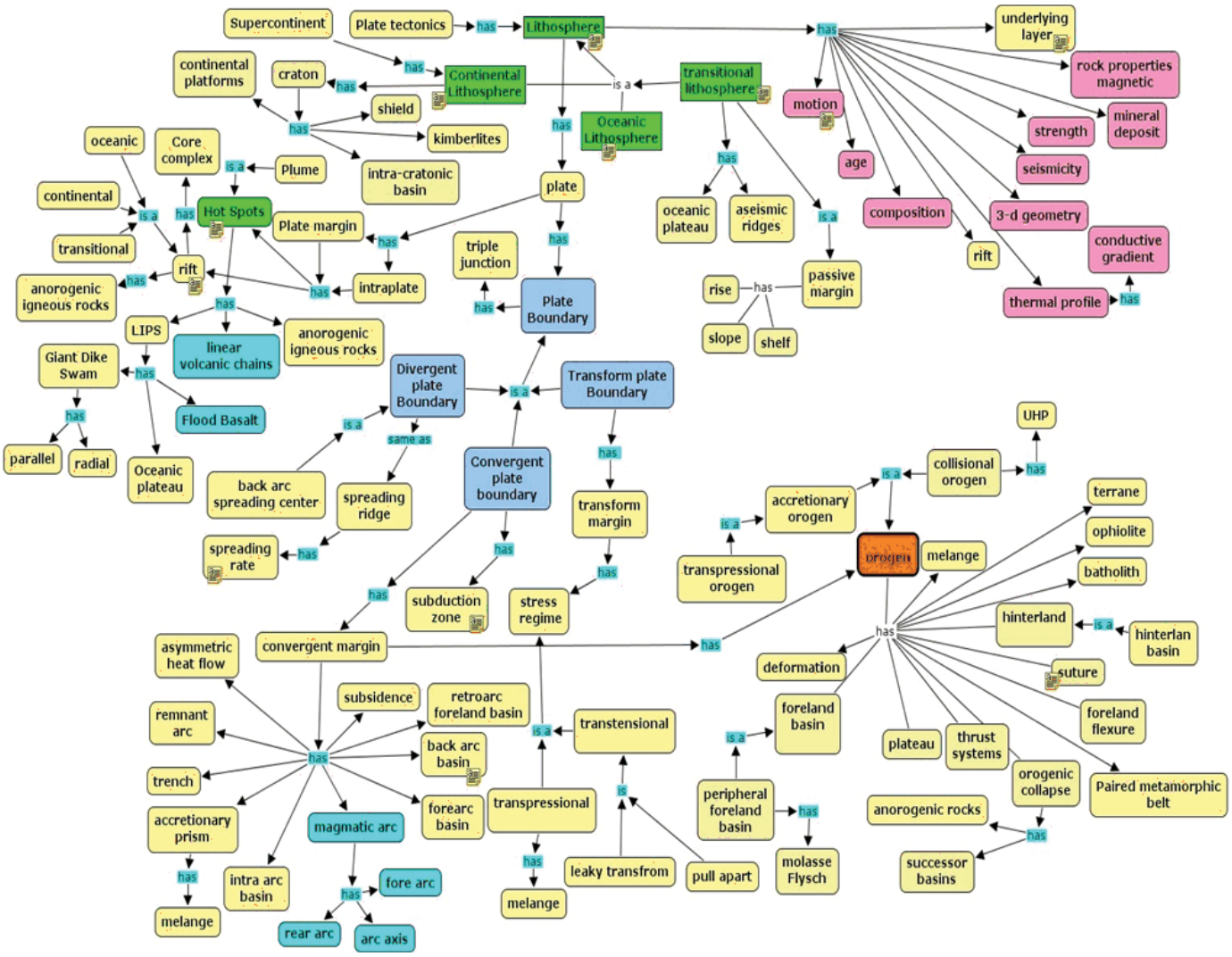

Figure 2. Class diagram relating plate tectonic features and their products.

centers, as well as measurable effects on global climate. A comprehensive categorization of types of volcanic activity and its episodic nature is a challenging scientific goal that requires access to many known types of geologic and sensor data. We gathered leading volcano and scientific ontology experts to capture necessary conceptual relationships between data and phenomena associated with volcanism (Sinha and others, 2006; McGuinness and others, 2006). Figure 1 depicts some of the results of this meeting in the form of a concept map. Our representation documents the terms and their relationships. Such representations are requirements for ontology-supported access to data and integration tools. Three high-level classes associated directly with volcanism were identified: volcanic systems, volcanic phenomena, and climate. The concept of volcanic systems was linked to both plate tectonic and geologic environments for providing links between eruptive style, magmatic composition, and the location of the volcanic field. Additional concepts related to volcanic systems include magma plumbing, eruption environment, and the three-dimensional geometry of the volcanic field. Deformation, eruption, landslides, biologic activity, atmospheric disturbance, hydrothermal alteration, magma motion, and earthquakes were considered to be subclasses of a higher level class called phenomena. The two classes—phenomena and volcanic systems—are linked through materials (in other words, magma and its products). This organization of concepts contains associations with all data types used by geologists to monitor active volcanic systems and to study volcanoes preserved in the geologic record.

Plate Tectonics Ontology

The dynamic interaction of lithospheric plates and their interaction with hot spots are closely related to all classes of volcanic phenomena. We gathered leading plate tectonics and scientific ontology experts and created a plate tectonics ontology that focuses on terminology expected to be used in semantic integration of data associated with volcanic activity and its plate tectonic setting (fig. 2). We established a high-level class

structure representing the concept of lithospheric plates, as well as the subclasses that are associated with such features. The concept of a plate was treated as an independent class to permit the association of plate boundaries with this concept. For example, a divergent plate boundary has organizational relationship with a back-arc spreading center and a spreading ridge. Similarly, the subclass of convergent plate boundary contains convergent margin, which in turn is the parent of features such as fore arc, arc axis, or back arc. These are some of the more common plate tectonic settings associated with volcanism, and its relationship to composition variability in volcanic products is one of the key research goals of SESDI. It was also established that plates have intraplate and plate margin settings which often contain volcanoes associated with possible hot spot activity.

Discussion and Conclusion

Ongoing climate modeling efforts (Robock, 1989, 1991) have ascribed gaseous emissions, especially sulfur dioxide, hydrogen sulfide, and hydrogen fluoride, as the most significant gases capable of changing climate over periods of decades. For example, ongoing NASA efforts of studying volcanic sulfur dioxide (SO_2) loading using a total ozone mapping spectrometer (TOMS; Volcanic Emissions Group, http://toms.umbc.edu/) show emission relationships between arc and non-arc volcanoes. Our work towards creating reference plate tectonics and volcano ontologies is aimed at facilitating scientific data integration in such interdisciplinary settings. Our newly developed high-level volcano and plate tectonic ontologies are being used to help clarify the relationships between total emissions and plate tectonic settings (noting in particular that these relationships are not unique). One conclusion is that the field requires a more sophisticated ontology for volcanoes and plate tectonics prior to extending current models associating climate change with volcanic activity.

Acknowledgment

Semantically-Enabled Scientific Data Integration (SESDI) is a semantic science data integration project sponsored by NASA Advancing Collaborative Connections for Earth-Sun System Science (ACCESS) and NASA Earth-Sun System Technology Office (ESTO) under award AIST-QRS-06-0016.

References Cited

McGuinness, D.L., Sinha, A.K., Fox, P., Raskin, R., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., and Lin, K., 2006, Towards a reference volcano ontology for semantic scientific data integration, *in* Proceedings of American Geophysical Union Joint Assembly, Baltimore, Md., May 23-26, 2006.

Robock, A., 1989, Volcanoes and climate, *in* Berger, A., Schneider, S., Duplessy, J. Cl., eds., Climate and geo-sciences: A challenge for science and society in the 21st century: Dordrecht, Kluwer, NATO ASI Series, Series C, Mathematical and physical sciences, no. 285, p. 309-314.

Robock, A., 1991, The volcanic contribution to climate change of the past 100 years, *in* Schlesinger, M.E., ed., Greenhouse-gas induced climatic change: A critical appraisal of simulations and observations: Amsterdam, Elsevier, p. 429-444.

Sinha, A.K., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., Fox, P., McGuinness, D.L., Raskin, R., and Lin, K., 2006, Towards an ontology for volcanoes, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Geoinformatics 2006—Abstracts: U.S. Geological Survey Scientific Investigations Report 2006-5201, p. 52.

USGS Energy Program Geoinformatics: From Data Management to Information Services

By David A. Ferderer[1], Gregory L. Gunther[1], Christopher C. Skinner[1], and Laura R.H. Biewick[1]

[1]U.S. Geological Survey, Denver, Colo.

The U.S. Geological Survey (USGS) Energy Resources Program (ERP) is responsible for generating publicly available, science-based assessments on the distribution, quantity, and quality of domestic and worldwide energy resources. Characterizing energy resources and their distribution promotes responsible use, helps sustain a dynamic economy, and supports balanced economic, energy, and environmental policy decisionmaking. The ERP also sponsors a Data Management Project to provide information stewardship in support of energy assessments; to develop information technology (IT), geographic information systems (GIS), and data management infrastructure; and expedites access to energy-resource information. This presentation offers an overview of ERP geoinformatic activities conducted by the ERP Data Management Project in support of information stewardship, Internet services, and future development of a service-oriented architecture.

Information management in the ERP involves integration and alignment of IT architecture, geospatial data and services, and data management processes to meet science requirements. These components must also meet project needs, foster data stewardship, and increase access to and discovery of products and support Internet map services and future service-oriented architecture designs. Some components utilized and developed include: (1) IT hardware infrastructure (Linux/Dell servers, Network Appliance file server, Oracle relational database management system, and gigabit networks); (2) GIS data and services (geodatabases, metadata documentation, desktop GIS,

map services); and (3) data management protocols (information planning, data mining, data warehousing, inventory, product access, and discovery). These activities have produced data organization schemas and warehouses, metadata, project and product work flows, search and discovery tools, and Enviromental Systems Research Institute map (image) services, to name a few. Some of the key project areas where these capabilities, services, and tools have been implemented include the National Oil and Gas Assessment, World Petroleum Assessment, and the Gulf Coast Geologic Framework Project.

The ERP is now transitioning its primary information management components into a more open, manageable, and flexible information service environment that is based on a service-oriented architecture. Successful information services and service-oriented architecture rely on metadata documentation, open-source standards, interoperability, catalogues, indexes for discovery, and by leveraging the Internet and portal technologies. Ultimately, these capabilities can foster advanced computing, advanced ontologies, and support knowledge integration and decisionmaking in a complex science environment. To this end, the ERP continues to develop more robust and functional metadata server architecture, consolidate and improve existing Internet map services, develop national and global-scale energy-resource service layers and capabilities, incorporate Open Geospatial Consortium standards, and is instituting plans to develop ERP portals highlighting services, catalogues, and service-oriented architecture capabilities.

Web Services for Geoscience Data: Experiences and Lessons

By Robert A. Arko[1], Andrew K. Melkonian[1], Suzanne M. Carbotte[1], Kerstin A. Lehnert[1], and Sri Vinayagamoorthy[2]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, N.Y.

[2]Center for International Earth Science Information Network, Columbia University, Palisades, N.Y.

The lack of accepted standards for data interoperability is a continuing challenge in the geoscience community, and hampers our ability to make research results broadly available. The Open Geospatial Consortium, Inc. (OGC; http://www.opengeospatial.org) is attempting to address this challenge by developing Web service standards through a consensus process among industry, government, and academic partners. These standards include the Web Map Service (WMS) to compose and display map images from underlying data sources, as well as the Web Feature Service (WFS) and Web Coverage Service (WCS) to provide direct access to geospatial data.

WFS provides a simple standard for serving geolocated vector data such as points and polygons. The request is expressed entirely in the URL, and the response is delivered as an XML object using the Geography Markup Language (GML). Such a service can advertise a wide array of useful geoscience data, including station locations, physical specimens, event catalogs, track lines, and so on. WFS is currently supported by numerous geographic information systems (GIS) server products, both commercial (ArcIMS (http://www.esri.com) and RedSpider (http://www.ionicsoft.com)) and open source (GeoServer (http://www.geoserver.org) and MapServer (http://mapserver.gis.umn.edu)). It is supported in GIS clients, such as uDig (http://udig.refractions.net) and GeoMapApp (http://www.geomapapp.org).

Deployment of WMS and WFS providers is underway throughout the geoscience community. University Navigation Signal Timing and Ranging (NAVSTAR) Consortium (UNAVCO; http://www.unavco.org), Incorporated Research Institutions for Seismology (IRIS; http://www.iris.edu), and the Marine Geoscience Data System (MGDS; http://www.marine-geo.org) recently reported results from ongoing collaborative work (fig. 1). The National Geophysical Data Center (http://www.ngdc.noaa.gov), Petrological Database of the Ocean Floor (PetDB; http://www.petdb.org), and LDEO Borehole Research Group (http://www.ldeo.columbia.edu/BRG) have also deployed WFS providers. The Marine Metadata Interoperability Project (MMI; htttp://www.marinemetadata.org) is pursuing activities, including a formal Open Geospatial Consortium (OGC) Interoperability Experiment for ocean-observing data. Examples from these projects will be described in detail.

WFS can provide extensive information for each feature instance, including identifier, location, time, elevation, URL (in other words, reference for further information), and any

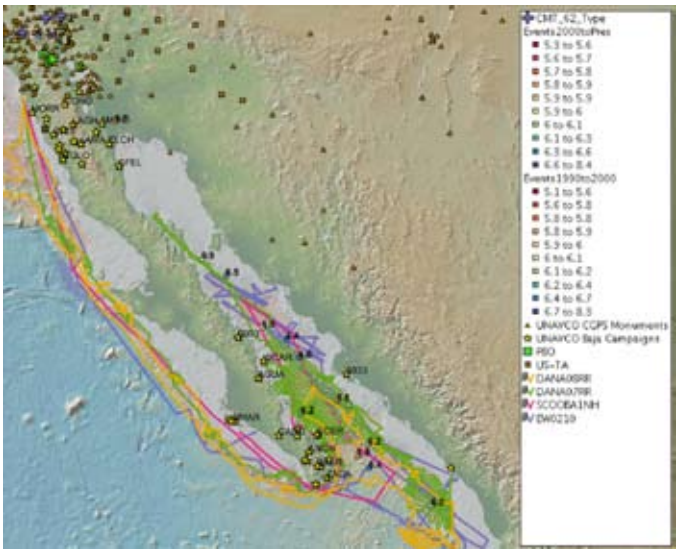

Figure 1. uDig GIS client displaying integrated results from UNAVCO, IRIS, and MGDS providers at Gulf of California (Baja) local study site. Features and layers include global positioning system campaign surveys and continuous stations, seismic stations, earthquake locations, seafloor bathymetry, and expedition tracks.

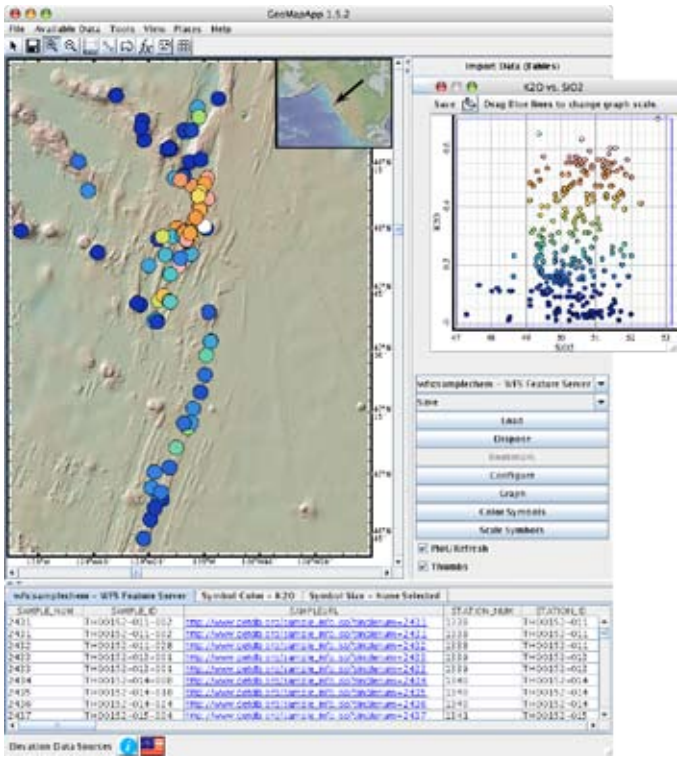

Figure 2. GeoMapApp GIS client displaying results from PetDB WFS provider. User selected Juan de Fuca mid-ocean ridge study site; plotted and colored distribution of potassium oxide (K_2O) versus silicon dioxide (SiO_2).

number of additional data attributes. For example, the PetDB WFS provides a sample feature with an extensive listing of geochemical analyses at each instance. A WFS-enabled client such as GeoMapApp can load and display PetDB samples on a map, and allow the user to color, plot, and intercompare different analytical values (fig. 2).

As WFS usage increases, several performance issues have become apparent. A request for a large number of feature instances can return a prohibitively large result that exhausts network or memory resources. The GeoMapApp client addresses this issue by restricting the user to a defined bounding box (world, ocean, or local study site, as appropriate), thus keeping the server request to a manageable size. Further, there is no accepted standard for attribute names, units, or ordering. GeoMapApp addresses this issue by offering a generic interface for plotting data values, allowing the user to select and color attributes according to domain knowledge.

Scientific Workflows for Ontology-Based Data Mining of Geological Data

By Ilkay Altintas[1], Sara Graves[2], Rahul Ramachandran[2], Dogan Seber[1], and A. Krishna Sinha[3]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[2]Information Technology and Systems Center, University of Alabama at Huntsville, Huntsville, Ala.

[3]Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Va.

As scientific research becomes more interdisciplinary and requires integrative approaches, domain scientists are in need of advance technology tools to be able to analyze and integrate diversified and voluminous datasets for their research activities. Although there are a variety of technology-based resources, such as data-mining tools, scientific-workflow-management systems, and portal frameworks, building a complete cyberinfrastructure framework that incorporates all these technologies is challenging and requires an extensive collaboration among domain scientists and technology developers. In this abstract, we explain our approach to develop strategic cyberinfrastructure technologies that will integrate workflow technologies with data-mining resources and portal frameworks in a semantically enabled work environment. This is a unique effort in the sense that each component will have to be well integrated into the system while giving sufficient flexibility for the developed services to be applicable in other scientific disciplines. Services to be developed in these efforts will enable scientists to manage large and heterogeneous datasets in a timely fashion and extract new knowledge from existing and future datasets. We chose geosciences as our demonstration domain because geoscience data are extremely heterogeneous and complex and there is significant expertise and resources available to be used in such activities.

Each cyberinfrastructure component identified in this effort is at a sophistication level such that they can now be used in an integrative environment. Workflows significantly improve data analysis, especially when data are obtained from multiple sources and (or) various analysis tools. Given their nature, workflows are effective in integrating different technologies and formalize the process; hence, they form a natural integration environment. Despite many existing efforts in workflow development, integrating workflows with ontology-based data mining provides unique challenges for developers and requires collaborative research efforts among technology developers. Application of traditional data-mining techniques (clustering and classification) has been around for a long time and has resulted in extracting novel information from large scientific databases (for example, atmospheric sciences and genetics), and helped to manage costs and design of effective sampling/experimental strategies. As scientific data become more complex however, as in the geosciences, it is important to relate data to the concepts within disciplines. Data-mining

services, including developing new ones that deal with sparse data, can be applied at different levels of abstraction and help the user discover more meaningful patterns leading to a more robust capability to answer scientific questions. For example, a geoscientist may want to study predicting volcanic eruptions using these types of resources or apply the technologies to identify the plate tectonic setting of a former volcano. As future steps, we plan to apply the ontology-driven data-mining approach to global geoscience datasets, such as GeoRoc, EarthChem, Petros, Pluto, and volcano databases from the U.S. Geological Survey and the National Aeronautics and Space Administration, towards discovering patterns and trends between plate tectonic settings and volcanism not recognized by individual scientists. We will utilize workflows in a portal environment to integrate semantic data management and data-mining technologies seamlessly to facilitate a more comprehensive understanding of the nature of volcanism and its plate tectonic settings.

The Australian Mineral Occurrence Data Exchange Model

By Adele Seymon[1], Lesley Wyborn[2], Bruce Simons[1], Oliver Raymond[2], Gary Andrews[3], Terry Denaro[4], Greg Jenkins[5], Peter Lewis[6], James Llorca[1], Marcus McClenaghan[7], Alistair Ritchie[1], Jafar Taheri[7], Ian Withnall[4], and Andrew Wygralak[3]

[1]GeoScience Victoria, Victorian Department of Primary Industries, Australia

[2]Geoscience Australia, Canberra, Australia

[3]Northern Territory Geological Survey, Northern Territory Department of Primary Industries, Fisheries and Mines, Australia

[4]Geological Survey of Queensland, Queensland Department of Natural Resources and Water, Australia

[5]Primary Industries and Resources South Australia, Government of South Australia, Australia

[6]Geological Survey of New South Wales, New South Wales Department of Primary Industries, Australia

[7]Mineral Resources Tasmania, Department of Infrastructure, Energy and Resources, Australia

Introduction

The Australian Mineral Occurrence Data Exchange Model has been collaboratively developed under the leadership of the Australian Government Geoscience Information Policy Advisory Committee (GGIPAC). Representatives from all Australian Federal, State, and Territory Geological Surveys contributed to the model.

Australian mineral occurrences information is stored in individual State and Territory Geological Survey databases. The individual organization's store and maintain information about mineral occurrences, such as commodities, historical production, endowment, reserves, resources, and mineral deposit classification. Each agency's database has its own format, attributes, and vocabularies, and each was developed using a variety of software platforms and versions of the software to meet the organization's individual business requirements.

Geoscience Australia developed a central database that provided a national overview of mineral occurrences, which was made accessible through the Australian Geoscience portal (http://www.geoscience.gov.au/geoportal/minocc/). To create this Web page, data from the States and Territories is currently sent to Geoscience Australia and then manually massaged and uploaded to the central database. As this database is not dynamically linked to the State databases, there are often inconsistencies between data at the Federal and the State and Territory level for the same deposit or occurrence, depending on how long it has been since an upload was completed.

Drivers for Development of a Mineral Occurrence Data Model

Web services offer an ideal, cost-efficient technology for removing both the inconsistencies and the need for data to be regularly uploaded to the national database. Web services also offer a chance to access the latest and most up-to-date data from the originating agency and return the data in a consistent format; however, any such Web service requires an agreed-upon data exchange standard, and none existed.

The Mineral Occurrence Data Exchange Model

The model is a high-level data exchange model for mineral occurrences represented in a unified modeling language (UML) that can be extended to cover all Earth resources (fig. 1). It will enable data on mineral localities to be delivered live to the Australian Geoscience Portal and will also facilitate data transfer between government, industry, and other organizations. It will enable real-time access to the latest data from each Survey. Because it is a standard data model, it will also enable a more formal structure for reporting resources and reserves that can comply with national and internationally accepted reporting codes. The model will require that standard vocabularies be compiled for each attribute, and this is work in progress.

The model is compatible with Geoscience Markup Language (GeoSciML), the International Union of Geological Sciences (IUGS) developed language for exchange of geological map features, and uses patterns and features common to GeoSciML. These patterns are based on International Organization for Standardization (ISO) and Open Geospatial Consortium (OGC) standards using Geographic Markup Language (GML) as an Extensible Markup Language (XML) encoding for geographic information. In the ISO model, "features," or real-world objects of interest, are classified into types on the basis of a characteristic set of properties. GML provides few

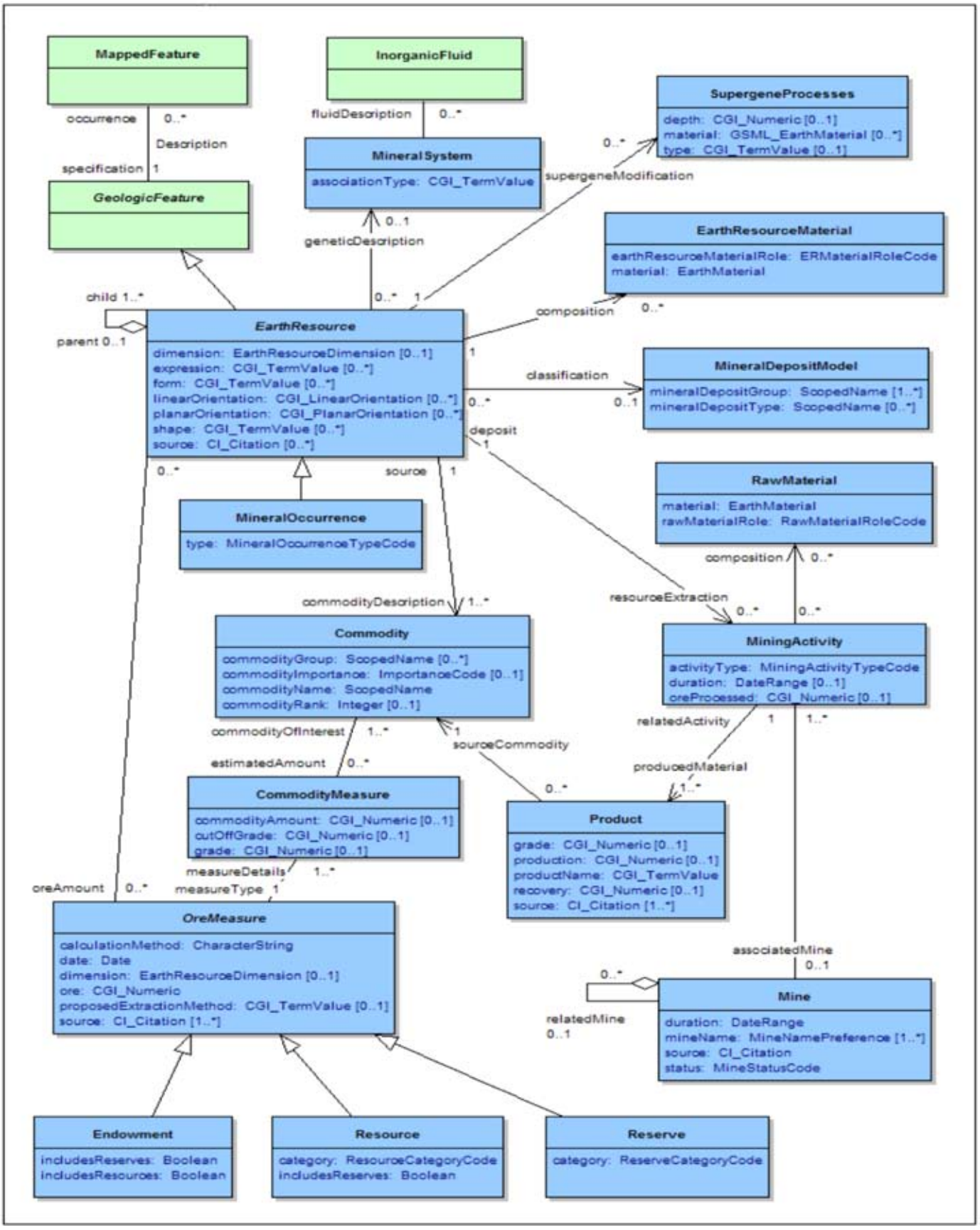

Figure 1. A UML diagram of the Mineral Occurrence Data Exchange Model.

concrete feature types directly, as these are intended to be created using the standard components in a domain-specific "GML Application Schema." "Mineral occurrences" is an example of a domain-specific schema. Model development took place in the graphical UML environment.

Below is a summary of the key points of the Australian Mineral Occurrence Model:

- The model describes earth resources independent of associated human activities (for example, mining).
- Caters to a description of earth resources using the following:
  - Mineral deposit models that describe the actual deposit type (encompassing the Cox and Singer classification);
  - Mineral systems that describe the processes associated with deposit formation; and
  - Supergene processes.
- Utilizes GeoSciML mapped feature to describe spatial representation.
- Utilizes GeoSciML earth material to describe host and associated materials.
- The model describes a mine as made up of a number of mining activities, each of which produce some commodity.
- The model provides the ability to describe commodity resources formally or informally, utilizing CRIRSCO (Committee for Mineral Reserves International Reporting Standards) and including basic JORC Code requirements (Joint Ore Reserves Committee—the 2004 Australasian code for reporting exploration results, mineral resources, and ore reserves).

Participating in Further Developments of the Model

The model can be accessed at https://www.seegrid.csiro.au/twiki/bin/view/Xmml/MineralOccurences.

Any interested parties are welcome to comment and perhaps participate in trying to extend and progress the model to becoming an international data exchange standard.

Emerging Web Services at the IRIS Data Management Center (DMC)

By Bruce R. Weertman[1], Joanna Muench[1], Linus Kamb[1], Rob Casey[1], and Tim Ahern[1]

[1]IRIS Data Management Center (DMC), Seattle, Wash.

IRIS Background

New demands for interdisciplinary science have broadened the data delivery mission for the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC). In the past, the DMC has been charged with archiving seismic waveform data and making it available to download primarily via batched mechanisms. In recent years we have seen the successful addition of a Common Object Request Broker Architecture-Remote Procedure Call (CORBA-RPC) mechanism named DHI (Data Handling Interface). The mission of IRIS has now been expanded to delivering more general data products, usable by nondomain experts. Many of the building blocks to create these data already exist, but in their current format are difficult to combine. At the DMC we are working to expose these existing capabilities and to develop new ones, linking them together with workflows to create a Web-based service-oriented architecture. The DMC began providing Web services two years ago, with pilot projects exposing data access and analysis through Web services. A new tool, SPADE (Searchable Product Archive and Discovery Engine), uses Web services both for data submission and discovery. Enabling interorganizational collaborations is one of the strengths of Web services, and IRIS has been working with the University Navigation Signal Timing and Ranging (NAVSTAR) Consortium (UNAVCO) and the Marine Geology Data System at Lamont-Doherty Earth Observatory (LDEO) on the Geoscience Web Services (GeoWS) project (http://www.geows.org). The GeoWS project aims to bring together datasets from the three organizations by means of Open Geospatial Consortium (OGC)-standard mapping technologies. Some of the DMC's existing capabilities that will be exposed as services include data format conversion, plotting, and phase analysis. New tools will include hypocenter and tomographic retrieval and visualization. Our long-term aim is to provide reusable, composable services with programmatic and interactive interfaces, enabling users to easily customize seismic data access. The flexibility of a service-oriented architecture will enable the IRIS DMC to respond effectively to technology changes and demands from the geosciences community.

IRIS Earthquake Hypocenter Web Service for use with the GEON IDV

The Geosciences Network (GEON) IDV (Integrated Data Viewer) (http://geon.unavco.org/unavco/IDV_for_GEON.html) is a powerful, free Java-based desktop application that allows 3D visualization of complex solid earth-science data. The GEON IDV can display earthquake hypocenters encoded in Network Common Data Form (NetCDF). We have developed a Representational State Transfer (REST)-style Web service for generating such data. The service can rapidly query our earthquake event tables for events in latitude-longitude bounding boxes and date ranges. Our event tables contain millions of events from multiple contributors and span the time range from 1964 to present. The Web service is coupled with an easy-to-use map-based Web application that allows users to quickly discover what events are in our tables and to download

the information in the NetCDF format. This should be a useful tool for educators, scientists, and students.

An Experiment Tool for the Multilayered Geoscience Ontology

By Kangping Sun[1] and Lei Wang[1]

[1]China University of Geosciences, Wuhan, China

One of the significant challenges towards the integration of geoscience datasets and query for the meaningful geological subjects (such as instances of the earth-material types or deposits) from a virtually integrated geoscience database, is the need to capture geoscience knowledge with which geologists will be comfortable. This paper reports on an experimental tool that supports compiling multilayered ontology (or concept space, as the community refers to it) and outputting geologic query concepts. The end users can browse the captured structural knowledge and query geological data in the integrated database via the geologic query concepts.

Basically, the tool supports the compilation of three kinds of geologic concept models. The first one is the hierarchical concept model, which defines what an independent geologic concept is and where it is within the hierarchical concept model. The definition methods of the model are very similar to the classification schemes of geoscience terminology. The classification scheme depends on the geologic term to be classified; thereafter, the definition methods form into the layers of the hierarchical model. The second one is the relational concept model, which establishes relationships among the independent geologic concepts with a more structural style. The last one is the query concept model, which builds up application-oriented subjects that take the concept localization or specification into account.

The tool can output "geologic concept queries" that look like the following: "Select Terrigenous-clastic material. Mudrock in the area Lat-long area.areaA" or "Select Wash-over-fan deposit with the description of lithofaceA or lithofaceB."

