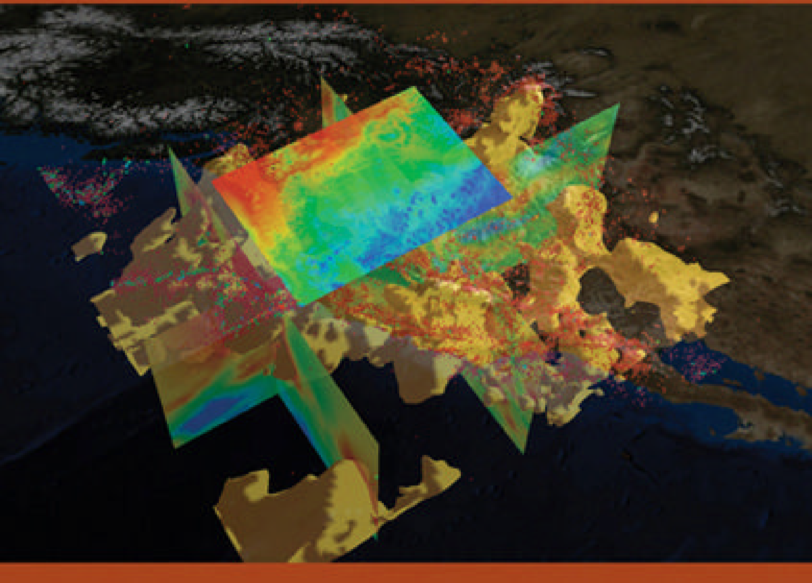
# Geoinformatics

## Cyberinfrastructure for the Solid Earth Sciences



EDITED BY
G. Randy Keller
and Chaitanya Baru

CAMBRIDGE

CAMBRIDGE
more information - www.cambridge.org/9780521897150

# GEOINFORMATICS

Cyberinfrastructure for the Solid Earth Sciences

Advanced information technology infrastructure is being employed increasingly in the Earth sciences to provide researchers with efficient access to massive databases and an ability to integrate diversely formatted information from a variety of sources.

A range of geoinformatics initiatives are enabling manipulation, modeling, and visualization of Earth Science data and are helping to develop integrated Earth models at various scales, and from the near surface to the deep interior.

This book provides a series of case studies that demonstrate the use of cyberinfrastructure across the Earth Sciences. Chapters are grouped thematically into sections that cover data collection and management; modeling and community computational codes; visualization and data representation; knowledge management and data integration; web services and scientific workflows.

Geoinformatics is a fascinating and accessible introduction to this emerging field for readers across the solid Earth sciences and is an invaluable reference for researchers interested in initiating new cyberinfrastructure projects of their own.

G. RANDY KELLER is a Professor of Geophysics and McCullogh Chair at the Mewbourne College of Earth and Energy, University of Oklahoma.

CHAITANYA BARU is a Distinguished Scientist at the San Diego Supercomputer Center, University of California-San Diego.

The two editors have collaborated since the emergence of geoinformatics as an organized scientific initiative in the USA in the late 1990s – helping to lead and organize the US Geoinformatics initiative and communicate its potential to colleagues around the world, both informally and through many appointments to advisory committees. Both Professor Keller and Dr. Baru are also Principal Investigators on the GEON (Geoscience Network) project that is a major effort funded by the National Science Foundation. While early geoinformatics programmes focused on database creation and on the development of highly functional software tools, these have since been merged with other efforts, such as high-performance computing and integrated earth-system modeling, to create a more extensive cyberinfrastructure for the geosciences. Dr. Baru's work at the San Diego Supercomputer Center has involved cyberinfrastructure activities across a range of scientific subject areas, while Professor Keller's research has focused on applications specific to the geosciences.

# GEOINFORMATICS

## Cyberinfrastructure for the Solid Earth Sciences

*Edited by*

G. RANDY KELLER
*University of Oklahoma*

CHAITANYA BARU
*University of California-San Diego*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge, CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521897150



First published 2011

Printed in the United Kingdom at the University Press, Cambridge

*A catalog record for this publication is available from the British Library*

ISBN 978-0-521-89715-0 Hardback

# Contents

The color plate section can be found between pages 114 and 115.

# Contributors

**Arun Agarwal,** University Center for Earth & Space Sciences, University of Hyderabad, Central University P.O., Hyderabad – 500 046, India

**Timothy K. Ahern,** IRIS Data Management Center, 1408 NE 45 St., Suite 201, Seattle, WA 98105, USA

**Michael Aivazis,** Center for Advanced Computer Research, California Institute of Technology, Pasadena, CA 91125, USA

**Raed Aldouri,** Regional Geospatial Service Center, University of Texas at El Paso, El Paso, TX 79968, USA

**M. Lee Allison,** Arizona Geological Survey, 416 W. Congress St., Suite 100, Tucson, AZ 85701, USA

**Ilkay Altintas,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Luis Armendariz,** Computational Infrastructure for Geodynamics, California Institute of Technology, Pasadena, CA 91125, USA

**J Ramón Arrowsmith,** School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287-1404, USA

**Hassan Babaie,** Department of Geosciences, Georgia State University, P.O. Box 4105, Atlanta, GA 30302-4105, USA

**Chaitanya Baru,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Ralf Bill,** University of Rostock, Geodesy and Geoinformatics 18051 Rostock, Germany

**H. P. Bunge,** Department of Earth and Environmental Sciences, Munich University, Theresienstr. 41, 80333 Munich, Germany

**James A. Boyden,** University of Sydney, Madsen Building (F09), Room 410, Sydney, NSW 2006, Australia

**John S. Cannon,** University of Sydney, School of Mathematics and Statistics (F07), Sydney, NSW 2006, Australia

**Robert Casey,** IRIS Data Management Center, 1408 NE 45 St., Suite 201, Seattle, WA 98105, USA

**Cinzia Cervato,** Iowa State University, Department of Geological and Atmospheric Sciences, 253 Science I, Ames, Iowa 50011-3212, USA

**Amit Chourasia,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**James A. Clark,** University of Sydney, Madsen Building (F09), Room 410, Sydney, NSW 2006, Australia

**Ryan J. Clark,** Arizona Geological Survey, 416 W. Congress St., Suite 100, Tucson, AZ 85701, USA

**Sierd Cloetingh,** Department of Tectonics, Faculty of Earth and Life Sciences, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

**Peter Cornillon,** Graduate School of Oceanography, University of Rhode Island, Narragansett, RI 02882, USA

**Daniel Crawl,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Christopher J. Crosby,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Nicholas Del Rio,** Department of Computer Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

**Peter Fox,** Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, 2nd floor, Troy, NY 12180, USA

**Aida Gandara,** Department of Computer Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

**Ann Gates,** Cyber-ShARE Center, Department of Computer Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

**Wolfgang Grunberg,** Arizona Geological Survey, 416 W. Congress St., Suite 100, Tucson, Arizona 85701, USA

**Linda C. Gundersen,** U.S. Geological Survey, 12201 Sunrise Valley Drive, MS 911, Reston, VA 20192, USA

**Michael Gurnis,** Seismological Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

**Richard Hughes,** British Geological Survey, Kingsley Dunham Centre, Keyworth, Nottingham, NG12 5GG, UK

**Hamish Ivey-Law,** Institut de Mathématiques de Luminy, Université de la Méditerannée Aix-Marseille II, Marseille, France

**Ian Jackson,** British Geological Survey, Kingsley Dunham Centre, Keyworth, Nottingham, NG12 5GG, UK

**G. Randy Keller,** University of Oklahoma, School of Geology and Geophysics, 100 E. Boyd Norman, OK 73019, USA

**Jens Klump,** Helmholz Centre Potsdam German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

**Sriram Krishnan,** San Diego Supercomputer Center, University of California-San Diego, MC0505, San Diego, CA 92093-0505, USA

**Walter Landry,** Computational Infrastructure for Geodynamics, California Institute of Technology, Pasadena, CA 91125, USA

**Matthias Lendholt,** Helmholtz Centre, Potsdam German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

**Qingsong Li,** Lunar and Planetary Institute, Houston, TX 77058, USA

**Mian Liu,** Dept. of Geological Sciences, University of Missouri, Columbia, MO 65211, USA

**Peter Löwe,** Helmholz Centre Potsdam German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

**Gang Luo,** Dept. of Geological Sciences, University of Missouri, Columbia, MO 65211, USA

**Ravi Madduri,** Argonne National Lab, 9700 South Cass Avenue, Building 221, Argonne, IL 60439-4844, USA

**David R. Maidment,** Center for Research in Water Resources, University of Texas at Austin, Austin, TX 78712, USA

**Deborah L. McGuinness,** Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, 3rd floor, Troy, NY 12180, USA

**Charles Meertens,** UNAVCO Inc., 6350 Nautilus Drive, Boulder, CO 80301-5554, USA

**R. Dietmar Müller,** University of Sydney, Madsen Building (F09), Room 406, Sydney, NSW 2006, Australia

**Viswanath Nandigam,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Omar Ochoa,** Department of Computer Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

**Paulo Pinheiro da Silva,** Cyber-ShARE Center, Department of Computer Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

**Stephen M. Richard,** Arizona Geological Survey, 416 W. Congress St., Suite 100, Tucson, AZ 85701, USA

**Peter M. Sadler,** Department of Earth Sciences, University of California, Riverside, 900 University Ave., Riverside, CA 92521, USA

**Leonardo Salayandia,** Cyber-ShARE Center, University of Texas at El Paso, El Paso, TX 79968, USA

**Leif Strand,** Computational Infrastructure for Geodynamics, California Institute of Technology, Pasadena, CA 91125, USA

**K. V. Subbarao,** University Center for Earth & Space Sciences, University of Hyderabad, Central University P.O., Hyderabad – 500 046, India

**Eh Tan,** Computational Infrastructure for Geodynamics, California Institute of Technology, Pasadena, CA 91125, USA

**Trond H. Torsvik,** PGP, University of Oslo, Physics Building, Sem Selands vei 24, Oslo NO-0316, Norway

**Mark Turner,** Tectonics Observatory, California Institute of Technology, Pasadena, CA 91125 USA

**Joachim Wächter,** Helmholtz Centre, Potsdam German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

**Rajeev Wankar,** Department of Computer and Information Sciences, University of Hyderabad, Central University P.O., Hyderabad – 500 046, India

**Robin J. Watson,** Norwegian Geological Survey, NGU, Postboks 6315 Sluppen, Trondheim NO-7491, Norway

**Stuart Wier,** UNAVC Inc., 6350 Nautilus Drive, Boulder, CO 80301-5554, USA

**Youqing Yang,** Dept. of Geological Sciences, University of Missouri, Columbia, MO 65211, USA

**Ilya Zaslavsky,** San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505, USA

**Huai Zhang,** Laboratory of Computational Geodynamics, Graduate University of the Chinese Academy of Sciences, Beijing 100049, P. R. China

# Preface

The idea for this book arose out of the development of Geoinformatics as a research emphasis in the Division of Earth Sciences at the U. S. National Science Foundation (NSF) that was fostered by Dr. Herman Zimmerman over a decade ago. This development was coincident with the creation of the Information Technology Research (ITR) for National Priorities program at NSF. Thus, the content of this book features several papers based on research that was inspired and/or funded as a result of these developments. Geoinformatics is certainly not a term or concept that is the invention of the U. S. earth sciences community, and this book also features chapters by authors from nine different countries. In addition, most chapters are the result of research in which geoscientists and computer scientists work together to solve complex scientific questions. This research involves topics such as data systems and models, data integration, advanced computer simulations, visualization, ontologies, workflows, webservices, and international cooperation.

Over most of the past decade, we have been supported by the National Science Foundation to pursue research in Geoinformatics primarily through the GEON (Geosciences Network) project, and we gratefully acknowledge this support. GEON was not created in a vacuum. A number of distributed and grid computing-based projects were in early stages at the time that GEON was originally conceived, including the Grid Physics Network (GriPhyN), funded by the US National Science Foundation and other agencies; the Biomedical Informatics Research Network (BIRN), funded by the National Center for Research Resources (NCRR) at the US National Institutes for Health; and the Southern California Earthquake Consortium's Common Modeling Environment (CME), which was also funded by an NSF ITR grant. From its inception, GEON recognized the need for "cross training" between earth and computer scientists. Dr. Margaret Leinen, then Assistant Director for Geosciences at NSF, proposed that one way to facilitate such cross training was by organizing summer institutes that would attract students, researchers, and faculty from both groups. This vision was realized, and the 7th

Cyberinfrastructure Summer Institute for Geoscientists (CSIG) was held in 2010. In addition, a series of meetings on Geoinformatics have been organized, and the most recent one was held in Potsdam, Germany in 2008; its proceedings are available at http://pubs.usgs.gov/sir/2008/5172.

The activities mentioned above have set the stage for programs such as the U.S. Geoscience Information Network (GIN) and OneGeology, which are represented in this book and which are now tackling the organizational issues (as opposed to purely technical ones) surrounding building metadata-based discovery and search across many organizations around the world.

Finally, we want to express our appreciation to the 58 authors who contributed to the 24 chapters in this book. Geoinformatics is a vibrant and dynamic field. It has taken over 2 years to finalize the contents of this book, and their patience is greatly appreciated. We also want to acknowledge the staff at Cambridge University Press who were helpful, knowledgeable, and effective at every step along the path that led to publication of this book.

# Introduction

# 1

# Science needs and challenges for geoinformatics

G. RANDY KELLER

## 1.1 What is geoinformatics?

Before we can begin to discuss geoscience informatics needs and challenges, we must first explain our use of the term geoinformatics for the purposes of this book. Over the past decade geoinformatics has become a term that has been independently employed by groups in several geospatial and geoscience fields around the world. In addition, this word appears in the title of several periodical publications. For example, there is an online magazine named *GeoInformatics* (www.geoinformatics.com) and an *International Journal of Geoinformatics* (www.j-geoinfo.net) that primarily focus on geospatial data and analysis within a geographic information system (GIS) framework. However, our emphasis in this book is on the data, software tools, and computational infrastructure that are needed to facilitate studies of the structure, dynamics, and evolution of the solid Earth through time, as well as the processes that act upon and within it from the near surface to the core. To approach such challenges, we must not only think and work in 3-D spatially, but we must include a 4th dimension, time. Time in this case ranges from seconds, such as in an earthquake, to millions of years, such as in plate movements over the Earth. *Here we have used **geoinformatics** to describe a variety of efforts to promote collaboration between computer scientists and geoscientists to solve complex scientific questions*. This book builds on the foundation of a book entitled *Geoinformatics: Data to Knowledge* (Sinha, 2006) that emphasized databases and their analysis, but here we emphasize topics such as web services, modeling of earth processes, visualization, and international developments.

At the U.S. National Science Foundation (NSF), geoinformatics has emerged as an initiative within the Earth Sciences Division to address the growing recognition that Earth functions as a complex system, and that existing information science infrastructure and practice within the geoscience community are inadequate to address the many difficult problems that must be overcome to understand this system (e.g., Allison *et al.*, 2002). In addition, there is now widespread recognition that successfully addressing these problems requires integrative and innovative approaches to analyzing, modeling, and developing extensive and diverse datasets.

Currently, the geoscience community is awash in data due to many new satellite observing systems that provide data to study phenomena such as changes in the Earth's surface via multi-band remote sensing (e.g., ASTER), the Earth's gravity field and small changes in it (e.g., GRACE), vertical movements of the Earth's surface (e.g., inSAR), the topography of the Earth (SRTM: Shuttle Radar Topography Mission), and the Earth's magnetic field (Maus *et al*., 2010). Also, massive amounts of seismological data are being archived in databases around the world. However, a lack of easy-to-use access to modeling and analysis codes are major obstacles for scientists and educators alike who attempt to use these data to their full potential, especially in a highly integrated fashion. However, recent advances in fields such as computational methods, visualization, and database interoperability provide practical means to overcome such problems and some examples are presented in this book. Thus, in *addition to the statement above*, ***geoinformatics*** *can be thought of as the field in which geoscientists and computer scientists are working together to provide the means to address a variety of complex scientific questions using advanced information technologies and integrated analysis*. This type of activity is also being called cyberinfrastructure.

## 1.2 Geoinformatics as a scientific tool is data driven

Open access to data from satellites is very common but spatial resolution is a limitation for many applications. In many cases, access to land-based or low-altitude measurements and even maps remains an issue in many countries due to government policies, but progress is being made on many fronts (e.g., gravity data, Aldouri *et al.*; seismic data, Casey and Ahern, this volume). Even though many useful datasets are emerging, discovering and accessing them is difficult if scientists wish to find the very best data for their particular application or research project. However, a very promising example of the development of an advanced data discovery and access system is the Global Earth Observation System of Systems (GEOSS) whose 10-Year Implementation Plan states that the purpose of GEOSS is "to realize a future wherein decisions and actions for the benefit of humankind are informed via coordinated, comprehensive and sustained Earth observations and information." GEOSS is seen by its participants as an important contribution to meeting United Nations Millennium Development Goals and to furthering the implementation of international treaty obligations (www.earthobservations.org).

In an ideal world, geospatial data developed by governmental agencies or by researchers using governmental support would be freely and openly available. However, crafting high-quality, easily accessible databases is expensive, especially if legacy data are to be converted to digital form. Thus in many cases, it is not possible for data to be accessible free of charge, but costs need to be low enough to make them available to a broad cross-section of users (e.g., Jackson and Hughes, this volume).

In order to understand the subsurface of the Earth, one has to be able to correlate known geological features with geophysical data and models. OneGeology (Jackson, this volume) is an example of international geological organizations banding together to produce a geological map of Earth's surface. Such a product would be invaluable to countless researchers, governmental agencies, environmental protection efforts, and planning efforts to name a few.

## 1.3 Geoinformatics as a scientific tool seeks to foster the development of community-based software

A guiding principle in geoinformatics is fostering community-based development of software that is open source and highly usable (e.g., Gurnis *et al.*, this volume). In the following chapter, Baru discusses the technical issues and developments that affect this and other technical challenges that affect geoinformatics, but below I discuss an example of major scientific need.

### *1.3.1 Building 3-D models*

Today, a major research goal in the geosciences is the construction of geologically realistic (i.e., as complex as in nature) 3-D models of earth structure and variations in physical properties such as seismic velocity (P-wave and S-wave), density, and electrical resistivity. The physical basis of many geophysical techniques is inherently scale-independent, so it is realistic to aspire to build models that range in scale from the near surface (environmental and groundwater studies), to geologic studies of features such as basins and fault zones, to studies of tectonic plates and their boundaries (e.g., Boyden *et al.* and Liu *et al.*, this volume), to mantle dynamics, to studies of the core and its boundaries. In order to construct such models, software that enables the integration of a wide range of geological and geophysical data is required. This software should also facilitate the application of empirical and theoretical relationships that provide constraints for integrated modeling via estimations of relationships between various physical properties (e.g., P-wave velocity, S-wave velocity, and density; Brocher, 2005), the effects of porosity (e.g., Mavko *et al.*, 1998), and the effects of pressure and temperature (e.g., Perry *et al.*, 2006).

One way to conceive of an ideal model would be for it to consist of geological structures and major discontinuities in physical properties that are represented by surfaces that bound layers and within which variations in multiple physical properties are associated with voxels, which need not be cubical in form. Since the resolution of geophysical techniques decreases with depth, it would make sense that the size of the

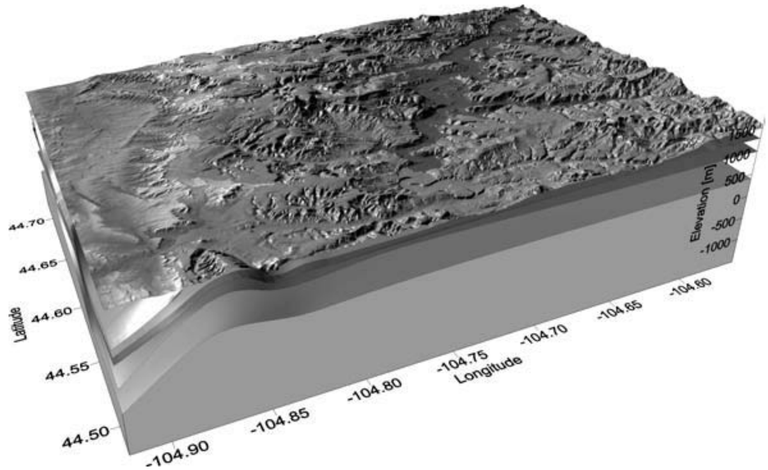



Figure 1.1. Example of a 3-D geological/geophysical model consisting of layers that are bounded by geologic interfaces that have been extracted from surface and subsurface geologic data. The lowest layer is the Precambrian basement. The interfaces are georeferenced and provide a framework for assigning physical properties to the layers between them. Image provided by Kevin Crain. See color plates section.

voxels would increase with depth. This type of model is shown in Figure 1.1, where the topographic relief and surfaces that represent the tops of a series of stratigraphic units are shown above the last surface, which is the top of the Precambrian basement. The concept is that these surfaces bound the stratigraphic layers and Precambrian basement that form the model. These layers can then be populated with voxels with associated physical properties based on studies of samples collected from exposures, data from drill holes, and geophysical surveys. In this ideal case, the resulting model would be structured in a form that would facilitate calculations such as various geophysical responses, fluid flow in the layers, and response to stress. Modeling a response to stress would be an example of adding the dimension of time to the analysis.

In most cases, seismic data have the highest spatial resolution (and cost) of subsurface imaging techniques, and many diverse techniques are available to process and analyze these data at various spatial and depth scales. Each type of seismic data has its own sensitivities and resolution and can constrain important aspects of earth structure. For example, tomographic modeling is based on voxels, seismic refraction/wide-angle reflection data produce models with interfaces and velocity values measured directly, and seismic reflection data produce images of earth structures from which surfaces and discontinuities such as faults can be extracted. It is intuitively obvious that, when a variety of seismic data are used together in a quantitative manner, the resulting earth model should be better resolved than in the typical approach of simply comparing results qualitatively. However, proving this inference mathematically is not easy. As constraints from geological and drilling data and other geophysical techniques are added, the resolution will improve further, which is also hard to prove mathematically. These extra data also make it possible to add non-seismic physical properties (e.g., density, electrical conductivity, magnetic susceptibility) to the model.

Tools for modeling seismic data and honoring independent constraints exist for 2-D approaches, and an example of some preliminary results from a large experiment in Central Europe (Figure 1.2) are shown in Figure 1.3. The final scientific results of the analysis of the long profile (CEL05, Fig. 2) are presented in Grad *et al.* (2006).

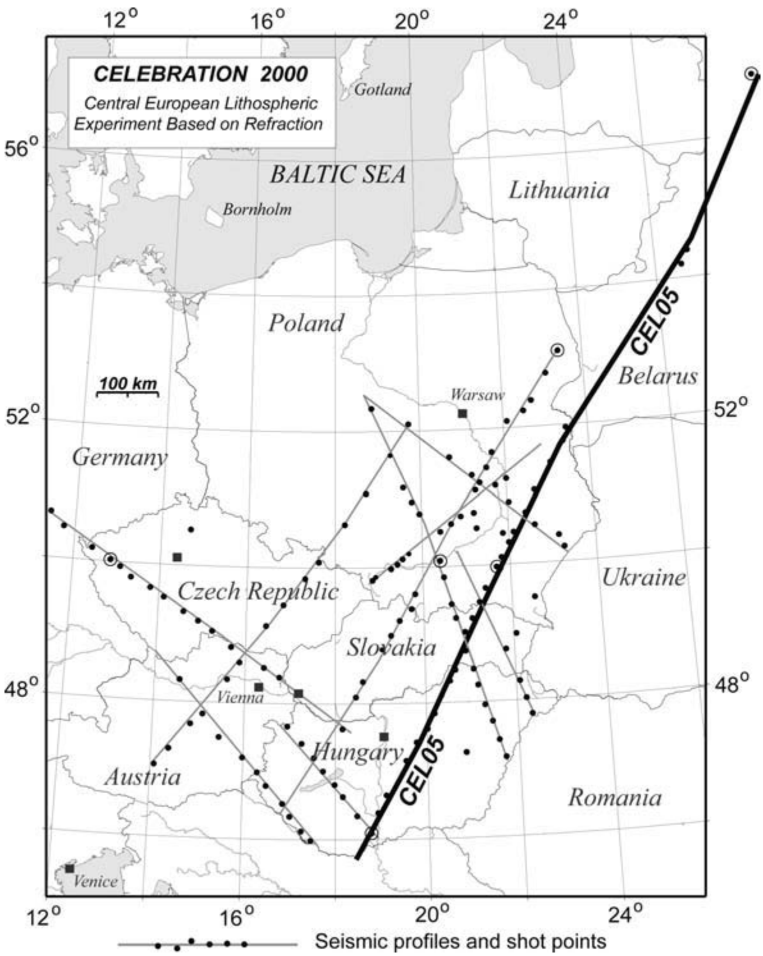



Figure 1.2. Index map of the CELEBRATION 2000 seismic experiment showing the location of the 1400 km long CEL05 profile (heavy black line). The gray lines indicate the location of other profiles that were recorded. The seismic velocity models shown in Figure 1.3 are for this profile.

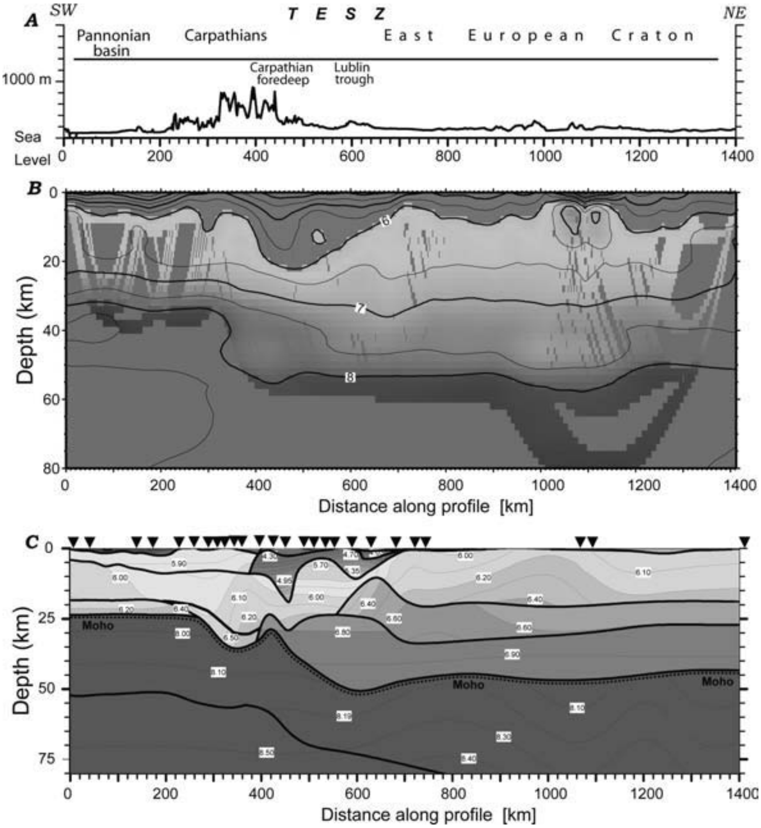



Figure 1.3. (*A*) Topographic profile showing the main geologic features present; (*B*) Preliminary seismic velocity model derived by tomographic inversion of the arrival times of the first seismic wave observed. The model is smooth and lacks the detail that is needed to make a suitable geological interpretation. The numbers in the model are P-wave velocities in km/s; (*C*) Seismic velocity model derived by ray trace forward (trail-and-error) modeling of all observed seismic arrivals. This approach has the advantage of providing more detail, but the formal analysis of certainty is difficult. The numbers in the model are P-wave velocities in km/s. Inverted triangles indicate the locations' shot points that produced the observed seismograms. See color plates section.

The tomographic result (*B*) shows the broad variations in seismic velocity based on voxels. Using the tomographic result as a starting point, modeling of waves reflected and refracted at interfaces from within the Earth add structural detail (*C*) that can be interpreted geologically. In turn, the upper few kilometers of the model could be further refined using geological, drilling, and other types of geophysical data. Presently, expanding this example of an analysis scheme to 3-D, quantitatively assessing resolution, and moving smoothly between modeling approaches are at best very challenging. The software tools that do exist for 3-D modeling (e.g., Hole, 1992) need further development, need to be interoperable, and need to facilitate integrated analysis.

