USGS
science for a changing world

GFZ
Helmholtz Centre
POTSDAM

Geoinformatics 2008—Data to Knowledge

Proceedings

June 11–13
Potsdam, Germany

Scientific Investigations Report 2008–5172

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

Cover. Image of Europe produced by the Moderate Resolution Imaging Spectrometer (MODIS) aboard the National Aeronautics and Space Administration's (NASA's) Terra satellite. This image is part of the U.S. Geological Survey's Digital Image Gallery (http://eros.usgs.gov/imagegallery/).

Geoinformatics 2008—Data to Knowledge

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

Scientific Investigations Report 2008–5172

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: 2008

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

Abstracts in this volume written by U.S. Geological Survey authors have been reviewed and approved for publication by the USGS. Abstracts submitted by researchers from academia and from other State, Federal, and international agencies are published as part of these proceedings but do not necessarily reflect the Survey's policies and views.

Suggested citation:
Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., 2008, Proceedings, Geoinformatics 2008—Data to Knowledge, Potsdam, Germany, June 11–13, 2008: U.S. Geological Survey Scientific Investigations Report 2008–5172, 76 p.

Preface

The Geoinformatics 2008 Conference brought together many of the leading researchers from many countries to share new insights into the status of informatics-based research, with the common goal of working towards meeting geoscience-based societal challenges. The vision of the participants of a fully integrated geoscience cyberinfrastructure was highlighted at the meeting. Information technology research presentations emphasized the significant gains in the implementation of collaborative environments, portals, workflows, visualizations, semantics, and Web-based engines for visualization and integration. The application of many of these technologies was identified within the ongoing, multinational OneGeology project. The emphasis on developing collaborations between institutions, agencies, and countries was the most significant outcome of the conference. Collaboration is likely to be the intellectual driver of the future as geoscientists and computer scientists join to meet the global challenges of resource discovery and management, climate change, and natural disasters.

Conference Sponsorship:

Geoinformatics Division of the Geological Society of America, U.S. Geological Survey, National Science Foundation, German Research Center for Geosciences, American Geophysical Union, British Geological Survey, Federal Institute for Geosciences and Natural Resources [Germany], American Geophysical Union (Earth and Space Science Informatics), European Geosciences Union (Earth and Space Science Informatics), Electronic Geophysical Year (eGY), and the Committee on Earth Observation Satellites' Working Group on Information Systems and Services.

Organizing Committee:

Kristine Asch, Federal Institute for Geosciences and Natural Resources, Hannover, Germany; Sally Brady U.S. Geological Survey, Reston, Va.; Peter Fox, National Center for Atmospheric Research, Boulder, Colo.; Linda Gundersen, U.S. Geological Survey, Reston, Va.; Ian Jackson, British Geological Survey, Nottingham, United Kingdom; Jens Klump, German Research Center for Geosciences, Potsdam, Germany; Peter Löwe, German Research Center for Geosciences, Potsdam, Germany; Bernd Ritschel, German Research Center for Geosciences, Potsdam, Germany; Márta Nagy-Rothengass, European Commission, Brussels, Belgium; Andrea Schwerdtfeger, German Research Center for Geosciences, Potsdam, Germany; A. Krishna Sinha, Virginia Polytechnic Institute and State University, Blacksburg, Va.; Dogan Seber, San Diego Supercomputer Center, La Jolla, Calif.; and Lesley Wyborn, Geoscience Australia, Canberra, Australia.

Financial support was provided by the U.S. Geological Survey, National Science Foundation Division of Earth Science, and the German Research Center for Geosciences.

Conference support was provided by Geological Society of America (Nancy Carlson, Eric Nocerino).

Contents

Geoinformatics 2008—Data to Knowledge

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

Oral Session I

Global Challenges and the Challenges for Geoscience Informatics (Keynote)

By Ian Jackson[1]

[1]British Geological Survey, Keyworth, Nottingham, United Kingdom.

At a meeting in late April 2008, at the Royal Society in London, the Chief Scientific Advisor to the United Kingdom (UK) Government set out the key global policy challenges facing mankind. The following list is predictable but no less worrying: population growth, urbanization, poverty alleviation, technological change, food supply, energy demand, water resources, security, infectious diseases and, compounding them all, climate change. Some of the predictions are frightening: 60 percent of the world's population will live in urban environments in 2030 as compared with only 30 percent in 1950—and that is 60 percent of a very much larger global population than exited in 1950.

Where does geoscience come in? Well, if we (society) want clean water, a house that won't fall down, fuel for our cars, and a safe site to dispose of our waste, then we need to know about the rocks and processes beneath our feet. If our urban or rural homes are in a part of the world where earthquakes, volcanoes, and tsunamis are possibilities, then the need to understand those rocks can be a matter of life or death. Unfortunately for society, information about the rocks isn't always up-to-date, joined-up, understandable, and (sometimes) even available at all in all parts of the world. As a result, the essentials of life, such as clean water, building materials, and precautions against natural disasters, are that much more difficult to provide. We in the geosciences have some serious challenges to contemplate.

Whether we work in universities, geological surveys, other government agencies, or commerce, contributing to finding solutions to the challenges above has to be our overriding mission. So how can we in the geoscience informatics domain contribute best and what are the specific challenges that this creates for us if we want to raise our game? What are the areas in our science and our approaches to them that we need to change, improve, and do more of?

Taking a lead from the word "challenge", these areas and approaches can (hopefully memorably) be defined by 10 words beginning with the letter "C" (and here I must acknowledge and apologize for taking liberties with the discussions at a recent summit on geoinformatics in Rome, and with the English language!).

- Communicate—We need to make sure our science is available to all the stakeholders whether they have a degree in geoscience or computing or not.
- Content—We need to give a higher priority to data management because, arguably, the biggest problem we face is the lack of quality data, not the lack of computing power.
- Collaborate—We need to be able to share and integrate our information; therefore, we need interoperability and much improved taxonomies and semantic control.
- Coordinate—We need to manage our efforts in a coherent and cost-effective way with minimal duplication and fewer turf wars.
- Consistency—We need to develop standards and best practices and then comply with them so we can cut effort and add value to otherwise insular data and models.
- Chart—We need to audit and map the data resources we have and make them discoverable.
- Currency—We need to exploit new technologies, in particular Web technologies, but we also need to focus on the problem and not get seduced by the technology.
- Competencies—We need to ensure that we have the right skills and therefore we also need to ensure that we are providing the proper education and training to meet the challenges.
- Contribute—We need to proactively share our know-how with each other and listen to and understand the needs and contexts of those in the developing world.
- Change—We need to be prepared to be more agile and more flexible; we need to accept that change (as well as the increased pace of change) is a fact of life and that more than any time in the past we probably won't be able to do as we have always done.

This presentation will, drawing on national and international examples, explore these challenges and the issues they raise.

Answers to Earth System Science Questions—The Evolution of Informatics at the U.S. Geological Survey

By Linda C. Gundersen[1]

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

The U.S. Geological Survey's (USGS's) science strategy (U.S. Geological Survey, 2007) identified data integration as one of its crosscutting strategic science directions and states the following: "The USGS will use its information resources to create a more integrated and accessible environment for its vast resources of past and future data. It will invest in cyber-infrastructure, nurture and cultivate programs in Earth-system science informatics, and participate in efforts to build a global integrated science and computing platform."

USGS is constructing a service-oriented architecture (SOA) for all USGS data and science applications, which is a complex challenge for a 129-year-old institution that has been collecting earth science data since its inception. This effort requires operating on many aspects of architecture creation simultaneously while dealing with extensive legacy analog and digital data. Projects are underway that include everything from building a federated database warehouse, developing Web services for discovery of data, creating community-specific data models, and building integrated-earth-system scientific applications. The complex issues of governance, platforms, standards, and active engagement in international and national informatics efforts require the development of the SOA to be collaborative, iterative, and experimental. The following provides highlights of projects underway to create this service-oriented architecture for earth-system science.

Data Discovery

As a first step, USGS is engaged in evaluating the agency's data holdings and is creating a searchable, spatially enabled database tentatively named the "National Digital Catalog (NDC)" This catalog will provide discovery tools, metadata, a geospatial interface, and other information related to the nature of the materials or data. A Web-based pilot was recently developed to test hardware and software technology. This pilot, the Geospatial Management Information System (GMIS), provides one-stop access to different sources of USGS data, including detailed information on the thousands of USGS science projects being conducted around the world. The system allows the user to search information by topic and geographic area. Another part of the NDC under development is a catalog service for the physical material collections of the State geological surveys and Department of Interior bureaus that will contain searchable metadata on data and materials such as core, rock samples, well logs, engineering data, and maps.

Map Services

The USGS has created (and is constantly revising and updating) national, regional, and topical map-based data in a wide range of scales and resolutions. The National Map (http://nationalmap.gov/) provides imagery from a variety of sources, elevation and hydrographic data, geographic names, and land-cover data. The Mineral Resources On-Line Spatial Data service (http://mrdata.usgs.gov/) provides access to national databases and maps of mines, historical mining, mineral occurrences, geochemistry of rocks and sediments, lithology, geology, and geophysics of the United States. Both the National Map and the Mineral Resources On-Line Spatial Data service include map browsers, download sites, and Web services.

Integrated Science Applications

The PAGER (Prompt Assessment of Global Earthquakes for Response) system is an automated system developed by the USGS (http://earthquake.usgs.gov/eqcenter/pager/) to integrate multiple data streams; rapidly assess the number of people, cities, and regions exposed to severe shaking by an earthquake; and inform emergency responders, government agencies, and the media about the scope of the potential damage. PAGER monitors the National and Global Seismological Networks; retrieves shaking intensities reported by people in the epicenter region via USGS's online "Did You Feel It?" system; generates a site-specific ground-motion amplification map; and computes the population affected at each intensity level. Within 15 to 30 minutes, depending on the location and size of the earthquake, PAGER produces regional ground shaking estimates using the reported intensities, the site-specific ground-motion amplification map, and seismic-wave-attenuation equations that account for the variations of seismic shaking intensity with magnitude, distance, and depth. Final information is distributed in multiple formats and through multiple media, including Google's online map display services, automatic e-mail alerts, and community standard Extensible Mark-up Language (XML) format.

Two significant USGS collaborative efforts that bring global communities together to address global-scale societal issues are the Famine Early Warning Systems Network (FEWS NET) and the Delta Research and Global Observation Network (DRAGON). The FEWS NET grew out of an effort started by the U.S. Agency for International Development (USAID) in the wake of the devastating 1985 famine in Ethiopia. Today it is a modern network of multiple agencies supplying multiple data streams from both ground observations and satellite remote sensing. These data feed into a series of Web services and programs to produce products ranging from key vegetation indexes to sophisticated daily flood models. The network identifies and provides early warnings for famine in sub-Saharan Africa,

Afghanistan, Central America, and Haiti. The USGS Earth Resources Observation and Science (EROS) Data Center (USGS EDC) works in cooperation with USAID, the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and Chemonics International, Inc., to provide the data, information, and analyses needed to support the FEWS NET activity. NASA and NOAA are responsible for the collection and processing of satellite data that provide the spatial coverage and temporal frequency necessary for monitoring both vegetation condition and rainfall. Chemonics maintains a staff of field representatives responsible for key field observations and monitoring regional and country-specific conditions. USGS EDC provides end-to-end data management, processing, and analyses; GIS and remote-sensing technical support; crop and flood modeling; and long-term data archive and distribution services. The Africa Data Dissemination Service (ADDS) provides additional Web services.

DRAGON is an effort recently initiated by USGS to create a global science framework for comparing, integrating, and predicting the key drivers and management practices in large delta ecosystems. Large delta ecosystems provide the habitat for a broad diversity of flora and fauna as well as life-sustaining agriculture, commerce, and fisheries for hundreds of millions of people. The project will require an extensive effort to (1) make large volumes of ecological, hydrological, geological, and biogeochemical information interoperable; (2) create a common data and discovery portal; and (3) develop community tools and models through a global "community of practice" in delta system management. The pilot program partners the USGS with the Chinese Qingdao Institute for Marine Geology to develop the conceptual frameworks for the Lower Mississippi Valley simultaneously with those for the Huang He River. The resulting joint comparative model will be expanded to include river deltas in the Netherlands, Russia, Vietnam, and other countries with similar deltaic systems.

Reference Cited

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

Using Remote Sensing and GIS Techniques to Monitor the State of the Greenland Ice Sheet

By Meredith C. Payne[1] and Anne W. Nolin[2]

[1]College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oreg.

[2]Department of Geosciences, Oregon State University, Corvallis, Oreg.

As we strive to tailor hypotheses related to global climate change while assessing the possibility of an anthropogenic driver, it becomes crucial to constantly monitor the world's most climatically sensitive areas. Examples of such areas include glaciers and ice sheets whose record melting is impacting communities on a global scale. In some cases, regions that rely upon glacial water as a principle source of fresh water are witnessing the rapid dwindling of resources. In other cases, rising sea level, to which the melting of glacier ice contributes, is threatening low-lying communities. Unfortunately, as is the case with the Greenland ice sheet, many such areas are remote and dangerous, making spatially and temporally comprehensive field measurements cost prohibitive. Hence, we must rely on remotely sensed measurements from aircraft and satellites in order to fill in our knowledge gaps left by sparse field measurements.

The Multi-angle Imaging SpectroRadiometer (MISR) instrument (operational since 2000) is on board the Earth Observing System (EOS) satellite, Terra, which is in a sun-synchronous polar orbit. MISR is uniquely suited for studying the poles because of the continuous, overlapping coverage of data taken by its nine pushbroom cameras that are arrayed at fixed angles ranging from 0° to 70.5° (from nadir) and symmetric about the nadir camera. Each camera has four filters: red, green, and blue (in the visible), and a near-infrared (NIR) at 866 micrometer (μm) wavelength. The multi-angle views in conjunction with the 275-meter (m) resolution (available at the visible red wavelength on all nine cameras) can be used to compute a proxy of surface roughness of the observed target on a scale that is comparable to that of the Moderate Resolution Imaging Spectroradiometer (MODIS) 250-m data (Nolin and others, 2002). We have elected to use MISR's red-filtered C-cameras (60.0° fore and aft) in order to define and investigate a proxy for ice surface roughness based on forward and backward scattered radiation, which we call the Normalized Difference Angular Index (NDAI). We define NDAI for Greenland to be fore C-camera red-channel values subtracted from the aft C-camera red-channel values divided by their sum. Because the forward-viewing camera is seeing forward-scattering radiation while the aft camera sees backscattered radiation (the sun is to the south), the forward scattering is associated with generally smooth surfaces and backward scattering dominates when an observed surface is rough (Nolin and Payne, 2007). Therefore, in an NDAI image, values range from -1 to 1 and rougher surfaces appear brighter.

As a case study of the NDAI proxy measurement, we chose to study a region in western Greenland encompassing Jakobshavn Glacier (69.2° N, 50.2° W, 40 m elevation), which is one of the fastest moving glaciers in the world and one that drains a significant percentage of the Greenland ice sheet. Its area is greater than 9,200 square kilometers (km^2) (Rignot and Kanagaratnam, 2006). Our study site extends upglacier in the inland ice to Summit (72.6° N, 38.5° W, 3200 m elevation), which is the highest point on the Greenland ice sheet. We reviewed all available MISR images of the Greenland ice sheet for blocks 30 to 35, paths 8 to 10 during the 2000 to 2007 sunlit seasons across this transect. We determined 2004 to be the year when our study site was least obscured by clouds.

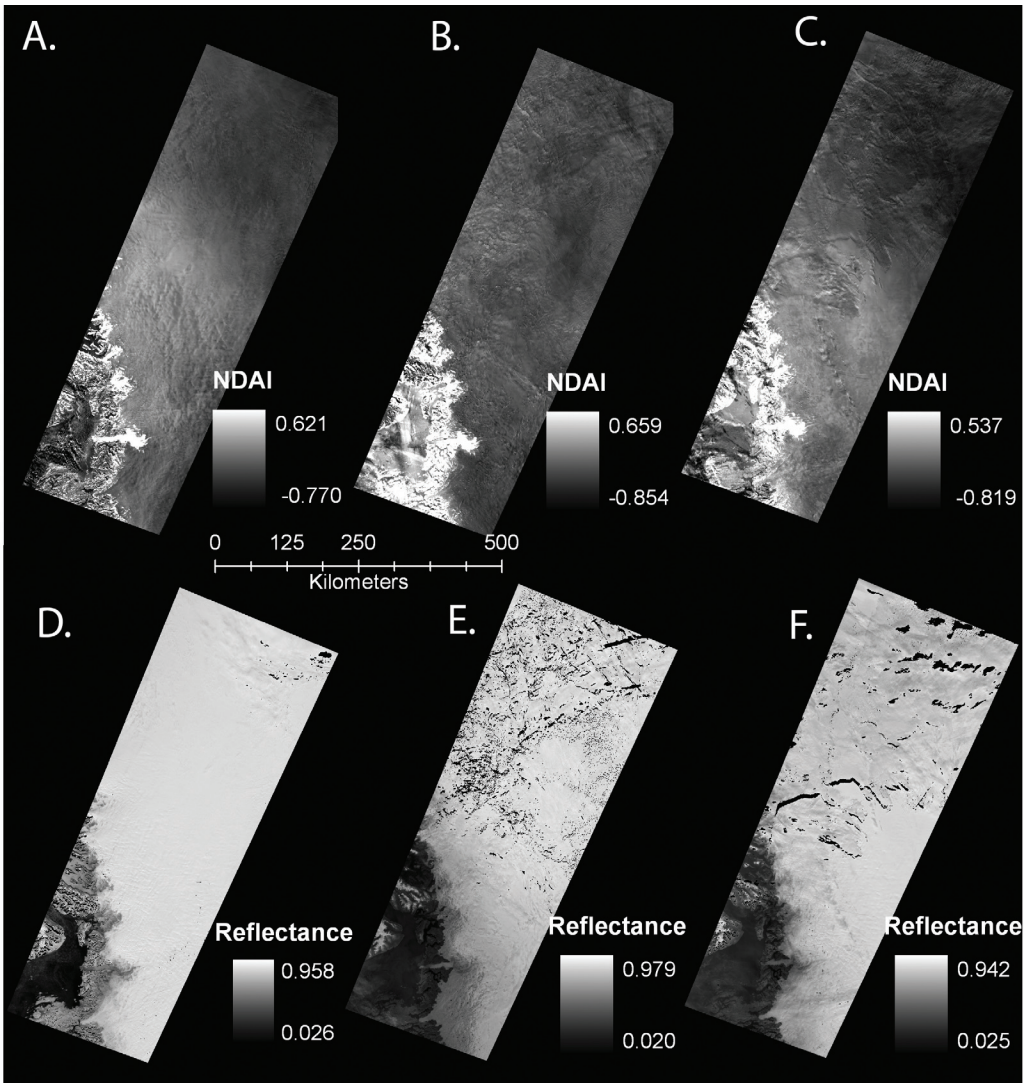


Figure 1. Images showing the ice-surface roughness of a portion of the Greenland ice sheet. Images A through C show the Normalized Difference Angular Index (NDAI), a proxy for ice-surface roughness, for the 2004 ablation season (April-May, June-July, and August-September composites, respectively). Higher NDAI values signify rougher surfaces. Images D through F are the composite reflectance images for April-May, June-July, and August-September, respectively. In images D through F, greater reflectance values signify smoother surfaces, which demonstrates the opposite relationship to the NDAI. Black pixels on the inland ice in images D through F indicate no data.

Nevertheless, completely cloud-free images over the entire region of study were impossible to come by. We investigated the application of the Radiometric Camera-by-camera Cloud Mask (RCCM) product, provided by the National Aeronautics and Space Administration's (NASA) Langley Atmospheric Science Data Center (ASDC) to the radiance images, but found the mask to be overly strict when distinguishing cloud pixels from ice pixels. Hence, using digital image processing along with geographic information system (GIS) techniques, we devised a method of creating composite images of NDAI and of top of the atmosphere (TOA) bidirectional reflectance factor (BRF) encompassing the early- (April and May), mid- (June and July), and late- (August and September) ablation season (fig. 1). These composite NDAI and reflectance images, along with their corresponding gradient images (mid-season composite minus early-season composite, and late-season composite minus mid-season composite) were examined to establish a pattern whereby the coastal regions are observed to grow progressively rougher throughout the ablation season. Ice surface roughness is intensified by (1) seasonal snow and ice in the ablation zone melting back to reveal underlying bedrock, (2) melt ponds forming upglacier in the wet-snow (slush) and percolation zones (as defined by Benson, 1960; Long and Drinkwater, 1994), (3) sastrugi (jagged erosional features caused by wind) morphology becoming more pronounced, and (4) melting and collapse of snow and ice bridges to reveal the highly irregular crevasse topography beneath. Although changes are not as dramatic upglacier towards Summit (melt ponds do not appear in the dry-snow zones), changes towards a rougher surface are observed mid-season in the percolation-zone, as NDAI pixels have greater values and become brighter compared with the early-season. As expected, there is little observed change in the near proximity of Summit over the sunlit season because it lies in the conjectured dry-snow zone as defined by Benson (1960) and Long and Drinkwater (1994). In the late-season images (August and September) after snowfall has resumed (especially over the wet-snow and percolation zones), NDAI values are observed to drop, but not fall as low as the early-season (April and May) values. Study of TOA reflectance images reveals the exact opposite relationship of pixel values throughout the time series: the pixels become darker (lower values) during mid-season and brighten

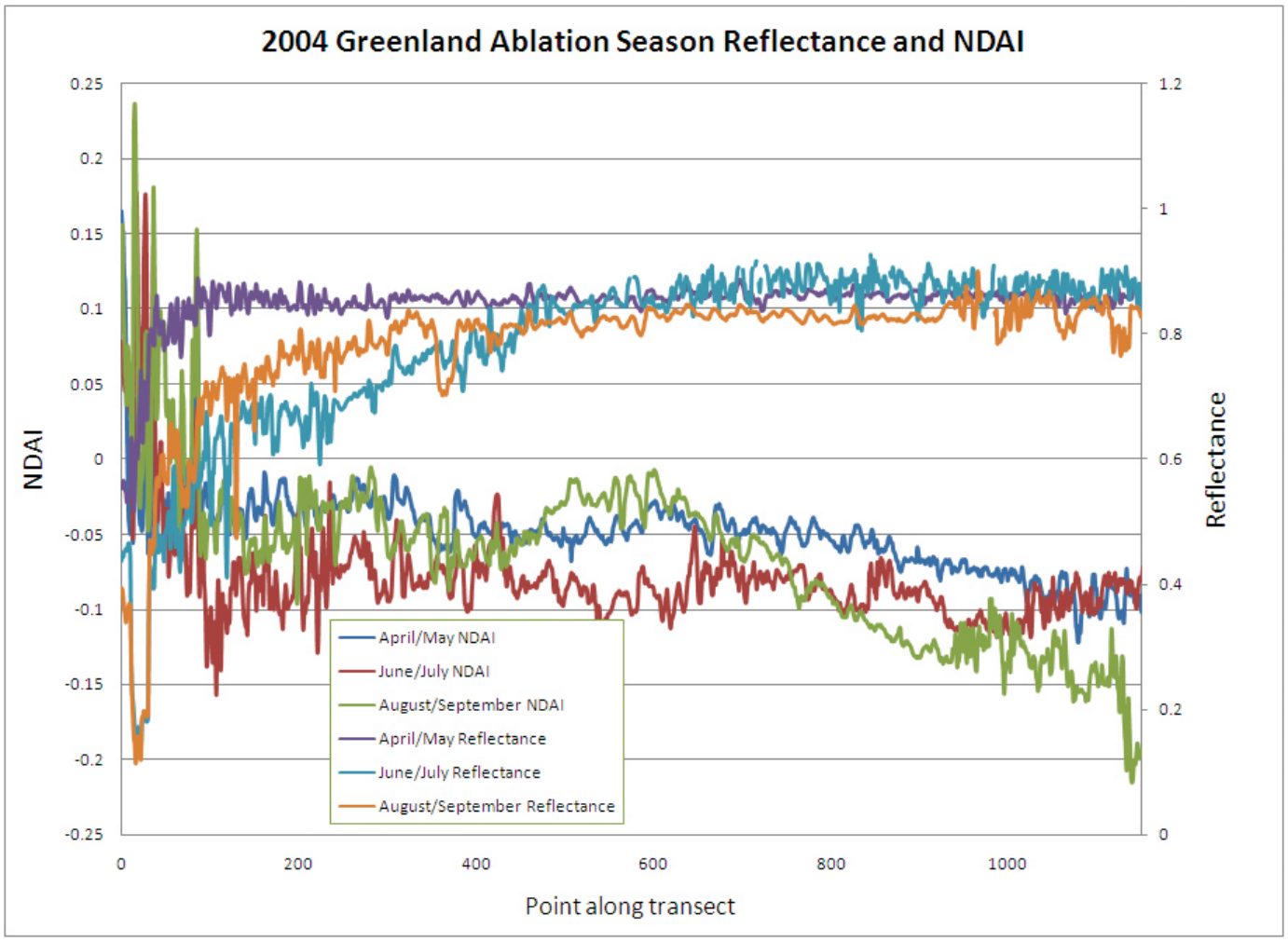


Figure 2. Graph showing Normalized Difference Angular Index (NDAI) values (left vertical axis) and reflectance values (right vertical axis) along a straight-line transect from Jakobshavn (left side of plot) to Summit (right side of plot) for the April-May, June-July, and August-September images as shown in figure 1. The contrasting relationship between NDAI and reflectance is noticeable, as is the trend of the ice surface growing rougher from spring to late summer, and then becoming smoother during late summer and early fall when snowfall recommences.

after fresh snowfall towards the end of the sunlit season. These relationships are illustrated in figure 2, in which NDAI and reflectance values from each of the three respective composite images are plotted along a straight transect from Jakobshavn to Summit.

We are encouraged enough by these results to proceed with production of similar NDAI composite images for the entire Greenland ice sheet for all years where enough low-cloud-percentage images are available. We hope to use these products to expand our analyses of the evolution of glacier zones during the operational lifetime of the MISR instrument to possibly include identification of glacier zones (such as the superimposed-ice zone) that are invisible to radar (Nolin and Payne, 2007). Furthermore, we believe that these products will enrich the already plentiful MISR dataset, which is publicly available for use in analyses.

References Cited

Benson, C.S., 1960, Stratigraphic studies in the snow and firn of the Greenland ice sheet: Pasadena, Calif., California Institute of Technology, unpublished Ph.D. dissertation, 213 p.

Long, D.L., and Drinkwater, M.R., 1994, Greenland observed at high resolution by the Seasat-A scatterometer: Journal of Glaciology, v. 40, p. 213–220.

Nolin, A.W., Fetterer, F.M., and Scambos, T.A., 2002, Surface roughness characterizations of sea ice and ice sheets—Case studies with MISR data: IEEE Transactions on Geoscience and Remote Sensing, v. 40, p. 1,605–1,615.

Nolin, A.W., and Payne, M.C., 2007, Classification of glacier zones in western Greenland using albedo and surface roughness from the Multi-angle Imaging SpectroRadiometer (MISR): Remote Sensing of Environment, v. 107, p. 264–275.

Rignot, E., and Kanagaratnam, P., 2006, Changes in the velocity structure of the Greenland ice sheet: Science, v. 311, p. 986–990.

Use of Remote Sensing Data in Searching for Hydrocarbon Deposits

By Yuri Baranov,[1] Sergey Kulapov,[1] Ekaterina Denisevich,[1] Maxim Vanyarkho,[1] and Denis Filatov[1]

[1]Space Information Lab, Gazprom Research Institute (VNIIGAZ), Moscow, Russia.

For an oil and gas company, providing the right to develop the planet's interior is preceded by efforts to identify economic expediency prior to licensing. Lack of studies of new oil- and gas-producing regions and limited access to existing information result in problems in expediting licensing for exploration and development of new regions.

Estimates of oil and gas in a new region can be performed using a complex analysis of widely available geological, geophysical, and remote-sensing data (Alexeev and others, 1988; Baranov and others, 1989; DeMers, 1999).

The advantage of using remote-sensing data is they provide unique information that cannot be retrieved by any other method. Remote-sensing is also a cost-efficient manner of gathering information while searching for liquid hydrocarbons, which in turn ensures efficient geological prospecting by means of innovative high-end technologies. Because remote-sensing methods (especially those that analyze landscapes) determine hydrocarbon geochemical anomalies observed in space images rather than the deposits themselves, they can be applied to any type of accumulations, including unstructured oil-and-gas fields. In addition to remote sensing techniques, other methods such as (1) analyzing spectral characteristics of images, (2) determining particular anomalies in the infrared spectrum, (3) lineament analysis, and (4) photogrammetry are also employed (Baranov and others, 2003).

In this paper, we show that the use of remote-sensing data substantially reduces the costs of geological prospecting, in particular because of the localization of potentially productive areas. We also demonstrate remote sensing to be extremely powerful from the standpoint of ecological monitoring and industrial safety of licensed areas, development objects, and hydrocarbon transport.

References Cited

Alexeev, A.S., Pyatkin, V.P., and Dement'ev, V.N., 1988, Automatic image processing of Siberian ecoterritories: Novosibirsk, Russia, Nauka, 224 p.

Baranov, Y.B., Grushin, R.V., Kruchkova, T.A., Baranova, L., and Feygin, A.E., 2003, Aerospace methods for monitoring of cryopedology conditions of northern oil-and-gas deposits of YANAO, *in* Proceedings, Gazprom—Maintenance of safety of infrastructure objects on permafrost soil territories, Tyumen, Russia, 2002: Moscow, Gazprom, p. 19–24.

Baranov, Y.B., Sokolovskiy, A.K., and Fedchuk, V., 1989, On some issues of regional geology of Precambrian of Aldan, *in* Proceedings, Russian Conference on Methods of Remote Sensing Data and Space Information Processing: Ryazan, Russia, p. 81–82.

DeMers, M., 1999, Geographical information systems—The basics: Moscow, Russia, Data+, 490 p.

Geoinformatics Mapping of Renewable Energy Resources and Systems in the Philippines

By Carlos Miniano Pascual,[1] Phebe Marcos Pasion,[2] and Irma Pascual Acebedo[3]

[1]Department of Agricultural Engineering, Mariano Marcos State University, Batac, Philippines.

[2]Management Information Services, Mariano Marcos State University, Batac, Philippines.

[3]Affiliated Non-Conventional Energy Center, Mariano Marcos State University, Batac, Philippines.

Accurate assessment of renewable energy resource data (such as for solar-, water-, wind-, and biomass-generated power) are important in order to (1) assess the availability of such resources, (2) mitigate global climate change, and (3) determine the size, cost, and life cycle of renewable energy systems technologies. Knowledge of the spatial distribution of the renewable energy resources allows for a more cost-effective design and operation of such systems. The goal of our study is to develop a renewable energy resource assessment for the Philippines that incorporates and builds upon the current understanding of the spatial distribution of each resource.

In order to assess solar energy, we used a high-resolution, global-satellite-derived, cloud-cover database for creating a climatological solar radiation model. In the case of hydrologic resources, the total volume can be specified according to the flow rate, and the effective elevation (head) can be measured through use of a digital elevation model (DEM).The availablility of the wind-power resource is defined in terms of the wind-power-density value, which is expressed in watts

per square meter. This value is based on wind speed and air density. To estimate biomass resources, some conservative assumptions were made in order to calculate a practical and reliable estimate; for example, outputs of straw and stalks, which were left over after a harvest, were calculated based on crop outputs and the ratio of grain production to stalk mass. These assumptions were related to the type of processing done for a particular commodity by researchers and planners. Other geographic data from land remote-sensing satellites, digital land-use and boundary maps, as well as hydrometeorological data were gathered (downloaded), converted, and compiled as input databases and base maps. ArcView geographic information system (GIS) software was used for the query-based spatial-data analysis. Visual Basic 6 programming was used to develop graphic user-interface programming to compile the georeferenced input data, as well as to create the graphic user interface that linked relational databases to GIS query modules. The long-range energy alternative planning system (LEAP) program was used to account for how renewable energy is consumed, converted, and produced in a given region or economy under a range of alternative assumptions on population, economic development, technology, price, and so on. An estimate of greenhouse-gas emissions will also be presented in order to quantify its mitigation effect on climate change. This assessment provides data that may be helpful to researchers, planners, developers, and investors in establishing successful commercial renewable energy technologies that can be adapted to mitigate climate change in the Philippines. A geoinformatics-based decision-support system was developed to build wealth of georeferenced data and information on renewable energy resources (sun-, water-, wind-, and biomass-generated) for policy research and development on energy resources. Georeferenced databases and thematic maps are major outputs that show various indicators of assessment, monitoring and evaluation, and efficiency that are useful for energy research, planning, and policy options. Such activities are necessary for both intermediate and long-term energy development plans for the country. The use of satellite remote sensing, geographic information systems (GIS), and global positioning systems (GPS) are associated with renewable energy resource management. Additionally, geoinformatics-based mapping tools and statistical analysis are required to share information through the World Wide Web. We present a graphical user interface using ArcView GIS, which includes a mapping system architecture to share information about renewable energy resources.

The combined use of satellite remote sensing, GISs, and GPSs with graphic user programming language has proven to be a very valuable and indispensable geoinformatics tool for gathering, organizing, retrieving, and storing georeferenced data that will be used for subsequent retrievals and analyses. Efforts to create a cadre of experienced geoinformatics professionals and students and to conduct advanced collaborative research projects, symposia, and partnerships at local and international levels in the Philippines also will be discussed.

Network of Research Infrastructures for European Seismology (NERIES)—Progress Review

By Rémy Bossu,[1] Torild van Eck,[2] Domenico Giardini,[3] Stefan Wiener,[3] and the NERIES Consortium[4]

[1]Euro-Mediterranean Seismological Center (EMSC), Bruyères le Châtel, France.

[2]Observatories and Research Facilities for European Seismology (ORFEUS), De Bilt, The Netherlands.

[3]Swiss Federal Institute of Technology, Zurich, Switzerland.

[4]NERIES, Royal Netherlands Meteorological Institute, De Bilt, The Netherlands.

NERIES (http://www.neries-eu.org) is a research infrastructure project addressing observational seismology in its broadest sense and providing opportunities far beyond the consortium members alone. Many elements within this Integrated Infrastructure Initiative (I3) European Commission project are currently being realized, and a wide-scale collaboration with other projects within Europe and the United States has been established. This presentation will provide an overview of the current status of the project, its ongoing and planned activities in 2008, and the opportunities provided.

NERIES so far has accomplished the following: (1) the Virtual European Broadband Seismic Network (VEBSN) has been extended to more then 250 broadband stations, (2) two deep-sea ocean-bottom seismometer systems have been operating in the Mediterranean for nearly one year, (3) homogeneous earthquake ground shaking maps (shakemaps) can currently be produced at several European observatories, (4) prototypes of portal elements have been launched for broadband-waveform retrieval services, earthquake-parameter services, historical earthquake data, and European tomography model-review and site-response software. In addition to numerous small meetings, NERIES also organized focused workshops and meetings to promote coordination on a European scale; these events addressed such topics as the acceleration of data exchange, European observatory coordination, software developments for Web application, and historical seismology. Grants for European earth scientists have been provided and will continue to be available for research visits at several institutes.

Project collaborations involving NERIES have been set up with the following: (1) the EarthScope program, for Web portal developments; (2) the U.S. Geological Survey, for rapid parameter exchange and shakemap developments; (3) the German-Indonesian Tsunami Early Warning System (GITEWS) program, for waveform and parameter handling; (4) the Seismic Early Warning For Europe (SAFER) program, for rapid warning; (5) the Southern California Earthquake Center (SCEC) for hazard assessment as a function of time;

and (6) other global partnerships, for standardization issues such as Extensible Mark-up Language (XML).

One of the goals of NERIES is to design and develop a Web portal, which would be the uppermost layer that provides rendering capabilities for the underlying sets of data. The portal would offer tools and services related to earthquake data to the earth-science community and to the public. The proposed portal is presented in separate posters with a demonstration of the alpha version.

OneGeology—The Global Context for a European E-Geoscience Project

By Ian Jackson[1]

[1]British Geological Survey, Keyworth, Nottingham, United Kingdom.

In February 2006, a deceptively simple concept was put forward. Could we use the International Year of Planet Earth (IYPE2008) as a stimulus to begin the creation of a digital geological map of the planet at a scale of 1:1,000,000? 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 geoscientist-friendly vehicle of creating a tangible geological map to in turn accelerate the 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 thereby 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, plus the opportunity to raise the profile of geoscience as part of IYPE2008, seemed to be more than enough reasons to take the proposal to the next stage.

In March 2007, in Brighton, United Kingdom, 81 geoscientists from 43 countries and 53 national and international bodies gathered together to consider whether they would be prepared to collaborate in order to create a global, interoperable, geological map dataset. The participants unanimously agreed to the Brighton "Accord" and kicked off "OneGeology," an initiative that now has the support of 78 nations. The agreed OneGeology mission is "to make Web-accessible the best available geological map data worldwide at a scale of about 1:1 million, as a Geological Survey contribution to the International Year of Planet Earth." (See http://www.onegeology.org/what_is/mission.html/.) The aim is to create dynamic, digital, geological map data for the world with an initial target scale of 1:1,000,000, but the project is pragmatic and accepts a range of scales and the best available data. The geological map data are being made available via a distributed Web service, using Web Map Service (WMS) and Web Feature Service (WFS). Geological surveys are dynamically "serving" the data for their territories to a Web portal. OneGeology is accelerating the global introduction of the foundation technologies necessary for the dynamic interchange of geoscience data and allows real-time access to the latest version of information and knowledge from the geological surveys of the world.

Since those early days in 2006, OneGeology has grown to be an international project that has progressed not only in its scientific and technical goals by launching the first version of its Web map portal with map data from many nations, but it has also attracted substantial scientific, public, and media interest around the world and spawned continent-wide activity and projects. A major project involving 29 partners and 20 nations has now been funded by the European Commission. The project—OneGeology-Europe—has a budget of 3.25 million Euros and will directly support the INSPIRE directive that is creating a spatial-data infrastructure for Europe. The project will start in September 2008. The project proposal has a very straightforward case: geological data are a key environmental dataset, which is essential to the health and wealth of society. Although rich geological data assets exist in the geological surveys of each individual Member State, they are extremely difficult to discover, to obtain, to use, and to integrate with each other. Geological spatial data are necessary for, among many other things, the prediction and mitigation of landslides, land subsidence, earthquakes, flooding, and pollution. INSPIRE spatial data themes are grouped into Annexes (I, II, III). Geology is a key dataset in INSPIRE's Annex II; it is also fundamental to the Annex III themes of natural risk zones, energy, and mineral resources. Geological spatial data are needed for (1) ground-water and soil protection directives, (2) the Global Monitoring for Environment and Security (GMES) program, (3) the Global Earth Observation System of Systems (GEOSS) program, and (4) the development of the Shared Environmental Information System (SEIS) by the European Environment Agency. The proposed OneGeology-Europe project will make geological spatial data held by the geological surveys of Europe more easily discoverable and accessible via the Web.

OneGeology-Europe will produce a Web-accessible, interoperable, geological spatial dataset for the whole of Europe at a scale of 1:1,000,000, which is based on existing data held by the pan-European geological surveys. The project will develop uniform standards for basic geological map data and make progress towards harmonizing the dataset (an essential first step to addressing integration at higher data resolutions). The project also will accelerate the development and deployment of a recent international interchange standard for geological data, Geoscience Mark-Up Language (GeoSciML), which will enable the sharing and exchange of the data within and beyond the European geological community. OneGeology-Europe will facilitate the re-use and addition of value by a wide spectrum of users in the public and private sector and will identify, document, and disseminate strategies for reducing the technical and business barriers to

re-use. The project plans to address the multilingual aspects of access through a multilingual discovery portal. By identifying user and provider communities and raising awareness within them, OneGeology-Europe will move geological knowledge closer to the end-user, where that knowledge will have greater societal impact and ensure fuller exploitation of the key data that have been gathered at huge public expense. The project intends to provide examples of best practice in the delivery of digital, geological spatial data to users, such as those in the insurance, property, engineering, planning, mineral resource, and environmental sectors. OneGeology-Europe will also help Europe play a leading and pivotal role in the development of a global geoscience spatial data infrastructure. Geoscience, like all environmental domains, is worldwide in its nature and reach, and through this project, Europe will make a crucial contribution to and advance the OneGeology-Global project.

In summary, OneGeology addresses head-on the challenges of interoperability and open standards and will improve access to, and exploitation of, rich and relevant digital content. In bringing together an extensive network—geological surveys with proven capacity and stability, plus users and stakeholders from a cross section of key sectors—and raising awareness within that network, OneGeology will assist with disseminating best practice while also identifying and addressing weaknesses, barriers, gaps, and opportunities. Although the problems and opportunities that geology raises are transnational, many of the issues are currently being tackled on a local basis and in a disconnected way by individual nations. OneGeology, at a global and European level, seeks to address that.

Implementation Plan for the Geoscience Information Network (GIN)

By M. Lee Allison,[1] Linda C. Gundersen,[2] Stephen M. Richard,[1] and Tamara L. Dickinson[2]

[1]Arizona Geological Survey, Tucson, Ariz.

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

Rationale for a Geosciences Information Network

Many of the challenges to creating an Earth science cyberinfrastructure are not technical but organizational and cultural in nature. Recent workshops have focused on how to achieve cooperation, integration, and community governance of a geoinformatics system. The critical stumbling blocks to creating a wide-reaching geoinformatics component of the cyberinfrastructure for the sciences are (1) agreements on common standards and protocols, (2) the engagement of a vast number of distributed data resources, (3) practices for recognition of and respect for intellectual property, (4) a simple data and resource discovery system (distributed integrated catalogs), (5) mechanisms to encourage development of Web service tools for analyses, and (6) business models for continuing the maintenance and evolution of information resources.