Initially, the tool is designed for the sedimentary materials: science language for their classification, description, and interpretation in digital geologic-map databases (North American Geologic Map Data Model Science Language Technical Team, 2004). Almost all the interested terms in it could be arranged into one hierarchy under the top-level geologic concept model. In fact, it can capture geologic knowledge only if the geoscientists agree upon the representations predefined in the tool. We believe the tool will be upgraded when more geoscience knowledge is captured.

Reference Cited

North American Geologic Map Data Model Science Language Technical Team, 2004, Sedimentary materials: science language for their classification, description, and interpretation in digital geologic-map databases, Version 1.0 (12/18/2004): Draft report posted on the North American Data Model Web site, http://www.nadm-geo.org/sltt/products.html, 595 p. (Accessed October 12, 2007.)

Data Independence and Geospatial Web Services

By Upendra Dadi[1] and Liping Di[1]

[1]Earth System Science, George Mason University, Fairfax, Va.

The importance of data independence in a database management system is well recognized in the database community. Data independence means that a database can be changed structurally without affecting programs processing the data. The concept of data independence can play a very useful role in the development of geospatial Web service architectures. We have created a prototype Web service platform on top of the Geographic Resources Analysis Support System (GRASS) geographic information systems (GIS) based on the concept of data independence. The paper presents the general architecture used in developing the Web services. This architecture consists of several layers of services. The interfaces of services at each layer are independent of the implementation of the lower level services. The geospatial modeler, who is also the developer of services at the topmost layers, is immune from having to learn the specifics of any one particular software package or service. The services at the higher levels can be cataloged and made discoverable for use in other higher level models. Some of the issues that we encountered—granularity of services, statelessness of services, and interoperability between services—are also discussed.

Layered Architecture for Geospatial Web Service Development

One of the important considerations in the design of the architecture is that the scientist or model developer should be immune to the details of the underlying implementation of the software or services. The model developer should be able to view the service interfaces available at the conceptual level without having to know much about the details of the implementation. At the same time, the flexibility and power of the underlying software should not be restricted when incorporated into Web services. In other words, a user of the services should be able to achieve all the functionality that can be achieved by using the software in a standalone manner. This

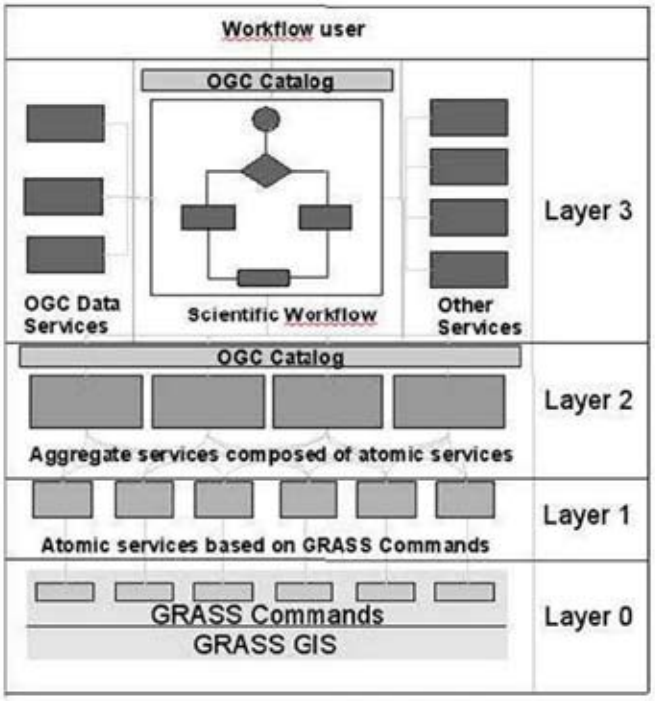

Figure 1. Layered architecture for Web service design.

idea of conceptual data independence has played a crucial role in the design and development of relational databases. Having a layered architecture for geospatial Web service development can play a similar role. At each layer are services composed by chaining services at the lower level or same level. The higher level services can be created in such a way that they completely hide specifics of the software and services used in the lower level. Figure 1 displays the layered architecture used in the development of Web Services on top of GRASS GIS. The reader is referred to for more details about GRASS GIS. GRASS GIS is at Layer 0 in the architecture.

Layer 1

Layer 1 consists of "atomic" Web services which are directly based on the GRASS commands. Each and every service interface emulates a GRASS command. Some simple mapping rules, depending on the pattern of the command line, were used when going from the command-line interface to the Web-service interface.

Most of the services in this layer are not self-contained, although they are independent of each other. In reality, however, Web services are supposed to be discrete units of code which are independent as well as self-contained. They are usually only loosely coupled with each other. In a later section, we will argue that services on top of session-based command-line software like GRASS are more naturally designed as Grid services which have state. The parameters used in the above GRASS services—DATABASE, LOCATION, and MAPSET—can be thought of as representing state information which is stored on client side and passed during each and every call to the service. We will see that in the next layer of the stack in figure 1, we can chain the atomic services, which can be either stateful or stateless, into higher-level Web services that are completely independent and self-contained. Specifics of GRASS can be completely hidden from the users of the Web services at this layer, unlike services in the present layer which require user-level knowledge of GRASS.

Layer 2

In this layer, several services from the layer below are chained together to create new services. These higher-level services are similar to GRASS scripts that are composed of several GRASS commands. The interfaces to these services are such that they can be looked at the conceptual level by the model developer, without regard to the implementation details. They follow directly from the definition of the parameter being modeled. This will be illustrated with a simple example. A service has been developed to calculate Normalized Difference Vegetation Index (NDVI) values from two raster data files—near infrared (NIR) and red images. The inputs to this service are two raster images—the first containing reflectance values from the visible range in Hierarchical Data Format-Earth Observing System (HDF-EOS) format and the second containing reflectance values from the NIR range, also in HDF-EOS format. The two images are assumed to have the same projection, resolution, and bounds. The output from this service is the image containing NDVI values in Portable Network Graphics (PNG) format. All the services used for creating this service—r_in_gdal, r_mapcalc, and r_out_png—are atomic services in Layer 1.

Layer 3

In this layer, services developed in Layer 2 are chained with other services to create even higher level services. The Layer 3 services can also consist of other Web services. The other Web services could be Web services based on other geospatial software. The OGC data services, such as WCS or WFS implementations, could be the data sources for the workflows at this layer. Now we will look at an example of a scientific workflow that is essentially a service at this layer. A simple model for calculating landslide susceptibility based on a few parameters, such as slope, aspect, NDVI, etc., has been implemented as a Layer 3 service. Figure 2 shows the model schematically.

The services developed in Layer 2 are chained with other services to create a workflow which represents the landslide model. In the image above, land cover is obtained from a Web Image Classification Service (WICS), which is a non-GRASS service. Each of the components in the model—NDVI, slope, aspect, landcover, and landslide susceptibility—are composed as independent Web services. These components form Level 2 services. They are chained together to create a workflow

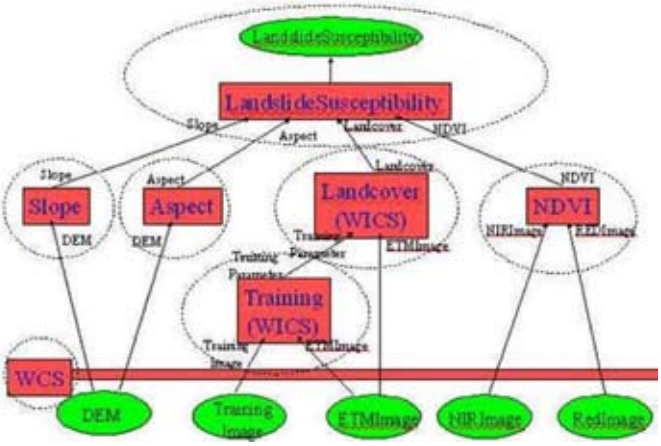

Figure 2. A simple landslide susceptibility model.

that represents a landslide model. While one can create the model from atomic services itself, to do that requires knowing the specifics of GRASS GIS and working at a lower level of abstraction. We used Business Process Execution Language (BPEL) for creating workflows.

Conclusion

This paper presents layered architecture for geospatial Web service development over GRASS GIS; however, similar architecture may be applicable when converting other geospatial software to Web services. Each geospatial software package has its own set of interfaces with which users can work. Usually these interfaces are designed for optimal functioning of the software. Bypassing these interfaces to develop coarse-grained services from the software may severely restrict the Web service developer from using the full capabilities of the underlying software; however, on the other hand, a model developer who uses these interfaces directly must learn the specifics of the software. A model developer does not and probably should not think about any specific software when developing a model; therefore, it is better to have multiple layers of users. At one layer is the user who is conversant with the interfaces. He or she can develop higher-level interfaces that completely hide the details of the underlying software and that are highly reusable across many different models. These service interfaces can be understood at the conceptual level by the model developer. The high-level interfaces would be much easier for a model developer to use than using implementation-dependent low-level interfaces. There is much scope for adding new functionality to the system developed so far. GRASS has many interactive commands. It is challenging to convert these commands to Web services. An intimate knowledge of the GRASS application programming interface (API) may be required. So far, there is no software that can act as a client to the services developed. For the end user to be able to use the services, the development of a general purpose and an extensible geospatial Web service client is needed.

Using WDO-it to Build a Geoscience Ontology

By Paulo Pinheiro da Silva[1], Leonardo Salayandia[1], and Ann Q. Gates[1]

[1]Department of Computer Science, University of Texas at El Paso, El Paso, Tex.

Introduction

Workflow-Driven Ontologies (WDOs) are an approach to ontology development based on scientist-level terminology such as “dataset” and “methods,” which claims to facilitate the scientific process of encoding knowledge from their domains (Salayandia, Pinheiro da Silva, Gates, and Salcedo, 2006). In addition, resulting ontologies produced from using the WDO approach may include properties that enable the automatic generation of suggested workflow specifications. These suggested workflow specifications, once refined and endorsed by scientists, can be used as a training tool, since they provide a graphical representation to which the scientist can easily relate. The endorsed workflow specifications can be refined into fully computable workflows that facilitate the discovery and integration of resources available over cyberinfrastructures.

WDO-it is a prototype tool developed in Java that supports the WDO approach. Ontologies created with the WDO-it tool are encoded in the Ontology Web Language (OWL). OWL is the standard ontology language proposed by the World Wide Web Consortium (W3C) and the main framework language used by the semantic Web community (OWL, 2004).

In this abstract, we discuss the creation of an ontology about gravity data processing as defined and used in the domain of geophysics. The ontology, called GravityWDO, has been created with the current version of the WDO-it tool available at http://trust.utep.edu/ciminer/software/wdoit/. The current state of WDO-it allows suggested abstract workflow specifications to be generated from the knowledge provided by scientists. We call these abstract workflow specifications Model-Based Workflows (MBWs), since they are instantiations on an abstracted workflow model. A graphical representation of the MBWs can be visualized and can serve as a training tool by itself. Additional work is underway that will allow the MBWs to be migrated to an executable workflow language, such as the Modeling Markup Language (MoML), the language that the Kepler scientific workflow engine uses to represent workflows (Ludäscher and others, 2005). The claim is that scientists can relate easier to MBWs than to real executable workflows because they describe only essential properties that are required at the scientist level. Nevertheless, WDO-it will provide mechanisms to use MBWs as the basis to create executable workflow specifications.

Building Ontologies Using WDO-it

WDO-it provides three basic modes for building ontologies: (1) brainstorming mode; (2) harvesting mode; and (3) relation elicitation mode. In the brainstorming mode, scientists have the opportunity to enter concepts from his or her domain of interest, where these concepts are classified as either information concepts or method concepts. WDO-it does not use the term "concept" in its user interface; instead, it provides a very simple interface where scientists can see that information concepts range from raw data (for example, concepts that represent data measured in the field) to products (for example, concepts that represent models or maps of interest to the scientist). Moreover, the interface provides a way for entering method concepts that represent the algorithms, applications, and tools that are used to retrieve or transform information concepts. For example, an application that retrieves gravity data from a database about a specified region of interest can be classified as an information-retrieval concept. A tool that employs the nearest-neighbor algorithm to create a grid of uniformly distributed data points from a collection of scattered points can be classified as an information- transformation concept. These concepts, however, can be already specified in some existing ontology that a scientist may want to reuse. In this case, concepts in the existing ontology can be imported and later classified into information and method accordingly. This is referred to as the harvesting mode.

Any time after at least a method is created in the brainstorm or harvesting mode, the scientist can switch to the relationship elicitation mode to identify relationships between the information concepts and the method concepts. These relationships are of the type "IsInputTo" and "IsOutputFrom," where a scientist indicates which information concepts are required as input to a method concept, and which information concepts are the output of a given method concept. Figure 1 shows a snapshot of the ontology relationship tab. Notice that the tool does not show all the relationships available in the ontology being created; instead, the user selects a method concept, and then the input-information concepts, so that the output-information concepts are shown for the selected method concept only. By focusing on one method at a time, the scientist can have better control of the relationships between concepts, instead of seeing a cluttered diagram that shows all relationships between all concepts, which is typical of other general-purpose ontology editor tools. Figure 1 shows the creation of a gravity ontology, where the scientist selects a method called gridding, and for which the information concept "CompleteBouguerAnomaly" is shown as its only input; the concept grid is shown as its output.

Additionally, the scientist can create new types of properties that can be used to customize relationships between concepts. For example, a scientist may create a "HAS" property that can be used in a relationship to indicate that a "SeismicEvent" has a location and a time concept related to it. This functionality is available through the Advanced View button shown in figure 1. Moreover, because WDOs are OWL ontologies, more generic ontology editors, such as Protégé (Gennari, and others, 2002) and Semantic Web Ontology Overview and Perusal (SWOOP; http://www.mindswap.org/2004/SWOOP/) can be used.

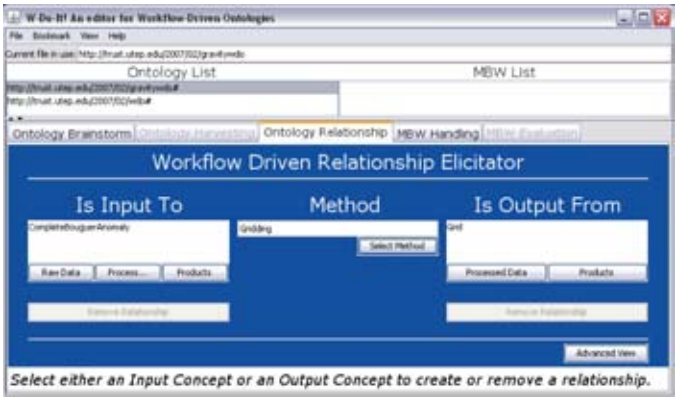

Figure 1. WDO-it tool; ontology relationship tab.

Generating Workflow Specifications through WDO-it

Once the scientist has built a WDO about their domain, WDO-it can use this WDO to automatically generate a suggested workflow specification for a given information concept of interest. For example, if a scientist is interested in obtaining a workflow that would describe the necessary steps to create a grid of gravity data, the scientist would choose the concept of interest from the information concepts available in the captured knowledge about gravity data (in other words, a gravity ontology). The WDO-it tool creates a suggested workflow specification (in other words, an MBW) based on the relationships available in the captured knowledge (Salayandia, Pinheiro da Silva, Gates, and Rebellon, 2006). The scientist is presented with a graphical representation of the workflow, and the workflow specification can be saved as an OWL file that is separate from the WDO. The MBW is not formally considered a workflow specification until it is endorsed by a scientist as an accurate representation of a process in the scientist domain. Corrections and refinements may be needed for the scientist to endorse the MBW. The WDO-it evaluation mode is responsible for enabling the scientist to critique, refine, and endorse suggested MBWs.

Figure 2 shows a snapshot of the workflow generator tab of the WDO-it tool and the resulting diagram generated for the grid-information concept, according to the knowledge captured in the loaded gravity ontology. Notice that there are some methods that have multiple inputs. Multiple inputs going into a method go through an "AND" method, indicating that all inputs are necessary for the given method to produce a given output. Exclusive-OR (XOR) operators are also used for the case where there can be different inputs to a method, but only one of them is needed for the method to create an output.

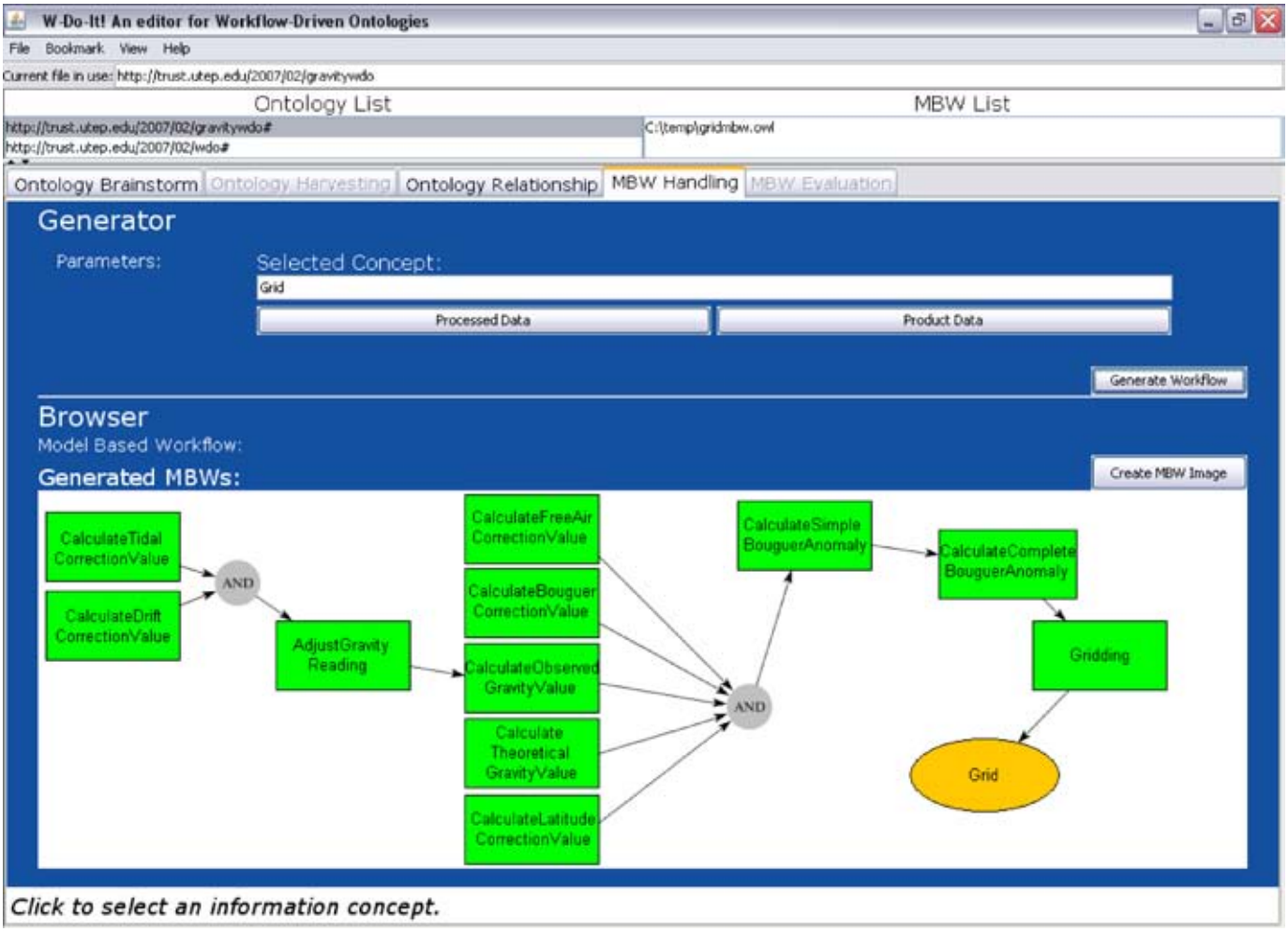

Figure 2. WDO-it tool; workflow generator tab.

References Cited

Gennari, J., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubezy, M., Eriksson, H., Noy, N.F., and Tu, S.W., 2002, The evolution of protégé: An environment for knowledge-based systems development, 32 p.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., and Zhao, Y., 2005, Scientific workflow management and the Kepler system, *in* Concurrency and computation: Practice and experience, special issue on workflow in grid systems: New York, John Wiley and Sons, Ltd., v. 18, no. 10, p. 1039–1065.

McGuiness, D.L., and van Harmelen, F., 2004, OWL Web ontology language overview: World Wide Web Consortium (W3C) Web site http://www.w3.org/TR/owl-features/. (Accesssed October 12, 2007.)

Salayandia, L., Pinheiro da Silva, P., Gates, A.Q., and Salcedo, F., 2006, Workflow-driven ontologies: An earth sciences case study, *in* Proceedings of the Second International Conference on e-Science and Grid Computing, Amsterdam, Netherlands, December 2006: Washington, D.C., IEEE Computer Society, 1 CD-ROM.

Salayandia, L., Pinheiro da Silva, P., Gates, A.Q., and Rebellon A., 2006, A model-based workflow approach for scientific applications, *in* Gray, J., Tolvanen, J.-P., and Sprinkle, J., eds., Proceedings of the Sixth OOPSLA Workshop on Domain-Specific Modeling (DSM'06): University of Jyvaskyla [Finland] Computer Science and Information System Reports, Technical Report TR-37, available online at http://www.dsmforum.org/events/DSM06/Papers/20-Salayandia.pdf.

WXGURU: An Ontology-Driven Chatbot Prototype for Atmospheric Science Outreach and Education

By Rahul Ramachandran[1], Sunil Movva[1], Xiang Li[1], Prashanth Anantharam[1], and Sara Graves[1]

[1]Information Technology and Systems Center, University of Alabama at Huntsville, Huntsville, Ala.

Chatbots are computer programs designed to simulate an "intelligent" conversation with one or more human users via auditory or textual methods. These programs typically scan user inputs for keywords and then extract a reply with the most matching keywords or the most similar wording pattern from a local database. Chatbots have been successfully used in industry as virtual customer-service assistants, guides, alternative to FAQs, and so on. One such chatbot is A.L.I.C.E. (Artificial Linguistic Internet Computer Entity), which uses Artificial Intelligence Markup Language (AIML) to encode conversations. AIML is an XML-compliant language with a formal specification and W3C XML schema. AIML contains elements to define patterns to match user inputs and the subsequent replies. WxGuru (Weather Guru) is a clone of A.L.I.C.E., designed as an educational chatbot for atmospheric science. WxGuru is unique because its conversational patterns have been coupled with a reasoner loaded with atmospheric ontology. This coupling allows WxGuru to provide useful and knowledgeable replies to user queries regarding any atmospheric science questions. WxGuru design and implementation will be presented in this poster.

The U.S. National Geologic Map Database

By David R. Soller[1], Harvey Thorleifson[2], and Nancy Stamm[1]

[1]U.S. Geological Survey, Reston, Va.

[2]Minnesota Geological Survey, University of Minnesota, St. Paul, Minn.

Since the mid-1990s, the U.S. National Geologic Map Database (NGMDB, http://ngmdb.usgs.gov) project has systematically addressed its Congressional mandate to develop scientific and technical standards and to provide a national archive of geoscience map information. Under this mandate in the Geologic Mapping Act of 1992, the U.S. Geological Survey (USGS) and the Association of American State Geologists (AASG) have made significant advances in the design and development of the NGMDB. (See yearly reports of progress at http://ngmdb.usgs.gov/Info/reports/.)

The NGMDB's goal is to help users find the information they need to address a variety of societal and research applications. Because our users' range in interest and expertise, from the general public to the geologic mappers and geographic information systems (GIS) specialists who prepare maps and databases, the NGMDB project began in 1996 by building a set of fundamental resources and databases that include the following: (1) a careful and responsive customer service capability; (2) a comprehensive geoscience map catalog of nearly 79,000 products by 350 publishers; (3) the U.S. Geologic Names Lexicon ("GEOLEX"), a standard reference for the Nation's stratigraphic nomenclature; and (4) proceedings from the 10 annual Digital Mapping Techniques (DMT) workshops. The DMT workshops have provided a unique venue for discussion of map and database-preparation techniques and Web delivery of geospatial information, and have facilitated convergence toward common practices and toward science and technical standards for the geosciences. These resources receive about 140,000 visits per month from 35,000 users.

Society, businesses, and private citizens commonly are faced with complex, multidimensional issues; in order to facilitate the use of geologic information and its integration with other types of information (for example, soils, engineering, hydrologic, and cultural), it must be presented in a form that is readily comprehensible to the "nongeologist." In other words, the presentation of geologic information must, to some extent, be standardized. Geological survey agencies produce individual maps, reports, and datasets in a wide variety of formats and layouts, each containing specialized scientific terminology. Without a doubt, these have proven immensely valuable to our users; however, with the advent of GIS and Web services, our users demand access to a more integrated, comprehensive set of geologic maps and reports. With this in mind, the NGMDB project has extensively collaborated with U.S. and Canadian agencies to develop essential standards; these include the North American Data Model (NADM) geologic map data model and lithologic terminologies, and the Federal Geographic Data Committee (FGDC) standard for map symbolization and for specifying the locational accuracy of mapped features. NGMDB also participates in the development of the emerging international standard for exchange of spatial geoscience information (Geoscience Markup Language; GeoSciML).

A principal goal for the NGMDB project is to design and build a database of richly attributed, standardized geospatial information in vector and raster formats. That database is intended to be a distributed system, with nodes hosted by the State geological surveys and the USGS, and integrated with the existing NGMDB databases described above. Because of this project's requirement to build a national archive that can contain geoscience maps from all geological surveys in the U.S., a preparatory period was required in order to (1) fully discuss among the many participating agencies the technical and scientific options for building such a database; (2) forge a collaboration among these agencies; and (3) agree upon the necessary standards. During this process, essential concepts were prototyped in order to support discussion on how to proceed (see http://pubs.usgs.gov/of/2001/of01-223/soller2.html and http://pubs.usgs.gov/of/2002/of02-202 /). These concepts included the requirement for the NGMDB system to

enable spatial analyses that combined data from the NGMDB and other types of databases; this was viewed as essential to the wider dissemination and use of geologic information by nongeologists who may need, for example, to compare and analyze demographic, hydrologic, engineering, and geologic data. The project is now building a prototype database and portal for this distributed system; we anticipate it will be publicly available in 2007.

A Data Integration and Interoperability Blueprint for USGS

By Kevin T. Gallagher[1], R. Sky Bristol[2], and Linda C. Gundersen[1]

[1]U.S. Geological Survey, Reston, Va.

[2]U.S. Geological Survey, Denver, Colo.

Overview

The U.S. Geological Survey (USGS) Geospatial Information Office is currently leading efforts to develop a long-term plan or "blueprint" for data integration, accessibility, discovery, and interoperability across the USGS. The data integration blueprint (fig. 1) will include projects that provide metadata, data content standards, infrastructure, and informatics that enhance scientific techniques, improve data access, provide management visibility, advance the strategic directions of the USGS science strategy, and connect the USGS to its partners and collaborators through participation in international efforts to develop a global science and computing platform for the 21st century. The plan will be comprehensive by design, incorporating the data integration and scientific tools development efforts of all USGS into a single framework with common practices and a seamless infrastructure. The USGS will work with its partners and national and international cyberinfrastructure activities to develop this framework.

Background

The USGS is a world leader in monitoring, assessment, and research related to the natural sciences. Coupled with a diverse multidisciplinary workforce, extensive monitoring networks, and national- and regional-scale approaches, the USGS has carved out a reputation for being the "authoritative source" of specific national datasets, such as water quality, cartographic bases, land cover and land use, biological resources, and geologic mapping. As the future unfolds, the USGS's ability to map and integrate this data will be critical for the advancement of all science directions. Some of the major mission activities that USGS engages in include the following:

- Collect and maintain long-term national and regional geologic, hydrologic, biologic, and geographic databases.

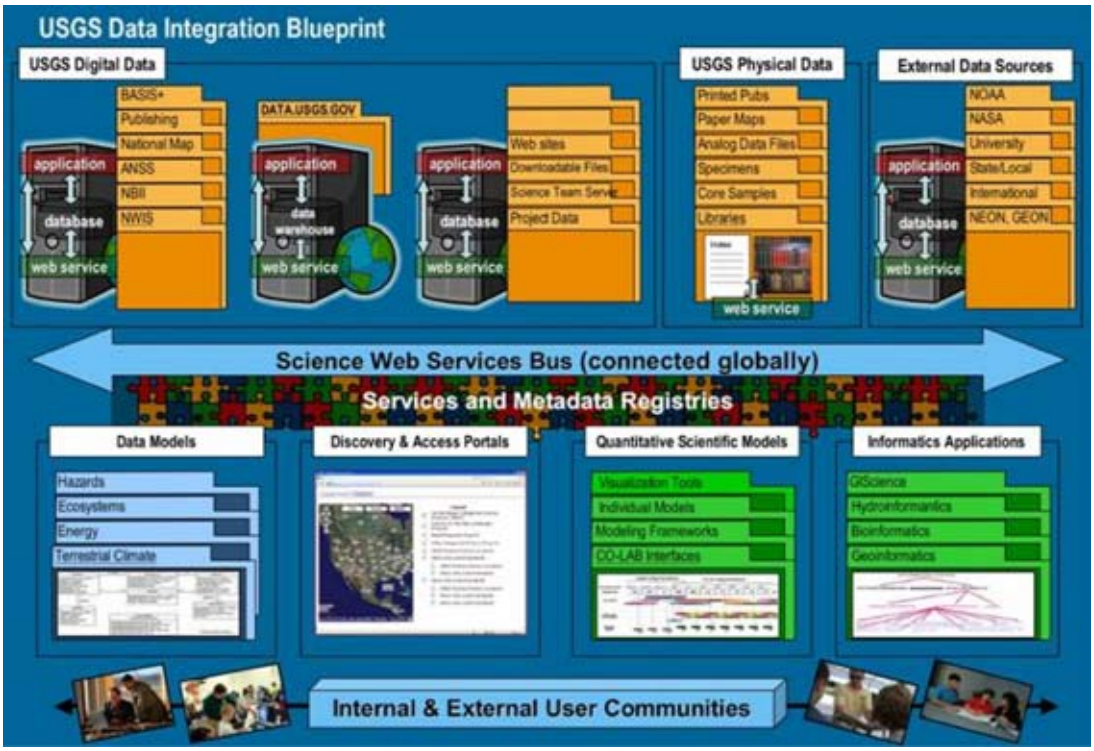

Figure 1. Schematic of a service-oriented architecture for integration of USGS data.

- Collect, process, and analyze earth and planetary imagery and remote sensing.
- Develop open-source models of complex natural systems and human interaction with those systems.
- Maintain national and global geologic, biologic, hydrologic, and geographic monitoring systems.
- Archive and preserve physical collections of earth materials, biologic materials, reference standards, geophysical recordings, and paper records.
- Develop standards of practice for the geologic, hydrologic, biologic, and geographic sciences.

The USGS maintains a large number of science datasets at local, regional, and national scales. The USGS ability to integrate this data is critical to the achievement of Department of Interior (DOI) mission objectives in resource protection, resource use, and serving communities, and the USGS national Federal mission of conducting science and serving earth and biological data. Development of a fully integrated science data environment will improve the accessibility of science data and information within the USGS, across the DOI, and with its scientific partners, collaborators, and customers in other Federal agencies and the public. Greater access to a broad range of integrated science data will spark new discovery and support a wider range of inquiry, better informing and enhancing the decisionmaking of managers, policymakers, and stewards of the Nation's resources.

Examples of some of the long-term national datasets maintained by USGS include the following:

- The National Map (topography, orthoimagery, hydrography, and so on);
- MRDATA (comprehensive source of mineral resource data);
- The National Geologic Map Database (a standardized community collection of geologic mapping);
- NWISWeb (the National Water Information System);
- The National Geochemical Database (collection of rock data, stream-sediment data, and data on other materials analyzed by the USGS);
- National Geophysical Database (aeromagnetic, gravity, and aeroradiometric data);
- National and Global Earthquake catalogs;
- North American Breeding Bird Survey;
- National vegetation and speciation maps;
- National Oil and Gas Assessment; and
- National Coal Quality Inventory.

The conduct of science is changing worldwide. There is widespread recognition that the Earth's complex natural systems are interrelated and that scientific inquiry must be equally integrated to develop new understanding of the implications for the environment, land management, resource utilization, and policymaking. Complex scientific questions require the analysis, integration, and modeling of science data and information from multiple disciplines, locations, and timeframes. The USGS and its partners, including industry, Federal, State, and local governments, universities and associations, as well as international scientific organizations, are beginning to connect and integrate the data and research techniques of the world's scientists, making them accessible to a global science community and transforming the way in which research, engineering, and education are conducted. Science data integration within the USGS is a prerequisite for joining these international efforts to develop a worldwide science collaboration and computing platform that can address future environmental science challenges.