In summary, scientific advances on many fronts face technical barriers that require a geoinformatics approach if they are to be overcome. In a lot of cases, there are large volumes of data to examine and mine, and in others, interoperability between analysis and modeling software is needed. Obviously, providing the "best" integrated model of earth structure possible with existing data is a goal that we are far from achieving, except in very special circumstances. Thus, geoscientists and computer scientists have many interesting and important problems that they can attack together in the future.

## References

Allison, M. L., Snyder, W. S., and Walker, J. D. (2002). Geoinformatics: A nascent revolution in the earth sciences. *GSA Today*, **12**: 17.

Brocher, T. M. (2005). Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bulletin of the Seismological Society of America*, **95**: 2081–2092.

Grad, M., Guterch, A., Keller, G. R. *et al*. (2006). Lithospheric structure beneath trans-Carpathian transect from Precambrian platform to Pannonian basin: CELEBRATION 2000 seismic profile CEL05. *Journal of Geophysical Research*, **111**, B03301, doi:10.1029/2005JB003647.

Hole, J. A. (1992). Nonlinear high-resolution three-dimensional seismic travel time tomography. *Journal of Geophysical Research*, **97**: 6553–6562.

Maus, S., Barckhausen, U., Berkenbosch, H. *et al*. (2010). EMAG2: A 2-arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. *Geochemistry Geophysics Geosystems*, **10**, Q08005, doi:10.1029/2009GC002471.

Mavko, G., Mukerji, T., and Dvorkin, J. (1998). *The Rock Physics Handbook: Tools for Seismic Analysis in Porous Media*. Cambridge: Cambridge University Press, 329pp.

Perry, H. K. C., Jaupart, C., Mareschal, J.-C., and Shapiro, N. M. (2006). Upper mantle velocity-temperature conversion and composition determined from seismic refraction and heat flow. *Journal of Geophysical Research*, **111**, B07301, doi:10.1029/2005JB003921.

Sinha, A. K., ed. (2006). *Geoinformatics: Data to Knowledge*. Geological Society of America, Special Paper 397, 283pp.

# 2

# Introduction to IT concepts and challenges

CHAITANYA BARU

Scientific applications have been at the forefront of driving computer and information technology since the early days: from the development of early computers for numerical computing, to the introduction in the USA of the NSFNET (which helped launch the Internet), and the subsequent invention of the World Wide Web. The geosciences, in particular, have been a long-standing user of such technologies, given the importance of applications related to weather, natural resources, natural hazards, and environmental monitoring. Scientific computing was focused initially on the need for fast computers to perform larger numbers of complex numerical calculations. The concerns more recently have turned towards the ability to manage the very large amounts of data that are being generated by a wide range of sensors and instruments, sophisticated observing systems, and large-scale simulations on large computer systems. Data rates of terabytes per day and petabytes per year are not uncommon (1 petabyte = terabytes) (Hey *et al.*, 2009, p. 9). Yet, computer science and information technology solutions must deal not only with the size and scale of data, but also the inherent richness and complexity of scientific data – especially when data are combined across multiple projects, institutions, and even multiple science disciplines and subdisciplines. The need to understand complex, interdependent, natural as well as anthropogenic phenomena has made science a *team sport*, requiring collaborations among multidisciplinary teams of scientists to process, analyze, and integrate extremely heterogeneous data.

The *e-Science* initiative in Europe and the *cyberinfrastructure* initiative in the United States were launched in the early 2000s to tackle these issues, by harnessing the power of advanced information technologies for scientific research and education. Scientific research, it has been suggested, has entered the *fourth paradigm* (Hey *et al.*, 2009). The first three being *empirical*: focused on observations and descriptions of natural phenomena; *theoretical*: focused on the development and use of models and generalization of scientific principles; and, *computational*: focused on simulations of complex phenomena using computers. This fourth paradigm is *data intensive*, focused on building unified theories of complex phenomena, but based on data exploration and integration using software tools and computer platforms capable of dealing with complex data and large data (Hey *et al.*, 2009, p. 177).

## 2.1 Cyberinfrastructure and geoinformatics

The study of complex phenomena in earth, ocean, and atmospheric sciences all require integration of heterogeneous data from a wide variety of sources and disciplines. As in every area of science, discovery in the geosciences is also driven by the ease and efficiency with which one is able to do this integration by manipulating and assimilating large, heterogeneous datasets. Remote sensing instrument and observing systems are able to generate rapidly large amounts of data, while large-scale computational models are able to generate increasingly large outputs that require post-processing, visualization, and eventually integration with other simulation, observational, and contextual data. A range of cyberinfrastructure capabilities is needed to provide such capabilities and to support scientific research and discovery at the frontiers of the earth sciences.

NSF's *Cyberinfrastructure Vision for 21st Century Discovery* describes the set of challenges and opportunities in computing systems, data, information resources, networking, digitally enabled sensors, instruments, virtual organizations, and observatories, along with an interoperable suite of software services and tools (NSF, 2007). As described in the report, this technology is complemented by the interdisciplinary teams of professionals who are responsible for its development, deployment, and its use in transformative approaches to scientific and engineering discovery and learning. The vision also includes attention to the educational and workforce initiatives necessary for both the creation and effective use of cyberinfrastructure. Figure 2.1 depicts the set of cyberinfrastructure components, from hardware platforms, systems software, middleware services, user services/functions, and a portal providing access to this environment.

As mentioned in Chapter 1, geoinformatics is the term used to describe the set of activities related to the development and use of cyberinfrastructure for the earth sciences. The area has been making rapid progress since the early 2000s, with the introduction by NSF of its cyberinfrastructure initiative and, subsequently, the *geoinformatics* program in the Earth Sciences Division (NSF EAR, 2010). Since then, major geosciences professional organizations have also recognized geoinformatics as a special area. Both the American Geophysical Union (AGU) and the European Geophysical Union have an Earth and Space Science Informatics focus area (AGU, 2010; EGU, 2010). The Geological Society of America created a Geoinformatics division, which defined geoinformatics as "*the science discipline that utilizes cyber-products, tools and discovery of data and*

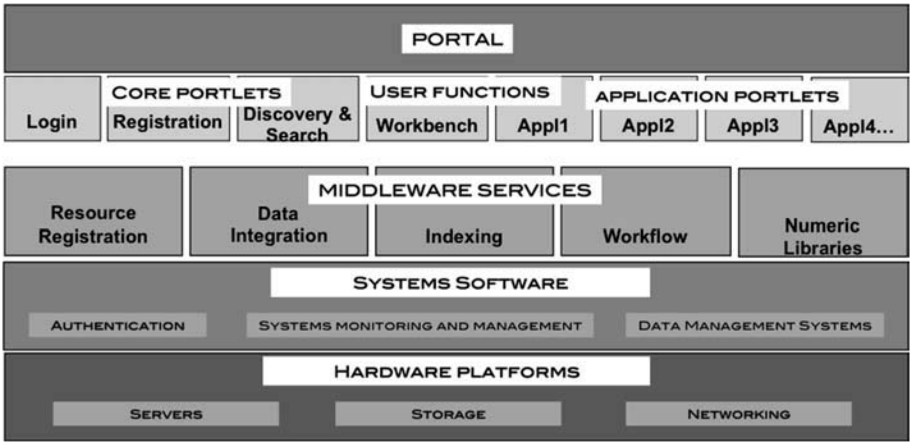



Figure 2.1. Cyberinfrastructure hardware and software layers.

*models towards exploring integrative solutions to earth and planetary systems"* (GSA, 2006).

Development and deployment of cyberinfrastructure at the "speed of research," so to speak, is an ongoing challenge. While the capabilities provided by information technology progress rapidly, so does research in the geosciences. The challenge is to keep the two connected so that the science activities benefit from advances in information technology while, at the same time, the right type of information technology is being created to help the science mission. The term *cyberinfrastructure* implicitly includes this two-way interaction between the science domains and computer science and information technology. Similar to internet-scale services such as, say, *email*, *web search*, and *street maps* that have now become an everyday phenomenon, the goal of cyberinfrastructure is to make the use of information technology equally ubiquitous in all aspects of research and education across all domains of study – natural sciences, engineering, social sciences, humanities, and medicine. For example, Chapter 3 (Fox and McGuinness) and Chapter 4 (Krishnan and Madduri) in this book provide overviews of cyberinfrastructure efforts in some other domains of science.

To achieve this goal, the systems developed must not only be essential to the mission at hand, but also intuitive and easy to use. An example of effective use of information technologies is from the exploration industry where, typically, groups of experts from a wide range of disciplines, e.g., geology, geophysics, engineering, and economics, meet to interrogate and integrate heterogeneous datasets to make decisions related to, say, further exploration at a given location. These sessions are essential to the conduct of business, but cannot happen without sophisticated software and visualization systems that help bring together the myriad data types needed for analysis in support of decision-making. There is necessarily interplay here between the technology and the science. Available technologies influence the nature of the solution – whether using FORTRAN versus C++ for programming; relational databases versus GIS for spatial data management; or, more recently, web services versus Grid computing for distributed processing. At the same time, the demands of the science push technologies – whether in the area of more advanced sensor systems, larger, 3-D visualization displays, or sophisticated software for 3-D and 4-D data integration.

## 2.2 Geoinformatics and IT concepts

Geoinformatics activities run the entire gamut from data acquisition and archiving, to database creation, data analysis, modeling and simulation, and integration and visualization. For example, large-scale projects such as EarthScope (2010), as well as individual PI-led efforts such as the High Lava Plains Experiment (2010) are engaged in deploying sensors and making field measurements. Established data archives such as the Incorporated Research Institutes for Seismology (IRIS, 2010), and UNAVCO (2010) provide repositories for such data. The EarthScope project has three large components – USArray, Plate Boundary Observatory (PBO), and the San Andreas Fault Observatory at Depth (SAFOD) – each of which has deployed a range of sensors (EarthScope Observatories, 2010). These data are made accessible via an EarthScope Data Portal (2010). Chapter 13 (Casey and Ahern) describes a set of web services that provide access to seismic data archives maintained by IRIS, which includes data from USArray. Typically, the sensor data have to be interpreted in the context of other extant data. Support for such information integration is a key challenge in geoinformatics.

Extant datasets may be available across a wide variety of databases, some maintained by independent (large or small) consortiums such as, say, CUAHSI (2010), NAVDAT (2010), and EarthChem (2010), and others maintained by individual groups or researchers, e.g., PGAP (2010) and the PaleoBiology database (PaleoDB, 2010). Chapter 12 (Zaslavsky and Maidment) describes the CUAHSI Hydrologic Information System (HIS), which provides access to distributed hydrologic data archives. Chapter 14 (Aldouri and Keller) describes a gravity and magnetic database that was developed by carefully assembling "primary" data from a variety of independent sources. Chapter 16 (Crosby *et al.*) describes a portal-based system that provides the capability to access and process large remote sensing datasets for high-resolution topography to a community of users. The datasets themselves are from other projects that have commissioned the surveys. Chapter 20 (Jackson) describes an international effort to assemble derived data for a global geologic map.

Other parts of the geoinformatics community develop software tools that are used for analysis. This includes complex computational codes that run on supercomputers and generate many terabytes of data, as well as desktop/laptop-based analysis tools that might operate on much smaller, though complex, data. The section on "Modeling software and community codes" (Chapter 5, Gurnis *et al.*), Chapter 6 (Liu *et al.*), and Chapter 7 (Boyden *et al.*) describes three different projects that are all focused on development of robust modeling software. Chapter 10 (Sadler and Cervato) describes tools for interpreting and analysing geologic time-related data. Development and maintenance of such "community codes" is a major geoinformatics activity.

With large and heterogeneous 3-D and 4-D (three dimensions of space + time) data, visualization of the data is essential for examining and properly interpreting the results. Many geoinformatics projects focus on development of such 3-D and 4-D visualization software as well as combined hardware and software environments for visualization. Chapter 8 (Chourasia) and Chapter 9 (Wier and Meertens) describe efforts in developing 3-D and 4-D visualizations of geoscience data. In the former, the data are from large-scale earthquake simulations conducted by the Southern California Earthquake Center (SCEC). In the latter, the chapter discusses a specific software system that supports integrated visualization of heterogeneous earth science data.

Once databases have been assembled and data archives deployed, there is the opportunity to provide a number of data access, analysis, and integration services for these data. A major need in the geosciences is for a data integration environment that would allow for easy incorporation of a variety of different datasets, representing different types of geoscience data from different subdisciplines, for a given spatial region specified by a spatial extent and depth/elevation. Such an integration environment would not only provide the ability to easily bring in different types of data, but would also allow for on-the-fly data processing, such as re-filtering or re-gridding of data, and even running models to evaluate fit between model outputs and other data derived from sensor or field observations (for example, see Youn, 2005), which describes a system for generating synthetic seismograms and comparing these with observed data). Chapter 15 (Altintas *et al.*) describes *scientific workflow* systems whose goal is to provide intuitive and easy-to-use analysis environments that remove the tedium of having to write a new computer program for each new analysis task, while providing powerful built-in capabilities such as fault tolerance and data provenance tracking. More advanced capabilities may be needed to represent, analyze, and interpret data within and across databases. Chapter 11 (Babaie) discusses the use of knowledge representation techniques, such as ontologies, for modeling geodynamic processes. Chapter 17 (Gates *et al.*) describes a system that provides infrastructure to track how results are derived, thereby helping users better understand the derived data products.

Significant geoinformatics activities are also underway at several institutions around the world. Chapter 18 (Jackson and Hughes) describes activities in the UK; Chapter 19 (Cloetingh *et al.*) describes a European project, TOPO-EUROPE, which is similar to the EarthScope project in the USA; Chapter 21 (Klump *et al.*) describes geoinformatics activities in Germany; and Chapter 22 (Subbarao *et al.*) describes a geoinformatics project in India. Whether in the USA or elsewhere, the largest geoscience data holdings are with government agencies. In the USA, that ranges from the U.S. Geological Survey (USGS) and state geological surveys, to agencies such as NASA, National Oceanic and Atmospheric Administration (NOAA), Environmental Protection Agency (EPA), Department of Energy (DOE), U.S. Department of Agriculture (USDA), U.S. Forest Service (USFS), and many others. These agencies have also been pursuing a vigorous geoinformatics agenda with internal initiatives and projects to enable easier discovery, access, and sharing of digital data. The U.S. Geological Survey and the various state geological surveys across the USA are also engaged in geoinformatics efforts aimed at easier discovery and access to their data holdings. Chapter 23 (Allison *et al.*) describes efforts at the USGS. Chapter 24 (Richard *et al.*) describes a DOE-funded effort to develop a National Geothermal Data System.

## 2.3 Challenges and the future

An NSF-funded workshop on "Envisioning a National Geoinformatics System for the United States," held in Denver, Colorado in March 2007 articulated a vision for geoinformatics that would facilitate "*. . . a future in which someone can sit at a terminal and have easy access to vast stores of data of almost any kind, with the easy ability to visualize, analyze and model those data.*"

Much progress has been made over the past several years in developing the cyberinfrastructure that would help realize such a future. Before long, one can envisage web clients from a web portal that would be able to query federated metadata catalogs and obtain detailed information about vast collections of geoscience data and tools. The environment would allow users to access datasets of their choosing, apply various processing routines to those data – perhaps using a Cloud computing platform or some other distributed processing platform, such as the NSF TeraGrid – and bring the data into online environments that would facilitate interrogation and integration of 3-D and 4-D geoscience data.

However, several challenges still remain in fully realizing this vision. With the rapid growth in data, the underlying cyberinfrastructure must be scalable and efficient in dealing with the large scale and wide variety of data. As the cyberinfrastructure enables access to data from a wide variety of distributed and heterogeneous sources, it will also need to assist users in properly interpreting data and dealing with *data semantics*. The suitability and usefulness of a given dataset must be interpreted in the context of the analysis that the user is attempting to perform. While standards have emerged for metadata describing resources and for the software interfaces to query the corresponding metadata catalogs, more work is needed. Standardized approaches to describing data collection *protocols* and data processing steps, especially for quality assurance and quality control (QA/QC), can facilitate better data access and sharing. Standards are needed for representing derived information, e.g., via the use of *controlled vocabularies* and *ontologies*, to minimize ambiguity when communicating data from one researcher (or system) to another. Indeed, several efforts are underway in the earth science and related disciplines for arriving at such standards, including the effort to develop a Geosciences Markup Language (GeoSciML, 2010).

As we make progress in bringing more digital data online and providing more online environments for processing, integrating, and visualizing these data, the focus will continue to increase on issues related to data quality, accuracy, data authenticity, provenance, and a host of related issues. One approach to addressing these complex issues is via the use of social networking techniques and technologies. Strategies such as community ranking of datasets can be used to obtain useful semantic information about datasets. Pre-existing levels of trust among individuals in a community can be used to evaluate the quality of a dataset or its suitability for a particular analysis. A user's comment about a dataset could then be interpreted within this context. Another key challenge for the cyberinfrastructure is to provide *repeatability*, i.e., the ability to repeat a set of processing steps with a given set of data to reproduce a previously published result. Such repeatability is essential for users to gain trust in the system. Finally, another major challenge, which is closely allied to the issue of repeatability, is that of data preservation. What technical, economic, and business-oriented strategies should be used to ensure that all of the important digital data that is being produced are preserved for the next generation of scientists and generations thereafter, so that future generations can correctly re-produce results from the past, and re-examine results and, perhaps, re-interpret them in a future context?

## References

AGU Earth and Space Science Informatics (2010), http://essi.gsfc.nasa.gov/.

CUAHSI: Consortium of Universities for the Advancement of Hydrologic Science, Inc. (2010), www.cuahsi.org.

EarthChem (2010), www.earthchem.org.

EarthScope (2010), www.earthscope.org.

EarthScope Data Portal (2010), http://portal.earthscop.org.

EarthScope Observatories (2010), http://earthscope.org/observatories.

EGU Earth and Space Science Informatics (2010), www.egu.eu/inside-egu/divisions-and-present-officers/division-earth-and-space-science-informatics/home.html.

GeoSciML Resource Repository (2010), www.geosciml.org.

GSA Geoinformatics Division Bylaws (2006), p. 1, March, www.geoexpertsintl.com/geoinformatics/pdf/geoinfoBylaws.pdf.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

HLP: High Lava Plains Experiment (2010), www.dtm.ciw.edu/research/HLP/.

IRIS: Incorporated Research Institutes for Seismology (2010), www.iris.edu.

NAVDAT: The Western North American Volcanic and Intrusive Rock Database (2010), www.navdat.org.

NSF Cyberinfrastructure Vision for 21st Century Discovery, National Science Foundation report (2007), www.nsf.gov/pubs/2007/nsf0728/index.jsp.

NSF EAR/IF (2010), NSF Earth Sciences Division: Instrumentation and Facilities, www.nsf.gov/pubs/2005/nsf05587/nsf05587.htm

PaleoDB: The Paleobiology Database (2010), http://paleodb.org.

PGAP: The Paleogeographic Atlas Project (2010), www.geo.arizona.edu/~rees/PGAPhome.html.

UNAVCO (2010), www.unavco.org.

Youn, C., Kaiser, T., Santini, C., and Seber, D. (2005). Design and implementation of services for a synthetic seismogram calculation tool on the Grid. In *Proceedings of the 5th International Conference (ICCS 2005)*, Atlanta, GA, USA, May 22–25, 2005. Part I, LNCS 3514. Berlin: Springer, p. 469.

# Part I

## Case studies from other disciplines

# 3

# Semantic cyberinfrastructure: The Virtual Solar-Terrestrial Observatory

PETER FOX, DEBORAH L. MCGUINNESS, AND THE VSTO[1] TEAM

## 3.1 Introduction

The utilization of now fairly well established information technology (IT) to provide access to science data and information is becoming more routine. The present foundational IT includes hardware services utilizing computers, disks, and networks along with software systems such as databases, web servers, portals, and so on. At the same time, scientific data are being collected or generated and maintained in digital form (repositories) in high volumes by many diverse research projects and groups. The need for access to, and interoperability between, these repositories is also growing, as research groups need to access their *own* increasingly diverse data collections. As investigations begin to include results from many different sources, researchers also need to access and utilize other research groups' data repositories from within a single discipline or, more interestingly, among multiple disciplines. Also, it is not simply trained scientists who are interested in accessing scientific data; nonspecialists are becoming interested in looking at trends in scientific data as well. Two notable examples are in the areas of natural hazards and environmental impact assessments. While the promise of true virtual interconnected heterogeneous distributed international data repositories is being realized in a number of areas there is still much work to be done. To address ultimate goals such as to provide support for both specialist and broader nonspecialist usage, including lay people from a rich set of science and information product, new approaches are being devised. One such approach is that of Virtual Observatories (VOs: Dalton, 2007). When viewed within one discipline, VOs are becoming popular and successful. For example, the NASA Virtual Observatories for Heliophysical Data (VOHD) program currently funds several VO efforts (King *et al.*, 2007). As these efforts move either beyond their discipline or to nonspecialist use, vocabulary challenges arise. Often vocabularies

[1] Luca Cinquini, Patrick West, Jose Garcia, James Benedict, Tony Darnell, Don Middleton, and Stephen Zednik.

differ, some are quite esoteric and jargon laden, sometimes similar terms have different meanings, and often there are multiple terms with different meanings, and multiple terms for the same phenomenon or process. These challenges present barriers to efforts that hope to use existing cyberinfrastructure in support of interdisciplinary data query and access, especially when the interdisciplinary applications must go beyond search and access to actual manipulation and use of the data. In addition, the user community now has a more diverse level of education and training and need.

One approach that has now gained acceptance and success in facing the above-mentioned challenges is that of addressing the semantics of the underlying science and data representations in a way that can be utilized by existing cyberinfrastructure; computers and people. Encoding formal semantics in the technical architecture of virtual observatories and their associated data frameworks is similar to efforts to add semantics to the Web in general (Berners-Lee *et al.*, 2006), workflow systems (e.g., Gil *et al.*, 2006; Ludaescher *et al.*, 2006), computational grids (e.g., De Roure *et al.*, 2005), and data mining frameworks (e.g., Rushing *et al.*, 2005).

In this chapter, we describe the Virtual Solar-Terrestrial Observatory project, which was funded by the U.S. National Science Foundation Office of Cyberinfrastructure (NSF/OCI) to explore the applications of semantic web to VOs in the areas of solar, solar-terrestrial, and space physics. We include our vision, design, and semantic web-enabled implementation. We highlight the methodologies and technologies we utilize and some production environment, which is built upon our previous cyberinfrastructure that started production in the summer of 2006 and has been continuously deployed since then.

## 3.2 New needs driven by use, not by technology

We believe that an essential and distinguishing feature of our approach was to find out how (and sometimes why) the diverse user base that we wished to support wanted to find, access, and use data. In particular, we wanted to let them find and use data and information that they would not normally have access to – due to lack of familiarity, vocabulary or discipline unfamiliarity, etc. We developed a series of use cases to reflect a set of diverse requirements (Fox *et al.*, 2007). We quickly determined, based on the use cases we collected, that the distributed multidisciplinary internet-enabled VSTO virtual observatory required a higher level of semantic interoperability than what had been previously required by most (if not all) distributed data systems or discipline-specific virtual observatories. In developing the use cases, we targeted subject matter experts as end users to elaborate their need to support the integration of multiple collections. This extended beyond providing basic access to search interfaces that were typically specialized and idiosyncratic.

Our initial science domains were those of interest to scientists who study the Earth's middle and upper atmosphere, the inner heliosphere, and the Sun. Our initial interdisciplinary virtual observatory is thus VSTO – the Virtual Solar-Terrestrial Observatory. Scientists in these areas must utilize a balance of observational data, theoretical models, analysis, and interpretation to make effective progress. Many data collections are interdisciplinary. In order to provide a scientific infrastructure that is usable and extensible, VSTO required contributions concerning semantic integration and knowledge representation, while requiring depth in a number of science areas. We chose a technology foundation that was based on a long history of artificial intelligence (AI) research set in the context of the modern world-wide-web (WWW) environment (Berners-Lee *et al.* 2006) because of the promise for a declarative, extensible, reusable technology platform. The value added by basic knowledge representation and reasoning is supporting both computer-to-computer and researcher-to-computer interfaces that find, access, and use data in a more effective, robust, and reliable way. What arose from this approach were both a semantic methodology as well as a number of semantic technology components. We developed and implemented the semantic methodology throughout the effort and found that it provided consistency as we met user requirements. While individual technology components might change, this did not affect our ability to deliver a capability that was useful and usable, especially by a broad range of people, some of whom will not be trained in all areas of science covered in the collection.

## 3.3 The pre-existing cyberinfrastructure

VSTO was originally intended to replace at least partly the middleware component of an existing set of data systems built at the High Altitude Observatory (HAO). Those data systems accommodated a large number (~15) of distinct data holdings while serving internal and external user groups ranging in size from 2–5 and up to thousands of users. In some cases, the pre-existing systems had evolved over decades (e.g., CEDAR and MLSO; see below) and been updated, especially as new internet-based technologies became available. A limiting characteristic of almost all of the previous implementations is that they required a significant amount of domain knowledge to formulate meaningful and correct queries.

We began with two distinct science communities represented by long-standing project communities: the Coupled Energetics and Dynamics of Atmospheric Regions (CEDAR) and the Advanced Coronal Observing System (ACOS) operated at the Mauna Loa Solar Observatory (MLSO). The CEDAR archive provides an online database of middle and upper atmospheric, geophysical index, and empirical and simulation model data. The ACOS/MLSO archive provides an online database (including many images) of solar atmospheric physics data.

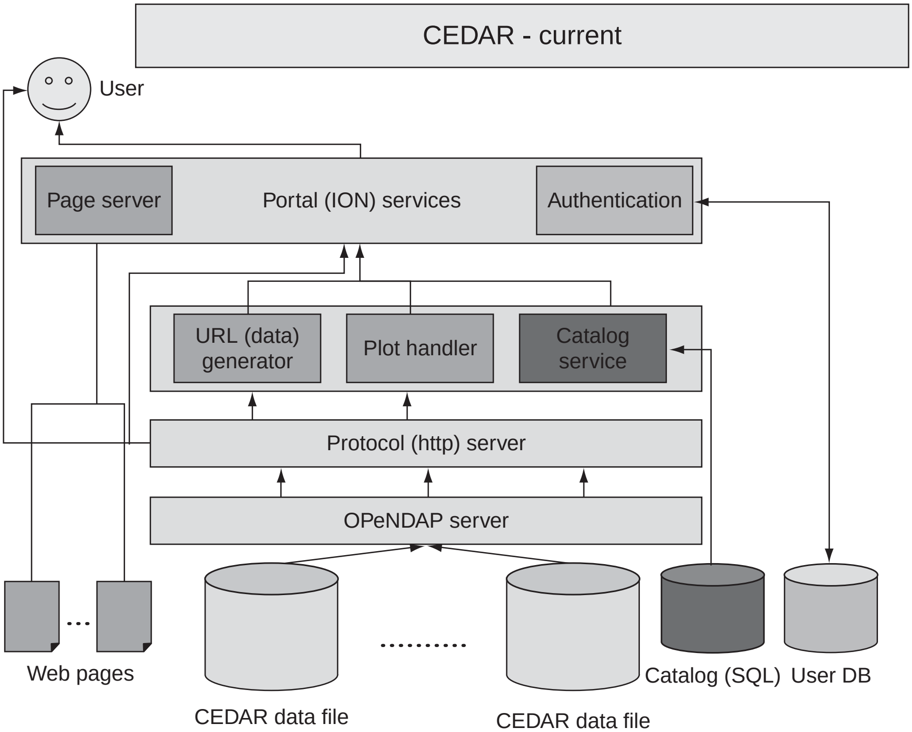



Figure 3.1. CEDARWEB version 3.2 architectural layout indicating use of technologies at the time.

Figure 3.1 shows a schematic of the CEDARWEB architecture around 2005. This version 3.x was built upon the ION Script (IDL On the Net) language, IDL (Interactive Data Language from ITT Visual Systems) with the interface content partly dynamically generated by accessing catalogs stored in a *mySQL* relational database. The data access layer used a customized version of the OPeNDAP server software (Garcia *et al.*, 2008). This version had in turn replaced a two-tier architecture (version 2.x; 1998–2001) that was developed in Perl using the Common Gateway Interface (CGI) methodology popular at the time. Version 1 of the CEDAR database (before it became a web-accessible application) was a single-tier application that ran in full screen VT100 mode in a terminal where users needed logins to the CEDAR computer at NCAR.