Geoscience Information Network (GIN)

The Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS) agreed in 2007 that "the nation's geological surveys develop a national geoscience information framework that is distributed, interoperable, uses open source standards and common protocols, respects and acknowledges data ownership, fosters communities of practice to grow, and develops new web services and clients" (Allison and Gundersen, 2007; Allison and Dickinson, 2008). The AASG and USGS subsequently formed an interagency steering committee to pursue the design and implementation of the Geoscience Information Network (GIN). The National GIN concept involves four modular components:

1. Agreement on open-source standards and common protocols through the use of Open Geospatial Consortium (OGC) standards.

2. A data-exchange model that, to begin with, will use Geoscience Mark-up Language (GeoSciML) (Cox and Richard, 2006; Richard and the Commission for the Management and Application of Geoscience Information Interoperability Working Group, 2007), which is based on the OGC-compliant Geography Mark-up Language (GML).

3. Prototype data discovery tools or catalogs such as the tentatively named "National Digital Catalog" (NDC), which is being developed under the USGS's National Geological and Geophysical Data Preservation Program (NGGDPP), and the National Geologic Map Database (NGMDB).

4. Data integration software tools developed or planned by a number of independent projects that can be used in various applications including meeting GIN goals.

The "lack of a national (U.S.) civil Earth information strategy" was noted by Gail and others (2007). They argue that the Global Earth Observation System of Systems (GEOSS) and the U.S. Group on Earth Observations fall short in addressing the Nation's Earth information needs. Instead, they call for the United States to "commit to a National Earth-Information Initiative to re-evaluate the national process of collecting and using civil Earth information, including the effectiveness of governmental organizations, the relationship between government functions and private sector activities, and the ability to effectively connect scientific developments to societal uses." We believe implementation of the GIN will effectively achieve this goal.

National Digital Catalog

As part of the NGGDPP, State geological surveys are compiling inventories of collections that they maintain, or that are outside the surveys but are available to be archived,

or that are at risk of being lost. Next year, the States will start compiling metadata records that describe, at the individual sample level, contents of these collections. Linking these data resources to the network using data interchange tools is a primary initial target of the GIN.

Completing GIN

The GIN implementation plan will enable basic network operation by (1) establishing service definitions, standard protocols, and best practices through community workshops, and (2) instituting the network architecture by means of a series of test bed systems. The first test bed will focus on services for serving interpreted geospatial features (for example, a geologic map) in the context of the International Union of Geological Sciences Commission for the Management and Application of Geoscience Information (IUGS CGI) Interoperability Working Group's GeoSciML development. Priorities for subsequent service development will be established by a steering committee; one high-priority candidate is to serve observation data recorded at point locations (for example, samples, chemical analyses, boreholes). Test-bed-network nodes initially will be implemented and tested on a single server; after a demonstration for the community, the service will be "rolled out" to other nodes in the network. We will seek expressions of interest from State geological surveys and individual USGS programs to participate in each test bed.

The network will use data discovery services that are being implemented as part of the NGGDPP and the NGMDB. Web services will enable integration of GIN data with other applications and data sources.

Sustainability

Like the Internet, a successful information network will create a tipping point at which users and providers will see the network as critical to their basic functions. When that happens, populating and maintaining that network become necessary costs of doing business. Few organizations are mandated to maintain a Web site, yet most realize that without one, they essentially do not exist in today's environment. We are quickly moving to a similar situation for sharing data in an interoperable manner.

The AASG-USGS workshop participants acknowledged the need to recognize that providing and using interoperable, Web-enabled information resources as part of their mission should be sufficiently compelling to support network maintenance and development just as they currently do for Web sites. Once the framework of GIN is built and the test beds are demonstrated successfully, we expect that other data providers and users will find compelling needs that will prompt the use of the network for a wide variety of specific tasks, which will in turn help fund the full implementation and expansion of the GIN. We also expect that each network participant will include costs for expanding their contributions to the GIN in their base operating costs and grant proposals in the same way costs for Web site activities are funded.

Education and Training

We plan a "circuit rider" approach wherein GIN technical staff members are dedicated to providing potential network participants with technical training or to actually carrying out the technical work themselves by "riding the circuit" among them for short durations. The original circuit riders were preachers in the late 1700s who rode a circuit through frontier regions of the United States to serve rural populations who had no churches. Similarly, our circuit riders will travel (in person or electronically) to organizations that want to join the GIN but need assistance or training. The circuit rider's services will be free but will need to be prioritized by the steering committee. Our goal is to give each State geological survey and USGS program the ability to write GeoSciML protocol "wrappers" to translate their datasets and to guide them on the server configurations that are necessary for the datasets to be discoverable by GIN users. For geological surveys or programs without the technical expertise to handle these chores, the circuit rider would carry them out either onsite or remotely as required. Various online services exist to facilitate a virtual environment for the circuit riders to work interactively in real time with network participants, including shared access to computers or servers while writing code or tutoring on code development.

A help desk will provide no-cost remote assistance to providers and users. The goal is to provide service not only to the initial geological survey data providers but also to other organizations that want to be early adopters of the GIN opportunities.

Mechanisms for Change and Adaptation in Technologies

The challenge to creating a dynamic, flexible, community-based network is defining and maintaining sufficient standards to make the network effective and reliable while keeping it open to new developments. The GIN will be defined by collections of service definitions, interchange formats, and vocabularies that are established (to the extent possible) independent of any particular hardware, operating system, or lower-level network protocols. Adoption of new technology will only require the implementation of network elements in a new environment, ideally with no change to any network service definitions or protocols. The architecture allows for the use of multiple conventions for different user groups.

Acknowledgments

This project is supported by National Science Foundation award 0723437 to the AASG, and by the USGS NGMDB, through support of S.M. Richard.

References Cited

Allison, M.L., and Dickinson, T.L., 2008, Final report—A Workshop on the Role of State Geological Surveys and U.S. Geological Survey in a Geological Information System for the Nation: Arizona Geological Survey Open-File Report 08–01, 22 p.

Allison, M.L., and Gundersen, L.C., 2007, Association of American State Geologists (AASG)-USGS Plan for a National Geoscience Information Network, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds, Proceedings, Geoinformatics 2007—Data to Knowledge, San Diego, Calif., May 17–19, 2007: U.S. Geological Survey Scientific Investigations Report 2007–5199, p. 76–77.

Cox, S.J.D., and Richard, S.M., 2006, A formal model for the geologic timescale and global stratotype section and point, compatible with geospatial information transfer standards: Geosphere, v. 1, p. 138–146.

Gail, W.B., Lane, N.F., and MacCauley, M.K., 2007, A U.S. Earth information strategy?: Imaging Notes, v. 22, no. 2, p. 45–46.

Richard, S.M., and the Commission for the Management and Application of Geoscience Information Interoperability Working Group, 2007, GeoSciML—a GML application for geoscience information interchange, *in* Soller, D.R., ed., Digital Mapping Techniques '06—Workshop proceedings: U.S. Geological Survey Open-File Report 2007–1285, p. 47–59.

GeosciNET—A Global Geoinformatics Partnership

By Walter S. Snyder,[1] Kerstin A. Lehnert,[2] Emi Ito,[3] Ulrich Harms,[4] and Jens Klump[4]

[1]Department of Geosciences, Boise State University, Boise, Idaho.

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

[3]Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minn.

[4]German Research Center for Geosciences, Potsdam, Germany.

GeosciNET is an emerging partnership of existing geoinformatics organizations that are collaborating to provide a more effective data system. Current members are: CoreWall (http://www.corewall.org), Geoinformatics for Geochemistry (GfG; http://www.geoinfogeochem.org), System for Earth Sample Registration (SESAR; http://www.geosamples.org), PaleoStrat (http://www.paleostrat.org), and the International Continental Drilling Program (ICDP; http://www.icdp-online.org). With the existing membership, GeosciNET can offer a comprehensive, integrated system for data acquisition in the field, data dissemination, archiving, visualization, and data integration and analysis. The system will enable a single researcher or a group of collaborators to keep track of, visualize, and virtually archive any type of geologic sample, the data produced from that sample, and subsamples taken from the original sample. Samples can be solid, liquid, or gas; can be from an outcrop or drilled or dug from a pit; or may even be virtual (for example, a spectral analysis of an outcrop detected by satellite imagery).

GeosciNET is envisioned as a model to advance the larger, ongoing process of building a global geoinformatics system and it is open to new partners that would expand the scope and impact of the partnership. Lehnert and others (2008) stressed that geoinformatics needs to be developed as a linked system of sites that provide to users a library of research data and tools to discover and access, integrate, manipulate, analyze, and model interdisciplinary data without corrupting the original data. The "grand challenge" for geoinformatics is to build such a linked system. Enforced networking of geoinformatics systems (the "top-down" approach) has not been perceived well by academics that tend to value bottom-up systems that can be more responsive to users' needs. There are major roadblocks to building networks within government and academic domains as well as between them partly because of perceived differences in their respective modes of operation and partly due to inadequate funding. Few links exist today among various geoinformatics efforts, so it is difficult for anyone to know where various data are (data discovery), to integrate them (interoperability), and to view diverse data in a synthetic and dynamic way (visualization). GeosciNET's aim is to eliminate the obstacles so that users can take advantage of geoinformatics resources and value their benefits. Once these benefits are understood by the user community, the barriers that currently exist in building a larger geoinformatics system will start to erode.

We are organizing GeosciNET to advance coordination, complementarity, and interoperability and to minimize both the duplication of efforts and the overlap of scope among the involved partner systems in order to streamline the development and operation of geoinformatics efforts. We believe that by advancing the development and data holdings of its member groups, the overall value of each site will be significantly enhanced and will better meet the needs of the users. We are jointly developing a plan that outlines the proposed interaction among the projects and their specific responsibilities for building a network of data, services, and tools. Our goal is to establish an integrated, apparently seamless network that can be offered to the community of users to support science and education programs. This collaboration is based on a memorandum of understanding predicated on the idea of mutual benefit. Specific responsibilities of partner projects are based on maximizing this mutual benefit and the technical capabilities of the partners. For example, the Antarctic Drilling Program (ANDRILL), ICDP, and others are work-

ing with CoreWall to enhance technologies that can then be implemented by their projects. One key need that has been identified by the user community is that CoreWall needs to be able to push data out to other databases, including but not restricted to PaleoStrat and GfG members. PaleoStrat is partnering with the International Congress on Carboniferous and Permian (ICCP), the Permian and Carboniferous Subcommissions of the International Commission on Stratigraphy (ICS), and GeoSystems (http://www.geosystems.org) to support the newly formed Upper Paleozoic Paleoclimate Working Group. EarthChem (http://www.earthchem.org), which is a core member of GfG, is building a Geochemistry Information Network and has established a consortium of international partners with the goal of accessing globally distributed collections of geochemical data via the EarthChem Portal. All of these activities will enhance the effectiveness of all GeosciNET partners.

A major focus for GeosciNET is to support individual researchers and projects that do not have their own dedicated data management, education, and outreach programs. One of the greatest challenges for geoinformatics lies in being perceived as a friendly resource by its users where they can easily link their observations and analyses to other users and integrate them with other data. These data, when viewed holistically, provide a continuum of information that allows the geoscience community to address fundamental questions of earth processes. The data can provide us with information about the environment or natural resources that we did not know we had or with questions we did not know to ask.

Despite the importance of data (legacy or otherwise), there currently are no convenient mechanisms that enable users to easily input their data into databases. Although the developers of some efforts such as the GfG databases, PetDB and SedDB have worked hard to compile such data, only the users' active participation can capture the major part of the overall legacy and new data. User participation requires the proper tools such as a translator that can recognize tags and parse the data accordingly, and incentives such as tools and more data for enhanced data synthesis and analysis. GeosciNET will be experimenting with these mechanisms. Efficient capture of legacy and new digital data is part of the larger data preservation "grand challenge" that includes physical samples; there are many government and academic drilling core and sample repositories that can benefit from SESAR, CoreWall, and other components of GeosciNET.

Reference Cited

Lehnert, K.A., Harms, U., Ito, U., Klump, J., and Snyder, W.S., 2008, Promises, achievements, and challenges of networking global geoinformatics resources—Experiences of GeosciNET and EarthChem: Geophysical Research Abstracts, v. 10, 2 p., available only online at http://www.cosis.net/abstracts/EGU2008/05242/EGU2008-A-05242-1.pdf/. (Accessed July 31, 2008.)

A Three-Dimensional GIS Model Based on Crystallographic Principles Dedicated to Spatial Analyses in Geosciences

By Benoit Poupeau,[1] Benoit Deffontaines,[1] and Olivier Bonin[2]

[1]Department of Earth Materials and Engineering Geology, University of Paris, Marne-la-Vallée, France.

[2]Object Design and Generalization of Topographical Information (COGIT) Laboratory, National Geographic Institute, Saint Mandé, France.

A three-dimensional geographic information system (GIS) enables the integration and coherence of multiple sources of data from different providers while respecting and representing the end user's choices in terms of geometry and topology. Current three-dimensional GISs use unique topological and geometrical modeling as a rule (Zlatanova, 2000; Coors, 2003). This feature of a three-dimensional GIS makes queries easier to compute from topological models such as the "close to close" way of relating geometric primitives of one object or its neighbor objects; however, this homogenization leads to the loss of the model's specificities, leading to heavy computations for conversion of the data. Additionally, it does not compensate in an automated way for issues arising from data acquisition and modeling.

This paper proposes a model for analysis based on a three-dimensional GIS. The proposed approach allows queries to be made on one object (intra-object analysis) or a set of objects (inter-objects analysis) even if the geometrical coherence between objects is not perfect. This model, which is based on the principles of crystallography, analyzes the symmetric features of each object in order to describe its structure, such as the way in which geometric primitives are arranged together. This first abstraction (that is, the structure of the object, which allows handling of the object independent of its geometry) gives a global overview of the object. One of the advantages of the structure is to make some queries easier, such as the roof extraction of a cavity or the three-dimensional building simplification.

A second abstraction, the bounding crystalline mesh or lattice unit in the terminology of crystallography, is obtained through the analysis of symmetric elements (plans, axes, or centers). The lattice unit is, in crystallography, the envelope of the smallest parallelepiped, which is a structure that preserves the geometric properties. The lattice unit is used like a three-dimensional bounding box adapted to the object shape and allows relationships of geographical objects, regardless of their geometric dimension. Relationships between geographical objects follow two principles:

1. Each geographical object is subject to gravity; therefore, emptiness is not allowed in the model. The main objective is to connect non-adjacent objects (for example, a house that is not in contact with a Digital Terrain Model) or to ensure relationships between the parts of different objects

(for example, connect the roof of a geological layer with those of an included cavity).

2. If a lattice unit intersects another lattice unit, we assume that these lattice units are adjacent. This principle is formalized with proximity spaces (Naimpally and Warrack, 1970). This mathematical theory constitutes, in topology, an axiomatization of notions of "nearness").

With the help of lattice units, two graphs are computed. The first one is an incidence graph, which describes relationships between objects and makes it possible to establish the interrelationship between them. The second one, a temporal graph, represents, for one object, the evolution of its relationship with its own environment.

This model has been used and validated in different applications, such as three-dimensional building simplification (Poupeau and Ruas, 2007), and in a context of a coal basin affected by anthropic subsidence due to extraction (Gueguen, 2007).

References Cited

Coors, V., 2003, 3D GIS in networking environments, *in* Environments and Urban Systems: Amsterdam, The Netherlands, Elsevier, p. 345–357.

Gueguen, Y., 2007, Etude des mouvements de surface en environnement minier à partir d'interférométrie radar et identification des origines des déformations—L'exemple du bassin Nord/Pas-de-Calais: Marne-la-Vallée, France, University of Paris, unpublished Ph.D. dissertation, 209 p.

Naimpally, S.A., and Warrack, B.D., 1970, Proximity spaces: Cambridge Tracts in Mathematics and Mathematical Physics 59, 128 p.

Poupeau, B., and Ruas, A., 2007, A crystallographic approach to simplify 3D building, *in* Reports of participants, Twenty-third International Cartographic Association Conference, August 4–10, 2007, Moscow, Russia: Moscow, Russia, Federal Geodetic and Cartographic Service, 1 CD-ROM.

Zlatanova, S., 2000, 3D GIS for urban development: Enschede, The Netherlands, International Institute for Geo-Information Science and Earth Observation, unpublished Ph.D. dissertation, 222 p. and appendix.

Avizo—Three-Dimensional Visualization Framework

By Peter Westenberger[1]

[1]Visualization Sciences Group, Mercury Computer Systems, Dusseldorf, Germany.

Avizo software is a powerful, multifaceted tool for visualizing, manipulating, and understanding scientific and industrial data. Wherever three-dimensional (3D) datasets need to be processed in material sciences, geosciences, or engineering applications, Avizo offers abundant state-of-the-art features within an intuitive workflow and an easy-to-use graphical user interface.

Avizo XGreen Package—3D Visualization Framework for Climatology Data

In these times, when climatology data have become more and more complex with respect to size, resolution, and the numbers of chronological increments, the German Climate Computing Center in Hamburg has chosen Mercury Computer Systems to develop a software extension called XGreen, which is based on their visualization framework Avizo (formerly known as "amira").

XGreen provides domain-specific enhancements for Avizo, such as the following:

- Network Common Data Form (NetCDF) CF-1.0 Reader—
  - Support for regular, rectilinear, rotated, and curvilinear grids.
  - Support for large data.
  - Streaming of time-dependent data.
  - Use of main memory for caching NetCDF data.
- Geographical projections—
  - Cylindrical, equidistant, spherical, Mollweide, and many other projections.
- Earth module—
  - High-level textures, three levels of detail.
  - Scalable elevation and bathymetry.
  - Continental outlines and country borders.
- Fast hardware-based bump shading for multiple scalar quantities on two-dimensional (2D) slices.
- Particle advection and trajectories.
- Volume rendering for rectilinear grids.
- Several other unique interactive visualization techniques.

Avizo Earth Edition—3D Visualization Framework for Geoscience Data

Avizo Earth Edition is the software suite that includes Avizo and all its extensions for interactive exploration, visualization, analysis, comparison, and presentation of geoscience

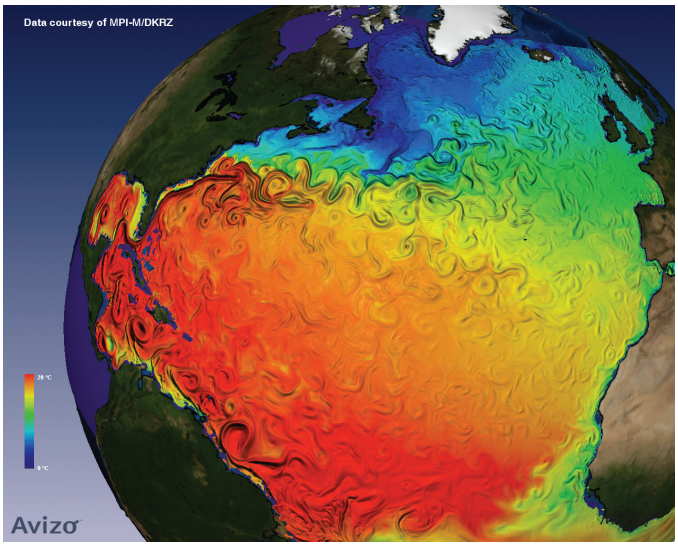


Figure 1. A visualization of 1 time step out of 1,440 in an ocean model simulation. The colors represent the temperature near the surface of the Atlantic Ocean, with blue being the coolest and red being the warmest. The embossing illustrates the current's velocity. Data courtesy of the Max Planck Institute for Meteorology and the German Climate Research Center in Hamburg, Germany.

data. This 3D visualization framework is an ideal tool, allowing the user to import, manage, interact with, and visualize geoscience data from multiple sources within a single environment.

Avizo Earth Edition includes the whole Avizo feature set plus an advanced SEG-Y (file format used by the Society of Exploration Geophysicists) data importer, and the XLVolume Pack, which manages and visualizes very large amounts of volume data, up to terabytes.

Avizo Earth Edition addresses several geosciences needs with the following features:

- Quality control of 3D seismic data.
- Multidiscipline project review.
- Final project presentation.
- Core sample analysis.
- Rapid application development

The Earth Edition delivers advanced technologies and extensions suitable for efficient, multi-data visualization of large quantities of geosciences data by means of the following features:

- Ultimate memory management technology to interactively explore out-of-core datasets.
- Versatile data management capabilities through the SEG-Y wizard, the OpenSpirit plug-in, and user-defined readers.
- Readers for horizons, fault sticks, and fault surfaces.
- Advanced seismic visualization by interactive investigation using inline, crossline, time slices, oblique slices, and fence slices (random slices); uses slice animation to identify hidden features.
- Interactive region of interest, with progressive image quality during 3D navigation.
- Embossing to enhance feature identification (fig. 1).
- Advanced colormap editor to interactively change colormap transparency curve "on the fly;" several ready-to-use colormaps for instantaneous attributes, velocity, and other geoscience data.
- Horizon visualization and quality control using time-depth colormap.
- "GeoBody" segmentation through Avizo's advanced segmentation editor.
- Data query from seismic volumes using PointProbe, LineProbe, and SplineProbe
- Multidiscipline, multiviewer project review—Visualize seismic volumes, wells, horizons, faults, and reservoirs in a single 3D environment, and manage exploration using a dedicated "project tree" view.
- Professional project presentations using a large set of tools, including DemoMaker, CameraPath, MovieMaker, 3D Annotation, virtual trackball, and high-resolution snapshots, and more.
- Project review in immersive theaters is also possible using the Avizo XScreen extension which provides scalability on clusters and virtual reality systems.

Tunisian Structural Extrusion Revealed by Numerical Geomorphometry

By Benoit Deffontaines,[1] Tarek Slama, Jr.,[2] Noamen Rebai,[2] and Mohamed Moncef Turki[2]

[1]Earth Materials and Engineering Geology Laboratory, University of Paris, Marne-la-Vallée, France.

[2]Department of Geology, University of Sciences of Tunis, Tunis, Tunisia.

Neotectonics may be revealed by detailed numerical geomorphic analyses of topography. Tunisia is an excellent case example of a country affected by active tectonics; numerous earthquakes have struck the region, and faults, folds, and associated structural features have been mapped by geoscientists. Previous geomorphic studies of the region and the development of new indicators in this current study

have lead us to propose a new structural scheme for analyzing the Tunisian tectonic setting. We propose herein an eastern extrusion model of central Tunisia based on the northward migration of the African plate toward Eurasia. This model is based on drainage analyses, drainage network classifications, specific analyses of the digital terrain model (DTM), summit-level surface analyses, and other analyses which were integrated in a geographic information system (GIS) (Deffontaines, 1990, 2000; Deffontaines and others, 1994) as well as previous studies (Sokoutis and others, 2000; Bouaziz and others, 2002).

We propose that the well-known diapir fault line (Medjerda-Tunis fault zone) corresponds to a major left-lateral, transtensive, northeastward-trending fault zone, which acts as the major northern boundary of the central Tunisia extrusion. The zone is identified in the field as a compressive structure that resulted from the continuous uplifting of the elongated northeast-southwest-trending salt diapirs that parallel this major transcurrent extrusion.. The zone is associated with the well-known northeast-southwest-tending Gafsa-Gabes fault zone, which is characterized by numerous en echelon folds and acts as a major right-lateral fault zone that bounds the southern part of the central Tunisia extrusion (Bouaziz and others, 2002). Within central Tunisia, the north-south axis of the zone (also known as the Al Abiod fault zone) appears to be a reactivated graben that is closely associated with the eastern extrusion of central Tunisia and differentiates the high Atlasic and low eastern Tunisian domains (Bouaziz and others, 2002). This geomorphic approach is limited by the difficulty of distinguishing the different tectonic phases within the long, northward migration of the African plate because the topography reveals cumulated effects; however, the development of a more advanced system is underway.

Geomorphometry appears to be an excellent tool for understanding, at a regional scale, the geodynamic setting of this northeastern part of northern Africa where the African-Eurasian collision is characterized by an eastern extrusion in the central part of Tunisia. Further work is proposed, such as studies of bathymetry, gravimetry, and magnetism; and a quick offshore seismic-reflection survey that would help locate the exact continuation of the Tunisian extrusion's major tectonic boundaries.

References Cited

Bouaziz, S., Barrier, E., Soussi, M., Turki, M.M., and Zouari, H., 2002, Tectonic evolution of the northern African margin in Tunisia from paleostress data and sedimentary record: Tectonophysics, v. 357, p. 227–253.

Sokoutis, D., Bonini, M., Medvedev, S., Boccaletti, M., Talbot, C.J., and Koyi, H., 2000, Indentation of a continent with a built-in thickness change—Experiment and nature: Tectonophysics, v. 320, 243–270.

Project Towards Simultaneous Visualization for Various Kinds of Geoscience Data on Google Earth

By Yasuko Yamagishi,[1] Hiromichi Nagao,[1] Seiji Tsuboi,[1] Katsuhiko Suzuki,[1] Hajimu Tamura,[1] Hiroshi Yanaka,[2] and Tadahiro Hatakeyama[3]

[1]Institute for Research on Earth Evolution, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan.

[2]Fujitsu Limited, Chiba, Japan.

[3]Okayama University of Science, Okayama, Japan.

Scope of the Project

Numerous types of data are available throughout the community of solid-earth scientists. New insight into the structure of and activity in the Earth's interior could be gained by a simultaneous interpretation of multidisciplinary data. The Japan Agency for Marine-Earth Science and Technology (JAMSTEC) also has been accumulating various bodies of geoscience data (obtained by observations both on land and in the oceans) related to the solid earth. In order for cross-disciplinary research to occur, it is necessary, as a first step, to visualize simultaneously these different types of data in a common platform. Google Earth is a powerful tool for the simultaneous visualization of data; it converts geoscience data into keyhole mark-up language (KML), a format that is expected to be available for everyone.

The Institute for Research on Earth Evolution (IFREE), which is a data collection center within JAMSTEC, adopted Google Earth as a common browser for databases produced by the solid-earth science community and has been promoting a project to develop the conversion tools to produce KML files easily and quickly. Here, we introduce the conversion tools developed in the project.

Released Conversion Tools from the Project

The present target data to be converted to KML files are seismic tomography data, geomagnetic data, geochemical data of rocks, and navigation data from JAMSTEC's research vessels. The conversion tool for the seismic tomography data has already been released as Web application software and is available from the Pacific 21 Website (http://www.jamstec.go.jp/pacific21/). It is possible to download, for example, a KML file of an arbitrary horizontal and vertical cross section of the seismic tomography model of the Earth's mantle proposed by Obayashi and others (2006), Isse, Suetsugu, and others (2006), and Isse, Yoshizawa, and others (2006), setting parameters such as which cross section and area is to be displayed using Google Earth.

A conversion tool for geochemical data from the following two databases is also available: Geochemistry of Rocks of the Oceans and Continents (GEOROC, http://georoc.mpch-mainz.gwdg.de/georoc/), and Petrological Database of the Ocean Floor (PetDB, http://www.petdb.org/), which archives analytical data (major element composition and isotope ratios, for example) for rock sampled from the ocean floor.

As a part of this project, we have been developing a conversion tool that enables us to view the global or local geomagnetic field on Google Earth. This tool makes it possible to generate a KML file from a geomagnetic field model, which is given as a table of spherical harmonic coefficients such as those used by the International (or Definitive) Geomagnetic Reference Field and the National Geophysical Data Center 720 models.

In this presentation, we show examples of KML converters for seismic tomography models, geomagnetic field models, the geochemical data of rocks, and navigational data for JAMSTEC vessels. We also show a simultaneous comparison of these data on Google Earth.

Summary

Cross-disciplinary cooperation between various fields in the solid-earth sciences becomes more and more important in order to understand the Earth's interior. A simultaneous visualization of different types of geoscience data is essential in order to achieve this purpose, and we consider Google Earth to be the best solution for a data browser. We plan to provide more conversion tools such as Web and (or) Java applications not only for the data presented here but also for other types of data. We believe our conversion tools will elevate the use of simultaneous visualization as a research tool and enable new discoveries about the Earth's interior.

References Cited

Isse, T., Suetsugu, D., Shiobara, H., Sugioka, H., Yoshizawa, K., Kanazawa, T., and Fukao, Y., 2006, Shear wave speed structure beneath the South Pacific superswell using broadband data from ocean floor and islands: Geophysical Research Letters, v. 33, 5 p.

Isse, T., Yoshizawa, K., Shiobara, H., Shinohara, M., Nakahigashi, K., Mochizuki, K., Sugioka, H., Suetsugu, D., Oki, S., Kanazawa, T., Suyehiro, K., and Fukao, Y., 2006, Three dimensional shear wave structure beneath the Philippine Sea from land and ocean bottom broadband seismograms: Journal of Geophysical Research, v. 111, no. B6, 13 p.

Obayashi, M., Sugioka, H., Yoshimitsu, J., and Fukao, Y., 2006, High temperature anomalies oceanward of subducting slabs at the 410-km discontinuity: Earth and Planetary Science Letters, v. 243, p. 149–158.

GIS-Morphometry—A GIS Framework for Digital Tectonic Geomorphology Studies

By Tarek Slama, Jr.,[1] Benoit Deffontaines,[2] Noamen Rebai,[1] and Mohamed Moncef Turki[1]

[1]Department of Geology, University of Sciences of Tunis, Tunis, Tunisia.

[2]Earth Materials and Engineering Geology Laboratory, University of Paris, Marne-la-Vallée, France

Digital terrain modeling and landform analysis are carried out by use of (1) general geomorphometry of a digital elevation model (DEM), (2) digital drainage network analysis, (3) digital image processing, (4) lineament extraction and analysis, (5) spatial and statistical analysis, and (6) three-dimensional surface modeling. These approaches define the growing field of digitally analyzed tectonic geomorphology. Most of its investigations include visualizing geology imagery with topography, raster calculations using map algebra procedures, and the integration of fault-model calculations to determine surface uplift. Analysis of DEMs by means of geomorphometry provides an efficient tool for recognizing fractures and for quantitatively characterizing the morphotectonics and the morphodynamics of an area.

Topographic attributes extracted from the DEM allow for the general and specific characterization of relief in terms of coupled and inherent interactions between surface processes and tectonic activity. The DEM manipulations illustrate the basic evidence of deformation and surface processes recorded in the topography. In addition, digital drainage network analysis and digital image processing allow, via morphometric parameters extraction and integration, for morphotectonic investigations of the landform. The large number of quantitative attributes, however, requires a well-organized digital system to ensure, numerically, systematic extraction and manipulation. The implementation of an adequate framework using available geographic information system (GIS) technology is the motivation for and the primary goal of our work.

A GIS architecture and database has been designed and developed to ensure many quantitative procedures and approaches. They are originally based on landform investigations and geomorphological characteristics which were translated into mathematical and numerical algorithms. The general framework of the developed GIS is composed of three core "morphomtric components" as follows: (1) DEM-morphometry, (2) digital drainage network (DDN)-morphometry, and (3) digital image (DI)-morphometry. A large set of morphometric parameters are extracted from these GIS substructures; however, classical operations and GIS functions also are ensured, such as data integration and interoperability, data management and spatial analysis, topology, and interactive visualization. This GIS-Morphometry framework was developed using the ArcGIS package, the ModelBuilder tool, and Python programming language. The GIS object-oriented technology was used to extract automatically most of the morphometric parameters

and attributes. A comprehensive geodatabase was structured for the purposes of data integrity and effectiveness of data manipulation. The system's completely digital nature ensures that it will be flexible so that it can grow and evolve as new data, processing procedures, and modeling and visualization tools become available. A set of morphotectonic and morphostructural maps of test sites in northern Tunisia was created. The neo-morphodynamic model of folded and faulted structures was particularly emphasized and mapped.

Geoinformatics and Nature Parks

By Peter Löwe,[1] Claudia Eckhardt,[2] and Ralf Löwner[1]

[1]German Research Center for Geosciences, Potsdam, Germany.

[2]Geo-Naturpark Bergstrasse-Odenwald, Lorsch, Germany.

Introduction—Google Earth Blazes the Trail

This paper describes the mutually beneficial relationship geoinformatics can have with nature parks—to provide location-based up-to-date information for park visitors based on decentralized data and knowledge repositories. Geoinformatics-based applications such as Google Earth have become widely accepted for everyday use during the last few years. They have been readily accepted by both laypersons and academia. In this paper we highlight how similar geoinformatics-driven approaches can be applied to nature parks. This approach enables nature park staff to help visitors understand global change processes on a local scale. The approach is based on new databases and services that are maintained and operated by online user communities and by new technology for local and regional applications.

Nature Parks—Potential and Challenge

Geology forms the foundation of any park's ecosystem by setting the stage and providing the context for its local natural and historic heritage; therefore, having visitors acknowledge the beauty of the geologic features and having park staff communicate information about the geologic setting of those features to the visitors are interactions that are central to any nature park concept. Some nature parks in Europe are members of the European Geoparks Network and the Global Geoparks Network, which is supported by UNESCO. These organizations have an obligatory, strong focus on informing the public about the parks' geological and cultural heritage and protecting both in order to achieve sustainable regional development. The following quote by Mike Soukup (Associate Director for Natural Resource Stewardship and Science, National Park Service, United States), as cited in Higgins (2007), highlights the current status: "Intuitive decision-making might have sufficed in the twentieth century, (but) it certainly will not ensure that the natural systems (the wildlife and the scenery) of national parks will be maintained unimpaired throughout the 21st century."

Good management of nature parks requires a sound understanding of the available physical settings and anthropogenic infrastructure, including the scientific aspects. They need to be appropriately communicated to park visitors and academia.

An ongoing dialogue between geoscientists and park resource-management staff about research needs, projects, and grant appeals is crucial in order to introduce the means for knowledge management, data access, and representation of the facts and findings (fig. 1); however, according to Higgins (2007), even within U.S. system of national parks, where landscape interpretation was first introduced as a means to communicate nature and science to the public, about 90 percent of the parks still lack an onsite geoscientist. This indicates the huge potential for contributions from the field of geoinformatics.

Geo-Naturpark Bergstrasse-Odenwald

We present the Geo-Naturpark Bergstrasse-Odenwald, in Lorsch, Germany, as a real-world scenario that depicts the potential benefits of applying geoinformatics to nature parks. The park covers the Odenwald mountain range between Frankfurt and Heidelberg and stretches from the Rhine River valley to the wine-producing region of Franconia. The park was the first nature park in Germany to achieve "Geopark" status on the national, European, and global level. The park promotes "protection by usage," preservation of heritage and knowledge, and the empowering of local enterprises (that is, sustainable tourism) under the crosscutting objective of sustainable regional development.

Real-World Examples of the Potential Impact of Geoinformatics

Apart from basic management tasks such as visitor statistics and trail management, which can be supported by off-the-shelf software solutions, geoinformatics can help to open fields of research and communication that otherwise could not be addressed with the available resources. We provide two examples.

Contemporary Regional Geology—A Challenge

Lots of regional field expertise and knowledge needs to be stored, managed, and communicated within a nature park. In the case of Bergstrasse-Odenwald, this is a significant challenge because the park's area is covered by geologic maps constructed by three different geological surveys, one each for the States of Hesse, Bavaria, and Baden-Württemberg.

The implementation of geoinformatics software to integrate the growing number of geodata infrastructures (GDIs) on both State and regional levels could be modeled after the efforts

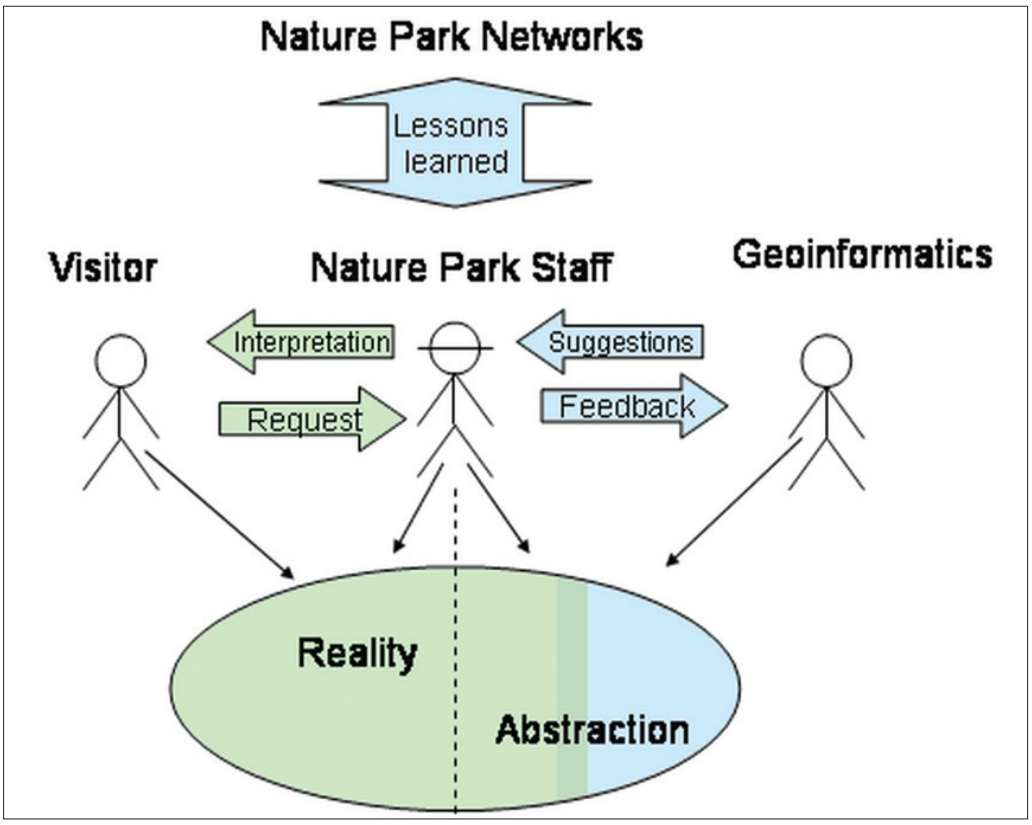


Figure 1. Interaction between park visitors, park staff, and geoinformatics advisors.

reported by Baru and Lin (2008). Despite the improvements in data access, the issue of data being kept up to date remains. For the Bergstrasse-Odenwald area, most of the available geologic data were charted more than a century ago. It will be a challenge for the future to integrate crowd-sourced contemporary observations into a regional geodatabase, but a method similar to that described by Ramm and Topf (2008) may be possible. "Crowd sourcing" describes an activity that is jointly undertaken by a community of individuals who share the same interests and pool their resources to accomplish the overall task. In our case, the mapping-related geological observations made by people spread out over the whole park area could be used to create new geologic or geomorphological map products.

Telling Stories From Data Queries

In geoinformatics, providing query results to the user is the last step in a process and typically yields a map product. Yet from the nature park perspective, the challenge to communicate scientific facts just begins at this stage. For communication "in the field," computer-independent activities (relying on vision, touch, smell, and so on) are still preferred when interacting with park visitors. Also, the challenge to communicate findings in a barrier-free manner needs to be addressed (Ludwig, 2005). In order to better convey the science of the park to the public, dedicated Web Processing Services (WPS) will help to enable the on-demand production of derivative products such as conventional maps, global positioning system (GPS) tracks, multi-thematic lenticular products, or even jigsaw-puzzles (fig. 2).

Figure 2. Using a three-dimensional jigsaw puzzle (Geocubes) to communicate remote sensing products to the public.

Conclusion

Recent developments in geoinformatics can make a significant contribution to (1) the management of a nature park's scientific information by enabling access by both park staff and visitors to local spatial data repositories using services such as Sensor Observation Service (SOS) and WPS, and (2)

the on-demand creation of adequate map-related products (fig. 2). The use of free and open-source software (FOSS) tools enables access to geoinformatics software without significant financial investments. Research results shared through park interpretive staff and interest groups will enhance the experience for all park visitors by helping to communicate the current scientific understanding of the park to the general public.

References Cited

Baru, C., and Lin, K., 2008, Mediating among GeoSciML resources: EGU Geophysical Research Abstracts, v. 10, 2 p., available online at http://www.cosis.net/abstracts/EGU2008/11390/EGU2008-A-11390.pdf. (Accessed August 5, 2008.)

Higgins, Bob, 2007, Geoscience research in U.S. National Parks: Eos, Transactions of the American Geophysical Union, v. 88, no. 21, p. 226.

Ludwig, Thorsten, 2005, Kurshandbuch Natur- und Kulturinterpretation: Bidungswerk interpretation Web site at http://www.interp.de/. (Accessed August 5, 2008.)

Ramm, Frederik, and Topf, Jochen, 2008, OpenStreetMap—Using and improving the free world map: Berlin, Germany, Lehmanns Media, 288 p.

Geoinformatics on the Front Lines—Purdue University's Inaugural Geoinformatics Course

By C.C. Miller[1] and Michael Fosmire[2]

[1]Earth and Atmospheric Sciences Library, Purdue University, West Lafayette, Ind.

[2]Physical Science and Engineering Libraries, Purdue University, West Lafayette, Ind.

In 2007, the National Science Foundation's (NSF) Cyberinfrastructure Council published their "Cyberinfrastructure Vision for 21st Century Discovery." In it, the NSF lays out a plan for directing and funding the initiatives designed to take the many and rapid advances in (mostly) large-scale computing architectures, communication and data transfer protocols and very large data storage capabilities and build from them a coordinated, integrated, interoperable cyberinfrastructure that, among other virtues, "serves as an agent for broadening participation and strengthening the nation's workforce in all areas of science and engineering" (NSF Cyberinfrastructure Council, 2007 p. 6). Much of what the NSF proposes addresses the challenges of distributed, high-performance computing at the scale and scope one would expect from an NSF response to a revolution; yet, they've taken care to acknowledge in their plan the fact that, although an electronic infrastructure is an increasingly vital component of scientific research, these machines still require engineers, of sorts: scientists who can care for the data drawn from and fed into the electronic machines that produce or consume them. The Council also states, "In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form" (NSF Cyberinfrastructure Council, 2007, p. 22). The Council goes on to state that "ongoing attention must be paid to the education of the professionals who will support, deploy, develop, and design current and emerging cyberinfrastructure" (NSF Cyberinfrastructure Council, 2007, p. 38).