For example, phenology is the study of periodic plant and animal life cycle events that are influenced by environmental changes, especially seasonal variations in temperature and precipitation driven by weather and climate. Phenological events record—immediately and empirically—the consequences of environmental variability and change vital to the public interest. Variability in phenological events, such as the beginning of the growing season, can have important environmental and socioeconomic implications for the economy, health, recreation, agriculture, management of natural resources, and natural hazards. Although phenology is a far-reaching component of environmental science, it is not well understood. The predictive potential of phenology requires a new data resource—a national network of integrated phenological observations. A USA National Phenology Network (USA-NPN) is currently being designed and organized to engage Federal agencies, environmental networks and field stations, educational institutions, and mass participation by citizen scientists. The initial phase will establish a continental-scale network focused on phenological observations of a few regionally appropriate native plant species and nationally cultivated indicator plants. The USGS must not only integrate its scientific data to support this effort, but must also integrate data from other monitoring activities, such as water availability and soil chemistry, to inform larger national issues, such as climate change and ecosystems restoration.

Some of the national and global monitoring systems that the USGS maintains include the following:

- National Stream Flow Information Program;
- Advanced National Seismic System;
- National Volcano Early Warning System;
- Debris Flow Warning System;
- Global Terrestrial Network for Permafrost;
- Landsat 5 and 7;
- Biomonitoring of Environmental Status and Trends;
- National Bird Banding Program;
- Land Cover/Land Change Monitoring;
- Famine Early Warning System; and
- National Water Quality Assessment Program.

Long-Term Vision

In 2006, the Director of the USGS chartered a team to develop a new USGS science strategy. That strategy, entitled "Facing Tomorrow's Challenges: USGS Science in the Coming Decade," was released in April 2007 and includes six major science goals and a special chapter on "New Methods of Investigation and Discovery" that provides the following

long-term vision for USGS data integration: The USGS supplies an information environment where diverse and distributed knowledge is accessed and used seamlessly by scientists, collaborators, customers, and the public to address complex natural science issues.

The USGS science strategy also lays out the following strategic actions to accomplish this long-term vision:

- Incorporate planning for long-term data management and dissemination into multidisciplinary science practices.
- Adopt and implement open data standards within USGS and contribute to the creation of new standards through international standards communities.
- Develop and implement a comprehensive scientific cataloging strategy that incorporates existing datasets, resulting in an integrated science catalog.
- Develop a sustainable data-hosting infrastructure to support the retention, archiving, and dissemination of valuable USGS datasets in accordance with open standards.
- Develop and enhance tools and methods that facilitate the capture and processing of data and metadata.
- Identify and support authoritative data sources within USGS programs and encourage development and adoption of standards.
- Build and strengthen the internal workforce augmented by external partnerships in environmental information science.
- Identify and leverage national and international efforts that promote comprehensive data and information management and foster greater sharing of knowledge and expertise.
- Partner with collaborators and customers to facilitate data integration across the worldwide science community.
- Partner with collaborators and partners in the development of informatics tools and infrastructure that contribute to the evolving global science computing and collaboration platform.

The last three strategic actions are key to successful creation of an international cyberinfrastructure for the sciences. One of the ways to achieve this collaboration is through the creation and participation in "communities of practice." A community of practice is not merely a community with a common interest, but it comprises practitioners who share experiences and learn from each other. They develop a shared repertoire of resources: experiences, stories, tools, vocabularies, and ways of addressing recurring problems. This takes time and sustained interaction. Standards of practice and reference materials will grow out of this experience; however, the critical benefits include the following: creating and sustaining knowledge; leveraging of resources; and rapid learning and innovation.

Constructing an International Geoscience Interoperability Testbed to Access Data from Distributed Sources: Lessons Learned from a GeoSciML Testbed

By O. Raymond[1] and Interoperability Working Group GeoSciML[2]

[1]Geoscience Australia, Canberra, Australia

[2]Commission for the Management and Application of Geoscience Information, International Union of Geological Sciences (IUGS)

Introduction

Geoscience data are being generated at exponentially increasing volumes, and it is no longer feasible to develop centralized warehouses from which data are accessed. Efficient access to such data online in real time from distributed sources is rapidly becoming one of the major challenges in building cyberinfrastructures for the earth sciences.

Extensible Markup Language (XML) coupled with Web-based data delivery is a proven technology which allows access to standardized data "on the fly" via the Internet. GeoSciML (Geoscience Markup Language) is a geoscience specific, XML-based, GML (Geography Markup Language) application that supports the interchange of geoscience information. It has been built from various existing geoscience data model sources, particularly the North American Data Model (NADM) and Extensible Mining Markup Language (XMML). It is being developed through the Interoperability Working Group of the Commission for the Management and Application of Geoscience Information (CGI), which is a commission of the IUGS. The working group (currently) consists of geology and information technology specialists from agencies in North America, Europe, Australia, and Asia.

The GeoSciML Testbed

In 2006, representatives from geological surveys in the United States, Canada, the United Kingdom, France, Sweden, and Australia came together to develop a testbed that would utilize GeoSciML to access globally distributed geoscience map data (Duffy and others, 2006).

Data was served from seven sites in six countries with several different Web Feature Service/Web Map Service (WMS/WFS) software solutions employed. Geological surveys in Canada, the United States, and Sweden used an Enviromental Systems Research Institute ArcIMS platform (and in one case a MapServer platform) with a Cocoon wrapper to handle queries and transformations of XML documents. The British and Australian Geological Surveys employed the open-

source GeoServer software to serve data from ArcSDE and Oracle sources. The French Geological Survey implemented a system using an Ionic RedSpider server for WMS and client, and a custom development to implement a WFS. Web clients were constructed in Vancouver, Canada, using Phoenix, and later in Canberra, Australia, using Moxi Media Internet Mapping Framework (IMF) software to test various use case for the WMS/WFS services. Generic Web clients, such as Carbon Tools' Gaia 2, were also used to test some use cases.

In addition to geologic map data, the testbed also demonstrated the capacity to share borehole data as GeoSciML. Two WFSs (French and British) provided borehole data to a client able to display the borehole logs.

System (Open Geospatial Consortium) Compliance

There are three important things to consider when establishing an Open Geospatial Consortium (OGC)-compliant interoperability testbed—compliance, compliance, and compliance; however, working at the cutting edge of WFS implementation in the GeoSciML Testbed strained existing WFS software implementations to the breaking point. Approaching the deadline of the public release of the Testbed, rigorous OGC standards compliance became an unrealistic goal. The focus of the project necessarily turned from OGC standards compliance to ensuring a useful degree of data exchange, including display, download, and some simple query functionality.

In comparison with other data types that have successfully used OGC services, GeoSciML deals with extremely complex data; thus, although the GeoSciML Testbed did prove that it is possible to make geoscience interoperable, it also showed that semantic compliance is not going to be a trivial exercise, particularly for the more descriptive components of earth sciences.

Proprietary vendor and open-source software that aims to fully support the detail of OGC Web-service specifications is still at the developmental stage. The complexity of both the WFS query framework and the XML implementation model make implementation of such software an onerous task. It may eventually be found that WFS as a generic query framework over an XML model of GeoSciML's complexity is not achievable; however, this can only be tested by presenting the OGC standards with well conceptually modeled schema in a real domain, such as geoscience's GeoSciML.

Support for a subset of WFS services was achieved in the GeoSciML Testbed, but there is no standard mechanism to expose or describe the set of functionality that is implemented. In time, vendor and (or) open-source software will likely provide more rigorous and powerful WFS software implementations. The GeoSciML Testbed proved an effective mechanism to push further development of software capability in this area.

Semantic Compliance

The GeoSciML Testbed highlighted firstly the importance of strict compliance to standard vocabularies of controlled concepts for true interoperability, and, secondly, the complexity of the concepts that we were tying to standardize and make readable by computers. Humans easily cope with a degree of fuzziness in data structures or ontologies. It is in our nature that many geologists cannot see the problem with attributing a sandstone as "cross bedded" or "cross-bedded;" however, it is vital to computer-based queries of digital data.

A lot of work is still to be done (and is underway) in the vocabulary arena to make data exchange and query more interoperable. A geologist knows that an "igneous extrusive" rock and a "volcanic" rock are the same thing, but a computer searching for volcanic rocks will not find rocks coded as "igneous extrusive" unless rules of equivalence and hierarchy are established in complex vocabularies. As with many other international initiatives for sharing information, the multilingual aspect has to also be taken into account in any vocabulary development.

So, you have a data model. It is entirely, scientifically logical and robust, with complex hierarchical structure and vocabularies to accommodate your complex and hierarchical data, but, just how practically interoperable is it?

Participants in the GeoSciML WFS Testbed all provided information on the age of the geological units that they served; however, the schematic flexibility of the GeoSciML data model allows services to provide their age information in fully compliant, yet in slightly different ways—as single terms, as multiple hierarchical terms, and as maximum and minimum terms. This meant that querying and reclassifying the data based on age information had to be done differently on each dataset without the ability to apply a single standard query to all the GeoSciML datasets.

Usability issues such as these will only be solved with the increasing maturity of emerging complex scientific spatial data models like GeoSciML. Use cases for data models and WFS services must be developed recognizing the capabilities of existing and future WFS/WMS and GIS software, as well as scientific user needs. However, as the GeoSciML Testbed showed, some of the limiting factors in a cutting edge project do not become apparent until the project is well underway.

Implications of the GeoSciML WFS Testbed

While WFS/WMS standards, data models and supporting software are still being developed, demonstrator projects such as the GeoSciML Testbed are vital to the progress of interoperability to show users that the technology can deliver access to distributed data sources in real time. Further testbeds for GeoSciML will result in more robust and functional WFS/WMS services that will become mainstream data delivery services in the near future. Above all, this testbed highlighted the complexity of geoscience data and showed that strict adher-

ence to controlled vocabularies is essential to making Geoscience data semantically interoperable.

Reference Cited

Duffy, T., Boisvert, E., Cox, S., Johnson, B.R., Raymond, O., Richard, S.M., Robida, F., Serrano, J.J., Simons, B., and Stolen, L.K, 2006, The IUGS-CGI International Geoscience Information Interoperability Testbed, International Association for Mathematical Geology, Eleventh International Congress, Liege, Belgium.

The GEON LiDAR Workflow as a Distribution Pathway for Community LiDAR Topography Datasets

By Christopher J. Crosby[1], Frank Efrat[2], J. Ramon Arrowsmith[1], Viswanath Nandigam[2], Han Suk Kim[3], Jeffrey Conner[1], Ashraf Memon[2], Chaitan Baru[2], and Newton Alex[1]

[1]School of Earth and Space Exploration, Arizona State University, Tempe, Ariz.

[2]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[3]Department of Computer Science and Engineering, University of California—San Diego, La Jolla, Calif.

Due to the growing awareness of the power of high-resolution topography from light detection and ranging (LiDAR) data for earth-science research, it is becoming increasingly common for these datasets to be acquired as a community resource. The forthcoming GeoEarthScope LiDAR topography acquisition is an excellent example of this trend. The GeoEarthScope acquisition offers an unprecedented opportunity to examine the strain field adjacent to major active faults in the western United States at time scales greater than those provided by the Plate Boundary Observatory geodetic instrumentation. Modeled after the recent B4 data acquisition, the GeoEarthScope LiDAR data is expected to provide digital elevation models (DEMs) of one meter or better spatial resolution with scientific-grade geodetic accuracy. These datasets will be exceptionally valuable for geologic slip rate studies, paleoseismic research, and as a pre-earthquake representation of the landscape should an event occur in the near future. These datasets will be utilized extensively and they must be available to the EarthScope community as quickly and as easily as possible.

Traditionally, access to community LiDAR datasets has been difficult because of the massive volumes of data generated by LiDAR technology. For example, the recently acquired B4 dataset covers nearly 1,000 kilometers of the southern San Andreas and San Jacinto faults and contains approximately 3.7 billion individual LiDAR returns. With the B4 dataset as a model, the tremendous volume of data generated by the forthcoming GeoEarthScope LiDAR acquisition effort could potentially be a significant barrier for user-community access and processing of these data.

In order to address the challenges posed by the distribution and processing of community LiDAR datasets, we have applied a geoinformatics approach that capitalizes on cyberinfrastructure developed by the GEON project (http://www.geongrid.org). The Internet-based resource we have developed, the GEON LiDAR Workflow (GLW), is designed to democratize access to these challenging datasets and provides tools to enable users to perform basic processing (for example, DEM generation) of the data. As a proof of concept, we have made four community LiDAR datasets available, including the B4 data, via the GLW. Our approach utilizes a comprehensive workflow-based solution that begins with user-defined selection of a subset of point data and ends with download and visualization of DEMs and derived products. In this workflow, users perform point cloud data selection, interactive DEM generation and analysis, and product visualization, all from an Internet-based portal. This approach allows users to carry out computationally intensive LiDAR data processing without having appropriate local resources.

In its proof-of-concept capacity, the GEON LiDAR Workflow has proven to be a valuable and innovative community gateway for accessing LiDAR topography. As of April 2007, the GLW had 126 users who processed over 10.5 billion LiDAR returns in 1,250 unique processing requests.

As a result of this success, the GLW has been selected as the distribution pathway for the forthcoming GeoEarthScope LiDAR datasets. In order to prepare for these new data, we are currently in the process of migrating the system from its current proof-of-concept implementation to a fully robust, production-level, community data portal. As part of this migration, we are working on a number of enhancements that include improving system stability, documentation, and portal usability, as well as adding processing capacity and providing new job-monitoring and job-archiving capability.

The distribution of GeoEarthScope LiDAR topography via the GEON LiDAR Workflow represents an excellent example of the utilization of cyberinfrastructure to facilitate access to computationally challenging community datasets.

Integrating Geologic Data in the NGMDB, a Standards-Based Approach Using GeoSciML and Open-Source Tools

By David Percy[1], Stephen Richard[2], David Soller[3], and Jon Craigue[4]

[1]Department of Geology, Portland State University, Portland, Oreg.

[2]Arizona Geological Survey, Tucson, Ariz.

[3]U.S. Geological Survey, Reston, Va.

[4]University of Arizona and U.S. Geological Survey, University of Arizona, Tucson, Ariz.

The National Geologic Map Database "phase three" prototype builds substantially on the project's 10-year effort to develop standards (for example, contributions to development of the North American Data Model, or NADM) for a common data structure and controlled science terminology for geologic map data. Recently, this effort has been incorporated into an international standard proctored by the IUGS Commission for the Management and Application of Geoscience Information (CGI) with the resulting Open Geospatial Consortium (OGC) GML-derived standard, GeoSciML.

GeoSciML is a transport mechanism and schema developed under the auspices of the CGI and demonstrated successfully in late 2006 at the International Association for Mathematical Geology conference in Liege, Belgium. In that demonstration, geologic databases from worldwide participants were queried by a desktop client (not browser-based) to show a consistent set of geologic data across disparate datasets from different countries or agencies.

The phase three prototype of the NGMDB integrates data from Arizona, the Pacific Northwest (Oregon, Washington, and Idaho), as well as from several national datasets. Additionally, we demonstrate interoperability with other standards-based services, such as the Natural Resources Conservation Service's (NRCS) soils database and National Aeronautics and Space Administration's Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data.

The prototype is enabled by a custom-built data-import tool that allows for matching data in fields from an input-database source, such as Oregon or Washington geologic compilations, to a database schema used as a back end for the Web map service (U.S. Geological Survey's National Geologic Map Database-lite or NGMDB-lite). The NGMDB-lite schema is a flat-file view of a subset of data from the NGMDB database design (Richard and others, 2004). Fields in the input table are matched to corresponding fields in NGMDB-lite. Subsequently, unique values from each input field are matched to corresponding terms in controlled vocabularies defined by the NGMDB.

After matching fields and terms from the input map database to the NGMDB-lite schema and vocabulary, a fully attributed Enviromental Systems Research Institute shape file is generated. This shape file is then appended to the an aggregated shape file master table for display in an open-source mapping framework developed at Portland State University and managed on SourceForge as the project Map-Fu (http://sourceforge.net/projects/map-fu/).

Map-Fu is an Asynchronous Java Script and XML (AJAX)/Asynchronous Java Script and JSON (AJAJ)-implemented browser front end to map data that uses JavaScript and PHP/Mapscript to interact with individual users to handle requests for map data. It includes specific tools for zooming, panning, querying, and rendering individual map layers. It is a "thick client" that runs in all modern Internet browsers; thus, it is inherently cross-platform.

On the server side, we implement an open-source stack that consists of Mapserver and PostGIS (a set of geographic information systems extensions for PostgreSQL, a mature open-source object-relational database) running atop Apache and Linux. Mapserver is configured via map files to display and symbolize map data from shape files for quick response to user requests. PostGIS is used to answer queries by the user and to enable interaction with the database so that registered experts can update the database through a GeoWiki.

The GeoWiki allows users to draw points or polygons on a map interface and update information in the database on paleontological, engineering properties, hydrogeology, or general comments. This facilitates a wider community's participation in making this resource useful for an even larger user group. Expert's data are stored directly in the NGMDB table structure.

As a proof of concept for integrating multiple datasets stored locally or remotely, we serve a local dataset that is compiled from all of the above-named sources (with the exclusion of Oregon) from a shape file located on the server at Portland State University. We have configured the Oregon data as a Web Feature Service (WFS) that responds to requests from remote servers. This is stored as a reference in the mapfile as if it were a remote source, and displayed along with the other data simply as another layer. Requests for GeoSciML that cross boundaries of local versus "remote" services (for example, the border of Washington and Oregon) still return standardized fields and science terminology.

This system is a model for aggregating multiple datasets from many agencies. Some organizations have sufficient resources to set up a WFS server and maintain their own GeoSciML-compliant data which can be integrated into our system. Some organizations lacking these resources, however, could simply provide the data to our project for hosting on the NGMDB site. A third option is to allow organizations access to a virtual server on our system, where they can have their own subdomain (for example, id.ngmdb.us - Idaho) and manage their own data as if it existed on their own local server. Each of these three scenarios is mediated by our custom Data Import Tool, which allows the expert geologist to map their data fields for a region to a common schema and the unique values contained within to controlled science terminology of the NGMDB.

The end user's process of data discovery and use will be further enabled by the ability to overlay standards-compliant WFS and WMS (or generically OWS) services in our mapping framework; however, since all of our services will be broadcast as OWS services, any other agency or organization could create their own "mashup" with data of their own choosing in whatever client they choose, including proprietary desktop clients such as ArcExplorer, or any of a number of clients, including virtual Earths that will be proliferating in the near term.

Reference Cited

Richard, S.M., Craigue, J., and Soller, D.R., 2004, Implementing NADM C1 for the National Geologic Map Database, *in* Soller, David R., ed., Digital Mapping Techniques '04—-Workshop Proceedings: U.S. Geological Survey Open-File Report 2004-1451, p. 111-144. (Also available online at http://pubs.usgs.gov/of/2004/1451/pdf/richard.pdf.) (Accessed June 22, 2007.)

Geospatial Interoperability: From Sensors to Decision Support

By George Percivall[1]

[1]Open Geospatial Consortium (OGC), Crofton, Md.

Geospatial interoperability standards are increasing the discovery, access, and use of sensed data in research and applications. It is critically important for informed decisions that the vast data and processing resources of the geosciences community become available as part of the Web.

The Open Geospatial Consortium (OGC) has developed Web-based interoperability extending from sensors to decision support services. OGC interoperability for sensors builds on the OpenGIS® Web Map Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS). It is now practical to fit sensors of virtually any type or connection to the Web. They can be controlled through open interfaces, and their data can be output for an array of uses.

This paper reviews previously developed OGC Web Services to set the stage for description of OGC's current development. Three OGC developments are reviewed: Sensor Web Enablement (SWE), Geo-Processing Workflow, and Geo-Decision Support Services (GeoDSS). These three developments provide the specifications necessary for the acquisition, processing, and tailoring of sensor data using Web services.

The OGC is an international voluntary consensus standards organization of more than 300 companies, government agencies, and universities. OGC members participate in a consensus process to develop publicly available geoprocessing interface and encoding standards that enable integration of geospatial content and services into enterprise systems and that "geo-enable" the Web, wireless and location-based services, and mainstream information technology.

The OGC documents described in this article are available at http://www.opengeospatial.org/specs/.

OGC Web Services

OGC Web Services (OWS) are open standards for geospatial interoperability on the Internet. The OWS specifications and architecture have been developed by the members of the OGC in an interoperability test and development program and adopted in a consensus specification program. Following are some of the adopted OpenGIS implementation specifications that are most relevant to Web-based interoperability, extending from sensors to decision support services: Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Web Map Context (WMC), Catalog Service (CAT), and Geography Markup Language (GML).

Previously developed OGC Web Services, such as WMS, WCS, and WFS, are now in the marketplace. Hundreds of thousands of map layers are available from WMSs around the world. Just as the World Wide Web opened up a whole new information space, OGC Web services opens up a vastly expanded geospatial Web.

Sensor Web Enablement

Members of the OGC are building a unique and revolutionary open platform for exploiting Web-connected sensors. For example, such sensors and devices include the following: flood gauges, air pollution monitors, stress gauges on bridges, mobile heart monitors, Webcams, and satellite-borne earth-imaging devices. In much the same way that HTML and HTTP standards enabled the exchange of any type of information on the Web, the OGC Sensor Web Enablement (SWE) Initiative is focused on developing standards to enable the discovery and exchange of sensor observations, as well as the tasking of sensor systems.

SWE standards that have been built and prototyped by members of the OGC include the following pending OpenGIS specifications: Observations & Measurements (O&M), Sensor Model Language (SensorML), Transducer Model Language (TML), Sensor Observation Service (SOS), Sensor Planning Service (SPS), Sensor Alert Service (SAS), and Web Notification Service (WNS).

The sensor Web standards infrastructure defined by these specifications constitutes a revolution in the discovery, assessment, and control of live data sources and archived sensor data. The goal is to make all types of Web-resident sensors discoverable, accessible, and where applicable, controllable via the World Wide Web.

GeoProcessing Workflow

Making sensor data widely available through the Web will allow better-informed decisionmaking. In many cases the sensor data must be processed to create the information specifically relevant to a decision. OGC has defined and initially developed a set of geoprocessing services that extract context-specific information from sensor data. Typically, several geoprocessing services must be applied to sensor data. Such chains of services define workflows that can be automated for reuse in a variety of communities. OGC has applied workflow technologies to the automation of geoprocessing service chains to support decisions based on sensor data.

Recent OGC Interoperability Program Initiatives have developed workflow for decision support. The Web Processing Service (WPS) is a generic interface for multiple types of geoprocessing. Several types of WPSs were used to process data from WFS and WCS access services. This service chain was automated using the Business Process Execution Language (BPEL).

OGC continues to develop geoprocessing services and to refine the practice of chaining the services using BPEL. The chaining of geoprocessing services is one method to create geospatial information to better inform critical decisionmaking. The ultimate challenge is to enable the geographic imagery collected from different sources to become an integrated digital representation of the Earth, widely accessible for humanity's critical decisions.

Geospatial-Decision Support Services (GeoDSS)

Traditionally, decision-support systems have been monolithic applications that run on workstation class computers. Decision Support Services (DSS) extends this previous body of work into the distributed services environment. In the Geospatial Decision Support Services (GeoDSS) Initiative, open specifications are being developed that enable decisionmakers to integrate geospatial data and services from a variety of sources into the operating environment that best supports optimal decisionmaking. Elements of GeoDSS include the following: Schema Tailoring and Maintenance, GeoSemantic Web, Symbology Management, and GeoDSS Integrated Client.

The key concept of GeoDSS is that a decisionmaker is able to sit down at a single workstation, identify any resource anywhere, access that resource, bring it into their operational context, and integrate it with other resources to support the decision process. All of this takes place in a global enterprise made up of many different organizations and many different information communities.

It is critically important for informed decisions that the vast data and processing resources of the geosciences and remote-sensing community become available as part of the Web. Application communities need Internet-enabled, interactive, and interoperable access to data from observations and models for decision support systems. Bringing data archives and active remote sensors online supports the flow of geospatial information supporting policymakers, operational managers (including governments at all levels), and the citizens of our world.

Achieving Interoperability in Geosciences

By David K. Arctur[1], Phillip C. Dibner[2], and David Schell[3]

[1]OGC Interoperability Institute, Austin, Tex.

[2]Ecosystem Science Programs, OGC Interoperability Institute, Los Altos, Calif.

[3]OGC Interoperability Institute, Wayland, Mass.

Introduction

Much has been said in the geoinformatics community about the need for interoperability. There are many scientific research centers compiling significant data holdings, and many of these recognize the need to bridge between multiple scientific disciplines. For example, at a recent National Science Foundation (NSF) workshop—Building a National Geoinformatics System (National Science Foundation, 2007)—the following was stated by Chris Paola, director of the National Center for Earth-Surface Dynamics (NCED): "Important, and often costly decisions concerning land management, restoration, and subsurface resources rely on outdated science and reasoning by analogy rather than process-based analysis. The necessary science comprises elements of geomorphology, ecology, hydrology, sedimentary geology, engineering, social sciences, and geochemistry, but is not any one of these. The foundation of a useful science of Earth-surface dynamics must be synthesis across disciplines and scales, and quantitative prediction."

Similarly, D.A. Miller of the Center for Environmental Informatics at The Pennsylvania State University describes the "Critical Zone, defined by the outer limits of vegetation and the lower boundary of ground water, which reflects a complex interplay between the physical, chemical, and biological realms and has become the focus of research among a community of scientists derived from disciplines including, but not limited to, ecology, soil science, biology, geochemistry, hydrology, and geomorphology. A currently funded NSF investigation at Penn State is building tools and infrastructure to promote interdisciplinary research in a developing consortium known as the Critical Zone Exploration Network (CZEN). CZEN is envisioned as a network of sites, people, tools, and ideas…"

These are just two of a very broad and impressive set of research initiatives including EarthRef, Geosciences Network, National Ecological Observatory Network, and many others that are, within their respective communities, seeking to bridge the differences in classification systems and ontologies, semantics, spatial-temporal scales and reference systems, tools, processes, and other barriers that have inhibited interdisciplinary collaboration and integrative research. This is clearly a substantial effort for any one research center or consortium to undertake, and it is not yet complete. The dream and goal remain to carry out integrative studies across multiple scientific data centers and consortia, comprising the broadest possible range of disciplines. This is essential if we are to understand, for example, the causes and implications of climate change. Facing and coming to terms with the barriers to interoperability will also lead to more useful and robust spatial data infrastructures (SDI) among regional, national, and global

agencies, which will, in turn, greatly improve the ability of our governments and other organizations to respond to natural and human-caused disasters and emergencies. It is no exaggeration to say that the stakes are high, and that no collection of geosciences research data can be considered exempt from the need to become interoperable with other scientific data for the purpose of interdisciplinary, integrative analysis.

Steps Toward Interoperability

The particular challenge here is to provide for unfettered exchange of information among a great many scientific endeavors, while still accommodating the unfettered requirements of investigators to organize and present their data as per the standards of their own research communities. Fortunately, such advances are being made. Many consortia have adopted the cause of interoperability within their subject domains. A growing number of research centers are striving now to stimulate interdisciplinary research. Bodies of research data that can be compiled through substantially automated means, such as with satellite imagery, synthetic radars, sensor networks, and so on, are often designed around large-scale database architectures that lend them to interoperability. Various common schemas based on XML are being developed by different consortia, each with a specific application focus to meet their varied needs. For data with location and temporal content, a growing number of community-level schemas are based on ISO DIS 19136 Geography Markup Language (GML; http://www.opengeospatial.org/standards/gml), such as GeoSciML (https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/GeoSciML; also NFS 2007, p. 71-73) and ClimateML (http://ndg.nerc.ac.uk/csml/).

Common sets of interacting Web services are also emerging that enable users to find and use the data they seek. The Open Geospatial Consortium (OGC; http://www.opengeospatial.org), which developed GML, has also developed open and international specifications for online data catalog services, Web mapping services (to transfer data as simple graphic images), Web feature services (for scalar and vector data), Web coverage services (for gridded or field-type data), and services for sensor data. What is important about these specifications are that they define interfaces for data exchange; they do not require restructuring of existing databases nor changes in custodial policy.

Functional interoperability among Web services is of limited use unless the data provided by these services can be understood by the greater audience of users and accommodated by an array of client software programs. In the context of the complex, multifaceted investigations that one encounters in interdisciplinary work, it is essential to integrate semantics across multiple disciplines. Fortunately, there has been much progress in recent years in developing the means for expressing semantic content, and in its application, by a great many scientific and other information communities. Well-defined, widely adopted ontologies now exist or are emerging in many disciplines.

But this is not an easy process. Cultivation of a productive consortium and engagement with the right stakeholders is not assured. The OGC, however, has had a number of successful collaborative experiments with its Interoperability Program. Since its first Interoperability Initiative in 1999, the OGC has evolved a process that confronts precisely the challenges outlined in this paper. The process to date has enabled collaborations of diverse stakeholders, mashups of data from multiple different sources, software development by teams of skilled programmers (sometimes from competing software vendors working together), and complex analyses using data from fields as diverse as atmospheric science, toxicology, hydrology, geology, and marine science.

The OGC Interoperability Institute (OGCII) became operational in 2006 to promote the use and benefits of the continuing development of interoperable geoprocessing with the broader scientific community, as well as public sector agencies and research organizations. In order to accomplish these objectives, OGCII works closely with the OGC Interoperability Program to involve researchers in both public and private sector in testbed and pilot activities that develop the techniques and specifications that enable interoperability.

It should not be necessary to mobilize vast new funding initiatives to address this issue; considerable funds are already in place for developing integrative, interdisciplinary datasets, tools, and processes. What is needed is the recognition by both the research centers and the national agencies of the importance of semantic, as well as functional interoperability, and support for ongoing collaboration to achieve this end.

Reference Cited

National Science Foundation, 2007, Building a National Geoinformatics System: A Community Workshop, Denver, Colo., March 14-15, 2007, available online at http://www.geoinformatics.info/Download%20Center.html.

DIGGS—Data Interchange Standard for Geotechnical and Geoenvironmental Data

By Daniel J. Ponti[1], Thomas E. Lefchick[2], and Marc Hoit[3]

[1]U.S. Geological Survey, Menlo Park, Calif.

[2]Federal Highway Administration, Columbus, Ohio

[3]Civil and Coastal Engineering, University of Florida, Office of Academic Affairs, Gainesville, Fla.

DIGGS—Data Interchange for Geotechnical and Geoenvironmental Specialists (http://www.diggsml.org)—is a developing international standard interchange format for geotechnical and geoenvironmental data. It is being developed under the auspices of the U.S. Department of Transportation Federal

Highways Administration (FHWA) through a collaboration of representatives and researchers from 11 State Departments of Transportation, the United Kingdom Highway Agency, U.S. Geological Survey, U.S. Environmental Protection Agency, U.S. Army Corps of Engineers, University of Florida, the Consortium of Organizations for Strong-Motion Observation Systems (COSMOS), and the geotechnical software industry. The goals of DIGGS are as follows:

- Facilitate data exchange among different databases within an agency or organization;
- Enable oversight and regulatory agencies to receive data from consultants in a standardized format;
- Facilitate exchange of data among practitioners and researchers over the Internet;
- Facilitate data quality assurance and quality control and promote preservation of valuable subsurface data and metadata;
- Facilitate the exchange of data between software packages and providers; and
- Promote the development of data analysis software products that are more standardized and compatible.

DIGGS consists of an Extensible Markup Language (XML) schema (Geographic Markup Language-compliant) that defines surface, subsurface, and substructure features and the associated geological, geotechnical, geoenvironmental, and geophysical data that are obtained from field observations and field and laboratory tests. Version 1 of the DIGGS standard was developed by reconciling and integrating existing geotechnical and geoenvironmental data dictionaries developed by the Association of Geotechnical and Geoenvironmental Specialists in the United Kingdom (AGS; http://www.ags.org.uk/aboutus/welcome.cfm), the University of Florida, Department of Civil Engineering, and by COSMOS, which developed a pilot XML-based exchange standard for its Geotechnical Virtual Data Center (GVDC; https://geodata.cosmos-data.org/). The current schema specifically handles borehole geologic and geophysical logs and deep foundations, including an extensive suite of associated in-situ and laboratory tests. The DIGGS structure is extensible, and planned expansion of DIGGS will ultimately cover a much wider range of geotechnical and geoenvironmental tests and features.