The CEDARWEB example of pre-existing capabilities is representative both within the HAO environment but also at other institutions and disciplines. As noted above, new needs were being driven by a more diverse set of users and intentions for the provided data. Our task was then to suitably leverage as much of the existing cyberinfrastructure as made sense, while reducing duplication and maintenance, with the goal of enabling more rapid evolution of capabilities and to incorporate and adapt to new technologies. We proceed to describe this process within the context of the semantic web.

## 3.4 Toward semantic data frameworks – knowledge encodings

Our approach to the vocabulary challenges mentioned previously was to provide a virtual observatory implementation (Szalay, 2001) that incorporated suitable background information about the terms used in the subject matter repositories. The primary difference between extant virtual observatories and what we developed was that we encoded this background information about terms and their meanings in ontologies (Gruber, 2003). We decided what terms to put in the background ontologies by developing and analysing motivating use cases (Cockburn, 2001). The ontologies included terms used in the data collections along with machine parsable formal definitions. Those definitions are then used to enable semantic search and interoperability. The use cases were drawn from the CEDAR and MLSO communities (Fox *et al.*, 2007; McGuinness *et al.* 2007), and these were used to scope the ontologies. The general form is "retrieve data (from appropriate collections) subject to (stated and implicit) constraints and display (plot) in a manner appropriate for the data." One of the very first examples from the CEDAR community was: Retrieve neutral temperature data taken by the Millstone Hill Fabry-Perot interferometer looking in the non-vertical direction from January 2000 to August 2000 and plot as a time series. This query chose neutral temperature as the parameter, Millstone Hill Fabry-Perot interferometer as the instrument, and time series as the data product and places non-vertical direction and a date range as the constraints. While seemingly very specific, due to the modeling approach required for the semantic web, i.e., an object design, abstractions or generalizations appeared very quickly.

Essentially we looked at the variables in the templates above and natural hierarchies in those areas (such as an instrument hierarchy), and important properties (such as instrument settings), and restrictions. We also looked for useful simplifications in areas such as the temporal domain. The CEDAR holdings also relied upon a controlled vocabulary, including terms related to observatories, instruments, operating modes, parameters, observations, etc. MLSO holdings also embodied a controlled vocabulary with significant overlap in concepts but with a very different presentation of data access to an end user from the CEDAR holdings.

Before proceeding to fully develop the ontology model and engineer it, we searched for existing ontologies in earth and space sciences and identified the Semantic Web for Earth and Environmental Terminology (SWEET: Raskin and Pan, 2005) ontology that was gaining acceptance in the earth sciences community, with sufficient overlap with our domains. SWEET version 1.0 itself was intended to be a mid-level ontology for earth and environmental science and covered much more than we needed in breadth, and not enough in depth in multiple places. SWEET used a conceptual decomposition for the ontologies, i.e., faceted or orthogonal ontologies such as Realm, Physical Property, Physical Process, etc., and integrative ontologies such as Human Activities. We also used terms from the ontology as much as possible and added depth in the areas we required.

We focused on domain areas where little or no ontology work had been done, with an eye towards adding the best leverage using semantics. Interestingly, these areas also have proven to be leveragable in applications outside of a solar-terrestrial focus. Our application into the disciplines of volcanic effects on climate has led us to re-use many of the ontology concepts we developed for VSTO (Fox *et al.*, 2007; McGuinness *et al.*, 2007). In developing what became known as the VSTO ontology, the first focus area was instruments. One significant challenge for the integration of scientific data from multiple instruments is in understanding the conditions under which the data were collected or generated. Important metadata may include the geographic location of its observatory installation, its operating modes and settings, stages of calibration, and so on. For the ontology model these concepts are added as properties on classes in the ontology and accurate modeling is needed to ensure that properties are associated on the correct classes at an ontologically consistent place in any class hierarchy. Scientists, or in our case, software that mediates the access and use for any user, that need to interpret data may need to know how an instrument is being used – for example, using an optical instrument such as a spectrometer as a photometer.[2] More specifically among the CEDAR instruments, the Davis Antarctica Spectrometer is a spectrophotometer and thus has the capability to observe data that other photometers may collect. An unfamiliar user would not necessarily know this but the multimodal expressiveness of an ontology language such as OWL make this easy and consistent. A schematic of part of the ontology is given in Figure 3.2.

## 3.5 The VSTO semantic data frameworks – developing the architecture and choosing technology

In the implementation phase, which occurred in prototype form in the first year of the project, we had to make design and architectural choices that suitably leveraged parts of the existing infrastructures. Because we were building semantics into and around the interfaces in the architecture, VSTO depended on a number of components and tools developed for the semantic web, including background ontologies, query languages, triple stores, and reasoners. From a development and maintenance perspective, the

[2] http://en.wikip edia.org/wiki/Spectrophotometry.

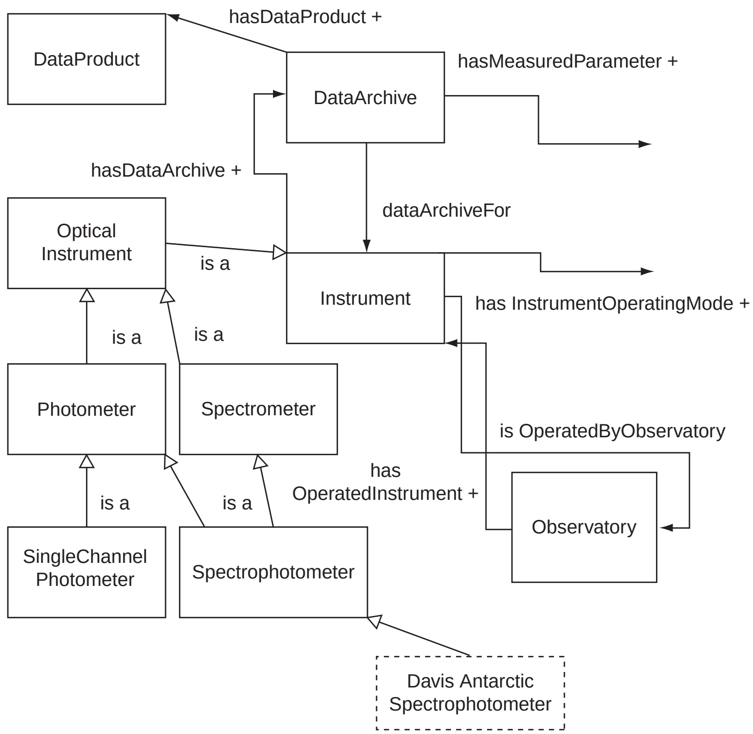



Figure 3.2. VSTO ontology fragment showing the expressiveness that is often required in scientific settings. Here the Spectrophotometer is a subclass of both Photometer and Spectrometer. The instance of such an instrument is shown in the dash box – Davis, Antarctica.

technology tools included ontology editors, validators, and plug-ins for code development. We limited the ontology design to the expressiveness of the middle "species" of OWL – OWL-DL (Description Logics). We did this so that we could leverage freely available reasoners available for OWL-DL, along with their better computational efficiency. Within OWL-DL, we had the expressiveness we needed for the ontology with the following three exceptions: support for numerics (representation and comparison), rules, and default values. The implementation for VSTO does not use default value encoding and the rules and numerical analysis are handled with special purpose query and comparison code (in Java). It turned out that computational efficiency was another factor due to the volume of numerical data, meaning that we needed special purpose handling anyway. Of particular note is the choice we made concerning how the time coverage was represented and implemented. The quantity of "date" data in the repositories we were considering was overwhelming. In other words, hundreds of millions of discrete time records are very common for observational data of this type. Thus, we chose a very simple representation for date-time and did not encode the time instances in OWL-DL. Instead, we supported a hybrid solution whereby a metadata service class was added, so that we had support functions for accessing date and time records directly from original (mySQL) relational catalogs instead of actually retrieving it into some cached or local triple store. Our solution used semantically enhanced web services to retrieve the data directly.

## 3.6 Developing the software and implementation aspects

In the first year, a small, carefully chosen six-person team developed and analyzed the use cases, built the ontologies, designed the architecture, and implemented an alpha release. We had our first users within the first eight months with a small ontology providing access to all of the data resources. Over the next two years, we expanded the ontology,made the system more robust, and increased domain coverage, developing two more prototypes before the current production capability was in place.

Early issues that needed attention in design included determining an appropriate ontology structure and granularity. Our method was to generate iterations initially done by our lead domain scientist and lead knowledge representation expert, vet the design through use case analysis and other subject matter experts, as well as the entire team. We developed minimalist class and property structures capturing all the concepts into classes and subclass hierarchies, only including associations, and class value restrictions needed to support reasoning required for the use cases. This choice was driven by several factors: (a) keeping a simple representation allowed the scientific domain-literate experts to view and vet the ontology easily; (b) complex class and property relations, while clear to a knowledge engineer, take time for a domain expert to comprehend and agree upon. A practical consideration arose from Protégé with automatic generation of a Java class interface and factory classes (see Fox *et al.* 2006 for details).

As we assembled the possible user-query workflows and used the Pellet reasoning engine, we built dependencies on properties and their values. If we had implemented a large number of properties and needed to change them, or moved properties to different class levels – as a result of adding new classes and, thus, "evolving" the ontology – the existing code would have had to be substantially rewritten manually to remove the old dependencies, since an automated refactoring mechanism/tool does not currently exist. Our approach preserved the existing code, automatically generating the new classes, and adding incrementally to the existing code, allowing more rapid development. The deployment cycles and updates to the ontology are thus released with no changes in the existing data framework, thereby benefiting both developers and users. Subsequent releases added new data sources, refinements and corrections to the ontology, and web services access.

We used only open-source free software for the project, which met most of our needs for ontology editing and reasoning. There were a few occasions when the support that is typically provided with commercial software would have been a benefit, especially in the early stages. However, we were able to garner sufficient support on mailing lists, and sometimes through personal communications with software authors/maintainers. Perhaps the most obvious gap in the software tool support, which persists to the current day, is the lack of a robust, industry-strength collaborative ontology evolution and source control system. Initially, the ontology development process was distributed, but collecting and processing the inputs was centralized because our early environment was fragile in terms of building the ontology and then generating robust functional Java code based on that. As the issues concerning the development environment eventually got resolved, our approach also evolved to the point where ontology development and maintenance was distributed, using modularization and social conventions.

We used the SWOOP (www.mindswap.org/2004/SWOOP), Protégé (http://protégé.stanford.edu), and CMap (Cañas *et al.*, 2004) editors for ontology development. The definitions in the ontologies were used to generate Java classes and a Java object model using the Protégé-OWL-Java API (application programming interface) initially for Protégé version 2.2 and later versions 3.3 and 3.4. We also used the Jena API and Protégé-Jena plug-ins for triple store management. We built Java services that use this Java code to access the catalog data services. We used the PELLET (www.clarkparsia.com/pellet) descriptions logic-reasoning engine to compute information that is implied, and also to identify contradictions. The user interface uses the Spring (www.springsource.org) framework for supporting workflow and navigation. The choice of Java and the well-integrated and free tools allowed a rapid prototype, evaluation, redesign, and redeployment cycle that greatly facilitated both implementing the use cases and involving users in the evaluation of the interfaces and the knowledge represented in the ontologies.

Figure 3.3 displays a combined schematic of the integrated semantic data framework implemented for VSTO. In the center are the key abstractions in the query workflow: that of *instrument*, *parameter*, and *date-time*. These are the primary components on both the web portal and the web service interfaces (which was developed later in the project). As input to these concepts – in essence populating the class hierarchies, associated properties and instances – is the VSTO ontology (upper right) along with possible semantic filters.

Figure 3.4 shows an example of the current VSTO web portal, whereby guided workflow selection is made available (see caption for details).

## 3.7 Web services

After the initial web portal implementation, it became clear that other virtual observatories wished to take advantage of VSTO's query and access capabilities. To enable this

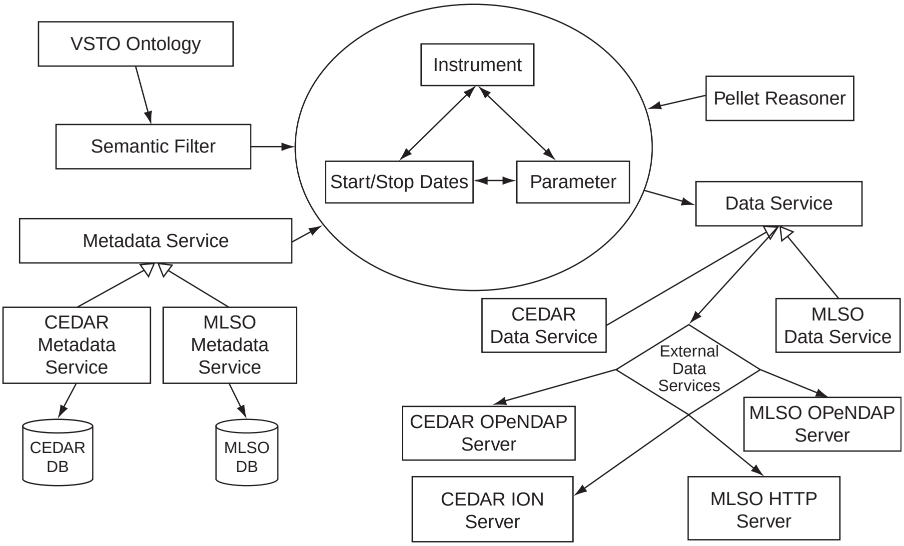



Figure 3.3. VSTO architectural components and query workflow.

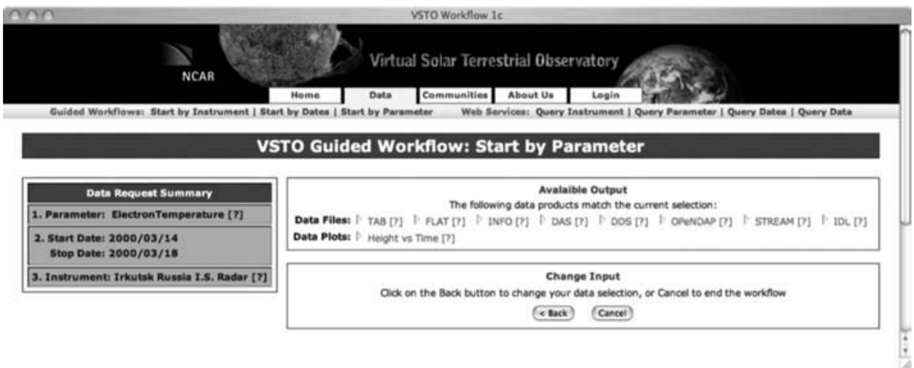



Figure 3.4. VSTO web portal screen shot indicating parameter, date, and instrument selections. The Available Output, including data product and visual data products, inferred from the ontology.

access, web services were an obvious choice, compared to each installation replicating the VSTO infrastructure. To implement the web service access we requested a series of use cases (Fox *et al.*, 2007). Due to the simplification of the data query workflow, the development of web service interfaces naturally followed from the functionality made available in the web portal interface. The three query services are: "by instrument," "by parameter," and "by start/stop date." In addition, the Data Service, which provides access to pointers to the data (in our case OPeNDAP URLs), was made available. The details of the web service are given in Fox *et al.* (2007).

The returned document of a web service call is encoded in OWL-DL using the VSTO ontology, similar to a query invoked from the VSTO web portal. Any consumer of such a service, either another service, or client application, may parse the OWL as XML without semantic meaning or directly (using the background ontology) and use their own reasoning engine (or VSTO's) to further work with the returned information.

The current implementation has two mandatory inputs: instrument and start/stop dates. The remaining choices of parameter class and data product are optional and are typically called in any order depending on the end use. While one typical use of the web services interface is to choose instances for instruments, a user may choose to use a class of instruments allowing more flexibility in the list of options retrieved. For example, a user may choose *OpticalInstrument* for an instrument (instead of choosing a specific optical instrument) and then use the web services to discover the optical instruments or to retrieve data from multiple optical instruments (subject to the other constraints in the query). Thus, the web service interface provides a much greater degree of flexibility for queries. The current portal implementation is presently being updated to include this level of arbitrary use of services, and order/combination of constraints.

## 3.8 Sustaining the developed capability and application extensions

Our ontology was designed to be extensible and, over time, we are finding that the design is indeed holding up to both extension within our project and reuse in other projects. We have investigated the reuse of our ontologies in our Semantically-Enabled Science Data Integration project that addresses virtual observatory needs in the overlapping areas of climate, volcano, and plate tectonics. We found that while seismologists use some instruments that solar-terrestrial physicists do not, the basic properties used to describe the instruments, observatories, and observations are quite similar.

As a result of our successful implementations, we continue to use and promote use-case-based design and extensions. When we plan for extensions, we begin with use cases to identify additional vocabulary and inferences that need to be supported. We have also used standard naming conventions and have maintained as much compatibility as possible with terms in existing controlled vocabularies.

Our approach to distributed multiuser collaboration is a combination of social and technical conventions. This is largely due to the state of the art, where there is no single best multiuser ontology evolution environment. We have one person in charge of all VSTO releases and this person maintains a versioned, stable version at all times. We also maintain an evolving, working version. The ontology is modular so that different team members can work on different pieces of the ontology in parallel.

## 3.9 Benefits

One very important aspect of our developments for VSTO and subsequent projects was to be clear on what the benefits and advantages of semantic web methods and technologies were (and are). Fortunately the benefits were significant (McGuinness *et al.*, 2007) and included the reduction in the number of "clicks" to the data when using the web portal, formalized syntactic and semantic support (eliminating the possibility of obtaining an inconsistent query result), and a broader range of users, especially beyond the immediate field of specialty. We found it extremely helpful to record baseline capabilities represented in the use case, i.e., metrics, where possible before we commenced implementation. In a few cases, at least 50%, the use cases were not implementable with existing (non-semantic) infrastructure without a substantial amount of one-time design and reprogramming. Perhaps one of the more surprising benefits was that we were able to unify the query workflows (by instrument, parameter, date-time) across several discipline data holdings. We had expected that we would build a discipline-specific portal for each discipline/community, much as the non-semantic virtual observatories were being built. However, as noted earlier, an unanticipated effect of the ontology modeling was that generalizations and abstractions of the concepts being modeled, e.g., instrument types, emerged early, but also the property (and inverse property) relations among them. The result was that a query workflow designed to begin a structured query with a complex list of compound concepts (e.g., best, non-vignetted, rectangular coordinate, Mark IV polarization brightness) could be searched using the important terms first with more accuracy and omitting irrelevant choices or delaying them to the last stage of query refinement. When coupled with our experiences in developing semantic web services for VSTO, we concluded that the ability to present facets of search to a user in any number, order or combination would be a powerful next step in demonstrating the benefits of semantic web. Interestingly, it was around this time that the mspace software appeared (Schraefel *et al.*, 2005) and we began to incorporate its concepts into related work such as for the Earth System Grid (Middleton *et al.*, 2006) and our follow-on work noted below.

## 3.10 Summary and ongoing work

The Virtual Solar-Terrestrial Observatory is a production, interdisciplinary virtual observatory. Semantic web methods and technologies were used to quickly design, develop, and deploy this integrated, virtual repository of scientific data in the fields of solar and solar-terrestrial physics. VSTO is being used in ways that the previous individual systems could not be conveniently used. A key aspect of the semantic design for VSTO was the balance between expressivity and implementability, i.e., between the level and depth of knowledge representation and what the current and evolving software and tools could support in a heavily used environment.

We demonstrated that, after a few iterations, we were able to design an extensible, reusable ontology for solar-terrestrial physics, which is compatible with controlled vocabularies that are in use in the most widely used relevant data collections. Potentially much more leveragable, the structure of the ontology was found to be reusable in multiple virtual observatory projects. We also reviewed the ontology with respect to needs for the NSF-funded Geosciences Network (GEON: www.geongrid.org) project, the NASA-funded Semantically-Enabled Science Data Integration (SESDI: http://tw.rpi.edu/portal/SESDI) project, and the NASA-funded Semantic Assistant for Mining (SAM: http://tw.rpi.edu/portal/SAM) project.

Our findings around ontology reuse, when developing the VSTO knowledge encoding, were that the initial SWEET ontology suite was simultaneously much too broad and not deep enough in our subject areas. As a result, we worked with the SWEET author to provide use cases for restructuring and modularizing the packaging (SWEET 2.0: http://sweet.jpl.nasa.gov/2.0). The overall intent was to be able to import only the portions of SWEET needed so that we could add appropriate extensions.

While the VSTO project enjoyed overall success, we encountered numerous challenges that had to be overcome (e.g., Fox *et al*., 2009a). The scope of the ontology is sufficiently broad that it is not possible for any single scientist to have enough depth in the subject matter to provide all of the concepts/content. Thus, the project had to be a collaborative effort, which turned out to be a major contribution to refining the use case and knowledge representation development. Logistically, smaller sets of experts were identified to be the main contributors to particular subject areas. Thus, while an ontology could be created by them, to achieve an extensible, evolving, widely reusable ontology, it was necessary to obtain broad community buy-in, including vetting and augmentation by the larger scientific community. Partly from this project, and from experience that we noted in related fields, we found that to maximize benefit the ontology needs usage from the broad community and also multiple publication venues.

Our initial implementation used fairly limited inference and supported somewhat modest use cases. This was intentional, as we wanted an initial implementation that was simple enough to be usable by the broad community with minimum training but that demonstrated the benefits of a semantic web approach. After several evaluations and architecture and ontology redesign, it was clear that additional inferential and query support were desirable. As VSTO evolved, we added those additional capabilities based on use cases and demonstrated or measureable benefits to users.

Our follow-up on initial informal evaluations in a workshop setting provided both general and specific answers and comments, as well as more quantitative yes/no or multiple-choice answers. These results reaffirmed the sense we obtained in the initial study that our efforts in applying semantic technologies led to an interdisciplinary virtual observatory that provides significant additional value for a spectrum of end users. It also provides significant additional value for the developers of both the VSTO and other federated VOs and data systems wishing to take advantage of the services that our VO provides. The implementation of newer use cases (e.g., for script/programming language access, synthesizing models and observations and new plotting options) continues to drive our knowledge representation and reasoning requirements.

We also commenced work on transparency and provenance within the VSTO framework in another NSF/OCI funded project entitled Semantic Provenance Capture in Data Ingest Systems (SPCDIS: Fox *et al.*, 2008; Fox *et al.*, 2009b). SPCDIS is leveraging the Proof Markup Language (Pinheiro da Silva *et al.*, 2006) – an Interlingua for representing provenance, information. In this extension, we capture content such as where the data came from. Once captured in PML, the Inference web toolkit (McGuinness *et al.*, 2004) is used to display information about why an answer was generated, where it came from, and how much the information

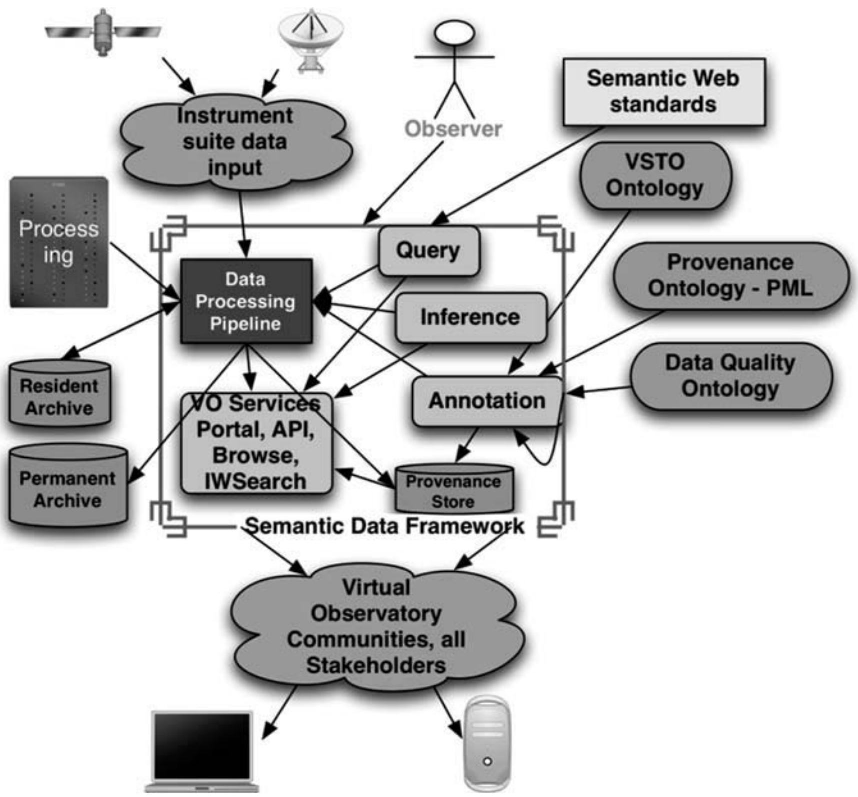



Figure 3.5. Augmented VSTO architecture with knowledge provenance.

might be believed and why. Figure 3.5 shows a schematic on the evolution of semantic capabilities.

Finally, as a result of the aggregate experience with VSTO and the follow-on projects on data integration and provenance, we have commenced a new effort also funded by the NSF to bring all of the developed semantic application capabilities (and more) into a toolkit form to support and advance science. The project, the Semantic eScience Framework (SESF: http://tw.rpi.edu/portal/SESF) strongly builds on our methods and technology developments over the last five years. The intent is to provide a configurable semantic data framework that is deployable in many disciplines, and especially for nonspecialist use. We look forward to reporting on the outcomes of this work in the future.

## Acknowledgements

The VSTO project was funded by the National Science Foundation, Office of Cyber Infrastructure under the SEI+II program, grant number 0431153. We thank David Fulker and Robert Raskin for stimulating conversations related to engaging community and leveraging existing ontology and CI resources. We also thank the editors of this volume for the invitation to present the results of our work.

## References

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Weitzner, J. (2006). Enhanced: Creating a science of the Web. *Science*, **313**(5788): 769–771, doi:10.1126/science.1126902.

Cañas, A. J., Hill, G., Carff, R. *et al*. (2004). CMAPTOOLS: A knowledge modeling and sharing environment. Concept Maps: Theory, Methodology, Technology. *Proceedings of the First International Conference on Concept Mapping*, ed. A. J. Cañas, J. D. Novak and F. M. González, Pamplona, Spain.

Cockburn, A. (2001). *Writing Effective Use Cases*. London: Addison-Wesley Professional, 270pp.

Dalton, R. (2007). Geophysicists combine forces. *Nature*, **447**(7148): 1037.

De Roure, D., Jennings, N. R., and Shadbolt, N. R. (2005). The semantic grid: Past, present, and future. *Proceedings of the IEEE*, Vol. 93, Issue 3, pp. 669–681, doi:10.1109/JPROC.2004.842781.

Fox, P., McGuinness, D. L., Middleton, D. *et al*. (2006). Semantically-enabled large-scale science data repositories. *Proceedings of the 5th International Semantic Web Conference (ISWC06)*, ed. I. F. Cruz *et al*. LNCS 4273. Berlin: Springer, pp. 792–805.

Fox, P., McGuinness, D. L., Cinquini, L. *et al*. (2007). Use cases for the Virtual Solar Terrestrial Observatory. *Proceedings of the Electronic Geophysical Year, Virtual Observatories in Geosciences (VOiG)*, June 2007.

Fox, P., Cinquini, L., McGuinness, D. *et al*. (2008). Semantic web services for interdisciplinary scientific data query and retrieval. *Proceedings of the Association for the Advancement of Artificial Intelligence*, Semantic e-Science workshop.

Fox, P., McGuinness, D. L., Cinquini, L. *et al*. (2009a). Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory Experience. *Computers & Geosciences*, **35**(4): 724–738.

Fox, P., McGuinness, D. L., Pinheiro da Silva, P. *et al*. (2009b). Knowledge provenance in science data pipelines; languages, tools and artifacts: What can we now answer? *Eos Transactions AGU*, **90**(22), Jt. Assem. Suppl., Abstract IN24A-06.

Garcia, J., Fox, P., West, P. *et al*. (2008). Developing service-oriented applications in a grid environment: Experiences using the OPeNDAP back-end server. *Earth Science Informatics*, **2**(1–2): 133–139, doi:10.1007/s12145–008–0017–0.

Gil, Y., Ratnakar, V., and Deelman, E. (2006) Metadata catalogs with semantic representations. In *International Provenance and Annotation Workshop 2006 (IPAW2006)*, Chicago, IL, USA, ed. L. Moreau and I. Foster. LNCS 4145, Berlin: **Springer**, pp. 90–100.