In other words, just as high-end computing architectures must be relied upon to process and manipulate and share data, there is an equally important amount of human finesse that goes into the preparation, consumption, interpretation of, and care for those data. To this end, faculty from Purdue University Libraries and the Department of Earth and Atmospheric Sciences cooperated to offer an inaugural course in Spring 2008 entitled "Geoinformatics." The course was designed to be a discipline-independent overview of emerging trends and issues in the geosciences that fall within the purview of geoinformatics. The course was intended to fill a need within the emerging universe of cyberinfrastructure that will "both demand and support a new level of technical competence in the science and engineering workforce and in our citizenry at large" (NSF Cyberinfrastructure Council, 2007, p. 37). The instructors intended the course to fit somewhere within that hotspot between (1) cyberinfrastructure and the Semantic Web and (2) the future of data and rapidly developing, increasingly geospatially savvy technologies. The intended focus of the course was on the "workforce" component of the NSF vision (NSF Cyberinfrastructure Council, 2007, chapter 5). The plan was to begin teaching our next generation of scientists about the stores of data available in online systems, the power and limitations of those networked tools and data structures, and the importance of "good data hygiene."

Data are only as interoperable as the scientists who are willing to understand the technologies, adhere to standards, and share their work with others. The provenance of data—who collected them, how they collected them, and whether and how they have been verified—is an important factor for researchers to consider before incorporating external data into their analyses, but it is just as important when it comes time to convey these data and the results of data analyses back into the community via the growing cyberinfrastructure-based sharing and dissemination systems and digital libraries. The course instructors therefore emphasized the concept of data management and sharing mechanisms throughout the course. All coursework was put into the context of the greater world of geodata and geodata issues. Course modules ranged from data collection and massaging (such as the development and use of a global positioning systems (GPS), or statistics and databases) to the more semantic concerns of metadata, ontological structures, and systems designed to assist with data stewardship and sharing.

In this presentation, one of the instructors (a geographic information system (GIS) librarian) of Purdue's "Geoinformatics" course will (1) discuss the factors that influenced the development of the course, (2) briefly describe the modules, assignments, and technologies that were introduced in the course, and (3) address the successes and failures of an attempt to introduce the gargantuan world of geoinformatics to a diverse (by background and technical skill) body of 13 students.

An infrastructure is only as good as the ability of societies and cultures to develop and deploy solutions upon it. Likewise, future scientists will only be willing to work within the bothersome restrictions necessitated by an adherence to standards and interoperability if they have been trained and have been convinced of the benefits of doing so. Stated in analog terms, perhaps one last time, even the fantastic technology of the book was lost on the illiterate. The "literacy skills" required of scientists in the interdisciplinary, high-grade cyberinfrastructure future proposed by the NSF and others are perhaps more nascent than the revolution itself. This is geoinformatics on the front lines: the lessons learned by the students ideally will have prepared them to move further into their respective domains with their eyes open to the opportunities that exist (or will exist) for advancing disciplinary or interdisciplinary research, and the lessons learned by the course instructors speak to the difficulties of moving geoinformatics itself into the future and the importance of doing so.

Reference Cited

National Science Foundation Cyberinfrastructure Council, 2007, Cyberinfrastructure vision for 21st century discovery: Arlington, Va., National Science Foundation, 57 p. (Also available online at http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.) (Accessed on August 5, 2008.)

Oral Session II

A Cyberinfrastructure-Based Portal for Topographic Data Access, Processing, and Community Interaction

By Christopher J. Crosby,[1] Viswanath Nandigam,[1] Chaitan Baru,[1] Newton Alex,[2] J. Ramon Arrowsmith,[2] and Ashraf Memon[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.

Over the past decade, there has been dramatic growth in the acquisition of publicly funded high-resolution topographic and bathymetric data for scientific, environmental, engineering, and planning purposes. Because of the richness of these datasets, they are often extremely valuable beyond the initial application that drove their acquisition and thus are of interest to a large and varied user community; however, because of the massive volumes of data produced by high-resolution mapping technologies such as light detection and ranging (LiDAR), it is often difficult to manage and distribute these datasets via the Internet. Furthermore, the datasets can be technically challenging to work with because they require specific software and computing resources that are not readily available to many users. Through a variety of cyberinfrastructure tools, we have launched an initiative to build an online portal (the Open Topography Portal, or OpenToPo, at http://www.opentopography.org) that provides integrated access to high-resolution topographic data and Web-based processing tools and enables the community of users to share knowledge, experiences, and resources. OpenToPo builds upon the cyberinfrastructure-based system developed in the Geosciences Network GEON LiDAR Workflow (GLW) project during four years of collaboration between earth scientists at Arizona State University and computer scientists at San Diego Supercomputer Center.

OpenToPo will use the GLW as its core cyberinfrastructure-based system to provide online access to multibillion-point, high-resolution, LiDAR topography datasets. To address the distribution of both LiDAR point data as well as standard, high-resolution digital elevation models (DEMs) produced from LiDAR and provided by the data vendor, we have developed multiple pathways for users to access data. We employ a Google Maps- or Google Earth-based interface to allow users to browse and download standard, tiled digital elevation data. For users who wish to explore the full potential of the LiDAR data, we provide access to the raw LiDAR point data as well as a suite of DEM generation tools to enable users to generate custom DEMs to best fit their science applications. Through

these multiple pathways, we are able to service various user communities and thereby democratize access to these challenging community datasets.

Given the diverse applications of these datasets, the relative inexperience of the user community, and the technical difficulty of working with the data, OpenToPo's goal is to offer not only access to data and processing tools but also to provide an environment for users to (1) learn about the datasets and data processing and (2) network and interact with fellow users. OpenToPo will use blogs, discussion forums, and wikis to encourage communication and interaction between users. We also hope to leverage the collective knowledge of the user community to build a system that provides processing recommendations and guidance based on what other users already have accomplished.

Currently, OpenToPo serves five datasets totaling over 7 billion data points and approximately 2.5 terabytes. This system has been selected as the primary distribution pathway for LiDAR data acquired by the GeoEarthScope component of the National Science Foundation-funded EarthScope project (which will entail more than 20 billion additional points and a significantly larger user community). OpenToPo's predecessor, the GLW, has over 180 users who have processed over 49 billion LiDAR returns and downloaded more than 6,000 DEM tiles. Future OpenToPo development includes expanding the dataset distribution and processing approach to develop a more generic workflow that will permit users to query, process, and calculate common derivatives for DEMs of various resolutions and origins. We are currently seeking collaborators to host additional datasets in the system.

Standardizing Interfaces for External Access to Data and Processing for the National Aeronautics and Space Administration's Ozone Product Evaluation and Test Element (PEATE)

By Curt Tilmes[1] and Albert J. Fleig[2]

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

[2]PITA Analytic Sciences, Bethesda, Md.

The National Aeronautical and Space Administration's (NASA) traditional science data-processing systems have focused on specific missions and on providing data access, processing, and other services to the funded science teams selected for those specific missions. Recently, NASA has been modifying this stance by changing its focus from "Missions" to "Measurements." Where a specific Mission has a discrete beginning and end, the Measurement considers long-term data continuity across multiple missions. Total column ozone, a critical measurement of atmospheric composition, has been monitored for decades on a series of Total Ozone Mapping Spectrometer (TOMS) instruments. Some important European Space Agency (ESA) missions also monitor ozone, including the Global Ozone Monitoring Experiment (GOME) and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY). With the U.S-European cooperative launch of the Dutch Ozone Monitoring Instrument (OMI) on NASA's Aura satellite, and the launch of ESA's GOME-2 instrument on its Meteorological Operational (MetOp) satellite, the ozone monitoring record has been further extended.

In conjunction with the U.S. Department of Defense (DoD) and the National Oceanic and Atmospheric Administration (NOAA), NASA is now preparing to evaluate data and algorithms for the next generation Ozone Mapping and Profiler Suite (OMPS), which will be launched as part of the National Polar-orbiting Operational Environmental Satellite System (NPOESS) Preparatory Project (NPP) in 2010. NASA is constructing a Science Data Segment (SDS) that will be used to evaluate the various NPP data products and algorithms.

The NPP SDS Ozone Product Evaluation and Test Element (PEATE) will build on the heritage of the TOMS and Ozone-Monitoring Instrument (OMI) mission-based processing systems. The overall measurement-based system that will encompass these efforts is the Atmospheric Composition Processing System (ACPS). We have extended the system to include access to publicly available datasets from other instruments where feasible, including non-NASA missions as appropriate. The TOMS and OMI systems were largely monolithic and provided only a very controlled processing flow from the raw data gathered by the satellite to the ultimate archive of specific operational data products. The ACPS will allow more open access using standard protocols, including Hypertext Mark-up Language (HTTP), Simple Object Access Protocol/ Extensible Mark-up Language (SOAP/XML), Really Simple Syndication (RSS), and various Representational State Transfer (REST) incarnations. Outside users can be granted access to various modules within the system, including an extended data archive, the ability to search metadata, and production planning and processing tools.

Data access is closely controlled. Certain datasets may be designated as being available to the public or restricted to groups of researchers or limited strictly to the originator. These finely tuned controls can be used, for example, to release one's best validated data to the public but restrict access to the version of data that may be processed with a newer, unproven algorithm until it is ready for release.

Similarly, the system can provide access to algorithms, both as modifiable source code (where possible) and as fully integrated, executable algorithm plug-in packages (APPs). This will enable researchers to download publicly released versions of the processing algorithms and easily reproduce the processing remotely, while interacting with the ACPS. The algorithms can be modified, which enables better experimentation and rapid improvement. The modified algorithms can be easily integrated back into the production system for large-scale bulk processing to evaluate the improvements.

The ACPS includes complete provenance tracking of algorithms, data, and the entire processing environment. The origin of any data or algorithm is recorded and the history of the processing chains are stored such that a researcher can understand the entire data flow. Provenance is captured in a form suitable for the system to guarantee scientific reproducibility of any data product it distributes, even in cases where the physical data products themselves have been deleted due to space constraints. We are currently working on Semantic Web ontologies for representing the various types of provenance information.

A new Web site consolidating information about measurements, processing systems, and data access has been established to encourage interaction with the overall scientific community. We will describe the system, its data processing capabilities, and methods the community can use to interact with the system.

The Open GeoSpatial Consortium Web Coverage Service Standard for Unified Sensor, Image, and Statistics Data Services

By Peter Baumann[1]

[1]School of Engineering and Science, Jacobs University, Bremen, Germany.

Motivation

In the modular geospatial and location-based services specification set of the Open GeoSpatial Consortium (OGC, http://www.opengeospatial.org), the foundation for handling coverages is laid down in the Web Coverage Service (WCS) standard (fig. 1); its current version 1.1.2 (Whiteside and Evans, 2008) was adopted by the OGC Technical Committee in April 2008. WCS offers basic services such as spatial and temporal subsetting, range (commonly also called band or channel) subsetting, scaling, and reprojection. Further services, such as Sensor Web Enablement (SWE) (Botts and others, 2006) and Web Processing Service (WPS) (Schut, 2005) build upon the coverage model introduced by WCS.

Actually, WCS is not just one specification, but a core of services with optional add-on services (extensions) that the implementer may offer. Among the extensions under construction are data format encodings, so-called "transactional" services for updating coverages (because WCS per se is purely focused on data retrieval), and a coverage processing extension that offers a processing request language with server-side evaluation.

In this contribution we present this processing extension—the Web Coverage Processing Service (WCPS) emerging standard (Baumann, 2008a,b). The author is co-chair of the WCS Working Group and the Coverages Working Group, and chair of the WCPS Working Group.

The OGC Coverage Model

OGC's conceptual model of a coverage is based on ISO 19123:2005 (International Organization for Standardization, 2005) and OGC Abstract Specification Topic 6 (Open Geospatial Consortium, Inc., 2007). Definitions in both of those documents are rather generic and include, for example, any currently known type of coverage. For practical reasons, the notion of a coverage in WCS refers only to gridded coverages (for the time being).

A coverage offered by a server consists of the following: (1) a locally unique identifier, (2) the value array itself, (3) its domain (a description of its spatial and temporal extent), (4) its range (the data type of the coverage's "pixel" elements), (5) a list of coordinate reference systems in which the coverage can be addressed, (6) an optional set of null values, (7) interpolation methods which can be applied to this coverage when needed by an operation, and (8) metadata (part of which are optional).

The WCPS Processing Language

WCPS offers an XQuery-oriented coverage request language for describing coverage manipulation. The design of the request language was guided by the experience gained in developing and formalizing database array query languages (Baumann, 1999).

The basic request structure consists of a loop over a list of coverages offered by the server, followed by an expression indicating the desired processing of each coverage, as shown below:

for var_1 in (coverage_1_1, coverage_1_2, ...), var_2 in (coverage_2_1, coverage_2_2, ...), ... [where filterPredicate(var_1, var_2, ...)] return processingExpr(var_1, var_2, ...)

For example, the difference between red and near-infrared channels in the coverage "ModisScene" encoded in a Tagged Image File Format (TIFF) file would be as follows:

for m in (ModisScene) return encode(abs(m.red - m.nir), "TIFF")

In the example above, the resulting TIFF file is returned to the client immediately. By using the "store()" function, a coverage result alternatively can be stored server-side for subsequent download by the client. The response in this case is a Uniform Resource Locator (URL, or Web address) under which the file is accessible. The following request implements this:

for m in (ModisScene) return store(encode(m, "TIFF"))

The list of operations allows for spatial and temporal subsetting, changing pixel values via arithmetic or further operations, scaling, reprojection, summarization, and derivation of new coverages (such as histograms or convolutions). The following example returns not a coverage, but a single number that represents the average of the squared differences between red and near-infrared channel of a Modis scene. Note that the result, one floating point number, does not need to be encoded:

for m in (ModisScene) return avg(sq(m.red - m.nir))

In practice, WCPS covers a range of statistics, images, and signal processing functionality. As an exercise, we have

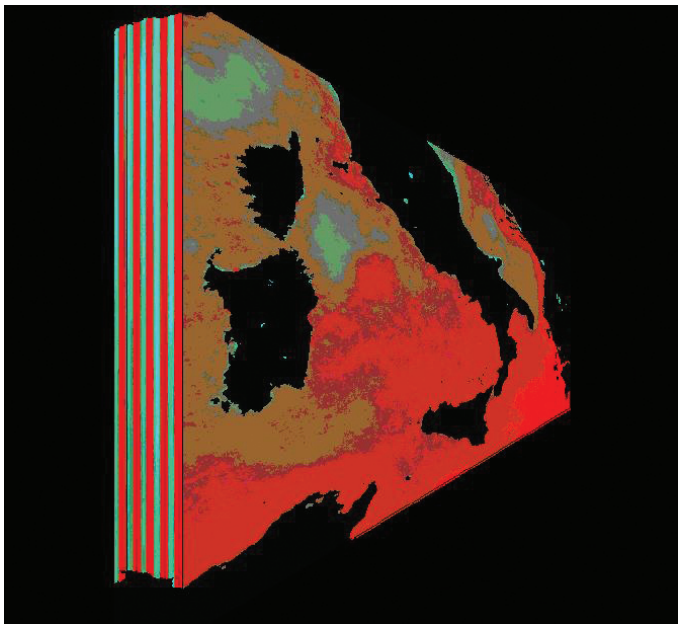


Figure 1. Web Coverage Service (WCS) extraction from a three-dimensional Advanced Very High Resolution Radiometer (AVHRR) satellite image. This image was retrieved from a time-series mosaic consisting of about 10,000 single images.

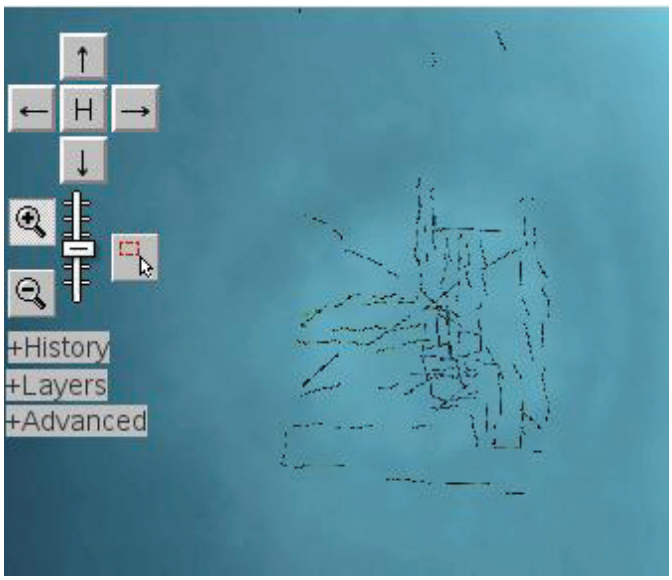


Figure 2. Web Map Service (WMS) access to the Hakon-Mosby underwater volcano area (one bathymetry layer, three submersible seafloor video mosaic layers)

implemented a Web Map Service (WMS) over WCPS. The language is "safe in evaluation," which means that no single request can ever block the server for unlimited time.

WCPS Implementation

The WCPS reference implementation is based on the multidimensional raster database management system "rasdaman" which stores multidimensional coverages of unlimited size in standard relational databases and adds a raster query language (similar to the standard database language SQL) for retrieving and manipulating them. Extensive server-side optimization (in particular, algebraic rewriting and hardware and software parallelization) is then performed. Next, raster objects are partitioned internally into tiles and are optionally compressed.

Rasdaman has been operational for several years and serves, for example, airborne image maps that are dozens of terabytes in size, three-dimensional geophysical data models, and four-dimensional climate simulation results. As proof of the concept, a Web Map Service (WMS) has been implemented on top of WCPS (fig. 2). An online demonstration is available at http://www.earthlook.org.

WCPS Service Embedding

The WCS Processing Extension (Baumann, 2008b) embeds WCPS into the WCS suite by defining an additional request type, ProcessCoverages, which specifies how clients can send WCPS requests to a server and how they receive the processing results.

Outlook and Future Work

Now that WCS has become a stable and mature specification, the standardization group's work is now concentrated on modularization and adding optional extensions to accommodate the different user communities. For example, atmospheric researchers from OGC's Geo-interface to Atmosphere, Land, Earth, and Ocean netCDF (GALEON) Network (see http://www.ogcnetwork.net/galeon) are actively contributing.

Among the extensions planned or already under development are (1) extensions to allow the incorporation of irregular grids (currently WCS, and therefore also WCPS, only consider regularly gridded or raster data), and (2) extensions that allow for any number of spatial and temporal dimensions (currently only two, three, and four dimensions can be considered). All in all, work is plentiful, hence, experts in the fields of, for example, computers or earth sciences are invited to join the standardization group and contribute to shaping the future standards of our communities.

References Cited

Baumann, Peter, 1999, A database array algebra for spatio-temporal data and beyond, *in* Proceedings, Fourth International Workshop on Next Generation Information Technologies and Systems (NGITS '99): Lecture Notes on Computer Science, v. 1649, p. 717.

Baumann, Peter, ed., 2008a, Web Coverage Processing Service (WCPS), version 1.0.0: Open Geospatial Consortium, Inc., document 08-068.

Baumann, Peter, ed., 2008b, Web Coverage Service (WCS) Processing Extension, version 1.0.0: Open Geospatial Consortium, Inc., document 08-059r2.

Botts, M., Robin, A., Davidson, J., and Simonis, I., eds. 2006, Sensor web enablement architecture: Open Geospatial Consortium, Inc., document 06-021r1.

International Organization for Standardization, 2005, ISO 19123:2005—Geographic information—Schema for coverage geometry and functions: Geneva, Switzerland, International Organization for Standardization, 65 p.

Jerosch, K., Ludtke, A., Schluter, M., and Ioannids, G.T., 2006, Automatic content-based analysis of georeferenced image data—Detection of Beggiatoa mats in seafloor video mosaics from the Hakon-Mosby mud volcano: Computers and Geosciences, v. 33, no. 2, p. 202–218.

Open Geospatial Consortium, Inc., 2007, Abstract specification topic 6—The coverage type and its subtypes: Open Geospatial Consortium, Inc., document 07-011, 67 p.

Schut, Peter, 2005, Web Processing Service Implementation Specification, version 1.0.0: Open Geospatial Consortium, Inc., document 05-007r7, 87 p.

Whiteside, A., and Evans, J.D., eds., 2008, Web Coverage Service Implementation Standard, version 1.1.2: Open Geospatial Consortium, Inc., OGC document 07-067r5, 133 p.

Orchestrating Grid-Computing-Enabled Web Processing Services

By Bastian Schaeffer[1] and Bastian Baranski[1]

[1]Institute for Geoinformatics, University of Münster, Münster, Germany.

Existing spatial data infrastructures (SDIs) are mainly focused on data retrieval, data processing, and data visualization. An SDI based on open standards, (for example, those of the Open Geospatial Consortium (OGC)) mostly supports the retrieval and visualization of data through Web services; however, the geospatial-data processing (otherwise known in the SDI community as "geoprocessing") is normally performed by humans with more or less proprietary and monolithic geographic information systems (GISs). With growing network capacity and processing power, some efforts were made to integrate stand-alone geoprocessing applications and their expert functionality into a Web service environment and, therefore, enable Web services to execute geoprocessing tasks. The OGC's Web Processing Service (WPS), which became an official standard in late 2007, is a major attempt to address this issue in a standardized way. The WPS specification defines a standardized interface to publish and perform geospatial processes over the Web. Such a process can range from a simple geometric calculation (for example, a simple "intersect" operation) to a complex simulation process (for example, creating a global-climate-change model). To speed up the processing of large amounts of data and perform complex calculations (for example, when doing a weather forecast simulation), the use of grid computing or related methods and technologies is a good choice for achieving high calculation performance, for improving service availability, and for guaranteeing a different quality of service.

Even though grid computing or distributed computing is not a new approach, only some research has been done on combining OGC's Web service (OWS) standards implementations in such a manner. The intrinsic complexity of geospatial data (also known as "geodata"), however, often requires the use of several processing steps to address a given problem. Grid computing can be applied to one dimension to improve the performance of each step, but to automate the whole business process, orchestrated geoprocessing workflows have to be built. This would create a second dimension and enable the creation of high-speed, fully automated, value-added geoprocessing workflows.

This paper will give an introduction to the new OGC WPS standards and will present an approach on how to combine this specification with grid computing. This approach will be combined with an introduction to the orchestration of these grid-based WPS in workflows.

Finally, the presented approaches will be validated by means of a real-world scenario. A geoprocessing workflow will be presented that solves a given problem in the field of fire protection in southern Spain. The Business Process Execution Language (BPEL) will be used to design this workflow, which will contain some grid-based WPS Web Services. Modeling, execution, and the visualization of results will all be performed in a single integrated environment and will prove the usefulness of these new approaches.

Building a Geospatial Web Portal Based on Service-Oriented Architecture

By Peisheng Zhao,[1] Liping Di,[1] Weiguo Han,[1] Yaxing Wei,[1] and Xiaoyan Li[1]

[1]Center for Spatial Information Science and Systems, George Mason University, Greenbelt, Md.

Geoscience research and applications often involve analysis of a large volume of geospatial data. Traditionally, scientists spent a lot of time installing and learning a variety of software on local machines, searching for and collecting the data from various sources, and preprocessing and analyzing the data on local machines. This "everything-locally-owned-and-operated" paradigm makes the analysis and application of geospatial data very expensive and time-consuming. Recent advances in Service-Oriented Architecture (SOA) are shifting the geospatial data and analysis from the "everything-locally-owned-and-operated" paradigm to the "everything-shared-over-the-Web" paradigm (a "Web-and-service-centric" paradigm). Currently, a

significant number of geospatial datasets are available that use Web services, such as the Web Coverage Services (WCS) of the Open Geospatial Consortium (OGC) (Whiteside and Evans, 2008). These services perform various geospatial analysis functions. By embracing geospatial content and capabilities within the context of the SOA, we have developed a SOA-based geospatial Web portal, which is a fully extensible portal system for discovering, retrieving, analyzing, and visualizing geospatial data and other types of data obtained from networks. The most distinguished characteristic of this portal is that it is designed to use distributed software, hardware, and applications to provide services from a number of different sources, thereby enabling the following: (1) a single point of access to geospatial information and processes over the Web, (2) Web-service-based analysis in which all functions are provided through interoperable Web services, and (3) user customization and collaboration by integrating and chaining together user-specified Web services.

The SOA uses loosely coupled and interoperable Web services to implement system requirements. All the services within the SOA are independent with self-described interfaces so that they can be accessed in a standard way without knowledge of how the service actually performs its tasks. Moreover, the SOA can support the integration and orchestration of distributed services into a composite service. The presented portal includes four main components: Web portal, catalog service, data service, and Web-processing service.

The Web portal follows the model-view-controller (MVC) design pattern, which is a commonly used software engineering architecture, to implement the following:

- User portal—Storing the current state of the portal in an OGC Web Map Context (WMC) (Sonnet, 2005) document that can be imported again later to restore the portal's state.
- Data management—Retrieving geospatial data from a remote service and temporarily storing them on the map server in a network-accessible location.
- Data analysis—Selecting or integrating a preferred processing service to perform data analysis.
- Workflow—Allowing the user to build a chain of services to perform a task.
- Data visualization—Allowing the user to set up preferences for displaying the data, such as the building a sequence of coverages, deciding which subsets of data to show, and creating an image palette.

Data services provide users with a common data environment in which they can use data in an interoperable manner. OGC's WCS provides intact multidimensional and multitemporal geospatial data as a "coverage" in order to meet the requirements of client-side rendering, scientific model inputs, and other clients beyond simple viewers. OGC's WFS (Web Feature Service) (Vretanos, 2005) supports the networked exchange of geographical vector data as "features" encoded in Geographic Mark-up Language (GML). OGC's Web Map Service (WMS) (De la Beaujardiere, 2006) provides geospatial data as a "map," which is generally rendered in spatially referenced pictorial image formats (such as PNG, GIF, or JPEG) that are dynamically generated from real geographical data.

In distributed computing environments, a catalog service plays the role of "directory" in helping with the registration and discovery of data and services. The OGC's Catalog Service for Web (CSW) (Nebert and others, 2007), an Electronic Business Registry Information Model (ebRIM) profile for Web-based geospatial catalog services, has been implemented to register, discover, and access a wide variety of distributed resources (for instance, geospatial data, applications, and services). CSW searches distributed catalog services such as the Group on Earth Observations' Global Earth Observation System of Systems (GEOSS) Clearinghouse, the National Aeronautical and Space Administration's (NASA) Earth Observing System Clearinghouse (ECHO), and George Mason University's GeoBrain catalog. With regard to the geospatial data metadata descriptions in ISO 19115 (International Organization for Standardization, 2003), the CSW makes some further extensions to accommodate the ISO model. The class "Dataset" has been added to the ebRIM to provide a flexible way to describe network-accessible data. The CSW supports a variety of classification methods to enable the service publisher to indicate the domain to which a service belongs at publication time, including the definitions from OGC specifications, ISO 19119 (International Organization for Standardization, 2005), and NASA's Global Change Master Directory (Olsen and others, 2007).

A Web Processing Service (WPS) provides a domain-specific computational model, which might be a simple spatial calculation or a complex global climate-change model, to enable users to perform data analysis over the network. For manipulating and analyzing vector and raster geospatial data, the Web portal we present here provides more than 20 built-in WPSs and more than 50 relevant operations that are based on Open Source Geospatial Foundations's Geographic Resources Analysis Support System (GRASS). The WPSs can also be chained together to perform more complex analysis task. Users can access these services to perform data analysis and data mining of any OGC-compliant online data source. Moreover, the portal is able to integrate new Web services dynamically in order to provide users with an open and fully extensible environment. If a user has his or her own geospatial processing service and would like to use it to perform data analysis, he or she can integrate that service into the portal to build a unique portal. If the service is registered into the catalog service, other portal users will benefit from it. Hence, the more users are involved, the more powerful the portal becomes.

References Cited

De la Beaujardiere, Jeff, ed., 2006, OpenGIS Web Map Server Implementation Specification: Wayland, Mass., Open Geospatial Consortium, Inc., document 06-042, 85 p., available only online at http://www.opengeospatial.org/standards/wms/. (Accessed August 25, 2008.)

International Organization for Standardization, 2003, ISO 19115—Geographic information—Metadata: Geneva, Switzerland, International Organization for Standardization, 140 p.

International Organization for Standardization, 2005, ISO 19119—Extensions of the service metadata model: Geneva, Switzerland, International Organization for Standardization, 4 p. [Amended in 2008.]

Nebert, Douglas, Whiteside, Arliss, and Vretanos, Panagiotis, eds., 2007, OpenGIS Catalogue Services Specification: Open Geospatial Consortium, Inc., document 07-006r1, 218 p., available only online at http://www.opengeospatial.org/standards/cat/. (Accessed August 25, 2008.)

Olsen, L.M., Major, G., Shein, K., Scialdone, J., Vogel, R., Leicester, S., Weir, H., Ritz, S., Stevens, T., Meaux, M., Solomon, C., Bilodeau, R., Holland, M., Northcutt, T., and Restrepo, R.A., 2007, NASA/Global Change Master Directory (GCMD) earth science keywords, version 6.0.0.0.0: Greenbelt, Md., National Aeronautics and Space Administration, available only online at http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html/. (Accessed August 25, 2008.)

Sonnet, Jerome, ed., 2005, Open Geospatial Consortium, 2005a, Web Map Context Implementation Specification: Wayland, Mass., Open Geospatial Consortium, Inc., document 05-005, 30 p., available only online at http://www.opengeospatial.org/standards/wmc/. (Accessed August 25, 2008.)

Vretanos, P.A., ed., 2005, Web Feature Service Implementation Specification: Wayland, Mass., Open Geospatial Consortium, Inc., document 04-094, 131 p., available only online at http://www.opengeospatial.org/standards/wfs/. (Accessed August 25, 2008.)

Whiteside, Arliss, and Evans, J.D., eds., 2008, Web Coverage Service (WCS) Implementation Standard: Wayland, Mass., Open Geospatial Consortium, Inc., document OGC 07-067r5, 133 p., available only online at http://www.opengeospatial.org/standards/wcs/. (Accessed August 25, 2008.)

Coming of Age—The Positive Legacy of Free and Open-Source Software Geographic Information Systems

By Peter Löwe[1]

[1]German Research Center for Geosciences, Potsdam, Germany.

Introduction

Software projects evolve during their development and usage lifecycles through various stages. For successful projects, this evolution results in improvements regarding code quality, reliability, and performance. Because of these processes, projects which were started in the past can be considered as "legacy" with respect to the very latest information technology approaches. Usually the term "legacy" is used in a derogative way. Several Free and Open-Source Software (FOSS) projects have had several iterations of these cycles and could be rightfully addressed as "legacy" while they still serve their intended purpose. The example of the Geographic Resources Analysis Support System (GRASS), which is a free and open-source geographic information system (GIS), is used in this presentation to showcase a software project that has a positive legacy. This kind of legacy is a valuable asset for software development. In this type of system, "cutting edge" technologies are incorporated once those technologies have withstood the test of time; at the same time, code functionality is ensured by the careful maintenance of existing code. We will show how the rise of integrated development environments (IDEs) will help to ensure the future: saving projects by employing a gradual yet strategic increase in the developer base.

Perceptions of Legacy

The online, user-written encyclopedia Wikipedia (http://www.wikipedia.org) describes two ways in which the term "legacy" can be applied to "software," two terms which seem to be almost mutually exclusive: On the one hand (from an innovation-friendly point of view), application programs are considered "legacy" when they continue to be used because the user does not want to replace or redesign them. On the other hand (from a problem-solving perspective), the term "legacy" can be applied to software that is actually performing its tasks effectively even for very large amounts of input data, which indicates a fully developed tool, which then makes the term sound more positive. We will show that when it comes to software development, there are at least two other factors related to the term "legacy."

The Legacy of GRASS GIS—Alive and Kicking

The GRASS GIS project has been evolving for more than two decades. The project was started by the U.S. Army Corps of Engineer's Construction Engineering Research Laboratories (CERL) and lasted from 1982 to 1995. Afterwards, the code was handed over to academia, where a new phase of development began. Licensing under the general public license (GPL) in 1999 resulted in a dramatic increase in its development. Over the years, the functionality evolved from a raster-based GIS onwards to include floating point operations, a topology-based vector model, volume support, and, finally, the inclusion of Open Geospatial Consortium-based services. Apart from its traditional use as a desktop GIS, GRASS is also used as part of the backend of other applications, such as Quantum GIS (QGIS) and the Java Geographic Resources Analysis Support

System (JGRASS) (developed using the uDig system). Both of these open-source GIS tools provide easy-to-use graphical user interfaces (GUIs) for standard desktop GIS tasks such as data queries or map production. They shield the user from most of the underlying complexity of GRASS GIS. For Internet-based mapping, GRASS can be used with the University of Minnesota's MapServer (UMN Mapserver) or as a standalone OGC-based Web Processing Service (WPS) such as George Mason University's GeoBrain Online Analysis System (GeOnAs) or pyWPS (a Python-language WPS available through Wald Intevention). These tools allow Web Mapping Services (WMS) or Web Processing Services (WPS) (for custom data processing) to be provided to their end-user communities; both services are available through the World Wide Web (WWW).

Coding Paradigm

The highly modular code paradigm follows the approach of (scriptable) Unix shell commands. The reference code base, which is hosted by the GRASS servers, consists of more than 300 ANSI C-modules, with additional add-on scripts and modules that are independently provided and hosted by user communities. The current version consists of about 500,000 lines of source code. "Write access" to the code repository has been controlled since 2006 by the GRASS Project Steering Committee.

Serial Code Development

A relatively small multinational group of constant contributors voluntarily adds new features and updates old ones. The majority of the GRASS software-development community is highly fluent in the code structure and uses text-based tools for development such as the text editors "vi" or "emacs." For newcomers, this situation results in a steep learning curve. Additionally, the constant overall development of libraries forces the contributors of add-on C-modules to adapt their code. As a consequence, many C-based add-on-modules become defunct over time, while shell-style scripts remain usable because the C-module interfaces remain unchanged despite the internal changes.

Arguing the Case for Extended IDE Usage

During the last several years, IDEs became widely available for code development. They allow for easy navigation in large code repositories and collaborative development. Many programmers consider the availability of IDEs as a given; therefore, it makes sense to apply the know-how regarding GRASS-development to IDEs, such as the Eclipse C/C++ Development Toolkit (CDT) (fig. 1). Also, due to the availability of code-tracking and refactoring tools, add-on modules can be much more easily updated to the latest standards by their developers. The decision of the GRASS GIS community to stick to the C-language code, in the spirit of a positive legacy, allowed for the successful deployment of a native version of GRASS for Microsoft Windows. Because IDEs such as Eclipse CDT are platform independent, active code can be developed on non-Linux systems, such as Microsoft Windows. With respect to the "dark side" of legacy, such as the pending retirements of the current pre-IDE developers and the resulting loss in knowledge and skills, it is even more important to enable IDE-based development to help document the code while there is time.

An IDE can be used as a convenient front end for source code development, maintenance, and building binary executable files. The issue of software legacy can be characterized by two distinctive aspects: (1) the overall quality of the source code itself, including knowledge preservation by means such as comments, and (2) the mechanisms ("toolchains") required to derive executable files, which are executed according to control files (such as a "makefile" for C/C++ programming). For the latter, GRASS GIS relies on a custom makefile structure. Standard approaches such as autoconfig/automake for the configure/make/install toolchain (or cmake as an alternative) are not supported (yet). This can be perceived as a case of positive legacy (stemming from the "if it works, don't fix it" approach), yet this approach could create a bottleneck for platform-independent development. Fortunately, the toolchain based on a custom makefile for building GRASS GIS binary executable files can be encapsulated in order to be used within Apache's Ant toolchain. Ant is commonly used for Java applications and is part of the Eclipse IDE. Use of Ant allows developers to manage the whole software development process of checking out the latest sources from the SVN, editing the code, and configuring and building the executable files in one IDE, thereby making the IDE truly platform independent.

Conclusion

Over the last two decades, GRASS GIS has proven to be a geographic information system for professional use. It continues to grow and is very much alive. The development approach of the community is basically conservative, yet it integrates additional technologies once they are considered mature and relevant. The usage of IDEs is expected to extend and rejunevate the developer community. IDEs will also broaden and speed up the development process because they enable active development non-Linux systems. The use of IDEs is an example of the positive, stabilizing effects of trusted software which "just works well" and can continue to be developed on various platforms.

Outlook

A tool to describe processing chains would be desirable to in order to document and manage the community-inherent knowledge about how to enable GRASS modules to cope with complex tasks. Usually, every task can be accomplished in various ways, some of them more efficient than others.

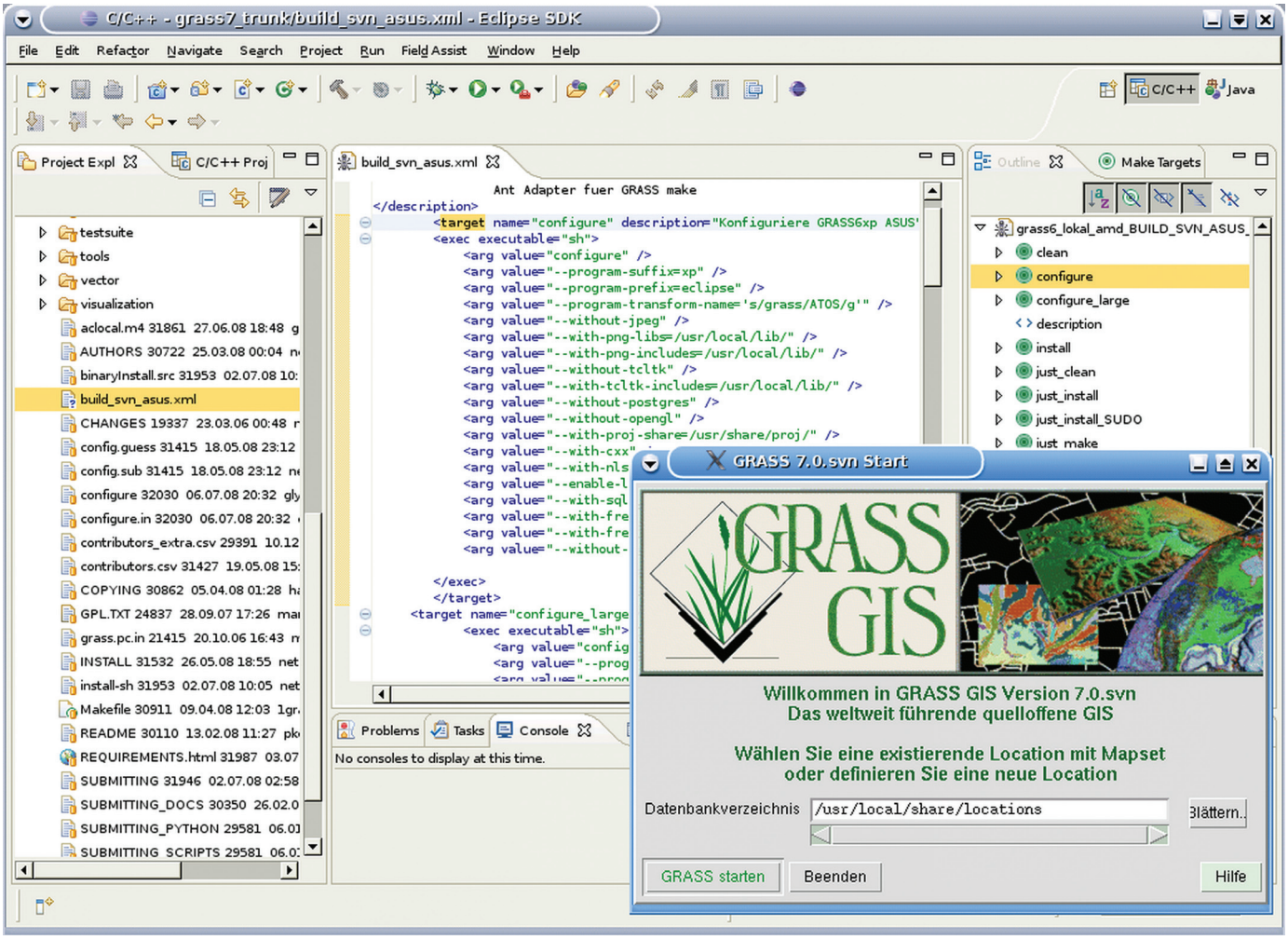


Figure 1. Screen capture showing the use of the Eclipse IDE with the C/C++ Development Toolkit, which provides an integrated development environment (IDE) for the C/C++ programming language. The view shows that the current Geographic Resources Analysis Support System (GRASS) geographic information system (GIS) source code has been downloaded from the project's software source code repository (based on CollabNet's Subversion (SVN) tool for software versioning) and has just been compiled via the Apache-Ant-wrapped GRASS-building chain. The resulting binary executable file has just been invoked as the last step of the processing chain.

Although knowledge of the solutions remains only marginally archived, the loss of experienced members from the user community will result in massive losses of knowledge and skills. A likely tool for helping the community retain knowledge and skills is CyberIntegrator (CI, developed by the University of Illinois' Image Spatial Data Analysis Group), which is a workflow-based system that supports interactive workflow creation, connection to external data and event streams, provenance tracking, and incorporation of workflow fragments and functionality from other systems and applications. Trials for the integration of CI and GRASS GIS are underway. Once such tools have been tied into the GRASS GIS environment, the knowledge and skills behind this community-driven FOSS Geoinformatics project will be saved and its continued development ensured for years to come.

Data Integration Using the Image Grand Tour

By Bradley C. Wallet[1] and G. Randy Keller[1]

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

Introduction

Geoscience, like many disciplines, is experiencing an overabundance of data. A geoscientist often has access to a large variety of data including gravity anomalies, magnetic anomalies, digital elevation grids, laser altimetry, multispectral imagery, geological maps, and seismic velocity images. The challenge is to take these varied and often disparate sources

and to integrate them together in a manner that increases knowledge and understanding of the underlying structures and processes in the Earth.