DIGGS version 1 is currently being reviewed by a wide group of stakeholders and is slated for public release in the fall of 2007. Concurrent with the release will also be a number of software tools to facilitate data translation and data display. Specifically, public domain software to translate AGS flat files to DIGGS XML is in development by the DIGGS consortium, and Web-based data previewers that will consume DIGGS XML and produce borehole geologic, geophysical, and cone penetrometer graphic logs are in development by the COSMOS GVDC. Several commercial geoscience software developers, including gINT, EarthSoft (EQuIS), and Keynetix (HoleBASE), are also making their software compatible with DIGGS.

Visual Representation of Seismic Data

By Amit Chourasia[1] and Steven Cutchin[1]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

The amount of geoscience data available to researchers has seen an exponential growth in the past few decades and continues to grow at this rate. This growth is a direct result of advancements in monitoring, observation, and recording of data through satellite and field-specific sensors. The drastic reduction in compute and storage costs over the same timeframe has spurred the majority of this growth. This data collected through observation and simulation enable analyses at significantly expanded scope and resolution. We present a few novel visualization techniques and case studies of simulated and observed seismic data, and also show how existing tools from the general animation domain can be applied to create visualizations in the geosciences. The case studies are visualizations of the TeraShake1&2 and Puente Hills seismic simulations, the 1906 San Francisco Earthquake, and a visualization of an experimental data captured from the physical shaking of a seven-story building.

Reference

Chourasia, Amit, 2007, Digital recreation of a seven-story building during an earthquake: The Association for computing Machinery, Inc. (ACM), Crossroads, Computer Graphics, Spring 2007-13.3.

The GEON IDV (Integrated Data Viewer) for Data Integration and Exploration in the Geosciences

By Stuart Wier[1] and Charles Meertens[2]

[1]University NAVSTAR (Navigation Signal Timing and Ranging) Consortium (UNAVCO), University of Colorado, Boulder, Colo.

[2]UNAVCO Facility, UNAVCO, Inc., Boulder, Colo.

The Geosciences Network Integrated Data Viewer (GEON IDV), provided and developed in part by the University Navigation Signal Timing and Ranging (NAVSTAR) Consortium (UNAVCO), is a freely available four-dimensional software tool for data visualization and exploration in the earth sciences. New complex datasets require a tool with full three-dimensional and temporal display capabilities to discover and interpret details in the data. Most any earth-mapped data, including data with depth extent below or above the surface, can be seen in the IDV. The IDV can show data for any area on the Earth, any map projection, and any vertical scale, with time animation controls (fig. 1).

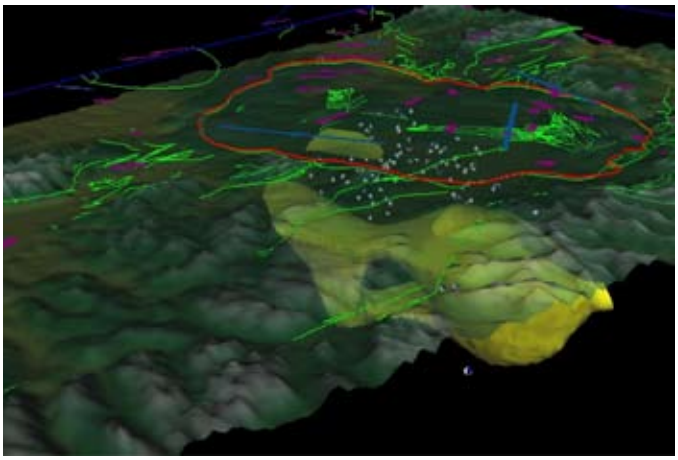

Figure 1. IDV oblique view of the geophysics of the Yellowstone National Park area, including three-dimensional surface relief, GPS velocity vectors (magenta), surface fault lines (green), caldera outline (red), epicentral focal mechanisms, and an isosurface of low P-wave velocity.

The GEON IDV (called the IDV here) was designed to enable data integration, and to foster inter-interoperability across the earth sciences. Diverse datasets from various distributed data sources can be viewed simultaneously in the same display. For example, the IDV can simultaneously display seismicity under St. Augustine volcano (Alaska) using data from real-time earthquake catalogs, and real-time 3D images of atmospheric ash clouds over the volcano, and using data from U.S. National Weather Service WSR88D Level II radar data, in true vertical scale, with a three-dimensional surface-relief image.

The IDV recognizes several data-source protocols, including local files, URLs, Web catalog servers such as THREDDS catalogs and OPeNDAP data servers, Web map server, and RSS feeds (such as the U.S. Geological Survey's (USGS's) "this week's earthquakes"). Catalogs need not be located at physical data locations, offering the possibility of distributed data sources, data services, and data users.

The IDV can be a key part of a Web-based data store and provision service. As an example, Incorporated Research Institutions for Seismology (IRIS) has recently built a Web portal to their catalog of earthquake epicenter locations and magnitudes. The user can specify earthquakes of interest by latitude, longitude, time, and magnitude range. The IRIS service generates a file with the data and provides the URL to the file. The IDV can use this service and URL as a data source. IRIS is now considering a Web service to convert seismic tomography data files, submitted online by their creators, for conversion to a format the IDV can use.

The IDV can be scripted to run unattended and make data display imagery, for example, on time as new data arrives, or on demand from a Web site. This allows a Web site to provide images of IDV displays of data specified by an online user. The Web service running the IDV need not be colocated with the data source.

The IDV relies on the power of the network Common Data Form (netCDF) file format, which stores multidimensional data, related metadata, and source information. NetCDF data files provide complete metadata, including geographic location, data units and scaling, time values and formats, variable descriptions, data provenance, publication references, data creator credits, affiliations, contact information, and information to support data discovery by geospatial, temporal, and keyword searches. All netCDF metadata is available in the IDV. The IRIS tomography data-format converter will convert ASCII files of tomography data to netCDF files.

The UNAVCO GEON Web site (http://geon.unavco.org/unavco/) provides a complete description of the GEON IDV, including how to download the software, a tutorial, help, and detailed guides to formatting data for the GEON and software tools for data conversion to netCDF.

The UNAVCO portion of this work is funded by the NSF Division of Earth Science through the GEON Information Technology Research project.

Acknowledgments

We thank the science investigators, including Mike Ritzwoller, Nikolai Shapiro, Alan McNamara, Francis Wu, Corne Kreemer, Bill Holt, Bob Smith, and Greg Waite, and the USGS for sharing their results and providing us with valuable feedback.

Demonstrating Hydrodynamic Data Assimilation with OpenGL Animations

By Zepu Zhang[1]

[1]University of Chicago, Center for Integrating Statistical and Environmental Science (CISES), Chicago, Ill.

We have developed Open Graphics Library (OpenGL) animated demonstrations for a hydrodynamic data-assimilation project. The project uses the Princeton Ocean Model to simulate wind-driven circulations in Lake Michigan. Hourly advection observations at 10 locations are assimilated into the model under mass conservation and coastal constraints. The demonstration employs a straightforward particle-tracking algorithm and is designed such that (1) it clearly shows both the magnitude and the direction of the advection; (2) it clearly compares the observed and modeled advections at the data sites; (3) it supports zooming, panning, fast forwarding and reversing, pausing, and "on-the-fly" adjustment of most parameters that control the appearance of the animated advection field; and (4) the demonstrated advection field is switched easily at any time between different scenarios (without

assimilation, with assimilation assuming exact data, and with assimilation considering measurement error).

From Caves to Optiportals: Evolution and Deployment of Visual Communication for Geoscientists

By Brian N. Davis[1], Christopher F. Polloni[2], Jason Leigh[3], and Luc Renambot[4]

[1]U.S. Geological Surevey, Sioux Falls, S. Dak.

[2]U.S. Geological Survey, Woods Hole, Mass.

[3]Electronic Visualization Laboratory, University of Illinois—Chicago, Chicago, Ill.

[4]Electronic Visualization Laboratory and the Department of Computer Science, University of Illinois—Chicago, Chicago, Ill.

Virtual Reality

The goal of virtual reality is to address the influence of human factors to allow the human brain to better process the data from virtual environments by presenting data in a visual form that the human brain has evolved to process. Such systems can enable scientists to experience models of reality "virtually," without the expense and difficulty of traveling to remote locations. For example, oceanographers can analyze and understand the geology and features of the ocean floor without traveling in a submarine, and geologists can visualize and hypothesize about the geology of Antarctica without having to experience subzero temperatures.

Consumer-Grade Visualization

Over time, virtual reality, led by developments at the Electronic Visualization Laboratory at the University of Illinois at Chicago, became more capable and realistic; however, while achieving additional capability, cost and complexity became barriers to deployment in settings usable by every-day earth scientists. Cave Automatic Virtual Environment (CAVE) technology eventually was ported to smaller, less-expensive configurations, but wide acceptance was not achieved until the GeoWall Consortium led the effort to make stereo display systems affordable and commonplace.

Software

Software is always the most difficult and expensive component of any computer system, including virtual-reality display systems. GeoWall technology did not become widely used until after commercial software routinely used by earth scientists became available. Recently, free Web browser-based software, such as Google Earth, has increased access to three-dimensional stereo visualization to an entirely new set of earth-science data consumers.

Deployment

Though the National Science Foundation-sponsored OptIPuter project (so named for its use of optical networking, Internet protocol, computer storage, and processing and visualization technologies) has developed and deployed leading-edge networking, grid computing, and visualization cyberinfrastructure technologies to again advance the capabilities of scientific visualization, use of these technologies by earth scientists will not become widespread until their deployment follows the evolution of GeoWall technology. This model determines that technology must be affordable, comprised of commodity hardware and software components, and widely available to enable scientists at different locations to collaborate using similar technology. However, the high-speed network access to visual "OptIPortals" developed by the OptIPuter project should not be discounted. As network bandwidths increase and become widely available, only real-time access to earth-science data using commonplace visual communications tools will make virtual reality practical and meaningful to earth scientists. Therefore, widespread deployment and use by geoscientists of visual communications systems must embrace the concepts developed by the OptIPuter project while evolving toward the cost model of the GeoWall Consortium.

Demonstration

OptIPuter technology, residing on the University of California at San Diego campus, will be reviewed during this talk, and then demonstrated in an informal setting at other times during the program. These technologies will include GeoWall, HiperWall, and Varrier displays used to prototype future visualization technology capabilities.

Visually Browsing Georeferenced Digital Libraries

By Angus Forbes[1] and Greg Janée[2]

[1]Map and Imagery Laboratory, Davidson Library, University of California—Santa Barbara, Santa Barbara, Calif.

[2]Institute for Computational Earth System Science, University of California—Santa Barbara, Santa Barbara, Calif.

We present a prototype visual browser for georeferenced digital libraries that allows a library to be navigated, and library content to be discovered, without explicitly formulating queries. The browser also provides a seamless transition from a synoptic view of library content to examination of individual library items.

A georeferenced digital library organizes information of all stripes, from raw data to textual documents, into collections of discrete items, each of which has a geographic region of relevance or footprint. Among the geosciences there are copious amounts of data that are georeferenced, and it has been estimated that over 70 percent of all textual documents contain relevant georeferences. The principal goal of a georefenced digital library is to take advantage of this metadata, and to take advantage of the near-universal reference frames provided by cartographic coordinate systems and geographic place names, in order to provide spatial search and navigation services and to reveal spatial context and relationships.

The dominant paradigm in information-retrieval systems is the query-result cycle: the user first formulates a query; submits the query and waits for results to appear (usually in the form of a linear list); examines the results (often limited to examining one at a time); and perhaps refines the query, thus repeating the cycle. The retrieval mechanisms offered by georeferenced digital libraries all follow this model (including the Alexandria Digital Library (ADL), Geospatial One Stop, and the Geography Network), as well as geographically enhanced Web search systems (including Google Local, Yahoo! Local, and MSN). These systems typically provide contextual base maps to support the query formulation phase, but support for visualizing query results is limited, if present at all. Also largely missing is the ability to see spatial relationships among query results and library content as a whole. ADL provides some synoptic graphics, but these are static and not integrated with the search system.

To address these limitations, we have developed a visual browser for georeferenced digital libraries that eliminates the query and result cycle. Instead, the browser displays reduced-resolution versions (in other words, thumbnails and iconic representations) of all items in the digital library over a base map. Map pan and zoom controls provide the means of navigating the library (fig. 1).

Of course, displaying all items immediately begs the question of how to display the items in a visually coherent fashion, especially when large numbers of items overlap and even coincide in geographic space. After analyzing and categorizing the types of overlaps encountered in digital library collections, we developed a suite of decluttering mechanisms:

- A custom-clustering system that groups items having sizes that are too large or too small or that are too dense to be compatible with the current zoom level. Clusters of items are represented by variable-sized icons to represent the number of library items within a region. The clusters are defined both by the geospatial range of a collection and the viewing dimensions of the browser as follows:
  - Two interface controls that dynamically control the opacity and size of item footprints.
  - A novel control widget that combines selection of zoom level with a histogram depicting the numbers of items visible at each zoom level. This device allows the user to determine at which zoom levels items are visible, which is of particular value for items that are too small and (or) dense for the current view, as well as for items that are too large for the current view (and, hence, effectively invisible due to the lack of visible footprint boundaries).

Our prototype browser is implemented using the Google Maps API, and is built on top of the ADL middleware server. To support the rapid response times needed for visual browsing, an automated preprocessing step, built on top of core ADL library services, computes and caches certain display-related information such as clusters and zoom-level histograms. A demo version of the browser operating over selected UCSB collections is available at http://clients.alexandria.ucsb.edu/ngda/.

Future development includes integration with other (nongeographic) query criteria; refinement of the aforemen-

Figure 1. Example of map available on Atlas of the Cryosphere Web site (Arctic view).

tioned clutter-reducing mechanisms; reimplementation using the Google Web Toolkit; and integration on top of the National Geospatial Digital Archive (NGDA), the funding project for this work.

References

Hill, L.L., 2006, Georeferencing: The geographic associations of information: Cambridge, Massachusetts, MIT Press, 272 p., available online at http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11007. (Accessed July 10, 2007.)

Hill, L.L.,, Janée, G., Dolin, R., Frew, J., and Larsgaard, M., 1999, Collection Metadata Solutions for Digital Library Applications, in Journal of the American Society for Information Science (JASIS), v. 50, no. 13, November 1999: p. 1169-1181, available online at http://www3.interscience.wiley.com/cgi-bin/abstract/66001477/ABSTRACT?CRETRY=1&SRETRY=0. (Accessed July 10, 2007.)

Janée G., Spatial Footprint Visualization, 2006, available online at http://www.alexandria.ucsb.edu/~gjanee/archive/2006/footprint-visualization/. (Accessed June 25, 2007.)

Janée, G., and Frew, J., 2002, The ADEPT Digital Library Architecture, in Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Portland, Oreg., July 14-18, 2002, p. 342-350, available online at http://portal.acm.org/citation.cfm?doid=544220.544306. (Accessed July 10, 2007.)

Data Fusion, Compression, and Visualization of Thermal and Visible Imagery for Remote Analysis of Geologic Surfaces on Earth And Mars

By Scott A. Nowicki[1]

[1]Environmental Studies, University of Portland, Portland, Oreg.

Global datasets of high-resolution thermal infrared (TIR) and visible to near-infrared (VNIR) satellite imagery provide opportunities for mapping planetary surface properties at resolutions that can be directly applied to field observations, with coverage that allows for analysis at regional-to-global scales. This information can be ideal for investigating geologic processes, climate variables, and surface history on the terrestrial planets, although advanced processing and analytical tools are needed to calibrate and display the data in a way that is useful for field application. To facilitate the use of these multispectral, multitemporal data in a wider scientific community, a new data-fusion technique developed using Mars Thermal Emission Imaging System (THEMIS) multispectral imagery and modified for use with the Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER) provides a data product that can be manipulated using commonly available software to map and interpret surface physical properties, such as sediment grain size, bedrock exposure, and water content of surface materials.

The term thermal morphology is used here to describe the combination of daytime visible reflectance with nighttime brightness temperature, in which the physical properties controlling the diurnal temperature and albedo can be directly interpreted. Daytime visible images produce scenes in which the reflectivity, topography, and surface roughness provide the majority of variation within a field of view (fig. 1*A*). Daytime thermal images are similar to daytime visible, in which the albedo and morphology dominate the temperature variation within a scene (fig. 1*B*). Nighttime thermal images display information related primarily to the thermal inertia of materials, in which albedo and topographic information is significantly subdued (fig. 1*C*). Thermal inertia represents the ability of near-surface materials to absorb solar energy during the day, conduct it into the subsurface, and then release that energy throughout the night. A combination of these datasets results in a striking image where colorized nighttime thermal information is draped over daytime data (fig. 1*D*).

Application to ASTER involves the compression of coregistered and calibrated three-band VNIR and five-band TIR radiance into a single, red-green-blue-color, byte image (GeoTIFF). In the thermal infrared emissivity and temperature separation, five-band nighttime observations are combined to produce a one-band real-number image. Given the normal range of diurnally varying temperatures of natural surfaces on Earth, the observed range can be linearly converted to byte data range (0-255), and retain 0.1ºC temperature resolution. Daytime observations are equally compressed in data volume to present the most thermophysically significant information in a compressed format. ASTER three-band VNIR data is integrated and converted to top-of-atmosphere calibrated reflectance, producing a single image of minimum data resolution in albedo of 0.004. Conversion of nighttime temperature to a standardized hue-based color gradient allows temperature to be draped over gray-scale visible reflectivity, resulting in a color image which displays the information from those two datasets. Retrieval of temperate and albedo values can be made visually, with the aid of a color gradient and gray scale. Digital separation of temperature and albedo can be performed by converting the RGB to Hue Saturation Intensity (HSI) format, in which hue is converted to temperature information and

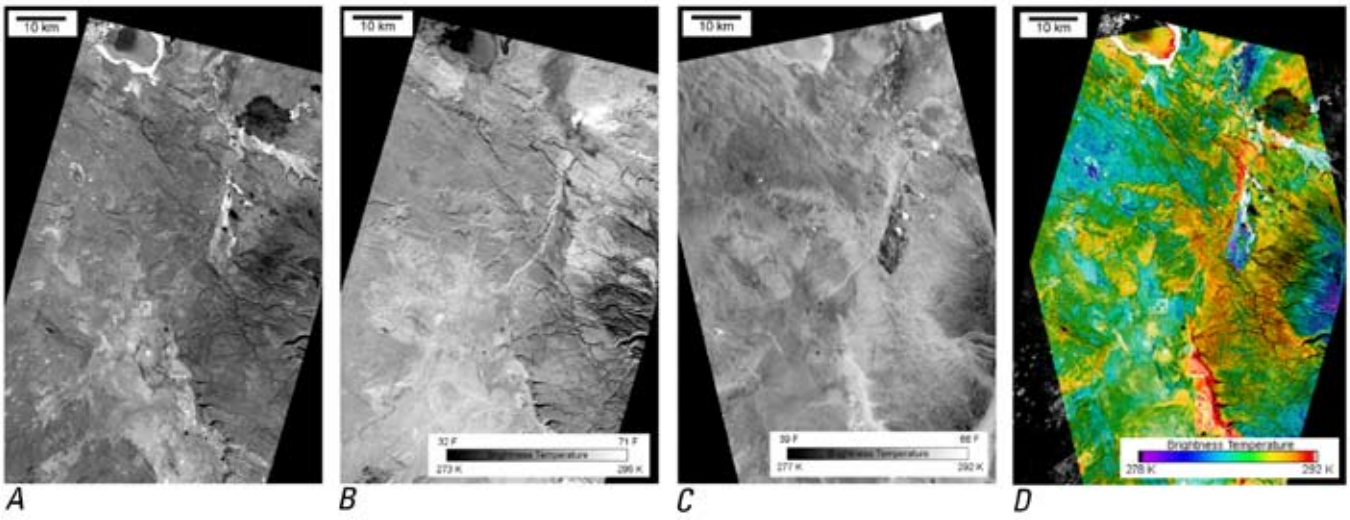

Figure 1. ASTER imagery from the Steen Mountains in southeastern Oregon, providing four perspectives used in thermophysical analysis. *A*, Visible albedo is controlled by the reflectivity, roughness, and topography of the surface. B, Daytime brightness temperature provides an image similar to VNIR reflectance, because it is primarily controlled by the albedo and insolation. *C*, Nighttime brightness temperature is primarily a function of the thermal inertia, which can be used to constrain the effective sediment grain size, material conductivity, and surface-water content. *D*, Thermal morphology is the combination of the gray-scale-visible (*A*) and colorized nighttime temperature (*C*) images.

intensity to albedo. This method allows eight bands of data to be compressed into a three-band byte image, and displayed in a format that can be interpreted in the field while retaining quantitative information for detailed analysis.

Sharing Earth Science Information Through Interoperable Approach and Cyberinfrastructure

By Myra Bambacus[1], Marge Cole[1], John Evans[1], Wenwen Li[2], Rob Raskin[3], Dick Wertz[4], Danqing Xiao[2], and Phil Yang[2]

[1]National Aeronautics and Space Administration Geoscience Interoperability Office, Goddard Space Flight Center, Greenbelt, Md.

[2]Joint Center for Intelligent Spatial Computing, College of Science, George Mason University, Fairfax, Va.

[3]Jet Propulsion Laboratory, Pasadena, Calif.

[4]The Federation of Earth Science Information Partnership, Boyce, Va.

Summary

Earth-science data and information are generated, collected, and archived at geographically dispersed locations and computers by different organizations, including government agencies, companies, and others. To leverage the legacy resources for discovering earth-science information and knowledge, we need a convenient access to the resources in an integral and timely fashion. This paper presents a joint effort in developing the Earth Information Exchange (EIE), an earth-science portal that supports this need by approaches based on interoperability and cyberinfrastructure. Earth Science Gateway (ESG), an interoperable portal, is used to provide interoperability support to access to heterogeneous resources. George Mason University (GMU) Grid, as part of the cyberinfrastructure, is utilized to support time-consuming preprocessing, modeling, and decision-support operating tools. Semantic search is utilized in bridging different domains for sharing cross-domain information and knowledge and for refining research results. The functions are integrated into the EIE developed by the partnership of National Aeronautics and Space Administration (NASA), The Federation of Earth Science Information Partnership (ESIP), and GMU to support the objective for facilitating the easily exchange of earth-science information. The ongoing effort will also provide a spatial Web portal to access and improve earth-science information holdings at different government agencies, educational and research institutions, and are non-governmental organizations.

Introduction

Earth-science data has been massively produced from satellite earth observations, insitu sensor detections, computational model simulations, and other sources for use by its extensive science and research community. These data are most useful when made easily accessible to earth-science researchers and scientists, to government agencies, and to society as a whole. To achieve an objective of apply-

ing observing and research results about Earth systems for improved knowledge gain and more effective decision support, government agencies and nongovernment organizations are working together to develop, promote, and implement interoperable architectures, and geosciences data-management and information-technology approaches to share geospatial resources among data producers, distributors, modelers, decision supporters, and decisionmakers. The objective drives the delivery and application of earth-system science research through integrated systems solutions. Successes have been achieved in various forms, as in the following examples: (a) the NASA Applied Science Program-sponsored decision-support projects using NASA data through system solutions (Birk and others, 2006), and the NASA Geosciences Interoperability Office (GIO) have developed the Earth Science Gateway (ESG; Evans and Bambacus, 2005); (b) GMU developed and deployed a grid and spatial Web portal platform to support geospatial applications (Center for Intelligent Spatial Computing, 2006); and (c) the ESIP Federation initiated the vision of an EIE (Federation of Earth Science Information Partnership, 2005).

The ESG is a standards-based, Web-services-enabled geospatial portal that allows users to discover and access geospatial data and services, and also a prototype for demonstrating and advancing geoscience interoperability. The GMU grid and spatial Web portal platform is connected with a community grid system, the Southeastern Universities Research Association Grid (SURAgrid) (Southeastern Universities Research Association, 2005) and a nationwide next generation research computer network (National LambdaRail (NLR), 2005). The ESIP Federation is a network of researchers and associated groups that collects, interprets, and develops applications for satellite-generated Earth observation information. One practical approach is to develop the EIE.

Approaches

The ESG was developed through a public consensus-driven process as a prototype Web services portal to advance the discovery, access, and use of NASA's earth-science data products. The ESIP Federation has identified requirements for this type of tool within its cluster communities and specific scientific research areas. These clusters and research areas directly contribute to the U.S. commitments within national and international initiatives, such as Global Earth Observation System of Systems (GEOSS). GIO, ESIP Federation, and GMU partner in an effort to leverage the three organizations' assets; utilizing emerging technologies on developing and advancing ESG and EIE in an interoperable way supported by the cyberinfrastructure, as illustrated in the ESG-EIE architecture (fig. 1): (1) meetings among communities are held to demonstrate ESG and EIE portal capabilities and solicit requirements to develop and improve portlets for EIE; (2) the development and evolution of application area portlets are lined up with ESIP application clusters and national applications; (3) workshops or telephone conferences are held to demonstrate ESG and EIE portal status; and (4) EIE is maintained at http://eie.cos.gmu.edu, for the time being. It will be changed to an ESIP domain name when it becomes operational.

EIE includes support to 12 national application areas and relevant technological functions of semantic search, data quality, service quality, resource integration, interface and

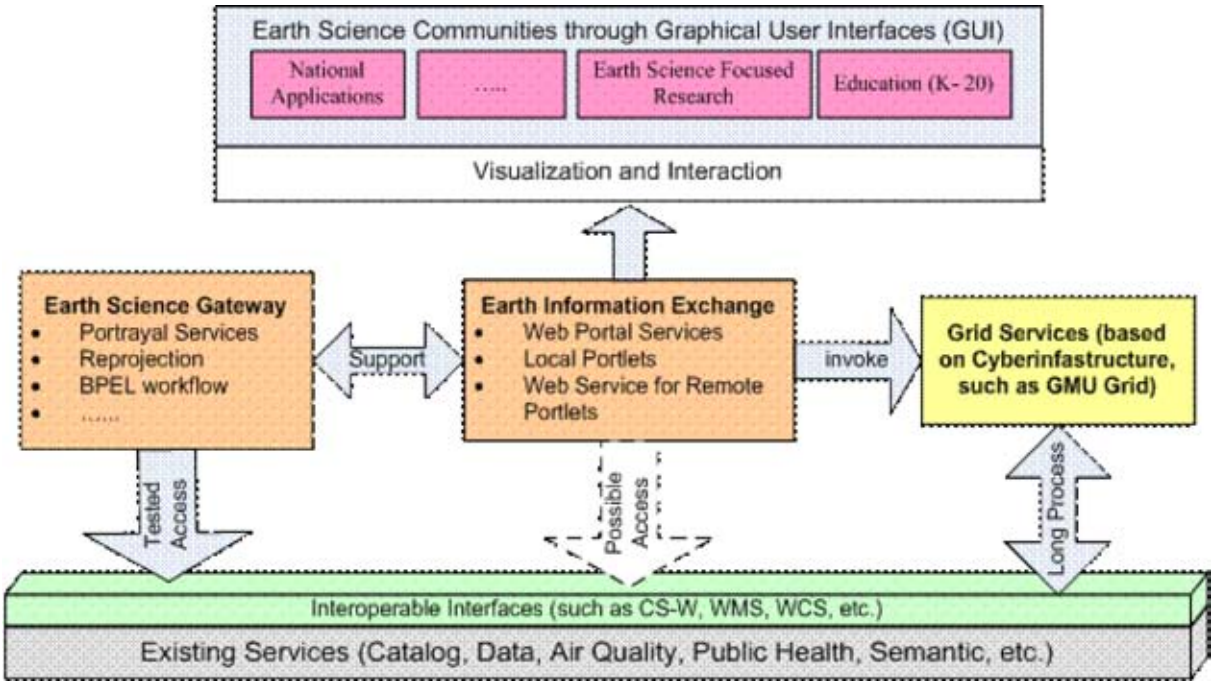

Figure 1. Interoperable (ESG) and Cyberinfrastructure (GMU Grid) supported EIE.

Relevant Organizations	EIE Application Areas	EIE functions
Environmental Protection Agency and National Aeronautics and Space Administration	Air Quality	Semantic search
National Oceanic and Atmospheric Administration	Disaster/Coastal/Water	Data quality
Environmental Protection Agency, National Aeronautics and Space Administration, and Department of Energy	Carbon/Energy/Educatopm	Service quality
Centers for Disease Control, Environmental Protection Agency, and National Aeronautics and Space Administration	Public Health/Ecological	Resource integration
U.S. Department of Agriculture and National Aeronautics and Space Administration	Agriculture/Climate	Interface and Visualization
Federal Aviation Administration, U.S. Department of Agriculture, and National Aeronautics and Space Administration	Aviation/Invasive Species	RSS filter

Table 1. Earth Information Exchange (EIE) application areas and functions.

visualization, and RSS filter, with each phase focused on one technology area and one to three application domains (table 1).

Prototypes

The EIE has been prototyped and is under construction at http://eie.cos.gmu.edu. The following figures illustrate selected functionalities supported through EIE. The EIE provides data from information, knowledge ranged from data, information, knowledge, and catalog, to services and applications.

Figure 2 depicts the EIE user interface. The left side illustrates portlet entry to the 12 national applications and education; earth-science research focus areas will be added thereafter. The middle column highlights the recent breakthroughs and phenomena within earth science. The right side provides a searching tool to find and navigate the system easily. The tabs on top provide overview, collaboration, resources, and a map client to view all these different application areas in detail.

The client also provides access to interoperable geospatial Web services to leverage heterogeneous geospatial resources through a service-oriented architecture (SOA), providing finding, binding, and chaining functions. It can be utilized to support EIE in prototyping earth-science applications in a fast fashion. The semantic search on refining searching results to different catalogs and support quality of services is under development and based on the Noesis tool developed by University of Alabama at Huntsville (Ramachandran and others, 2005).

Discussion and Conclusion

This paper presents a joint effort among GMU, ESIP Federation, and NASA GIO in leveraging the cutting-edge information technologies (such as grid computing and Web services) and geosciences interoperability in designing, developing, and operating the EIE. EIE leverages the cyberinfrastructure (Ian and Kesselman, 2004), geosciences interoperability (Evans and Bambacus, 2005), and spatial Web portal (Yang and others, in press) technologies to interoperably access heterogeneous geosciences resources, computing power, and existing knowledge to meet the needs of different levels of earth-science information users from educators to earth scientists.

Acknowledgment

The research and development reported is sponsored by NASA under grant # NNX07AD99G.

References Cited

Birk R., Frederick, M., Dewayne, L.C., and Lapenta M.W., 2006. NASA's Applied Sciences Program: Transforming Research Results into Operational Success: Earth Imaging Journal, v. 3, no. 3, p. 18-23.

CISC, 2006, The joint Center for Intelligent Spatial Computing, George Mason University, available online at http://landscan.scs.gmu.edu:8080/. (Accessed June 25, 2007.)

ESIP, 2005, Earth Information Exchange, the Federation of Earth Science Information Partnership, available online at http://www.esipfed.org/. (Accessed June 25, 2007.)

Evans, J., and Bambacus M., 2005, NASA's Earth-Sun System Gateway: An open standards-based portal to geospatial data and services, in Proceedings of IEEE IGARSS, Seoul, Korea: p. 4228-4231.