Gruber, T. (2003). It Is What It Does: The Pragmatics of Ontology. Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee, Smithsonian Museum, Washington, D.C., March 26, 2003.

King, T., Walker, R., Merka, J., Joy, S., and Narock, T. (2007). The architecture of a multi-tiered virtual observatory. *Earth Science Informatics*, **1**(1): 21–28, doi:10.1007/s12145–008–0006–3.

Ludaescher, B., Altintas, I., Berkley, C. *et al*. (2006). Scientific workflow management and the Kepler System. *Concurrency and Computation: Practice & Experience*, **18**(10): 1039–1065.

McGuinness, D. L., and van Harmelen, F. (2004). OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. www.w3.org/TR/owl-features/.

McGuinness, D. L., Fox, P., Cinquini, C. *et al*. (2007). The Virtual Solar-Terrestrial Observatory: A deployed semantic web application case study for scientific research. *Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07)*, Vancouver, British Columbia, Canada, July 22–26, 2007.

Middleton, D. L., Bernholdt, D. E., Brown, D. *et al*. (2006). Enabling worldwide access to climate simulation data: The Earth System Grid (ESG). In SciDAC 2006, Scientific Discovery through Advanced Computing, 25–29 June 2006, Denver, CO, USA, ed. William M. Tang. Journal of Physics: *Conference Series*, **46**: 510–514.

Pinheiro da Silva, P., McGuinness, D. L., and Fikes, R. (2006). A Proof Markup Language for semantic web services. *Information Systems*, **31**(4–5): 381–395, June–July 2006. Previous version, KSL Tech Report KSL-04–01.

Raskin, R. G., and Pan, M. J. (2005). Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, **31**(1): 119–125.

Rushing, J., Ramachandran, R., Nair, U. *et al*. (2005). ADaM: A data mining toolkit for scientists and engineers. *Computers & Geosciences*, **31**: 607–618.

Schraefel, M. C., Smith, D. A., Owens, A. *et al*. (2005). The evolving mSpace platform: Leveraging the semantic web on the trail of the Memex. *Hypertext (HT05)*, 174–183.

Szalay, A. S. (2001). The National Virtual Observatory. In *Astronomical Data Analysis Software and Systems*, ed. F. R. Harnden, Jr., F. A. Primini and H. E. Payne. ASP Conference Series, Vol. 238. San Francisco: ASP, p. 3.

# 4

# Cyberinfrastructures for life sciences and biomedicine

SRIRAM KRISHNAN AND RAVI MADDURI

## 4.1 Introduction

Grid computing is defined as the ability, using a set of open standards and protocols, to gain access to applications and data, processing power, storage capacity, and a vast array of other computing resources over the Internet. Under the umbrella of Grid computing, mechanisms are provided for single sign-on, job submission, and data transfer, in order to allow the coupling of distributed resources in a seamless manner. The concept of a "cyberinfrastructure" encompasses advanced scientific computing as well as a more comprehensive infrastructure for research and education based upon distributed, federated networks of computers, information resources, online instruments, and human interfaces (Atkins *et al.*, 2003). Cyberinfrastructures represent the evolution of high-performance and Grid computing, making these technologies truly usable by all the nation's scientists, engineers, scholars, and citizens. To achieve the goal of increasing research productivity and effectiveness of education, the cyberinfrastructures must provide effective tools to end users for online collaborations, access to computing resources and ability to launch computational tasks, and sharing of data and other resources with others in a given community. Over the past decade, an extensive amount of time and effort has been invested in building such cyberinfrastructures for various scientific and engineering communities, such as:

- Life sciences: the National Biomedical Computation Resource (NBCR), the Biomedical Informatics Research Network (BIRN), the cancer Biomedical Informatics Grid (caBIG)
- Geosciences: the Geosciences Network (GEON)
- Metagenomics: the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA);
- Earthquake engineering: the Network for Earthquake Engineering Simulation (NEES)
- Meteorological research: the Linked Environments for Atmospheric Discovery (LEAD) project

One of the key challenges of building cyberinfrastructures is how to expose all the functionality supported at the back-end to the scientific end users in an easy-to-use manner, in such a way that it is easy to implement, maintain, and manage given the dynamic nature of scientific research. This can be broken up into two distinct, important issues. First, from the perspective of an end user, the cyberinfrastructure should provide an intuitive user interface for leveraging all the applications of interest, preferably from a single point of entry, and without the need to install complex software tools. Second, from the perspective of a software architect or a developer, the cyberinfrastructure should allow the addition of new applications, data sources, and collaboration tools, without having to make large-scale changes to the software stack.

Over the past several years, with the advent of the Open Grid Services Architecture (OGSA) (Foster *et al.*, 2002) and the Web Services Resource Framework (WSRF: www.globus.org/wsrf), service-oriented architectures (SOA) and web service technologies have been embraced in the field of scientific and Grid computing. Services reduce complexity by encapsulating the service implementation, and providing simple well-defined interfaces. With the help of open standards, they ensure interoperability across systems and architectures, i.e., a client from one system can easily access a service implemented within another system, as long as it understands the protocols and the interfaces used by the service implementation.

In the past, SOAs have only been used to build Grid middleware. For instance, the Globus Toolkit (Foster and Kesselman, 1997) has been redesigned to be consistent with latest web service technologies. This is great for middleware developers who can use the platform-independent WSRF-based APIs to access standard Grid functionality, such as job submission and data transfer. However, developers of scientific tools are not experts in Grid technologies, irrespective of the simpler and more consistent APIs being provided. From their perspective, the services that are most relevant are those that perform a scientific operation, where the semantics of the operations are defined in terms of the domain science. For example, a Blast service can provide biologists with the capability to compare multiple DNA sequences against each other or against standard, publicly available, datasets. The particular infrastructure used for its calculation is irrelevant. The tool developers would rather focus on the algorithms and the appropriate interfaces to present to the scientific end users. More recently, Grid technologies have been envisioned for accelerating the development and adoption of "service-oriented science" by supporting a separation of concerns between discipline-specific content and domain-independent software and hardware infrastructure (Foster, 2005). Enabling access to scientific applications and tools using a service-oriented approach allows the developers of scientific tools to focus on the functionality of the tool itself, while delegating the management of the complex back-end resources to others who are more proficient in Grid middleware.

In this chapter, we focus on cyberinfrastructures that enable service-oriented science in the areas of life sciences and biomedicine. We provide some of the scientific background, followed by a detailed description of the middleware. In particular, we discuss the cyberinfrastructure for the National Biomedical Computation Resource (NBCR) and the Cancer Bioinformatics Grid (caBIG$^{TM}$) in Sections 4.2 and 4.3, respectively, followed by a summary in Section 4.4.

## 4.2 The National Biomedical Computation Resource (NBCR)

### *4.2.1 Goals*

The goals of the NBCR cyberinfrastructure are twofold. Firstly, to provide transparent access to the emerging Grid-based computational infrastructure by "grid-enabling" biomedical codes and providing access to distributed biological and biomedical databases; and secondly, to enable integration of applications across different scales, e.g., atomic to macro-molecular, molecular to cellular, tissue to organ. For instance, a user may wish to utilize the highly accurate quantum models to calculate atomic charges, but opt for the less accurate but significantly more computationally tractable classical molecular dynamics approach for determining protein-ligand docking. Independent developer communities have developed these capabilities over long periods of time. The goal of NBCR is to provide a mechanism for scientific users to discover and leverage these capabilities, in a standard easy-to-user manner, possibly with the use of freely available software tools.

### *4.2.2 Architecture overview*

Figure 4.1 shows the multitiered architecture for enabling easy access to scientific applications on Grid resources. The bottom tier consists of the Grid resources where the jobs are scheduled. Accessing these resources via Grid schedulers can be quite complicated for the end user. The end user would have to learn about mechanisms and tools to manage Grid credentials using the Grid Security Infrastructure (GSI) (Foster *et al.*, 1998), access Grid schedulers such as Condor and the Sun Grid Engine (SGE), and manage data transfers using tools such as GridFTP (Allcock *et al.*, 2005). Furthermore, installation and deployment of scientific software on the heterogeneous Grid resources is no easy task in itself, and end users may spend countless hours compiling and installing their complex scientific tools. Hence, in the middle tier, programmatic access to scientific applications and security services running on Grid resources is enabled via simple web service wrappers. The web service wrappers address many of the problems described above. In particular, scientific applications are installed once, and are available to a set of authorized users. Users are shielded from the complexities of Grid schedulers and data management software.

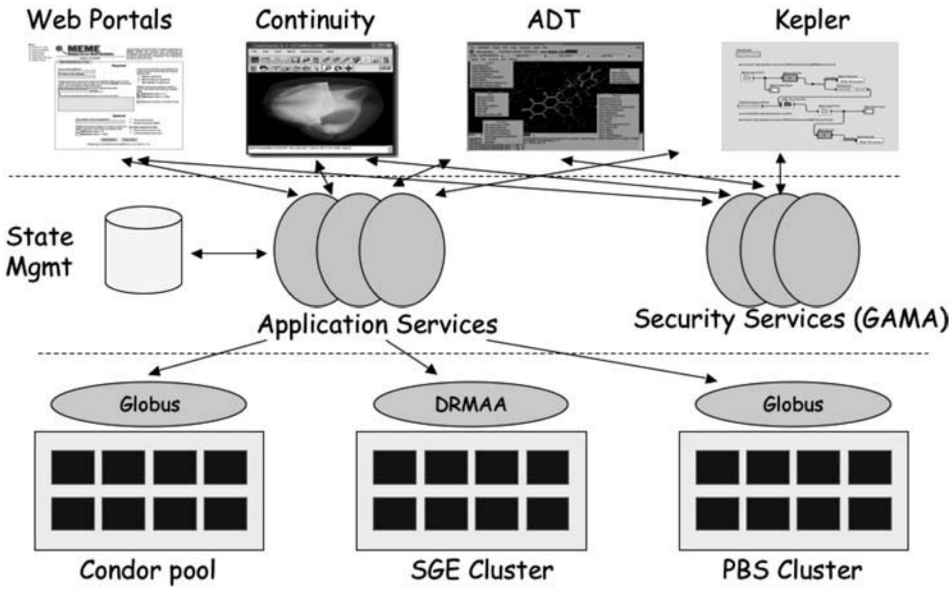



Figure 4.1. NBCR SOA architecture. See color plates section.

Internally, the services use APIs such as the Globus GRAM (Czajkowski *et al.*, 1998) and the Distributed Resource Management Application API (DRMAA: www.drmaa.org) to leverage Grid schedulers; however, this is completely transparent to the end user. The Grid Account Management Architecture (GAMA: Bhatia *et al.*, 2005), as described in Section 4.2.3, simplifies access to Grid security mechanisms.

Various client tools, available in the top tier, leverage the easy-to-use web services APIs to securely access these applications running on the complex back-end infrastructures. Using the clients, the scientific end users are completely shielded from the complexity of the back-end resources, and even from the simpler web services APIs. In fact, they are mostly oblivious to where applications are being run, and interact with the applications as if they are running locally. In Section 4.2.3, we present a list of available client interfaces.

The current set of deployed applications services in the middle tier consists of key scientific software plus numerous secondary resources and utilities, and includes a combination of handwritten and autogenerated services based on the Opal toolkit (Krishnan *et al.*, 2005). Some of the important available scientific services include those for GAMESS, a general atomic molecular electronic structure package which uses general *ab initio* quantum chemistry techniques; APBS, a classical modeling package for electrostatics computations, important in several areas of biomolecular simulation; PDB2PQR, a utility for converting protein files from the Protein Data Bank (PDB) format to PQR format used by applications such as APBS; AutoDock and AutoGrid, a suite of automated molecular docking software tools used in applications such as X-ray crystallography and structure-based drug design; MEME and MAST, software tools for discovery and search of motifs (highly conserved regions) in groups of related DNA or protein sequences; and Continuity, a toolkit for multi-scale modeling in bioengineering and physiology, especially cardiac biomechanics, transport, and electrophysiology.

### 4.2.3 Software tools

Apart from the application-specific services for the NBCR SOA, generic tools to help build SOAs for other scientific communities have also been built. We discuss some of these tools in this subsection.

*The Opal toolkit*

The Opal toolkit enables application providers to automatically wrap legacy scientific applications as web services. Application providers configure the web service wrapper using configuration files, and the Opal toolkit automatically deploys the application as a web service fully integrated with the local cluster scheduler and with GSI-based authentication (if desired). This enables the rapid deployment of legacy applications as web services on the Grid without having to write a single line of source code. The 1.0 version of the Opal toolkit was released in September 2006, and has been used effectively by a number of projects in the Grid community apart from NBCR, e.g., the CAMERA project sponsored by the Moore Foundation, which addresses marine microbial genomics (CAMERA, 2010). The current stable version of the Opal toolkit is 1.9.4, and is available via SourceForge (OPAL, 2010). The latest version of Opal includes new features such as a Dashboard, which provides usage statistics for all the deployed services, and automatically generated web interfaces for services that provide optional metadata for command-line arguments.

*GAMA security*

The Grid Account Management Architecture (GAMA) provides the security infrastructure for the SOA. In general, Grid systems use a GSI-based security mechanism to administer resources. However, they are known to be difficult for administrators to deploy, and end users to use. GAMA provides a simple mechanism for administrators to deploy the security services in the form of Rocks Rolls (Papadopoulos *et al.*, 2001), and a web services API for user interfaces (on various platforms) to interact with. Users have the option of retrieving the Grid credentials from the GAMA web services for use in a stand-alone mode. However, in most cases, lightweight GAMA clients are incorporated into various clients shown in the top tier of Figure 4.1.

### Client interfaces

No single user interface typically suffices for all types of users. With that in mind, our architecture does not impose a single user interface on the end users. Some of the user interfaces supported are as follows:

(1) Gemstone, a rich client application (http://gemstone.mozdev.org) built on top of the Mozilla Firefox platform. Gemstone runs on the user's desktop, and provides an interface for discovering and accessing remote application web services.
(2) MGLTools, a suite of Python tools developed at The Scripps Research Institute (TSRI) for visualization and analysis of molecular structures. This includes the Python Molecular Viewer (PMV) for displaying molecular surfaces and advanced volume rendering; AutoDockTools (ADT), which is a graphical front-end for setting up and running AutoGrid and AutoDock; and the accompanying visual programming tool called Vision for the creation of scientific workflows for tasks such as molecular dynamics simulations and energy minimization calculations.
(3) Kepler (Ludäscher *et al.*, 2006), which is a scientific workflow toolkit being developed at the University of California, San Diego (UCSD), and presented here in Chapter 15. Kepler provides a visual interface to create a scientific pipeline, and provides a standard set of tools to interact seamlessly with emerging Grid-based resources.
(4) GridSphere (http://www.gridsphere.org), which provides an open-source portlet-based web portal framework, also being developed at UCSD. The portlet model gives users a flexible easy-to-use interface, and it gives portal developers a model to create pluggable and dynamic application support.
(5) Continuity, which is a problem-solving environment for multi-scale modeling in bioengineering and physiology. Continuity consists of a GUI client, which runs on a user's desktop, and a set of computational components, which may run locally or on remote clusters via web services.

## 4.3 Cancer Biomedical Informatics Grid (caGrid)

### *4.3.1 Goals*

The Cancer Biomedial Informatics Grid, or *caGrid*, is the service-based Grid software infrastructure of the National Cancer Institute (NCI) cancer Biomedical Informatics Grid (caBIG™) program. The caBIG™ program was initiated in 2004 with participants from cancer centers and research institutions. The overarching goal of caBIG™ is to improve cancer research and accelerate the application of research results towards curing cancer by creating a network of cancer centers, research institutions, and investigator laboratories. The program was established in response to the increasing and challenging informatics needs of the cancer research community. These needs include support for integration of information from diverse data sources and data types, for coordinated and secure management of information across multiple institutions (e.g., multi-institutional clinical trials), and for analysis of large volumes of data. They present challenging problems because of the paucity of interoperable applications and databases, and mechanisms for efficient and secure access to resources at disparate locations. In order to address these challenges, the caBIG™ program has been developing common applications and tools, information standards, data and analytical resources, and the caGrid software infrastructure, which will link the applications and resources and facilitate sharing of information within the guidelines and policies accepted by the caBIG™ community.

### *4.3.2 Architecture overview*

We describe caGrid version 1.0 (caGrid 1.0), which was released to caBIG™ participants and the research community at large in December 2006. The first version of caGrid was version 0.5, released to the caBIG™ community in September 2005. It provided a reference implementation of the basic Grid architecture of the caBIG™ program and an initial suite of tools and services. It was intended as a test-bed infrastructure for the caBIG™ community to evaluate the architecture and provide feedback and additional requirements through implementation of several reference applications. caGrid 1.0 represents a significant improvement over caGrid 0.5, with new features and enhancements based on the feedback from early adopters of caGrid 0.5 and on additional requirements collected from the caBIG™ community. It is being used as the production Grid environment in caBIG™. It provides a set of core services, toolkits for the development and deployment of community-provided services, and APIs for building client applications.

caGrid 1.0 adds several important features on top of the basic Grid services architecture to better address the informatics needs of cancer research:

(1) It is built on a Model Driven Architecture (MDA) best practice, with emphasis on syntactic and semantic interoperability across heterogeneous resources. Interoperability is achieved through use of common data elements, published information models, controlled vocabularies, and well-defined application programming interfaces. Client and service APIs in caGrid represent an object-oriented view of data and analytical resources, and operate on data types that are based on common data elements and controlled vocabularies. caGrid identifies two kinds of services – Data services and Analytical services. Data services are Grid services providing access to data in the form of common data elements to clients. Analytical services make available to clients different analytical routines in the form of Grid services.

(2) It facilitates metadata-driven discovery of resources by taking advantage of structural and semantic descriptions of data models and services. It provides middleware infrastructure and client APIs to create, manage, and perform searchers on rich, standardized, service metadata. In caGrid 1.0, a client can discover resources based on data models and semantic information associated with them.
(3) It provides a comprehensive security system, referred to as the Grid Authentication and Authorization with Reliably Distributed Services (GAARDS) infrastructure (Langella, 2008). The GAARDS infrastructure includes Grid-wide identity management support, support for management of Grid-wide trust fabric, and Grid Grouper for virtual organization and Grid-wide group management support.
(4) It provides a unified service-authoring toolkit, called Introduce (http://dev.globus.org/wiki/Incubator/Introduce), for service providers to easily develop and deploy caGrid compliant data and analytical services. Introduce is an extensible workbench with graphical user interfaces and high-level functions that abstract away the complexities of Grid technologies. It capitalizes on the rich collection of metadata that caGrid makes available, by providing such features as data type discovery to facilitate reuse and interoperability, and automatic service metadata generation.
(5) It enables federated access to distributed data sources via a common data service and federated query infrastructure. The federated query support enables distributed aggregations and joins over multiple data services.
(6) It implements service support for orchestration of Grid services using the industry standard Business Process Execution Language (BPEL). This is represented as part of the Coordination Services, especially the workflow services. caGrid 1.0 provides a workflow management service, enabling the execution and monitoring of BPEL-defined workflows in a secure Grid environment. The future versions will also support orchestration of Grid services via the Taverna workbench (Oinn *et al.*, 2004) from the myGrid team in the UK, which is an alternative solution for managing scientific workflows compared with, for example, the Kepler system mentioned earlier, and discussed in Chapter 15. The higher-level services like Federated Query Processing (FQP) and the workflow services orchestrate data and analytical services in meaningful patterns.
(7) It provides support for bulk data transport to allow uniform access to large datasets from caGrid compliant services. This work currently supports access via the existing specifications of WS-Enumeration, WS-Transfer, and GridFTP.

The design of caGrid is motivated primarily by use cases from the cancer research community. However, the architecture of caGrid and many of the common functions provided by the system are generic enough to be applicable in other biomedical research domains, and other distributed computing environments. In fact, several of

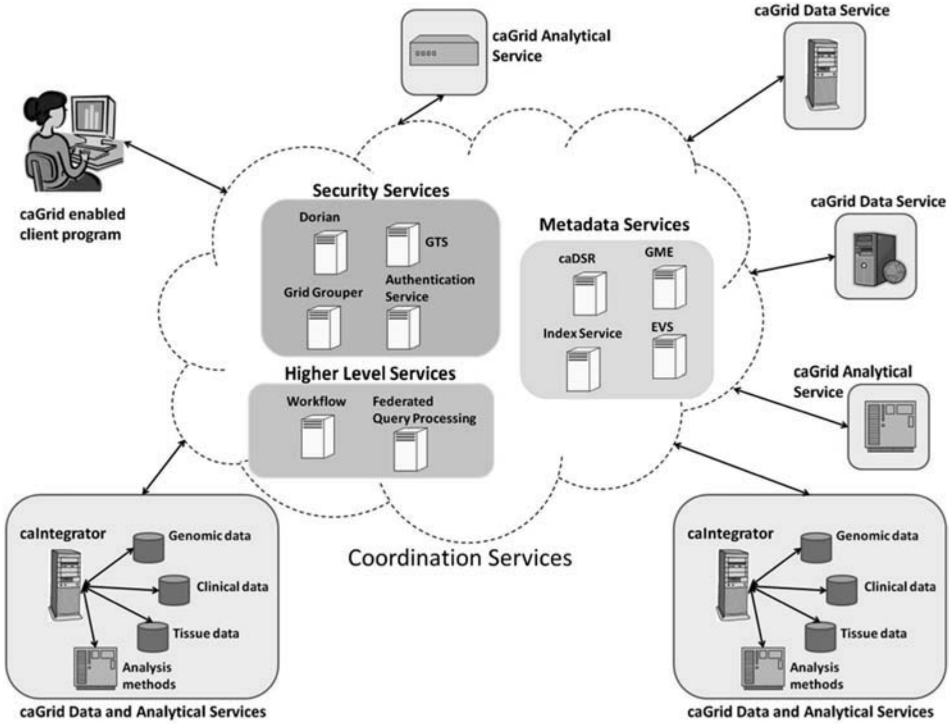



Figure 4.2. caGrid architecture.

the components of caGrid have been contributed back to Globus for future incorporation into the toolkit. caGrid is not only built on open-source components, it is itself made publicly and freely available under a liberal open-source license for use both in and outside of caBIG$^{TM}$. Additional information and downloads of the software can be found at the project web site (http://cabig.nci.nih.gov).

## 4.4 Conclusions

Science and engineering communities are becoming increasingly dependent upon cyberinfrastructures for their research and education. We strongly believe that service-oriented architectures (SOA) are the future of all cyberinfrastructures, due to flexibility, ease of implementation and use, and the ability to leverage a diverse set of information sources at the back-end. In this chapter, we discussed the SOA-based cyberinfrastructures for two projects in the areas of the life sciences and biomedicine. We believe that our experiences in building cyberinfrastructures for these areas would prove equally useful to other fields as well – and, in particular, to the field of geoinformatics, to which this book caters.

## Acknowledgements

NBCR is supported by the National Institutes of Health (NIH) through a National Center for Research Resources program grant (P41 RR 08605). The caGrid has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400. The authors would like to thank Peter Arzberger, Wilfred Li, and the rest of the NBCR team at the University of California, San Diego, The Scripps Research Institute (TSRI), and Washington University in St. Louis (WUSTL); and Ian Foster, Joel Saltz, and the caGrid team at Ohio State University and the University of Chicago.

## References

Allcock, W. *et al*. (2005). The Globus Striped GridFTP Framework and Server, Super Computing Conference.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I. *et al*. (2003). Revolutionizing Science and Engineering through Cyberinfrastructure. Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure, National Science Foundation, 84pp.

Bhatia, K., Chandra, S., and Mueller, K. (2005). *GAMA: Grid Account Management Architecture*. 1st IEEE International Conference on e-Science and Grid Computing, pp. 420–428, doi:10.1109/E-SCIENCE.2005.45.

CAMERA: Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (2010). http://camera.calit2.net.

Czajkowski, K. *et al*. (1998). A Resource Management Architecture for Metacomputing Systems, IPPS/SPDP 98, Workshop on Job Scheduling Strategies for Parallel Processing.

Foster, I. (2005). Service-oriented science. *Science*, **308**(5723): 814–817.

Foster, I. and Kesselman, C. (1997). Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications*, **11**: 115–128.

Foster, I., Kesselman, C., Nick, J., and Tuecke, S. (2002). Grid services for distributed system integration. *Computer*, **35**(6).

Foster, I., Kesselman, C., Tsudik, G., and Tuecke, S. (1998). *A Security Architecture for Computational Grids*, ACM Conference on Computers and Security.

Krishnan, S., Steam, B., Bhatia, K. *et al*. (2005). Opal: Simple web services wrappers for scientific applications. In *Proceedings of Grid 2005, 6th IEEE/ACM International Workshop on Grid Computing*, 8pp.

Langella, S. *et al*. (2008). Sharing data and analytical resources securely in a biomedical research Grid environment. *Journal of the American Medical Informatics Association*, **15**(3): 363–373.

Ludäscher, B. *et al*. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice & Experience*, **18**(10): 1039–1065.

Oinn, T. *et al*. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**(17): 3045–3054.

OPAL (2010). OPAL Toolkit. http://opaltoolkit.sourceforge.net.

Papadopoulos, P., Katz, K., and Bruno, G. (2001). NPACI rocks: Tools and techniques for easily deploying manageable Linux clusters, *Concurrency and Computation: Practice & Experience*, Special Issue.

# Part II

## Modeling software and community codes

# 5

# Development, verification, and maintenance of computational software in geodynamics

MICHAEL GURNIS, WALTER LANDRY, EH TAN, LUIS ARMENDARIZ, LEIF STRAND, AND MICHAEL AIVAZIS

## 5.1 Introduction

Research on dynamical processes within the Earth and planets increasingly relies upon sophisticated, large-scale computational models. Improved understanding of fundamental physical processes such as mantle convection and the geodynamo, magma dynamics, crustal and lithospheric deformation, earthquake nucleation, and seismic wave propagation, are heavily dependent upon better numerical modeling. Surprisingly, the rate-limiting factor for progress in these areas is not just computing hardware, as was once the case. Rather, advances in software are not keeping pace with the recent improvements in hardware. Modeling tools in geophysics are usually developed and maintained by individual scientists, or by small groups. But it is difficult for any individual, or even a small group, to keep up with sweeping advances in computing hardware, parallel processing software, and numerical modeling methodology.

We will focus on the challenges faced by computational geophysics and the response of a community initiative in the United States called the Computational Infrastructure for Geodynamics (CIG). Instead of reviewing all of the activities CIG has been involved with, we will focus on just a few so as to describe the multiple ways that a virtual organization developed and used software within the rapidly evolving backdrop of computational science. We will focus on the scientific topics of mantle convection, tectonics, and computational seismology, although CIG has also been deeply involved with magma dynamics and the geodynamo.

Mantle convection is at the heart of understanding how the Earth works, but the process remains poorly understood at best. Progress on fundamental questions, such as the dynamic origin of tectonic plates, layering and stratification within the mantle, geochemical reservoirs, the thermal history of the Earth, the interpretation of tomography, and the source of volcanic hotspots, all require an interdisciplinary approach. Numerical models of mantle convection must therefore assimilate information from a wide range of disciplines, including seismology, geochemistry, mineral and rock physics, geodesy, and tectonics.

The technical challenges associated with modeling mantle convection are substantial (Moresi *et al*., 2000). Mantle convection is characterized by strongly variable (i.e., stress-, temperature-, and pressure-dependent) viscosities (Schubert *et al*., 2001). In addition, the mantle is chemically heterogeneous, is replete with silicate melts and volatiles, and has numerous pressure- and temperature-induced structural (phase) changes that affect its dynamics (Davies, 1999). Equally importantly, we have learned that it is not practical to model mantle convection using a wholly theoretical or "first-principles" approach, even if we had the hardware resources at our disposal. Two decades of experience have taught us that "first-principle" mantle convection models cannot replicate the unique history of the Earth, starting from arbitrary initial conditions (Bercovici *et al*., 2000). Accordingly, mantle convection models need to be more closely integrated into the "real world" through direct data assimilation and by direct testing against a variety of observations (Davies, 1999).