Visualizing the integrated sources is inherently difficult due to the high dimensionality of the data. When visualizing the data as an image, there are a number of different color models that will support the display of multidimensional spatial data (the most common model being red-green-blue, or RGB); however, most color-display models only allow for a maximum of three different components. Although there has been work in visualizing more than three dimensions of spatial information, these newer approaches generally add only a few additional dimensions. Furthermore, the increased ability to visualize dimensions comes at a cost of increased complexity and difficulty in analysis and understanding.

Dimensionality Reduction

A common method of addressing the difficulties of visualizing multidimensional data is to reduce the dimensionality by methods such as linear projections. Perhaps the most popular method of doing this is principal component analysis (PCA). PCA is a common technique in remote sensing and is often applied to multispectral imagery. Guo and others (2006) applied it to spectral decomposition of seismic data with excellent results; however, PCA has some serious limitations as its goal is to maintain the maximum variance in the projected data. The reliance upon variance is troubling in many remote sensing applications because the feature of interest may only be a small portion of the dataset. In such a case, the variance of the overall dataset is likely to be dominated by the background or noise. As such, PCA may be optimally the wrong method because it may maximize the noise. Additionally, the global nature of PCA means that local dependencies are generally ignored in favor of global trends. The most interesting limitation of PCA is the lack of involvement of the implied spatial nature of structure in spatial data. Organization of the data in the spatial view of the data is not considered in PCA.

Image Grand Tour

The Image Grand Tour (IGT) is an interactive method for reducing dimensionality when the data are spatial in nature and can be visualized as an image. The IGT involves defining a smooth trajectory in the set of all possibly linear projections. This trajectory may be chosen according to a number of criteria, including denseness or maximum coverage in a finite time. The data are then viewed in image form in a smooth manner. The result is a form of "data movie." Geoscientists can then visualize the data from all possible angles and look interactively for interesting views that offer insight into the data. The IGT has been applied to a number of areas, including medical imaging, land mine detection, and multispectral imaging. Wallet and Marfurt (2008) demonstrated the value of applying the IGT to interpreting spectral decomposition information from seismic data. They demonstrated that the IGT yielded information that was not readily apparent using PCA alone. In doing this, they constructed single, grayscale views that showed features that increased their understanding of the data.

Application

We illustrate the IGT technique by applying it to a land seismic survey acquired over south Texas in the United States. Spectral decomposition was calculated on the seismic volume resulting in 85 spectral components ranging from 5 cycles per second (hertz, or Hz) to 90 Hz. We recognized a horizon containing a fluvial-deltaic system and flattened the data. We then applied PCA to these 85 images and retained the first eight PCA images (fig. 1). We noted that several channel features can be seen in the first six principal components, or eigenspectra, while only random noise (either geological or seismic) can be seen in eigenspectra 7 and 8.

Examining other eigenspectra revealed that significant information was present in images other than the first three (fig. 1); however, by the time the seventh eigenspectrum was examined, little information appeared to be visible. We thus chose to run the IGT using the first six eigenspectrum because they appeared to capture all of the values in the dataset.

Running the IGT revealed structures that were not readily apparent when examining just the first three eigenspectra. We stopped our tour any time we identified a feature of geologic interest. The tops of figures 2*A* and 2*B* show the six coefficients applied to each eigenspectra at the current tour location. Several small channels appeared that were difficult to see in figure 1. Figure 2*A* presents various meandering channels, several of which were not clearly visible, or were absent, in the first three eigenspectra. Figure 2*B* clearly shows a distinct view of a single meandering channel. Although this channel was visible in previous images, this view provides better, more localized views of the channel edges. These results illustrate the value of combining information gathered from multiple tour projections.

Conclusions

The IGT is a valuable method for interactively integrating multiple data sources. Because the geoscientist controls the projections, the process is geared towards views that are interesting from a heuristic definition rather than predefined criteria related to variance. Using the IGT, it is possible to construct views of data that reveal insight that is not apparent using other methods.

Acknowledgments

We thank Anadarko Petroleum for permission to use its seismic data in this study. This study was partly supported by the National Science Foundation via the Geosciences Network (GEON) project.

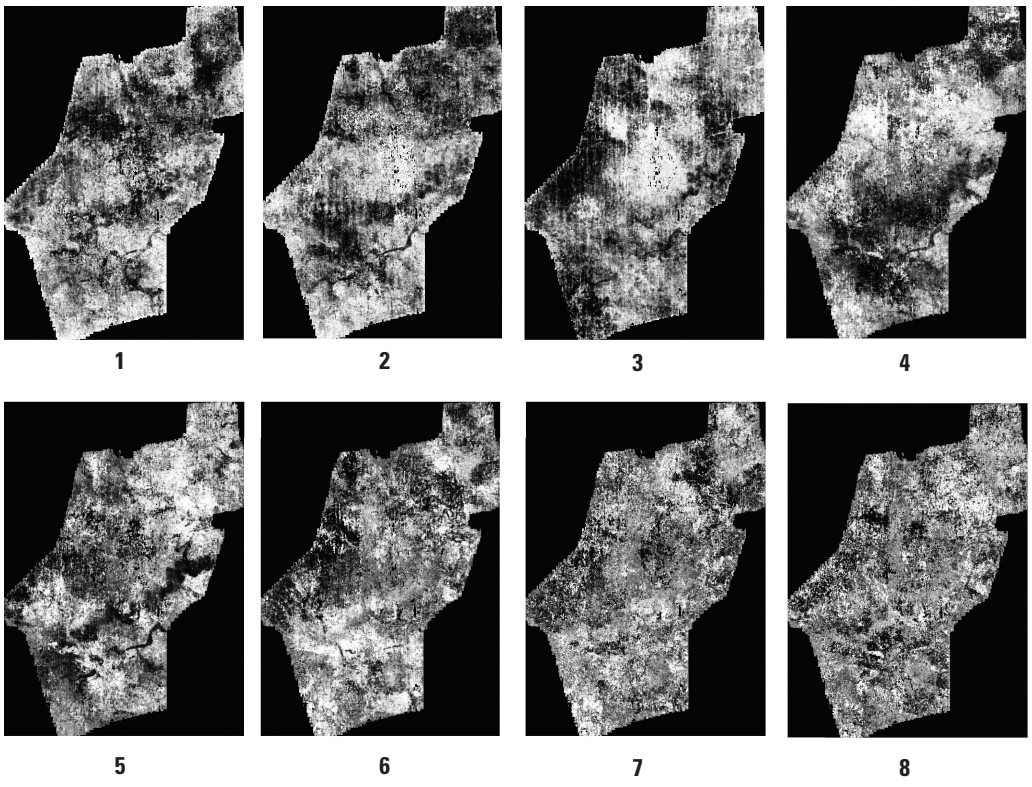


Figure 1. An example of output from a principal component analysis (PCA) that was input to the Image Grand Tour from a portion of a seismic survey from south Texas. The first eight eigenspectra (principal components) represent the vast majority of the energy in the 85 spectral components. North is oriented towards the top. Note that different channel features appear stronger in different components. Eigenspectra 7 and 8 show very little channel information. Eigenspectra 2 and 3 are quite sensitive to the north-south-trending acquisition footprint.

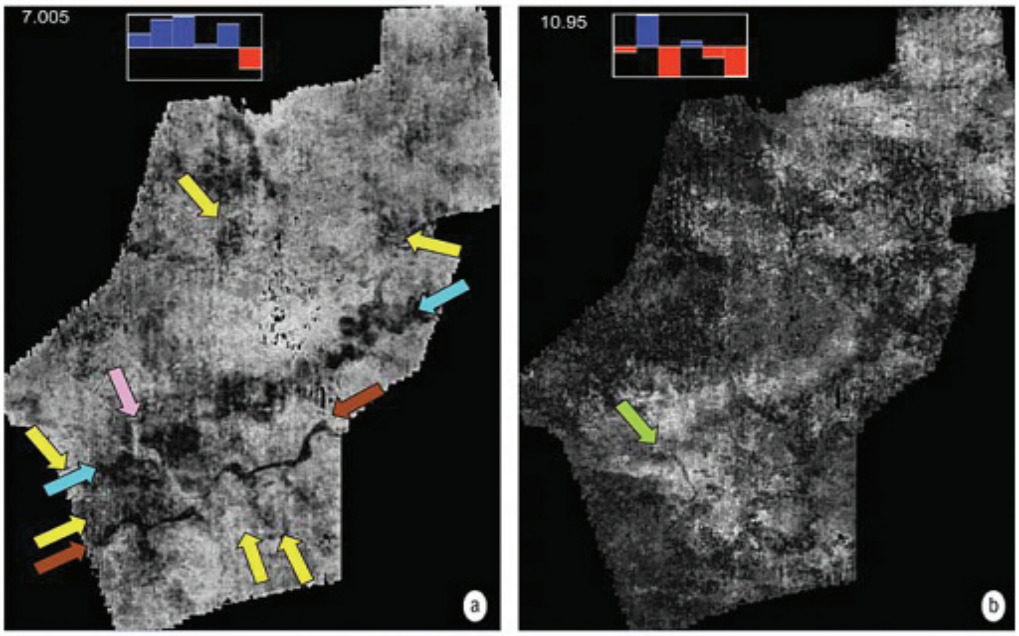


Figure 2. Image Grand Tour (IGT) projections. *A*, An IGT projection that illuminates very narrow meandering channels (yellow arrows) and a northwest-trending major channel that corresponds to deeper valley fill (magenta arrow). *B*, An IGT projection that is generally featureless except for the channel indicated by the green arrow.

References Cited

Guo, H., Marfurt, K.J., Liu, J., and Dou, Q., 2006, Principal components analysis of spectral components: Society of Exploration Geophysicists Expanded Abstracts, v. 25, p. 988–992.

Wallet, B.C., and Marfurt, K.J., 2006, A grand tour of multispectral components—A tutorial: The Leading Edge, v. 27, p. 334–341.

A Collaborative Environment for Climate Data Handling

By Stephan Kindermann[1] and Martina Stockhause[2]

[1]German Climate Computing Center, Hamburg, Germany.

[2]Max Planck Institute for Meteorology, Hamburg, Germany.

The Collaborative Climate Community Data and Processing Grid (C3Grid) project, which began in September 2005 as part of the German grid initiative (D-Grid), is currently building a climate-data handling infrastructure to support scientists in finding, analyzing, processing, and sharing climate datasets. The purpose of the infrastructure is to track the entire data cycle, from the discovery of input data to the publication and archiving of the results.

The approach consists of three layers: common data discovery layer, data access layer, and data manipulation layer:

- Data discovery is based on harvesting ISO 19139 (International Organization for Standardization, 2007) metadata descriptions into a central metadata catalog.
- The underlying complexities of data access, provider-specific data extraction, and pre-processing steps are hidden by a simple Web service interface.
- Data processing can be triggered in a collaborative grid environment that provides computing resources and both short- and long-term data storage components.

There are three main problems the project encountered while building the infrastructure:

1. Establishing a consistent security layer—On the one hand, there is a clear need for a federated data authentication and authorization (AA) infrastructure, which prevents user identity and role management outside of the user's home organization. On the other hand, the collaborative environment uses grid technology with a public key infrastructure (PKI) and a virtual organization- (VO-) based AA. The C3Grid approach merges Shibboleth's Security Assertion Mark-up Language (SAML) information with grid- and VO-based AA mechanisms. After a prototype phase with plain Web services, C3Grid is now moving toward the implementation of Web-service resource framework (WSRF) services.
2. Generation and quality control of ISO 19139-compliant metadata—The discovery of information currently is handled mostly by large data providers. Yet, substantial support is necessary to allow (1) small data providers to join the infrastructure and (2) workflow providers to generate appropriate metadata for the resulting data (including data provenance information). Therefore, within C3Grid, tools were developed to provide automatic data provenance tracking, semiautomatic data archiving, and quality checks of discovery metadata.
3. Enable flexible but modular processing—Complex scientific workflows, which are composed of predefined modules, need to be supported by the infrastructure. Support requires both the availability of sufficient workflow descriptions and additional description-of-use details for the data. The handling and generation of such metadata can be built on collective experience and developed tools, but also needs additional agreements and information services.

In general, a major challenge in the project is to find or develop legal agreements that reflect an elaborate balance between technical progress and manageable effort. The established data and computing-service providers want to re-use their current implementations in order to minimize the maintenance of their software and the labor required to adapt to changes that are necessary when building the infrastructure. Yet, integrating collaborative environments always requires the creation of prototypes and the adoption of not-yet-established technologies. Different technological pathways have to be merged with respect to the specific needs of the existing scientific community and the future needs of intercommunity cyberinfrastructures. In this presentation, the key experiences and design decisions within the C3Grid project are discussed.

Reference Cited

International Organization for Standardization, 2007, ISO/TS 19139:2007—Geographic information—Metadata—XML schema implementation: Geneva, Switzerland, International Organization for Standardization, 111 p.

Globalization of Geoscience Information—Developing Collaboration to Sustain Growth (Keynote)

By Kristine E. Ch. Asch[1]

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

Universities, geological surveys, and other agencies around the world have similar goals and activities, issues and challenges in the field of geoscience information. The rapid development of technology (especially Web services), the massive explosion of data, and the increasingly diverse requirements of users across all sectors (governments, scientists, commerce, and the public) introduce new demands and challenges. Add to this the global and transnational issues—sustainable energy resources, mineral resources, agriculture, ground water, transport, catastrophic natural hazards such as earthquakes and tsunamis, and last (but not least) climate change—where geoscience has a critical role to play and it becomes obvious that there is a pressing need for us to work

together to sustain growth. Although the field of geoscience informatics provides many good examples that exploit developments in information technology to further global cooperation, we still have a long way to go in efficiently sharing information, experience, and expertise so we can analyze and synthesize data and knowledge across political and continental boundaries with a minimum of barriers and work in international teams on research projects that add value.

There already are some excellent examples of collaboration: (1) global initiatives such as the U.S. Environmental Protection Agency's Global Earth Observation System of Systems (GEOSS), the European Information Services for Environment and Security's Global Monitoring for Environment and Security (GMES) project, the European Union's Infrastructure for Spatial Information in the European Community (INSPIRE), and the International Union of Geological Science Commission for the Management and Application of Geoscience Information's (IUGS-CGI's) Geoscience Mark-up Language (GeoSciML); (2) cooperation between governmental and nongovernmental organizations such as the European Geoscience Unions and the Commission of the Geological Map of the World (CGMW); and (3) specific international projects such as OneGeology and Australia-based eWater. Several of these initiatives and groups have made good progress and the challenges above are very much part of their discussions and agendas; however, significant questions remain: How effective are the current collaborations? Given the enormity of the challenges listed in the first paragraph, can we do better? This presentation will use the approach of a "Strengths, Weaknesses, Opportunities, and Threats (SWOT)" analysis to examine that question and, within that analysis, closely examine two specific examples: IUGS-CGI and its development of GeoSciML, and the European Union's INSPIRE.

Strengths and Opportunities

The strengths of the actors (participants) in this domain include our rich and diverse expertise, different viewpoints, extensive datasets, state-of-the-art computing applications, and access to high-performance computing. There are existing projects, networks, and collaborations that provide a foundation for future integration.

The opportunities are considerable because information and informatics are what joins the sciences together and there has never been a more opportune need for multidisciplinary science. The wish to bring together and integrate data means that there are opportunities for the development of standards for interoperability and harmonization. Some of these strengths and the opportunities will be illustrated by reviewing the work on GeoSciML and INSPIRE. GeoSciML is being taken forward by the IUGS-CGI through its active working group on interoperability. INSPIRE is an example of the positive growth of spatial data infrastructures; ultimately, it will create an infrastructure for environmental spatial data across Europe that will (1) set forth data discovery methods and data and network specifications, and (2) enable sharing of both data and systems.

Weaknesses and Threats

In spite of our successes, we must acknowledge weaknesses, too. Because there are so many actors, it is difficult to know just who is working in our domain and what they are doing. Can we identify and map them? The different cultures and viewpoints of these actors and their different goals often lead to "re-inventing the wheel," a duplication of effort, and ultimately, a defense of territory. A subsidiary question must be asked as well: Does the academic reward system, with its emphasis on individual merit and esteem through papers and citation indexes, actually discourage collaboration? Finally, the geoscience information domain has many different actors, but there is little or no clear collective vision or clear leadership behind which all these actors can unite. Can we (or indeed, do we want to) change that which is now, essentially, a liberal system?

Perhaps the most significant threat is inertia—we do nothing and all the weaknesses persist. We do take advantage of the potential opportunities, and so our branch of the sciences remains marginal. Some may also see it as a threat that, unless we (academics, and governmental and nongovernmental agencies) take a lead, commercial companies and the market will move on and ignore us and our work.

The Results of the SWOT Analysis

So what do we need to do to make the collaboration more effective (that is, what are our opportunities)? We can begin by addressing our weaknesses and exploiting our strengths. One of these weaknesses, and perhaps the one that needs to be addressed first, is to simply understand who is doing what in our domain. There are an amazing number of groups and initiatives in the area of geoscience information and geoinformatics. With each meeting like this, we seem discover more new initiatives, groups, and individuals who are working in the same fields and on the same problems. Second, we need to embrace more enthusiastically the work on spatial data infrastructures (SDIs), which is work ignored in some scientific quarters because it is lead by geographers, not earth scientists. Third, the major scientific unions need to take a stronger lead in their involvement with global projects such as GEOSS, and they need to learn to interact more with regional political bodies like the European Commission. Many in our field feel there is a need for these organizations to give geoscience informatics a much higher priority and profile than they have to date. Without the lead of these organizations to provide the "mutual" discovery and "glue money" (funding for meetings, publications, and others means of sharing information), then we will make much less progress than we deem is necessary or society expects of us.

Our greatest opportunity for improving the situation lies in our own hands. That means that we in the geoscience information and informatics community should proactively look to develop collaboration that transcends the different sectors of our own domain (academia, geological surveys, and agen-

cies), use our transferable technology and techniques to extend beyond that into the different sciences, and last (but not least) work even more intensively across political boundaries.

The German Research Center for Geosciences' Information System and Data Center—Portal to Geoscientific Data, Information, and Knowledge

By Bernd Ritschel,[1] Vivien Mende,[1] Hartmut Palm,[1] Lutz Gericke,[1] Sebastian Freiberg,[1] Ronny Kopischke,[1] and Christian Bruhns[1]

[1]German Research Center for Geosciences, Potsdam, Germany.

The German Research Center for Geosciences's Information System and Data Center (ISDC) portal is integrating important data management services for satellite missions such as German Research Center for Geosciences' Challenging Mini-Satellite Payload (CHAMP), Germany's TerraSAR-X (an X-band synthetic aperture radar satellite), and the Gravity Recovery and Climate Experiment (GRACE, a joint partnership between the German Aerospace Center and the National Aeronautics and Space Administration (NASA)), as well as for international geodetic projects like the International Association of Geodesy's Global Geodynamic Project (GGP), and the German Research Center for Geoscience-Dresden Technical University's global positioning system (GPS) data reprocessing (GPS-PDR) project (Flechtner and others, 2003; Reigber and others, 2003; Ritschel and others, 2003, 2006, [2008]).

The main components of the ISDC portal system are the portal framework, a content management system, a user and product management application, as well as detailed monitoring and statistics software. The new ISDC product philosophy describes the usage of product-type and data-product-related metadata documents written in Directory Interchange Format Extensible Mark-up Language (DIF XML) (Mende and others, 2007, 2008; Ritschel and others, 2007) for a standardized documentation of product type and data-product-specific information.

The entrance to the new ISDC portal (http://isdc.gfz-potsdam.de) is defined by a three-fold graphical user interface (GUI) as shown in figure 1. The portal framework is based on the open-source software PostNuke (http://postnuke.com). In addition to the portal framework and standard components (such as user registration, content management, and a user forum), most of the components were developed by the ISDC team.

At present, the ISDC is managing almost 300 different product types covering various geoscience domains such as geodesy, geophysics, atmospheric physics, and ionosphere physics. One third of these product types are accessible by public users and user groups. The other product types are only for operational, restricted, and internal uses. More than 11 terabytes of data and 16 million data products are in long-term storage and are accessible online through the ISDC product archive. There are currently 1,760 international users and user groups, a number that has been increasing exponentially since the start of the new ISDC portal in March 2006. Besides Germany, most of the users and user groups are from China, followed by United States, India, and Japan.

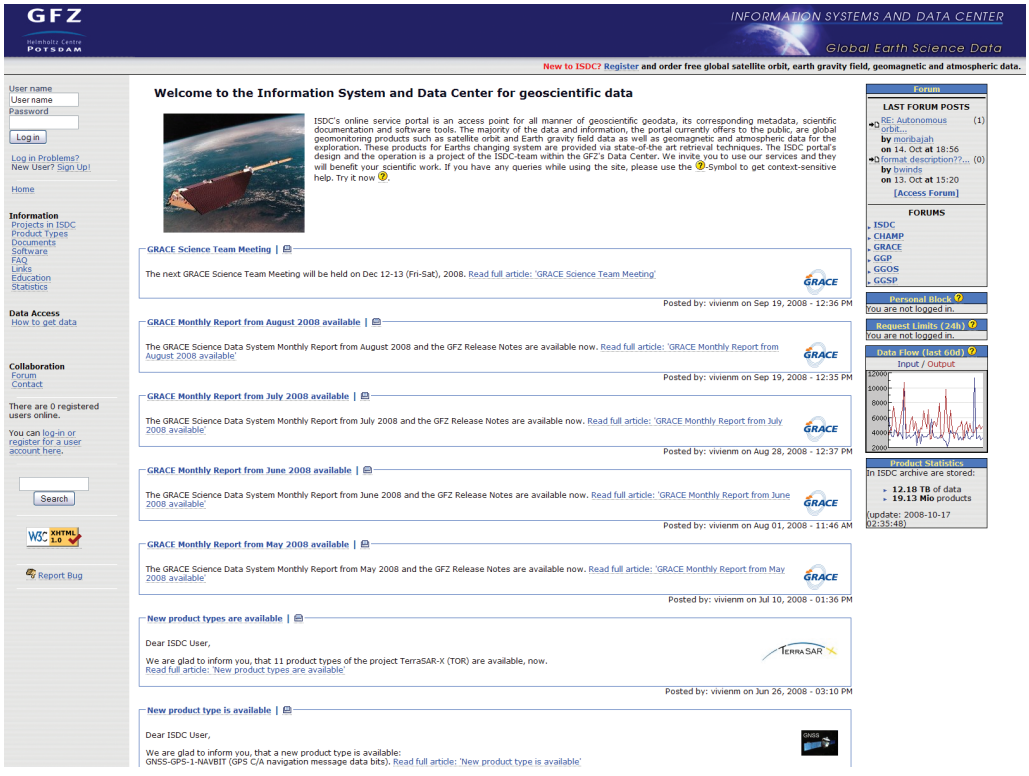


Figure 1. The German Research Center for Geosciences' Information System and Data Center graphical user interface. The left panel is the navigation portlet with a list of links to the main parts of the portal, such as projects, product type description, the Content Management System CMS, and so on. The middle panel contains links to the main content. The right panel contains links to the auxiliary information portlet.

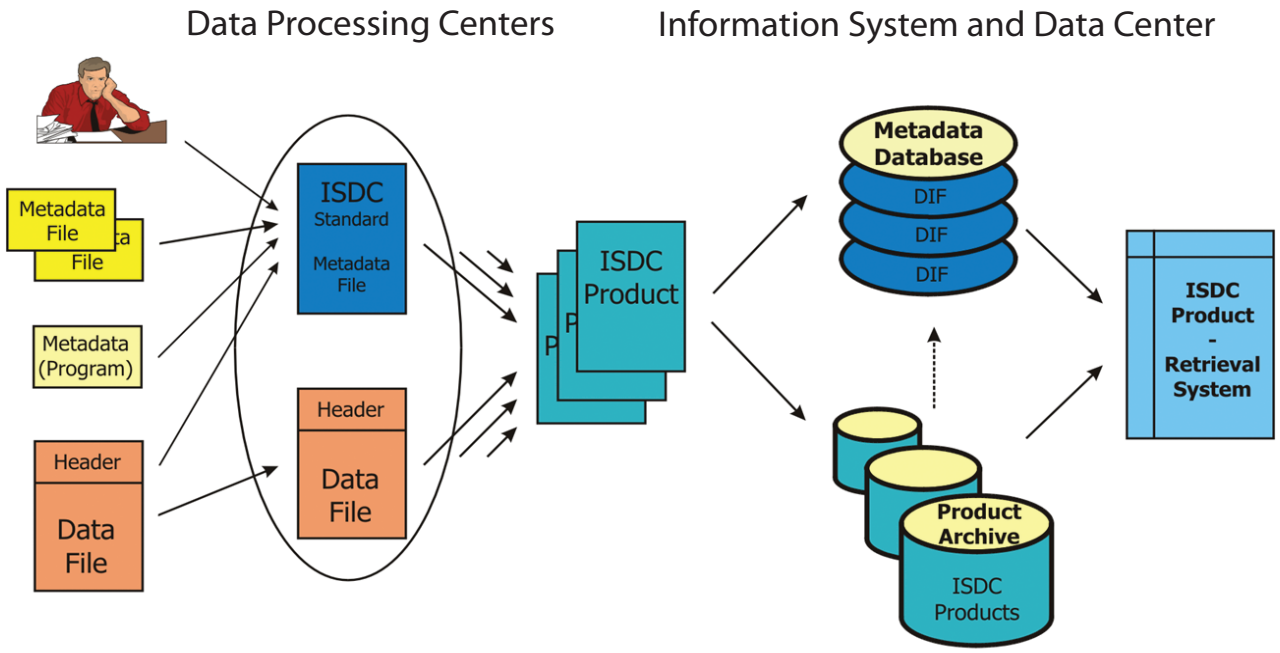


Figure 2. Diagram illustrating product philosophy and metadata processing at the Information System and Data Center (ISDC). ISDC products are generated in different processing centers. Important sources of information about the content and structure of data files are often stored in data-file headers, or in processing software (programs), or in proprietary metadata files, or, sometimes, it is stored only in scientists' heads; therefore, not only is it necessary to collect and store this information, but it is also necessary to be able to read the wide variety of different data formats for different data products. The ISDC is dealing with almost 300 different product types and many of them use different approaches for the storage of metadata. The only way that ISDC can overcome these challenges is with its own product philosophy, which describes a method of using a standardized metadata schema for the storage of metadata and the creation of an ISDC product. Standardized ISDC products contain one data file (or a small collection of data granules) and one appropriate DIF metadata file. In this manner, the management of ISDC products is realized by software that collects metadata by parsing only the standardized DIF metadata files. There is no need to look into the header of different types of data or to deal with additional metadata files. The collected metadata are stored in a database, whereas the complete products are in long-term storage in the ISDC data archive.

The ISDC is managing almost all parts of a scientific data product's lifecycle. Because of the standardized ISDC product philosophy, almost 300 different product types can be managed by a uniform process. As mentioned already, all product types are described by standardized parent DIF (NASA's Directory Interchange Format for the Global Change Master Directory, or GCMD) metadata documents (fig. 2) (Mende and others, 2007, 2008, this volume; Ritschel and others, 2007). This means that all product-type-dependent information (such as entry identification, entry title, parameters or keywords, topic category, data center, summary, personnel, instrument, and quality) and the temporal and spatial coverage of the whole data set are recorded in DIF files, which describe product-type and data-product-related metadata in one file only. In order to deal with data products (single data files) of a specific product type in a more efficient way, a further development of the DIF standards was necessary. In addition to the "old" ISDC metadata standards, the product-dependent metadata of all new ISDC product types are stored in separate data-product DIF metadata documents. Detailed information about the new ISDC product philosophy based on XML structures is found in Mende and others (2008, this volume).

The data input and output management is operated by ISDC's file transfer software components ("data pumps"), which are not only designed for the transfer of data but also for importing information to the ISDC product catalog (fig. 2). In order to keep ISDC data and information sustainable, science-driven data review processes are necessary; unfortunately, they are not occurring on a regular basis.

In order for improvements to (1) the interoperability of the relational database-based ISDC catalog (which consists of information about product types and data products), and (2) the XML document-based ISDC metadata collection to occur, the additional use of standardized service-oriented-architecture- (SOA-) driven concepts would be helpful. Using XML for metadata related to ISDC data and data products provides an opportunity for the smooth transformation of metadata documents from one standard to another. An example of this benefit is the Extensible Stylesheet Language (XSL) transformation from ISDC's metadata written as DIF XML to Open Geospatial Consortium- (OGC-) and ISO 19115-compliant metadata for importing into Web-based systems such as the open-source Catalogue Service for Web (CSW) software "deegree" (http://www.deegree.org) (Braune and others, 2003;

International Organization for Standardization, 2003; Voges and Senkler, 2005; Burgess and others, 2006).

Most of the ISDC interfaces are based on committee-driven standards and techniques. In addition, there are many community-driven activities and developments, which have been used in composing the interactive Web 2.0. Currently, the ISDC team is studying such Web 2.0 techniques as tagging and social navigation for use at the ISDC, and appropriate user interfaces already are in development. The combination of Web 2.0 techniques with Semantic Web languages such as Web Ontology Language (OWL) and the Simple Knowledge Organization System (SKOS) is offering new ways to represent data, information, and knowledge stored at the ISDC. Detailed information about ISDC's research in Semantic Web technologies is given by Ritschel and others (2008, this volume).

References Cited

Braune, S., Czegka, W., Klump, J., Palm, H., Ritschel, B., and Lochter, F.A., 2003, Applications of metadata in conformity with ISO 19115 for catalogue services dealing with environmental and geoscientific geodata: Zeitschrift für Geologische Wissenschaften, v. 31, no. 1, p. 37–44.

Burgess, P., Palm, H., Ritschel, B., Bruhns, C., Freiberg, S., Gericke, L., Kase, S., Kopischke, R., Loos, S., and Lowisch, S., 2006, Implementing modern data dissemination concepts in the ISDC Portal, *in* Observation of System Earth from Space, Status Seminar, Bonn, Germany, September 18–19, 2006: Potsdam, Germany, Geotechnologien, available only online at http://isdc.gfz-potsdam.de/. (Accessed August 11, 2008.)

Flechtner, F., Ackermann, C., Meixner, H., Meyer, U., Neumayer, K., Ritschel, B., Schmidt, A., Schmidt, R., Zhu, S., and Reigber, C., 2003, Development of the GRACE Science Data System, *in* Program and abstracts, Observation of System Earth from Space, Status Seminar, Munich, Germany, June 12–13, 2003: Geotechnologien Science Report 3, p. 48–50.

International Organization for Standardization, 2003, ISO 19115—Geographic information—Metadata: Geneva, Switzerland, International Organization for Standardization, 140 p.

Mende, Vivien, Ritschel, Bernd, Freiberg, Sebastian, Palm, Harmut, and Gericke, Lutz, 2008, Directory Interchange Format (DIF) metadata and handling at the German Research Center for Geosciences' Information Systems and Data Center, *in* Proceedings, Geoinformatics 2008—Data to Knowledge, Potsdam, Germany, June 11–13, 2008: U.S. Geological Survey Scientific Investigations Report 2008–5172, p. 43–46 (this volume).

Mende, V., Ritschel, B., Palm, H., Gericke, L., and Freiberg, S., 2007, ISDC metadata management, *in* Program and abstracts, Observation of System Earth from Space, Status Seminar, Munich, Germany, November 22–23, 2007: Geotechnologien Science Report 11, p. 9–11.

Reigber, C., Schwintzer, P., Lühr, H., Massmann, F., Galas, R., and Ritschel, B., 2003, CHAMP mission science data system operation and generation of scientific products, *in* Program and abstracts, Observation of System Earth from Space, Status Seminar, Munich, Germany, June 12–13, 2003: Geotechnologien Science Report 3, p. 129–131.

Ritschel, B., Behrends, K., Braune, S., Freiberg, S., Kopischke, R., Palm, H., and Schmidt, A., 2003, CHAMP/GRACE-Information System and Data Center (ISDC)—The user interfaces for scientific products of the CHAMP and GRACE mission, *in* Program and abstracts, Observation of System Earth from Space, Status Seminar, Munich, Germany, June 12–13, 2003: Geotechnologien Science Report 3, p. 132–133.

Ritschel, B., Bruhns, C., Burgess, P., Freiberg, S., Gericke, L., Kase, S., Kopischke, R., Loos, S., Lowisch, S., and Palm, H., 2006, The integration of CHAMP and GRACE products as well as associated scientific services in the new ISDC portal, *in* Observation of System Earth from Space, Status Seminar, Bonn, Germany, September 18–19, 2006: Potsdam, Germany, Geotechnologien, available only online at http://isdc.gfz-potsdam.de/. (Accessed August 11, 2008.)

Ritschel, B., Bruhns, C., Kopischke, R., Mende, V., Palm, H., Freiberg, S., and Gericke, L., 2007, The ISDC concept for long-term sustainability of geoscience data and information, *in* Proceedings, PV 2007 Conference—Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data, Munich, Germany, October 9–11, 2007, available only online at http://www.pv2007.dlr.de/main/results_en.htm/. (Accessed August 12, 2008.)

Ritschel, B., Mende, V., Palm, H., Kopischke, R., Bruhns, C., Gericke, L., and Freiberg, S., [2008], ISDC services—Data management, catalog interoperability and international cooperation: Potsdam, Germany, German Research Center for Geosciences, available only online at http://isdc.gfz-potsdam.de/. (Accessed August 12, 2008.)

Ritschel, Bernd, Pfeiffer, Sabine, Mende, Vivien, and Freiberg, Sebastian, 2008, Semantic web technologies for value-added services at the German Research Center for Geosciences' Information Systems and Data Center, *in* Proceedings, Geoinformatics 2008—Data to Knowledge, Potsdam, Germany, June 11–13, 2008: U.S. Geological Survey Scientific Investigations Report 2008–5172, p. 66–69 (this volume).

Voges, U., and Senkler, K., eds., 2005, OpenGIS Catalogue Services Specification 2.0—ISO19115/ISO19119 Application Profile for CSW 2.0.: Wayland, Mass., Open Geospatial Consortium, Inc., available only online at http://www.opengeospatial.org/standards/deprecated/. (Accessed August 12, 2008.)

Scientific Application Portal Development for Research and Education in Cyberinfrastructure

By Choonhan Youn,[1] Chaitan Baru,[1] and Nancy Wilkins-Diehr[1]

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

In recent years, the Internet has become an integral resource in the classrooms and homes of teachers and students. Widespread Web access to data and analysis tools by means of a cyberinfrastructure enhances the opportunities for teaching and learning. The concept of a cyberinfrastructure encompasses advanced scientific computing as well as a more comprehensive infrastructure for research and education, all of which are based on distributed federated networks of computers, data collections, information resources, online instruments, visualization tools, and human interfaces. Science communities increasingly are becoming dependent on such cyberinfrastructure for their research. In order to effectively train future scientists to make use of today's cyberinfrastructure, educators must embrace the same technologies. Portals (or science gateways) provide tools for end users to use for online collaborations, access to computing resources, the ability to launch computational tasks, and sharing of data and other resources with others in a given community. The accessibility of Web interfaces means that students in a variety of locations and with a variety of backgrounds can all make use of an advanced cyberinfrastructure. Enhancing portals that we designed for high-end science so that they are also suitable in a variety of educational settings is a major contribution to workforce development.

The Geosciences Network (GEON, http://www.geongrid.org), which is a project funded by the United States' National Science Foundation (NSF), is an

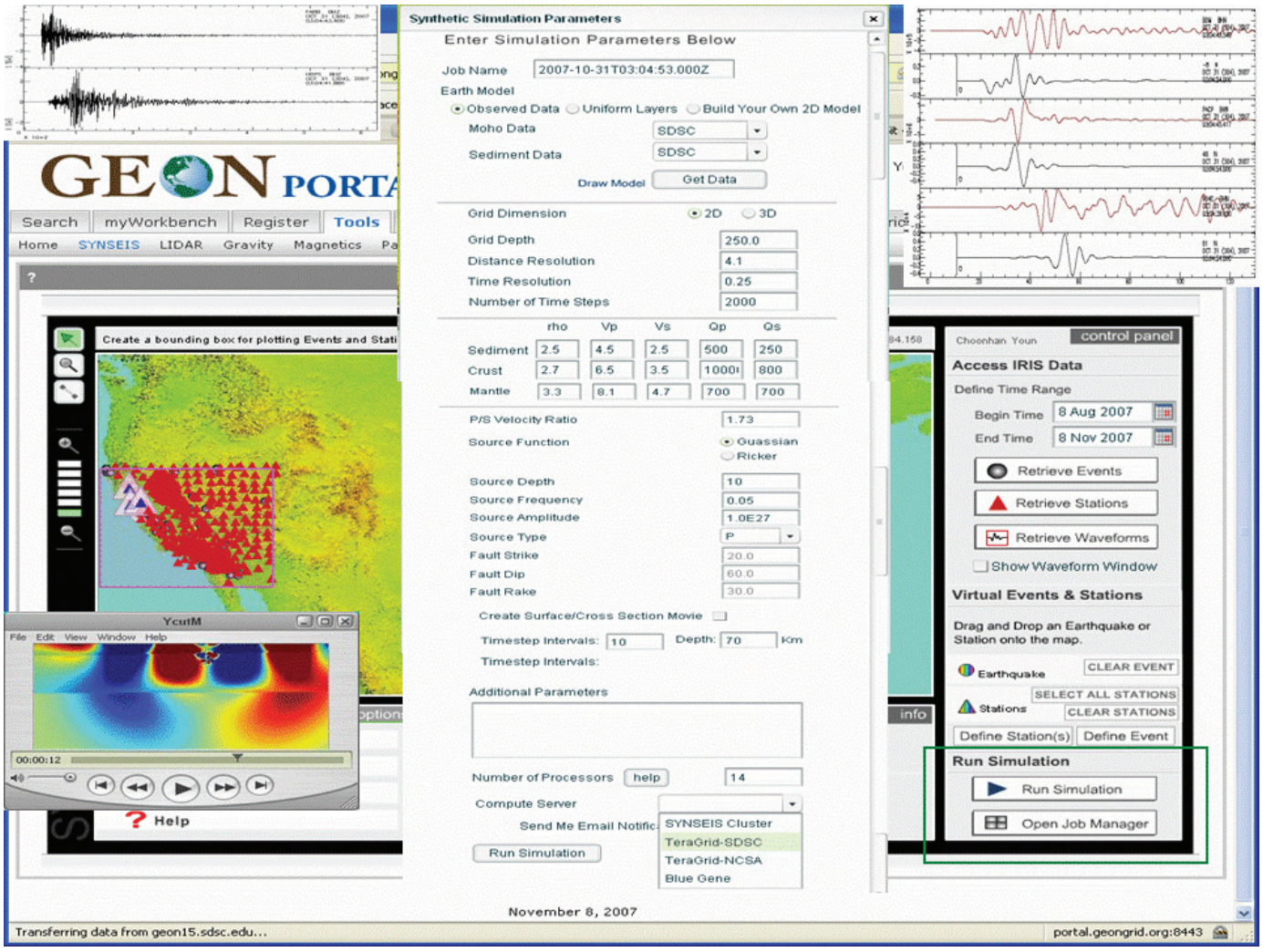


Figure 1. Screen capture showing the Geosciences Network (GEON) Synthetic Seismogram (SYNSEIS) application being used to access data in the Incorporated Research Institutions for Seismology (IRIS) archives.

open collaborative project that is developing a cyberinfrastructure for the integration of three- and four-dimensional earth science data. The focus is on building data-sharing frameworks, developing tools and services, and identifying best practices with the objective of dramatically advancing geoscience research and education. These developments in infrastructure seek to extend access to data and to complex modeling tools from the hands of a few researchers to a much broader set of users. The GEON Synthetic Seismogram (SYNSEIS) application, for example, provides an easy-to-use interactive data-access and computing environment, using resources in TeraGrid (a network coordinated by the University of Chicago's Grid Infrastructure Group) or GEON to study the three-dimensional lithospheric structure. SYNSEIS enables users to access the Incorporated Research Institutions for Seismology (IRIS) data archives by using an interactive map interface and retrieve data on earthquakes, seismic stations, and corresponding seismic waveforms (fig. 1). A three-dimensional crustal structure model is then obtained from a different Web service. The user defines earthquake source parameters and then generates a synthetic seismogram using validated software running on a remote supercomputer.

To support classroom use of this tool, we have developed a class account management system that instructors can use to easily create group accounts. Our goal is to combine Catalogue Service for Web (CSW) 2.0 concepts with conventional cyberinfrastructure to create virtual scientific and education communities. In this presentation, we will describe how the myProjects collaboration tools available in the GEON portal can be used, along with tagging, to allow users to review and vote on submitted contents (including parameter settings and job outputs) and support group discussions among the class. Sharing of such information may potentially help users avoid the unnecessary (and expensive) execution of computer coding and may provide them with a more effective way of sharing and testing possible solutions. As the use of collaboration tools and cyberinfrastructure matures, tools such as myProjects will have the potential for significant impact on education and research.

The TeraGrid project has recently launched the TeraGrid Pathways initiative. The goal of the initiative is to broaden the use of the high-end computing, data, and visualization resources provided by the NSF's Office of Cyberinfrastructure. One component of this initiative is the adaptation of "science gateways" for use by educators. GEON has been selected to serve as a prototype for this work. The work will focus on the extension of GEON's educational tools for use at the community college level, because there are many underserved communities who use community colleges as a springboard to higher education. The geosciences present tangible, visual concepts that lend themselves well to a variety of educational settings. In addition, the Native American population, through tribal colleges, has expressed interest in using science gateways for education on remote reservations. These connections will be pursued through the work with GEON.

Neptune—Developing a Digital Information Infrastructure for Micropaleontology in the 21st Century

By David Lazarus,[1] Cinzia Cervato,[2] Douglas Fils,[2] and Patrick Diver[3]

[1]Paleontology Section, Museum of Natural History, Berlin, Germany.

[2]Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa.

[3]DivDat Consulting, Wesley, Ark.