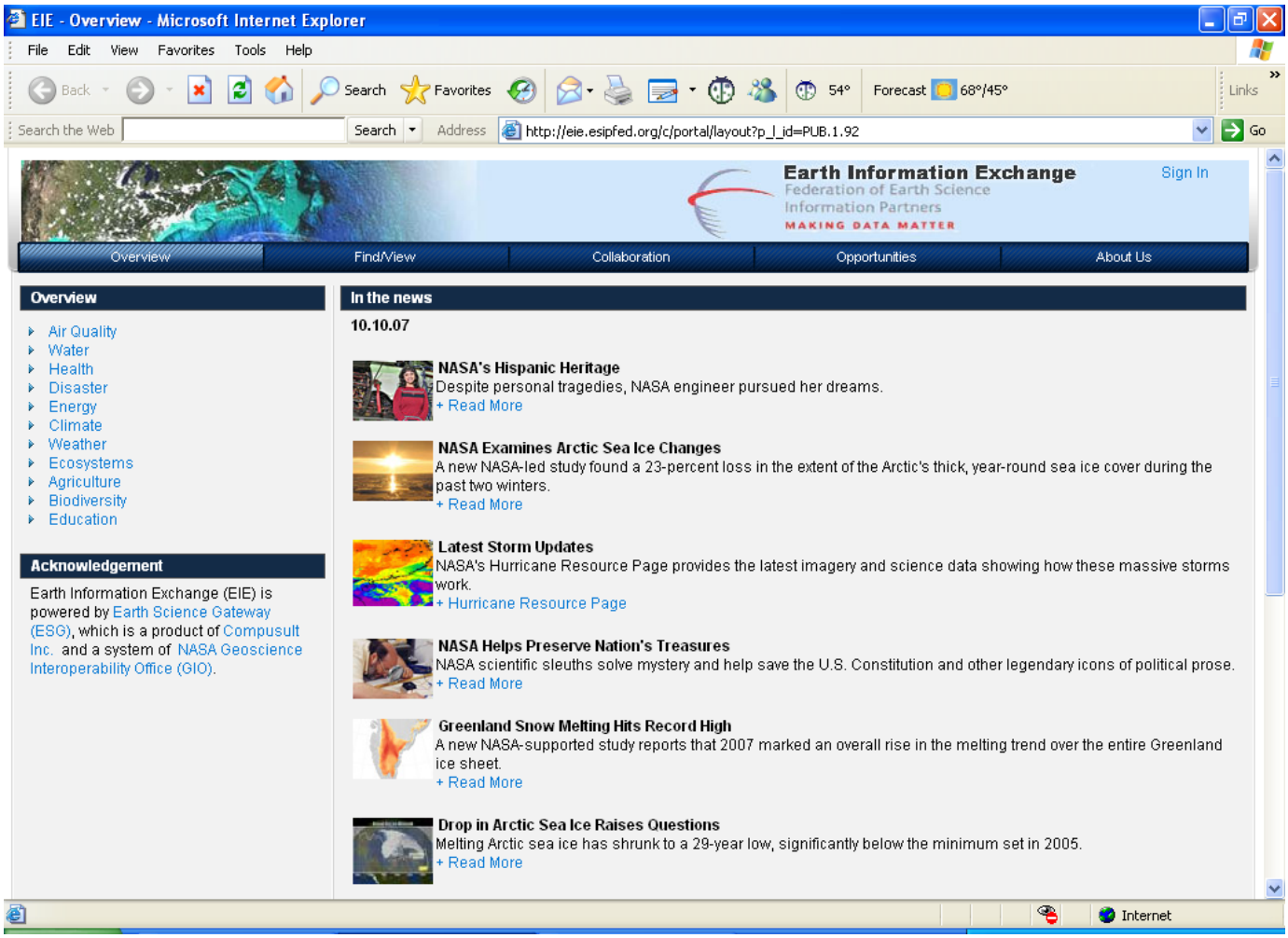

Figure 2. Earth Information Exchange interface.

Ian, F., and Kesselman, C., 2004, The Grid: Blueprint for a New Computing Infrastructure: Morgan Kaufmann Publishers,

National LambdaRail, 2005, National LamdaRail, available online at http://www.nlr.net/. (Accessed June 25, 2007.)

Ramachandran R., Movva S., Graves S., and Tanner, S., 2005, Ontology-based semantic search tool for atmospheric science, available online at http://ams.confex.com/ams/pdfpapers/102272.pdf. (Accessed June 25, 2007.)

SURA, 2005, The grid platform of the Southeastern Universities Research Association, available online at http://www.sura.org/programs/sura_grid.html. (Accessed June 25, 2007.)

Yang, P., Evans, J., Cole, M., Marley, S., Alameh, N., and Bambacus, M., in press, The emerging concepts and applications of the spatial Web portal: Photogrammetric Engineering & Remote Sensing (PE&RS), v. 73, no. 6.

A Community Workshop and Emerging Organization to Support a National Geoinformatics System in the United States

By G. Randy Keller[1], David Maidment[2], J. Douglas Walker[3], Lee Allison[4], Linda C. Gundersen[5], and Tamara Dickinson[5]

[1]School of Geology and Geophysics, University of Oklahoma, Norman, Okla.

[2]Deptartment of Civil Engineering, University of Texas—Austin, Austin, Tex.

[3]Department of Geology, University of Kansas, Lawrence, Kans.

[4]Arizona Geological Survey, Tucson, Ariz.

[5]U.S. Geological Survey, Reston, Va.

At the request of the Earth Sciences Division of the National Science Foundation a meeting was held in March of 2007 to explore what direction the geoinformatics community

in the United States should be taking in terms of developing a National Geoinformatics System. It was clear that developing such a system should involve a partnership between academia (in particular, efforts supported by the National Science Foundation), government, and industry that should be closely connected to the efforts of the U.S. Geological Survey and the State geological surveys that were discussed at a workshop in February 2007.

The March 2007 meeting had three main goals: (1) define the content of a National Geoinformatics System for the United States; (2) identify the technology by which such a system could be created; and (3) create a process for moving forward to jointly plan and develop such a system. The meeting was designed to be flexible and to emphasize breakout group sessions and plenary sessions that encouraged open discussion, brainstorming, and forward thinking.

The major conclusion of the meeting was that the geoinformatics community should proceed to investigate setting up a formal organization that is a community of informatics providers and scientists whose aim is to enable transformative science across the earth and natural sciences by doing the following:

- Fostering communication and collaboration;
- Enabling science through informatics;
- Engaging other communities (science domains and other informatics groups);
- Helping its members work to be more effective science information providers;
- Sharing resources and expertise;
- Enabling interoperability;
- Sustaining service to the community over the long haul; and
- Providing a mechanism for our community to speak with a united voice.

Association of American State Geologists (AASG)-USGS Plan for a National Geoscience Information Network

By M. Lee Allison[1] and Linda C. Gundersen[2]

[1]Arizona Geological Survey, Tucson, Ariz.

[2]U.S. Geological Survey, Reston, Va.

National Geoscience Information Network

In early 2007, the Nation's geological surveys agreed to the development of a national geoscience information network 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.

Geological surveys have unique resources and mission-specific requirements that include the gathering, archiving, and dissemination of data. Together these data represent one of the largest, most extensive long-term information resources on the geology of the United States. Currently, however, these data are available in disparate systems which require time and resources to explore, extract, and reformat. Using modern information technology and a virtual "service-oriented architecture" that provides common discovery tools and standards, the surveys and the general science community will benefit in multiple ways. First, online data and other informational products from each survey will be more readily available to the world audience and will be more valuable because they will be interoperable. Second, data and applications from external sources, such as the U.S. Geological Survey's (USGS's) more than 1,000 databases, catalogs, and inventories, will be readily accessible and integratable with each participating Survey's own data system. Third, a large federated data network will create inestimable opportunities for the broader community, including academia and the private sector, to build applications utilizing this huge data resource, and integrate it with other data. The work of each geological survey will be enhanced by access to these new data and applications.

By demonstrating national cooperation for data access and interoperability among the Federal and State geological surveys, we may be able to serve as a model for broader cooperation in geoinformatics across the entire earth-science community and linkages to other scientific disciplines, especially those with a geospatial aspect. We intend to coordinate the development of this network with other efforts, including the National Science Foundation's "Cyberinfrastructure Vision for 21st Century Discovery," and the emerging academic and international efforts in informatics. This "community of practice" approach means that we will learn, develop, evolve, and coordinate the building of the network with each other and our partners.

When initiated among the geological surveys, any user may go to a geological survey (or other participating) Web site, enter a distributed science data catalog, and view available data. Because all these data will use a common markup language, the user can select and download needed data and load them into any number of their own applications, including in-house, freeware, and proprietary commercial products. The interface would be seamless and near-instantaneous and the original data source would be credited with the download. The information network will expand as others participate.

Roles of Geological Surveys

On one hand, geological surveys have dual requirements to collect, archive, and disseminate data for their stakeholders and customers to use, and on the other, to access data held by others that will enable the surveys to better carry out their

analytical and research duties. Data accessibility enhances this two-way exchange of information.

The geological surveys are primary geoscience data providers and have mandated responsibilities to collect, organize, and distribute this information to the public. Currently, information assets exist in many databases and in many forms. Similarly, organizations have implemented a wide variety of solutions to manage, process, and support research and data stewardship requirements. Some organizations have integrated their data to provide products to the public, and others have developed accessible Internet map services. Because of the large investment in these distributed systems, the emerging service architecture must build on existing systems and use protocols, standards, and services to help integrate the information systems and scientific information.

Geological surveys have unique resources and functions as institutions with statutory mandates to collect, archive, and disseminate data permanently. These missions can complement and facilitate development of a national geoscience information network as well as benefit greatly from the result. Geological surveys also contribute to the building of standards of practice and fundamental baseline geologic information, such as lexicons, geologic maps, and time scales. These contribute directly to the overall geoinformatics efforts. The breadth and depth of survey-based data are so large that collectively they constitute one of the largest, if not the largest, data resources in the geosciences—in essence, a national data "backbone."

Vision for a National Geoscience Information Network

The surveys agreed to the following principles and activities to be undertaken in the next few years to achieve the vision:

- Develop a coordinated national geoscience information network to access and integrate State survey and USGS information resources (databases, maps, publications, methods, applications, and data services).
- Function as a "community of practice" in development of geoinformatics and the geoscience network.
- Develop prototypes (pilots and testbeds) to show proof of concept and to determine realistic levels of effort, and 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 framework should be distributed and leverage existing systems, map services, and data, with local autonomy but using standards to enable interoperability.
- Review and adopt standards and protocols for developing the network (including metadata).
- New and existing systems should communicate with open-source (for example, Open Geospatial Consortium-based) protocols to promote interoperability.
- Test and consider accepting GeoSciML (Geoscience Markup Language) as a protocol and consider proposing as a standard to Federal Geographic Data Committee (FGDC).
- Recognize there are priority data for which we have mission requirements and inherent partnerships amongst the geological surveys. Review these and adopt service definitions and protocols as appropriate:
  - Geologic maps, hazard data and maps, topographic data, existing map services;
  - Publications and bibliographies.
  - Observations and analytical measurements, samples, and site information.
  - Applications and methods, and analytical tools.
  - Legacy analog data.
  - Resource data and maps (minerals, energy, water, and so on).
- Encourage clients and services to be developed and facilitate participation and implementation by others, preferably with low overhead, while improving business models 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, using "branding" in data services to acknowledge data sources;
- Develop usage measurements and utilize them in clients and Web services;
- Develop a database citation format; and
- Policy—Acknowledge that geological surveys need to recognize interoperable, Web-enabled information resources as part of their mission. Seek partnerships to leverage resources, and to develop and implement the vision.

References

Kumar, M., 2006, Geoinformatics 2006: Eos, v. 87, no. 44, p. 481.

Allison, L., Dickenson, T., and Gundersen, L., 2007, Role of State geological surveys and USGS in a geophysical system for the Nation: A report to the AASG and USGS on the results of a workshop conducted February 21-22, 2007, available online at http://www.geoinformatics.info/Download%20Center.html. (Accessed June 25, 2007.)

The Cultural and Social Challenges of Developing Geoinformatics: Insights from Social, Domain, and Information Sciences

By Kerstin A. Lehnert[1], Paul N. Edwards[2], Steven Jackson[2], and Geoffrey Bowkers[3]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, N.Y.

[2]School of Information, University of Michigan, Ann Arbor, Mich.

[3]Center for Science, Technology and Society, Santa Clara University, Santa Clara, Calif.

Geoinformatics represents an effort to link a vast, geographically distributed, diffuse, and relatively uncoordinated set of existing projects into a genuine infrastructure that is envisioned to provide highly reliable, widely accessible capabilities and services in support of scientific work, with the aim to facilitate distributed collaboration, democratize the research environment, and empower cross-boundary and interdisciplinary scholarship. While the vision of cyberinfrastructure is becoming increasingly comprehensive and well defined, the path to achieve this vision is still vague and its controls poorly understood.

Insights gained by historians and social scientists from the analyses of other kinds of infrastructure, such as railroads, telephony, and the Internet, can help guide Geoinformatics practitioners to more effectively advance the growth of geoinformatics and cyberinfrastructure in general. The recently released workshop report "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" (Edwards and others, 2007) explains the dynamics, tensions, design challenges, and navigation strategies that emerge as common patterns and practices of infrastructural development. The report emphasizes the relevance of social and organizational factors in infrastructure development, and concludes that "robust cyberinfrastructure will develop only when social, organizational, and cultural issues are resolved in tandem with the creation of technology-based services." The workshop recommends strategic collaborations between social, domain, and information scientists to assist the design of effective navigation strategies that will help realize the vision of cyberinfrastructure. A primary target of these collaborations should be the study of existing cyberinfrastructure projects to reveal key factors in success and failure.

We present here initial insights from analyzing the development of successful geoinformatics systems for geochemistry. Geochemistry is a discipline characterized by a culture of independent research in the form of small- to medium-scale projects, in which data are acquired by human "observers" rather than by sensors, often through idiosyncratic data-collection practices and in idiosyncratic formats. Due to the large personal effort involved in generating the data, geochemical data are considered private intellectual property, and are shared only through publications in the scientific literature that guarantee the appropriate credit for data authors. This practice has led to a wide dispersion of data in the literature, making it difficult for the broad geoscience community to access and efficiently use the full range of available data. Data publications are frequently missing contextual information describing the complex processes of data gathering that is needed in order for other data users to interpret the data.

In the mid-1990s, domain scientists in the U.S. and in Europe independently recognized the need for more efficient access to data to support new research endeavors (for example, the National Science Foundation-funded RIDGE program or the Geochemical Earth Reference Model Initiative) and the potential of emerging technologies, such as relational databases and the Web, and started to develop geochemical databases (PetDB, GEOROC, NAVDAT, and EarthRef) that were publicly accessible on the Internet. These database projects were of limited scope, motivated by the scientists' personal research agendas (PetDB focused on mid-ocean ridge basalts and abyssal peridotites; GEOROC initially focused on ocean-island basalts). They were rapidly embraced by relevant parts of the community because they provided substantial benefits to researchers and educators who no longer had to expend significant efforts to produce their own data compilations, excruciatingly typing data from the literature into spreadsheets. The new online databases matched the scientific working environment and workflow, allowing researchers to easily access the data, and provided tools to integrate data from hundreds of publications into customized datasets within minutes. Since the databases went online, several hundred scientific articles have cited these databases as the source for datasets used to create or test new hypotheses, providing evidence for their utility and success.

The development of the various databases can be assigned to the "System Building Phase" in infrastructure development, which is characterized by the successful design of technology-based services. Even though systems were not yet implemented in a sustainable manner, this phase was critical to provide a proof of concept to the community that broadened support for an advanced data infrastructure in geochemistry, and increased awareness within the community about deficiencies in the data culture, such as the inconsistent and incomplete reporting of data-quality information in publications.

The database projects next moved into a phase of "System Growth and Stabilization," during which the systems were migrated to more professional and sustainable IT environments, with expert teams that supported development and operations. The community increasingly accepted the systems as part of their research infrastructure. Finally, the projects entered the "Networking and Consolidation Phase" when they founded the EarthChem consortium, with the objective to better link and integrate the independent data collections, nurturing synergies among projects, minimizing duplication of efforts, and sharing tools and approaches. Tensions regarding ownership, control, and design approaches that arose dur-

ing the Networking Phase were surpassed by the substantial benefits of the collaboration, such as the broader impact of technical or organizational developments, including standards and policies, and ultimately led to a more stable and well-considered implementation. The EarthChem "network" is now expanding with new partners, and has attained a leadership role in the field, advancing a culture change in the geochemistry community and working with other geoinformatics projects, societies, editors, and the science community at large toward standards for data sharing and data reporting.

According to our analyses, the following factors have been key for the success of the geochemical databases: (a) The initial proof of concept systems did not rely on contributions from the community; only investigators experienced the benefits of the systems and thus were more readily accepted and supported the systems. (b) The systems offered capabilities that did not exist before, and that an individual could not achieve. (c) Early collaboration among the database systems led to compatible data models and metadata schemes, increasing the impact of the more widely applicable system designs. (d) The teams provided the necessary organizational and "marketing" components to successfully advance the projects (for example, the "wizard-maestro-champion" combination noted by historians as a common combination of system-building teams).

Based on the lesson learned from the geochemistry systems, we see three roles that projects need to fulfill in order to successfully advance and implement geoinformatics: (a) Service provider—The needs and requests of the users need to be given the highest priority. It is critical to understand that systems are operations, rather than pure research and development efforts. (b) Competent partner—Both data authors and data users need to trust the service provider, who needs to understand and respond to domain science concerns. (c) Team player—In order to advance the infrastructure development via networking, project members have to be willing to collaborate, share expertise and experiences, and acknowledge others' achievements.

Community concerns regarding issues of intellectual property (credit to data authors), and the impact of geoinformatics on core science funding, as well as the broad implementation of new practices and procedures (for example, unfunded mandates of metadata generation) still represent major challenges that need to be overcome.

Reference Cited

Edwards, P.N., Jackson, S.J., Bowker, G.C., and Knobel, C.P., 2007, Understanding infrastructure: Dynamics, tensions, and design *in* Report of a workshop on history and theory of infrastructure: Lessons for new scientific cyberinfrastructures, January 2007, available online at http://www.si.umich.edu/InfrastructureWorkshop/. (Accessed June 25, 2007.)

Hydroseek—A Search Engine for Hydrologists

By Bora Beran[1] and Michael Piasecki[1]

[1]Civil, Architectural and Environmental Engineering, Drexel University, Philadelphia, Pa.

Search engines have changed the way we see the Internet. The ability to find the information by just typing in keywords was a big contribution to the overall Web experience. While the conventional search engine methodology worked well for textual documents, locating scientific data remains a problem since they are stored in databases not readily accessible by search engine bots.

Considering different temporal, spatial, and thematic coverage of different databases, especially for interdisciplinary research, it is typically necessary to work with multiple data sources. These sources can be Federal agencies which generally offer national coverage or regional sources which cover a smaller area with higher detail; however, for a given geographic area of interest there often exists more than one database with relevant data. Being able to query multiple databases simultaneously is a desirable feature that would be tremendously useful for scientists. Development of such a search engine requires dealing with various heterogeneity issues. In scientific databases, systems often impose controlled vocabularies that ensure homogeneity within themselves, thus heterogeneity becomes a problem when more than one database are involved. Having controlled vocabularies at individual database level defines the boundaries of vocabulary variety, making it easier to solve the semantic heterogeneity problem than with the conventional search engines that deal with free text.

Structural, syntactic, and information system heterogeneities emerge as types of additional incompatibilities that these systems have to resolve. Structural heterogeneity is generally defined as different information systems storing their data in different document layouts and formats. In the current state of hydrologic data providers, we can speak of HTML tables, XML documents, or text files, where the file format alone does not guarantee homogeneity since data output can be organized in many different ways. Syntactic heterogeneity is the presence of different representations or encodings of data. Date and time formats can be given as an example, where common differences are local time versus coordination universal time, 12-hour clock versus 24-hour clock, and standard date format versus Julian day, which is common in Ameriflux data. Whereas information system heterogeneity requires methods of communication specifically tailored to interact with each data providers' servers due to the difference in interfaces (for example, Representational State Transfer (REST) services versus Simple Object Access Protocol (SOAP) services), it also encompasses the difficulties from the difference of arguments that each service requires. Sometimes even responses and requests have different formats. In the U.S. Environ-

mental Protection Agency's STOrage and RETrieval System (STORET), data requests (through available REST services) require dates to be provided in Dublin Julian days (days since the noon between December 31, 1899, and January 1, 1990) while the server returns Gregorian dates with the data.

We have developed a search engine (http://cbe.cae.drexel.edu/search/) that enables querying multiple data sources simultaneously and returns data in a standardized output, despite the aforementioned heterogeneity issues between the underlying systems. This application relies mainly on metadata catalogs or indexing databases, ontologies, and Web services with virtual globe and Asynchronous JavaScript and Extensible Markup Language (XML) (AJAX) technologies for the graphical user interface. Users can trigger a search of dozens of different parameters over hundreds of thousands of stations from multiple agencies by providing a keyword, a spatial extent (in other words, a bounding box), and a temporal bracket.

CUAHSI Cyberinfrastructure for Hydrologic Sciences

By Ilya Zaslavsky[1], David Valentine[1], and David R. Maidment[2]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[2]Center for Research in Water Resources, University of Texas—Austin, Austin, Tex.

The mission of the CUAHSI HIS (Consortium of Universities for the Advancement of Hydrologic Science, Inc., Hydrologic Information System) project is development of cyberinfrastructure supporting advanced hydrologic research and education. It is a collaborative project that involves several research universities and the San Diego Supercomputer Center as the technology partner. Over the last three years, the CUAHSI HIS team has been researching, prototyping, and implementing Web services for discovering and accessing a variety of hydrologic data sources, and for developing applications for the desktop and for the Web.

The CUAHSI HIS system architecture is envisioned as a component of a large-scale environmental observatory effort, which emerges as a network of seamlessly integrated data collection, information management, analysis, modeling, and engineering endeavors implemented across disciplinary boundaries. The HIS design follows the open services-oriented architecture model; that is, it relies on a collection of loosely coupled, self-contained services that communicate with each other and that can be called from multiple clients in a standard fashion. The core of the system is WaterOneFlow Simple Object Access Protocol (SOAP) services, providing uniform access to multiple heterogeneous repositories of hydrologic observation data, both remote and local. The services follow a common Extensible Markup Language (XML) messaging schema named CUAHSI WaterML, which includes constructs for transmitting observation values and time series, as well as observation metadata including information about sites, variables, and networks. The information model of WaterML follows that of Observation Data Model (ODM, http://www.cuahsi.org/his/odm.html), while extending it to handling of observation data available as fields. The currently available services provide access to the U.S. Geological Survey's (USGS's) National Water Information System (NWIS), the U.S. Environmental Protection Agency's (EPA's) STORET, National Climatic Data Center Automated Surface Observing System (NCDC ASOS), Daymet, Moderate Resolution Imaging Spectroradiometer (MODIS), and North American Mesoscale 12 Kilometer (NAM12K) data, as well as to data maintained by users in the ODM format.

The Web services are accessed from different types of clients, including a Web browser, a range of desktop applications such as Matlab, ArcGIS, and Excel, and several programming languages (NET and Java), which were exposed as the primary desktop client environments by the CUAHSI user-needs assessment. The general organization of the HIS System is shown in figure 1. The Web browser-based client is developed in collaboration with Enviromental Systems Research Institute and relies on ArcGIS Server 9.2 for online mapping functionality.

At the physical level, the system being deployed now includes a central HIS server and a networked collection of workgroup HIS servers. The central node contains observation data catalogs for nationwide hydrologic observations repositories maintained at USGS, EPA, NCDC, and other agencies. The catalogs are accessible via the Web services supporting GetSiteInfo and GetVariableInfo requests, while the data series are returned via GetValues SOAP Application Program Interface (API) calls. The organization of Workgroup HIS nodes is similar, with additional ability to import local observation data into ODM instances, configure the ODM and respective

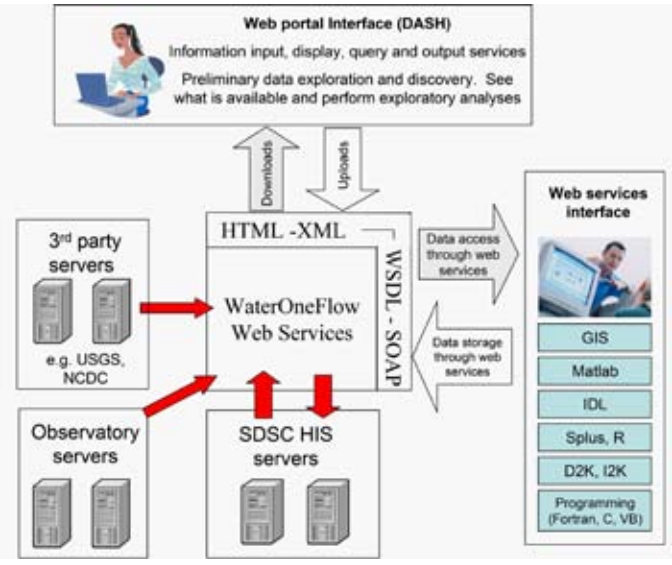

Figure 1. The general organization of the HIS system.

Web services, and serve the newly registered data along with the national hydrologic observations. The software stack for both the Central and the Workgroup HIS servers is based on Windows 2003 server and includes Structured Query Language (SQL) Server 2005, ArcGIS Server 9.2, Visual Studio 2005, as well as a series of tools and databases developed by the HIS team.

In the first stage of deployment, the workgroup HIS servers are designed to be implemented at 11 hydrologic observatory testbeds as part of a National Science Foundation Water and Environmental Research Systems (WATERS) initiative. This will let us test the HIS tools on diverse hydrologic datasets collected by the testbeds and further help develop the Cyberinfrastructure for Hydrologic Sciences. More information about the project is available at http://www.cuahsi.org/his.

Design and Implementation of CUAHSI WaterML and WaterOneFlow Web Services

By David Valentine[1], Ilya Zaslavsky[1], Thomas Whitenack[1], and David R. Maidment[2]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[2]Center for Research in Water Resources, University of Texas—Austin, Austin, Tex.

WaterOneFlow is a term for a group of Web services created by and for the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) community. CUAHSI is an organization representing more than 100 U.S. universities and is supported by the National Science Foundation to develop infrastructure and services for the advancement of hydrologic science. CUAHSI Web services facilitate the retrieval of hydrologic observations information from online data sources using the Simple Object Access Protocol (SOAP) protocol. CUAHSI WaterML (below referred to as WaterML) is an Extensible Markup Language (XML) schema defining the format of messages returned by the WaterOneFlow Web services.

Background

Beginning in 2005, the CUAHSI HIS project team implemented a variety of Web services providing access to large repositories of hydrologic observation data, including the U.S. Geological Survey's (USGS's) National Water Information System (NWIS) and the U.S. Environmental Protection Agency's (EPA's) STORET (Storage and Retrieval) database of water-quality information. The services gave access to station and variable metadata, and observations data stored at these sites. As the services were written for each data source individually, their inputs and outputs were different across data sources. The services developed in this ad hoc manner do not scale well. As the number and heterogeneity of data streams to be integrated in CUAHSI's hydrologic data access system increased, it would become more and more difficult to develop and maintain a growing set of client applications programmed against the different signatures and keep track of data and metadata semantics of different sources. As a result, WaterML was developed to provide a systematic way to access water information from point observation sites.

Point Observations Information Model

In parallel with Web service development, CUAHSI has been developing an information model for hydrologic observations that is called the Observation Data Model (ODM). Its purpose is to represent observation data in a generic structure that accommodates different source schemas. While based on the preliminary set of CUAHSI Web services, WaterML was further refined through standardization of terminology between WaterML and ODM, and through analysis of access syntax used by different observation data repositories, including USGS NWIS, EPA STORET, National Climatic Data Center Automated Surface Observing System (NCDC ASOS), Daymet, Moderate Resolution Imaging Spectroradiometer (MODIS), and the North American Mesoscale 12 Kilometer (NAM12K) System.

According to the information model, a data source operates one or more observation networks; a network is a set of observation sites; a site is a point location where water measurements are made; a variable describes one of the types of measurements; and a time series of values contains the measured data, wherein each value is characterized by its time of measurement and possibly by a qualifier that supplies additional information about the observation. Figure 1 demonstrates the main components of the model and respective Web services.

WaterML Concepts

The goal of the first version of WaterML was to encode the semantics of hydrologic observations discovery and retrieval and implement WaterOneFlow services in a way that creates the least number of barriers for adoption by the hydrologic research community. In particular, this implied maintaining a single common representation for the key constructs returned on Web service calls related to observations, features of interest, observation procedures, observation series, etc.

An observation is considered an act of assigning a number, term, or other symbol to a phenomenon, and a result of such assignment. Hydrologic observations are performed against many different phenomena (properties of different features of interest), and shall be associated with time measurements (time points or time intervals). The features of interest common in hydrologic observations may include points (gauging stations and test sites), linear features (steams and river

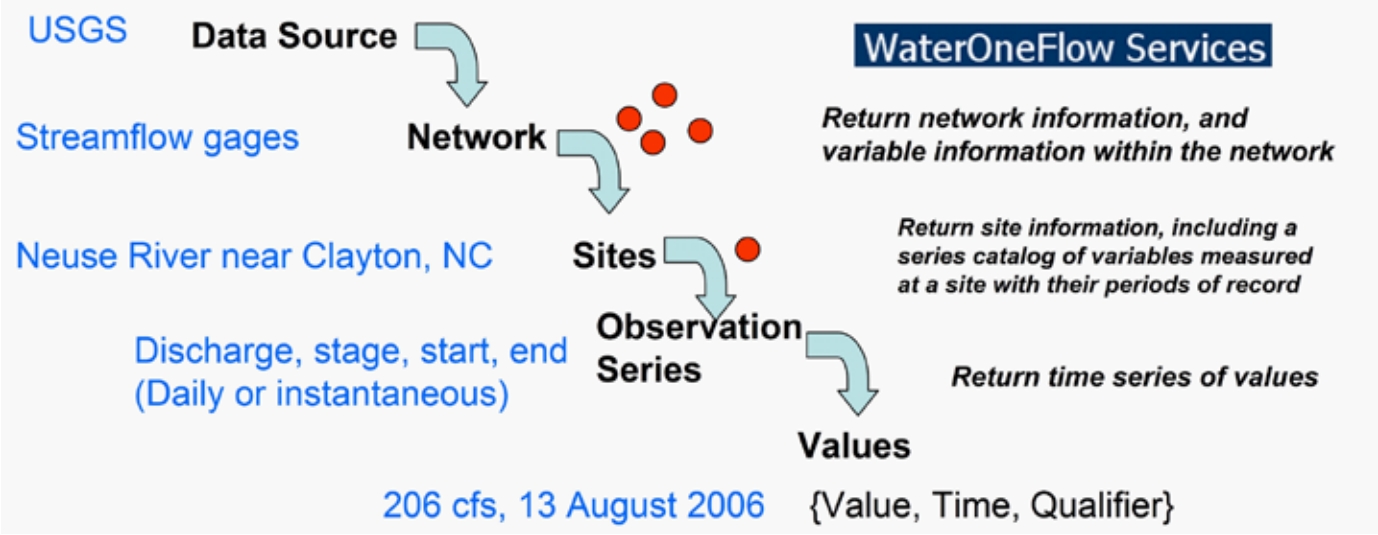

Figure 1. CUAHSI Point Observations Information Model, and corresponding Web service methods.

channels), or polygon features (catchments and watersheds). Spatial properties of the features of interest may be further expressed in two or three dimensions, in particular via vertical offsets against common reference features. The observations are made in a particular medium (water, air, and sediments) using a procedure. The procedure may represent a multistep processing chain, including an instrument (sensor), algorithms for transforming the initially measured property (for example, "partial pressure of oxygen in the water" may be transformed into a measure of "dissolved oxygen concentration"), and various techniques for censoring and for quality control of the value assignment, including multiple scenarios for assignment of no value. Individual observations are organized into observation series (regular sequences of observations of a specific variable made at a specific site), which are in turn referenced in series catalogs. A series catalog is an element of an observation network, which represents a collection of sites where a particular set of variables is measured. A responsible organization can maintain one or more observation networks.

In addition to point measurements described in the ODM specification, hydrologic information may be available as observations or model outcomes aggregated over user-defined regions or grid cells. While USGS NWIS and EPA STORET exemplify the former case, sources such as MODIS and Daymet are examples of the latter. In this latter case, as in the case of other remote-sensing products or model-generated grids, the observation or model-generated data are treated as fields, and sources of such data are referenced in WaterML as datasets, as opposed to sites.

The practice of hydrologic observations provides ample evidence of complications beyond this general treatment. These complications are related to the complex and often incompatible vocabularies used by several Federal hydrologic observation systems, to different and not always documented contexts of measurement and value assignment, to often ambiguously defined features of interest, and to complex organizational contexts of hydrologic measurement, transformation, aggregation, etc. It is in response to this complexity that the CUAHSI WaterML is primarily designed. While some of this complexity may be captured within the standards being developed under the OGC's Sensor Web Enablement (SWE) activity, the flexibility inherent in such standards may itself be a barrier to adoption when the target audience is not computer scientists.

Implementation Context

WaterML is primarily designed for relaying fundamental hydrologic time-series data and metadata between clients and servers, and to be generic across different data providers. Different implementations of WaterOneFlow services may add supplemental information to the content of messages; however, regardless of whether or not a given WaterML document includes supplemental information, the client shall be sure that the portion of WaterML pertaining to space, time, and variables will be consistent across any data source.

Depending on the type of information that the client requested, a WaterOneFlow Web service will assemble the appropriate XML elements into a WaterML response and deliver that to the client. The core WaterOneFlow methods include the following:

- GetSiteInfo—For requesting information about an observations site; the returned document has a root element of SiteInfoResponse type.
- GetVariableInfo—For requesting information about a variable; the returned document has a root element of VariableResponse type.
- GetValues—For requesting a time series for a variable at a given site or spatial fragment of a dataset; the returned document has a root element of TimeSeriesResponse type.

The provisional services are available from http://river.sdsc.edu. The CUAHSI WaterML description has been submitted as a discussion paper to the Open Geospatial Consortium (OGC), and is available from the OGC portal. The project Web site is http://www.cuahsi.org/his.