The lithosphere, with the embedded crust, represents the main thermal boundary layer of the Earth's heat engine and, as such, encompasses a wide range of pressure and temperature conditions with a diversity of deformational mechanisms (Ranalli, 1995). On million-year timescales, the crust is a storehouse of observational constraints drawn from structural geology, tectonics, metamorphic petrology, stratigraphy, etc., that have been acquired for more than a century. Recently, deep seismic profiling, receiver-function analysis, and magnetotelluric sounding have greatly increased our understanding of crustal and lithospheric structure, and the USArray component of EarthScope is flooding us with new observations. Numerical modeling has become an essential step in the integration of these data into process-orientated models of orogenesis, lithospheric stretching, sedimentary basin genesis, and plate boundary deformation (Karner *et al*., 2004).

A rapidly advancing area of crustal geodynamics, one of great societal importance, is the problem of the physics of the earthquake cycle. Because of the recent development of the capability for high-accuracy geodetic monitoring of crustal deformation in real time using GPS and InSAR, this field, long starved for data, is now a burgeoning observational science. Recent observations made with high-precision space geodesy indicate that displacements caused by slow aseismic motions following earthquakes can be comparable to co-seismic displacements, demonstrating substantial post-seismic evolution of strain and stress in addition to co-seismic changes. Existing GPS arrays in the Pacific Northwest, southern California, Japan, and Taiwan, and the Plate Boundary Observatory component of EarthScope, are providing heretofore-unavailable observational constraints on the stress and strain changes through the earthquake cycle.

Deformation of the lithosphere presents a number of challenges to numerical simulation. The lithospheric mantle encompasses a differential temperature of up to 1400 °C and an effective viscosity contrast of many orders of magnitude. The continental crust exhibits compositional heterogeneity and low-temperature domains dominated by frictional plastic deformation. The complex physics of frictional materials is particularly challenging because it involves strain-localization, time- and rate-dependent yield strength, and strain softening. Faults are ephemeral on a million-year timescale and thus form or vanish according to the evolution of stress and strain. Crustal deformation is a free-surface problem and sensitive to the complexities of the Earth's surface, including physical and chemical erosion, mass transport by rivers and ocean currents, and deposition of sediment. Ultimately, assimilation of data such as the state of stress in the crust or fault rupture histories from paleoseismic data will provide important constraints on the dynamics of the system. Thus, modeling fault dynamics promises new breakthroughs in earthquake hazard mitigation by uniting various earthquake observations (data assimilation) into comprehensive treatments at the regional (fault zone) level. This integration must cross timescales of seconds to weeks (fault afterslip) to years (transient and steady viscoelastic relaxation) to geologic time (development of fault systems), as well as spatial scales of hundreds of kilometers (fault systems) to centimeters (fault gouge zones).

Finally, seismology is in a position akin to crustal geodynamics: It is being buried by data. The tremendous success of programs like IRIS (Incorporated Research Institutions for Seismology) in the acquisition and archival of seismic data now provides more than 25 Tbytes of high-quality digital seismograms available instantly over the Internet. New data are now coming in at a rate of 8 Tbytes/year with the USArray component of EarthScope. Seismology has moved from a data-limited field to an analysis-limited field. The foundation of computational seismology is the generation of synthetic seismograms, used in the modeling of earth structure, earthquake rupture, and aspects of wave propagation. In the past, considerable simplifications were made in the calculation of seismic rays, usually made using ray theory that limited their applicability. Recently, an important advancement for global and regional seismology was made with the spectral element method (SEM) (Komatitsch *et al.*, 2002) where the propagation of seismic waves through the Earth can now be modeled accurately. The SEM takes into account heterogeneity in earth models, such as three-dimensional variations of seismic wave velocity, density, and crustal thickness. Although the full solution to the wave equation is computationally expensive, it has been implemented on parallel computers (Komatitsch *et al.*, 2002). This combination of hardware and software enables the simulation of broadband seismograms without intrinsic restrictions on the level of heterogeneity or the frequency content. These 3-D codes will revolutionize seismology, allowing for a direct investigation of countless geodynamic topics (such as the fate of subducted lithosphere, existence of mantle plumes, lithospheric structure, plate boundary zone complexity) that are poorly suited for traditional "1-D" codes.

In the decades since the plate tectonics paradigm provided a framework for understanding the Earth, the earth sciences have changed markedly, as geoscientists have developed an increasing understanding of, and appreciation for, the range of scales in space and time that influence the planet's evolution. Advances in technology, as well as more and better instrumentation, have led to breakthroughs in our understanding of the Earth's structure and dynamics. Growing seismological databases have led to new, high-resolution models of the Earth's interior, providing detailed images of structure from the crust to the inner core (Romanowicz, 2008). In the past decade, satellite geodesy has allowed us to observe active deformation within the crust on the human timescale. We are now being inundated by vast quantities of observational data on the Earth's composition, structure, and dynamics. EarthScope is providing high-resolution seismic imaging of the North American continent through USArray, while the PBO component of EarthScope provides high-resolution, continuous geodetic observations of deformation along the western boundary of the North American plate.

## 5.2 Emerging from hero codes

Computer simulations came into wide use in geophysics during the decade after the plate tectonic revolution. Combining the new tectonic paradigm with high-quality observations allowed many problems to be addressed in novel ways by using computer models. Individual investigators responded with the development of self-contained research codes written from scratch for a narrow range of problems. Solution schemes and numerical algorithms that developed in other areas of science, most notably engineering, fluid mechanics, and physics, were adapted with considerable success to geophysics. Because the software was often largely the product of an individual effort, this style has informally been referred to as "heroic." This approach has proven successful as is evident from the range of innovative papers using computer models starting in the 1970s; indeed, many of the examples presented above resulted from this programming model. For specific individual projects, this style of code development is certainly easiest, and will likely remain useful into the future. Nevertheless, its strength for solving problems of interest is now starting to show its limitations as we try to share codes and algorithms, or when we want to recombine codes in novel ways to produce new science.

As an example, consider *Citcom*, one of several sophisticated codes for solving mantle convection problems (Moresi and Gurnis, 1996; Moresi and Solomatov, 1995). Since its initial development phase in the early 1990s, *Citcom* is entirely self-contained and written in *C* with its own unique data structures, mesh generators, input and output routines, and equation solvers for both heat transport and incompressible flow. Other convection, solid deformation or magma-migration codes have similar components but use different algorithms, physics, rheologies, or data structures, many of them tuned to meet the needs and interests of the individual investigator. Every one of these codes contains real gems of science and software. However, the ability to extract them, compare them, and recombine them into new science is currently daunting due to coding inconsistencies, language barriers, or poor documentation that were never designed to be publicly supported. It is clear to most modelers, that if we could change the way we do code development to a more modular, reusable approach, it would enable individual scientists to more readily use advances of other scientists while concentrating on the novel components of most interest to them.

This approach becomes crucial if we want to tackle multi-scale and multi-physics problems or combine additional techniques to calculate the observable consequences of dynamic problems. For example, few individual modelers have the time, background, or resources to write a comprehensive code that could calculate mantle convection and its seismic and geochemical signature. Moreover, there is a large community of users who would like to use simulation software for data interpretation, data assimilation, and hypothesis testing, but are not specialists in numerical methods. To make progress in these areas, train the next generation of modelers and geoscientists, and move the field forward as a whole requires that we change the way we develop scientific computations. This is not trivial and is unlikely to evolve spontaneously from the heroic model of code development. Given this backdrop, in 2005 the US National Science Foundation (NSF) sponsored such an effort – to move computational geodynamics forward – with the creation of a virtual organization, the Computational Infrastructure for Geodynamics (CIG).

## 5.3 The Computational Infrastructure for Geodynamics (CIG)

Computational Infrastructure for Geodynamics is an NSF center that develops computational software and maintains related services in support of earth sciences research. CIG is a community-governed organization with 46 Member Institutions (mostly research universities) and 10 Foreign Affiliates, under the control of an elected Executive Committee and administered by the California Institute of Technology. With a small team of software engineers, CIG develops, supports, and disseminates community-accessible software for the greater geodynamics community. The software is being developed for problems ranging widely from mantle and core dynamics, crustal and earthquake dynamics, magma migration, seismology, and related topics.

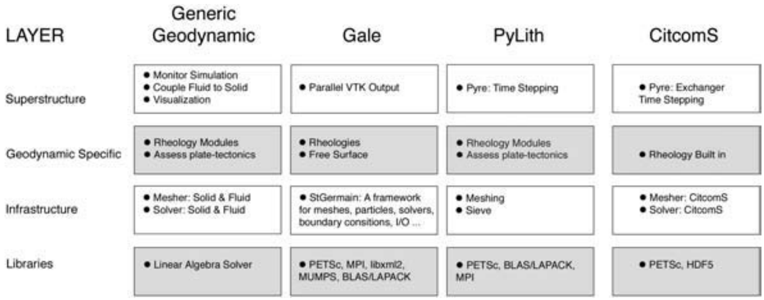



Figure 5.1. Hierarchy of software used generally by CIG, as well as how this approach has been used in three community codes: *Gale*, *PyLith*, and *CitcomS*.

With a high level of community participation, CIG has attempted to leverage the state of the art in scientific computing into a suite of open-source tools and codes. The infrastructure consists of a number of different components. At its core is a coordinated effort to develop reusable, well-documented, and open-source geodynamics software. The development is made possible by a central system facilitating software development, including a repository, a bug tracking system, and automatic regression testing (described in the next section). The CIG software is organized in a hierarchy with the basic building blocks – an infrastructure layer – of software by which state-of-the-art modeling codes can be assembled (Figure 5.1). On top are extensions of existing software frameworks to interlink multiple codes and data through a superstructure layer. In addition to this hierarchy, we have also found it necessary to develop a Science Gateway to allow users to initiate and monitor simulations on the TeraGrid via the Web. CIG is more than software, and involves strategic partnerships with the larger world of computational science, as well as specialized training and workshops for both the geodynamics and larger earth-science communities.

CIG is a virtual organization governed by an Executive Committee (EC). The structure of CIG recognizes member institutions, which are educational and not-for-profit organizations with a sustained commitment to CIG objectives, and a number of foreign affiliate members. CIG has a Science Steering Committee (SSC) that consists of eight elected members that have been fully engaged in a dialog with the user community and active users of CIG software. The committee has a balance of expertise in both geoscience and computational sciences and provides guidance within all of the subdisciplines of computational geodynamics. Their principal duties are to assess the competing objectives and needs of all the subdisciplines covered by CIG, provide initial assessment of proposals submitted to CIG, and revise the Five-Year Strategic Plan. Recommendations from the SSC are passed on to the EC. Concepts and plans for CIG activities have come directly from the community, member institutions, working groups, and their elected committees.

Ideas and plans will move from members to the Science Steering Committee and finally to the Executive Committee. As part of the development of the Strategic Plan each year with a running five-year window, the SSC formulates a prioritized list of tasks for software development for the coming year, how these tasks are both interrelated and related to the broader needs of the community, and then transmits this as a recommendation to the EC.

CIG has established a small team of dedicated software architects and engineers whose work is guided by scientific objectives formulated by the scientific community. The Software Development Team (SDT) provides software services to the community in terms of programming, documentation, training, and support. Guidance for the programmers comes from the SSC.

## 5.4 How CIG develops software

In a virtual organization like CIG, the priorities for software development result from a dynamic balance between scientific needs, resources, and what is feasible technically. We hope to convey this balance below, where we describe specific software development activities. Here we focus on the infrastructure that has facilitated software engineering, including development, verification, and maintenance of the software.

An important aim is to introduce good software design practices into the software development efforts at CIG. This includes techniques for automated build and test procedures, development of benchmarks and test cases, and documentation. The software repository and attendant web site are central to CIG's objectives of facilitating collaboration and sharing of validated open-source software and reusable components. The repository is critical to bring modern software engineering practices to our community and CIG's software development team. Originally, CIG used a single repository for developer and community use that managed multiple developers working concurrently on modular software components shared through the repository. We use the open-source package Subversion (*SVN*) for the main CIG software repository, which contains most of CIG's codes. For the magma dynamics project (*MADDs*), CIG started to use the Mercurial repository (*hg*), although this is beyond the scope of this chapter. The entire contents of both repositories are navigable from our web site. In addition, CIG provides a bug-tracking database (*Roundup*) to allow developers and external participants to register and comment on bugs and requests for new functionality in CIG software that can then be worked on by the developers of a program. Finally, for each of the subdisciplines, we maintain a Listserve as well as editable web pages through our *Plone* site (like a Wiki site).

A key problem that faces any dynamic software repository is ensuring that "nothing breaks" despite frequent dynamic changes needed to meet the evolving scientific goals of the community. CIG uses agile computing to minimize the risks of software development for continuously evolving requirements. In particular, the repository uses unit and regression testing. Building and testing in the *SVN* repository occurs either nightly or automatically in response to a software commit using *CIG-Regresstor*, a collection of *Python* codes we wrote. This software uses *Buildbot*, extended by us, and the results of the testing are both stored in a database and made available interactively on our web site (Figure 5.2). Nightly regression testing generates an electronic report that contains the build and test failures (including the platforms on which they occurred). Regression testing allows the SDT to rapidly identify when a change in a repository component or platform has caused an error or inconsistency. Regression testing gives users of the repository confidence in the robustness of the software. We also extended *Buildbot* so that executable binaries for common platforms are automatically generated.

Implementing a comprehensive program of software verification has been an important aspect of CIG. This is a complex and rich topic, but had not been uniformly executed within computational geophysics. Software must first be verified, meaning that the software works as expected and that the equations solved give the expected results. However, the software must also give valid results, meaning that the physics and algorithms embodied in the software reproduce what occurs in either experiment or nature. CIG has mostly focused on software verification because much of software validation is related to the core of geodynamics research. Whether or not a certain computer code that has already been extensively verified adequately represents the physics of the underlying problem for which an algorithm has been designed has been a topic beyond CIG.

We used a multilevel verification plan. The solution from each *geophysical* solver (such as for Stokes flow) is compared against known analytical solutions to the governing equations (see Zhong *et al*., 2008, for an example). For each solver that CIG produces, we make available at least one analytic solution that the solver can be automatically compared against. The solutions and prior results of these tests on different computers are published in our manuals and web page. In some cases, we have participated in community activities in which a range of numerical codes attempt to match the results of laboratory experiments.

To facilitate error analysis, benchmarking, and code verification, we developed *Cigma* (CIG Model Analyzer) that consists of a suite of tools for comparing numerical models. The current version of *Cigma* is intended for the calculation of L2 residuals in finite element models, but can be extended to compare arbitrary functions. In error analysis, *Cigma* calculates both the local error and a global measure of the differences between two solutions by performing integration over a discretized version of the domain. This comparison can take place even when the underlying discretizations do not overlap. In benchmarking, we hope that *Cigma*

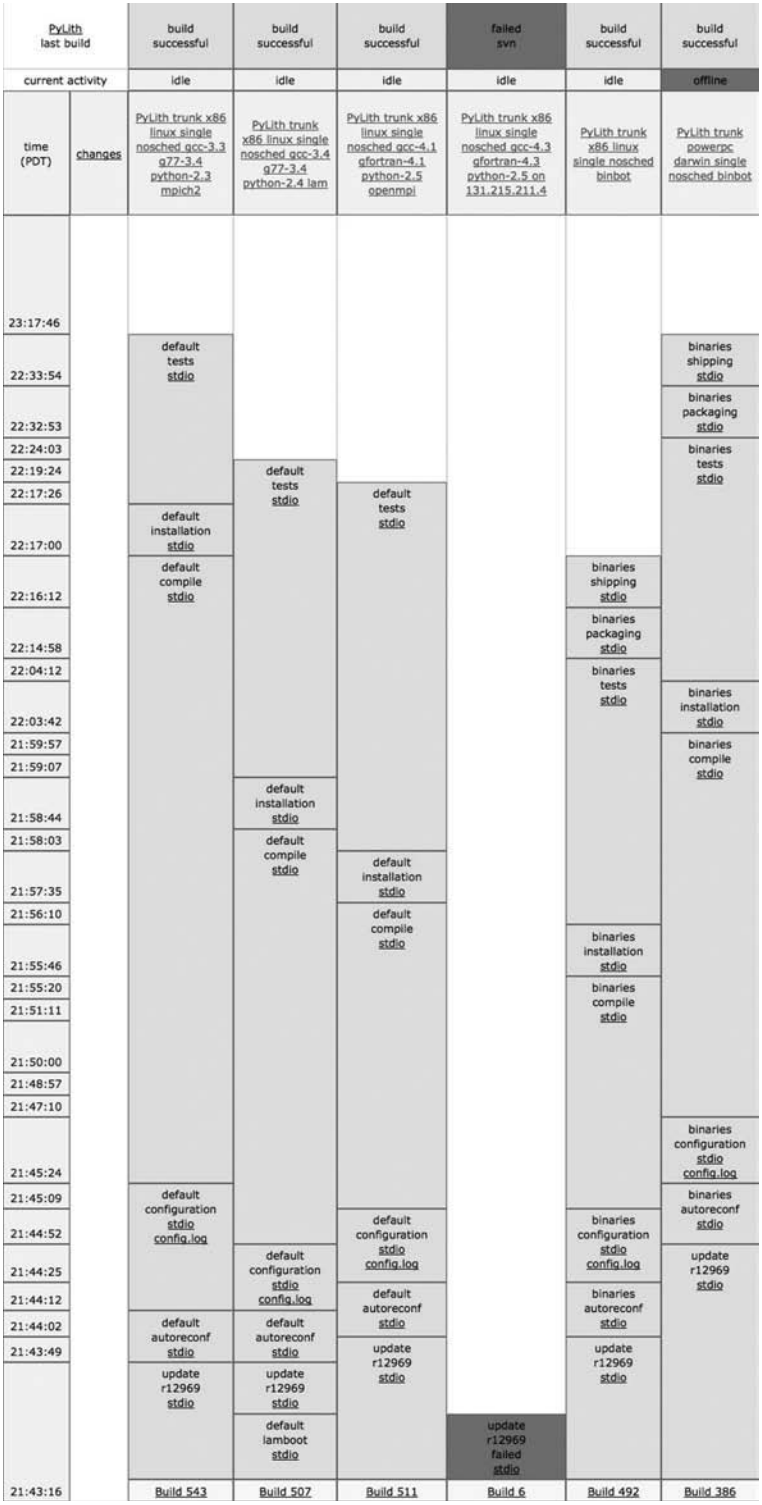



Figure 5.2. *Buildbot* automatically builds and tests code on a variety of platforms each time a change is made to the source. Here, *Buildbot*'s waterfall display indicates the status of some *PyLith* builds. Each column corresponds to a single platform.

will help the geodynamics community agree on a standard solution to specific problems by facilitating the process of comparing different numerical codes against each other. Lastly, as an automated tool, *Cigma* can help application developers create regression tests to ensure that software changes do not affect the consistency of the results. CIG developed *Cigma* in response to demand from the short-term tectonics community for a simple tool that can perform rigorous error analysis on their finite element codes. In the longer term *Cigma*, can be used for nearly all geodynamics modeling codes. *Cigma* relies on libraries. The Finite element Automatic Tabulator (FIAT) Python Library (FIAT, 2010) supports generation of arbitrary order instances of the Lagrange elements on lines, triangles, and tetrahedra. It can also generate higher-order instances of Jacobi-type quadrature rules on the same element shapes. The Approximate Nearest Neighbor (ANN) Library (ANN, 2010), written in *C++*, supports data structures and algorithms for both exact and nearest-neighbor searching in arbitrarily high dimensions. Both of these libraries extend and generalize *Cigma*'s functionality so it can handle other types of elements, and provide the ability to compare vector fields.

Of the tools we made available to the community to facilitate interaction, we found that activity was dominated by our Listserves. Few communities used our bug tracking software or our editable web pages, although there were notable exceptions. Several years into the life of CIG, we discovered that the ease-of-use of our software increased substantially if we distributed binaries for the packages, especially use that emerged from training sessions. We were able to streamline the production of binaries by automating their generation through our *Buildbot* system (a component that we called *Bin-bot*) (Figure 5.2).

## 5.5 Divergent development approaches

We quickly realized that even geodynamics, which one would have thought of as a rather small and homogeneous community, was in fact not so homogeneous. We found that each of the subdisciplines, as least within the United States, were distinct in terms of their technical expertise, their ability to develop codes, their reception of software engineering, and how close they were to achieving their goals with existing software. Some subcommunities were small but had enough experience in developing hero codes so that strong links with CIG never developed. While at another extreme, little experience existed to develop computational codes so that they were willing to collaborate with computational scientists. Most subcommunities fell between these extremes.

What we hope to convey are some details not only with software, but with community interactions and software development. The mantle convection community has a long history at developing and using codes. Two individual tectonics communities illustrate the different outcomes possible when they are comfortable with a few individuals taking the lead in developing entirely new codes. Finally, we will describe our experience with computational seismology where we developed a Science Gateway using existing codes for broad community use.

### *5.5.1 An emerging community code in mantle convection*

The mantle convection community has a long history of hero code development in which the products of software development have been passed on by generations of professors and graduate students. This community viewed the formation of CIG with suspicion since they nominally had the ability to develop codes and subsequently modify them for application to specific projects. CIG's primary role has been to bring standard well-proven codes under community control and then subsequently to add new features. CIG is now engaged in the development of a new package with adaptive mesh refinement (AMR) that transcends current capability (Burstedde *et al*., 2008). We will review the functionality of the community code and give a brief history of its development.

Much of CIG's development for the mantle convection community has been with *CitcomS*, a finite element code written in *C* that solves for thermal convection within a spherical shell. The code is capable of solving many different kinds of convection problems using the flexibility of finite elements. The normal sequence of steps for the solution of convection problems starts with an initial temperature field. First, the momentum equation is solved. The solution of this equation gives us the velocity from which we then solve the advection–diffusion equation, giving a new temperature solution. Variable viscosity, including temperature-, pressure-, position-, composition-, and stress-dependent viscosity are all possible. This code uses an iterative solution scheme to solve for the velocity and pressure and, as such, a converged solution cannot be guaranteed. Nested inside the iterative scheme, the code uses either a conjugate gradient solver or a full multigrid solver to solve the discretized matrix equations.

The original development of *Citcom* in the early 1990s proceeded in much the same way as any hero code (Moresi and Solomatov, 1995). The original code turned out to be quite modular and easily extensible, and the fundamental finite element infrastructure is still in place and forms the basis for most versions of the software. A number of features were quickly added by a distributed user community: a three-dimensional Cartesian version (Moresi and Gurnis, 1996), a parallelized version using message passing routines on a limited release Intel supercomputer (Zhong *et al*., 1998), a spherical version of the code named *CitcomS* (Zhong *et al*., 2000) and a Beowulf implementation (Conrad and Gurnis, 2003), and many others. We quickly found ourselves with a plethora of versions of *Citcom* around the world with different capabilities. Consequently, by 2002, there were so many different versions of the code that some rationalization was in order. The software was migrated into a version control and two versions, a fully spherical or regional model, were reengineered through the former GeoFramework project. By 2004, in order to increase the functionality of *CitcomS*, the developers began to reengineer the code into an object-oriented environment so that it could work with a *Python*-based modeling framework called *Pyre* (Cummings *et al*., 2002) (Figure 5.1). The release of the software became known as *CitcomS.py*, and allowed multiple simulations to dynamically interact, such as a regional model within a global flow model (Tan *et al*., 2006).

At this point, CIG took over the maintenance of the software for the community, but we found ourselves with a wide range of versions since the entire community neither used the same repository nor accepted the utility of software frameworks. Development of *CitcomS*, following the formation of CIG, proceeded in a somewhat similar fashion as occurring earlier with the GeoFramework project. Several members of the community had developed significant enhancements to the code that were not in the main repository trunk. Most notable was a particle tracking method that worked on top of the global mesh (McNamara and Zhong, 2004). Another was an alternative means by which to solve the equations for compressible flow. An example of combining the particle tracking with compressible convection is shown in Figure 5.3a. To increase functionality using common components, as illustrated in Figure 5.1, we incorporated the use of HDF5 (a parallel version of the Hierarchical Data Format). The most recent release of *CitcomS* (3.0) contains many new features, including two implementations of compressible convection; the ability to resume computation from previous checkpoints; multi-component chemical convection; a fixed non-Newtonian solver; and an exchanger package for solver coupling. By advocating the use of software repositories and allowing technically proficient users access to them, we have cut the number of alternative versions of the software down as well as shortening the time that features developed by others can be merged back into the main trunk.

CIG has also had an influence on the mantle convection community through workshops. We saw an important change in outlook between the two mantle convection workshops that we sponsored between 2005 (Boulder, CO) and 2008 (Davis, CA) (Table 5.1). What we observed during the first was a workshop dominated by the scientific questions, in which the computational methods were only reviewed. However, three years later, we saw an equally dynamic scientific discussion, but with discussion of new equation solvers and detailed attention paid to verification (benchmarking). We attribute this fundamental shift back to basics to the two mantle convection workshops and community organization between them, as shown in Table 5.1. A workshop on compressible mantle convection in Purdue, Indiana

Table 5.1. *Mantle convection and technical workshops*

| Title | Location | Date | Attendance |
|---|---|---|---|
| 2005 Mantle Convection Workshop | Boulder, CO | June 19–24, 2005 | 65 |
| 2006 Compressible Mantle Convection Workshop | Purdue, IN | March 27–28, 2006 | 14 |
| Workshop on Challenges and Opportunities at the Interfaces of Scientific Computing | Austin, TX | October 16–17, 2007 | 64 |
| AMR Tutorial Workshop | Boulder, CO | October 24–27, 2007 | 25 |
| 2008 Workshop for Advancing Numerical Modeling of Mantle Convection and Lithospheric Dynamics | Davis, CA | July 9–11, 2008 | 90 |
| Workshop on Mathematical and Computational Issues in the Solid Earth Geosciences | Santa Fe, NM | September 15–17, 2008 | 55 |

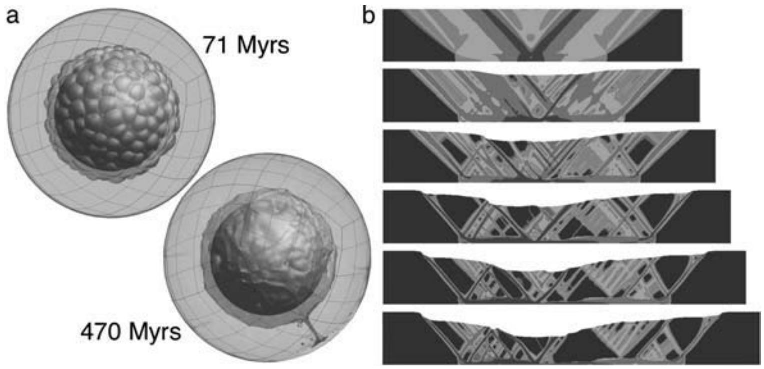



Figure 5.3. (a) Two time-instances in a calculation using *CitcomS* 3.0. The initial layer thickness is 250 km, and Ra = $1.4 \times 10^9$. The density of normal mantle increases 50% from the surface to the CMB due to compressibility. The anomalous mantle has 8% larger bulk modulus and is 3.8% denser at the surface and 1% denser at the CMB. The viscosity is temperature and depth dependent, varying 400× due to temperature and is 10× larger for the lithosphere and the lower mantle. 120M tracers and 64 × 64 × 64 × 12 = 3.1M elements and uses 48 processors. (b) Pictured is the strain-rate invariant for calculations of the GeoMod 2004 extension benchmark. The images range from 0 to 5 cm of extension, top to bottom. This model had a mesh of 512 × 32 and was computed on 32 processors using *Gale 1.3.0*. See color plates section.

focused on the development of a community code with new solvers, while a workshop on Challenges and Opportunities at the Interfaces of Scientific Computing in Austin, Texas brought investigators involved with new solver packages and meshing software into the broader discussion with geophysicists on how to advance the science. We viewed this change in community outlook as essential, since detailed discussions of the computational challenges were essential for progress on the geophysical questions.

### *5.5.2 Starting from scratch in tectonics*

Software development in tectonics resulted in two new codes. However, the software development proceeded in different ways and reflected different thrusts and technical abilities in two subcommunities. From the short-timescale tectonics (earthquakes) community, an existing working group with several individuals with computational experience teamed with CIG engineers to develop a new code from scratch. However, for long-timescale tectonics (orogenesis, basin formation), the community articulated its needs but was unable to take the lead developing the software to bring their vision to fruition. In that case, CIG teamed up with a group in Australia to produce a new code with common components.