Marine microfossil occurrences are used extensively for geologic age determination and for paleoceanographic or other paleoenvironmental research. They are less commonly used for studies of evolution, despite having one of the best-preserved records of evolutionary change in the entire fossil record. Although marine microfossils have been widely studied from rock formations on land, several decades of scientific deep-sea drilling also have yielded a large archive of information on marine microfossil occurrences, particularly for the pelagic unicellular plankton groups: diatoms, radiolarians, coccolithophores ("nannofossils"), and planktonic foraminifera. Despite many published (mostly monographic) comparisons of occurrence information for selected individual, biostratigraphically important species, there has been no general synthesis of the fossil data collected by deep-sea drilling, nor any appropriate tools such as taxonomically and age-controlled occurrence databases, which are necessary for the effective synthesis of fossil occurrence data. There are databases (such as the Sepkoski database, developed at the University of Chicago, for marine invertebrate fossils) that have played a central role in the development of invertebrate paleontology for many years. Beginning in the early 1990s, a new database system—Neptune—was developed to address the need for deep-sea marine microfossil synthesis tools. Neptune originally was developed at the Swiss Federal Institute of Technology—Zürich and subsequently as part of the Chronos project (funded by the United States' National Science Foundation) at Iowa State University.

Neptune is a relational database and a set of external tools that link raw occurrence data for marine microfossils, as given in several hundred selected original range charts of deep-sea drilling science reports, to the essential scientific information needed to effectively retrieve and synthesize these data. These essential scientific data additions include (1) numeric geologic ages for every occurrence (which are based on quantitative age models for every drill hole in the system) and (2) master taxonomic name lists that link synonyms for the same taxa concepts to each other and distinguish different taxonomic data quality records from each other (for instance, records with clearly identified taxa versus records with "cf" or "?" observations). Neptune thus allows data to be retrieved from this

important archive in a form suitable for large-scale syntheses of the deep-sea marine microfossil record and provides tools for summarizing the information. More recently, Neptune has been linked to the successor of the Sepkoski database—the Paleobiology Database (PBDB), which is an NSF-funded project currently hosted by University of California at Santa Barbara—thereby allowing microfossil data from land sections to be combined with data from marine sections. The system is currently being used to study large-scale patterns of Cenozoic evolutionary change in the plankton and as an age model and taxonomic reference library for other users of deep-sea drilling sections.

The current implementation of Neptune is as a relational database, which uses Post-Ingres Structured Query Language (PostgreSQL) and is hosted on the Chronos server stack at Iowa State University. Neptune is searchable through the Chronos portal and is seamlessly integrated with Chronos' Java-based versions of the original Mac True Basic Age-Depth Plot and the Age-Range Chart applications written at the Swiss Federal Institute of Technology.

The analysis of large, heterogeneous datasets inevitably raises problems of mixed data quality, with data gaps, uneven sampling, age outliers, and incorrectly entered primary observations all affecting the validity of results. By linking with PBDB, Neptune analyses can make use of PBDB's large library of paleobiologic tools, such as "range-through" and "subsampling" methods, for dealing with unevenly sampled data. Currently, tools are being developed for dealing with age outliers in taxon ranges, which were created due to taxonomic errors in the original data, reworking of fossils, or age model errors that resulted from poorly resolved or mutually inconsistent primary chronostratigraphic information.

Future development of Neptune is envisioned as part of a gradually evolving network of digital resources in marine micropaleontology (fig. 1, next page). These include stronger links to (1) primary deep-sea sediment core databases, such as the Janus system of the Integrated Ocean Drilling Program (IODP); (2) biostratigraphic and lithologic data from all IODP sites; (3) digital taxonomic catalogs of species images and descriptions (one such link, to Chronos' own digital catalog system, has already been developed by Chronos); and (4) major collections of marine microfossil materials held in museums and other institutions around the world, such as the Micropaleontological Reference Centers' network of deep-sea marine microfossil slides. Effective networking of these resources will require the development of funding mechanisms to maintain and regularly update a central registry of the key shared field data: the taxonomic and age model information. The benefits for research, however, will be substantial and will include (1) major increases in data synthesis capacity, particularly for studies of long-term global processes, and (2) an improved efficiency in data retrieval and analysis for many other individual micropaleontologic research projects.

The EarthScope Data Portal

By Ashraf Memon,[1] Chaitan Baru,[1] Knut Behrens,[2] Rob Casey,[3] Ben Hoyt,[4] Linus Kamb,[3] Kai Lin,[1] Bruce Weertman,[3] and Charley Weiland[5]

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

[2]International Continental Scientific Drilling Program, German Research Center for Geosciences, Potsdam, Germany.

[3]Incorporated Research Institutions for Seismology (IRIS), Data Management Center, Seattle, Wash.

[4]University NAVSTAR (Navigation Signal Timing and Ranging) Consortium, Boulder, Colo.

[5]Department of Geophysics, Stanford University, Palo Alto, Calif.

The EarthScope data portal, which is now in its alpha release, is being developed to provide a unified, single point of access to EarthScope data and products from (1) USArray (a continent-wide seismic observatory that is a component of EarthScope), (2) the University NAVSTAR (Navigation and Signal Timing and Ranging) Consortium's (UNAVCO's) Plate Boundary Observatory (PBO), and (3) the San Andreas Fault Observatory at Depth (SAFOD), which is funded by the National Science Foundation. The portal features basic search and data access capabilities to allow users to discover and access EarthScope data using spatial, temporal, and other metadata-based (data-type, station-specific) search conditions.

In this presentation, we will describe the features, design, and future improvements of the portal. This portal is being developed by a team consisting of the Geosciences Network (GEON, http://www.geongrid.org), Incorporated Research Institutions for Seismology (IRIS), NAVSTAR (Navigation and Signal Timing and Ranging) Consortium (UNAVCO), Stanford University, and the German Research Center for Geosciences' International Continental Scientific Drilling Program. The portal search module invokes Web services developed by IRIS, UNAVCO, and Stanford to search for EarthScope data in the archives at each of these locations. The Web services provide information about all resources (data) that match the specified search conditions. Users can select from the returned datasets, add selected data to a "data cart," and request the selected data to be packaged for download to the user. The services also are defined for "station discovery" (finding which stations are available for specified spatial and temporal parameters) and "data discovery" (finding the datasets that are available from the stations). The returned resulting datasets are organized in a hierarchical structure that is categorized based on the data type so that users can browse at ease and subsequently choose specific datasets, which are then assembled in a user "workspace" and available for download.

The EarthScope data portal has taken advantage of the significant portal development efforts of the GEON

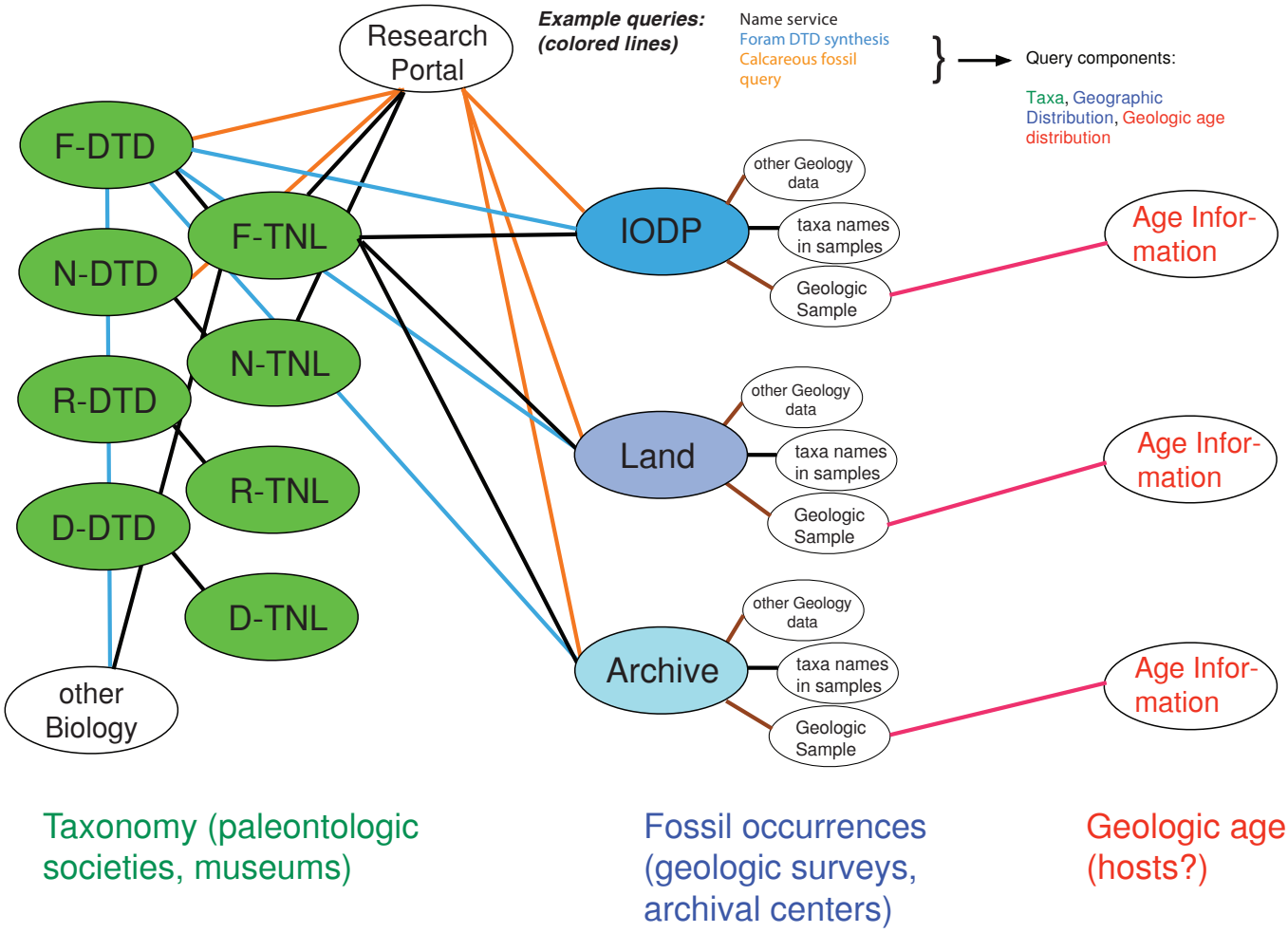


Figure 1. Diagram showing structure of a possible future federated network of micropaleontology databases and related data sources, in which the current unitary content of Neptune is distributed between different science organizations. Individual databases or sections of databases shown by ovals, logical links between data types by lines. Organizations supporting each type of database shown in parentheses at bottom of figure. Examples of queries are given by colored text at top middle and right of figure. Earth science research questions about microfossils typically require information about the occurrences of species or other taxa in geologic samples, together with the geologic ages and geographic distribution of the samples in which they are found. Colored lines on left of figure show examples of how databases query each other according to different needs. Other than simple name service resolution (black), most queries involve several different data types (taxa, geologic sample information, and geologic age); queries are resolved into their components which are handled by different databases in the network. Taxonomy, databases (green, on left side of figure) for each different microfossil group, created and maintained by groups of specialist paleontologists, can resolve the complex way in which species and other taxonomic concepts are recorded by taxonomic names. The organizations that collect geologic materials (blue, middle of figure) can provide the needed geographic information about the samples, and they or a scientific data center's archives store and query the actual reports of taxonomic name occurrences in geologic samples. "Land" refers to paleontologic databases from nonmarine sections, such as the Paleobiology Database. The geologic age of the samples (right side of figure) is a complex scientific interpretation linked to individual samples and sections. Most organizations do not maintain geologic age information with an accuracy appropriate for scientific research. Policy change or a dedicated chronology center may be needed to solve this problem. Abbreviations are as follows: IODP, Integrated Ocean Drilling Program; DTD, digital taxonomic dictionary or "catalog" that describes and illustrates taxonomic concepts; TNL, taxonomic name list that links names to concepts and serves as key field for linking database content together (name services). Microfossil groups are abbreviated as follows: F, foraminifera; N, calcareous nannofossils; R, radiolarians; D, diatoms.

project at the San Diego Supercomputer Center (SDSC, http://portal.geongrid.org), and the development of Web service interfaces at IRIS, UNAVCO, and Stanford. The portal is implemented using the open-source portal infrastructure software, GridSphere, which supports the well-known Java portlet interface, JSR 168, or the Portlet application programming interface (API). It uses a set of "core" portlets that have been developed in GEON for data registration, searches, and workspace services.

In this presentation, we will provide a report on the current state of development of the EarthScope data portal. So far, a preliminary deployment of the portal software has been conducted on systems at SDSC; initial designs have been accomplished for the StationDiscovery, DataDiscovery and DataPackaging services; and IRIS, UNAVCO, and Stanford have implemented the alpha version of the corresponding Web services, which runs on servers at their respective locations. The beta version of these Web services will be demonstrated during the presentation.

Enhancing Core Drilling Workflows Through Advanced Visualization Technology

By Yu-Chung Chen,[1] Jason Leigh,[1] Andrew Johnson,[1] Luc Renambot,[1] Emi Ito,[2] Paul Morin,[3] Sean Higgins,[4] Frank Rack,[5] Richard Levy,[5] and Josh Reed[6]

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

[2]Limnological Research Center, University of Minnesota, Minneapolis, Minn.

[3]Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minn.

[4]Consortium for Ocean Leadership, Washington, D.C.

[5]Department of Geosciences, University of Nebraska—Lincoln, Lincoln, Nebr.

[6]Antarctic Geological Drilling Science Management Office, University of Nebraska—Lincoln, Lincoln, Nebr.

Everywhere in the sciences, modern information technologies change the way people work. New tools and equipment are constantly being developed to help scientists to process a huge amount of data and to observe detailed phenomenon that they could not see before. We present the design and development of an initial visual core description tool with collaboration and annotation features that may be used for core drilling expeditions. By observing the use of the tool during real core drilling expeditions, we have learned how scientists make use of it and how it fits into modern core drilling workflows.

The CoreWall Suite is a set of tools designed to aid real-time stratigraphic correlation, create initial core descriptions, and provide data visualization for various core-drilling communities. Corelyzer is the initial visual core description tool developed for the CoreWall suite. Corelyzer allows scientists to collaborate over huge data visualizations on a desktop workstation using one or more monitors with great interactivity and scalability. The main software architecture was developed using Java language with a native scene-graph library. The user-interface module and data-retrieval module were written in pure Java. The scene-graph library was developed in native C language with standard Open Graphics Library (OpenGL) for efficient rendering.

Scalability

Corelyzer was designed to be scalable. A graphics system that employs level-of-detail (LOD) control and texture paging (a method to conserve the amount of memory an image needs to load) was implemented inside Corelyzer. This graphics system allows scientists to load and interact smoothly with data representing thousands of meters of geological cores; one kilometer of core data produces roughly 30 gigabytes (GB) of raw imagery.

Visualization Capability

Corelyzer supports hardware setups that range from a single screen on a laptop computer to six liquid crystal display (LCD) panels connected to a single desktop workstation. The system scales core images with different formats and resolutions to match the physical core sample size. The main user interface provides major data visualization capabilities for core drilling, such as high-resolution core imagery, numerical core logging data, lithologic diagrams, smear slides, thin sections, and user-generated free-form or structured annotations.

Software extensibility

The Corelyzer source code was released under an open-source license and uses plain Extensible Mark-up Language (XML) file formats. Anyone can take this code and make modifications to fit his or her needs. For example, with a simple exporter module, the Drilling Information System (DIS) can export core data along with core imagery as a Corelyzer session file format, which enables all the core data to be loaded seamlessly into Corelyzer. Corelyzer also provides a plug-in framework, which allows third-party developers to extend its functionalities and capabilities; for example, support for lithologic diagrams was developed by Josh Reed, who is the information technology manager of the Antarctic Geological Drilling project (ANDRILL, a third-party entity). Moreover, for standardized core data (metadata) distribution, a "core feed" plug-in was designed to allow users to subscribe to core data description feeds defined in the standard syndication format. Users can look up the available feeds and subscribe to interesting core data, in the same manner as "podcasts." The

feed provides the metadata required to download and interpret actual imagery and numerical core log datasets.

Deployment and Usage

The CoreWall prototype has been used since 2006 by the National Lacustrine Core Repository at the University of Minnesota and by the Lamont-Doherty Earth Observatory at Columbia University. At the end of 2006 and 2007, Corelyzer was used in core drilling expeditions by the ANDRILL project. In the 2006 season, ANDRILL deployed with two CoreWall workstations with 30-inch LCD displays (fig. 1). The workstations mainly were used as follows:

1. During the night shift, a CoreWall workstation was placed alongside the physical cores on the tabletop to help with core description. The visualization capability that allows zooming into the high-resolution images beyond a core's physical scale while still maintaining details made the setup act like an electronic microscope for the cores. This capability made it easier to do more accurate and detailed observations. For example, Dr. Franco Talarico (University of Siena, Italy) used to manually draw all of the core clasts on paper in order to classify them. With CoreWall, he now conducts research more efficiently with modern tools and techniques.

2. In the morning briefing, the other workstation was used for progress report explanations and tours of the core imagery alongside physical core samples. This workstation was set up in a public discussion area in order to help the researchers conduct context-sensitive discussions that benefited from being able to see the images on the screen.

Although there were only two CoreWall workstation setups for the entire science team, all involved personnel were encouraged to install Corelyzer on their laptop computers in order to easily access related data. The comments from the scientists have been positive, and in the 2007 season, ANDRILL increased the number of CoreWall workstations to six for the entire science team. One CoreWall workstation was set up right at the drill site to help the drillers make on-the-spot drilling decisions based on collected data.

Figure 1. A Corelyzer set-up, which is running on an Apple Mac Pro computer and uses two Apple 30-inch cinema-display monitors.

An Analysis of Landscape Change Based on Remote Sensing and Geographic Information Systems in the Jinghe Basin, China

By Yonghua Zhao[1]

[1]College of Earth Science and Land Resources Management, Chang'an University, Xi'an, China.

Using digital Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper (ETM+) imagery from 1986, 1995, and 2000 and a geographic information system (GIS), landscape changes were interpreted and analyzed in the Jinghe basin (a region in China that is experiencing serious soil erosion problems) in order to provide basic data for local decisionmaking and for sustainable land use and management. The results showed that between 1986 and 2000, most of the area covered in the basin was classified as grassland. The second largest area was cropland, the third was shrubland, and the fourth was forestland.

Because they cover the most area, grassland and crops probably have the most important effect on the direction of landscape change, ecological and environmental change, the safety of the regional ecology, and so on, in this region. Forests have an important function in maintaining environmental quality, preventing soil erosion, or maintaining ecological balance in the region; however, the combined area of all forests (including scattered forested areas) and shrubland was less than 11 percent of the total basin area. Only the area classified as built-up land always showed an increase from 1986 to 2000. Areas of crops, forests and scattered forests, and unused land increased between 1986 and 1995 and decreased between 1995 and 2000. Bidirectional change of all landscape types was more obvious between 1995 and 2000, and landscape change was more obvious between 1986 and 1995. Above all, these changes showed that the landscape developed continuously and obviously was transformed before 1995; the landscape became regulated after 1995.

Among the types of areas that showed an increase, crop areas increased the most, by about 4,165 hectares (ha) from 1986 to 1995. The second greatest increase was in areas classified as shrubland, by about 2,207 ha between 1995 and 2000. Unused land increased by about 849 ha between 1986 and 1995, and grassland increased by about 816 ha between 1995 and 2000. The increase in the amount of area covered by water was less than 94 ha from 1995 to 2000.

Among the types of areas that showed a decrease, grasslands decreased the most, by about 3,125 ha from 1986 to

1995. The second was shrubland, by about 2,781 ha between 1986 and 1995. Between 1995 and 2000, the area classified as crops decreased by about 2,097 ha, and scattered forests decreased by 1,086 ha. Other forested areas decreased by less than about 200 ha between 1995 and 2000.

The changes in the areas classified as crops, grassland, shrubland, and unused land dominated the changes in the Jinghe basin and influenced the direction and rate of the total landscape change. The GIS analysis further indicated that five land-use conversions were prominent: (1) crops to grassland, (2) crops to built-up land, (3) grassland to crops, (4) grassland to shrubland, and (5) grassland to scattered forest. The amount of change from crops to grassland was greater between 1986 and 1995 compared with the period between 1995 and 2000. The above results showed that the landscape changed more between 1986 and 1995 than between 1995 and 2000, and the landscape heterogeneity and fragmentation increased and the landscape connectivity decreased at regional scale.

Changes in land use or changes to the landscape can be measured by using either a land-use dynamic index (LUDI) (Liu and Buheaosier, 2000) or a landscape-departure index (LDI) (Wang and others, 2004), which quantify the rate of change and can help to predict the future trend of a change. The LUDI of cropland and grassland was lower and that of other forestland was higher during two periods. The LDI refers to the extent of departure from the original landscape resulting from of the effects of human activity. The index value is obtained by totaling the acreage covered by land classified as built-up land, crops, and other forested land in the research area and dividing that sum by the total acreage of the research area. In the Jinghe basin, the LDI increased from 0.436 in 1986 to 0.448 in 1995 and then decreased to 0.443 by 2000, which showed that the effect of human activities on the landscape increased from 1986 to 1995, but then decreased after 1995.

Remote-sensing data are widely used in many types of landscape-analysis studies. Data from the larger satellites such as Landsat's TM, ETM+, Multispectral Scanner (MSS), and Moderate Resolution Imaging Spectroradiometer (MODIS) mostly have been used by researchers who were studying regional land-use and land-cover change, global environmental change, and other studies that required large areas to be covered. Data from the larger satellites had some disadvantages; for example, data were expensive, satellite costs were too high, run cycles took too long, resolution was too low, and so forth. With recent advances in technology, more and more smaller satellites have been launched and have yielded the following advantages: data are inexpensive, satellites costs are low, run cycles are shorter, resolution is higher, the satellite signal is easier to obtain, and more. Therefore, we thought that a small satellite would be a better choice to obtain data related to landscape ecology, land use, and other landscape-related issues. In future landscape-change research, data from smaller satellites will be more widely used by researchers because of these advantages.

Reference Cited

Liu, Jiyuan, and Buheaosier, 2000, Study on spatial-temporal feature of modern land use change in China—Using remote sensing techniques: Quaternary Sciences [China], v. 20, no. 3, p. 229–239.

Wang, Zongming, Zhang, Shuqing, and Zhang, Bai, 2004, Effects of land use change on values of ecosystem services of Sanjiang Plain, China: China Environmental Science, v. 24, no. 1, p. 125–128.

The GEOROC Database as Part of a Growing Geoinformatics Network

By Baerbel Sarbas[1]

[1]Department of Geochemistry, Max Planck Institute for Chemistry, Mainz, Germany.

Since its introduction in 1999, the geochemical database GEOROC (Geochemistry of Rocks of the Oceans and Continents, http://georoc.mpch-mainz.gwdg.de), which is hosted by the Max Planck Institute for Chemistry in Mainz, established itself as a major online resource available to the scientific community. GEOROC provides geochemical data that have been published for volcanic whole rocks and glasses, minerals and inclusions from ocean islands, large igneous provinces, convergent margins, Archean greenstone belts, and rift and intraplate volcanic regions. As of April 2008, the database provides about 350,000 analyses published in about 7,300 papers.

The Web interface of GEOROC allows the user to select samples by bibliographic, tectonic, geographic, petrologic, and chemical criteria. As part of the bibliographic query, the search for the GEOROC_Reference_Number permits an easier reproduction of published compilations created with the help of the database. A Google Maps-based search has been added which allows users to query by geographic location. A new service for GEOROC users is a discussion forum. It allows comments and suggestions to be entered while simultaneously working with the database and allows users to flag typographical errors that were made either by the database team or were in the original published report.

To get a quick idea of the geochemical signature of samples from certain localities or of a special rock type, GEOROC includes "ready-made" compilations of published data. In addition to these unvalued data, we offer so-called expert datasets. These datasets are compiled by non-GEOROC scientists and can include more than just the measured data (for example, normalized data or element ratios).

GEOROC joined with the Petrological Database of the Ocean Floor (PetDB, which is hosted by Columbia University) and the North American Volcanic and Intrusive Rock Database (NAVDAT, administered by University of Kansas) to initiate

the National Science Foundation-funded EarthChem consortium (http://www.earthchem.org/), with the goal of increasing the synergy between the three geochemical database efforts. The EarthChem portal (http://geoportal.kgs.ku.edu/earthchem/jtest/) offers a seamless search across the three databases.

Publications cited in GEOROC are cross-linked with the geochemical database GeoReM (Geological and Environmental Reference Materials database; http://georem.mpch-mainz.gwdg.de/), which provides access to reference materials and isotopic standards. The detailed information about the analytical conditions that are described for the reference materials available in GeoReM enables users of GEOROC to estimate the quality of the analyzed rock samples. Similarly, it is possible to go directly from GeoReM to the respective reports in GEOROC to get the analyses of the studied samples.

Directory Interchange Format (DIF) Metadata and Handling at the German Research Center for Geosciences' Information System and Data Center

By Vivien Mende,[1] Bernd Ritschel,[1] Sebastian Freiberg,[1] Hartmut Palm,[1] and Lutz Gericke[1]

[1]Information System and Data Center, German Research Center for Geosciences, Potsdam, Germany.

The Information System and Data Center (ISDC) is managing more than 11 terabytes (TB) of geoscience data and information. Currently, these data are coming from 11 missions, including the German Research Center for Geosciences' Challenging Mini-Satellite Payload (CHAMP), the Gravity Recovery and Climate Experiment (GRACE, a joint partnership between the German Aerospace Center (DLR) and the National Aeronautics and Space Administration (NASA)), Germany's TerraSAR-X (an X-band synthetic aperture radar satellite), the International Association of Geodesy's Global Geodynamic Project (GGP), and others that have yielded nearly 300 product types and approximately 16 million products, which have been made available to more than 1,700 users. This paper gives a short overview about the development and use of metadata in the ISDC. Each product type that results from a geoscience mission or project consists of a set of products. A product is composed of a data file (or files) and a metadata document. Figure 1 shows the three product types resulting from the TerraSAR-X.

In order to describe and manage the products (data file(s) and metadata), we are using an evolution version 9.x of NASA's Directory Interchange Format (DIF) standard. The manager of NASA's Global Change Master Directory (GCMD), Lola Olsen, defines metadata as follows: "Descriptive information that characterizes a set of quantitative and/or qualitative measurements and distinguishes that set from other similar measurement sets." (See http://gcmd.nasa.gov/Aboutus/standards/.) For the management of ISDC product types, DIF is an excellent choice.

The ISDC base schema of the product type DIF XML documents is defined in the "base-dif.xsd" file. The ISDC Extensible Mark-up Language (XML) schema has been defined on the basis of the GCMD XML schema definition found online at http://gcmd.nasa.gov/Aboutus/xml/dif/dif_v9.7.1.xsd. In order to describe single products, it was necessary to extend the DIF standard and to modify the GCMD XML schema. Although the structure of the ISDC DIF XML schema is different from the GCMD schema, the ISDC product type DIF XML documents are still valid in relation to the GCMD schema. Additionally, the ISDC is using a product type-data child DIF combination. The metadata of product types are described in associated product type DIF XML files according to the "base-dif.xsd" schema. The product type DIF XML files are validated and stored in an Oracle XML database. The product- (data file-) specific metadata are documented in data-child DIF XML files. Each product type has its own schema for the data-child DIF XML files. Data-child DIF documents are used to describe the data-file-specific properties. The complex XML type <Data_Parameters> in the data-child DIF XML document provides the specific extension of the product type DIF XML structures. <Data_Parameters> includes specific metadata of the product, such as the data file name, data file size, revision, satellite identification, and other information. In order to implement this data model, we are using the redefined XML technique for the definition of complex XML types for the <Data_Parameters>. By redefining the ISDC "base-dif.xsd" schema, all data-child DIF XML documents are derived. Using the GCMD XML schema, this approach would not be possible because of the definition of XML reference structures.

The extended metadata in the data-child DIF XML documents are parsed by a Perl (an open-source software) script. If the data structure is correct, the extended metadata are stored in product-type-related tables in a relational database. The connection between the data-child DIF XML files and the product type DIF XML document is given by the equality of parts of the <Entry_ID> element in both the product type and the related product metadata documents. Additionally, the content of the <Parent_DIF> element in the data-child DIF XML document refers to the appropriate product-type DIF document. The relation between the schemata, the XML metadata files, and the storage structures is shown in figure 2.

Using the product-type DIF XML structures, it is possible to conduct a thematic content-based search of the different product-type documents as well as provide interoperability with other catalog systems. It is now possible to transform the XML DIF files into ISO 19115-standard documents (International Organization for Standardization, 2003) in order to use Open Geospatial Consortium-compliant Web services such as deegree's (www.deegree.org) Catalogue Service for Web 2.0. Furthermore, achieving harmony with other catalog systems

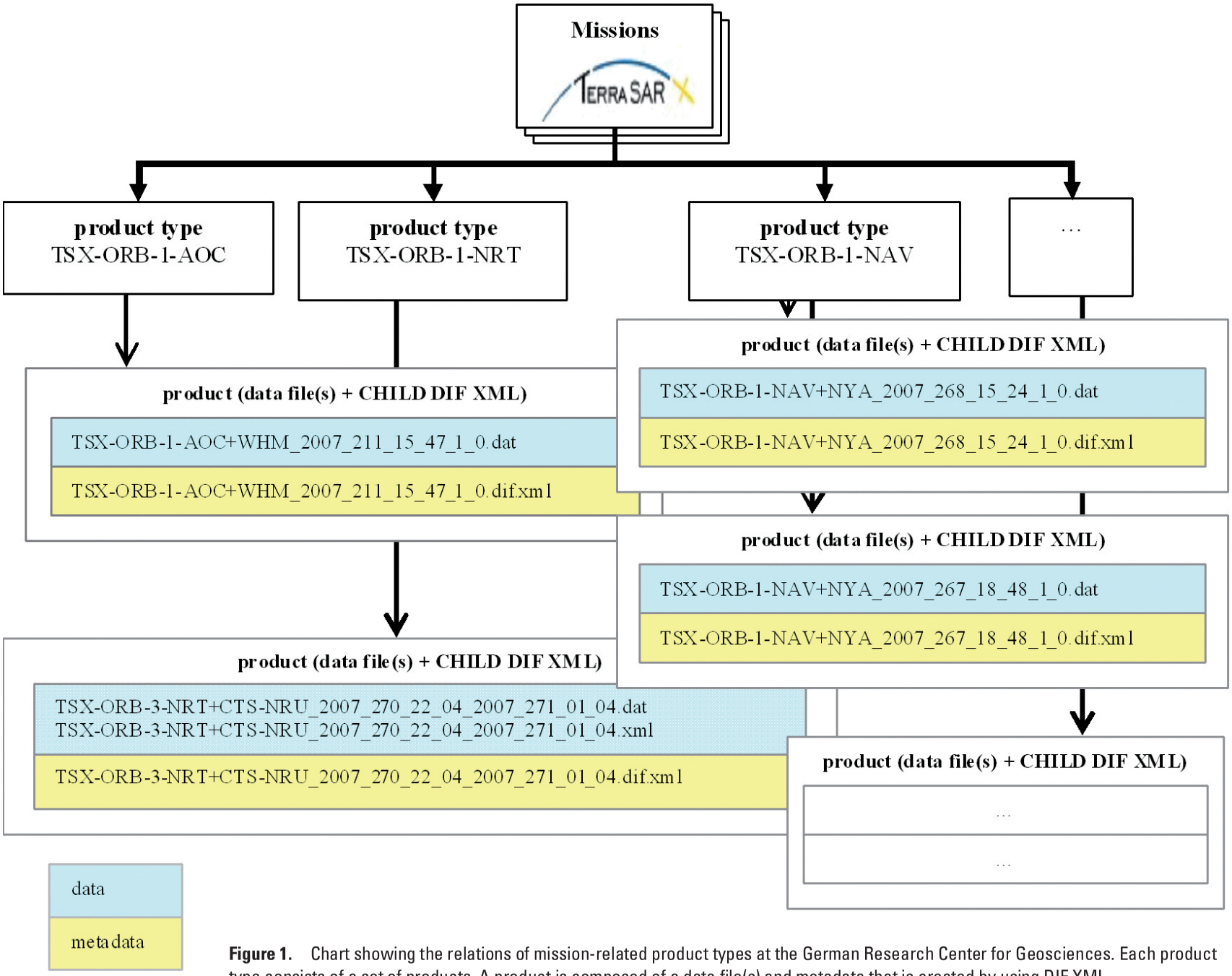


Figure 1. Chart showing the relations of mission-related product types at the German Research Center for Geosciences. Each product type consists of a set of products. A product is composed of a data file(s) and metadata that is created by using DIF XML.

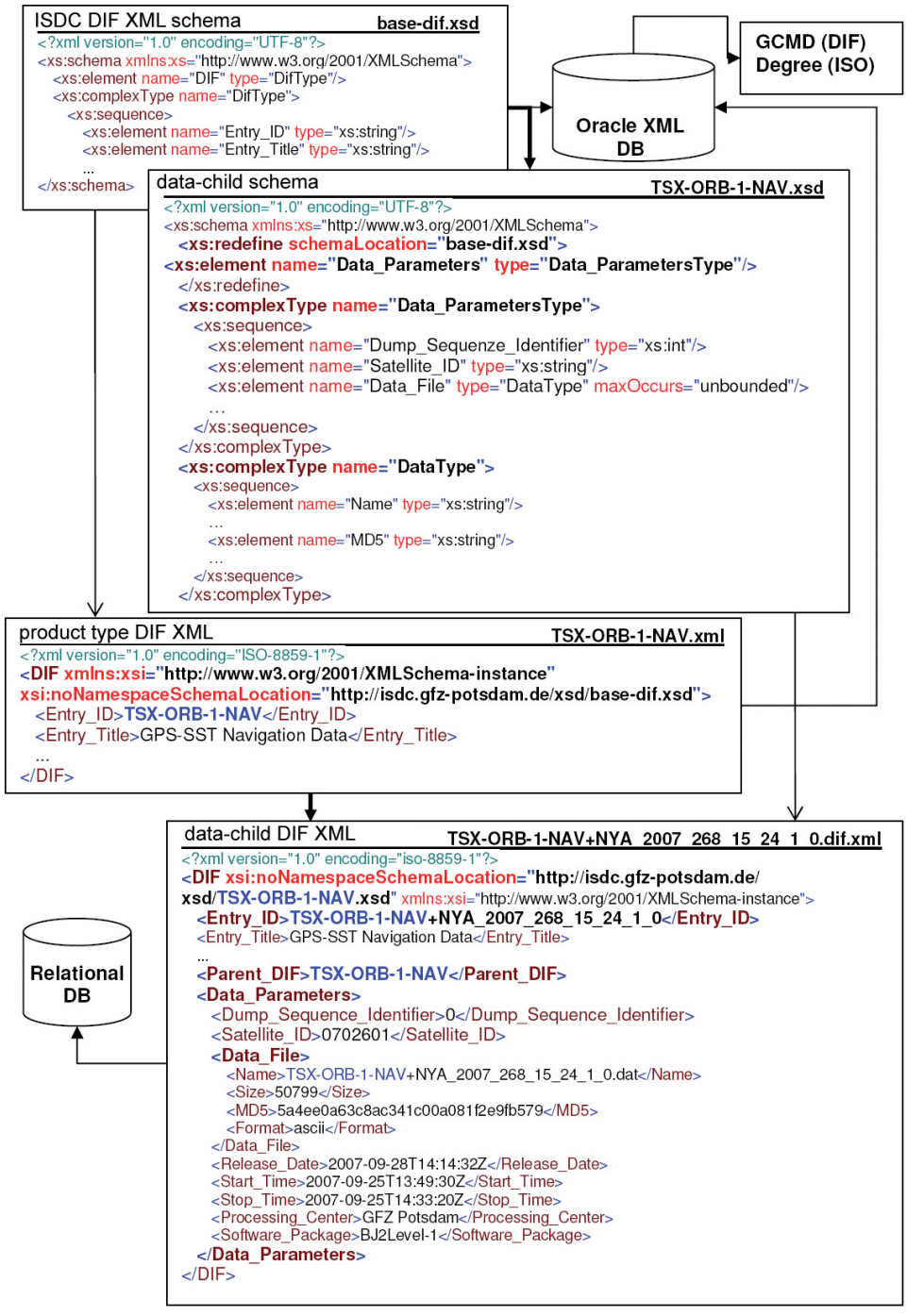


Figure 2. Chart showing the relation between schemata (base and the data-child schema) and the metadata (product-type metadata and child-metadata). The storage mechanisms, such as relational and XML-based databases, also are shown.

is possible by using international standards. The structure of XML easily allows extending the DIF standard in future. Using the parent-child DIF concept, only a small amount of mandatory metadata must be included in both the product type and data-child DIF XML documents.

Reference Cited

International Organization for Standardization, 2003, Geographic information—Metadata: Geneva, Switzerland, International Organization for Standardization, 140 p.

Network of Research Infrastructures for European Seismology (NERIES)—Web Portal Developments for Interactive Access to Earthquake Data on a European Scale

By Alessandro Spinuso,[1] Luca Trani,[1] Sergio Rives,[2] Phetaphone Thomy,[2] Fabian Euchner,[3] Danijel Schorlemmer,[4] Joachim Saul,[5] Andres Heinloo,[5] Rémy Bossu,[2] and Torild van Eck[1]

[1]Observatories and Research Facilities for European Seismology (ORFEUS) Data Center, Royal Netherlands Meteorological Institute, De Bilt, The Netherlands.

[2]European-Mediterranean Seismological Center (EMSC), Bruyères-le-Châtel, France.

[3]Swiss Seismological Service, Swiss Federal Institute of Technology, Zurich, Switzerland.

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

[5]German Research Center for Geosciences, Potsdam, Germany.

The Concept

The Network of Research Infrastructures for European Seismology (NERIES) is a European Commission (EC) project whose focus is networking together seismological observatories and research institutes into one integrated European infrastructure that provides access to data and data products for research. One of the goals of NERIES is to design and develop a Web portal that acts as the uppermost layer of the infrastructure and provides rendering capabilities for the underlying sets of data

Seismological institutes and organizations in European and Mediterranean countries maintain large, geographically distributed data archives. By using the NERIES portal, the broader earth science community and the general public will be able to access earthquake data archives from a single interactive Web site, which will make use of the proper tools and services. This scenario suggested a design approach based on the concept of an internet service oriented architecture (SOA) to establish a cyberinfrastructure for distributed and heterogeneous data streams and services. Recently, this approach has been tested and implemented in Europe by the NERIES consortium and in the United States by the EarthScope consortium. Here we present the current developments within NERIES.

Implementation

Key data formats

The Web services that are currently being designed and implemented will deliver data that have been adapted to appropriate formats. The parametric information about a seismic event is delivered using a seismology-specific Extensible Mark-up Language (XML) format called QuakeML (https://quake.ethz.ch/quakeml), which has been formalized and implemented largely within the NERIES project and in coordination with global earthquake-information agencies. Uniform Resource Identifiers (URIs) are used to assign identifiers to (1) seismic-event parameters described by QuakeML, and (2) generic resources such as authorities, location providers, location methods, adopted software, and so on, described by use of a data model constructed with the resource description framework (RDF) and accessible through a dedicated Web service.

In order to facilitate the exchange in Europe of the parametric information that is specific to seismic events, the European-Mediterranean Seismological Center (EMSC) has implemented a unique event identifier (UNID) that will create the seismic event URI used by the QuakeML data model. The UNIDs and URIs play an important role within the portal developments, and will be used to link and integrate the information retrieved from the different Web services. Access to data such as waveform files (broadband waveform and accelerometric data) will be provided through a dedicated set of Web services that will act as the Hypertext Transfer Protocol (http) interface, which would be the layer on top of the dedicated middleware used by the seismological observatory or institute that is supplying the data.

Software Technologies and Graphical User Interfaces

In order to achieve both the advantages of using a graphical user interface (GUI) and programmatic access to the data, all the Web services are integrated to create different interactive applications. Each single application consists of a Java-based JSR-168-standard portlet (often provided with interactive maps for data discovery) plugged into the centralized NERIES portal implemented through the use of open-source products belonging to the Apache Software Foundations's Portals Project (http://portals.apache.org), such as Jetspeed-2 and WSRP4J.

In specific cases, it will be possible to distribute the deployment of the portlets among the data providers, such as seismological agencies, because of the adoption within the distributed architecture of the NERIES portal of the Organization for the Advancement of Structured Information Stan-

dards' (OASIS') Web Services for Remote Portlets (WSRP) as the standard for presentation-oriented Web services. This approach has been already implemented between EMSC and Observatories and Research Facilities for European Seismology (ORFEUS)

The purpose of the GUI is to provide to the user his own environment where he can surf and retrieve the data of interest that are stored in the portal and managed by the user using the concept of the shopping cart found on some retail Web sites. This approach involves having the user interact with dedicated tools in order to compose personalized datasets that can be downloaded or combined with other information available either through the NERIES network of Web services or through the user`s own cart. For example, an event catalog composed by a user through a specific portlet also will be available to other applications found through the portal in order to obtain derivative measurements such as moment tensors or waveform data volumes.

Administrative applications also are provided to perform monitoring tasks such as retrieving service statistics or scheduling submitted data requests. An administrative tool is included that allows the RDF model to be extended, within certain constraints, with new classes and properties.

Semantically Enabled Registration and Integration Engines (SEDRE and DIA) for the Earth Sciences

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

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

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

We present both the justification and a development initiative to design and implement a pair of service engines that use ontologies for semantically enabled discovery and integration of structurally heterogeneous earth-science data. We also emphasize that the capabilities of these engines are likely to transform earth-science research and education. Our motivation in developing these engines is based on the recognized need to acquire knowledge through advanced semantic capabilities that are able to bridge the science disciplines.