Hydrologic Information System Server: The Software Stack and the Initial Deployment Experience

By Thomas Whitenack[1], David Valentine[1], Ilya Zaslavsky[1], and Dean Djokic[2]

[1]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[2]Environmental Systems Research Institute, Inc., Redlands, Calif.

Description of the Software Stack

One of the main outcomes of the CUAHSI Hydrologic Information System project is the development of a Windows-based software stack to support importing, registering, and serving hydrologic observation data. The Hydrologic Information System Server (HIS Server) organizes observation databases, geographic data layers, data importing and management tools, and online user interfaces into a flexible multitier application for serving both national-level and locally maintained observation data. The main components of the distributable software stack (fig. 1), in a typical deployment scenario, are a mix of commercial off-the-shelf technologies (COTS), and a custom-developed code for Web services, databases, and sophisticated online data access. The COTS stack includes a Windows 2003 Server operation system with Internet Information Server (IIS), a Structured Query Language (SQL) Server 2005 for storing observation databases, an ArcGIS 9.2 and ArcGIS Server for managing and serving observation site maps and other spatial data, and Visual Studio 2005 for impersonating and recompiling the map application when required. The following are components developed by the HIS Team:

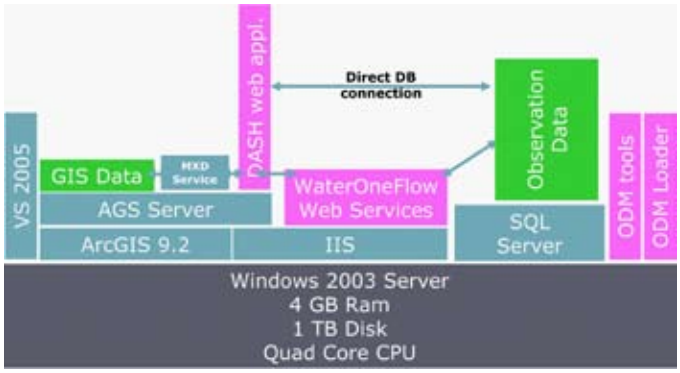

Figure 1. Components of the HIS software stack.

- WaterOneFlow Web services—Services providing uniform query access to remote repositories of observation data and data catalogs (U.S. Geological Survey's National Water Information System (NWIS), U.S. Environmental Protection Agency's Storage and Retrieval System (STORET), National Climatic Data Center's Automated Surface Observing System (ASOS), and so on) and to local Observation Data Model (ODM)-compliant databases. In addition, the software stack includes an online application for testing local Web services.
- Observation data loader—Software for importing observation data and catalogs into ODM instances.
- ODM tools—Software for exploring and editing data imported into ODM.
- Data Access System for Hydrology (DASH)—An ASP.NET application developed over ArcGIS Server 9.2 that represents a mapping, querying, and data-retrieval interface over observation and GIS databases, and Web services.

In addition, the deployed HIS server contains tutorials and instructions for several workflows supported by the system, specifically related to importing, curating, and serving observation data on the Internet.

Importing and Registering Observation Datasets

The critical capability that the HIS software stack supports is importing and registering user-collected hydrologic observation data. The steps in adding a new observation network are as follows (see fig. 2.):

1. Using the ODM DataLoader, the user (information manager in a Water and Environmental Research Systems (WATERS) testbed) loads a hydrologic observation dataset from an Excel or text file into a blank ODM instance in a local SQL Server 2005. As a result, a set of ODM-compliant tables will be created and populated with observation data and metadata, including the Sites, Variables, and SeriesCatalog tables upon which the DASH application relies. In addition, the user may explore and edit the newly created ODM instance using the ODM Tools desktop application.
2. The user creates a copy of the WaterOneFlow Web Services template and edits its "web.config" file to point to the newly created ODM instance (specifying the name of the instance and the connection string). Once the new ODM Web service is deployed, the user can test it with a Web service testing application to ensure that the new service works as expected.
3. The user creates a point layer of observation stations, as either a feature class in a geodatabase, or as a shape file, importing point coordinates from the new ODM's Sites table directly, or via GetSites Web service call.

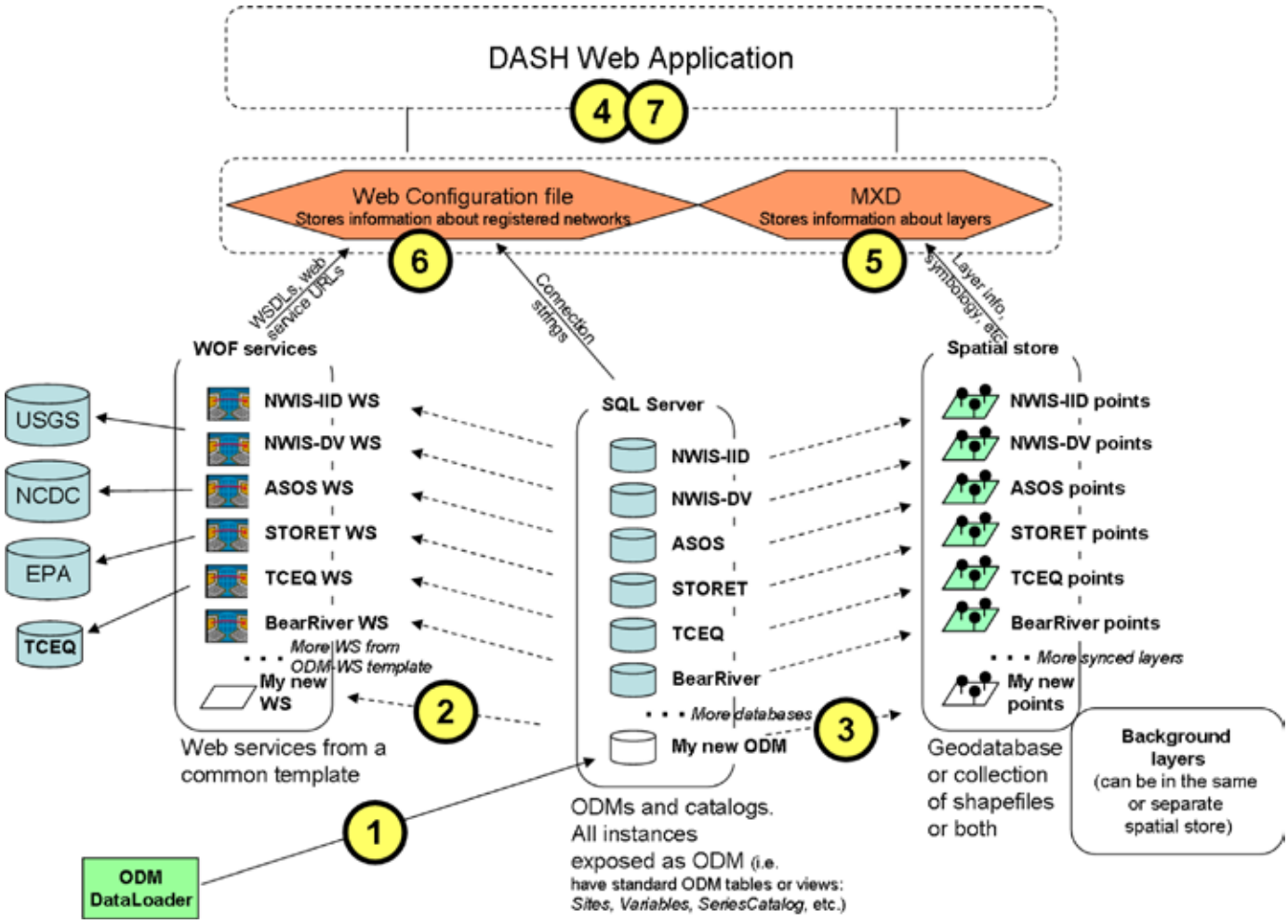

Figure 2. The HIS Server architecture and steps for adding observation networks.

4. To configure the DASH application with the added observation network, the DASH must be stopped.
5. Using ArcGIS 9.2, the user adds the new point layer to the map configuration file (an .MXD document) used by DASH. At this step, the layer's symbology, scale-dependent rendering, labeling, and other components of the map display can by modified.
6. The user edits the Web configuration file of the DASH, adding information about the newly created observation network database. For each new network, this information includes: NetworkID, NetworkCode, NetworkName, NetworkConnectionString, NetworkServicesWSDL, NetworkServicesURL, NetworkServicesGetValuesURL, CreatedDate, LastUpdateDate, ResponsibleParty, as well as Disabled and DisabledDate flags. In other words, this block of information contains network metadata and pointers to the associated ODM instance and Web services. The NetworkServicesGetValuesURL parameter is included to distinguish between situations where only the site catalog is imported into ODM while observation values are retrieved from remote repository (for example, national repositories such as USGS NWIS, EPA STORET, or regional repositories such as TCEQ (Texas Commission on Environmental Quality) or TCOON (Texas Coastal Ocean Observation Network)) versus observation data being imported into ODM instance as well.
7. The user restarts the DASH Web application with the updated .MXD and web.config files, and tests it.

Current Status

The user interface of the DASH Web application in its current version allows online users to query observation networks by location and attributes, selecting stations in a user-specified area where a particular variable was measured during a given time interval. Once one or more stations and variables are selected, the user can retrieve and download the observation data for further off-line analysis.

At the time of this writing, the first version of the software stack is being finalized, and the San Diego Supercomputer Center team is configuring servers to be shipped to WATERS testbed projects. At the conference, the initial deployment experience will be reported.

Automated Ground Motion Characterization Using Satellite Imagery

By Alan Yong[1], Susan E. Hough[1], Chris J. Wills[2], and Michael J. Abrams[3]

[1]U.S. Geological Survey, Pasadena, Calif.

[2]California Geological Survey, Sacramento, Calif.

[3]Jet Propulsion Laboratory, Pasadena, Calif.

Estimation of local site effects requires precise information about geomorphic and geologic conditions. On the basis of our ongoing effort (Yong and others, 2007), we report on our progress using multiresolution, remote-sensing data to provide preliminary site characterizations for estimating ground-motion intensities in future earthquakes. Use of remote sensing to determine site effects is especially important in many regions of the world, such as Pakistan, Mozambique, and Turkey, where local geologic information is either sparse or not readily available. Even where local geologic information is available, details in the traditional maps often lack the precision or the level of information required for effective microzonation. We use readily available satellite-based data and state-of-the-art object-based image analysis methods to address this problem. Our datasets consist of (1) optical imagery that includes regions in the short-wave infrared (SWIR) and thermal infrared (TIR) domains, and (2) relative digital elevation models (DEMs), based on stereoscopic-correlation methods, derived from National Aeronautics and Space Administration's ASTER (Advanced Space-borne Thermal Emission and Reflection Radiometer) sensors. On the basis of geomorphology and geology, we apply automated-feature extraction methods to determine the local terrain. Then, on the basis of the site-classification schemes from the Wills and others (2000) and Wills and Clahan (2006) maps of California, we assign shear-velocity-to-30-m-depth (Vs30) values to selected regions in California, Mozambique, Pakistan, and Turkey. We compare our results to available empirical data and discuss the applicability of our site-class assignments in each region and the implications of its effects on seismic hazards assessment.

References Cited

Wills, C.J., and Clahan, K.B., 2006, Developing a map of geologically defined site-condition categories for California: Bulletin of the Seismological Society of America, v. 96, p. 1483-1501.

Wills, C.J., Petersen, M., Bryant, W.A., Reichle, M., Saucedo, G.J., Tan, S., Taylor, G., and Treiman, I., 2000, A site conditions map for California based on geology and shear wave velocity: Bulletin of the Seismological Society of America, v. 90, p. S187-S208.

Yong, A., Hough, S.E., Wills, C., and Abrams, M., 2007, Site characterization using satellite imagery: A Progress Report, in Proceedings of the 2007 SSA Annual Meeting, Kona, Hawaii, April 11-13, 2007, p. 272.

QuakeML—XML for a Seismological Data Exchange Infrastructure

By Danijel Schorlemmer[1] and Fabian Euchner[2]

[1]Department of Earth Sciences, University of Southern California, Los Angeles, Calif.

[2]Swiss Seismological Service, ETH Zurich, ETH Hoenggerberg, Zurich, Switzerland

We report on the progress of the development of QuakeML. QuakeML is a flexible, extensible, and modular Extensible Markup Language (XML) representation of seismological data that is intended to cover a broad range of fields of application in modern seismology. QuakeML is an open standard and is developed by a distributed team in a transparent, collaborative manner. The first part of the standard, QuakeML—Basic Event Description, will be subjected to a Request for Comments process. The standardization process for inventory information and resource metadata is also underway. The flexible approach of QuakeML allows further extensions of the standard in order to represent waveform data, macroseismic information, location probability density functions, moment tensors, slip distributions, shake maps, and others.

QuakeML is developed in parallel with a UML representation of its data model. This allows an elaborate software development strategy which uses the unified modeling language (UML) class model together with a custom UML profile as the basis for automated code generation. With this technique, a library of C++ classes is generated which can be serialized either to XML (QuakeML) or to Structured Query Language (SQL) for persistent storage in a relational database. The XML Schema description is created automatically from the UML model with the help of tagged values, which describe the mapping from UML class attributes to XML representation. The library approach makes it easy for application developers to include QuakeML support in their products, since no one source code has to be written. Serialization of objects to and from QuakeML format will be supported by the Application Program Interface (API). It is possible to use the QuakeML library from other object-oriented programming languages (for example, Java and Python) using wrappers.

The QuakeML language definition is supplemented by a concept to provide resource metadata and facilitate metadata exchange between distributed data providers. For that purpose, we propose a Uniform Resource Identifier (URI)-based format for unique, location-independent identifiers of seismological resources that are assigned by approved naming authorities. QuakeML-Resource Metadata defines a Resource Description Framework (RDF) vocabulary for resource metadata description, covering the resource's identity, curation, content, temporal availability, data quality, and associated services. We propose to set up a network of registry institutions that offer Web services for resolving resource identifiers into corresponding RDF/XML metadata descriptions, and that additionally provide means for resource discovery by offering services for searches against resource metadata.

Currently, the QuakeML development team is bringing together people from University of Southern California Eidgenossische Technische Hochschule (ETH) Zurich, GeoForschungsZentrum Potsdam (GFZ Potsdam), U.S. Geological Survey, and Incorporated Research Institutions for Seismology (IRIS). QuakeML will be used in the Network of Research Infrastructure for European Seismology (NERIES) framework in Europe by the SeisComp3 software, by the European-Mediterranean Seismological Centre (EMSC), and by the Southern California Earthquake Consortium (SCEC) Collaboratory for the Study of Earthquake Predictability; it is also under consideration by the Advanced National Seismic System (ANSS). We are confident that, in combination with further standardization efforts, the concept of QuakeML can contribute to facilitate data exchange and interoperability of seismological data providers.

The EarthScope Plate Boundary Observatory Distributed Data Management System

By Greg Anderson[1], Mike Jackson[1], and Charles M. Meertens[2]

[1]University NAVSTAR (Navigation Signal Timing and Ranging) Consortium (UNAVCO), University of Colorado, Boulder, Colo.

[2]UNAVCO Facility, UNAVCO, Inc., Boulder, Colo.

EarthScope is an ambitious, multidisciplinary project funded by the National Science Foundation to explore the structure and dynamics of the North American continent. The Plate Boundary Observatory (PBO) is EarthScope's geodetic component, and will measure the four-dimensional strain field resulting from active tectonic deformation in the western United States. The University Navigation Signal Timing and Ranging (NAVSTAR) Consortium (UNAVCO) is installing and will operate the PBO network of more than 1,000 continuous global positioning system (GPS) borehole and laser strainmeters, seismometers, and tiltmeters. As of February 2007, 561 of these stations have been installed.

The flow of data from these stations is managed from our Boulder, Colo., Network Operations Center (NOC), located at UNAVCO Headquarters. Automated systems at the NOC retrieve data from our stations at least daily, monitor the status of the network and alert operators to problems, and pass data on for analysis, archiving, and distribution. Real-time network status can be found at http://pboweb.unavco.org/soh_map.

PBO's analysis centers generate high-quality derived data products from PBO raw data. Central Washington University and the New Mexico Institute of Mining and Technology process raw GPS data to produce initial PBO GPS products, including network solutions and station position time series; the Analysis Center Coordinator at Massachusetts Institute of Technology (MIT) combines these products into the official PBO GPS products. Staff of UNAVCO and the University of California—San Diego, process data from the PBO borehole and laser strainmeter networks and produce cleaned time series of shear, areal, and linear strain, Earth tides, pore fluid pressure, and other parameters.

The UNAVCO Facility archives and distributes all PBO GPS data products and runs a secondary archive offsite; currently, these centers hold over 2.5 terabytes (TB) of PBO products. The Incorporated Research Institutions for Seismology (IRIS) Data Management Center and Northern California Earthquake Data Center archive and distribute all PBO strainmeter data products, and IRIS archives all PBO seismic data products; more than 160 GB of data products are available from these archives. These same centers also archive other EarthScope seismic and strain data.

The PBO Web site (http://pboweb.unavco.org) provides centralized access to PBO products stored in our distributed archives. GPS products may be accessed from http://pboweb.unavco.org/gps_data and strain data products from http://pboweb.unavco.org/strain_data. In addition, the individual archives provide access to their holdings, both for PBO and other networks, through a variety of discipline-specific tools.

The most exciting development still to come in providing access to EarthScope data products will be the creation of the EarthScope Portal. This system will be based on Web services operated by the EarthScope components that provide access to holdings at the EarthScope archives and that are linked to a central Web portal. This system will provide a unified system for discovery and access to EarthScope digital data products, and is planned to be operational by October 2008.

EarthRef.org in the Context of a National Cyberinfrastructure for the Geosciences

By Hubert Staudigel[1], Anthony Koppers[2], Catherine Constable[1], Lisa Tauxe[2], and John Helly[3]

[1]Scripps Institution of Oceanography, University of California—San Diego, La Jolla, Calif.

[2]Scripps Institution of Oceanography, La Jolla, Calif.

[3]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

EarthRef.org is the information technology (IT) umbrella covering four independent grass-roots Cyberinfrastructure (CI) initiatives in distinct disciplines of earth and life sciences and education: (1) The Geochemical Earth Reference Model initiative (GERM, http://EarthRef.org/GERM/) is a consortium of leading geochemists who promote understanding of Earth chemistry on a planetary scale, by sponsoring scientific workshops and general improvements in CI. GERM was started more than a decade ago, and shortly after its inception, it played a pivotal role in the creation of the widely used GERM Web site and the development of the American Geophysical Union's electronic journal, G-cubed. GERM has also served as a forum for the continued development of several independent databases, such as PetDB and GeoROC. The GERM Web site offers a range of resources, including compositional data, partition coefficient data, and computational tools. It is also a source for geochemical publications, such as GERM conference circulars and abstract volumes. (2) The Seamount Catalog (http://EarthRef.org/SC/) is a data repository for seamount geographic information, maps, and a wide range of other seamount data. It is also the central Web site of the recently founded Seamount Biogeoscience Network (SBN; http://EarthRef.org/SBN/) and it aims to provide an equally useful CI for disciplines that range from marine geophysics to microbiology to fisheries, with the goal of helping to integrate the research from these diverse disciplines on seamounts. (3) The Magnetics Information Consortium's (MagIC; http://EarthRef.org/MAGIC/) goal is to provide a data archive and interactive visualization Web site for paleomagnetism and rock magnetism. This is a community-driven database, where users can upload, search, and use newly derived scientific data as well as data from peer-reviewed legacy publications. (4) The Enduring Resources for Earth Science Education (ERESE; http://EarthRef.org/ERESE/) project is part of the National Science Digital Library (NSDL) and offers a wide range of educational resources, in particular, for the teaching of plate tectonics. ERESE combines teachers' professional development with the development of digital library resources, and provides Web seminars in collaborations with the National Science Teacher's Association (NSTA). These four initiatives have a broad range of clients, with dramatically different database and IT needs, distinct working styles, and variable levels of comfort with CI and IT. However, combining these diverse CI components under the EarthRef.org umbrella has been quite cost-effective, and allows each effort to use common IT resources and to share software development work. This philosophy has also proved rather appealing for its end users: EarthRef.org now has over 2,500 registered users, has received 188,000 unique users over the last three years, and expects to receive roughly 120,000 unique users in 2007.

EarthRef.org has a number of key common features that reach across the GERM, MagIC, SBN, and ERESE initiatives, many of which we consider important to establish functional CI in Geosciences. Some of these features follow:

Collaboration between IT developers and leading members of the science community—Much of EarthRef.org has been developed by, or in close collaboration with, active Earth scientists who produce and use data, advise students, and have substantive current publication and science-funding activity. Their personal and practical experience and perspectives on how Earth science is done successfully are a major asset in these developments and contribute directly to the resulting EarthRef.org Web environment.

Integrating IT development with science conferences—EarthRef.org organizes top-level scientific conferences with invited speakers that synthesize the state-of-the-art in a given field. Leading scientists involved in major recent scientific developments are chosen as keynote speakers for these conferences. We use these science conferences to help a community hone its science vision, with an eye on the CI needs that will ultimately address and solve their "grand challenges."

Robust hardware-software requirements—EarthRef.org is hosted by the San Diego Supercomputer Center. This hosting guarantees deep archiving and safeguarding of all scientific data uploaded to the EarthRef.org databases. It also provides EarthRef.org with continuous upgrades for the underlying hardware, ensuring that its databases and Web sites operate with the best technology standards available.

Laying the foundation for a large-scale cooperation to keep scientific content up-to-date—Much care has been taken to enable the EarthRef.org users to participate (and ultimately take over) data population for legacy as well as primary (new) data. To that end we have created various data formats and upload wizards that are easy to use via the EarthRef.org Web site. Data uploaders have the option of a proprietary hold where unpublished data are not accessible to the public. In other words, they can keep their data private and view them within the context of the existing EarthRef.org databases. Users can also provide group names and passwords in order to authorize limited groups of coworkers, students, teachers, or even reviewers to query and visualize their private data.

Archiving original records for legacy data for user-based quality control and quality assurance—Legacy data entries into EarthRef.org are accompanied by several file types, including the original scanned image of data tables, and data files generated by optical character recognition (OCR) of the scanned image. This allows a user to trace data to the original source and to explore scanned images as a cause for errors.

Establishing an information continuum that ranges from top science to basic education—Earthref.org, and in particular ERESE, work to bridge the gap between science and general education through a genuine collaboration between educators and scientists, and by making science database contents accessible for education and public outreach. Some of this is accomplished through specific contents developed for the educational community, but there are also some basic metadata that allow educators to screen for specific database contents. To this end, all EarthRef.org data contents carry metadata on

the minimum expert level needed to be able to work with a particular digital object. This scale ranges from one (the most basic primary school level) to nine (the expert scientist level). This relative scale allows educational users to extract objects for use in the classroom or curriculum design by browsing based on a specified expert level.

Provide a set of basic CI building blocks that may be used by diverse CI efforts—A wide range of features are common to all EarthRef.org initiatives, including an online earth-sciences address book, a digital library archive, and an earth-science reference database. The address book allows us to keep track of users uploading new data; however, more importantly, because it is entirely Google-friendly, it allows users to easily find contact information for their colleagues. The EarthRef Digital Archive (ERDA) is a multi-purpose digital library that can archive any arbitrary digital object (ADO) and does provide the basic machinery for deep storing of ADOs and for linking data files to the GERM, MagIC, SBN, and ERESE portals. The EarthRef.org Reference Database (ERRD) contains close to 100,000 references from earth-science publications as provided directly by the publishers. This is an invaluable resource that allows us to confidently link all the user-provided data to the original publisher and to the publications on their respective Web sites.

Geospatial referencing through Google Maps—Geospatial parameters and data type are amongst the top-rated search parameters for geosciences data. The Google Maps interface offers an intuitive representation of any parameter. We have begun the development of a Google Maps interface for seamounts based on a combination of the multibeam data and satellite altimetry data stored in Earthref.org. However, this interface is equally useful for searching the SBN, MagIC, and ERDA databases and has been implemented within each of these search portals.

EarthRef has been built by scientists and educators for scientists and educators, and our experience has been that much can be done with well-known IT components focusing on key practical and some more visionary matters. Practical issues include the ease and efficiency of information and data acquisition to use, reuse, modeling, and visualization and a close link between science and education at all levels. Vision issues focus on anticipating future directions of education, science, and, in particular, multidisciplinary science integration. This type of vision has to translate into a specific CI design with a meaningful metadata structure and ontologies that allow us to combine data in new ways. There is a general consensus that a successful CI will create an environment that advances science in profound ways, helping achieve new levels of understanding and addressing the grand challenges. Key to such a CI will be community buy-in and ownership. We argue that this ownership has to start right from the beginning of the development, whereby the "science process" is profoundly integrated with CI development.

Phanerozoic Earth and Life: The Paleointegration Project

By Allister Rees[1], John Alroy[2], Christopher Scotese[3], Ashraf Memon[4], David B. Rowley[5], Judith Totman Parrish[6], David B. Weishampel[7], Emil Platon[8], Maureen A. O'Leary[9], and Mark A. Chandler[10]

[1]Department of Geosciences, University of Arizona, Tucson, Ariz.

[2]Paleobiology Database, NCEAS, University of California—Santa Barbara, Santa Barbara, Calif.

[3]PALEOMAP Project, University of Texas at Arlington, Arlington, Tex.

[4]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

[5]Department of Geophysical Sciences, University of Chicago, Chicago, Ill.

[6]Department of Geological Sciences, University of Idaho, Moscow, Idaho

[7]Department of Cell Biology and Anatomy, The Johns Hopkins University School of Medicine, Baltimore, Md.

[8]Energy and Geoscience Institute, University of Utah, Salt Lake City, Utah

[9]Department of Anatomical Sciences, Stony Brook University, Stony Brook, N.Y.

[10]Center for Climate Systems Research, Columbia University, New York, N.Y.

The Paleointegration Project (PIP) within the Geosciences Network (GEON) (http://www.geongrid.org/) is facilitating interoperability between global-scale fossil and sedimentary rock databases, enabling a greater understanding of the life, geography, and climate of our planet throughout the Phanerozoic. The key elements of PIP are databases, paleomapping tools, and Web services.

The following databases are presently in the system (see table 1 for temporal distributions): PBDB—The Paleobiology Database (~630,000 occurrences of marine invertebrates, vertebrates (including dinosaurs), plants, and microfossils from 69,000 collections); DINO—From "The Dinosauria" Encyclopedia, 2004 (~4,200 occurrences from 1,200 localities); GCDB—Graphic Correlation Database (~108,000 fossil occurrences from 2,000 localities, with over 700 interpreted localities that define taxon ranges); PGAP—The Paleogeographic Atlas Project (~135,000 sedimentary rock occurrences from 47,000 localities); CSS—Climatically significant sedimentary rocks (~13,000 occurrences from 3,600 localities);

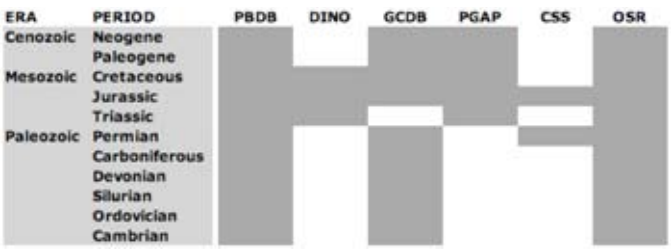

Table 1. Temporal distributions of databases within the Paleointegration Project (PIP).

OSR—Organic-rich sedimentary rocks (~2,000 occurrences from 1,600 localities, including geochemical data).

The databases are text- as well as map-searchable, through the use of age and geography ontologies, linked to geographic information systems (GIS) mapping tools. Results are viewable on present-day maps as well as paleomaps, and can also be downloaded for further detailed analyses. An important feature of PIP is the calculation of locality paleocoordinates "on the fly," based on modern latitude and longitude as well as locality age. This was achieved by developing a module (Auto Point Tracker, or APT) that enables dynamic calculations of locality paleocoordinates and automated plotting on the paleogeographic maps. PIP was designed to ensure fast data retrieval, making it especially useful for extensive as well as multiple simultaneous searches (for example, in a classroom setting). Through the use of Web services developed at the San Diego Supercomputer Center (SDSC), the PBDB server in Santa Barbara is searchable dynamically within PIP, alongside the other databases hosted at SDSC. In addition to the PIP features, each PBDB locality record is linked back to the original database, enabling the user to either explore and analyze the data further within that system (for example, using the PBDB statistical tools) or to continue within the PIP interface.

The IT architecture of PIP (fig. 1) shows how the different databases and components interact to produce an integrated result and map. Researchers send a query for data using the PIP user interface and this query is parsed by the PIP middleware. It is then decomposed into several different queries sent to the databases, using Java Database's Applications Program Interface (API) for the ones hosted on SDSC servers and using Web Services for querying data hosted at remote locations (for example, the PBDB). Once all the results are accumulated, they are passed to the APT Web Service for "on-the-fly" calculations of locality paleocoordinates so that they can be plotted on the paleogeographic maps.

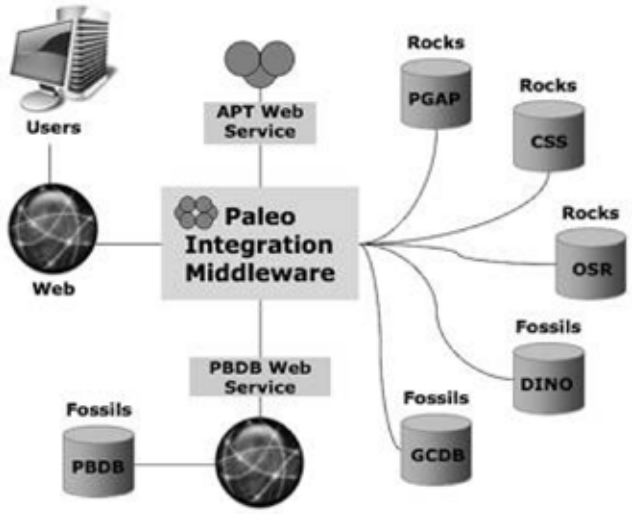

Figure 1. The information technology (IT) architecture of the Paleointegration Project (PIP).

The Paleointegration Project is already proving useful to researchers, teachers, and students. Anyone can now access data and tools that were only available previously to a few specialists. It should also prove to be an excellent resource for a new generation of projects that assimilate both paleoclimate models and data for more detailed views of the Earth's climate history. Complex computational tools like Global Climate Models (GCMs) simulate details of the Earth's atmosphere, oceans, and land-surface processes that are beyond what proxy interpretation alone can provide; however, modelers require detailed paleogeographic data, which is used to construct the type of boundary conditions used as GCM input. The PIP databases will also be useful to paleoclimate modelers needing access to proxy climatological data for use in model verification. Accessibility to GCMs is now improving through programs like the Educational Global Climate Modeling Project (see EdGCM at http://edgcm.columbia.edu); thus, accessibility to paleo databases is crucial to the development of quality data/model integration projects. We emphasize that PIP does not replace specialist expertise. It does, however, provide another means whereby researchers can develop their own scientific queries.

We envisage continuing to develop PIP with the addition of new datasets, tools, and services. The next phase will include the MorphoBank Database, which contains phylogenetic systematics of morphological data (~2,300 anatomical images and 23 phylogenetic matrices). The PIP will enable phylogeneticists currently using MorphoBank to search seamlessly for relevant fossil taxa in the PBDB and to generate, for example, survivorship plots of diversity through time. It will also include the online PaleoReefs Database (PARED) developed by Wolfgang Kiessling (Humboldt-University of Berlin), which contains data for 3,550 reef complexes throughout the Phanerozoic (with details about paleontology, architecture, environmental setting, and petrography). Some 35,000 reefal taxonomic occurrences from PARED are stored in the PBDB, so seamless integration of the additional reef architectural and environmental details will be very useful. More tools and services will be required as additional databases are integrated, to ensure that diverse user needs and interests are addressed. We plan to work with projects like PaleoStrat to develop some of these new tools and databases within PIP.

A more complete understanding of the interactions between Earth and life through time also requires the integration of geochemical, geophysical, igneous, and metamorphic data. From this broader "geoinformatics community" perspective, the Paleointegration Project demonstrates that disparate geologic databases residing on different servers can be searched seamlessly using embedded modules, Web services, and GIS—structures and tools that will greatly facilitate future collaborative efforts.

MOAS: Geoinformatic Database for Integration and Synthesis in Coastal Modeling

By Clifford J. Hearn[1], Matthew I. Streubert[1], and Michael A. Holmes[2]

[1]ETI Professionals, St. Petersburg, Fla.

[2]U.S. Geological Survey, Tampa, Fla.