While the planning for CIG was in its infancy, two members of the community, Brad Aagaard (USGS) and Charles Williams (RPI) began working towards integrating their individual codes, *EqSim* (Aagaard *et al.*, 2001) for dynamic rupture, and a version of *Tecton* (Melosh and Raefsky, 1980, 1981) for quasi-static problems), into the *Pyre* framework, with the ultimate goal of developing highly modular codes for the simulation of earthquake dynamics. A significant amount of commonality was identified between the codes, so that Aagaard and Williams then coordinated their development with a plan to merge their codes into a single suite of modules, *PyLith*. *PyLith* uses all of the levels in the software hierarchy that we first envisioned (Figure 5.1). As part of the general toolkit needed for the solution of many of the problems that CIG encounters, Matthew Knepley at Argonne National Laboratory (ANL) developed *Sieve*. *Sieve* is infrastructure for parallel storage and manipulation of general finite element meshes and can be used so that a developer avoids many of the complexities associated with parallel processing. *Sieve* is an integral component of the *PyLith*.

Developments in the long-timescale community proceeded quite differently. At a Tectonic Modeling workshop held in Breckenridge, Colorado, in June 2005, members of the tectonics community urged the development of a new open-source software code that could handle large deformations with viscoplastic rheologies. In particular, they advocated developing a code that uses the Arbitrary Lagrangian Eulerian (ALE) method. The ALE method uses a Eulerian grid to solve the Stokes flow problem and a Lagrangian grid to track material properties. Although such methods had long been used for 2-D problems, the expertise did not exist within the US community to develop a scalable 3-D code.

We realized that we could use technology developed by our partners, the Victorian Partnership for Advanced Computing (VPAC) in Australia, and develop such a code with common components. The end result was *Gale*, a parallel, two- or three-dimensional, implicit finite element code. The basic equations that *Gale* solves are the same as for mantle convection: Stokes and the energy equation. The development of *Gale* was jump-started by building on top of *Underworld* (Moresi *et al.*, 2003). *Gale* uses a hybrid particle-in-cell scheme, which combines a deformable mesh of computational points and a dense arrangement of mobile material points. The boundaries of the deformable mesh conform to the boundaries of the material as the simulation progresses, but the interior is constrained to remain as regular as possible. The particles track history-dependent properties such as strain for strain-softening materials. An example of a compression problem using *Gale* is shown in Figure 5.3b.

### *5.5.3 Production in a stable environment: An alternative for seismology*

Development for the seismology community was very different than that for either mantle convection or tectonics. Arguably, for computational seismology the major underlying algorithms have already been developed and engineered into highly scalable codes (Komatitsch *et al.*, 2002). The frontier in this field, a field being deluged with data, is to bring the computational tools to the observational seismologist's workbench. Consequently, CIG developed a Science Gateway to allow observational seismologists to harness the power of computational codes running on remote supercomputers.

The CIG Seismology Web Portal enables the user to request synthetic seismograms for any given earthquake, selecting from an assortment of 3-D and 1-D earth models. Simulations are performed on the TeraGrid platforms. Upon completion of a simulation, the user receives a notification email directing them back to the web portal, where they can download the resulting seismograms in ASCII and SAC format (Figure 5.4). Visualizations include a graphic that depicts the earthquake's source mechanism, and maps showing the locations of the earthquake and the seismic stations.

The portal runs 3-D simulations using SPECFEM3D GLOBE, which simulates global and regional (continental-scale) seismic wave propagation using the spectral element method (Komatitsch *et al.*, 2002). A typical SPECFEM simulation runs from two to three hours, using 150 to 216 processors. The portal's 1-D simulations are performed by the serial *Mineos* code, which uses normal mode summation. To simulate an earthquake, the portal needs source information in Harvard CMT format. To obtain these input data, the user can search for events in

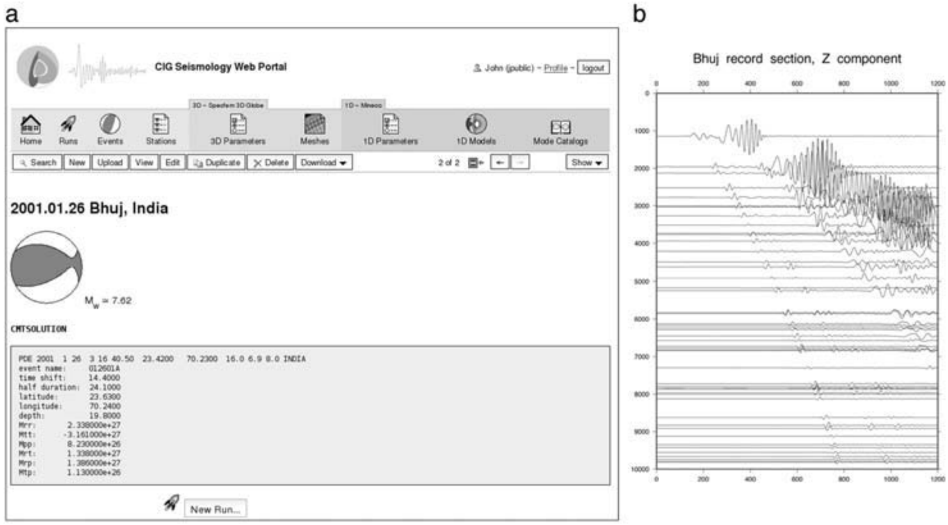



Figure 5.4. The CIG Seismology Science Gateway. (a) Screen shot. (b) Data created from the gateway, downloaded to a workstation, and displayed as a typical record section.

the database provided by the Global CMT Project. The database is integrated into CIG's web portal, allowing the user to select an earthquake by simply pointing and clicking. Alternatively, the user may upload custom CMT data to the portal. The portal provides a default set of seismic stations; the user may also upload a custom set.

The web site is written in *Python*, and built upon the *Django* web framework. Data persistence is achieved using an SQLite database. The site runs on top of Apache, which provides secure https connections. The web site is passive: it does not initiate connections to carry out tasks. Instead, the portal is powered by a separate daemon script that works in the background. The daemon constantly polls the web site looking for work to do. When a simulation request is posted, the daemon springs into action. First, it downloads input files for the simulation from the web site. Next, it connects to the TeraGrid cluster using GSI-Enabled OpenSSH and MyProxy credentials; then, it schedules and monitors an LSF batch job. Finally, it transfers simulation output files to a web server. Throughout this process, the daemon posts status updates to the web site using http POST, so that the user can monitor the progress of their request. The daemon script and other back-end helper scripts are also written in *Python*, but using the *Pyre* framework. The "beachball" and map visualizations are generated on the fly by the web server using GMT. The server to optimize performance caches the images.

## 5.6 Conclusions and future opportunities

As of this writing, CIG has finished its initial five years of community building and software development. An important component of the planning for the future was an assessment of the impact that CIG has had on geodynamics. In 2009, the geodynamics community produced an extensive document of science abstracts and statistics that painted a picture of major scientific advances using community developed software (Gurnis *et al.*, 2009). Research using CIG software has, for example, focused on enabling direct links between mantle convection and seismic observations, mineral physics, earth tides and gravity, and observations of surface vertical motions and sea level change, as well as elucidating the dynamics of plate tectonics, subduction, plumes, and the interior of other planets. CIG software has been used to improve images of the Earth's interior through adjoint tomography, directly test tomographic models with seismic waveforms, and improve and automate moment tensor inversions of earthquakes. It has also been used in some of the first three-dimensional models of the initiation and growth of faults in extensional tectonic environments. CIG software has facilitated studies of the interplay of crustal extension and melting, hill slope failure, and surface inflation associated with volcanic intrusion.

As the CIG community effort goes forward, it is essential that we modify how we do business, develop software, and interact as a community. Perhaps the most important change now taking place is to develop software using a common set of libraries. Adaptive mesh refinement has emerged as an important area where more emphasis is needed and where important progress has been made over the last several years. One possibility is to exploit the methods used for large mantle convection problems using octree meshes, which can be scaled to $10^5$ processors and allow resolutions as small as 1 km here needed (Burstedde *et al.*, 2008), to be applied broadly for convection, crustal dynamics, and magma dynamics problems. As the CIG community effort moves its central site from Caltech to the University of California, Davis, we are optimistic that the rapid pace of advancement in the use and application of community-developed and applied geodynamics software will continue.

## References

Aagaard, B., Heaton, T., and Hall, J. (2001). Dynamic earthquake ruptures in the presence of lithostatic normal stresses, implications for friction models and heat production. *Bulletin of the Seismological Society of America*, **91**: 1765–1796.

ANN (2010). www.cs.umd.edu/~mount/ANN/.

Bercovici, B., Ricard, Y., and Richards, M. A. (2000). The relation between mantle dynamics and plate tectonics: A primer. In *History and Dynamics of Global Plate*

*Motions*, ed. M. A. Richards, R. G. Gordon and R. D. van der Hilst. Geophysical Monograph 121. American Geophysical Union, pp. 5–46.

Burstedde, C., Ghattas, O., Gurnis, M. *et al*. (2008). Scalable adaptive mantle convection simulation on petascale supercomputers. International Conference for High Performance Computing, Networking, Storage, and Analysis, ACM/IEEE Supercomputing.

Conrad, C. P. and Gurnis, M. (2003). Seismic tomography surface uplift, and the breakup of Gondwanaland: Integrating mantle convection backwards in time. *Geochemistry Geophysics Geosystems*, **4**: 1031, doi:1010.1029/2001GC000299.

Cummings, J., Aivazis, M., Samtaney, R. *et al*. (2002). A virtual test facility for the simulation of dynamic response in materials. *Journal of Supercomputing*, **23**: 39–50.

Davies, G. F. (1999). *Dynamic Earth: Plates, Plumes and Mantle Convection*. Cambridge: Cambridge University Press.

FIAT (2010). www.fenics.org/wiki/FIAT.

Gurnis, M. *et al*. (2009). *Expanding Computational Infrastructure: The First 5 Years of CIG*. Pasadena, CA: Computational Infrastructure for Geodynamics, http://geodynamics.org/cig/proposalsndocs/documents/cig-first-5-years.

Karner, G. D., Taylor, B., Driscoll, N. W., and Kohlstedt, D. L. (2004). *Rheology and Deformation of the Lithosphere at Continental Margins*. New York: Columbia University Press.

Komatitsch, D., Ritsema, J., and Tromp, J. (2002). The spectral-element method: Beowulf computing and global seismology. *Science*, **298**: 1737–1742.

McNamara, A. K. and Zhong, S. (2004). Thermochemical structures within a spherical mantle: Superplumes or piles? *Journal of Geophysical Research*, **109**: doi:10.1029/2003JB002847.

Melosh, H. J. and Raefsky, A. (1980). The dynamical origin of subduction zone topography. *Geophysical Journal of the Royal Astronomical Society*, **60**: 333–354.

(1981). A simple and efficient method for introducing faults into finite element computations. *Bulletin of the Seismological Society of America*, **71**: 1391–1400.

Moresi, L. N. and Gurnis, M. (1996). Constraints on the lateral strength of slabs from three-dimensional dynamic flow models. *Earth and Planetary Science Letters*, **138**: 15–28.

Moresi, L. N. and Solomatov, V. S. (1995). Numerical investigation of 2D convection with extremely large viscosity variations. *Physics of Fluids*, **7**: 2154–2162.

Moresi, L. N., Dufour, F., and Muhlhaus, H.-B. (2003). A Lagrangian integration point finite element method for large deformation modeling of viscoelastic geomaterials. *Journal of Computational Physics*, **184**: 476–497.

Moresi, L., Gurnis, M., and Zhong, S. (2000). Plate tectonics and convection in the Earth's mantle: Toward a numerical simulation. *Computing in Science and Engineering*, **2**: 22–33.

Ranalli, G. (1995). *Rheology of the Earth*. New York: Chapman and Hall.

Romanowicz, B. (2008). Using seismic waves to image Earth's internal structure. *Nature*, **447**: 198–201.

Schubert, G., Turcotte, D. L., and Olson, P. (2001). *Mantle Convection in the Earth and Planets*. Cambridge: Cambridge University Press.

Tan, E., Choi, C., Thoutireddy, P., Gurnis, M., and Aivazis, M. (2006). GeoFramework: Coupling multiple models of mantle convection within a computational framework. *Geochemistry Geophysics Geosystems*, **7**, Q06001, doi:10.1029/2005GC001155.

Zhong, S., Gurnis, M., and Moresi, L. (1998). The role of faults, nonlinear rheology, and viscosity structure in generating plates from instantaneous mantle flow models. *Journal of Geophysical Research*, **103**: 15 255–15 268.

Zhong, S., McNamara, A., Tan, E., Moresi, L., and Gurnis, M. (2008). A benchmark study on mantle convection in a 3-D spherical shell using CitcomS. *Geochemistry Geophysics Geosystems*, **9**, Q10017, doi:10.1029/2008GC002048.

Zhong, S., Zuber, M. T., Moresi, L. N., and Gurnis, M. (2000). The role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *Journal of Geophysical Research*, **105**: 11 063–11 082.

# 6

# Parallel finite element modeling of multi-timescale faulting and lithospheric deformation in western USA

MIAN LIU, YOUQING YANG, QINGSONG LI, GANG LUO, AND HUAI ZHANG

## 6.1 Introduction

The lithosphere, which includes the crust and the uppermost part of the mantle, is the mechanically strong outer shell of the Earth. The lithosphere is broken into a dozen or so pieces, called the tectonic plates. Constant motions between these tectonic plates cause lithospheric deformation, which leads to mountain building, earthquakes, and many other geological processes, including geohazards. Hence understanding lithospheric deformation has been a major focus of modern geosciences research.

Most studies of lithospheric dynamics are based on continents, where the lithosphere is relatively soft and hence deforms strongly, and where direct observation is relatively easy. One major observational technique developed in the past decades is space-based geodesy (e.g., Gordon and Stein, 1992; Minster and Jordan, 1987). The Global Positioning System (GPS), the Interferometric Synthetic Aperture Radar (InSAR), and other space-based geodetic measurements are providing unprecedented details of deformation at the Earth's surface, thus revolutionizing the studies of lithospheric dynamics.

As space geodetic data grew rapidly in recent years, it has become increasingly clear that interpreting the tectonic implications of these data may not be straightforward. In many places, space-geodetic measurements of crustal deformation differ significantly from that reflected in geological records (Dixon *et al.*, 2003; Friedrich *et al.*, 2003; He *et al.*, 2003; Liu *et al.*, 2000; Pollitz, 2003; Shen *et al.*, 1999). This could be due to many factors; an important one is timescale-dependent lithospheric rheology and deformation (Jeanloz and Morris, 1986; Liu *et al.*, 2000; Pollitz, 1997, 2003; Ranalli, 1995). Multi-timescale deformation is clearly expressed in active faults where slip on the fault and deformation in the surrounding crust vary periodically over the seismic cycles consisting of co-, post-, and inter-seismic phases (Chlieh *et al.*, 2004; Savage and Burford, 1973; Weldon *et al.*, 2004).

Space geodetic data are acquired typically every few years, reflecting mainly interseismic deformation. Additional information, and various assumptions, are often necessary to relate the geodetic results to geological records of fault slips (Becker *et al.*, 2005; Meade and Hager, 2004) and crustal deformation (Flesch *et al.*, 2000; Kong and Bird, 1996).

Numerical modeling has been a major avenue to study lithospheric dynamics, but modeling multi-timescale lithospheric deformation is challenging. Including seismic cycles in models of long-term lithospheric deformation, for example, is computationally intensive because of the need to use small time steps in a long-time simulation. Furthermore, lithospheric rheology is timescale dependent. It is close to elastic or viscoelastic for short-term crustal deformation associated with seismic cycles (Meade and Hager, 2004; Okada, 1985; Savage, 1983), but long-term lithospheric deformation would be better represented by viscous rheology (Bird and Piper, 1980; England and McKenzie, 1982; Flesch *et al.*, 2000; Yang *et al.*, 2003). These problems have forced numerical models of lithospheric dynamics to separate artificially into two groups: those focusing on short-term deformation, usually constrained by space geodetic data, and those focusing on long-term lithospheric deformation, based mainly on geological history.

Rapid development of computer hardware and computational algorithms in recent years provides an opportunity to narrow the gap between the short- and long-term lithospheric dynamic models. The Beowulf PC clusters and other low-cost parallel computers make supercomputing power affordable to researchers, and the lithospheric geodynamic communities all over the world are making progress in tapping into the power of parallel computers to attack problems that were computationally prohibitive no long time ago. We have been developing a parallel computing system for lithospheric dynamics, and have applied it to explore some of the multi-timescale lithospheric deformation in western USA (Liu *et al.*, 2007). Here we present an updated overview of our ongoing studies with new models and results.

## 6.2 Tectonic background of western USA

The western United States is marked by strong lithospheric deformation (Figure 6.1). During the Mesozoic and the early Cenozoic, subduction of the oceanic Farallon plate under the North American plate elevated much of today's western USA and produced the Rocky Mountains (Atwater, 1970; Atwater and Stock, 1998; Axen *et al.*, 1993; Burchfiel *et al.*, 1992; Burchfiel and Davis, 1975; Coney, 1978; Dickinson, 2002; Lipman *et al.*, 1972; Sonder and Jones, 1999; Stewart, 1978; Wernicke, 1992; Zoback *et al.*, 1981).

The subducting Farallon plate has gradually disappeared under the North American plate except in the Pacific Northwest, where the Juan de Fuca plate,

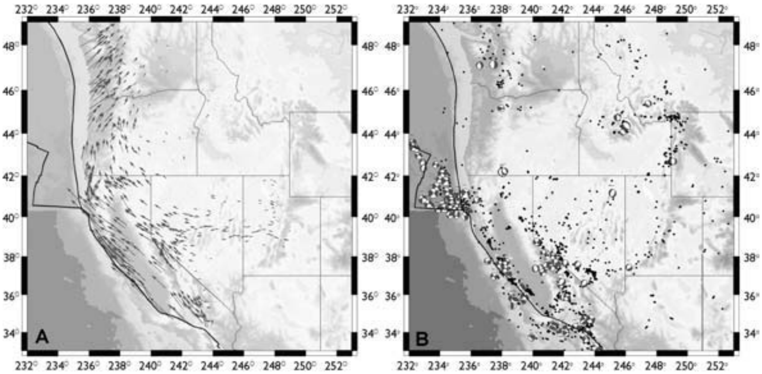



Figure 6.1. (A) Topography (background) and GPS velocity (vectors) relative to stable North America. GPS data are from USGS (http://quake.wr.usgs.gov/research/deformation/gps/qoca/index.html). (B) Seismicity (dots, from USGS) and focal mechanisms solutions (from Harvard catalog) in western United States. See color plates section.

which is the residue of the Farallon plate, continues to subduct under the North American plate (Figure 6.1). However, such a subduction process has not produced any significant mountain building in the Cascadia through the Cenozoic (Gabrielse and Yorath, 1992). On the other hand, extensional tectonics initiated in a belt of thickened crust that extended from southern British Columbia all the way to northern Mexico (Armstrong and Ward, 1991; Coney and Harms, 1984; Wernicke *et al.*, 1987). Extension has become widespread in southwestern USA since ~30 Ma, when the Farallon-North American convergent plate boundary started to be gradually replaced by the San Andreas Fault transform boundary (Atwater, 1970). In the Basin and Range province, the total amount of extension is up to 300% (Hamilton and Myers, 1966; Wernicke, 1992; Zoback *et al.*, 1981), and the extension continues to the present day (Bennett *et al.*, 2002; Wernicke *et al.*, 2000) (Figure 6.1). As a result of the extension, the crust in southwestern USA is abnormally thin, less than 30 km in places (Benz *et al.*, 1990), and the high elevation of southwestern USA is presently supported by the abnormally hot mantle (Liu and Shen, 1998; Sonder and Jones, 1999; Thompson *et al.*, 1989; Wernicke *et al.*, 1996).

The San Andreas Fault (SAF), the plate boundary between the Pacific and the North American plates (Figure 6.1), accommodates 60–80% of the Pacific–North American relative plate motion by fault slip; the rest is partitioned over a broad plate boundary zone that includes many secondary faults and the Eastern California Shear Zone, which currently accommodates up to 25% of the relative plate motion (Dokka and Travis, 1990; Gan *et al.*, 2000; McClusky *et al.*, 2001; Miller *et al.*, 2001a; Savage *et al.*, 2001).

In the past two decades space-based geodetic measurements, especially the GPS, have greatly improved the kinematics of crustal deformation in western USA (Argus and Gordon, 1991; Bennett *et al.*, 2002; Bennett *et al.*, 1999; Bennett *et al.*, 2003; Davis *et al.*, 2006; Dixon *et al.*, 2000; Dixon *et al.*, 1995; Miller *et al.*, 2001b; Minster and Jordan, 1987; Murray and Lisowski, 2000; Thatcher *et al.*, 1999; Ward, 1990; Wernicke *et al.*, 2000). The results show spatially diffuse and variable deformation. South of latitude 36ºN, much of the relative motion between the Pacific and North America Plates is taken up by dextral shear along the southernmost San Andreas Fault system. At the latitudes of the northern Basin and Range province, a significant portion of the Pacific–North America relative motion is accommodated by broad deformation across the Basin and Range province. Further north in the coastal regions of the Pacific Northwest, crustal deformation is dominated by NE compression related to the subduction of the Juan de Fuca plate.

Whereas the GPS-measured crustal deformation is broadly consistent with late Cenozoic tectonics in western USA, some significant discrepancies are noticeable. For example, in the Cascadian coastal region, geological records do not show strong northeastward crustal shortening as indicated by the GPS data. Instead, the moment tensor of small earthquakes in this region indicate E–W extension (Lewis *et al.*, 2003), consistent with evidence of structural geology (Wells and Simpson, 2001; Wells *et al.*, 1998). For much of the southwestern USA, the present-day deformation is dominated by dextral shear along the SAF plate boundary zone, contrasting to geological records of predominantly extensional tectonics in the Basin and Range province.

Some of the discrepancies between these datasets are expected because space-geodesy measures only short-term strain rates, whereas geological records show lithospheric deformation over millions of years. Timescale matters because lithospheric deformation is timescale dependent; furthermore, tectonic boundary conditions and driving forces, and lithospheric structure, may change through time. For example, crustal extension in western USA throughout the Cenozoic may have reduced the gravitational potential energy that drives extension (Liu, 2001; Sonder and Jones, 1999), so the driving forces and tectonic boundary conditions, averaged over late Cenozoic as recorded in geological data, differ from those of today. To understand the cause of lithospheric deformation in western USA and the control of strain distribution, we need a better understanding of how lithosphere deforms at different timescales. To do so, we need advanced numerical models and high computing power.

## 6.3 Parallel finite element modeling (FEM)

Numerical modeling has been a powerful way to investigate lithospheric dynamics. It generally involves solving for displacement of the model lithosphere with specified loading and rheological structures. The finite element method (FEM) is commonly used for simulating lithospheric dynamics. By discretizing the model domain into a finite number of elements with simple geometry, FEM turns a system of partial differential equations (PDE) with complex initial and boundary conditions into a system of algebraic equations that can be solved by computers.

Parallel computers, especially the Beowulf class PC- and workstation-clusters, have provided affordable supercomputing power to researchers in recent years. While parallel computing has been commonly used in many fields, such as atmospheric dynamics, its application in lithospheric dynamics has been rather limited. This is mainly because parallel finite element computation requires sophisticated algorithm development. The difficulty is easier to overcome in scientific communities where resources can be pooled to develop common community models; the examples include the general global circulation models for atmosphere and ocean, such as those developed in NASA (http://aom.giss.nasa.gov/general.html), and the MM5 parallel computing software packages (www.mmm.ucar.edu/mm5/) for weather forecasting. The lithospheric dynamic problems, however, are so diverse that no general-purpose models can meet all the needs. In recent years, community effort has produced an array of finite element models for specific problems in lithospheric dynamics, such as the *Pylith* (www.geodynamics.org/cig/software/packages/short/pylith) finite element codes for quasi-static and dynamic deformation of the lithosphere, and the GeoFEM project (http://geofem.tokyo.rist.or.jp/) that provides a multi-purpose, multi-physics parallel finite element solving environment. Commercial finite element packages, such as PDE2D (http://members.aol.com/pde2d/) and Finite Element program GENerator (FEGEN) (www.fegensoft.com), allow users to generate finite element codes using high-level scripts (modeling languages) in the input files (drivers) that specify the PDEs and model-specific properties.

We have been developing an automated parallel computing system (Zhang *et al.*, 2007). Instead of providing static packages of specific models, this system allows users to define their problem in high-level scripts, and then generates machine-independent FORTRAN source codes for parallel finite element computing. Details of this system are given in Zhang *et al.* (2007). In this system, users specify the partial differential equations, the algorithms, and solvers in a high-level script (Figure 6.2). Whereas the PDEs vary with problems, in the FEM they are eventually turned into a system of algebraic equations, so a large part of the finite element codes are similar, some even identical, for different FEM models. We refer to these codes as the static parts; in our system they are written as standardized FEM segment source codes, or program stencils. For parallel codes, these program stencils are designed using the domain decomposition method (Quarteroni and Valli, 1999). This method divides the modeling domain into sub-domains; by doing so the original problem is turned into a group of relatively smaller boundary value

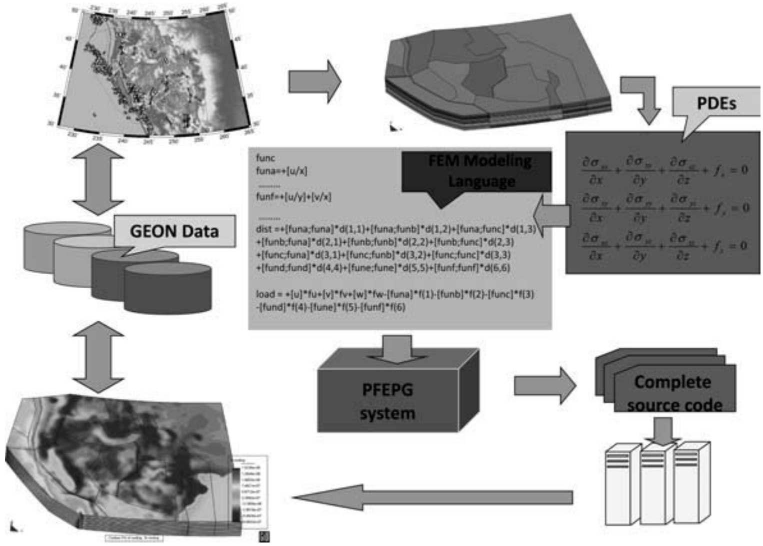



Figure 6.2. Workflow of the automated parallel FEM system. See color plates section.

problems that can be solved simultaneously on different nodes of a parallel computer. Various preconditioners (Ainsworth, 1996) can be used to treat the inner boundaries among the sub-domains, and a suite of solvers, such as the multigrid (Lv *et al.*, 2006) and multilevel (Axelsson and Larin, 1998) solvers, are included. Depending on the user's input, the system assembles a complete finite element program. The workflow of this modeling system is illustrated in Figure 6.2.

## 6.4 Modeling multi-timescale lithospheric deformation in the western United States

In this section, we present some of our recent parallel finite element modeling of multi-timescale lithospheric deformation in western USA. These studies focus on: (1) the cause of diffuse crustal deformation in the southwestern USA, (2) short- and long-term crustal deformation across the Cascadian subduction zone, and (3) fault slip along the San Andreas Fault and the associated crustal deformation at different timescales.

### *6.4.1 Short- and long-term crustal deformation in the southwestern USA*

The western United States is one of a few places in the world where crustal deformation diffuses far from the plate boundary (Figure 6.1). Such widespread

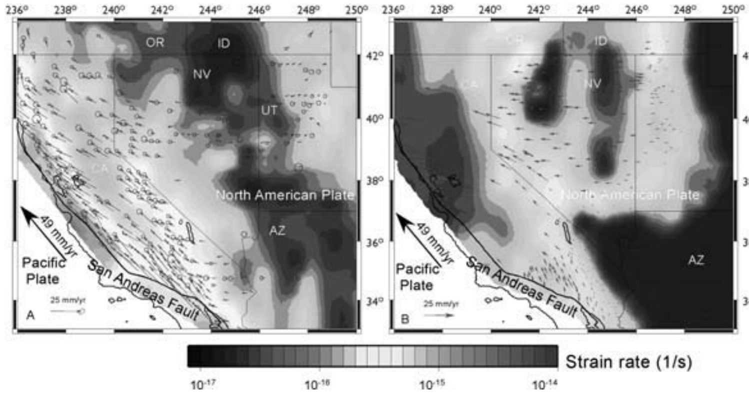



Figure 6.3. The GPS (A) and geologically (B) derived surface velocities (vectors) and strain rates (background color) in the southwestern United States. Source geologically determined surface velocity: McQuarrie and Wernicke (2005). See color plates section.

crustal deformation does not readily fit into the paradigm of plate tectonics, and the cause remains uncertain. Some workers suggest a dominant role of excessive gravitational potential energy (Jones *et al.*, 1996; Liu, 2001; Liu and Shen, 1998; Sonder and Jones, 1999), others see a strong role of shear coupling along the San Andreas Fault (Choi and Gurnis, 2003; Thatcher *et al.*, 1999; Zoback and Zoback, 1981), or traction on the base of the North American plate (Bokelmann, 2002; Liu and Bird, 2002; Silver and Holt, 2002).