Scientific studies of the Earth and solar system have resulted in massive volumes of data; however, most of the datasets are isolated from each other, and the ability to use these heterogeneous, discipline-specific data to generate "new knowledge" has been limited. In our ongoing research to facilitate the seamless exchange of heterogeneous data, we have developed a Web-based system called Discovery, Integration, and Analysis (DIA) that enables scientists to use ontologies to discover, integrate, and analyze earth-science data (Rezgui and others, 2007). In this paper, we also present the Semantically-Enabled Data Registration Engine (SEDRE), which is a system that complements DIA by enabling scientists to use ontologies to "advertise" their datasets so that they may be automatically discovered by others in the earth-science community. We first summarize our efforts in ontology development for the earth sciences and then present SEDRE to show how it uses ontologies for data registration.

Ontology Development for the Earth Sciences

The role of ontologies for enabling semantic integration is well established (Malik, Rezgui, Sinha, and others, 2007) as it allows a community to associate well-defined, commonly accepted definitions of scientific terms with data. In recent years, several research efforts have recognized the potential of ontologies in promoting data integration in the earth sciences (Sinha and others, 2007). Several ontologies currently are being developed, for instance, the National Aeronautics and Space Administration's (NASA) Semantic Web for Earth and Environmental Terminology (SWEET) and the more data-oriented Earth and Planetary Ontology (EPONT) developed by Malik, Rezgui, and Sinha (2007). EPONT imports and inherits properties from existing ontologies. The availability of these data ontologies is likely to have significant impact in promoting intra- and interdisciplinary interoperability

Overview of DIA

Geoscientists have generated massive volumes of earth science data for decades. Most of the produced data, however, remain as isolated "knowledge islands"; the ability to find, access, and properly interpret these large data repositories has been very limited because of the absence of data-sharing infrastructures that could be used to advertise the data and a common language to properly interpret other providers' data. As a result, it is difficult to answer complex questions that require data from several sources. To address this problem, we have developed DIA, which provides a collaborative environment where scientists can share their resources for the discovery and integration of data by registering them through well-defined ontologies (Malik, Rezgui, and Sinha, 2007).

The DIA engine (Rezgui and others, 2007) provides three functions: discovery, integration, and analysis. Data discovery enables users to retrieve data sets, while data integration enables users to query multiple resources using common attributes to generate previously unknown information: a data product.

DIA's architecture consists of five components:

- User interface—An ArcGIS Server .NET map viewer Web application.
- Two Web servers—The first 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.

- Map server—An ArcGIS map server that provides maps to DIA's query processor.
- Registry servers—Provide directory functionalities (registration of data and tools, indexing, and so on) that providers use to advertise their resources on registry servers.
- Query Processor—Produces the results for users' queries and delivers them to the Web server.

Overview of SEDRE

Semantic data registration is a precursor to improved data discovery and integration. To date, the majority of integrative solutions has been hindered because of the adoption of personal acronyms, notations, and so on, which makes it difficult for other scientists to correctly understand the semantics of the produced data. To address this concern, SEDRE was developed with the goal of allowing researchers to associate one or more ontologies with their data files. For instance, if a researcher wants to obtain data about carbon dioxide, SEDRE will access the data from any dataset where either "carbon dioxide," or "CO2" was used.

SEDRE facilitates discovery through resource registration at three levels:

- Keywords-based registration—Discovery of data resources (for instance, gravity meansurements, geologic maps, and so on) requires registering the use of high-level index terms.
- Ontological class-based registration—Discovering item-level databases requires registration at data-level ontologies.
- Item-level detail registration—Item-level detail or "fine-grained" registration consists of associating a column in a database to a specific concept or attribute of ontology, thus allowing the resource to be queried using concepts instead of actual values. This level of registration is a requirement for semantic integration (that is, for the automatic processing by tools of shared data).

Figure 1 shows the "wiring diagram" for SEDRE. We also show a small subsection of the Planetary Material package (part of EPONT) so that individual datasets containing geochemical analyses with locations (from Planetary Location ontology package of EPONT) can be mapped to terms defined in the ontologies. We recognize that data registration through ontologies is a time-consuming process, and as such, SEDRE has been developed as a downloadable service, where data owners can connect to SEDRE's online repository only to upload the data-ontology mappings. This feature allows data owners to register their data to ontology mappings at their own convenience, while keeping ownership of their data.

SEDRE is designed to be used as a desktop application. Figure 2 shows an example using sulfur-dioxide data. As shown in the figure, sulfur-dioxide data gathered on any given date can be registered to the concept of sulfur dioxide in the EPONT ontology. The conceptual mapping of locations and the analyzed element abundances of liquids, gases, or solids can be captured through the SEDRE user interface. DIA accesses these semantically registered datasets for integration and analyses. We suggest that semantic interoperability challenges can be easily overcome through the deployment of SEDRE and DIA in a Web environment.

Acknowledgments

This research is supported by National Science Foundation award EAR 0225588 to A.K. Sinha and by a subcontract from a NASA award for SEDRE to Peter Fox (University Corporation for Atmospheric Research (UCAR).

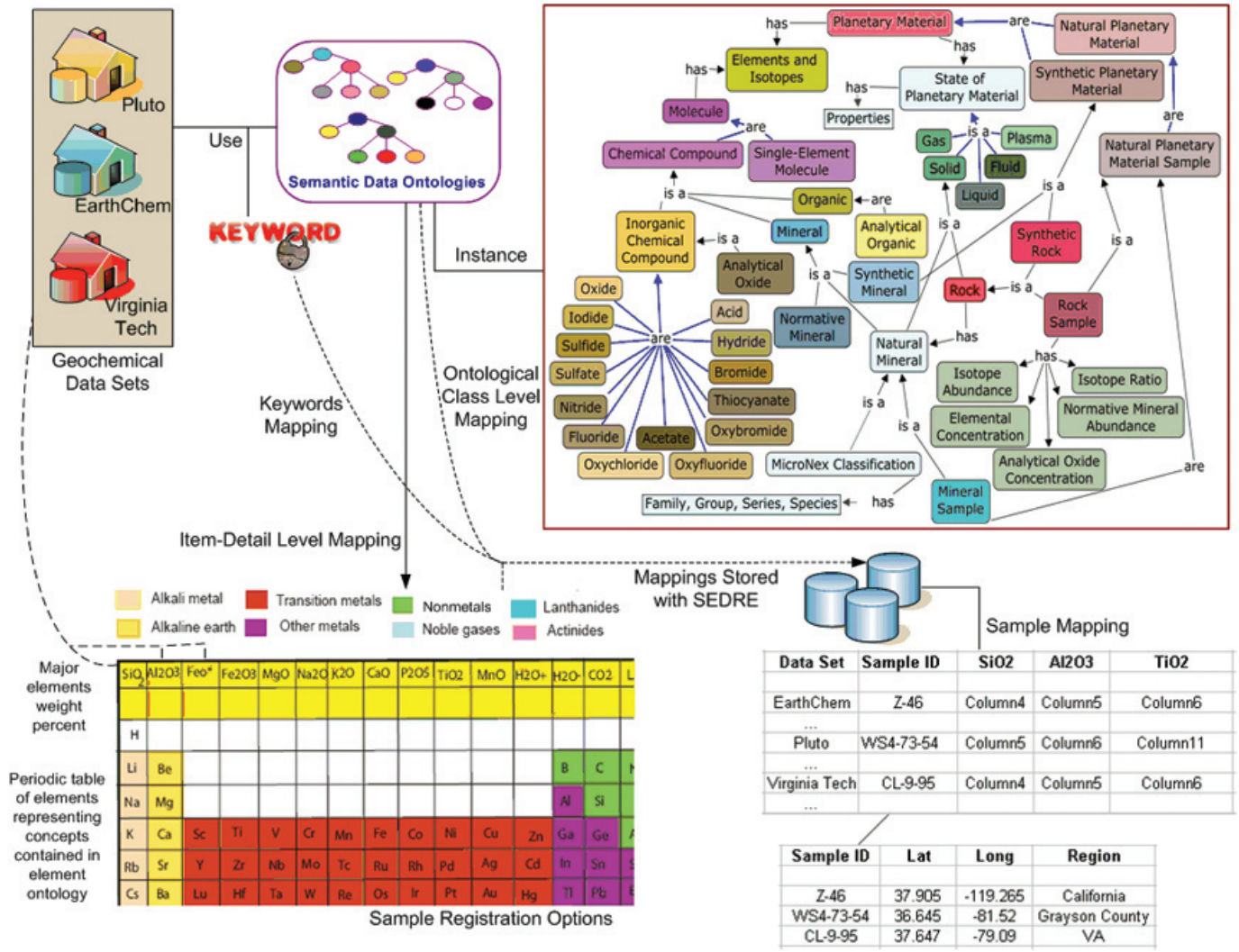
References Cited

Baedecker, P.A., Grossman, J.N., and Buttleman, K.P., 1998, National geochemical data base: PLUTO geochemical data base for the United States: U.S. Geological Survey Digital Data Series DDS-47, 1 CD-ROM.

Malik, Z., Rezgui, A., Sinha, A.K., Lin, K., and Bouguettaya, A., 2007, DIA—A Web services-based infrastructure for semantic integration in geoinformatics, *in* Proceedings, International Conference on Web Services, Salt Lake City, Utah, July 9–13, 2007: New York, N.Y., Institute of Electrical and Electronics Engineers, p. 1,026–1,023.

Malik, Zaki, Rezgui, Abdelmounaam, and Sinha, A.K., 2007, Ontologic integration of geoscience data on the semantic web, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2007—Data to Knowledge, San Diego Calif., May 17–18, 2007: U.S. Geological Survey Scientific Investigations Report 2007–5199, p. 41–43.

Rezgui, Abdelmounaam, Malik, Zaki, and Sinha, A.K., 2007, DIA engine—Discovery, integration, and analysis of earth science data, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2007—Data to Knowledge, San Diego Calif., May 17–18, 2007: U.S. Geological Survey Scientific Investigations Report 2007–5199, p. 15–18.



Data Set	Sample ID	SiO2	Al2O3	TiO2
EarthChem	Z-46	Column4	Column5	Column6
Pluto	WS4-73-54	Column5	Column6	Column11
Virginia Tech	CL-9-95	Column4	Column5	Column6

Sample ID	Lat	Long	Region
Z-46	37.905	-119.265	California
WS4-73-54	36.645	-81.52	Grayson County
CL-9-95	37.647	-79.09	VA

Figure 1. Flow diagram showing the main components of SEDRE. When the data provider publishes the data, SEDRE tracks their origin, assigns a registration number, and monitors their provenance. Data with a corresponding location can be registered through SEDRE. Geochemical data sets such as PLUTO (Baedecker and others, 1998) can be registered through a graphical user interface (linked to ontologies) that allows a user to map a geochemical measurement (for example, sulfur dioxide) to a corresponding concept shown in Ontological Class Level Mapping. The mapped product is also shown in the bottom-right corner.

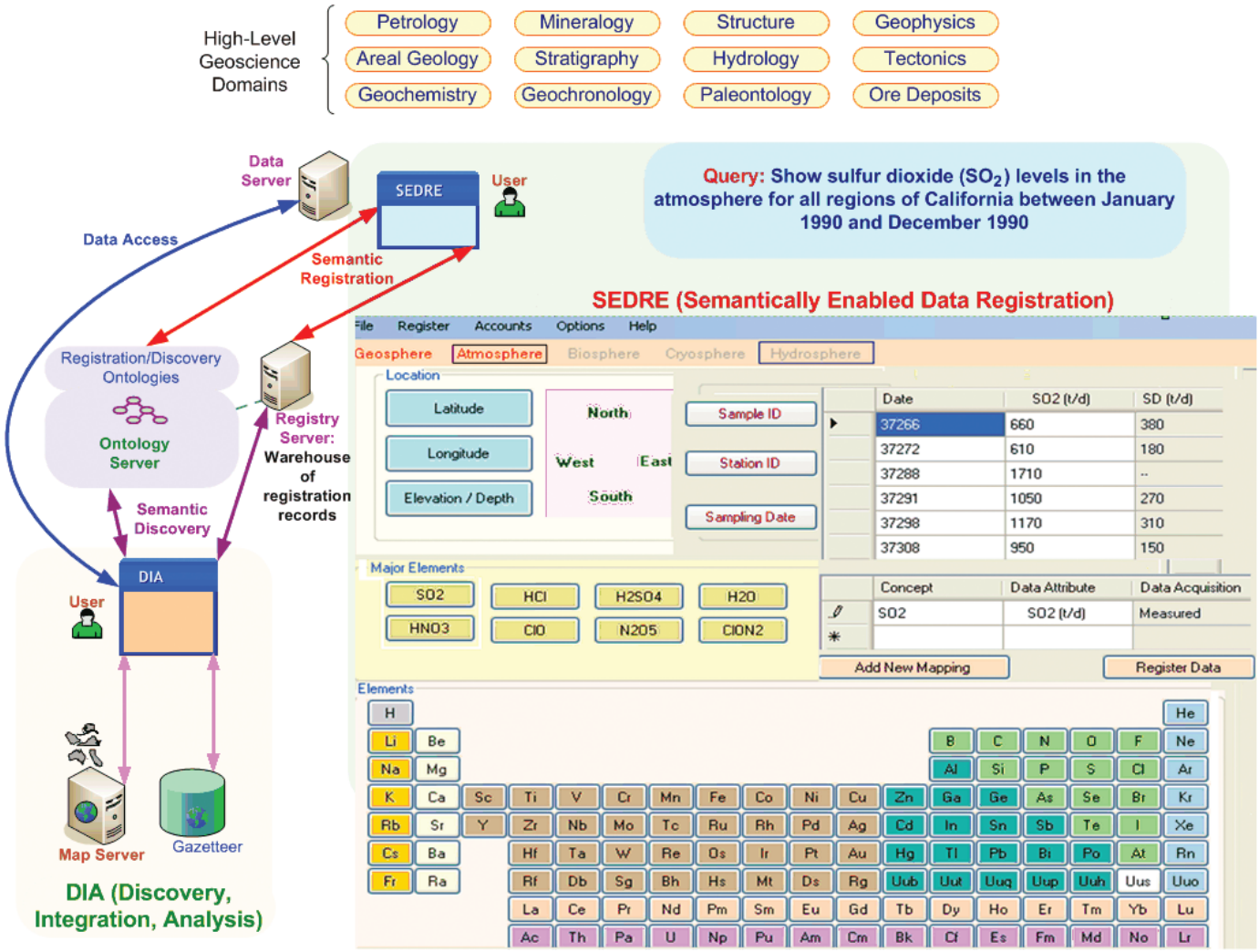


Figure 2. Schematic representation of showing the registration of data through SEDRE and their discovery and integration through DIA. To answer a query about sulfur dioxide, a user (data provider) registers sulfur-dioxide measurements (tons per day) to create an association (a map) between the ontologic concept and the data (SEDRE). The mapped data can be discovered by the DIA system, which uses the same ontological infrastructure.

Oral Session III

Semantic Provenance for Image Data Processing

Peter Fox,[1] Deborah L. McGuinness,[2,3] Paulo Pinheiro da Silva,[4] Stephan Zednik,[1] Jose Garcia,[1] Li Ding,[2] Nicholas Del Rio,[4] and Cynthia Chang[2]

[1]High Altitude Observatory, Earth and Sun Systems Laboratory, National Center for Atmospheric Research, Boulder, Colo.

[2]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, N.Y.

[3]McGuinness Associates, Latham, N.Y.

[4]Department of Computer Science, University of Texas—El Paso, Tex.

Introduction

In some virtual observatories, many of the functions that an end-user wishes to perform (in our terminology: use cases) are difficult to implement. This is a limiting factor when trying to make diverse image datasets available to a broad user base. Some real-life use cases are as follows:

1. What algorithms or methods were used in the process to obtain the Coronal Helium I Imaging Photometer (CHIP) solar limb "image of the day" for January 26, 2005 at the Mauna Loa Solar Observatory?

2. What were the cloud cover and atmospheric seeing conditions during the local morning of January 26, 2005 at the Mauna Loa Solar Observatory?

3. Why does this image look bad?

In these sample use cases, it is important to note that the required provenance information was either not collected from different stages in the data-processing pipeline, or it was collected but it was not propagated (because the pipeline is not fully integrated and does not have the necessary software instrumentation, or because some information was generated manually and was not annotated). In a Semantic Web context, the metadata about information acquisition is called "knowledge provenance." We describe the provenance requirements that have emerged in our previous work on virtual observatories as well as requirements that we identified from a series of use cases collected from scientific data users and instrument scientists. We also describe our progress in the context of solar physics image-data-processing pipelines and discuss the general applicability of our work to data ingest in general.

Virtual observatories are examples of the growing trend of supporting distributed interdisciplinary scientific research; with a wider range of users come additional requirements and changing priorities. We present the provenance requirements from work on virtual observatories (described by Del Rio and others, 2007, and McGuinness and others, 2007) in several scientific communities.

We use knowledge provenance in a broad sense to include the origins of knowledge in any virtual system. Knowledge provenance includes sources of raw data, experiments used to generate data, processing applied to the data, and so on. In this work, which was conducted as part of the Semantic Provenance Capture in Data Ingest Systems project (SPCDIS, http://spcdis.hao.ucar.edu), we provide an extensible representation for provenance in data systems in general and for the Virtual Solar-Terrestrial Observatory (http://www.vsto.org) in particular. To understand which elements we need to capture, along with the use cases, we also need to collect and assemble relevant documentation.

An example of how this documentation is collected is shown in figure 1 for the CHIP instrument. The central part of the diagram indicates the specific artifacts created from previous artifact(s) by the processes (left). The left side of the diagram indicates processes that are performed and additional information that is available or added. The right side indicates the specific people who have primary responsibility for the stage, as well as which portion takes place in Hawaii or in Boulder.

Inference Web and the Proof Markup Language

Inference Web (McGuinness and others, 2003, 2004) is a knowledge provenance infrastructure that supports comprehensive explanation capabilities. Those capabilities include interoperable explanations of sources (that is, sources published on the Web or accessible from files), assumptions, learned information, and answers (for example, scientific results) that are associated with inferred or stated conclusions. This knowledge provenance information may be used to improve users' trust regarding those conclusions and thus may make systems that include knowledge provenance support upon which we are able to provide answers to questions that a user wants to ask. The tools provided by the Inference Web suite are focused on scientific information and knowledge provenance for scientific workflow; its data may include datasets, visualizations, and simulations, all of which expand the range of data types that the IW infrastructure must support.

Inference Web provides the Proof Markup Language (PML) (Pinheiro da Silva and others, 2004; McGuinness and others, 2007) to encode justification information about any kind of response produced by agents. PML justifications are graphs with the edges always pointing towards the final justification conclusion and they store provenance about the associated information sources.

Probe-It! is a browser that graphically renders the provenance information associated with results coming from both inference engines and scientific workflows. Probe-It! assumes that a user has provided existing provenance resources that may be viewed. Here our main interest is to display in text or visualize the provenance for scientists to better understand several aspects of the quality of CHIP images.

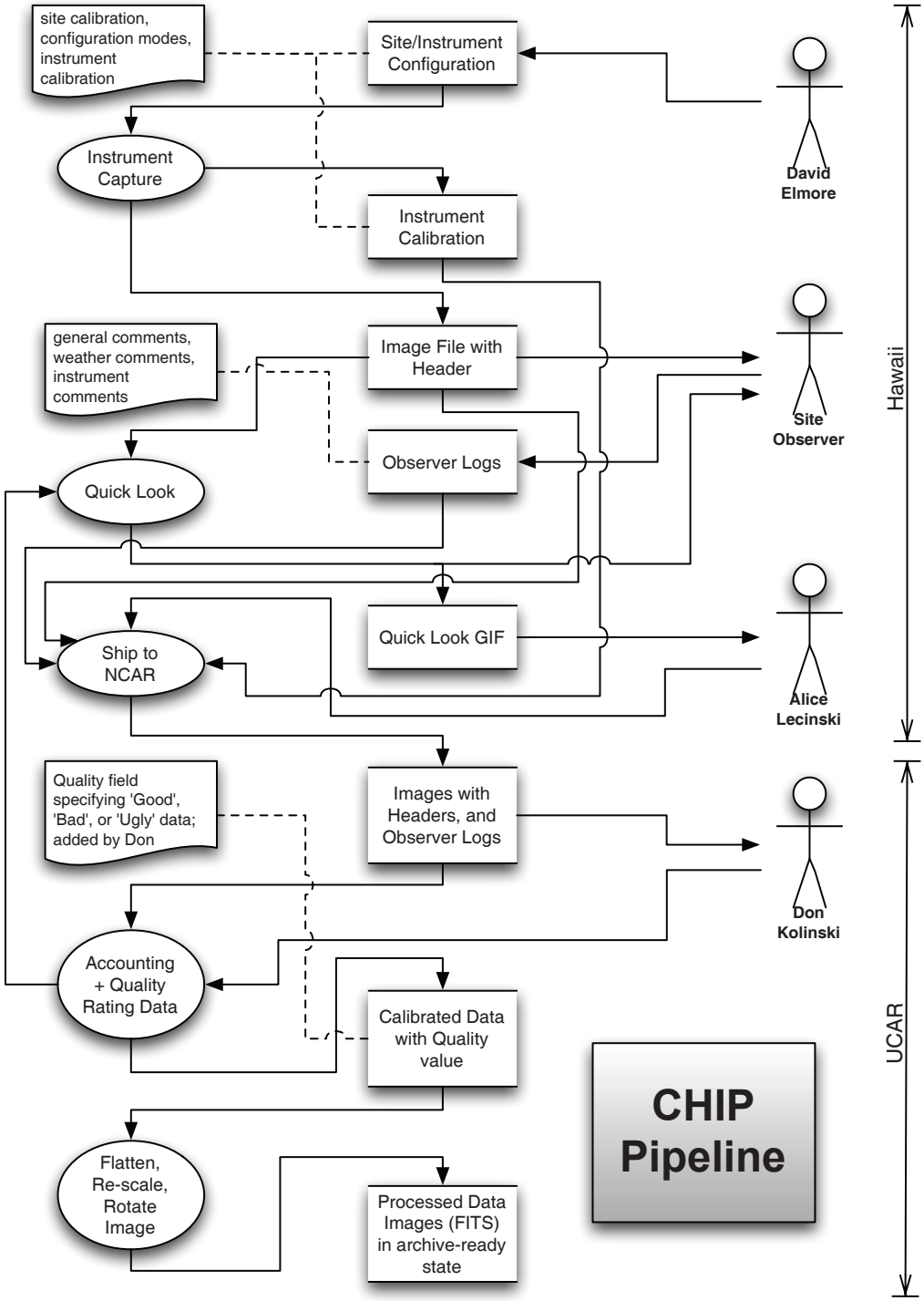


Figure 1. Graphical representation of the CHIP data pipeline, including key artifacts, actors and the roles they play, algorithmic processes applied along the pipeline, and ancillary data added to the artifacts as they progress through the pipeline.

Because provenance associated with results from small scientific workflows can become large and incomprehensible as a whole, Probe-It! provides multiple views suited to different elements of provenance: query, global, and local. The conclusion is encoded in an Extended QuickLook (EQL) format, which we have developed, and the views show how the conclusion was derived. The resultant PML documents contain both workflow provenance and workflow lineage (workflow results generated by scientific Web services); thus the browser is capable of rendering maps, tables, and other scientific artifacts.

Provenance—Identification and Mark-Up

The provenance information we collect depends on the processing stage. For instance, in figure 1, the middle part of the pipeline from the site observer to Alice Lecinski involves artifacts and processes in Hawaii and Boulder for the creation of quick-look (QL) images. We needed specific augmentation of the artifacts that were obtained while producing the QL images. The images are used (1) on the mountain in Hawaii for observers to judge image quality and instrument perfor-

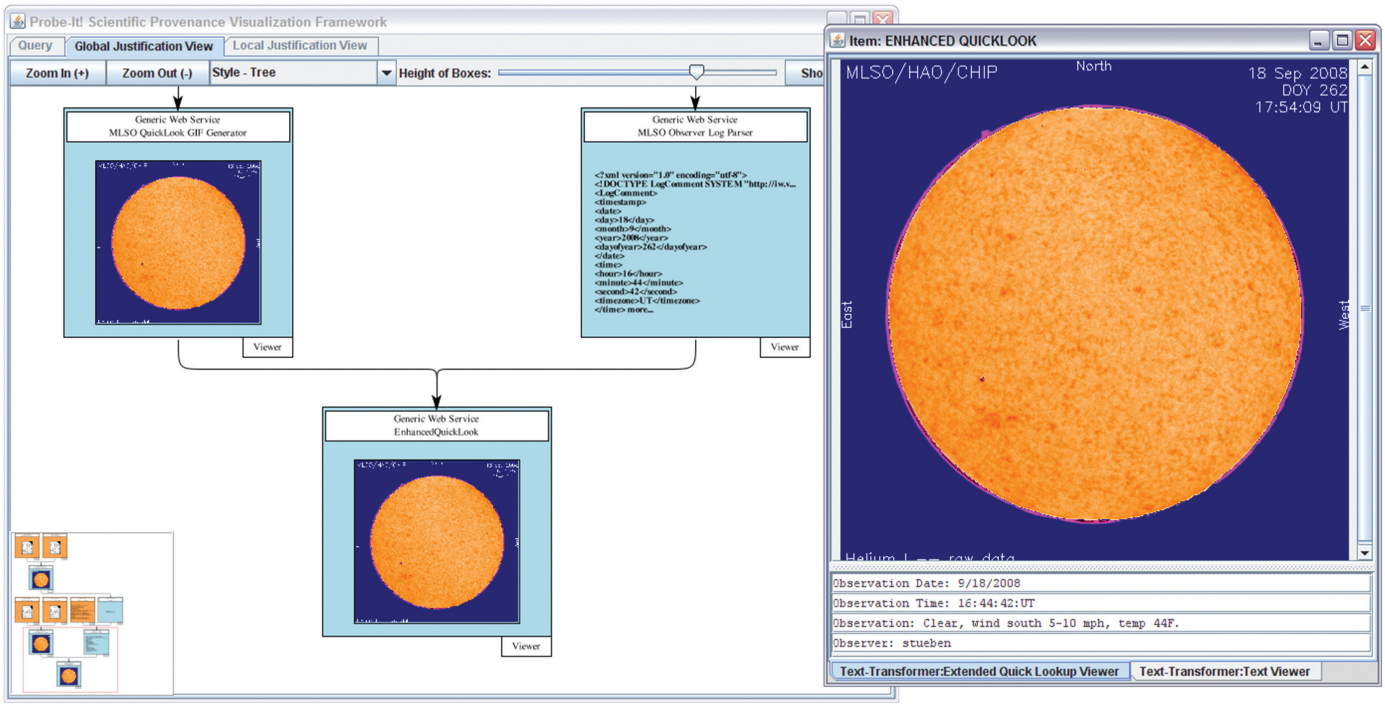


Figure 2. An example of the use of the Probe-It! browser application. The full provenance trace is in the lower left corner. The main part of screen shows an enlarged view of the lower portion of the provenance trace. The image of the Sun is enlarged and allows annotations by the data provider.

mance, or to note features of interest on the Sun; and (2) in Boulder for placement on the Mauna Loa Solar Observatory (MLSO) home page (http://mlso.hao.ucar.edu).

Figure 2 shows a result from the browser application Probe-It! when it is applied to the EQL images. Information has been extracted from the observer log and presented in a structured form to indicate who added the comments, when they were added, and that the sky was clear.

Discussion and Conclusion

We have provided a valuable addition to one part of an image-processing data pipeline for solar images taken by an instrument at the MLSO. The creation of EQL images in support of answering questions such as, "What were the weather and observing conditions for this QL image?" is providing significant added value to those users who monitor the data pipeline and instrument performance. We have introduced structured observer log information into the explanation documentation (not previously available) and created PML instances by using two sets of tools to browse and search these explanations. The next stage of development will involve other parts of the pipeline to benefit instrument designers, project scientists, analysts, and end users.

Acknowledgments

The Semantic Provenance Capture in Data Ingest Systems project is funded by the National Science Foundation (NSF) Office of Cyberstructure's Software Development for Cyberinfrastructure program grant OCI-0721943. The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research with substantial sponsorship from the NSF. We also received support from NSF grant HRD-0734825 to the Center of Excellence for Sharing Resources for the Advancement of Research and Education (Cyber-ShARE) at the University of Texas—El Paso.

References Cited

Del Rio, Nick, and Pinheiro da Silva, Paulo, 2007, Probe-it! Visualization support for provenance, *in* Bebis, George, Boyle, Richard, Parvin, Bahram, Darko, Koracin, Paragios, Nikos, Syeda-Mahmood, Tanveer, Ju, Tao, Liu, Zicheng, Coquillart, Sabine, Cruz, Neira, Carolina, Müller, Torsten, and Malzbender, Tom, eds., Advances in Visual Computing, Proceedings of the Third International Symposium on Visual Computing, Lake Tahoe, Nev., November 26–28, 2007: Lecture Notes in Computer Science, v. 4842, p. 732–741.

Del Rio, Nick, Pinheiro da Silva, Paulo, Gates, A.Q., and Salayandia, L., 2007, Semantic annotation of maps through knowledge provenance, *in* Fonseca, Frederico, Rodríguez, M.A., and Levashkin, Sergei, eds., Geospatial semantics, Proceedings of the Second International Conference on Geospatial Semantics (GeoS), Mexico City, Mexico, November 29–30, 2007: Lecture Notes in Computer Science, v. 4853, p. 20–35.

McGuinness, D.L., Ding, Li, Pinheiro da Silva, Paulo, and Chang, Cynthia, 2007, PML 2—A modular explanation interlingua, *in* Roth-Berghofer, Thomas, Schulz, Stefan, Bahls, Daniel, and Leake, D.B., eds., Explanation-aware computing, Papers from the 2007 Association for the Advancement of Artificial Intelligence Workshop, Vancouver, British Columbia, Canada, July 22–23, 2007: Association for the Advancement of Artificial Intelligence Technical Report WS–07–06, p. 49–55.

McGuinness, D.L., Fox, Peter, Cinquini, Luca, West, Patrick, Garcia, Jose, and Benedict, J.L., 2007, The Virtual Solar-Terrestrial Observatory—A deployed semantic web application case study for scientific research, *in* Cheetham, William, and Goker, Mehmet, eds., Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence, Vancouver, British Columbia, Canada, July 22–26, 2007: Menlo Park, Calif., Association for the Advancement of Artificial Intelligence Press, p. 1,730–1,737.

McGuinness, D.L., and Pinheiro da Silva, Paulo, 2004, Explaining answers from the Semantic Web—The Inference Web approach, *in* Sycara, Katia, and Mylopoulous, John, eds., Proceedings, Second International Semantic Web Conference, Sanibel Island, Fla., October 20–23, 2003: Journal of Web Semantics, v. 1 no. 4, p. 397–413.

Pinheiro da Silva, Paulo, McGuinness, D.L., and Fikes, Richard, 2006, A proof markup language for Semantic Web services, *in* Bell, David, Bussler, Christoph, and Yang, Jian, eds., The Semantic Web and Web services: Information Systems, v. 31, no. 4-5, p. 381–395.

Effective Future Use of Current Remotely Sensed Data Sets to Study Long-Term Climate Changes

By Albert J. Fleig[1] and Curt Tilmes[2]

[1]PITA Analytic Sciences, Bethesda, Md.

[2]Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Md.

Determining changes of a geophysical parameter over time requires having a way to estimate the parameter values at both ends of the time range under consideration. Measurements being made now will provide the beginning of the data record for studies to be done in the future. Unfortunately, present standards for archiving remotely sensed datasets do not cover all the information that future scientists will need to precisely understand and effectively use the current data. Even more unfortunately, the scientists who made the current data sets will be long gone and unable to answer the resulting questions. The problem is that any given measurement system has both a relative accuracy or precision and an absolute accuracy. Typically the precision of a "validated" measurement system is much better known than the absolute accuracy. Trends can be estimated from precise measurements as long as the same measurement system is continued in use and the fundamental geophysical situation does not change from that present when the dataset was validated; however, no instrument lasts forever and improved measurement techniques often lead to fundamentally different ways of measuring the same parameter. Over long time periods, geophysical values that were assumed as inputs for a given measurement system (input as static climate values rather than measured quantities) may also change. These changes in instruments, measurement and algorithm technology, geophysical situations, and knowledge of ancillary data might not be a great problem if the absolute accuracy of both systems is known and if the trend to be investigated involves changes that are substantially larger than the difference between the absolute accuracy of the current and future measurements. If this is not the case, then it would be possible for a future scientist to estimate the size of these changes and their impact on the resulting long-term trend if (but only if) he or she knew exactly how the original dataset was made. Unfortunately, current practice does not require this information to be archived. Journal articles, subject to page limitations, describe the approach but not the details of the implementation. Algorithm Theoretical Basis Documents currently produced for many of the measurement systems at the National Aeronautics and Space Administration (NASA) are longer and more detailed; however, they are written before launch and are not updated at the end of the mission to document the changes between what was intended and what was actually done. The only thing that is sure to be consistent with the archived measurements is the source code that was used to make the measurements. If the source code is augmented with specific identification and archiving of all of the inputs to the processing (for every production run) and the exact details of the processing system itself, then it becomes possible to replicate and understand the results.

We have developed and are currently running a production system, described by Tilmes and Fleig (2008, this volume), that automates the collection and storage of all the provenance information described above. In order for a future scientist to do sensitivity studies of the impact of various input and algorithm changes, there are several more things that need to be automatically collected and stored. They include detailed information about how tables used in the algorithm were developed and an explanation of what the code was doing at every step of the process. There are structural changes in the current data-management approach and best-practice procedures that can minimize the above problems.

Scientists who understand and have made current datasets will not be available to explain them forever. It is time to decide whether the information necessary to understand exactly how a dataset was produced should be provided as a general matter of science policy. Decisions about requiring the necessary changes and best practices should be made as the result of conscious evaluation of their cost and a consideration

of the impact of being unable to understand exactly how the past data were made, rather than being left by default to the current approach.

Reference Cited

Tilmes, Curt, and Fleig, A.J., 2008, Standardizing interfaces for external access to data and processing for the National Aeronautics and Space Administration's Ozone Product Evaluation and Test Element (PEATE), *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Geoinformatics 2008—Data to Knowledge: U.S. Geological Survey Scientific Investigations Report 2008–5172, p. 21–22 (this volume).

Peer-Reviewed, Open Data Publication as a Means of Data Quality Management in Research

By Hans Pfeiffenberger[1] and David Carlson[2]

[1]Alfred Wegener Institute, Bremerhaven, Germany.

[2]International Program Office, International Polar Year, Cambridge, United Kingdom.

A few years ago in the earth-science data community, a discussion began that sought to answer the following questions: How can data be cited, should the data be published, and, if so, how could they be published? At the same time, the community debated whether and why data can or should be openly accessible. The newly launched journal entitled "Earth System Science Data" is possibly the first journal devoted exclusively to the peer-reviewed publication of datasets. The journal's publication policy attempts to provide some specific answers to these questions in order to address data-quality management (especially in publicly funded research).

Organizations (such as exploration companies or public meteorological services) that collect data as an essential part of their business have instituted internal instruments and procedures for data-quality management. This is to be expected where financial or other success directly depends on data quality, especially if the collected data can be reused and can thus constitute "capital."

By contrast, publicly funded research is almost exclusively driven by the requirement to "publish or perish." In many public organizations, there is no financial or other type of bonus for data-quality management. Consequently, in most disciplines and institutions, data-quality management is regarded as overhead—and that's if an institutional awareness or a policy about data-quality management even exists. At the level of an individual researcher, there are more facets to the problem: the researcher will regard collected data as the basis for the next, or even many more publications, and thus the data become his personal "capital." This motivation, however, will not result in a systematic drive towards neutral data-quality management practices that must typically involve third parties.

The way that public institutions traditionally handle quality management led to the belief that publishing data by means of peer-reviewed journal articles as well is a concept that ideally unites a well-known and respected instrument with the necessary incentive (namely, an increase in the publication count of individuals or institutions by publishing a dataset). The special form of two-stage open peer review has been practiced by a number of high-impact journals in earth science topics for some years. Owing to this specific method, it is necessary that the published datasets be openly available to reviewers, commentators and, finally, the journal readers. Public peer review combines the power of peer review with the well-known purging effect seen elsewhere in the world of open-source software.

We will discuss the emerging rules and practices (for example, authors' and reviewers' guidelines) of Earth System Science Data and try to predict their effectiveness and limits regarding data-quality management in research. Finally, we will discuss (1) the role of this journal with respect to data centers or data repositories (and the people who contribute to and operate them) and (2) the relations and technical links between the journal articles and the datasets they are about.

Long-Term Availability of Geoscience Data

By Jens Klump[1]

[1]German Research Center for Geosciences, Potsdam, Germany.

Introduction

In the last decade, research in the geological sciences has produced vast amounts of new data. In some cases, it is the enormous volume of data that poses a technical challenge; in other cases, it is their semantic complexity. Regardless of the volume and format, geoscience data are characterized by their origin in a heterogeneous and dynamic research environment. Workflows in a business or administrative context are characterized by their transactional behavior, whereas scientific workflows may require ad-hoc changes that become necessary through the incorporation of new results into experimental working hypotheses (Barga and Gannon, 2007).

To the individual scientist, data curation is not at the focus of scientific work and there are few incentives for scientists to make data accessible for re-use or re-purposing. Only few science-funding agencies ask grant recipients to make their data accessible, and even fewer journals make data access a prerequisite for publication. Furthermore, the roles and responsibilities in long-term curation of scientific data still need to be resolved (Lyon, 2007). This situation leads to deficits in data management that put large portions

of our scientific heritage at risk of loss. The inaccessibility of data might have a negative impact on the quality of research (Nature, 2006).

Achieving sustainable, long-term accessibility and reusability of research data requires a combination of organizational and technical measures. On the organizational side, data curation needs to become an integral part of good scientific practice; at the same time, the geoinformatics community has to develop tools that facilitate the tasks needed for efficient and sustainable data curation.

Organizational Strategies

Several declarations by governmental, nongovernmental, and scientific bodies have called for open access to data and for better accountability for the long-term preservation of data, but with little success. Several studies (Lyon, 2007, and Klump, 2008, among others) have investigated the requirements needed for this effort. These studies also report on best-practice examples from existing data repositories.

A key to devising effective and sustainable strategies for the long-term preservation and accessibility of research data is to define "levels of persistence" in the data-curation process and its supporting technical architecture. The domains of collaboration and publication, with respect to research data, are not discrete but rather form the end-points of a "curation continuum." The implementation of data-curation processes, however, requires the definition of a boundary between the two domains in order to distinguish the roles and responsibilities of the actors in the data-curation processes. The idea is to distinguish the domain of active research, where curation is the responsibility of the scientists (collaboration domain), from the domain of long-term preservation (publication domain), where responsibility and expertise lie with the "memory institutions" such as a library or data center (Treloar and others, 2007). The diagram in figure 1 shows the two data-curation domains and the "curation boundary" with its interface between a university's research groups and its memory institution.

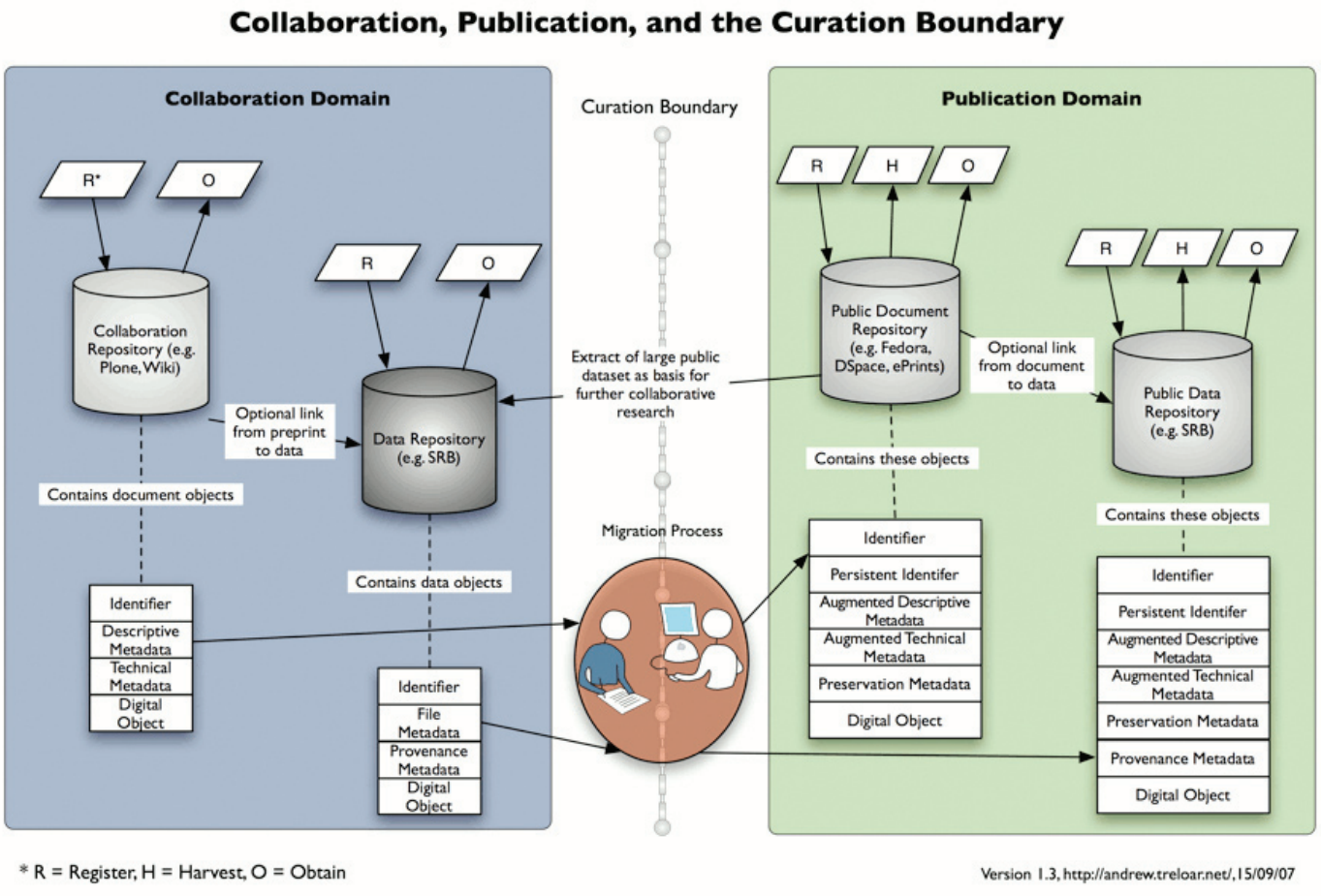


Figure 1. Schematic diagram showing data curation in the collaboration domain and in the publication domain. Note that objects in the publication domain require comprehensive metadata; however, in the collaboration domain, metadata are available only as implicit information. From Treloar and others (2007, used with permission).