Within the last decade, integrated modeling of coastal systems has developed as a new methodology through advances in the technologies of data acquisition, data synthesis, and simulation modeling. Processing of field and model data in multidisciplinary, integrated science studies is a vital part of synthesis modeling. Collection and storage techniques for field data vary greatly between the participating scientific disciplines due to the nature of the data being collected, whether it is in situ, remotely sensed, or recorded by automated data-logging equipment. Spreadsheets, personal databases, text files, and binary files are used in the initial storage and processing of the raw data. Network Common Data Form (NetCDF) files are created as output from intermediate processing procedures for portability and machine-independent sharing of time-series and array-oriented datasets. In order to be useful to scientists, engineers, and modelers, the data need to be stored in a format that is easily identifiable, accessible, and transparent to a variety of computing environments. The Model Operations and Synthesis (MOAS) database and associated Web portal were created to provide such capabilities. The industry-standard relational database is comprised of spatial and temporal data tables, shape files, and supporting metadata, accessible over the network through a man-driven, Web-based portal, or spatially accessible through ArcSDE connections from the user's local geographic information systems (GIS) desktop software. A separate server provides public access to spatial data and model output in the form of attributed shape files through an ArcIMS Web-based graphical user interface. Structured externally to MOAS is a geospatial grid termed the "Universal Grid." This grid, along with temporally referenced shape files, reside in an Enviromental Systems Research Institute (ESRI) .MXD file and serve as the framework for data visualization, manipulation, and utilization, which are essential components of the integration and synthesis of costal models and data. The Universal Grid provides a user-friendly GIS interface to the data within MOAS through point and click menus for easy navigation, display, and exportation of data. Each cell or group of cells within the grid can be linked directly to the database using the "add data" function within ArcGIS, then interrogated, analyzed, and compiled within the interface. The data can then be extracted in database, spreadsheet, or text-file format. The interface supports rapid display of attributes without the necessity of direct user manipulation of color bars, parameters, and fields. The interface also supports "on-the-fly" projection changes, and point-click navigation, as well as many other tools under development. This approach serves as a viable solution to the ever-growing need to access large amounts of data for modeling, mapping, and synthesis. Users of the Universal Grid require only minimal knowledge or training in GIS to take advantage of its vast capabilities.

The Tampa Bay Integrated Coastal Model is used as a case study in this paper, building on the wealth of published research and modeling information staged within the estuary. The integration and storage of physically collected, synthesized, and modeled data are described in studies concerning hydrodynamic, biogeochemical, sedimentary, and hydrological processes.

Automated Multidisciplinary Collection Building

By Caryn Neiswender[1], Stephen P. Miller[1], Dru Clark[1], John J. Helly[2], Don Sutton[2], and John Weatherford[2]

[1]Scripps Institution of Oceanography, University of California—San Diego, La Jolla, Calif.

[2]San Diego Supercomputer Center, University of California—San Diego, La Jolla, Calif.

Geologic disciplines often need to combine both historic and current data, observations, and interpretations. If data collections exist, they tend to be homegrown, often drawing on the isolated expertise of investigators, with minimal technological support. Expensive to acquire at $25,000 per day, and generally impossible to recreate, shipboard data are potentially valuable for a wide range of disciplines, far beyond the award that funded the original expedition; however, data need to be discovered before they can be used, and appropriate metadata need to be generated to support effective, wide community access.

Scripps Institution of Oceanography Explorer (SIOExplorer) presents, as a scalable solution, a multidisciplinary digital library of over 50 years of shipboard data. Based upon an extensible metadata scheme (fig. 1) and implemented with technology from the San Diego Supercomputer Center, SIOExplorer enables discovery of data collected onboard Scripps Institution of Oceanography (SIO) research vessels. The SIOExplorer digital library consists of over 700 SIO cruises, with more than 100,000 digital objects, including datasets, documents, and images.

The efforts are being extended to the collections of the Woods Hole Oceanographic Institution (WHOI) to include cruises, Alvin submersible dives, and remotely operated vehicle (ROV) lowerings. The technology also supports the efforts of thousands of scientists from dozens of nations with the Site Survey Data Bank of the Integrated Ocean Drilling Program (IODP).

Streamlined procedures have been developed to stage the data, extract metadata from data files, perform qual-

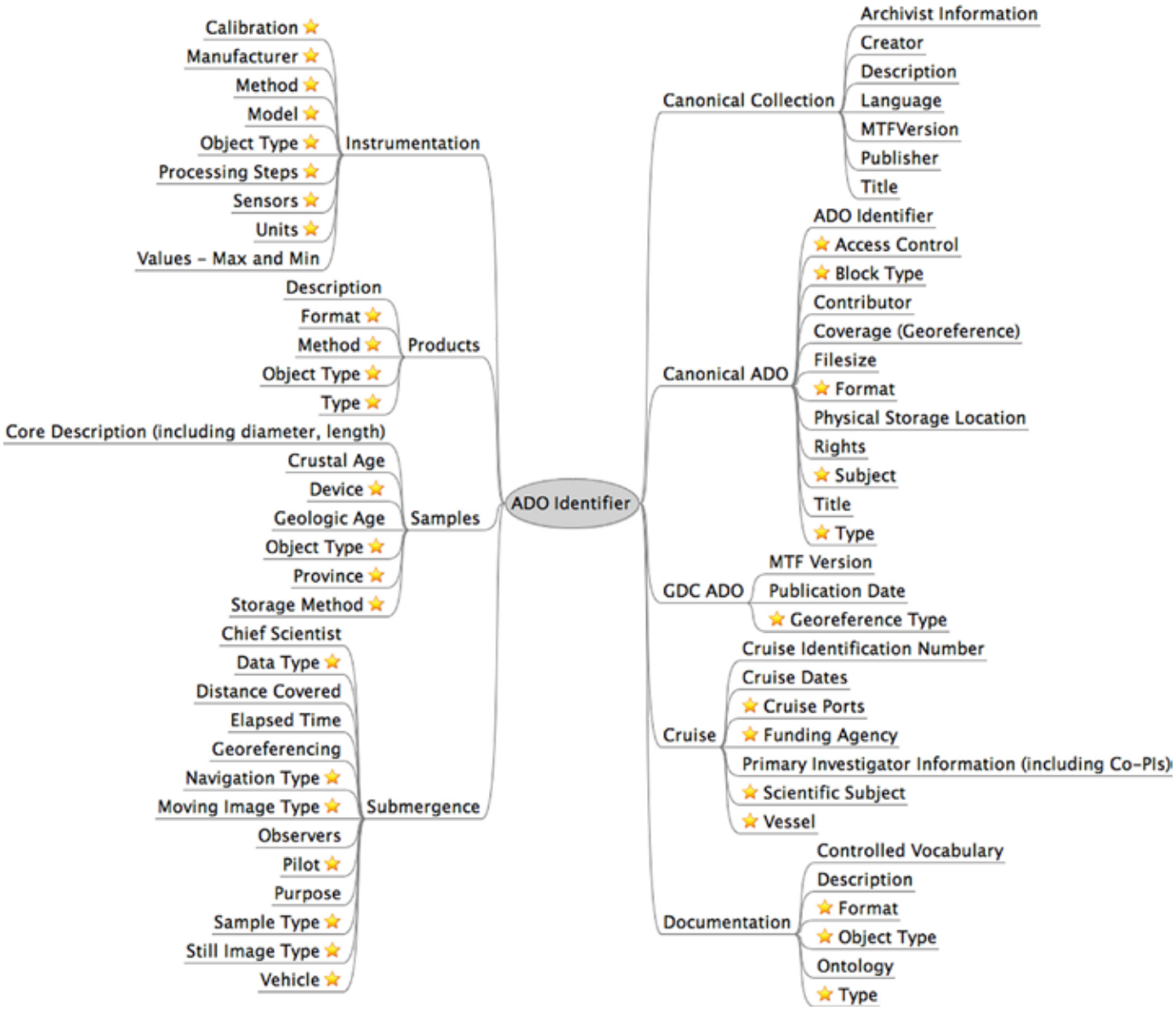

Figure 1. SIOExplorer metadata scheme. Every arbitrary digital object (ADO) is associated with a metadata structure complete with descriptive information. Starred entries indicate fields that are managed with controlled vocabularies.

ity control and error correction, and publish metadata and data in a searchable digital library. Discovery tools search a PostgreSQL database with metadata, and deliver relevant objects from a Storage Resource Broker (SRB) instance. The system provides both text-based Asynchronous Java Script and XML (AJAX) Web-form and interactive geographical Java interfaces.

Information Infrastructure

The SIOExplorer is a unique collaboration between scientists, technologists, and educators. The lifecycle of information within the digital library includes data creation, quality control and data packaging, publication via the Web, and data access/analysis use. To support each of these activities, a scalable scientific and technological infrastructure has been developed.

The infrastructure for the digital library is defined by a metadata scheme. Currently broken into eight sections, or blocks, the SIOExplorer metadata template file (mtf) file comprehensively describes the scientific framework in which the digital object was created.

Interoperability and Controlled Vocabularies

Interoperability has long been a goal of digital library implementations. To facilitate collaboration, tools are developed that can be utilized by other projects, and (or) adapted

to specific implementation needs. One example is the recent development of the controlled vocabulary dictionary. In addition to a list of acceptable values for metadata parameters, the controlled vocabulary dictionary incorporates a human-readable explanation of each allowed value. Scientists, technologists, and project managers can access this controlled vocabulary dictionary, which enables accurate interpretation of the metadata.

As a member of the Marine Metadata Interoperability (MMI) project, SIO promotes the exchange, integration, and use of marine data through enhanced data publishing, discovery, documentation, and accessibility. The MMI community works to develop best practices and guidance on project development and implementation. Community members include metadata experts, scientific researchers, and project managers.

Automatic Harvesting in an Imperfect World

Over the years, and across projects and disciplines, there has been a tendency for descriptive terminology within metadata to wander. Some of the variation is due to evolution in sensor technology, but some may be due to odd abbreviations, typographical errors on rolling decks, institutional practices, or a momentary inspiration to use a new term. As a consequence, we now face challenges in searching digital collections, and in designing reusable tools that can be applied to multiple institutions.

Practical experience with SIOExplorer has enabled the development of techniques to assess variations in metadata values across collections. The assessment helps to guide the development of controlled vocabularies, which in turn can be used to enable automatic detection of metadata errors, and in some cases automatic correction.

Controlled vocabularies underlie an emerging set of tools that support Web-user interfaces, large-scale automatic harvesting of metadata and data, project status assessment, workflow management, and overall quality control. They are a key resource for user upload code in the IODP Site Survey Data Bank, prompting and enforcing appropriate metadata values for ocean-drilling proposal-support data. Compared to previous generations of hard-wired code, the access to controlled vocabularies allows a project to evolve with flexibility, and the code to be ported from one project to another.

Related links

- SIOExplorer—http://SIOExplorer.ucsd.edu
- SSite Survey Data Bank—http://ssdb.iodp.org/
- SMarine Metadata Interoperability—http://ssdb.iodp.org/
- SStorage Resource Broker—http://www.sdsc.edu/srb/index.php/

Sensor Web Enablement—Its Impact on Continuous Digital Workflows

By Peter Loewe[1], Jens Klump[1], and Joachim Waechter[1]

[1]GeoForschungsZentrum-Potsdam (DRZ), Potsdam, Germany

A major challenge for the use of scientific data is caused by the absence of continuous digital workflows. Research in the earth sciences can span several dimensions and orders of magnitude in time and space, crossing scientific domain boundaries, which results in the fields' semantical richness. This research led to the accumulation of a tremendous amount of literature, data, and sample collections. So far, all the findings have been separated and underutilized due to the absence of continuous (digital) workflows providing sustainable data management. The advent of information technology and Internet-based Web services has created new options to improve this situation. They offer ways to integrate literature, data, and samples from the source of data in the field or laboratory, all the way to their interpretation in the literature. Yet, numerous breaks between data capture and data processing remain to be closed. The Sensor Web Enablement (SWE) standards currently developed by the Open Geospatial Consortium can be used to close the gaps. They provide integrative approaches to environmental monitoring and the capturing of real-time data in earth observing systems. The paradigm of Sensor Web Enablement extends beyond its application in environmental sensor networks. Its ability to model any data source or process as a sensor that takes in data and puts out processed data, in conjunction with directory and system management services, makes it a universal tool for earth-science data. In this paper, we look at the roles of SWE software tools provided by the open-source software communities for the creation of continuous digital workflows in the earth sciences. These workflows can be used for monitoring and data capture, data processing and modeling, creation of added value chains, and new forms of publication.

A System for Fast Spatial Searches on the Earth or Sky Using the Hierarchical Triangular Mesh

By Gyorgy Fekete[1]

[1]Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, Md.

Spatial searches represent the most frequent search patterns, both in astrophysics and in earth sciences. In each of these areas, scientists use a multitude of different coordinate systems, describing the distribution of both point-like and extended objects over the sphere. Many cutting-edge science projects today involve multiple sources, and thus require a

detailed fusion and federation of these datasets. Very often drop-outs (objects that are not detected in one of datasets) are particularly interesting, since they could represent objects that are quite "small." Such searches require a detailed knowledge of what area of the globe has been covered by a particular instrument (Earth- or satellite-bound). Once the common area (the intersection of the respective footprints) has been determined, we need to find the relevant objects in this area, and perform a spatial cross match, and find all detections that may correspond to the same physical object. All this requires a rather sophisticated set of tools. Requirements are rather stringent; satellite orbits consist of long stripes over the sky, and the required accuracy is determined by the resolution of the best detectors. We need to track objects over hundreds of degrees with a resolution of a few milliarcsecs. We need a framework that can be used in several different contexts, like stand-alone user applications, Web services, and Web applications, and that can also be embedded into existing commercial database systems. We have been working on this problem for over 12 years. In the mid-1990s we wrote a simple C-based package called the Hierarchical Triangular Mesh (HTM) that is still being used by several observatories and NASA centers. We then substantially rewrote the library to C++. This package is currently in use at multiple institutions and projects worldwide.

We address two separate mathematical issues. The first is about the correct abstract mathematical representation of complex spherical regions, and about implementations of various set operations (union, intersection, and difference), morphological functions (dilation and erosion), and scalar functions (area). Beyond the optimal algorithms for these operations, there are also serious challenges related to numerical precision; for instance, determining whether two planes exactly parallel. The second issue is a discrete pixelization of the sphere. For fast spatial searches we use data structures suitable for fast search algorithms. We use the sphere quadtree to build a hierarchy of triangle-shaped pixels (trixels) on the globe, thus the name—Hierarchical Triangular Mesh (HTM).

There are many ways to represent shapes on the surface of the unit sphere. We chose a system where there are no singularities at any pole, employing three-dimensional Cartesian unit vectors to describe locations on an idealized (unit radius) sphere. The most basic building block is the half space, a directed plane that splits the three-dimensional space into two halves. It is parameterized by a normal vector "n" and a scalar "c" (distance from the origin). When intersected with the unit sphere, a half space describes a spherical cap. Inclusion of a point inside such a cap is decided by a simple dot-product computation.

A so-called "convex" is the intersection of many half spaces. A rectangle is described by a convex of four half spaces, the sides of the rectangle. A region, the most general shape, is simply a union of convexes.

The normal form of a region is, therefore, a union of intersections. Very often the shapes are more conveniently represented by other well-known primitives, such as rectangles, spherical polygons, circles, and the convex hull of a point set; using these familiar terms, we provide a full complement of shape functions that convert descriptions into the normal form.

As always, the devil is in the details. Floating-point computations are subject to rounding errors; a number of mathematical issues arise from these inherent uncertainties in our project. What two points or half spaces should be considered identical? When is a point exactly on an arc? Computer geometry libraries working in Euclidean space sometimes avoid these issues by utilizing exact arithmetic, mathematical formulas on fractional numbers represented by integer numerators and denominators; however, this slower work-around is not an option for solving the spherical geometry due to the normalization constraint, which involves taking the square root of the coordinates. We use the IEEE-754 standard, double precision numbers, and derive the limitations from the actual number representation. At the core of many formulas is the cross product to find perpendicular directions. Colinear vectors have vanishing cross products and numerically this can be tested in a robust way by comparing the largest coordinate to the double precision. If it is too small, its square root cannot be taken to normalize the vector and the indefinite perpendicular direct means colinearity. A similar problem is solving for the roots of two circles on the sphere. If the computation is inaccurate, the roots will not be on the circles numerically. The sweet spot for double precision numbers is at about the tolerance level of 10^8 radians, which corresponds to a few thousandths of an arc second. This is about a foot in size on the surface of the Earth.

There are two basic kinds of spatial constraints in a query: (1) "Is this object within some distance of a target?" and (2) "Is this object inside some region?"

Both kinds involve costly geometrical computation, so we devised a mechanism that eliminates objects from consideration quickly. Crude boxing techniques, which elect candidates by boxing latitude and longitude values, work well only if the region of interest is shaped like a rectangle aligned with the latitude and longitude grid; however, they are less effective for other shapes, like those that contain the pole or narrow stripes that have a high inclination to the primary great circles. The main idea is that we implement a coarse filter whose job is to reject all objects that are certain to fail the spatial constraint, and use the fine filter for only the few false positives. The goal is to make the coarse filter as good as possible but also as fast as possible.

In a database that has a user-defined function called AngularDistance and a table of Objects that have columns named ObjID, Lat and Lon, a simple query that selects objects within half a degree of a point (Lon = 12, and Lat = -30) is as follows:

```
select ObjID from Objects o where
o.AngularDistance(12, -20, o.Lon, o.Lat) < 0.5 .
```

The costly AngularDistance function would be evaluated for each object in the table. Instead, if we implemented the coarse filter as a function that produces a table that would be

efficiently computable on the fly, then the query is augmented with a join as follows:

```
select ObjID from Objects o join CircleCover(12, -20,
0.5) c on o.HtmId between c.lo and c.hi

  AND AngularDistance(12, -20, o.Lon, o.Lat) < 0.5 .
```

The second example presumes the existence of an HtmID column and the CircleCover table-valued function that returns rows of low, high values that constrain the possible HtmId values. Only those objects that pass the first constraint are given to AngularDistance for the precise calculation. The function CircleCover here produces a table that represents a circular region of interest centered on the given location and radius; however, our methods are capable of expressing arbitrary shapes with the full richness of the Region-Convex-Half space paradigm. The essence of the idea is that the inside of every trixel is turned into a range query in a database over the HtmId values. We can exploit the relational database engine's ability to use the HtmID as an index for very rapid operation.

With the advent of modern silicon-based detectors, the way observational science is done is changing rapidly. To take full advantage of the avalanche of data, we need scalable information systems that provide reliable access via interoperable interfaces and efficient search capabilities. For a low-cost scientific data analysis infrastructure, all of the above components are required at the same time. Our approach is to wed state-of-the-art database and Internet technologies to novel algorithmic developments in order to enable fast and seamless access to collaborative science archives. The framework serves both the earth and space sciences communities.

SODA—Self-Service Online Digital Archive for Unloved Scientific Data

By Rex Sanders[1]

[1]U.S. Geological Survey, Santa Cruz, Calif.

SODA (Self-service Online Digital Archive) is a project under development at the U.S. Geological Survey (USGS) for archiving "unloved" scientific data.

Background

USGS collects many thousands of gigabytes of new data annually. Many data types have well-defined processing and archiving paths, but many do not—our so-called "unloved data." Unloved data types usually fall into two classes: data types that have not traditionally shown national significance (for example, marine sediment analyses), and data types created from new technology and research (for example, airborne and land-based light detection and ranging (LiDAR) surveys). Unloved data are difficult to find, difficult to access, and often vanish completely upon the retirement or departure of key scientists and technicians.

Scientists with the best intentions frequently fail to archive their data well enough. One scientist carefully created and labeled three copies of digital core photos on a total of 90 CD-R discs. Three years later, none of these disks were readable because the label adhesive had corrupted the data layer. As a nonprofessional archivist, he did not know the risk associated with using sticky labels on CD-R discs.

Goal and Use Cases

The SODA project wants to make archiving unloved scientific data easier than having to burn another CD-R disc.

We are building our system around two use cases—archiving data and finding data.

To archive data:

1. Point your Web browser to "soda.usgs.gov," and click the "Submit Data" button.
2. Fill out a form with data type, format, minimum metadata (or more), and select a public release policy.
3. Upload your data.
4. Get a "permanent" link pointing to the data and metadata, with initial internal-only access.
5. Archivist reviews the data, metadata, and release policy.
6. If review passes, archivist enables public access to data and metadata according to release policy. If review fails, archivist contacts you for corrections.

To find data:

1. Point your Web browser to "soda.usgs.gov," and click the "Find Data" button.
2. Fill out form to search for data using any metadata fields, including geographic region, data type, or author.
3. Get links to download the data and metadata directly to your desktop.

Design Features

SODA will have other features, including:

- Design for users—Using SODA will be easier than burning another CD-R;
- Design for reuse by other Web-based tools, including ArcGIS, Geospatial One-Stop, MRIB (http://mrib.usgs.gov), InfoBank (http://octopus.wr.usgs.gov/infobank), and so on;
- Design for longevity using open standards and simple technology;
- Design to scale to large numbers of very large files;
- Separate searches for USGS-only and public data and metadata; and

- Archivists can easily add new data types, data formats, metadata forms, and release policies.

Benefits

Some of the anticipated benefits of SODA include the following:

- Improved access to data and metadata by USGS scientists and the public;
- Scientists and technicians will be able to archive data easily and immediately;
- Improved data preservation;
- Reduced data rescue;
- Scientists will be able to cite permanent links in published papers; and
- Scientists and technicians will not need to respond to data requests.

SODA is intended to be the scientific data archive of last resort, dependent on the cooperation of overworked scientists and technicians to keep valuable data from being lost forever. As such, we cannot require very much effort to archive the data; the process must be simple and self-explanatory. Our reduced metadata and approval requirements disappoint many mainstream data archivists, but capturing more data with some metadata is better than capturing no data.

SODA is not intended to replace any other USGS data-archiving mechanism, including open-file reports, data series reports, or online databases such as NWIS (http://waterdata.usgs.gov/nwis).

Technical Design

SODA technical design is based on the following principles:

- SODA servers will be organizationally and geographically distributed, and locally run, but centrally searchable.
- SODA servers will be easy to set up and run by local information technology (IT) personnel, using inexpensive hardware and software.
- SODA is much more than hardware and software—SODA will include processes and procedures to ensure the longevity of the data archive.
- A SODA "cookbook" will enable IT personnel to set up and run a SODA server with little outside support.
- A central SODA server will enable searches and retrieval across the distributed SODA servers.

Current Status of the SODA Project

SODA has been under development with minimal funding since early 2006. After surveying commercial and open-source projects, we concluded that writing our own software and designing our own system would best meet our needs. We have a core developer team with three members, and an e-mail-based advisory group with about 45 members, all working at the USGS.

We have an initial, nonarchival prototype running. We are using the prototype to work out many technical, user-interface, process, and procedural issues.

We anticipate release of our first production SODA server by the end of 2007. A few months after that release, we anticipate release of the SODA "cookbook" to enable other sites to set up and run local SODA servers. Development of the central search system and other SODA features is unscheduled, dependent on acquisition of further resources.

We will consider joint development of SODA with non-USGS partners.

From Flight Data to Knowledge of the Atmosphere's Chemistry: An Example from INTEX-NA

By V.E. Delnore[1], J.H. Crawford[1], A.A. Aknan[2], and C.C. Brown[2]

[1]Science Directorate, National Aeronautics and Space Administration Langley Research Center, Hampton, Va.

[2]Science Systems and Applications Incorporated (SSAI), Hampton, Va.

Abstract and Concept

This paper describes how proper management of airborne data contributes to an increase in knowledge of the chemistry of Earth's atmosphere. Before the mission, historical data are used by mission scientists to design the airborne campaign. During the field phase, flight planners use new data from each flight to lay out following flights. Post mission, the newly acquired data are archived for maximum accessibility to a wide range of interested parties, to include educators and the interested public, as well as the mission participants. Throughout all phases of the mission, data specialists maintain close cooperation with mission scientists, flight planners, principal investigators, and potential users to ensure that the data become useful knowledge. This paper focuses on this process as successfully implemented for INTEX-NA (defined below), and cites an example benefiting the National Oceanic and Atmospheric Administration (NOAA).

The Field Mission

INTEX (the Intercontinental Chemical Transport Experiment) is an atmospheric chemistry field mission seeking to understand the transport and behavior of gases and aerosols on transcontinental and intercontinental scales and their impact on air quality and climate. INTEX is a National Aeronautics and Space Administration (NASA) contribution to a larger

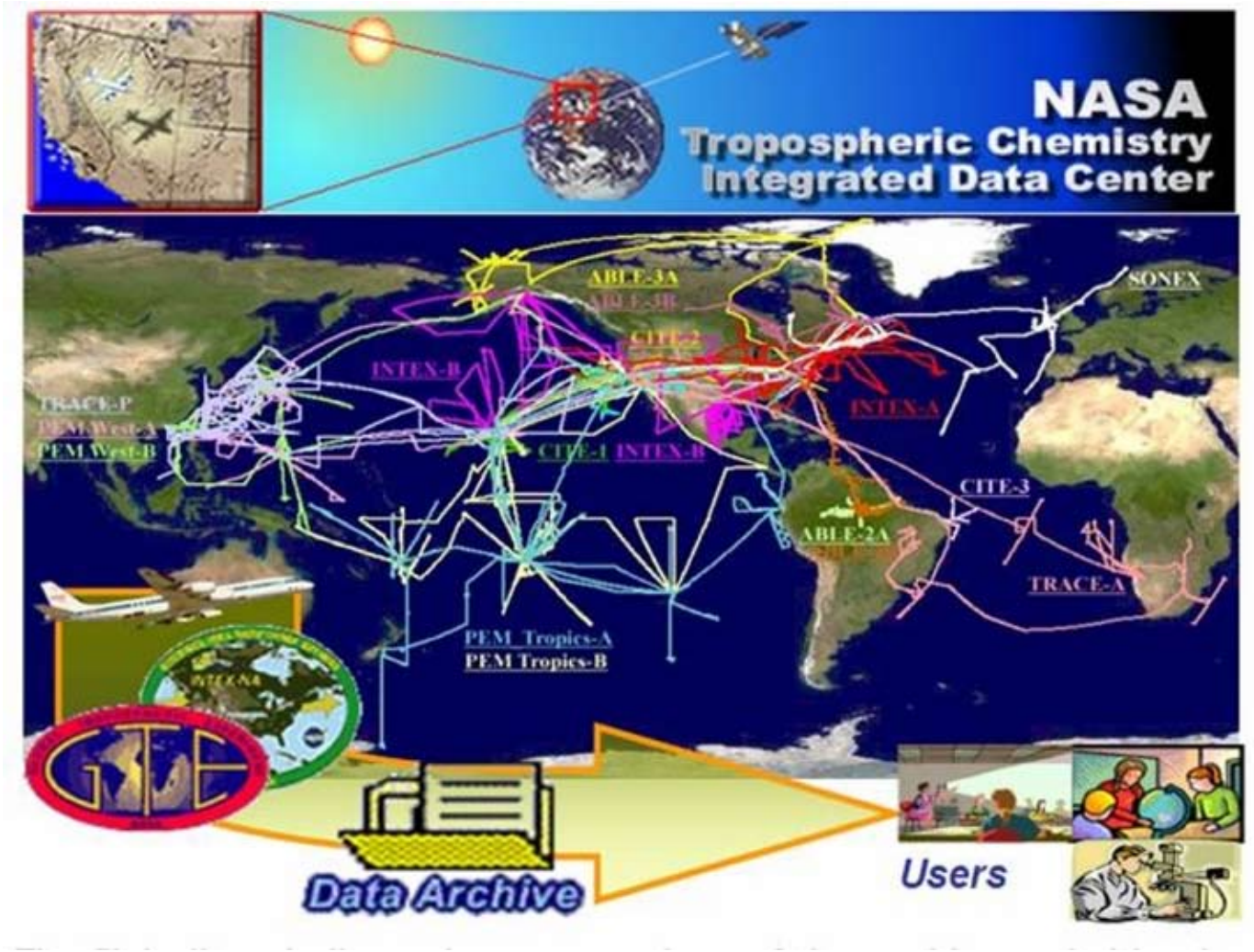

Figure 1. The flight lines indicate the many regions of the world sampled during the heritage Global Tropospheric Experiment (GTE) and the recent INTEX campaigns. The broad arrow represents the flow of data through the Langley Research Center's archive facility and, ultimately, to scientific and public users.

global effort—International Consortium for Atmospheric Research on Transport and Transformation (ICARTT). A particular focus of the 2004 phase of INTEX ("INTEX-NA," for North America) is to quantify and characterize the inflow and outflow of pollution over that continent. The red lines in figure 1 below show the track lines of the INTEX-NA flights, designed to sample a wide variety of chemical species over and around the continental U.S.

Data Collection and Archiving

Data management for the final phase of INTEX-NA has been completed. The Web-based Principal Investigator Data Registration system and Web-based data archive proved to be invaluable to mission scientists in making revisions to their data both in the field and after returning home, as they refined instrument calibrations and algorithms. NASA-sponsored aircraft serving as instrument platforms for the hundreds of datasets included the DC-8, SkyResearch J-31, and Proteus. The archive also hosts datasets from ground stations, satellites, national lightning-sensor and ozonesondes networks, and air-mass trajectory and model calculations. These data are now publicly available on the Web at http://www-air.larc.nasa.gov/missions/intexna/intexna.htm. The data have also been incorporated into a plotting tool at http://www-air.larc.nasa.gov/cgi-bin/2Dplotter and into a digital atlas tool at http://www-air.larc.nasa.gov/cgi-bin/datlas. The latter application displays altitude profiles (as well as statistical summaries) of chemical species measured on the DC-8 for the entire INTEX-NA mission. The archive also hosts merge products standardized to a common time base for many species. Sixty-second merges were created during the field phase of the mission and used for planning subsequent flights. For the preliminary and final phases of the mission, merges were created for a very wide variety of chemical species as needed by mission scientists.

Links to the INTEX-NA mission summaries, data archive, and analysis tools can be accessed at http://www-air.larc.nasa.gov/missions/intexna/intexna.htm. Among these is the Satellite Predictor Tool at http://www-air.larc.nasa.gov/tools/predict.htm, which yields the subtracks of an orbiting spacecraft, along with the footprint paths of sensors selected by the user. This was particularly useful for planning flights to support validation of satellite instruments.

Data Management and Formatting

To support INTEX-NA, the Langley Research Center's Research, Education, and Applications Solution Network Cooperative Agreement (REASoN-CAN) team worked very closely with NOAA's Aeronomy Lab, also an ICARTT mission partner, in developing the ICARTT data protocol and formats. The Langley group then took the lead in developing a Web-based data scanning and archiving system consisting of several tools that worked together to achieve full automation. Mission scientists received instant feedbacks that proved very useful for uploading their data files. The archiving system was then expanded to archive data for many additional platforms sampled during the ICARTT mission. Future refinements of this system will be undertaken with the cooperation of international data users' working groups.

Enabling Science: An Example of Moving Data to Knowledge

Making use of data merge and overlay tools developed by the Langley REASoN team, a NOAA researcher, Owen Cooper, combined data from the INTEX ozonesonde network (IONS), the National Lightning Detection Network, and Measurement of Ozone on Airbus In-service Aircraft (MOZAIC) observations from commercial aircraft. These data were interpreted by Cooper with the FLEXPART dispersion model to establish the link between lightning emission of nitrogen oxides and the large increases in ozone observed over North America during the ICARTT experiment.

Acknowledgments

The REASoN-CAN Award—Through its REASoN-CAN grant, "Synergistic Data Support for Atmospheric Chemistry Field Campaigns," Langley Research Center provides data management to support flight planning during tropospheric chemistry field missions. After the field phase of a mission, Langley's researchers gather corroborating data from many sources and provide these data to the scientific and educational community in forms that encourage further research and analysis, and to support NASA's outreach programs. The award is provided through NASA's Program Executive for Data Systems, Science Mission Directorate.

Patch Reef Analysis Using LiDAR-Derived Metrics at Biscayne National Park, Florida

By Monica Palaseanu-Lovejoy[1], John Brock[2], Amar Nayegandhi[1], and Wayne Wright[3]

[1]ETI Professionals, St. Petersburg, Fla.

[2]U.S. Geological Survey, St. Petersburg, Fla.

[3]National Aeronautics and Space Administration Wallops Flight Facility, Goddard Space Flight Center, Wallops Island, Va.