The spatial distribution of crustal deformation provides the most important constraints for the driving forces (England and Molnar, 1997; Flesch *et al.*, 2000; Kong and Bird, 1996). Recent GPS measurements have greatly refined the crustal kinematics in the southwestern United States with unprecedented details (Figure 6.1a). While the GPS results reconfirm the geologically observed crustal extension across the Basin and Range province, the strongest crustal motion, however, is near and subparallel to the San Andreas Fault, indicating strong influence from the relative plate motion through shear coupling across the SAF. This deformation pattern differs significantly from that reconstructed from relative motions of crustal blocks in the past four million years (Figure 6.3). The long-term crustal deformation is dominated by E–W directed extension, the highest extension rates are around the margins of the Basin and Range province. The rates and the spatial pattern of extension are more consistent with gravitational spreading (Jones *et al.*, 1996; Liu and Shen, 1998) than with shear coupling along the plate boundary.

Why does the short-term crustal deformation, measured by the GPS in the past decade, differ so much from the geological records in the past few million years? Does this mean that gravitational spreading was important in the past, but its role has now

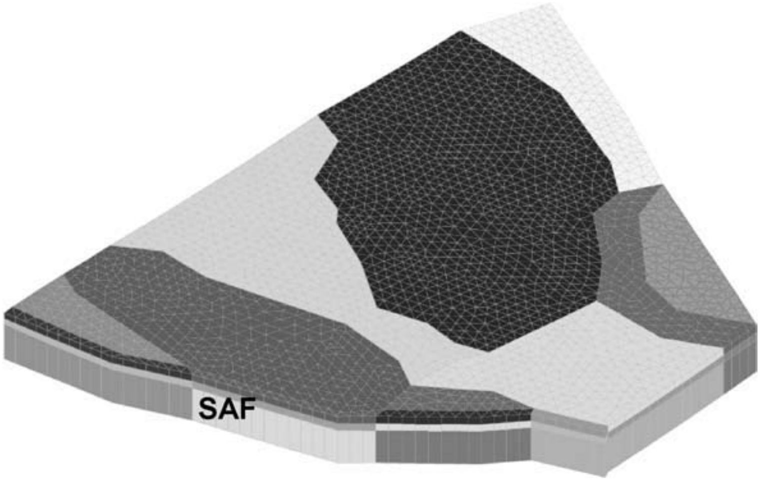



Figure 6.4. The finite element model. The blocks represent major tectonic units in southwestern USA. SAF: San Andreas Fault. See color plates section.

been replaced by shear coupling along the SAF? To explore the driving forces, we developed a finite element model for the southwestern USA (Yang and Liu, 2010) (Figure 6.4). The model is similar to previous viscous thin-sheet model for large-scale continental deformation (Bird and Piper, 1980; England and McKenzie, 1982; Flesch *et al.*, 2000) in that it assumes the rheology of viscous flow and is based on continuum mechanics. The major difference from viscous thin-sheet models include the following: (1) our model is fully three-dimensional, so lateral and vertical heterogeneities of lithospheric rheology can be included; (2) stress can be explicitly applied, and constrained, on the base of the model domain; (3) faults can be included in the model as rheological weak zones using the Goodman method (Goodman *et al.*, 1968); and (4) it can simulate viscoelastic deformation by increasing the viscosity of some part of the model (usually the upper crust) to $>10^{25}$ Pas so that the mechanical behavior of the stiff rocks is effectively elastic (Liu and Yang, 2003).

The model domain is fixed on the eastern side (the stable North America), so the predicted velocities are directly comparable with GPS or geological velocities relative to stable North America. The western side is bounded by the San Andreas Fault, on which proper traction forces are sought through iteration in the model. The northern and southern sides of the model domain are also traction boundaries on which the optimal tractions are sought through regression. In most cases, the optimal condition is for the southern side to be at the lithostatic pressure; on the northern side, 22 MPa sinistral shear is added to the lithostatic pressure, reflecting the change in tectonic boundary conditions north of the Basin and Range province (Figure 6.1). The model domain extends to 100 km depth. The surface is free. The bottom is free in horizontal directions but constrained in the vertical direction; in some cases, traction is applied to the bottom. The surface is based on the ETOPO5 topography, and the base of the crust is based on the CRUST2.0 model (http://mahi.ucsd.edu/Gabi/rem.dir/crust/crust2.html). Crustal density is assumed to be 2800 kg/m$^3$ for all regions; extra buoyancy forces needed to support the topography are provided by density deficiency in the mantle (Jones *et al.*, 1996). The model domain is divided according to the main tectonic units (Bennett *et al.*, 2003). Each unit can have its own rheological values.

Typical to most geodynamics problems, we have to solve for two unknowns (force and rheology) from one constraint (deformation field). We designed an algorithm to systematically search for the optimal solutions. For an assumed rheological structure, the driving forces can be determined through least-square fitting of the predicted and observed crustal displacement and stress orientations. The rheological structure is then modified, using a genetic algorithm of evolution and iteration, to seek for the optimal rheological structure that minimizes the residual errors (Yang and Liu, 2009). The power-law rheology for lithosphere (Brace and Kohlstedt, 1980; Kirby and Kronenberg, 1987) is nonlinear:

$$\tau = B\dot{E}^{1/n-1}\dot{\varepsilon} \tag{6.1}$$

where $B$ is stress coefficient, $n$ is power index, $\tau$ is deviatoric stress tensor, and $\dot{\varepsilon}$ is strain rate tensor. Define the effective strain rate as

$$\dot{E} = (2/3\dot{\varepsilon}_{ij}\dot{\varepsilon}_{ij})^{1/2}; (i = 1, 2, 3; j = 1, 2, 3) \tag{6.2}$$

We linearize the constitutive relation as $\tau = \eta\dot{\varepsilon}$, where $\eta$ is the effective viscosity: $\eta = B\dot{E}^{1/n-1}$. The linearized rheology allows us to seek for the optimal tectonic loading through linear regression.

Figure 6.5a compares the GPS velocities with the surface velocities predicted by a model with a "flat" western United States continent. The lateral variation of gravitational potential energy is excluded, and the model domain is loaded only by traction applied on the SAF. For a uniform effective viscosity of $5 \times 10^{22}$ Pa s, the predicted surface-velocity field fit the GPS data reasonably well to within 3.0 mm/yr. Hence the relative plate motion, through shear coupling across the SAF, is the main cause of the present-day crustal deformation in western USA as measured by the GPS. The optimal traction along the SAF is 25 MPa. Including the gravitational potential energy can slightly improve the averaged misfit to 2.8 mm/yr.

Here we used the viscous flow model to reproduce GPS data of surface deformation. Whereas this approach has been used by previous workers (England and Molnar, 1997; Flesch *et al.*, 2000), crustal deformation over the timescale of typical GPS measurements, usually a few years, is closer to elastic or viscoelastic than to

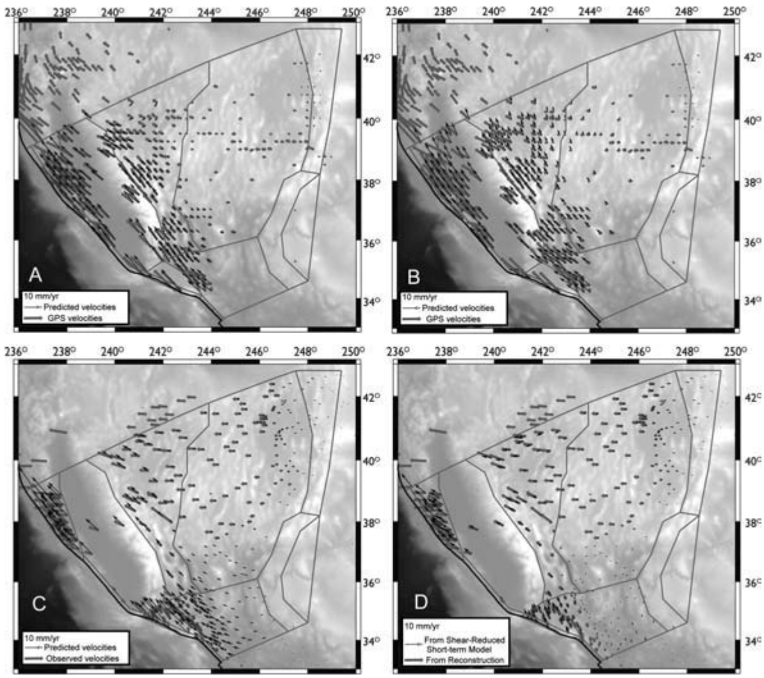



Figure 6.5. (A) Comparison of the GPS velocities with those predicted by a viscous model with plate boundary traction as the sole tectonic force. (B) Comparison of the GPS velocities with those predicted by an elastic model. (C) Comparison of the geologically derived crustal motion with the predictions of the short-term deformation model constrained by the GPS data. (D) A similar comparison but with reduced shear traction on the SAF in the model. See color plates section.

viscous (Meade and Hager, 2004; Savage and Burford, 1973). In one experiment we increased the viscosity of the upper crust to $1 \times 10^{25}$ Pa s so it effectively deforms as an elastic layer. The influence of gravitational potential energy diminishes with the increase of viscosity (Liu and Yang, 2003). With a uniform traction of 25 MPa, we obtained similar results (Figure 6.5b). These results indicate that the GPS velocity in southwestern USA represents largely the inter-seismic crustal deformation driven by shear coupling across the SAF.

The same driving forces, however, cannot explain the long-term crustal deformation constructed from geological records (Figure 6.5c). Such discrepancy cannot be simply attributed to uncertainties associated with the geological reconstruction – not only the predicted crustal motion is much faster than geological rates, their directions are significantly different.

We recalculated the driving forces using the geologically derived crustal motion as the primary constraints. We found that the model constrained by the GPS data can fit to the long-term crustal motion and stress orientations by lowering the traction forces on the SAF from 25 MPa to 17 MPa (Figure 6.5d). Further improvement is obtained by modifying the boundaries of the tectonic units in the model to better reflect tectonic patterns in the past few million years, assuming a relatively high viscosity for the central Basin and Range province and the Great Valley–Sierra Nevada block, and use a lower traction of 13 MPa on the SAF.

The lower traction on the SAF for long-term crustal deformation than that indicated by the GPS data is not surprising; the two datasets reflect different time-scales over which traction on the SAF is expected to vary. The GPS measures mainly the inter-seismic deformation when the San Andreas Fault is locked on most segments, so the shear traction is high. When the stress becomes large enough to cause sudden slips (earthquakes), much of the elastic strain is released, and only the plastic strain, a small portion of the total strain fields measured by the GPS, is preserved to lead to long-term crustal deformation. This explains why most of the discrepancies between the GPS and geologically inferred crustal deformation are found near the SAF. The stress difference between the best-fitting models for the short- and long-term deformation is 8–12 MPa, comparable with the typical stress drop for large inter-plate earthquakes (Stein *et al.*, 2003).

### *6.4.2 Short- and long-term crustal shortening in the Cascadia*

Along the coastline of the Pacific Northwest, the oceanic Juan de Fuca plate subducts under the North American plate; the tectonic setting is the same as that along the western margin of the South American plate (Figure 6.6). The plate convergence is causing ~10–20 mm/yr crustal shortening in the Cascadia, similar to that in the Andes. However, while the subduction-induced crustal shortening has produced the massive Andean mountain belts in South America (Allmendinger *et al.*, 1997; Isacks, 1988), no significant mountain building has occurred in the Cascadia since the Eocene (Gabrielse and Yorath, 1992).

The apparent disjunction between short-term crustal shortening and long-term mountain building in the subduction zones has not been well understood. Previous numerical models have focused either on short-term deformation, using elastic or viscoelastic models to simulate transient strain evolution during seismic cycles without linking the strain to long-term mountain building (e.g., Bevis *et al.*, 2001; Hu *et al.*, 2004; Wang *et al.*, 2003), or on long-term mountain building with viscous or viscoplastic models that assume continuous crustal shortening (e.g., Pope and Willett, 1998; Sobolev and Babeyko, 2005; Wdowinski and Bock, 1994).

To link the short-term crustal shortening to long-term mountain building, we have developed a two-dimensional viscoelastoplastic finite element model (Luo and Liu, 2009) (Figure 6.7). The model simulates strain evolution during the cycles of trench earthquakes, so it is focused on short-term crustal shortening. However, by

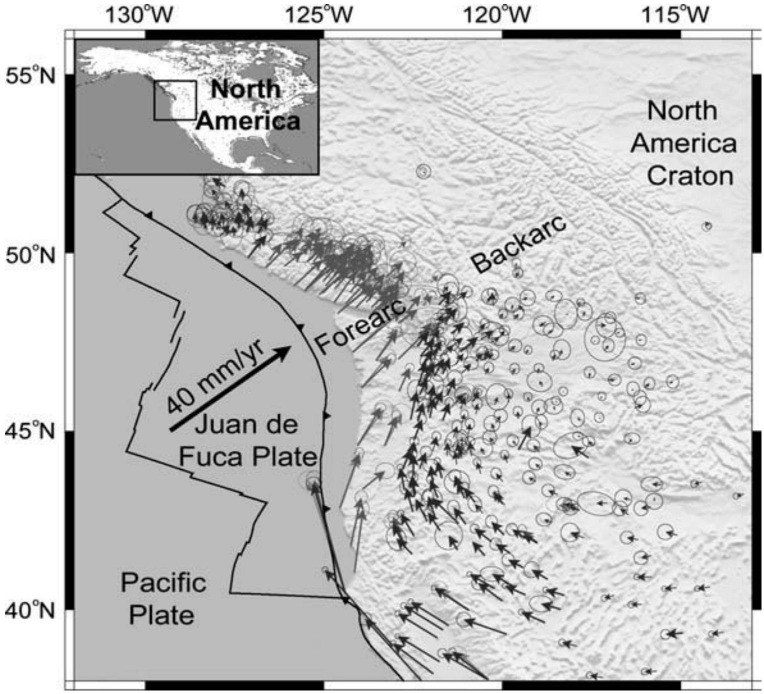



Figure 6.6. GPS velocities and topographic relief in the Cascadia subduction zone. Sources for the GPS data: Miller *et al.* (2001b) (green), Mazzotti *et al.*(2003) (red), and McCaffrey *et al.* (2007) (blue). See color plates section.

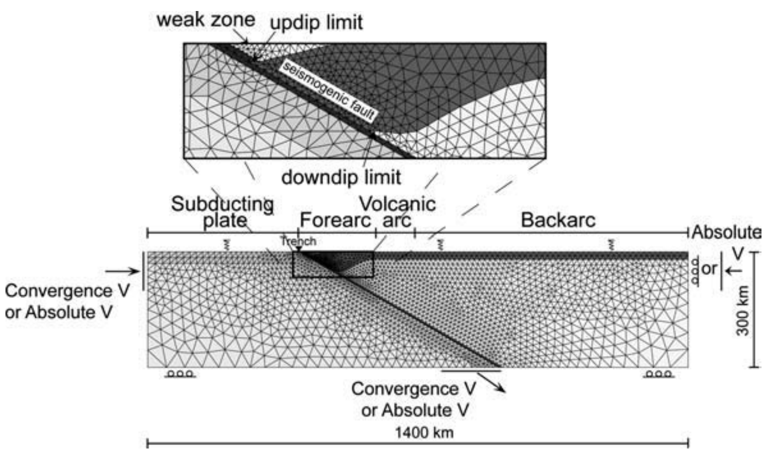



Figure 6.7. Mesh and boundary conditions of the finite element model.

exploring strain partitioning and the critical conditions for producing permanent (plastic) strain through the cyclic trench coupling, this model allows the short-term crustal deformation to be linked to long-term mountain building.

In the model, the subduction interface is modeled with two layers of special finite elements: the bottom layer with depth-variable low viscosity to simulate inter-seismic creeping, and the top layer with elastoplastic rheology to simulate the stick-slip seismogenic zone (Figure 6.7). The plastic deformation, both within the continental crust and on the subduction interface, is simulated with the Mohr-Coulomb yield criterion and non-associated flow rule (Jaeger *et al.*, 2007). To avoid the influence of unconstrained initial conditions, we run the models until they reach a quasi-steady state. We then calculate strain partitioning during seismic cycles. The intensive computation associated with calculating stress and strain evolution during multiple cycles of trench earthquakes is enabled by taking advantage of parallel computers.

The predicted displacement and strain evolution during one seismic cycle is shown in Figure 6.8. The inter-seismic displacement across the overriding plate (Figure 6.8a) shows distributed crustal shortening. During a trench earthquake or an aseismic slip, more than half of the inter-seismic crustal shortening may be restored

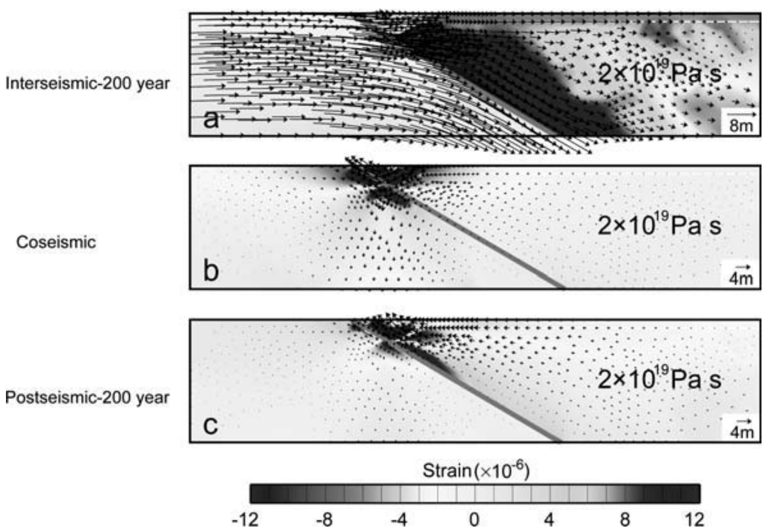



Figure 6.8. Predicted displacement (vectors) and horizontal strain (background color) across the plate boundary during one seismic cycle. (a) Accumulative inter-seismic displacement and strain for 200 years. (b) Co-seismic displacement and strain. (c) Accumulative post-seismic displacement and strain due to 200 years of viscous relaxation. Heavy gray lines show the subduction interface. White lines in the overriding plate show the bottom of the elastic or elastoplastic upper crust. Positive strain is compressive. See color plates section.

by co-seismic elastic rebound; the rest is restored by post-seismic viscous relaxation. In this case no plastic deformation, hence no mountain building, occurs in the overriding plate. Such trenchward co- and post-seismic rebound of the overriding plate has been observed in the Cascadia subduction zone (Dragert *et al.*, 2001). If all the inter-seismic crustal shortening is transient and is restored in later trench earthquakes and aseismic slip events, no long-term crustal shortening, hence mountain building, can be expected.

We explored the factors that control the partitioning between transient (viscoelastic) and permanent (plastic) strains across the subduction zone, and found that the critical factor is the relative magnitude of the strength of mechanical coupling on the trench plate interface and the yield strength of the overriding plate. Numerous lines of evidence indicate that trench coupling in the Cascadia subduction zone is weak: (1) force balance and forearc heat-flow data argue for low shear stress (<20 MPa) on the Cascadia subduction interface, contrasting to >37 MPa in the Peru–Chilean trench (Lamb, 2006; Wang and He, 1999; Wang *et al.*, 1995); (2) The thick (2–3 km) Cascadia trench sediments (Flueh *et al.*, 1998) imply lubrication, high pore fluid pressure, and low shear stress (Lamb, 2006), and the stress state in the Cascadia forearc indicated by small earthquakes is consistent with weak plate coupling (Wang *et al.*, 1995). The weak trench coupling limits compressive stress transferred to the overriding plate, so little plastic deformation occurred there. In this case most inter-seismic strain is transient, and is nearly fully recovered by the cycles of trench earthquakes and slip events (Figure 6.9).

### 6.4.3 Short- and long-term slips along the San Andreas Fault

In Section 6.4.1 we suggested that temporal variations of shear coupling along the San Andreas Fault has a major impact on short-term crustal deformation in western USA. In this section, we model the multi-timescale slips on the SAF and associated crustal deformation (Figure 6.10).

Because of the complex physics involved in faulting at different timescales, previous studies usually choose to focus on either short-term or long-term fault processes. The examples of short-term fault processes include dynamic fault rupture, static co-seismic and post-seismic stress evolution following fault ruptures (Bhat *et al.*, 2007; Freed *et al.*, 2007; Pollitz *et al.*, 2006; Xing *et al.*, 2004). The studies of long-term fault behaviors mainly focus on long-term steady-state fault slip, and fault evolution (Bird and Kong, 1994; Cooke and Kameda, 2002; Du and Aydin, 1996; Lavier and Buck, 2002; Li and Liu, 2006; Li *et al.*, 2009; Parsons, 2002). These models normally assume a viscous or creeping behavior of the fault. This separation of short- and long-term faulting is artificial, and it has been well recognized that short- and long-term fault processes are tightly connected with each

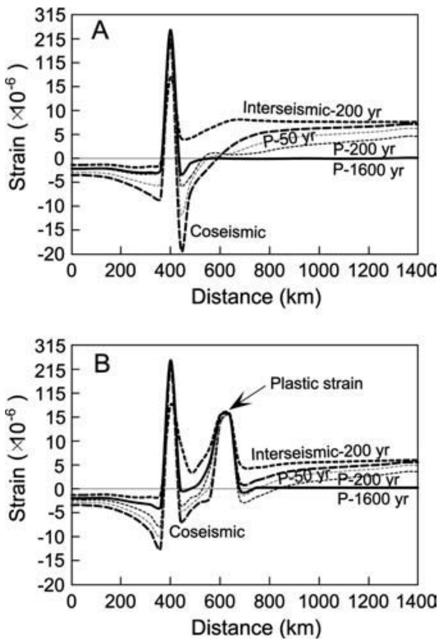



Figure 6.9. Predicted horizontal strain evolution on the surface during one seismic cycle. (A) Results of weak trench coupling. All strain is transient and will be eventually restored. (B) Results of strong trench coupling that leads to plastic strain, which is preserved during the co- and post-seismic rebound. The curves labeled P-50 yr are results for 50 years into the post-seismic relaxation; the meaning is similar for other curves.

other. The long-term fault slip determines the background stress field (Freed *et al.*, 2007; Li and Liu, 2006), which can be critical for studying single fault ruptures (Anderson *et al.*, 2003; Duan and Oglesby, 2006; Luo and Liu, 2010). On the other hand, it is the accumulation of slip events and the associated stress evolution that lead to long-term fault evolution. The computing power of parallel computers now permits exploration of some aspects of the multi-timescale faulting and crustal deformation.

We have developed a preliminary three-dimensional visco-elasto-plastic finite element model to integrate the short- and long-term fault processes (Li and Liu, 2006; Li and Liu, 2007; Liu *et al.*, 2007). We use a sudden drop of the plastic yield strength to simulate static fault rupture when the stress on a fault patch (element) exceeds the yield strength of the fault element. Failure of one fault element may cause stress increase in neighboring elements, triggering more elements to fail. When all the rupture ends, the displacement field is solved to meet the strength criterion on all the fault elements. The stress and strain fields are solved in a single time step through iteration to converge to the new yield strength. The model then

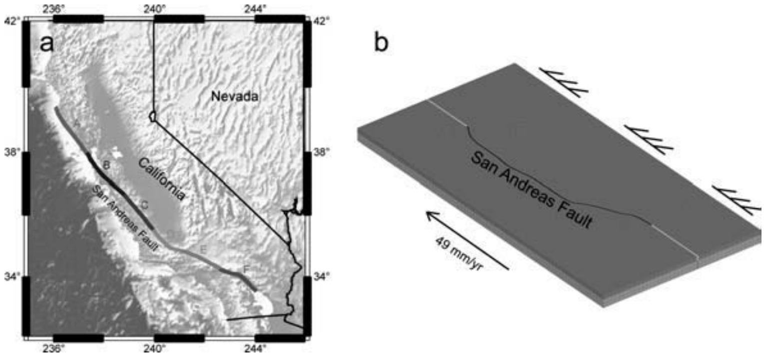



Figure 6.10. (a) Topographic relief and the San Andreas Fault (SAF). In the model, rupture patterns along each segment of the SAF are examined. (b) Numerical mesh and boundary conditions of the finite element model for the SAF system. The entire San Andreas Fault (black line) is explicitly included in the model. See color plates section.

enters an inter-seismic period, in which the fault strength is restored to the original value. If a fault element reaches the failure criterion, the rupture processes repeat. Without incorporating acceleration, the modeled rupture process is semi-static. By simulating a number of spontaneous fault rupture cycles and inter-seismic stress and strain evolution, the model can show the accumulative effects of long-term fault processes.

Figure 6.11 shows the cyclic stress changes, from co-seismic stress drop to post-seismic and inter-seismic stress buildup, for three sampling points located along central SAF. The points are within the fault block, and the maximum shear stress at these locations varies because of the variable fault traces (Li and Liu, 2006). The gradual post-seismic stress increase is due to tectonic loading and viscous relaxation in the lower crust and upper mantle, and the stress jump-ups are caused by fault rupture in neighboring fault segments. The fault ruptures, shown by stress drops at each location, are more regular along central SAF (Figure 6.12a) where the fault trace is relatively straight. Conversely, the rupture pattern in southern SAF over the Big Bend is more irregular (Figure 6.12b). These results imply that large earthquakes along central SAF segments may be more periodic than in southern SAF where the fault geometry is more complex. The current seismic record in the SAF, however, is not long enough to test this prediction.

After ~40 000 years of model time, the model enters a quasi-steady state, in which stress in the model domain fluctuates around the mean values. The imposed loading from the model boundary is released through both fault ruptures and plastic yielding in the off-fault crust. The results show relatively high shear stress around the transpressive fault bends, and low shear stress along the relatively straight fault

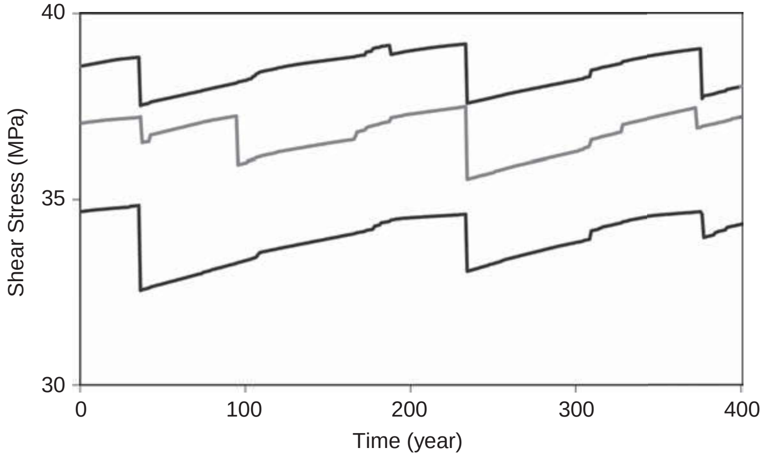



Figure 6.11. Predicted stress evolution at three sample points in the central segment of the SAF.