Technological Strategies

Some scientific disciplines already have repositories for their data; however, for the majority of researchers, no data repositories exist. One example where best practices for disciplinary data repositories are used is at the World Data Centers (WDC) of the International Council for Science (ICSU). The use of repositories for data curation on an institutional level is still a relatively recent idea (Lyon, 2007; Treloar and others, 2007).

In most cases, the existing disciplinary data repositories are not integrated into the scientific workflow, which leads to only a small proportion of the data being archived in disciplinary repositories. This break in the workflow also is reflected in the problems observed in the generation and curation of metadata. More research needs to be done to determine (1) which kind of metadata are needed at which level of data curation (Treloar and others, 2007) and (2) how metadata can be generated automatically in the data-curation processes (Robertson, 2006).

The heterogeneity of data in the geological sciences requires special attention to data and file formats. Not all formats that are popular among scientists are suitable for long-term preservation (Lormant and others, 2005). This also means that preservation metadata need to encode more of the data format than, for instance, just their Multipurpose Internet Mail Extension (MIME).

Re-use and Re-purposing of Data

Data curation and long-term preservation of digital research data are motivated by the goal of both scientists and institutions to re-use and re-purpose research data that already exist. This goal will be achieved only if the use and citation of data become part of scientific culture. Without demand from scientists, none of the data repositories can be operated on a sustainable basis. This effort requires that a scientist's or institution's data holdings can be found through catalogs and portals and that the published data can be cited in future work (Klump and others, 2006).

Uniform Resource Locators (URLs) are transient and, therefore, are not suitable as a means of referencing data for the purpose of citation. The shortcomings of URLs are overcome by the use of persistent identifiers, such as the Digital Object Identifier (DOI) and Uniform Resource Names (URNs) (Altman and King, 2007).

Conclusions

The introduction to the Open Archival Information Systems reference model (International Organization for Standardization, 2003) describes a digital archive as "an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community." Data curation and long-term preservation of scientific data are, therefore, not only technical issues; they also need an appropriate organizational framework.

Successful approaches to long-term availability of data need to recognize the roles and responsibilities in the data-curation process. The identification of actors in the process is needed in order to identify the right tools and incentives that are necessary components of technical and organizational strategies for long-term availability of data.

References Cited

Altman, Micah, and King, Gary, 2007, A proposed standard for the scholarly citation of quantitative data: D-Lib Magazine, v. 13, nos. 3-4, available only online at http://dlib.org/dlib/march07/altman/03altman.html/. (Accessed August 18, 2008.) (doi:10.1045/march2007-altman)

Barga, Roger, and Gannon, D.B., 2007, Scientific versus business workflows, *in* Taylor, I.J., Deelman, E., Gannon, D.B., and Shields, M., eds., Workflows for e-science: London, United Kingdom, Springer-Verlag, p. 9–16.

International Organization for Standardization, 2003, ISO 14721:2003, Space data and information transfer systems—Open archival information system—Reference model: Geneva, Switzerland, International Organization for Standardization, 141 p.

Klump, Jens, 2008, Requirements of e-science and grid technology for scientific data archiving: Nestor Studies 9, 50 p.

Klump, Jens, Bertelmann, Roland, Brase, Jan, Diepenbroek, Michael, Grobe, Hannes, Höck, Heinke, Lautenschlager, Michael, Schindler, Uwe, Sens, Irina, and Wächter, Joachim, 2006, Data publication in the Open Access Initiative: Data Science Journal, v. 5, p. 79–83. (doi:10.2481/dsj.5.79)

Lormant, Nicolas, Huc, Claude, Boucon, Daniele, and Miquel, Christine, 2005, How to evaluate the ability of a file format to ensure long-term preservation for digital information?, *in* Proceedings, Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data (PV 2005), Edinburgh, United Kingdom: Bath, United Kingdom, United Kingdom Office for Library and Information Networking, 11 p., available only online at http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf. (Accessed August 18, 2008.)

Lyon, Liz, 2007, Dealing with data—Roles, rights, responsibilities and relationships; consultancy report: Bath, United Kingdom, United Kingdom Office for Library and Information Networking, 65 p., available only online at http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf/. (Accessed August 18, 2008.)

Nature, 2006, A fair share (editorial): Nature, v. 444, no. 7120, p. 653–654.

Robertson, R.J., 2006, Evaluation of metadata workflows for the Glasgow ePrints and DSpace services: Glasgow, United Kingdom, University of Strathclyde, 48 p., available only online at http://eprints.cdlr.strath.ac.uk/2374/01/Robertson_DAEDALUS_Metadata_Workflow_Evaluation.pdf. (Accesed August 18, 2008.)

Treloar, Andrew, Groenewegen, David, and Harboe-Ree, Catherine, 2007, The data curation continuum—Managing data objects in institutional repositories: D-Lib Magazine, v. 13, nos. 9-10, available only online at http://dlib.org/dlib/september07/treloar/09treloar.html/. (Accessed August 18, 2008.) (doi:10.1045/september2007-treloar)

Towards an OpenEarth Framework (OEF)

By Chaitan Baru,[1] G. Randy Keller,[2] David R. Nadeau,[1] and John L. Moreland[1]

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

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

Integrating Data and Toolkits

In 2002, the Geosciences Network (GEON) brought together 16 institutions to develop an infrastructure for managing distributed collections of large, heterogeneous, multidisciplinary datasets. In the years since then, this infrastructure has expanded to include open-source software for integrating, analyzing, and visualizing these datasets. We call this software suite the OpenEarth Framework (OEF) because it is a community-driven set of open standards for data models and services. A principal focus of this work is integration that spans the following:

- Data types—Light detection and ranging (LiDAR), satellite, and other types of imagery; digital elevation models (DEMs), borehole samples, velocity models from seismic tomography, gravity measurements, and simulation results.
- Data storage schemes—File systems, databases, and archiving systems such as the Storage Resource Broker (SRB).
- Data delivery methods—Local files, database queries, Web services (such as Web Map Service (WMS) and Web Feature Service (WFS)), and services for new data types, such as large tomographic volumes.
- Data formats—Shapefiles, Network Common Data Format (NetCDF), Geostationary Earth Orbit Tagged Image File Format (geoTIFF), and other formal and commonly practiced standard formats.
- Data models—From two- and three-dimensional geometry to semantically richer models of features and relationships between those features.
- Coordinate systems—Including two- and three-dimensional spatial representations as well as coordinate systems for time scales that may span hundreds of millions of years.

There are several good toolkits that address portions of this space, including GeoTools, GeoTIFF libraries, Shapefile parsers, and many Web services toolkits. The OEF seeks to integrate them within a common framework. By spanning multiple toolkits, the OEF can grow as the sum of these toolkits grows. The OEF can also remain "toolkit agnostic" and expand to support additional toolkits as they emerge.

Visualizing Data

The OEF also is addressing gaps in these toolkits. Chief among these is a lack of support for interactive three-dimensional visualization. While many three-dimensional visualization tools exist, they often have built-in assumptions that limit their use within geoscience contexts. Google Earth, for instance, is a fascinating tool for exploring surface features, but it has no support for diving beneath the surface. Tools tuned for time lines on the scale of earthquake cycles may be insufficient to handle deep time for exploring the evolution of the lithosphere and linkages between such phenomena as orogeny and climate. Other tools may do very well when datasets are small enough to fit entirely in memory, but they stumble on large high-resolution data spanning continents and millions of years.

With the OEF's focus on integrating data that span the geosciences, it is important to develop an open software architecture and corresponding software that can properly manipulate and visualize the integrated data. Because of this requirement, the OEF's software stack extends from deep within data archives available through the Web outwards to interactive visualization tools running on the user's desktop or laptop computer.

Figure 1 shows three types of services that have been developed to accomplish interactive three-dimensional visualization and analysis. At the deepest level, Dataset Access Services manage and deliver stored data and metadata. These services hide storage details, such as the storage medium, Internet location, administrative domain, access authentication, data replication, and storage optimization. Stored metadata characterizes registered data, including its spatial and temporal extent, resolution, and history. Of particular importance is a data derivation tree that links together original

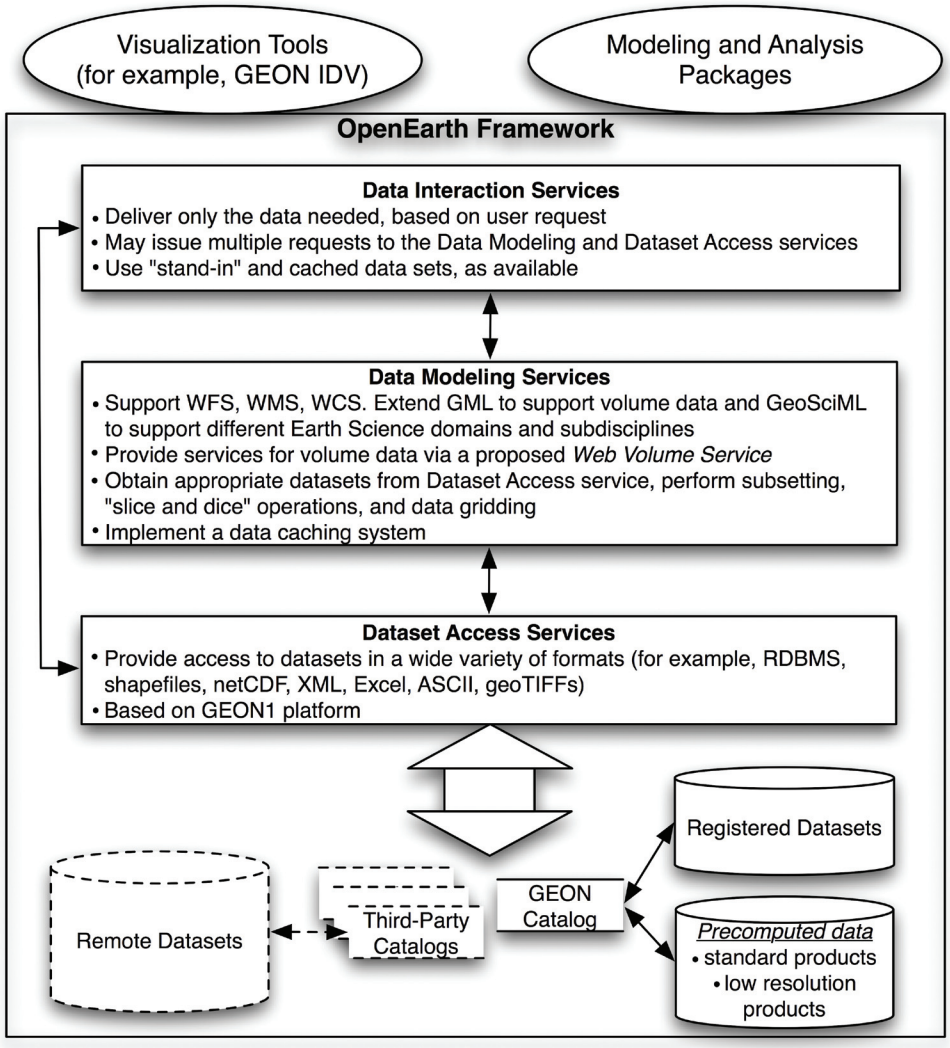


Figure 1. Layered service components of the OpenEarth Framework consisting of a Data Access Service, Data Modeling Service, and Data Interaction Service. The external visualization tools and the modeling and analysis packages can gain access to remote, multidimensional geoscience data using this framework. Abbreviations are as follows: ASCII, American Standard Code for Information Interchange; GEON1, Geosciences Network Project—Phase 1; GEON IDV, the Geosciences Network's Integrated Data Viewer; geoTIFF, Geostationary Earth Orbit Tagged Image File Format; GML, Geography Mark-up Language; GeoSciML, Geoscience Mark-up Language; netCDF, network Common Data Form; WFS, Web Feature Service; WMS, Web Mapping Service, WCS, Web Coverage Service; RDBMS, Relational Database Management System; XML, Extensible Mark-up Language.

data and data derived through format conversion, subsetting, resampling, or other analysis.

At the next higher level, Data Modeling Services provide on-demand and preprocessing operations on archived data. These operations may automatically subset data to extract only the specific data that would satisfy a Web services query. Services may cache the extracted data for future re-use, pre-extract data of expected interest, and otherwise manage the data to enable fast and fluid data delivery. It is at this level that the OEF implements Web services for the delivery of images, features, or volumetric data.

Above this level are Data Interaction Services, which are designed to support rapid visualization of integrated data sets. For instance, services here create multiresolution models that enable visualization tools to zoom smoothly into data by swapping low-resolution data for higher resolution data "on the fly." These services also subdivide data to better support progressive changes to the display as the user pans through large data or reveals additional details of interest. Derived data may be cached, staged at closer network locations, or downloaded in the background to the user's computer. Although three-dimensional rendering and interaction is performed on the user's own computer, these services help reduce delays as the user explores the deeper parts of the data archives.

Finally, the OEF's visualization tools run on the user's computer and use three-dimensional graphics acceleration hardware to display points, lines, polygons, volumetric data, animations, isosurfaces, cutting planes, and so forth. In keeping with the spirit of the inclusive style of the OEF, the open architecture supports multiple visualization tools authored throughout the community. GEON's Integrated Data Viewer, for example, provides an existing mature platform that is being extended to use OEF's layered data services. Additional visualization tools are being developed to drive the creation of higher level OEF data services and to explore new visualization techniques and user interface styles for interacting with integrated datasets.

OEF visualization tools will provide user interfaces that support spatial and temporal queries sent off to one or more data archives. Query results will be presented within integrated views that combine, for instance, surface elevations

derived from LiDAR, surface colors from satellite imagery, borehole paths, cutting planes through seismic tomography data, and other subsurface structures derived from analysis and simulations.

Although the various OEF data services described above will certainly be used, they are optional. Not all data of interest are in a published data archive. OEF visualization tools also must be able to visualize new unpublished data that are locally stored. These local datasets may be loaded along with published data from academic archives, thus enabling comparisons between new and established views of subsurface structure.

Developing Open-Source Software

Admirable visualization tools already exist in specific commercial domains. These tools can produce beautiful imagery that is tuned to the needs of those domains; however, the tools may be less suitable for supporting current research because they cannot provide a flexible test bed for new data models and visualization ideas, nor can they be integrated with academic data archives. Open-source software is needed that can provide the necessary flexibility for academic research. Open source software benefits also include community participation and contribution and the creation of a robust developer and user community. In the end, developing open-source software ensures both the flexibility and longevity of the software base, thereby creating a lasting community asset.

Project Plans

We plan to begin with a sample set of heterogeneous datasets for a given geographic region of interest where a variety of data is currently available. Using these data as a test case, we will develop software to enable visualization of the combined information and the ability to interactively access and manipulate the underlying data.

Integration of Hydrologic Observations from Government and Academic Data Collections with the Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI) Hydrologic Information System

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

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

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

The Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI) Hydrologic Information System (HIS) project is a multi-year, multi-university effort to develop a cyberinfrastructure for advanced hydrologic research and education. The HIS creates a virtual organization and a technical foundation that enables publication, discovery, retrieval, analysis, and integration of hydrologic information across multiple distributed sources to generate a comprehensive picture of hydrologic observations for the entire United States. As a university-based effort, the HIS project develops an infrastructure for publishing academic data collections. At the same time, several recent surveys (for example, Bandaragoda and others, 2005) have shown that hydrologic research often relies on selected Federal data sources: the National Water Information System (NWIS) from the U.S. Geological Survey (USGS), the Storage and Retrieval (STORET) system from the Environmental Protection Agency (EPA), the Automated Surface Observing System (ASOS) from the National Climatic Data Center (NCDC), Snowpack Telemetry (SNOTEL) from the Department of Agriculture (USDA), and others; some State data repositories also house large collections of hydrologic data. Integration of governmental data collections into the HIS has been an important direction of the project.

The main challenges to integrating observational data across agencies and academic projects include (1) heterogeneity across the information systems; (2) a lack of standard and widely adopted information models, data exchange protocols, and agreed-upon semantics for data interchange; and (3) incompatible policies for data serving, data retention, security, funding, and so on. Within the CUAHSI HIS project, these challenges have been addressed by the following activities:

- Developing a common information model for observation data collected at stationary points (measurement stations) that would be uniform across government and academic sources.
- Implementing the common information model as (1) a relational schema, the Observations Data Model (ODM), that supports publication of observational data collections developed as part of academic projects (Horsburgh and others, 2008), (2) a series of databases storing observation data catalogs describing agency repositories, and (3) a standard Extensible Mark-up Language (XML) schema for exchanging water observations, called Water Mark-up Language, or WaterML (Zaslavsky and others, 2007).
- Developing Web services with a common set of method signatures to retrieve WaterML-compliant information about observation stations (GetSites, GetSiteInfo), variables (GetVariables, GetVariableInfo), and values (GetValues). These services, called WaterOneFlow services, are implemented as XML wrappers over Web-based data access systems maintained by Federal and State agencies, such as NWISWeb. In the last year, these services have been

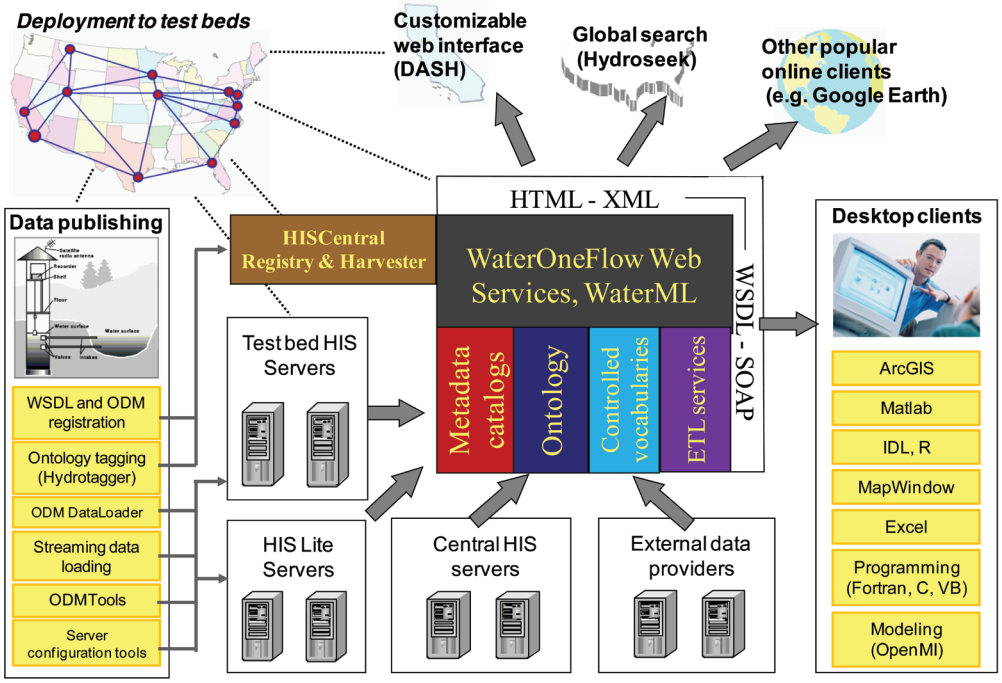


Figure 1. Main components of the CUAHSI HIS service-oriented architecture. Abbreviations are as follows: DASH, Data Access System for Hydrology; ETL, Extract-Transform-Load; HIS, Hydrologic Information System; HTML, Hypertext Mark-up Language; IDL, Interactive Data Language; ODM, Observations Data Model; OpenMI, Open Modeling Interface; SOAP, Simple Object Access Protocol; WaterML, Water Mark-up Language; WSDL, Web Service Description Language; XML, Extensible Mark-up Language.

recoded to take advantage of the Web services developed by our partner agencies. For example, the USGS NWIS team has published a beta version of a Web service that provides programmatic access to NWIS Daily Values data in a WaterML-compliant form. A similar effort has been undertaken at NCDC to publish ASOS data following the WaterML schema, while the EPA has developed the Water Quality Exchange (WQX) framework for sharing water-quality data and submitting them to the STORET data warehouse. Mapping of WQX elements to WaterML is the basis for recoding WaterOneFlow services for the STORET repository.

- Managing varying semantics by mapping water quantity, water quality, and other parameters collected at government agencies to a common vocabulary. This task is essential owing to the size and heterogeneity of available parameter codes (for example, both USGS NWIS and EPA STORET have more than 15,000 listed parameter codes) and the differences in naming conventions adopted at different agencies and research groups. The mapping supports cross-database searches; users navigate a parameter ontology and find variables at each observation network that have been associated with the search term. For example, a search for nitrate measurements in a given area may uncover a range of stations maintained by USGS, EPA, other agencies, and academic projects where nitrate-related variables were measured. The online mapping system for cross-database searching and retrieval is called Hydroseek (Beran and Piasecki, in press).

- Creating an observation-data publication environment where local data managers can load observation data they collected into ODM, validate them, publish them as Web services, configure them to be presented via an online mapping interface (Data Access System for Hydrology, or DASH), associate variable names with common ontology terms, and register the Web services at a central HIS site to make the data available through Hydroseek. The components of the data publication workflow are part of the HIS server, which has been deployed over the last year at hydrologic observatory test beds to support publication of local observation data.

- Developing online user interfaces that combine disparate data into common spatial and temporal representations.

The components mentioned above are organized in a service-oriented architecture (fig. 1). The HIS includes software stacks for HIS Server and HIS Server Lite (the latter is based on free software components only), which are being deployed to the 11 National Science Foundation-supported hydrologic observatory test beds in order to enable uniform publication of local observational data from mostly academic sources. The central HIS site at the San Diego Supercomputer Center (SDSC) serves observation data catalogs that contain sufficient information for formulating data retrieval requests made to agency data repositories.

Until recently, CUAHSI Web services mostly have worked by wrapping respective agency Web sites (NWIS, STORET, and so on) into XML wrappers. This caused a major bottleneck as the services were sensitive to changes in page layout, and the information had to be relayed via SDSC serv-

ers. In addition, the use of Web service wrappers to harvest observation data catalogs from agency Web sites was an error-prone process. As collaboration with these agencies on Web services development intensified in the last year, this situation has changed. The HIS project now receives database snapshots for building observation data catalogs, and connects to newly developed WaterML-compliant or other Web services that are hosted at agency servers, which enables faster data discovery and retrieval. The same model is being extended now to State agencies, as the States of Florida, Texas, and Idaho are implementing their HIS systems.

Acknowledgments

We gratefully acknowledge the support of National Science Foundation award EAR-0622374 (David R. Maidment, principal investigator). We gratefully acknowledge cooperation, insightful discussions, and help provided by partner agency personnel from USGS (R. Hirsch, K. Lins, D. Briar, D. Coyle, M. Hamill, and other members of the Water Resources Discipline), EPA (C. Spooner, M. Hamilton, D. Young and the STORET team), and NCDC (R. Baldwin).

References Cited

Bandaragoda, C.J., Tarboton, D.G., and Maidment, D.R., 2005, User needs assessment, chapter 4 *in* Maidment, D.R., ed., Hydrologic Information System status report, version 1—September 15, 2005: Washington, D.C., Consortium of Universities for the Advancement of Hydrologic Sciences, Inc., p. 48–87, available only online at http://www.cuahsi.org/docs/HISStatusSept15.pdf/. (Accessed August 19, 2008.)

Beran, Bora, and Piasecki, Michael, in press, Engineering new paths to water data: Computers and Geosciences.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., and Zaslavsky, Ilya, 2008, A relational model for environmental and water resources data: Water Resources Research, v. 44, no. 5, citation number W05406, available only online at http://www.agu.org/pubs/crossref/2008/2007WR006392.shtml/. (Subscription may be required.) (doi:10.1029/2007WR006392)

Zaslavsky, Ilya, Valentine, David, and Whiteaker, Tim, eds., 2007, CUAHSI WaterML: Wayland, Mass., Open Geospatial Consortium, Inc., document OGC 07–041, available only online at http://www.opengeospatial.org/standards/dp/. (Accessed August 19, 2008.)

Metadata and Semantics in the Astronomical Virtual Observatory

By Norman Gray[1]

[1]Department of Physics and Astronomy, University of Leicester, Leicester, United Kingdom.

The astronomical virtual observatory (VO) shares many of the goals of the geophysics and space-science VO, but has a different history and faces different challenges. I will review the astronomical VO's present situation and its current solutions, discuss the roadmap for this VO's "standards body" (the International Virtual Observatory Alliance, or IVOA), and pay particular attention to the semantic technologies which are being used or anticipated in this domain.

The Astronomical VO—Contexts and Goals

The astronomical VO—which we may take to cover night-time astronomy, plus radio, X-ray and solar astronomy (plus a little solar terrestrial physics)—is characterized by a large number of independent data and image archives, which contain collections ranging in size from megabytes to (soon) petabytes and with a mixture of curation styles ranging from semiformal to highly professional. Despite this heterogeneity, these archives have significant overlaps in terms of file formats, coordinate systems, and objects of interest; for instance, both X-ray and radio astronomers will be interested in a supernova remnant, both will refer to it with a right ascension and declination, and both will be able to produce a flexible image transport system (FITS) file containing relevant data. The range of resources available is large enough that scientists may not be aware of all the resources that might be of use to them, or they may not know how to use interesting resources in different wavelength ranges.

This shared technology and interest means that the astronomical VO should be in a prime position to take advantage of processes already developed by VOs in other scientific disciplines and to make significant progress towards domain-wide interoperability. The range of archive sizes means that this VO has notable data discovery problems and significant social problems in bringing a wide range of actors to common agreement.

IVOA Responses

The International Virtual Observatory Alliance (IVOA) is a consortium of consortia and acts as a coordinator for multiple national astronomy VO projects. Its processes are explicitly modeled on those of the World Wide Web Consortium (W3C).

The IVOA has been successful, both in brokering high- and low-level agreements on protocols and formats and (a softer, but equally challenging process) in establishing itself as the single forum which coordinates the astronomical VO, thus acting as a "nursery" for other spin-off VO developments. I will review the history of this process.

Semantic Technologies Within the IVOA

The IVOA's principal accomplishments in coordinating metadata output have been the establishment of a VO-wide resource registry, the development of a small set of common and extensible data models, and support for a few more sophisticated semantic experiments. The registry consists of references to image servers, catalog servers, and Web services, along with their associated metadata. The data models that IVOA has agreed to so far cover coordinate systems and catalog coverage information. Both models were substantially harder to agree upon than was initially expected, for interesting and informative reasons.

The semantic technologies explored to date include a basic system for describing astronomical data (Unified Content Descriptors, or UCDs), an ontology of astronomical object types, the early stages of a system for linking serialized data to data models, and the development of interoperable controlled vocabularies.

Other Projects

Several other semantics-oriented projects are being developed using IVOA technologies:

- The Explicator project's goal is to avoid the expense and complication of creating consensus data models by helping data centers first make their data available in a data model that is natural to them (and thus inexpensive to define and maintain) and then formally declare mappings to well-known data models.
- The Semantic Knowledge Underpinning Astronomy (SKUA) project is creating a prototype of a distributed network of semantically aware, shared, annotated services in the form of resource description framework (RDF) databases. This semantic layer will support a cluster of applications, which will either directly support users in finding and recovering useful resources, or indirectly support them by supporting user-facing applications, including a Facebook-like astronomical virtual research environment (VRE).

Comparison of Different Land-Use Object Classes by Means of Semantic Similarity Measurements

By Chris Schubert,[1] Ilonka Wolpert,[1] and Ingrid Christ[1]

[1]Delphi IMM, GmbH, Potsdam, Germany.

Introduction

Land-use and land-cover data are often required for public tasks on a regional, a national, and even on a European level. Currently, some datasets using different thematic and geometric accuracy are available for these tasks, for example, the Coordination of Information on the Environment (CORINE) Land Cover (CLC) project, the Authoritative Topographic-Cartographic Information System (ATKIS), or the recently established Global Monitoring for Environment and Security (GMES) services.

The CLC and GMES classification systems contain substantial, object-oriented descriptions of land-use and land-cover classes, which are produced via satellite in order to provide environmental monitoring on a European level. In contrast, ATKIS is an information system whose goal is to provide a German topographic land survey with a high geometric accuracy and progressive actualization degree for urban areas; however, for many specialized applications in environmental studies, this nationwide dataset does not provide the required information accuracy. Moreover, for mapping and updating services, it is necessary to combine these heterogeneous and already existing datasets. This action is technically feasible, but an important semantic problem remains. To tackle this problem, two steps must be taken: (1) formalize the description of object classes, and (2) set up a similarity measurement by using a knowledge-based model.

These approaches are developed by Delphi Information Model Management GmbH (Delphi IMM) in order to contribute to the German research project DeCOVER (Deutschland (Germany) Land Cover), which is a joint project of 11 partners, funded by the Federal Ministry of Economics and Technology and managed by the German Aerospace Center. DeCOVER was initiated to conceptualize and demonstrate innovative and cost-efficient geoinformation services based on semantic interoperability and remote-sensing data with the goal of updating the existing land-cover datasets in Germany.

Knowledge Representation

Datasets consisting of different objects, which are based on heterogeneous classification systems, can be exchanged and integrated only if they are comparable. Figure 1 depicts sport and leisure objects classified in ATKIS and CLC. These objects are nearly congruent, although the terms of the object classes are different; for instance, compare "ATKIS-2201

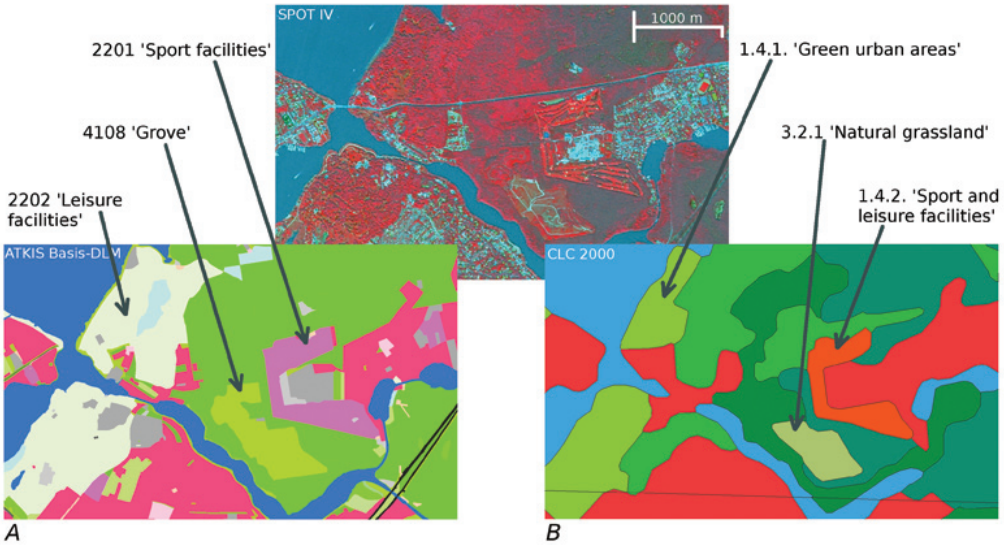


Figure 1. Land-cover maps derived from Spot Image satellite imagery. *A*, Land-cover map using the ATKIS classification system. *B*, Land-cover map using the CLC 2000 system. The minimum mapping unit for ATKIS is 1 hectare and for CLC is 25 hectares. A comparison of *A* and *B* shows different feature demarcations, which are based on different feature contents.

Sports facilities" with "CLC-1.4.2 Sport and leisure facilities." Instead, other objects in the catalog are clearly deviating from each other; for instance, compare "ATKIS-4108 Grove" with "CLC-3.2.1 Natural grassland."

In fact, these differences are caused by distinct conceptualizations and object descriptions. Consequently, to make object classes comparable, they have to be set at the same definition level by means of ontologies. An ontology is an explicit description of a common "world outlook" (Gruber, 2003). In order to make the knowledge of classification systems explicit, Delphi IMM created a multilevel process and applied it to the ATKIS and CLC catalogs as follows:

- Extracting the required information from the existing mapping instructions of the catalogs,
- Creating a basic knowledge model, and
- Defining all object classes as an application ontology.

The basic model deals with concepts, object properties, and relations for a general object class. For each object class of a classification system, there has to be a more specific application ontology based on taxonomy and relation of the basic knowledge model (Lutz and others, in press). Concept expressions and property restrictions characterize application ontologies. The main components—"land cover" (for example, vegetation, water, or urban area) and "land use"—are universal properties. Other parameters are the "location" with respect to the sea, the "characteristic neighborhood," or the "genesis" (natural or manmade). In addition, each object class can be expressed by specific properties depending on land cover. A vegetated object, for instance, can be specified by information about soil moisture, which actually implies that there might be a swamp present.

The following process of reasoning is an automated process, which is lead by logical conclusions and classifying to an automatic subsumption—a depiction of the hierarchy of object classes. The ontology-based reasoning serves as a validation of formalized application ontologies within one domain or catalog; for example, the object class "mixed forest" consequently is a subclass of "forest." The opposite of this very simple and easily realized automated reasoning (comparing the object classes from different catalogs) is significantly more difficult. Here, only equal or more comprehensive properties have to be taken into account; however, the similarity between different object classes must also be considered. For this reason, a similarity measurement was developed.

Similarity Measurement

The similarity of two object classes can be qualified in either a network model or a feature model. The network model describes the distance between nodes in a hierarchical tree structure where special attention is given to the number of edges between two nodes within a multidimensional area (Rada and others, 1989).The feature model measures the similarity by comparing the common properties of two object classes (Tversky, 1977). We used a combination of both models: the feature model permits a rough estimation of similarity regarding the number of common properties and the network model outlines the refinement. The distance between the two similar properties is based on the knowledge model.

We distinguish between substantial similarity and mapping similarity and therefore developed two different algorithms. Both are based on a combined feature-network model. Substantial similarity considers two object classes that need to be compared as symmetric objects; therefore, a one-time evaluation has to be executed for each pair of properties, regardless of which of the two object classes presents the origin or the destination (table 1). Finally, the calculated similarity shows the proximity of two object classes, but not the opportunity to correlate one object with another object. The intent of the mapping similarity is to show the potential transfer from an origin object class to a destination object

Table 1. Substantial similarity of "green spaces" object classes from the CLC and ATKIS systems.

Object class	ATKIS_2201_Sport facilities (percent)	ATKIS_2202_Leisure facilities (percent)	ATKIS_4108_Grove (percent)
CLC_141_Green_urban_areas (percent)	88	92	63
CLC_142_Sport_and_leisure_facilities (percent)	92	92	63
CLC_321_Natural_grassland (percent)	83	83	70

Table 2. Mapping similarity of "green spaces" object classes from the CLC system to the ATKIS system.

Destination class / Origin class	ATKIS_2201_Sport facilities (percent)	ATKIS_2202_Leisure facilities (percent)	ATKIS_4108_Grove (percent)
CLC_141_Green_urban_areas (percent)	65	67	31
CLC_142_Sport_and_leisure_facilities (percent)	76	76	36
CLC_321_Natural_grassland (percent)	57	57	28

Table 3. Mapping similarity of "green spaces" object classes from the ATKIS system to CLC system.

Destination class / Origin class	CLC_141_Green_urban_ areas (percent)	CLC_142_Sport_and_leisure_ facilities (percent)	CLC_321_Natural grassland (percent)
ATKIS_2201_Sport facilities (percent)	79	86	20
ATKIS _2202_Leisure facilities (percent)	86	86	20
ATKIS_4108_Grove (percent)	21	50	28

class (tables 2 and 3). This asymmetric examination only considers properties from the source class and gives the opportunity to map a source object class into a destination object class.

The following example shows how substantial and similarity mapping works, using the information found in tables 1 through 3. The substantial similarity of object class "CLC 141 Green urban areas" and "ATKIS 2201 Sport facilities" measures 88 percent (table 1). This high value indicates that the two classes are very similar, but this method does not show whether the class "CLC 141 Green urban areas" could be mapped into "ATKIS 2201 Sport facilities." To determine whether it could, we use the method of mapping similarity. Table 2 shows an opportunity of 65 percent to map "CLC 141 Green urban areas" into "ATKIS 2201 Sport facilities" and obversely the validation is 79 percent. In contrast to substantial similarity, the values of mapping similarity are lower. Actually, these differing similarity values can be explained by means of the properties. Figure 2 shows a section of a knowledge-based model which is used for all object classes. "ATKIS 2201 Sport facilities" has the property called "Sport" and "CLC 141 Green urban areas" has the property called "Recreation" or "Culture" for the land-use parameter in this model, connected respectively with one edge.

Conclusions

The results are run in a Web application where heterogeneous catalogs of land use and land cover are automatically compared. The framework, Jena (developed by SourceForge, Inc.; see http://sourceforge.net for more information), supports the extraction of several ontology concepts.

Delphi IMM has achieved a semantic comparison of different land-use classes at the catalog level by means of ontologies and a similarity measurement algorithm. We developed our basic knowledge model as well as the similarity measurement algorithm in an open-ended way so that any other land-use catalog can be added and compared. Future work will focus on a feature-based search and especially on the mapping at an object level, which eventually will ensure the opportunity to map spatial data classified in one catalog to spatial data of another catalog without needing to consider geometry and topology.

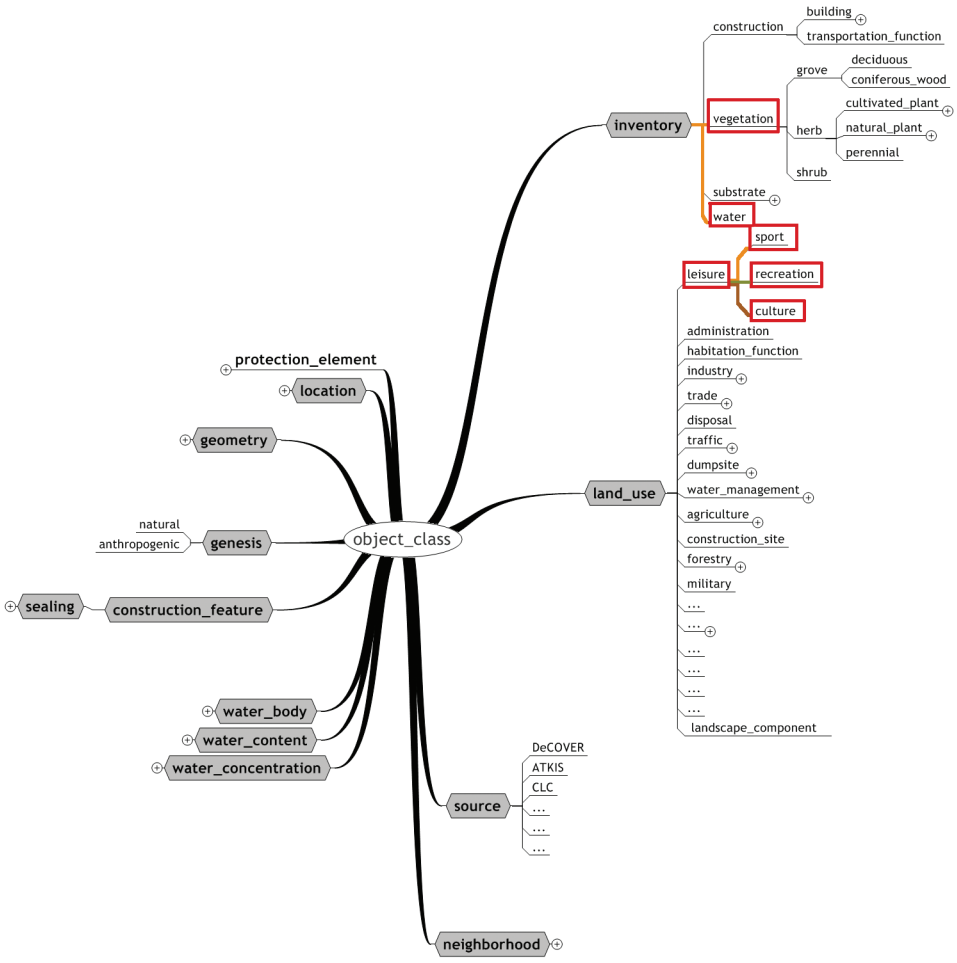


Figure 2. Diagram showing the section of the knowledge-based-model that demonstrates the distance between a property pair.

References Cited

Gruber, Tom, 2003, It is what it does—The pragmatics of ontology for knowledge sharing, *in* Proceedings, Sharing the Knowledge, International CIDOC CRM Symposium, Washington, D.C. March 26–27, 2003: International Committee for Museum Documentation Conceptual Reference Model Web page at http://cidoc.ics.forth.gr/docs/symposium_presentations/gruber_cidoc-ontology-2003.pdf. (Accessed September 9, 2008.)

Lutz, M., Christ, I., Schubert, C., Klien, E. and Hübner, S., in press, Overcoming semantic heterogeneity in spatial data infrastructures: Computers and Geoscience.

Rada, R., Mili, H., Bicknell, E., and Blettner, M., 1989, Development and application of a metric on semantic nets: IEEE Transactions on Systems, Man and Cybernetics, v. 19, no. 1, p. 17–30.

Tversky, A., 1977, Features of similarity: Psychological Review, v. 84, no. 4, p. 327–352.