This study uses submerged topographic data and coral reef rugosity estimates derived from National Aeronautics and Space Administration's (NASA's) Experimental Advanced Airborne Research LiDAR (EAARL) for Biscayne National Park, Florida (Brock and others, 2004, 2006). The purpose is to evaluate the capability of NASA EAARL to describe patch reef variability and habitat complexity. Over 1,000 patch reefs were analyzed using mean neighborhood statistics and reclassification of LiDAR and slope data. The area of each reef was divided into the following categories: (1) base or reef footprint, (2) side of the reef, and (3) top of the reef. Different mean metrics were derived for each reef category from the submerged topography and rugosity data. Scatter plots of reef depth versus rugosity, relative relief, shape index, perimeter, area and volume, respectively, suggested that the multivariate data is bimodal. A mixture of two log-normal distributions suitably approximated the reef depth distribution. The depth at which the two log-normal distributions intersect was used to divide the patch reefs in shallow (less than 7.77 m) and deeper (7.77 to 14 m) reefs, respectively. The results showed that shallow patch reefs had a tendency to be bigger with a smaller relative relief than the deeper patch reefs. Topographic complexity, or rugosity, increased with depth for shallow reefs. In contrast, for deeper reefs, rugosity decreased with depth.

An independent component analysis was carried out on principal components derived from the patch reef metrics to determine if depth was the single most important factor to influence reef physical variability and habitat complexity. Principal components, although uncorrelated, are only partly independent (Hyvarinen and Oja, 2000). Two distinctly different independent components emerged from the analysis of seven principal components that described over 95 percent of data variability. We demonstrate that one independent component can be a function of patch reef rugosity while the other independent component is most likely a function of reef geometry and depth. These two independent components divide the patch reefs population in the following three depth classes: (1) from 2 to 6 m, (2) from 6 to 9.5 m, and (3) from 9.5 to 14 m, respectively. The deepest class correlates with the tail data not modeled by the log-normal mixture distribution.

Independent component analysis is more sensitive than simple multivariate analysis in assessing data variability. Multivariate analysis confirmed two major different popula-

tions—shallow and deeper reefs—with divergent rugosity correlations, but similar behavior of other reef metrics, such as perimeter, area, and volume. Independent component analysis suggests that three classes may be more appropriate to describe patch reef variability and habitat complexity in Biscayne National Park.

References Cited

Brock, J.C., Wright, C.W., Clayton, T.D., and Nayegandhi, A., 2004, LIDAR optical rugosity of coral reefs in Biscayne National Park, Florida: Coral Reefs 23, p. 48-59.

Brock, J.C., Wright, C.W., Kuffer, I.B., Hernandez, R., and Thompson, P., 2006, Airborne lidar sensing of massive stony coral colonies on patch reefs in the northern Florida reef tract: Remote Sensing of Environment, v. 104, p. 31-42.

Hyvarinen, A., and Oja, E., 2000, Independent component analysis: Algorithms and applications: Neural Networks, v. 13, nos. 4-5, p. 411- 430.

Internet GIServices for Homeland Security

By Ming-Hsiang Tsou[1], Tong Zhang[1], and John Kaiser[1]

[1]Department of Geography, San Diego State University, San Diego, Calif.

Homeland Security and Web GIS

This paper will illustrate the potentials of applying Internet Geographic Information Services (GIServices) to improve homeland security intelligence works. In the post-911 era, homeland security has been one of the primary missions for all levels of U.S. governments. The homeland security intelligence is of critical importance in preventing terrorist attacks, responding to the natural or human disasters, and recovering from the hazard damages. A Web-based spatial decision support system can combine dynamic geospatial information and Web mapping services from multiple Web servers located in Federal, State, and local agencies. The dynamic integration of Web-mapping services and information can provide more accurate and effective information for decisionmaking processes. Such information exchange is essential to pre-emergency planning, critical first-response actions, relief efforts, and community recover. Furthermore, such information can greatly enhance daily operations and cooperation among agencies in meeting homeland defense responsibilities.

The NASA REASoN Project Showcase

This paper will introduce a National Aeronautics and Space Administration (NASA) REASoN (Research, Education and Applications Solution Network) project, called "A Border Security Spatial Decision Support System," as a showcase of Internet GIS applications for homeland security tasks. This project is the collaboration between San Diego State University and the San Diego Sector of U.S. Border Patrol Agency (http://geoinfo.sdsu.edu/reason). This project seeks to establish

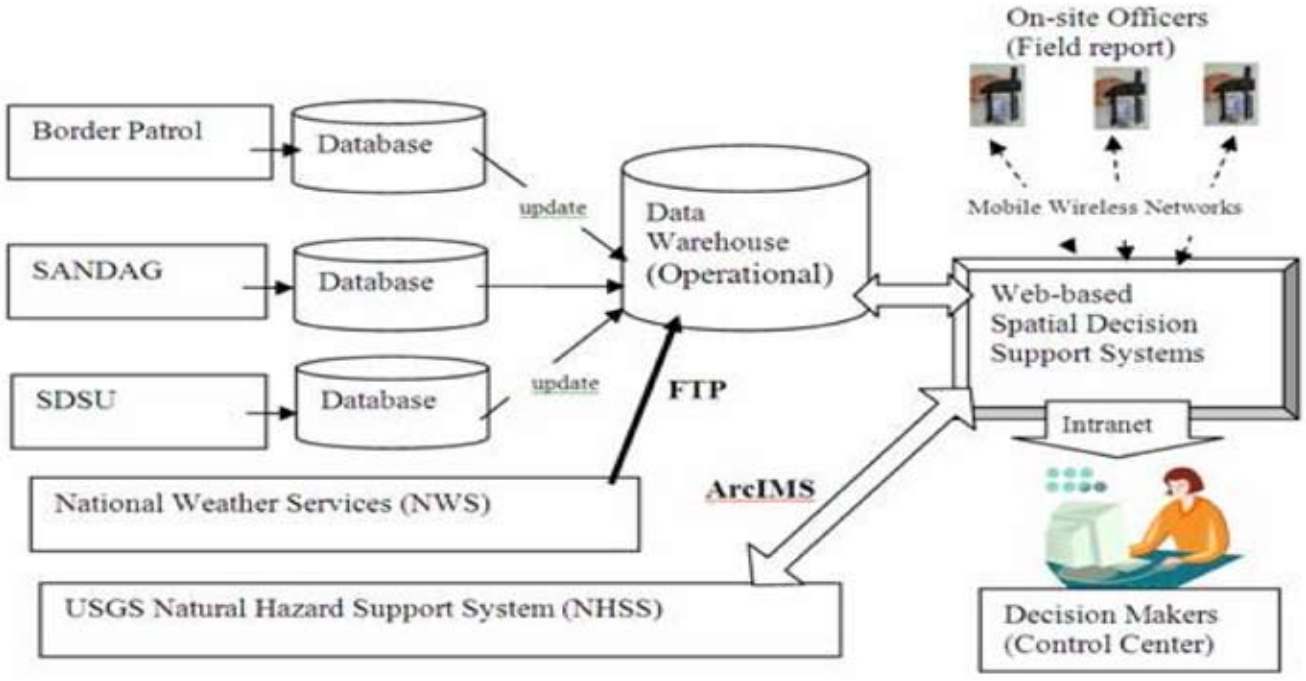

Figure 1. The interoperable database framework for Web-based spatial decision support systems.

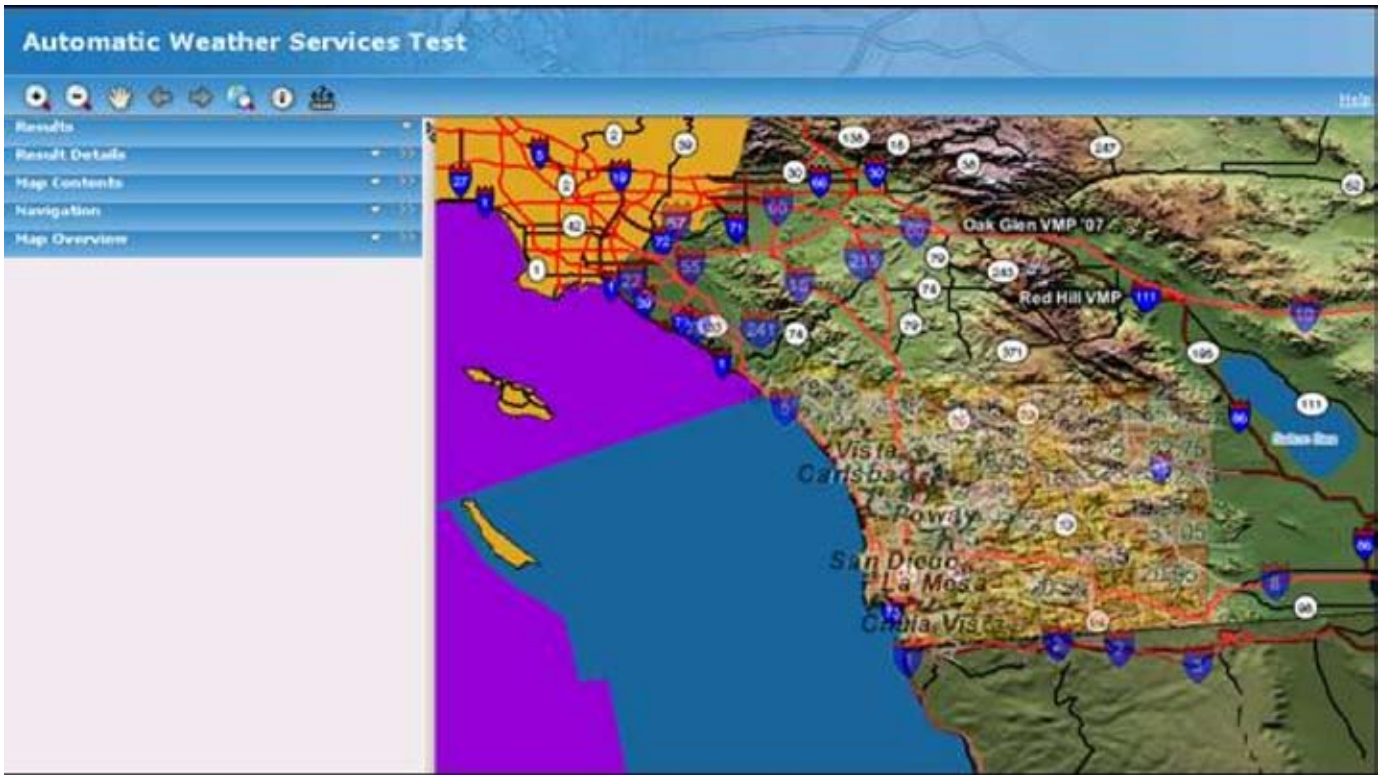

Figure 2. The integrated Internet GIServices (combining NWS and NHSS with local San Diego GIS layers).

and implement an integrated Web-mapping service that will allow rapid integration of geospatial data among participating agencies. By utilizing a standardized Web-mapping interface (Open Geospatial Consortium's Web Map Server Interfaces Implementation Specification) and popular vendor-based frameworks (Enviromental Systems Research Institute's ArcIMS services), individual participating agencies can implement their own data systems and services while maintaining an aggregated, system-wide interoperability through multiple data warehouses and Web-based decision support systems (fig. 1).

Real-Time Web Mapping from Multiple Resources

To test and evaluate the feasibility of the proposed framework, the REASoN project selected the National Weather Service's (NWS's) (http://www.nws.noaa.gov) and U.S. Geological Survey's (USGS's) Natural Hazard Support System (NHSS) (http://nhss.cr.usgs.gov) to combine with local San Diego Border Region GIS layers. Real-time or near-real-time data from NWS and NHSS can be rapidly fetched and distributed to the Web GIS application and be viewed by decisionmakers. Figure 2 illustrates an ArcIMS mapping service screenshot which combines NWS, NHSS, and local GIS data in the San Diego region to demonstrate the potentials of Web GIS-spatial decision support systems.

Discussion and Future Directions

Beyond GIS data and mapping services, Internet GIServices can support customizable spatial-analysis functions in the future. The newly developed ArcGIS server by ESRI or open-source GIS tools offers powerful GIS analysis tools at the server side. It is clear that desktop GIS tools are migrating to the Web platform. It is totally applicable to develop comprehensive GIS analysis online tools by either using commercial Application Development Framework or by adopting a lower level of programming models, such as Java-based applications. Another future direction of Internet GIServices for homeland security is the adoption of mobile GIS with wireless communication devices. Mobile GIS can provide critical geospatial information on the ground and can be sent to the Internet GIS server for data update and announcement in a timely manner. The servers can also respond to the inquiries made by the client devices to assist in-field actions.

To summarize, the needs of enhancing homeland security provide great opportunities for the future development of Internet GIServices. The recent progress of GIS technologies and Web services can facilitate the easy adoption of Internet GIServices in all major homeland security tasks. Besides the technology development of Internet GIServices, it is also important to promote the awareness of geospatial technology in various levels of decisionmaking and to facilitate the collaboration between GIS researchers and homeland security staff.

An Automated Parallel Computing System in the GEON Grid: Applications to Multiscale Crustal Deformation in the Western United States

By Mian Liu[1], Huai Zhang[1], Youqing Yang[1], and Qingsong Li[1]

[1]Department of Geological Sciences, University of Missouri—Columbia, Columbia, Mo.

In the past decade, supercomputing power has become available to most researchers in the form of affordable Beowulf clusters and other parallel computer platforms. However, to take full advantage of such computing power requires developing parallel algorithms and related codes, a task that is often too daunting for geoscience modelers whose main interest is in geosciences. As part of the Geosciences Network (GEON) effort, we have been developing an automated parallel computing system built on open-source algorithms and libraries. Users interact with this system by specifying the partial differential equations, solvers, and model-specific properties using a high-level modeling language in the input files. The system then automatically generates the finite element codes that can be run on distributed or shared-memory parallel machines. This system is dynamic and flexible, allowing users to address a large spectrum of problems in geosciences. We demonstrate this modeling system with a suite of geodynamic models that simulate multiscale crustal deformation in the western United States, ranging from timescale-dependent faulting in the San Andreas fault system to strain localization and active deformation in the western U.S. Cordillera. We show that this system may facilitate integration of high-performance computing with distributed data grids in the emerging geoscience cyberinfrastructures.

Disk-Based Gridding for Large Datasets

By Steven Zoraster[1]

[1]Austin Subsurface Modeling, Inc., Austin, Tex.

Abstract

As computer speed and memory increase, the amount of data to be processed grows even faster, requiring more sophisticated algorithms to run on the new hardware. Terrain- and subsurface-formation modeling have experienced this data explosion, along with a demand for larger output models. In this paper, I present a mathematically sophisticated, grid-based surface-modeling algorithm that makes little demand on computer random-access memory (RAM). The algorithm is competitive with other gridding algorithms on small- or medium-size datasets. For large problems, those with tens of millions of input data points and output grids with billions of nodes, this algorithm produces a solution where many others are just starting or have already crashed.

Introduction

The algorithm presented here, named Basin Gridding, is a disk-based, mathematical algorithm for gridding large datasets. The algorithm is an extension of an in-core, iterative, B-Spline algorithm (Lee and others, 1997). My implementation follows the workflow of the original algorithm, and retains the accuracy of that algorithm. Like that algorithm, Basin Gridding implements a one-directional, multilevel solution, working with coarse grids during early iterations and increasingly finer resolution grids during later iterations. The difference is that Basin Gridding requires minimal RAM. Because of the conservative use of RAM, Basin Gridding increases the size of datasets that can be modeled and the size of the grids that can be produced. The price paid for conservative use of RAM is reliance on disk-based storage and the associated time penalty. Central-processing unit CPU and wall-clock times for this algorithm are shown in table 1.

Previous Work

A working assumption for all two-dimensional gridding is that the data is sampled from an elevation function F, mapping R2 -> R1, and that the derived model approximates F to some degree of accuracy over R2. This problem has been studied for decades. It is possibly impossible to describe all types of gridding algorithms. Reviews can be found in papers and books by Foley and Hagen (1994), Franke and Nielson (1991), and Jone sand others (1986). A recent, general purpose algorithm for solving large gridding problems using matrix formulations and iterative techniques has been presented by Billings and others (2002).

A tradeoff rarely addressed for gridding algorithms is dependence on RAM and dependence on hard-disk memory.

Control Points	Grid Nodes (millions)	CPU (minutes)	Wall Clock (minutes)
89,000	.3	.2	.3
250,000	4.6	.4	.5
250,000	18.5	.5	1
5,700,000	8.3	5.7	11
5,700,000	133.0	17	78
30,000,000	380.0	50	205
145,000,000	4,876	170	600

Table 1. Algorithm central-processing unit (CPU) and wall-clock times.

In related domains, such as computational geometry, these tradeoffs are considered. For example, sweep-line algorithms for finding intersections between members of a set of line segments are both memory efficient and a natural way to organize computations. Fortune's (1989) sweep-line based, two-dimensional triangulation is an algorithm that could be used for surface modeling that would make relatively small demands on RAM. I do not know if it is being used anywhere for that reason.

For Basin Gridding, the criteria for algorithm selection was an ability to model large datasets and create large models in small amounts of RAM, without sacrificing mathematical properties such as smoothness, C2 continuity, and accuracy. I wanted, at least in theory, an algorithm that could solve a gridding problem of any size. Basin Gridding is the result. It is an extension of an in-core, hierarchical B-Spline gridding algorithm. By trading computational time for disk access, I created an algorithm with the good qualities of the in-core algorithm with reduced dependence on RAM.

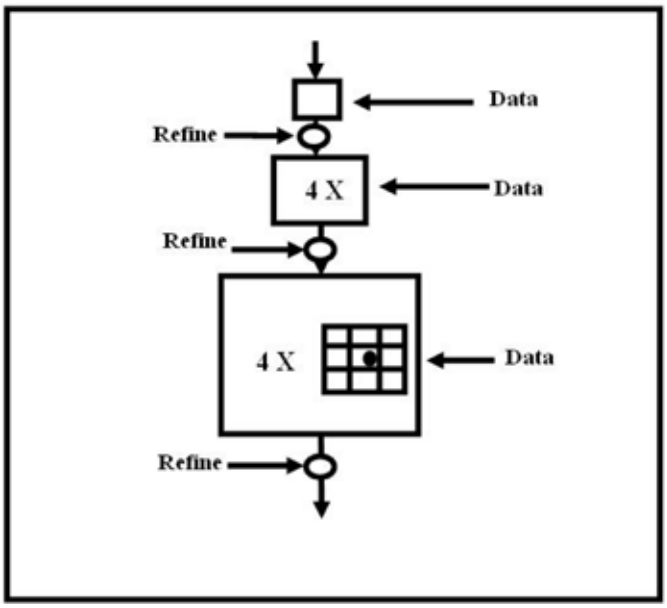

Figure 1. Algorithm general work flow showing embedded 4-x-4 grid.

Hierarchical B-Spline Gridding

In-memory algorithm

Creating a grid-based surface model with B-Splines requires the calculation of a lattice of control points over a grid at the same (delta x, delta y) resolution. A multigrid approach is often followed, starting with a coarse 4-x-4 grid that is refined at the end of every iteration until the target grid row and column spacing are achieved. Lattice values are derived so that the evaluation of a local set of control lattice values, used to weigh a set of third-order basis functions, produces an estimate of surface values in each grid cell. At the final grid resolution, this compact support avoids global influence from purely local data distributions or local surface roughness. In addition, B-Spline surfaces using third-order basis functions are C2 continuous, which promises high-quality interpolation between regions with different data densities.

This general hierarchical B-Spline workflow is shown in figure 1. Each square block represents an intermediate grid at some row and column spacing in the multigrid progression. Each refinement, represented by an oval, creates a new grid with twice as many rows and columns at one-half the spacing of the previous grid's rows and columns. Before updating, a feature of the refinement process is that the lattice values of the new finer grid represent the same surface model as the grid that was refined. So refinement supplies a more localized way to change lattice weights to better fit the available data at small cost.

As shown in figure 1, updating for each data point involves changing local lattice weights stored at the nodes of a 4-row-by-4-columns subset of the entire grid centered at the cell holding the data point. Each individual update makes the implicit surface better fit that data point's z value. For sparse data relative to grid row and column spacing, this algorithm interpolates control point z values exactly. Even at the final, densest grid resolution, however, there may be multiple control points within two grid increments of many grid nodes; therefore, the final grid will seldom interpolate all the input data exactly. Instead it will pass a smooth, representative surface model through such dense collections of data.

Basin Gridding Algorithm

The local 4-x-4 updating of lattice values is the algorithm feature that allows it to store most columns of the current grid model on disk during each iteration. For any one control point, updating modifies nodes in only four adjacent grid columns. Grid columns to the left of these four, and columns to their right, might as well be kept on disk. If the data is sorted on its x coordinate, then lattice updates will flow across the grid from left to right. As each new control point is reached, the four grid columns of interest will be either the four currently in memory, or a set of four adjacent columns to their right. This local focus, combined with sorting of the input data, breaks the RAM constraint. The focus on a few columns is indicated by the narrow rectangles inserted in each grid block in figure 2. These rectangles represent grid columns that are in memory at any one time. These processing rectangles move across the grid from left to right. Columns of the grid to the left of the window have been updated and can be written to disk. Columns to the right are still on disk, not yet updated, ready to be read into memory.

The "moving window" scheme discussed above is also applied to the "refinement" step. Refinement also proceeds from left to right. At any instant, a few columns of the grid at

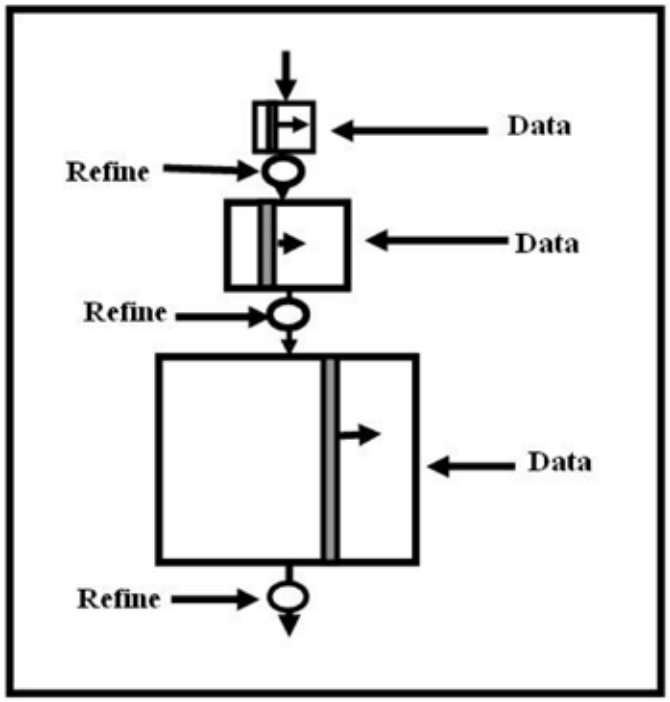

Figure 2. Disk-based work flow.

the "input" resolution will be in memory while being used to generate higher density columns of the output grid. The output grid columns have twice as many nodes as the input columns. This is not a problem because only three input columns need to be read into memory to create two output columns. Again, the output columns are written to disk after creation, the refinement process moves to the right, and another input grid column is read into memory.

In fact, the description above simplifies algorithm data structures and mathematical operations. Because multiple data points may influence the updating of a single grid node, the original in-core algorithm performs the summation of update information in two auxiliary grids, and only at the end of an iteration does it update the grid lattice values. That means that Basin Gridding must keep four columns from three grids in memory at one time. The actual updating of the lattice values in a left-hand column of the lattice grid is computed from the left-hand column of the auxiliary grids, just as the examination of a new control point reveals that one or more new (right-side) columns of the control lattice need to be read from disk, and left-side lattice columns need to be updated and written to disk. New columns for the auxiliary grids are created as needed and node values in those two columns start as zeros. Left-hand columns of the auxiliary grids that are no longer needed are simply deleted from RAM.

Basin Gridding incorporates other enhancements. Grid refinement is a distinct operation. That means there is a lattice grid read from and written to disk twice for every iteration level. I avoid this by interleaving grid refinement with imposing data on the grid. As grid columns leave the processing window—as the window moves right—the algorithm keeps two extra left-side columns in memory, does refinement in RAM, and then writes two new columns of the refined grid lattice. The result of all the enhancements discussed above is a gridding algorithm constrained only by free space on the computer's hard disks.

Discussion and Conclusion

An algorithm has been presented for creating rectangular grid-based surface models using disk storage to increase the size of the datasets that can be modeled and the size of the models that can be produced. The method applies multigrid B-Spline gridding in a novel way such that internal storage space is required for only a few columns of the target grid and one x, y, z triple of data at any one instance. The general work-flow and mathematical details for memory-based hierarchical B-Spline gridding are honored, so there is no loss of accuracy from this use of disk storage.

References Cited

Billings S.D., Beatson, R.K., and Newsam, G.N., 2002, Interpolation of geophysical data using continuous global surfaces: Geophysics, v. 67, p. 1810-1818.

Foley, T., and Hagen, H., 1994, Advances in scattered data interpolation: Surveys on Mathematics for Industry, v. 4, p. 71-84.

Fortune, S., 1987, A sweepline algorithm for Voronoi diagrams: Algorithmica, v. 2, p. 153-174.

Franke, R., and Nielson, G.M., 1991, Scattered data interpolation of large sets of scattered data: International Journal of Numerical Methods in Engineering, v. 15, p. 1,691-1,704.

Jones, T.A., Hamilton, D.E., and Johnson, C.R.,1986, Contouring geologic surfaces with the computer, *in* Computer methods in the geosciences: New York, Van Nostrand Reinhold, 315 p.

Lee, S., Wolberg, G., and Shin, S.Y., 1997, Scattered data interpolation with multilevel B-Splines: IEEE Transactions on Visualization and Computer Graphics, v. 3, p. 228–244.

Geospatial Cyberinfrastructure Solution: Open Source or COTS?

By Tong Zhang[1]

[1]Department of Geography, San Diego State University, San Diego, Calif.

Internet GIServices and Geospatial Cyberinfrastructure

Internet-based Geographic Information Servcies (GIServices) are definitely important components within the geospatial cyberinfrastructure. The goal of providing interoperable, intelligent, and scalable Internet GIServices is consistent with the underlying driving force of building the geospatial cyberinfrastructure; therefore, the research on building the well-performed geospatial cyberinfrastructure will pose challenges on the enhancement of current Internet GIS technologies. Internet-based GIService solutions are primarily two strategies—open source and commercial off-the-shelf (COTS). From both open-source and commercial-off-the-shelf development contexts, there are advantages and disadvantages demonstrated in the development of geospatial cyberinfrastructure. The methodology to integrate the advantages of two solutions while downplaying the downsides should be an interesting research topic.

The Geospatial Cyberinfrastructure Vision

The National Science Foundation is leading the research toward the cyberinfrastructure, which is an unprecedented initiative to construct underlying and supporting facilities in order to advance scientific research, decisionmaking, and collaboration by integrating any available and useful resources. The cyberinfrastructure initiative can be implemented in Geography and Geosciences domains as well. The so-called geospatial cyberinfrastructure is a vision that will be used as the blueprint of the future geospatial technology in the next few decades. Similarly, the geospatial cyberinfrastructure should be able to allow the researchers to conduct high-end research supported by a variety of computing resources in a collaborative manner. The goals may involve facilitating geographic decisionmaking, promoting productivity, preserving the environment, securing the homeland, and advancing geosciences education.

Internet GIServices and Their Roles in the Geospatial Cyberinfrastructure

Internet GIServices evolved from the rudimentary, static Hypertext Markup Language (HTML), prerendered map images, to interactive Web mapping and toward distributed GIServices. Current Internet GIServices are rapidly moving toward providing high-performance complex data and analysis services. This transition makes Internet GIServices a perfect technology to create global-scale geospatial cyberinfrastructure. The geospatial cyberinfrastructure heavily relies on the advancement of Internet GIServices.

Open-Source Solutions and COTS Solutions

Both the open-source and COTS Internet GIS solutions can be traced back to the early stage of static Web mapping. The commercial products, such as Enviromental Systems Research Institute's (ESRI's) ArcIMS, have been widely used; the open-source packages have also gained popularity in the last few years.

A Review of Open-Source Strategy

Open-source software development aims to produce computer programs which are free of charge. More importantly, the source codes are open to the public. Internet GIServices has long been an active area of open-source development. Lately, open-source Internet GIServices have gained a tremendous rise in popularity. The open-source packages vary, from the Internet GIServer and geographic information systems (GIS) database management system, to GIS analysis tools. The Open Geospatial Consortium (OGC) specifications work as the de facto standards for the entire open-source GIS community.

A Review of COTS Strategy

Commercial Internet GIServices solutions still dominate the market. COTS Internet GIServices software has been extensively using the computing technologies—from Common Gateway Interface (CGI), ActiveX, Java, and dynamic HTML (DHTML), to Asynchronous Java Script and Extensible Markup Language (XML) (AJAX) and Web services. Commercial Internet solutions also allow flexible customization of interfaces and functions.

A Comparison

As for costs and development transparency, open-source solutions prevail. Open-source software becomes competitive for GIS vendors in terms of reliability and functionality; however, COTS software still has advantages over the open-source counterparts, especially from the perspective of end users. The reluctance of going open source can be contributed to the fact that open-source software is relatively more difficult to configure and customize. The technical supports, product liability, as well as complete documentation, explain the larger portions of market share by the COTS solution. Given these challenges, the recent creation of the Open Source Geospatial Foundation could be seen as a major event for joint efforts compared to previously uncoordinated endeavors. A comparison will be given for the most popular open-source and COTS products covering the aspects of system architectures, costs, platform neutral, Open Geospatial Consortium (OGC) compatibility, performance, spatial data formats, customization and development, installation, administration, and configuration, Web

mapping effectiveness, spatial analysis functions, security, and reliability.

The Integration of Internet GIS in the Geospatial Cyberinfrastructure

The development of the geospatial cyberinfrastructure requires the integration of Internet GIS technologies with other supporting components.

Current Practices

Current geospatial cyberinfrastructure practices are extensively using Internet GIS to deliver a variety of services, including metadata cataloging, Web mapping, GIS data processing, and visualization, as well as high-performance GeoComputation and geocollaboration. Both open-source and COTS software can be used in the current geospatial cyberinfrastructure development. As an example, take the Geosciences Network (GEON) ArcIMS is used as an online mapping tool.

Challenges and Discussion

Geospatial cyberinfrastructure involves many computing technologies that interact with Internet GIS. Integration should be a big concern. To seamlessly work with other components, Internet GIS may have to comply with international standards from both GIS and computer science communities. In addition, some emerging technologies, such as Grid computing, may give rise to further problems. The large-scale scientific research projects demand stable, reliable, secure, and scalable Internet GIServices. The complexity of geospatial problems and decisionmaking will continue to introduce obstacles in developing domain-specific applications.

A Hybrid Methodology

Simply put, the hybrid methodology selects the most suited solution for a given task within geospatial cyberinfrastructure. The COTS strategy has been leading the technical development of Web-mapping user interface, server administration, and configuration, as well as service security. The high reliability of commercial products makes them better candidates to deliver the most frequently used GIS functions, such as Web mapping, data management, and simple geovisualization and analysis. The standardized and well-defined tools can be developed by commercial packages. However, some specific applications and requirements (for example, performance, interoperability, security, and reliability) would demand that developers go into a lower level of development; in these instances, open-source solution could be selected.

Based on this methodology, a Web portal prototype design is presented with specific software tools (fig. 1). The design is feasible and applicable with all available technologies.

The Prospect

The future geospatial cyberinfrastructure will promote the development of Internet GIS by providing both opportunities and challenges. Neither open-source solutions nor COTS solutions will prevail; they will continue to compete with each other.

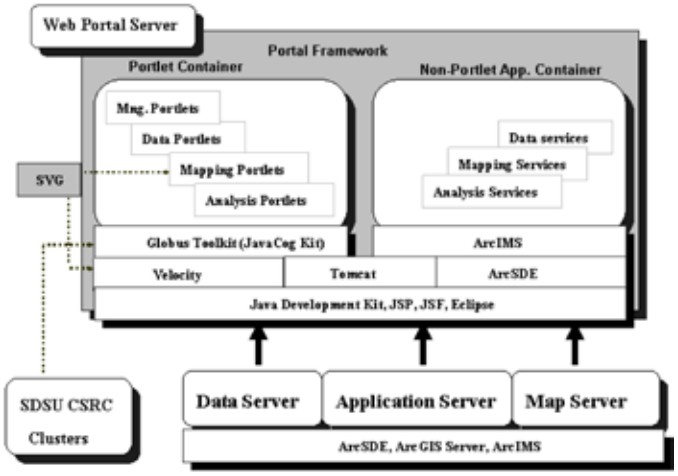

Figure 1. A Web portal prototype design based on the hybrid methodology.