Semantic Web Technologies for Value-Added Services at the German Research Center for Geosciences' Information System and Data Center

By Bernd Ritschel,[1] Sabine Pfeiffer,[1] Vivien Mende,[1] and Sebastian Freiberg[1]

[1]Information System and Data Center, German Research Center for Geosciences, Potsdam, Germany

The German Research Center for Geosciences' Information System and Data Center (ISDC) portal provides retrievable earth observation data, information, and knowledge about the geosciences using a metadata-based catalog system. Although searchable metadata related to a data product are stored in tables, the metadata that are dependent on the product type are represented using the National Aeronautics and Space Administration's (NASA's) Global Change Master Directory (GCMD) Directory Interchange Format (DIF) documents, which are written in Extensible Mark-up Language

(XML). General and project-specific information on ISDC's portal (including information about its architecture, operation, and the philosophy behind the use of the metadata) is found in a companion paper (Ritschel and others, 2008, this volume).

Semantic Web technology (Matthews, 2005) is used in order to provide new and extended access to data, information, and knowledge at the ISDC; it also makes references and correlations between different classes of metadata (documents) visible. In addition to these functions, interoperable ISDC portal Web services, such as the Catalogue Service for Web (CSW), the Web Map Service (WMS), or Web-based discovery and registry services based on semantic relations, can be realized using standardized metadata documents, structures, concepts, and languages, such as Open Geospatial Consortium's metadata based on ISO 19115 (International Organization for Standardization, 2003), XML, Resource Description Framework (RDF), Simple Knowledge Organization System (SKOS), or Web Ontology Language (OWL). Multi- and interdomain collaboration services are possible only with the addition of validated semantics in controlled vocabularies, such as NASA's universal GCMD science keywords and associated directory keywords, or the mostly marine-science vocabulary developed by the Natural Environment Research Council's Data Grid (NDG). In addition to organization- or committee-driven controlled vocabularies, generally accepted free vocabularies or "folksonomies" (community-driven classifications that have developed as the result of the transition to a Web-based culture) are becoming more and more important. (For more information on controlled vocabularies, consult David Rieckes' Web site, http://www.controlledvocabulary.com. For an introduction to the Semantic Web, see Palmer, 2001.)

The Semantic Web is built using a layered architecture. The lowest layers are based on Universal Resource Identifiers (URIs) and XML. URIs are used in order to identify the semantic scope of the content that is to be modeled. The lowest layers support the RDF and related RDF schemas. They provide the techniques for the representation of semantics within structured information sources, which consist of relations between subjects, predicates, and objects.

OWL and SKOS (Isaac and Summers, 2008) are appropriate and standardized languages for the representation and processing of knowledge. Whereas SKOS is used for the modeling of controlled vocabularies, different OWL specifications, such as OWL Lite or OWL Full, may be used to describe complex interdomain semantic relations. Protègè, Altova SemanticWorks, Semantic Web Ontology Overview and Perusal (SWOOP), and CmapTools are some of the tools for the creation, management, and processing of ontologies. CmapTools has been used for the modeling the new ISDC DIF-standard-compliant metadata concept.

As mentioned already, the ISDC product-type-dependent metadata are stored in DIF-standard-compliant XML documents. One product type is described by one metadata document from the metadata class product-type DIF. Inside this class, there are eight mandatory attributes (such as entry identification, entry title, and parameters) and more than 25 optional attributes (such as a summary or a reference). The analysis of the content and the structure of the GCMD product-type DIF metadata class shows some attributes (such as project, platform, instrument, and institution) that also can be used to extend the simple concept of autonomous metadata classes. This means that, in addition to the product-type class, new metadata classes and relations can be created. For the description of the unique and discrete data products (data files), an extension of the DIF standard is used, which is described in detail in Mende and others (2008, this volume).

Figure 1 shows the proposed structure of the new ISDC DIF-standard-related metadata concept. The concept consists of concept nodes, which in our case are metadata classes or attributes of classes, and the relations between two concept nodes. These relations are visualized by concept arrows and the corresponding linking phrases in figure 1. Two concept nodes and the appropriate relation are always reflecting one of the different ISDC metadata schemes. Although the different shapes and colors of the concept nodes represent specific features of our concept nodes, not all ISDC metadata classes, attributes, and relations are represented in the concept diagram (fig. 1); for example, there is no way to distinguish between mandatory and optional classes and attributes.

The main ISDC metadata product-type class is located at the center of the concept diagram (fig. 1). Important features of the product type are referenced by attributes such as "parameter" (science keyword), "reference," "citation," "free_keyword," "entry_title," "summary," and "entry_id." A relation, consisting of the two concept nodes "product_type" and "entry_id" and the associated linking phrase ("has unique"), can be read in the following way: "A product type has a unique identifier." Another relation is the following: "A product type has a citation," which reflects the relation between a producer of data products and the type of data products. The attribute "reference" is used for the relation between product type and the usage of related data products within scientific publications. Other major attributes are "project," "platform" (observatory or instrument carrier), and "instrument" (sensor), which are represented by autonomous classes, too. In order to keep the concept clear and simple, attributes such as "data_center," "personnel," "quality," and others are not shown here but are part of the metadata concept.

In addition to the ISDC metadata classes and their relations, the concept also represents concept nodes for controlled and free vocabularies and the different sources of those vocabularies. The content of the product-type attribute "parameter" as well as the content of the main attributes and classes (such as "project," "platform," and "instrument") are represented by controlled vocabularies. The ISDC metadata concept uses the GCMD's science keywords and associated directory keywords as controlled vocabularies. Sources of free vocabularies are generated by data providers or users; these vocabularies can be used in addition to controlled vocabularies through the attribute "free_keyword." In addition to the input of free keywords, a user-driven extension of the existing classification of keywords is possible. An example of a new

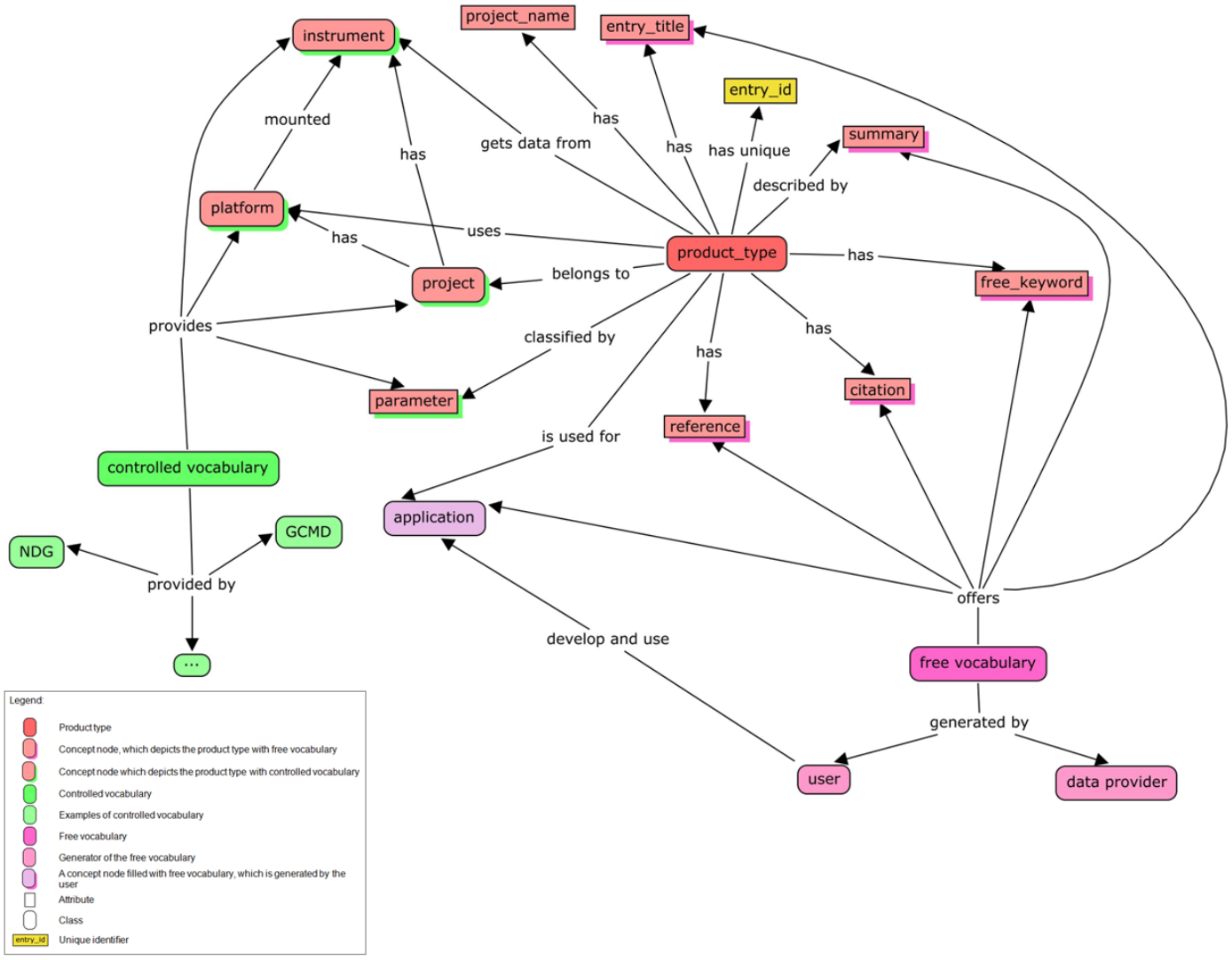


Figure 1. Diagram showing the relations between classes and attributes in the Information System and Data Center (ISDC) Directory Interchange Format (DIF) metadata concept.

non-DIF-standard-compliant keyword type is "application," which describes the usage of data products in a specific application. Generally accepted free vocabularies can be supported via Web 2.0 technologies such as collaborative or social tagging. Other semantic information often is hidden in data about social navigation (the analysis of user activity). A proposed activity can be prompted, as in the following example: If user A always downloads product types X and Y together, then the proposal for user B, who only downloads product type Y, could be "Why not download product type X, too?

The new ISDC DIF metadata concept also can be used in order to create new and more abstract product-type classes, which, for example, have a different scope. As an example, ISDC product types related to the orbit of objects in space (such as Rapid Science Orbit, Predicted Orbit, or Precise Orbit, which are derived from different satellite missions or projects) can be described in a general way by the new product type "Orbit." The name of the new product type not only allows users to network together the different unique product types, but also to create a more generalized, searchable index for orbit products within the ISDC portal in the future.

References Cited

International Organization for Standardization, 2003, Geographic information—Metadata: Geneva, Switzerland, International Organization for Standardization, 140 p.

Isaac, Aaron, and Summers, Ed, eds., 2008, SKOS Simple Knowledge Organization System primer—W3C working draft 21 February 2008: World Wide Web Consortium Web site at http://www.w3.org/TR/skos-primer/. (Accessed August 20, 2008.)

Matthews, Brian, 2005, Sematic Web technologies: Joint Information Systems Committee Techwatch Report TSW 05-02, 20 p., available only online at http://www.jisc.ac.uk/media/documents/techwatch/jisctsw_05_02bpdf.pdf/. (Accessed August 20, 2008.)

Mende, Vivien, Ritschel, Bernd, Freiberg, Sebastian, Palm, Hartmut, and Gericke, Lutz, 2008, Directory Interchange Format (DIF) metadata and handling at the German Research Center for Geosciences' Information System and Data Center, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2008—Data to Knowledge, Potsdam, Germany, June 11–13, 2008: U.S. Geological Survey Scientific Investigations Report 2008–5172, p. 43–46 (this volume).

Palmer, Sean, 2001, The Semantic Web—An introduction: Web site at http://infomesh.net/2001/swintro/. (Accessed August 20, 2008.)

Ritschel, Bernd, Mende, Vivien, Palm, Hartmut, Gericke, Lutz, Freiberg, Sebastian, Kopischke, Ronny, and Bruhns, Christian, 2008, The German Research Center for Geosciences' Information System and Data Center—Portal to geoscientific data, information, and knowledge, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2008—Data to Knowledge, Potsdam, Germany, June 11–13, 2008: U.S. Geological Survey Scientific Investigations Report 2008–5172, p. 33–35 (this volume).

Sensor-Based Landslide Early Warning System (SLEWS)—Development of a Spatial Data Infrastructure With Integrated Real-Time Sensor Data as a Basis for Early Warning Systems Exemplifying Landslides

By Stefanie Hass,[1] Kai Walter,[2] Frank Niemeyer,[2] Christian Arnhardt,[3] Kristine Asch,[1] Raffig Azzam,[3] Ralf Bill,[2] Tomas Fernandez-Steeger,[3] Stefan Daniel Homfeld,[4] and Hartmut Ritter[4]

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

[2]Department of Geodesy and Geoinformatics, Rostock University, Rostock, Germany.

[3]Department of Engineering Geology and Hydrogeology, RWTH Aachen University, Aachen, Germany.

[4]ScatterWeb GmbH, Berlin, Germany.

Accessible geographic information becomes more and more important for making decisions. Spatial data infrastructures (SDIs) allow the distribution and access of sensor and geographic data by using Web services developed by the Open Geospatial Consortium (OGC). The Sensor-based Landslide Early Warning System (SLEWS) is a joint project of the RWTH Aachen University, University of Rostock, the Federal Institute for Geosciences and Natural Resources (Hannover), and ScatterWeb GmbH (Berlin) that uses open standardized Web service specifications and data formats provided by OGC's Sensor Web Enablement (SWE) initiative to establish a monitoring and warning system for landslides. SLEWS is built around a self-organizing wireless sensor network, hosting sensors such as inclinometers, accelerometers, and displacement and pressure detectors to obtain real-time data (fig. 1).

The Sensor Web Enablement initiative offers methods to discover, access, and filter sensor data, which allows for effective data management and analysis. One major requirement of early warning and risk management systems is the ability to immediately extract reliable information from the collected data. This process takes place in the information tier (fig. 1) by data modeling, which involves complex series of algorithms and analytic processes. The sensor data are collected and standardized within the data tier and then forwarded to the information tier where the data are analyzed, modeled, and specifically processed for various end-user applications.

When a critical event occurs, OGC services such as the Sensor Alert Service (SAS) and Web Notification Service (WNS) automatically send a warning message. The Sensor Mark-up Language, SensorML, is used to formalize and standardize the sensor processes and hardware. SWE services can be combined with already existing OGC Web services such as the Web Feature Service (WFS), Web Coverage Service (WCS), and Web Map Service (WMS). When long-term planning is needed, digital maps are automatically created that provide support for decisions and resource deployment planning to respond to a critical event. Moreover, within an SDI, the functions of classical geoinformation systems can be used by the Web Processing Service (WPS). These functions allow the intersection of themes, such as detailed and visualized information about roads, buildings, or other crucial infrastructures inside the hazard zone. The geographic and sensor information is made available by means of an internet browser or other mobile applications such as a personal digital assistant (PDA) or a cell phone with advanced features.

The development of an SDI that uses OGC Web Services is an innovative technology that meets the requirements of early warning and risk management systems. The SDI is able to encapsulate heterogeneous data and to transform the measured data into relevant information. OGC Web Services (OWS) and Sensor Web Enablement open up new possibilities for real-time monitoring and provide information for decision-making by enabling the user to collect data stored in different locations. This newly developed, advanced technology provides local authorities with essential hazard information that may protect life and help mitigate potential damage caused by landslides and rockfalls.

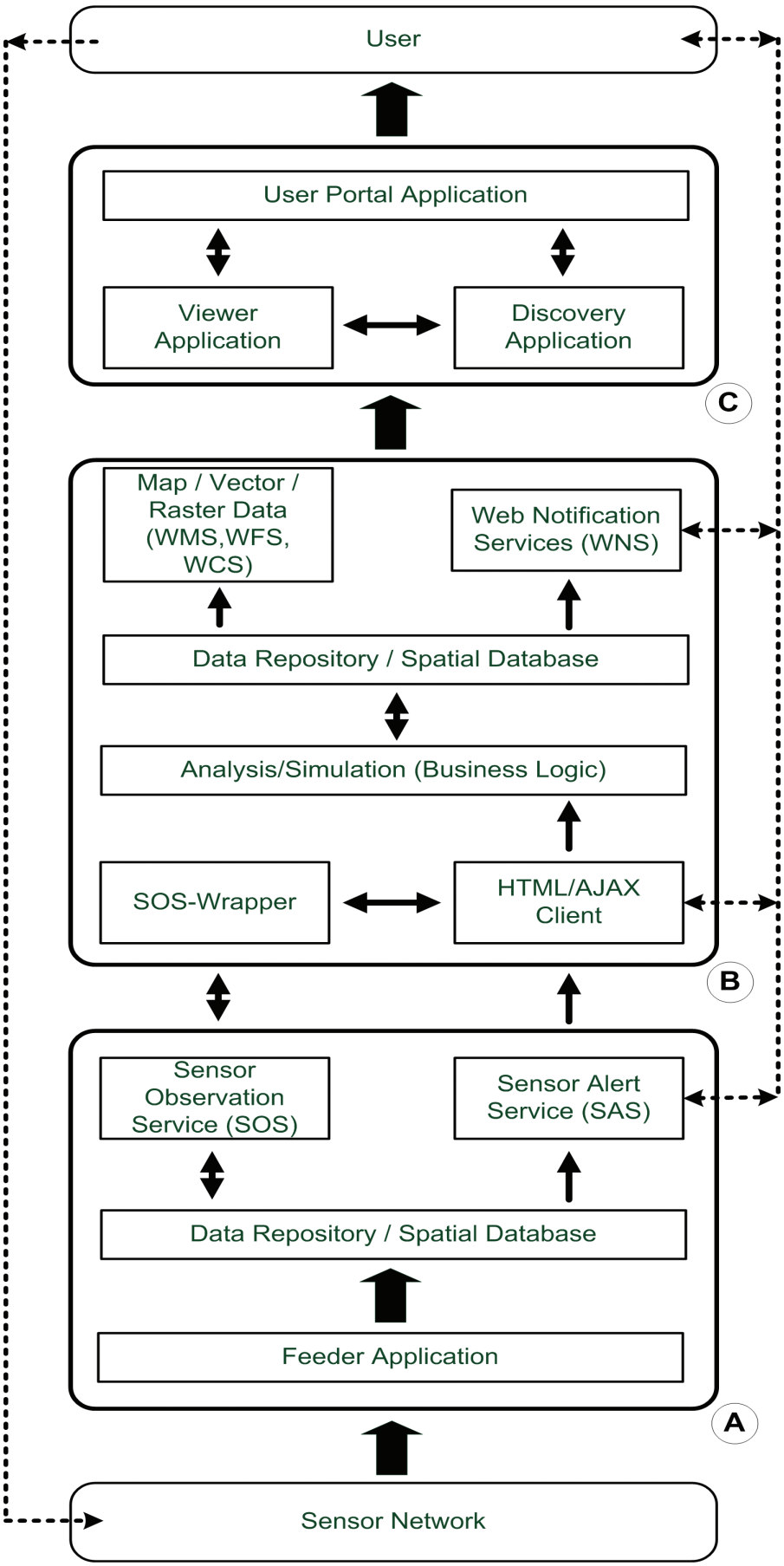


Figure 1. Diagram showing the planned system infrastructure for the Sensor-based Landslide Early Warning System (SLEWS), from data measurement to information distribution. Section A shows the sensor network and data management; section B shows data analysis, information management, and distribution; and section C shows user management, visualization, and discovery. Abbreviations are as follows: AJAX, asynchronous JavaScript and XML (Extensible Mark-up Language); HTTP, Hypertext Mark-up Language; WCS, Web Coverage Service; WFS, Web Feature Service; WMS, Web Map Service.

Ontological Geosciences

By Kangping Sun[1]

[1]Reserch Center for Space Science and Technology, China University of Geosciences, Wuhan, China.

The study of geosciences may be viewed as two separate activities: (1) gaining knowledge about the Earth through the traditional geoscience disciplines of geophysics, geochemistry and so forth and (2) determining how to seek, obtain, and infer new geoscience knowledge. Since the end of 2007, the author has led a team of graduates and senior undergraduates at China University of Geosciences (CUG) to experiment with the use of geoscience concept models to represent theme-oriented concepts in geoscience literature. One of the objectives of this project is to integrate geoscience data using a subject-oriented concept model. Another objective is to explicitly characterize, within the concept model, the knowledge that is related to or implied in the theme-oriented concept but is extracted from published literature (other than the literature that was used in order to develop the concept model) in order to gain further insight of the mechanisms a concept model should possess. This project is still ongoing; however, there are two conclusions we have been able to draw from the progress made so far:

1. The knowledge related to and implied in any single geoscience term, conclusion, or theory is vast, as it relates in some way to almost every other concept in geosciences; and

2. Geosciences contain various recognition patterns, including those that can be used to seek or infer new geoscientific knowledge.

Conventional automated computer systems may not be able to correctly comprehend and operate a geoscience concept model that is capable of recognizing new geoscientific knowledge patterns. Furthermore, it does not seem feasible that one can just use the limited number of geoscience concept patterns already known to us to consistently provide adequate representations of such complex and ever-evolving knowledge. For geoscientists, the knowledge represented by the new concept models derived only from the limited patterns already known to us may be especially hard to accept. For these reasons, we propose to develop a subdiscipline of the geosciences called "ontological geosciences."

We refer to ontological geosciences as that which is based on geoscience data and encompasses all the contents of the traditional geosciences. The essential characteristic of the ontological geosciences is that it is composed of formal discrete geoscience elements. Hence, the ontological geosciences possess better correlation, integrity, and consistency than the traditional geosciences. Although they are similar to the traditional geosciences, the ontological geosciences must first be recognizable and operable by people who want to use it, especially the geoscientists; second, the content that makes up the basis for the ontological geosciences must also be programmable (that is, recognizable and operable insofar as computers are concerned, which is unique for the data-based geosciences). Ontological geosciences, however, should be neither considered nor developed as an ordinary computer system; instead, they ultimately should be developed as a programmable knowledge system.

The traditional geosciences are mainly concerned with the knowledge of the structure, composition, and evolution of the Earth. Ontological geosciences focus on the way the structure, composition, and evolution of the Earth are represented on computers, thus providing an underlying structure for the internally integrated representation of geoscience knowledge. The main challenges to the ontological geosciences are as follows:

1. Extraction and standardization of geoscience knowledge patterns from the traditional geoscientific literature and the use of those derived knowledge patterns to represent that part of geoscience knowledge that has not been conclusively proven or generally recognized.

2. A determination of the extent to which one can surmise, generalize, and unify these knowledge patterns.

3. The knowledge about the Earth may not be mature enough to be integrated into a consistent model.

4. The structure, composition, and evolution of the Earth are beyond the control of human beings. Our knowledge of the Earth, therefore, has not been and will never be complete; all the geoscience knowledge we have acquired through the gradual process of studying the Earth is, in theory, incomplete. When scientists observe new natural phenomena, they often provide explanations by proposing theories or hypotheses, which are often personal and subjective. The verification process of these theories and hypotheses may take a long time because the Earth is an extraordinarily complex entity and scientific observational tools and studies are very expensive. On the other hand, in order to get a consistent interface to access geoscience data, ontologies or "standard dictionaries" must be defined. Consequently, the geoscience principles should be included because the meaning of geoscientific terms are often significant within a given context, or theory. In short, if we assume that we know the geoscience structures, then we can use them to "teach" computers to understand geoscientific concepts. We hope to use general or commonly recognized knowledge patterns in order to develop just one computer system to support the whole of the geosciences.

Our solution to the aforementioned challenges is to design a hierarchical ontological geosciences computer system that is developed by users through collaboration with others. The architecture of the proposed system is illustrated in figure 1. In order to construct such a system, two concepts must be accepted. First, the ontological geosciences consist of

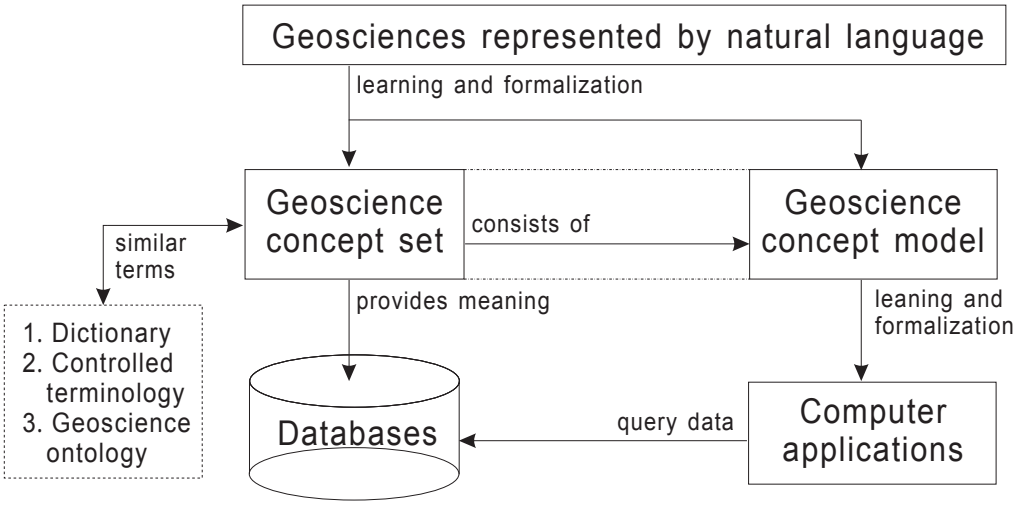


Figure 1. Diagram showing the architecture of the ontological geosciences.

a system of hierarchies of geoscience knowledge. The hierarchy of the architecture is not limited within the geoscience concept model; it extends to the geoscience knowledge levels, including the levels of the knowledge base that can be used to design the model or applications. Second, it is a system that is self-improving through the collaborative efforts between system users and computers. The computer network facilities provide a convenient mechanism for communication and sharing knowledge among the system users. The users gain knowledge from the system and improve it according to their knowledge of geosciences. Some of the most important learning and improving processes take place between the hierarchical levels. The users learn the knowledge patterns from the processes occurring in the upper layer and transform them into the lower-level model or applications. Last, the middle layer of the architecture can be regarded as "dictionary-based geosciences," which is that aspect of the geosciences that is only represented by the terminologies provided in the dictionary. The dictionary provides the meaning of terms for geoscience data in the databases as well; therefore, the ontological geosciences can be built on the basis of geoscience data via the dictionary.

In summary, ontological geosciences is a structured geoscience subdiscipline that is based on geoscience data and is recognizable and operable on computer systems. In the current environment of research and development, we regard it as being essentially a "human-being-centered" geoscience. It is the people who learn and generalize the knowledge patterns, who determine what geoscience content is appropriate for computers to represent, and who teach the computers to understand the ontological geosciences that in turn support geoscientific research.

Acknowledgments

The author would like to thank China Geological Survey for their support and financial contribution to this research project.

A Volcano Erupts—Semantic Data Registration and Integration

By Peter Fox,[1] A. Krishna Sinha,[2] Deborah L. McGuinness,[3,4] Rob Raskin,[5] and Abdelmounaam Rezgui[2]

[1]High Altitude Observatory, Earth and Sun Systems Laboratory, National Center for Atmospheric Research, Boulder, Colo.

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

[3]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, N.Y.

[4]McGuinness Associates, Latham, N.Y.

[5]Jet Propulsion Laboratory, Pasadena, Calif.

We present a progress report in a research effort (Semantically-Enabled Scientific Data Integration, or SESDI) into the application of Semantic Web methods and technologies to the challenging problem of integrating heterogeneous volcanic and atmospheric chemical-compound data, which are used to assess the atmospheric effects of a volcanic eruption. One requirement to accomplish this is the semantic registration of datasets to domain and integrative ontologies. We demonstrate how ontologies are implemented by leveraging existing distributed semantic technology frameworks.

Introduction

The goal of our project is to enable the next generation of interdisciplinary and discipline-specific data and information systems to answer many challenging science questions requiring data from widely distinct fields. Our initial focus was on the integration of volcanic and atmospheric data sources in

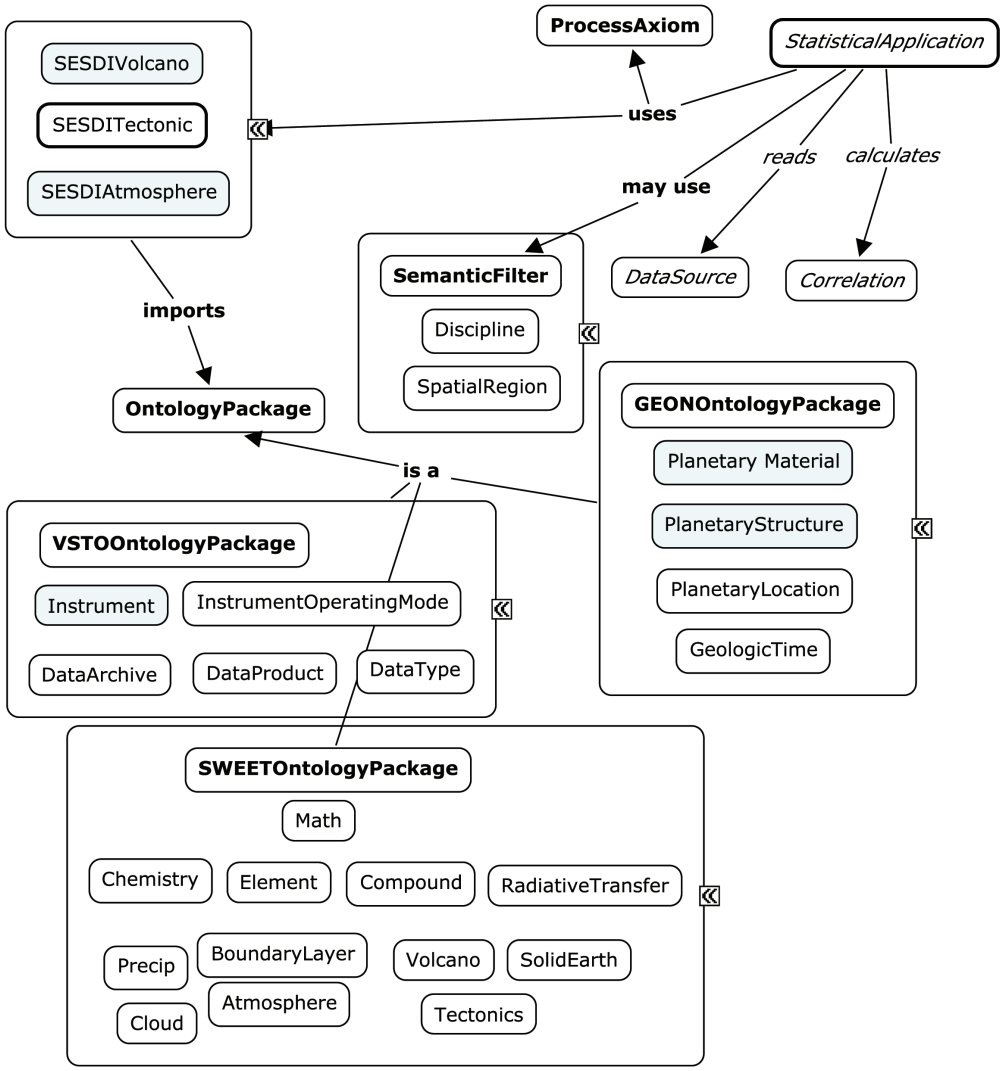


Figure 1. Schematic diagram showing the packaging (that is, referencing and importing of ontologies) to make the necessary knowledge concepts and relations known to the application program. The concept-mapping tool CMAP was used to generate this figure. CMAP allows the embedding of more detailed concepts within a grouping (for example, the SemanticFilter and SWEETOntologyPackage). Abbreviations and symbols are as follows: SWEET, Semantic Web for Earth and Environmental Terminology; << (to the right of the boxes), allows the box to be "closed" and only display a single box with the higher level name; boxes can be expanded by clicking on a corresponding >> icon.

support of investigations into relationships between volcanic activity and global climate (McGuinness and others, 2006, 2007; Fox, McGuinness, and others, 2007; Fox, Sinha, and others, 2007; Sinha and others, 2007). Another goal was to facilitate search and retrieval using an underlying framework that contains information about the semantics of the scientific terms that are used in the search. We also focused on the registration of disciplinary datasets in order to fully facilitate the integration of the volcanic and atmospheric data sources. We developed a tool to aid data providers with registering the data without explicitly knowing about the underlying ontologies.

Semantic Data Integration Methodology

We followed a methodology reported in previous work (Benedict and others, 2007) because our effort depended on machine-processable specifications of the scientific terms that are used in the study of volcanoes and the atmosphere. We identified specific ontology modules that need construction in the areas of volcanoes, plate tectonics, atmosphere, and climate, which draw heavily on existing ontologies. We used ontologies in the form of modules from the Jet Propulsion Laboratory's Semantic Web for Earth and Environmental Terminology (SWEET), Virtual Solar-Terrestrial Observatory (VSTO), and Geosciences Network (GEON). Our attention has been focused on the "Atmosphere-Volcano Use Case," whose goal is as follows: To determine the statistical signatures of both volcanic and solar forcings on the height of the tropopause.

We convened small workshop groups along these topic lines. We started with use cases and elements of the existing vocabularies or ontologies (where available) and proceeded to develop the knowledge representation. We used CmapTools, a concept mapping tool from Institute for Human and Machine Cognition (IHMC, http://cmap.ihmc.us/coe) that reads and writes Ontology Web Language- (OWL-) based ontologies and provides OWL-based predicate assistance for adding relations between concepts. Figure 1 shows how the ontologies were packaged, indicating the direction of importing and package dependency (see figure caption for more details).

We leveraged the VSTO framework (Fox and others, 2006, McGuinness and others, 2007) by replacing the solar-terrestrial-

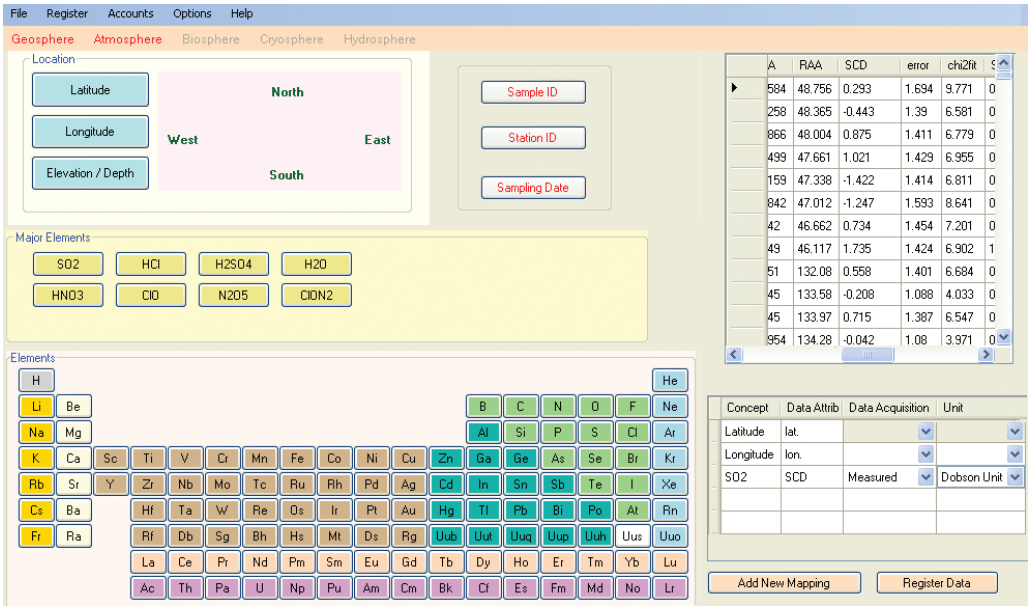


Figure 2. The SEDRE software tool as shown on the computer. There are four main panes in the window: (1) the upper left shows key level-1 and level-2 concepts such as location and related metadata for observations; (2) the lower left shows common compounds, the periodic table, and oxides (those that the user selects and associates with elements in the data table); (3) the upper right shows a preview of the data, where column headers are selectable so as to be associated with the concepts on the left two panes; and (4) the lower right is a display of the accumulated set of mapped relations (for example, SCD (Slant Column Density) is mapped to sulfur-dioxide as well as units, and so on).

specific ontology and data sources with appropriate volcano and atmospheric ontologies and data and catalog sources.

Our work required an even more modular approach to ontology re-use; the result was that we were able to conceive a new conceptual decomposition starting with SWEET 1.2 (http://sweet.jpl.nasa.gov). This effort will lead to the next version of the framework, which was based on broad community input and participation guided by the principle of re-use by other applications.

Data Registration

We base our data registration effort on the work from GEON (http://www.geongrid.org) and VSTO (http://www.vsto.org). The data registration sensibly consists of three levels:

1. Discovery of data resources, which requires registration through use of high-level index terms.

2. Discovery of item-level databases, which requires registration at data-type-level ontologies.

3. Item-detail-level registration (required for semantic integration).

We developed a software application known as the Semantically-Enabled Data Registration Engine (SEDRE) to implement the registrations. The tool is intended to be used by people with a variety of skill levels. We are using an ontology for the registration workflow for SEDRE for levels 1, 2, and 3 and the two "disciplines (+ sub-disciplines)": Geosphere+Geochemistry and Atmosphere+Atmospheric Chemistry.

Figure 2 shows one phase of registration of sulfur-dioxide data from a level-2 swath product from the European satellite-mounted Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY). A user opens a file and selects Atmosphere > Atmospheric Chemistry and the screen in figure 2 is displayed.

Discussion and Conclusion

We have presented the latest progress in an effort that uses modular ontologies to capture meanings of terms in distinct but related science domains with the goal of facilitating research into relationships between the domains and re-use of the modular ontologies. We have leveraged the existing starting points for reference ontologies in atmospheric science and partly developed ontologies for volcanoes and plate tectonics. We have held workshops to vet the ontologies among the multiple communities and completed a series of ontology mapping and merging exercises to arrive at the current modularization. The key element of data registration using developed ontologies is a new capability within the scientific Semantic Web community. The SEDRE tool we have developed is still evolving and we have only limited experience with user testing, but to date, the results and feedback are encouraging.

Acknowledgments

This work is supported by the National Aeronautics and Space Administration's (NASA's) Advancing Collabora-

tive Connections for Earth System Science (ACCESS) and Advanced Information Systems Technology (AIST) programs.

References Cited

Benedict, J.L., McGuinness, D.L., and Fox, Peter, 2007, A Semantic Web-based methodology for building conceptual models of scientific information: Eos, Transactions of the American Geophysical Union, v. 88, no. 52, Fall Meeting Supplement, Abstract IN53A–0950, p. F774.

Fox, Peter, McGuinness, D.L., Middleton, Don, Cinquini, Luca, Darnell, J.A., Garcia, Jose, West, Patrick, Benedict, James, and Solomon, Stan, 2006, Semantically-enabled large-scale science data repositories, *in* Cruz, Isabel, Decker, Stefan, Allemang, Dean, Proest, Chris, Schwabe, Daniel, Mika, Peter, Uschold, Mike, and Arroyo, Lora, eds., The Semantic Web—ISWC 2006, Proceedings, Fifth International Semantic Web Conference, Athens, Ga., November 5–9, 2006: Lecture Notes in Computer Science, v. 4273, p. 792–805.

Fox, Peter, McGuinness, D.L., Raskin, Rob, and Sinha, Krishna, 2007, A volcano erupts—Semantically mediated integration of heterogeneous volcanic and atmospheric data, *in* Proceedings of the ACM First Workshop on Cyberinfrastructure—Information Management in eScience, Lisbon, Portugal, November 9, 2007: New York, N.Y., Association for Computing Machinery, p. 1–6.

Fox, Peter, Sinha, Krishna, Raskin, Rob, McGuinness, D.L., Ammann, Caspar, Venezky, Dina, and Schwander, Florian, 2007, Semantic mediation and integration of volcanic and atmospheric data—In search of statistical signatures: Eos, Transactions of the American Geophysical Union, v. 88, no. 52, Fall Meeting Supplement, Abstract IN240A–05.

McGuinness, D.L., Fox, Peter, Cinquini, Luca, West, Patrick, Garcia, Jose, and Benedict, J.L., 2007, The Virtual Solar-Terrestrial Observatory—A deployed semantic web application case study for scientific research, *in* Cheetham, William, and Goker, Mehmet, eds., Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence, Vancouver, British Columbia, Canada, July 22–26, 2007: Menlo Park, Calif., Association for the Advancement of Artificial Intelligence Press, p. 1,730–1,737

McGuinness, D.L., Sinha, A.K., Fox, Peter, Raskin, Rob, Heiken, Grant, Barnes, Calvin, Wohletz, Ken, Venezky, Dina, and Lin, Kai, 2006, Towards a reference volcano ontology for semantic scientific data integration: Eos, Transactions of the American Geophysical Union, v. 87, no. 36, Joint Assembly Supplement, Abstract IN42A–03, also available online at http://www.agu.org/. (Accessed August 21, 2008.)

McGuinness, D.L., Fox, Peter, Sinha, A.K., and Raskin, Robert, 2007, Semantic integration of heterogeneous volcanic and atmospheric data, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2007—Data to Knowledge, San Diego, Calif., May 17–19, 2007: U.S. Geological Survey Scientific Investigations Report 2007–5199, p. 10–13.

Sinha, A.K., McGuinness, D.L., Fox, Peter, Raskin, Robert, Condie, Kent, Stern, Robert, Hanan, Barry, and Seber, Dogan, 2007, Towards a reference plate tectonics and volcano ontology for semantic scientific data integration, *in* Brady, S.R., Sinha, A.K., and Gundersen, L.C., eds., Proceedings, Geoinformatics 2007—Data to Knowledge, San Diego, Calif., May 17–19, 2007: U.S. Geological Survey Scientific Investigations Report 2007–5199, p. 43–46.

Manuscript approved for publication September 25, 2008.
Prepared by Reston Publishing Service Center.
Editing by Elizabeth D. Koozmin.
Photocomposition and design by Cathy Y. Knutson and Anna N. Glover.
For more information concerning this report, please contact
Shailaja R. Brady, U.S. Geological Survey, 911 National
Center, Reston, VA 20192, srbrady@usgs.gov.

ISBN 978-1-4113-2279-0