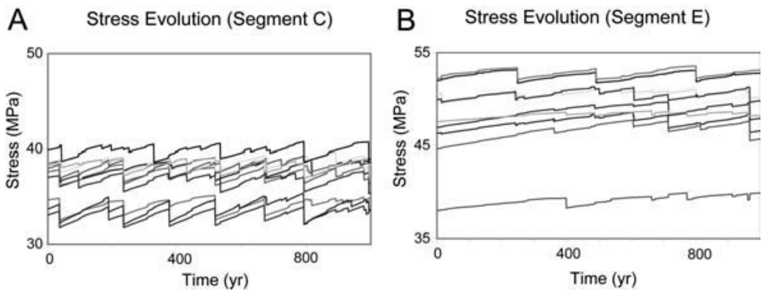



Figure 6.12. (A) and (B) Predicted rupture patterns in two segments of the SAF, shown by the stress evolution of a group of sample points along each segment. See color plates section.

segments (Figure 6.13a). The predicted regions of high shear stress are spatially correlated with seismicity in California. Note that the cumulative stress pattern differs significantly from the typical "butterfly" pattern of Coulomb stress changes associated with a single earthquake (e.g., Stein *et al.*, 1997). Furthermore, the stress variation in the off-fault crust is in tens of MPa, in comparison with the less than one MPa of Coulomb stress change of a single rupture event. Hence a better understanding of future earthquake hazards requires evaluating not only the Coulomb stress changes following a recent event, but also the background shear stress and the cumulative stress changes associated with all major earthquakes in the records.

It is encouraging to see that the cumulative stress field of the current model is similar to the steady-state shear stress field predicted by our earlier long-term

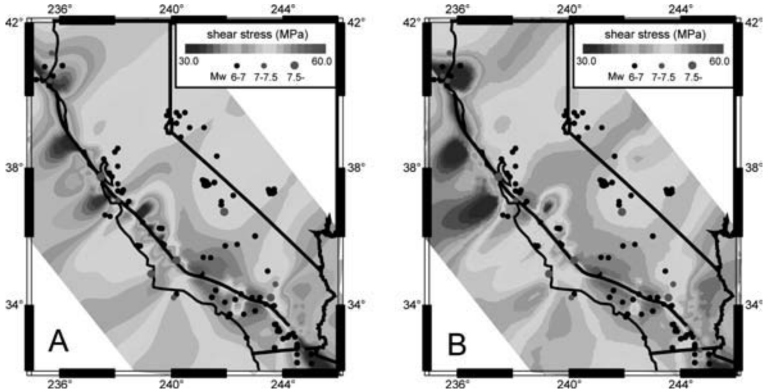



Figure 6.13. Predicted background shear stress. (A) Cumulative shear stress after 40 000 model years of repeated ruptures along various segments of the SAF. (B) Steady-state background shear stress predicted by the long-term faulting model where fault slip is simulated by continues plastic creep (Li and Liu, 2006). Dots are earthquakes. See color plates section.

models (Figure 6.13b). In those models, continuous plastic creeping is used to simulate long-term slip on the SAF. Hence our multi-timescale faulting model has taken a successful step toward integrating short- and long-term fault processes.

## 6.5 Discussion and conclusions

Although lithospheric dynamic models focusing on a specific spatial and temporal scale has been proven helpful in understanding the basic physics of the complex lithospheric dynamics, this approach becomes inadequate when the problem involves interaction of different physics operating at different timescales. As we have shown here, one such problem is faulting.

Faults are essential for lithospheric deformation. In the paradigm of plate tectonics, deformation results mainly from plate interaction at plate boundaries, which are lithosphere-cutting faults. Deformation within the lithosphere is dominated by faults in the upper crust, from where most direct observational data are derived. On the other hand, faults pose a fundamental challenge to numerical models, which are mostly based on continuum mechanics. While many numerical schemes have been devised to simulate faults in a continuum mechanical model, most of such schemes are limited to capture fault behavior at a specific timescale.

The need for integrating faulting process at different timescales has long been realized, and the rapid growth of space-based geodetic data makes it imperative to gain a better understanding of how short-term crustal deformation relates to long-term tectonics. This daunting task can be made somewhat easier by the supercomputing power provided by affordable Beowulf PC clusters and other parallel computers, which allow intensive computations that are prohibitively difficult for traditional computers.

With the help of parallel computing, we have explored some aspects of the multi-timescale lithospheric deformation in western USA. We found that considering the timescale-dependent driving forces and lithospheric rheology is helpful in understanding the nature of lithospheric dynamics and reconciling apparent discordances between the geodetic and geological observations of crustal deformation. These models are rudimentary, with many features yet to be developed. The subcontinental scale model for western USA has the potential of incorporating detailed 3-D lithospheric structures and crustal kinematics that will emerge from the EarthScope (www.earthscope.org) and other ongoing research studies (Liu *et al.*, 2007). However, the current model (Yang and Liu, 2010) does not include plastic deformation, thus cannot fully simulate strain localization, an important feature of continental tectonics. We have since improved this model to include plastic strain in off-fault crust (Liu *et al.*, 2010). Our multi-timescale faulting model takes a major step toward integrating short- and long-term faulting models. However, some important processes, such as dynamic rupture propagation and state and rate dependence of friction (Dieterich, 1979, 1994), are not included in the current model. This is because some of these processes are computationally costly, even with parallel computing; others are not well understood or poorly constrained.

Our major findings of multi-timescale lithospheric deformation in the western United States may be summarized as following:

(1) The timescale-dependent traction along the San Andreas Fault and lithospheric rheology may explain much of the apparent discordance between crustal deformation in western USA as indicated by the GPS data and that reconstructed from geological record in the past few million years. Inter-seismic locking along the SAF transmits high shear tractions into the western margin of the North American plate. Over the short term of GPS measurements, the lithospheric rheology is close to being viscoelastic, so the influence of gravitational potential energy is limited. The dominant driving force is the plate boundary force on the San Andreas Fault, hence the GPS data show strong dextral motion of western USA over the broadly defined SAF plate boundary zone. However, stress drops during earthquakes and aseismic slips. Hence over geological timescales (in millions of years), the average plate boundary force is lower than that required to explain the GPS data. Furthermore, the long-term lithospheric rheology is closer to being viscous than viscoelastic, which means a greater influence of gravitational potential energy, and the transient short-term strain would not show. This explains the

dominantly extensional deformation in western USA as indicated by the geological records.

(2) The present-day crustal shortening across the Cascadia subduction is almost entirely transient, and will be restored by future earthquakes and asiesmic slip in the subduction zone. The lack of permanent crustal shortening is primarily due to the weak trench coupling, perhaps because of the thick sediments in the trench that, by releasing water and volatiles, lubricate the subduction zone and reduce the effective normal stress on the plate interface.

(3) The present-day crustal deformation and stress state in the San Andreas Fault plate boundary zone are the cumulative results of long-term tectonic loading and short-term stress perturbation by previous slip events. Thus the GPS velocity field is a snapshot of the evolving crustal strain field that results from long-term tectonic loading, inter-seismic locking, and post-seismic relaxation from previous earthquakes. The geometry of the SAF has an important influence on the regional stress distribution and rupture patterns, with more regular ruptures along the relatively straight segments of the central SAF and more complex patterns along the restraining bends. These restraining bends, especially the Big Bend in southern California, causes high background shear stress around them. The importance of such background stress variations is indicated by their spatial correlation with the distribution of earthquakes in southern California.

## Acknowledgements

This work was supported by NSF grants EAR0225546, EAR0635574, and EAR0948620. Liu's collaboration with the Graduate University of Chinese Academy of Sciences (CAS) is supported by the CAS and the NSF/PIRE program. We thank Randy Keller for his editing.

## References

Ainsworth, M. (1996). A preconditioner based on domain decomposition for h-p finite-element approximation on quasi-uniform meshes. *SIAM Journal on Numerical Analysis*, **33**: 1358–1376.

Allmendinger, R. W., Jordan, T. E., Kay, S. M., and Isacks, B. L. (1997). The evolution of the Altiplano-Puna plateau of the Central Andes. *Annual Reviews of Earth and Planetary Science*, **25**: 139–174.

Anderson, G., Aagaard, B., and Hudnut, K. (2003). Fault interactions and large complex earthquakes in the Los Angeles area. *Science*, **302**: 1946–1949.

Argus, D. F. and Gordon, R. G. (1991). Current Sierra Nevada-North America motion from very long baseline interferometry: Implications for the kinematics of the Western United States. *Geology (Boulder)*, **19**: 1085–1088.

Armstrong, R. L. and Ward, P. (1991). Evolving geographic patterns of Cenozoic magmatism in the North American Cordillera: The temporal and spatial association of magmatism and metamorphic core complexes. *Journal of Geophysical Research*, **96**: 13 201–13 224.

Atwater, T. (1970). Implications of plate tectonics for the Cenozoic tectonic evolution of western North America. *Geolical Society of America Bulletin*, **81**: 3513–3536.

Atwater, T. and Stock, J. (1998). Pacific North America plate tectonics of the Neogene southwestern United States: An update. *International Geology Review*, **40**: 375–402.

Axelsson, O. and Larin, M. (1998). An algebraic multilevel iteration method for finite element matrices. *Journal of Computational and Applied Mathematics*, **89**: 135–153.

Axen, G. J., Taylor, W., and Bartley, J. M. (1993). Space-time patterns and tectonic controls of Tertiary extension and magmatism in the Great Basin of the western United States. *Geological Society of America Bulletin*, **105**: 56–72.

Becker, T. W., Hardebeck, J. L., and Anderson, G. (2005). Constraints on fault slip rates of the southern California plate boundary from GPS velocity and stress inversions. *Geophysical Journal International*, **160**: 634–650.

Bennett, R. A., Davis, J. L., Normandeau, J. E., and Wernicke, B. P. (2002). *Space Geodetic Measurements of Plate Boundary Deformation in the Western U.S. Cordillera: Plate Boundary Zones*. Vol. 30. Washington, D.C.: American Geophysical Union, pp. 27–55.

Bennett, R. A., Davis, J. L., and Wernicke, B. P. (1999). The present-day pattern of Werstern U.S. Cordillera deformation. *Geology*, **27**: 371–374.

Bennett, R. A., Wernicke, B. P., Niemi, N. A., Friedrich, A. M., and Davis, J. L. (2003). Contemporary strain rates in the northern Basin and Range Province from GPS data. *Tectonics*, **22**: doi:10.1029/2001TC001355.

Benz, H. M., Smith, R. B., and Mooney, W. D. (1990). Crustal structure of the northwestern Basin and Range province from the 1986 PASSCAL seismic experiment. *Journal of Geophysical Research*, **95**: 21 823–21 842.

Bevis, M., Kendrick, E., Smalley, R. *et al*. (2001). On the strength of interplate coupling and the rate of back arc convergence in the central Andes: An analysis of the interseismic velocity field. *Geochemistry Geophysics Geosystems*, **2**: doi:10.129/2001GC000198.

Bhat, H. S., Dmowska, R., King, G. C. P., Klinger, Y., and Rice, J. R. (2007). Off-fault damage patterns due to supershear ruptures with application to the 2001 Mw 8.1 Kokoxili (Kunlun) Tibet earthquake. *Journal of Geophysical Research*, **112**: doi:10.1029/2006JB004425.

Bird, P. and Kong, X. (1994). Computer simulations of California tectonics confirm very low strength of major faults. *GSA Bulletin*, **106**: 159–174.

Bird, P. and Piper, K. (1980). Plane-stress finite. *Physics of the Earth Planetary Interiors*, **21**: 158–175.

Bokelmann, G. H. R. (2002). Convection-driven motion of the North American Craton: Evidence from P-wave anisotropy. *Geophysical Journal International*, **148**: 278–287.

Brace, W. F. and Kohlstedt, D. L. (1980). Limits on lithospheric stress imposed by laboratory experiments. *Journal of Geophysical Research*, **85**: 6248–6252.

Burchfiel, B. C., Cowan, D. S., and Davis, G. A. (1992). Tectonic overview of the Cordilleran orogen in western United States. In *The Cordilleran Orogen: Conterminous U.S.*, ed. B. C. Burchfiel, M. L. Zoback and P. W. Lipman. Volume G-3. Boulder, CO: Geological Society of America, pp. 407–480.

Burchfiel, C. B. and Davis, G. A. (1975). Nature and controls of cordilleran orogenesis, western United States: Extensions of an earlier synthesis. *American Journal of Science*, **275-A**: 363–396.

Chlieh, M., de Chabalier, J. B., Ruegg, J. C. *et al*. (2004). Crustal deformation and fault slip during the seismic cycle in the North Chile subduction zone, from GPS and InSAR observations. *Geophysical Journal International*, **158**: 695–711.

Choi, E. S. and Gurnis, M. (2003). Deformation in transcurrent and extensional environments with widely spaced weak zones. *Geophysical Research Letters*, **30**: doi:10.1029/2002GL016129.

Coney, P. J. (1978). Mesozoic-Cenozoic Cordilleran plate tectonics: *Memoirs of the Geological Society of America*, **152**: 33–50.

Coney, P. J. and Harms, T. A. (1984). Cordilleran metamorphic core complexes: Cenozoic extensional relics of Mesozoic compression. *Geology*, **12**: 550–554.

Cooke, M. L. and Kameda, A. (2002). Mechanical fault interaction within the Los Angeles Basin: A two-dimensional analysis using mechanical efficiency. *Journal of Geophysical Research – Solid Earth*, **107**.

Davis, J. L., Wernicke, B. P., Bisnath, S., Niemi, N. A., and Elosegui, P. (2006). Subcontinental-scale crustal velocity changes along the Pacific-North America plate boundary. *Nature*, **441**: 1131–1134.

Dickinson, W. R. (2002). The basin and range province as a composite extensional domain. *International Geology Review*, **44**: 1–38.

Dieterich, J. (1979). Modeling of rock friction: 1. Experimental results and constitutive equations. *Journal of Geophysical Research*, **84**.

Dieterich, J. (1994). A constitutive law for rate of earthquake production and its application to earthquake clustering. *Journal of Geophysical Research*, **99**: 2601–2618.

Dixon, T. H., Miller, M., Farina, F., Wang, H., and Johnson, D. (2000). Present-day motion of the Sierra Nevada block and some tectonic implications for the Basin and Range Province, North American Cordillera. *Tectonics*, **19**: 1–24.

Dixon, T. H., Norabuena, E., and Hotaling, L. (2003) Paleoseismology and Global Positioning System: Earthquake-cycle effects and geodetic versus geologic fault slip rates in the Eastern California shear zone. *Geology (Boulder)*, **31**: 55–58.

Dixon, T. H., Robaudo, S., Lee, J., and Reheis, M. C. (1995). Constraints on present-day Basin and Range deformation from space geodesy. *Tectonics*, **14**: 755–772.

Dokka, R. K. and Travis, C. J. (1990). Role of the eastern California shear zone in accommodating Pacific-North American plate motion. *Geophysical Research Letters*, **17**: 1323–1326.

Dragert, H., Wang, K., and James, T. S. (2001). A silent slip event on the deeper Cascadia subduction interface. *Science*, **292**: 1525–1528.

Du, Y. and Aydin, A. (1996). Is the San Andreas big bend responsible for the Landers earthquake and the Eastern California shear zone? *Geology*, **24**: 219–222.

Duan, B. and Oglesby, D. D. (2006). Heterogeneous fault stresses from previous earthquakes and the effect on dynamics of parallel strike-slip faults. *Journal of Geophysical Research*, **111**: doi:10.1029/2005JB004138.

England, P. and Molnar, P. (1997). Active deformation of Asia: From kinematics to dynamics. *Science*, **278**: 647–650.

England, P. C. and McKenzie, D. P. (1982). A thin viscous sheet model for continental deformation. *Geophysical Journal, Royal Astronomical Society*, **70**: 295–321.

Flesch, L. M., Holt, W. E., Haines, A. J., and Shen, T. B. (2000). Dynamics of the Pacific-North American plate boundary in the Western United States. *Science*, **287**: 834–836.

Flueh, E. R., Fisher, M. A., Bialas, J. *et al*. (1998). New seismic images of the cascadia subduction zone from cruise SO 108-ORWELL. *Tectonophysics*, **293**: 69–84.

Freed, A. M., Ali, S. T., and Burgmann, R. (2007). Evolution of stress in southern California for the past 200 years from coseismic, postseismic, and interseismic processes. *Geophysical Journal International*, **169**: 1164–1179.

Friedrich, A. M., Wernicke, B. P., Niemi, N. A., Bennett, R. A., and Davis, J. L. (2003). Comparison of geodetic and geological data from the Wasatch region, Utah, and implications for the spectral characteristics of Earth deformation at periods of 10 to 10 million years. *Journal of Geophysical Research*, **108**: doi:10.1029/2001JB000682.

Gabrielse, H. and Yorath, C. J. (1992). *Tectonic Synthesis: Geology of the Cordilleran Orogen in Canada*, ed. H. Gabrielse and C. J. Yorath. Vol. G-2. Ottawa: Geological Survey of Canada, pp. 679–705;.

Gan, W. J., Svarc, J. L., Savage, J. C., and Prescott, W. H. (2000). Strain accumulation across the Eastern California Shear Zone at latitude 36 degrees 30′ N. *Journal of Geophysical Research*, **105**: 16 229–16 236.

Goodman, R. E., Taylor, R. L., and Brekke, T. L. (1968). *A Model for the Mechanics of Jointed Rock*. Vol. 94. New York: American Society of Civil Engineers, pp. 637–659.

Gordon, R. G. and Stein, S. (1992). Global tectonics and space geodesy. *Science*, **256**: 333–342.

Hamilton, W. and Myers, W. B. (1966). Cenozoic tectonics of the western United States. *Reviews of Geophysics*, **5**: 509–549.

He, J., Liu, M., and Li, Y. (2003). Is the Shanxi rift of northern China extending? *Geophysical Research Letters*, **30**: doi:10.1029/2003GL018764.

Hu, Y., Wang, K., He, J., Klotz, J., and Khazaradze, G. (2004). Three-dimensional viscoelastic finite element model for postseismic deformation of the great 1960 Chile earthquake. *Journal of Geophysical Research – Solid Earth*, **109**: doi:10.1029/2004JB003163.

Isacks, B. L. (1988). Uplift of the central Andean plateau and bending of the Bolivian orocline. *Journal of Geophysical Research*, **93**: 3211–3231.

Jaeger, J. C., Cook, N. G. W., and Zimmerman, R. W. (2007). *Fundamentals of Rock Mechanics*. Oxford: Blackwell Publishing, pp. 261–262.

Jeanloz, R. and Morris, S. (1986). Temperature distribution in the crust and mantle. *Annual Reviews of Earth and Planetary Science*, **14**: 377–415.

Jones, C. H., Unruh, J. R., and Sonder, L. J. (1996). The role of gravitational potential energy in active deformation in the southwestern United States. *Nature*, **381**: 37–41.

Kirby, S. H. and Kronenberg, A. K. (1987). Rheology of the lithosphere: Selected topics. *Reviews of Geophysics*, **25**: 1219–1244.

Kong, X. and Bird, P. (1996). Neotectonics of Asia: Thin-shell finite-element models with faults. In *The Tectonic Evolution of Asia: World and Regional Geology Series*, ed. A. Yin and T. M. Harrison. Cambridge: Cambridge University Press, pp. 18–36.

Lamb, S. (2006). Shear stresses on megathrusts: Implications for mountain building behind subduction zones. *Journal of Geophysical Research*, **111**: doi:10.1029/2005JB003916.

Lavier, L. L. and Buck, W. R. (2002) Half graben versus large-offset low-angle normal fault: Importance of keeping cool during normal faulting. *Journal of Geophysical Research*, **107**: doi:10.1029/2001JB000513.

Lewis, J. C., Unruh, J. R., and Twiss, R. J. (2003). Seismogenic strain and motion of the Oregon coast block. *Geology*, **31**: 183–186.

Li, Q. and Liu, M. (2006). Geometrical impact of the San Andreas Fault on stress and seismicity in California. *Geophysical Research Letters*, **33**, L08302, doi:10.1029/2005GL025661.

Li, Q. and Liu, M. (2007). Initiation of the San Jacinto Fault and its interaction with the San Andreas Fault. *Pure and Applied Geophysics*, **164**: 1937–1945.

Li, Q., Liu, M., Qie, Z., and Sandvol Eric, A. (2006). Stress evolution and seismicity in the central-eastern United States: Insights from geodynamic modeling. In *Continental*

*Intraplate Earthquakes: Science, Hazard, and Policy Issues*, ed. S. Stein and S. Mazzotti. Special Paper 425. Boulder, CO: Geological Society of America, pp. 149–166, doi:110.1130/2007.2425(11).

Li, Q., Liu, M., and Zhang, H. (2009). A 3D visco-elasto-plastic model for simulating long-term slip on non-planar faults. *Geophysical Journal International*, **176**: 293–306.

Lipman, P. W., Prostka, W. J., and Christiansen, R. L. (1972). Cenozoic volcanism and plate tectonic evolution of the western United States: Part I. Early and Middle Cenozoic. *Royal Society of London Philosophical Transactions*, **271**-A: 217–248.

Liu, M. (2001). Cenozoic extension and magmatism in the North America Cordillera: The role of gravitational collapse. *Tectonophysics*, **342**: 407–433.

Liu, M. and Shen, Y. (1998). Crustal collapse, mantle upwelling, and Cenozoic extension in the North American Cordillera. *Tectonics*, **17**: 311–321.

Liu, M. and Yang, Y. (2003). Extensional collapse of the Tibetan Plateau: Results of three-dimensional finite element modeling. *Journal of Geophysical Research – Solid Earth*, **108**: 2361, doi:10.1029/2002JB002248.

Liu, M., Wang, H., and Li, Q. (2010). Inception of the Eastern California Shear Zone and evolution of the San Andreas Fault plate boundary zone: From kinematics to geodynamics. *Journal of Geophysical Research*, **115**, B07401, doi:10.1029/2009JB007055.

Liu, M., Yang, Y., Li, Q., and Zhang, H. (2007). Parallel computing of multi-scale continental deformation in the western United States: Preliminary results. *Physics of the Earth and Planetary Interiors*, **163**: 35–51.

Liu, M., Yang, Y., Stein, S., Zhu, Y., and Engeln, J. (2000). Crustal shortening in the Andes: Why do GPS rates differ from geological rates? *Geophysical Research Letters*, **27**: 3005–3008.

Liu, Z. and Bird, P. (2002). North America Plate is driven westward by lower mantle flow. *Geophysical Research Letters*, **29**: doi:10.1029/2002GL016002.

Luo, G. and Liu, M. (2009). Why short-term crustal shortening leads to mountain building in the Andes, but not in the Cascadia? *Geophysical Research Letters*, **36**: L08301, doi:10.1029/2009GL037347.

Luo, G. and Liu, M. (2010). Stress evolution and fault interactions before and after the 2008 Great Wenchuan Earthquake. *Tectonophysics*, **491**, 127–140, doi:110.1016/j.tecto.2009.1012.1019.

Lv, X., Zhao, Y., Huang, X. Y., Xia, G. H., and Wang, Z. J. (2006). An efficient parallel/unstructured-multigrid preconditioned implicit method for simulating 3D unsteady compressible flows with moving objects. *Journal of Computational Physics*, **215**: 661–690.

Mazzotti, S., Dragert, H., Henton, J. *et al*. (2003). Current tectonics of northern Cascadia from a decade of GPS measurements. *Journal of Geophysical Research*, **108**: doi:10.1029/2003JB002653.

McCaffrey, R., Qamar, A. I., King, R. W. *et al*. (2007). Fault locking, block rotation and crustal deformation in the Pacific Northwest. *Geophysical Journal International*, **169**: 1315–1340.

McClusky, S. C., Bjornstad, S. C., Hager, B. H. *et al*. (2001). Present day kinematics of the Eastern California Shear Zone from a geodetically constrained block model. *Geophysical Research Letters*, **28**: 3369–3372.

McQuarrie, N. and Wernicke, B. P. (2005). An animated tectonic reconstruction of southwestern North America since 36 Ma. *Geosphere*, **1**: 147–172.

Meade, B. J. (2005). Block models of crustal motion in southern California constrained by GPS measurements. *Journal of Geophysical Research*, **110**, B03403, doi:10.1029/2004JB003209.

Meade, B. J. and Hager, B. H. (2004). Viscoelastic deformation for a clustered earthquake cycle. *Geophysical Research Letters*, **31**, L10610, doi:10.1029/2004GL019643.

Miller, M. M., Johnson, D. J., Dixon, T. H., and Dokka, R. K. (2001a). Refined kinematics of the Eastern California shear zone from GPS observations, 1993–1998. *Journal of Geophysical Research*, **106**: 2245–2263.

Miller, M. M., Johnson, D. J., Rubin, C. M. *et al*. (2001b). GPS-determination of along-strike variation in Cascadia margin kinematics; implications for relative plate motion, subduction zone coupling, and permanent deformation. *Tectonics*, **20**: 161–176.

Minster, J. B. and Jordan, T. H. (1987). Vector constraints on Western U.S. deformation from space geodesy, neotectonics, and plate motions. *Journal of Geophysical Research – Solid Earth and Planets*, **92**: 4798–4804.

Murray, M. H. and Lisowski, M. (2000). Strain accumulation along the Cascadia subduction zone. *Geophysical Research Letters*, **27**: 3631–3634.

Okada, Y. (1985). Surface deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America*, **75**: 1135–1154.

Parsons, T. (2002). Post-1906 stress recovery of the San Andreas fault system calculated from three-dimensional finite element analysis. *Journal of Geophysical Research – Solid Earth*, **107**.

Pollitz, F. F. (1997). Gravitational viscoelastic postseismic relaxation on a layered spherical Earth. *Journal of Geophysical Research – Solid Earth and Planets*, **102**: 17 921–17 941.

Pollitz, F. F. (2003). The relationship between the instantaneous velocity field and the rate of moment release in the lithosphere. *Geophysical Journal International*, **153**: 596–608.

Pollitz, F. F., Banerjee, P., Burgmann, R., Hashimoto, M., and Choosakul, N. (2006). Stress changes along the Sunda trench following the 26 December 2005 Sumatra-Andaman and 28 March 2005 Nias earthquake. *Geophysical Research Letters*, **33**: doi:10.1029/2005gl024558.

Pope, D. C. and Willett, S. D. (1998). Thermal-mechanical model for crustal thickening in the central Andes driven by ablative subduction. *Geology*, **26**: 511–514.

Quarteroni, A. and Valli, A. (1999). *Domain Decomposition Methods for Partial Differential Equations*. Oxford, New York: The Clarendon Press (Oxford University Press), 376pp.

Ranalli, G. (1995). *Rheology of the Earth*. London: Chapman & Hall, 413pp.

Savage, J. C. (1983). A dislocation model of strain accumulation and release at a subduction zone. *Journal of Geophysical Research*, **88**: 4984–4996.

Savage, J. C. and Burford, R. (1973). Geodetic determination of relative plate motion in central California. *Journal of Geophysical Research*, **78**: 832–845.

Savage, J. C., Gan, W. J., and Svarc, J. L. (2001). Strain accumulation and rotation in the Eastern California Shear Zone. *Journal of Geophysical Research*, **106**: 21 995–22 007.

Shen, T. B., Holt, W. E., and Haines, A. J. (1999). Deformation kinematics in the Western United States determined from Quaternary fault slip rates and recent geodetic data. *Journal of Geophysical Research – Solid Earth and Planets*, **104**: 28 927–28 956.

Silver, P. G. and Holt, W. E. (2002). The mantle flow field beneath western North America. *Science*, **295**: 1054–1058.

Sobolev, S. V. and Babeyko, A. Y. (2005). What drives orogeny in the Andes? *Geology*, **33**: 617–620.

Sonder, L. J. and Jones, C. H. (1999). Western United States extension: How the west was widened. *Annual Reviews of Earth and Planetary Science*, **27**: 417–462.

Stein, R. S., Barka, A. A., and Dieterich, J. H. (1997). Progressive failure on the North Anatolian fault since 1939 by earthquake stress triggering. *Geophysical Journal International*, **128**: 594–604.

Stein, S., Tomasello, J., and Newman, A. (2003). Should Memphis build for California's earthquakes? *Eos Transactions AGU*, **84**: 177, 184–185.

Stewart, J. H. (1978). Basin and Range structure in western North America: A review. In *Cenozoic Tectonics and Regional Geophysics of the Western Cordillera*, ed. R. B. Smith and G. L. Eaton. Vol. **152**. Boulder, CO: Memoirs of the Geological Society of America, pp. 1–31.

Thatcher, W., Foulger, G. R., Julian, B. R. *et al*. (1999). Present-day deformation across the Basin and Range Province, Western United States. *Science*, **283**: 1714–1718.

Thompson, G. A., Catchings, R., Goodwin, E. *et al*. (1989). Geophysics of the western Basin and Range province. In *Geophysical Framework of the Continental United States*, ed. L. C. Pakiser and W. D. Mooney. Vol. **172**. Boulder, CO: Memoirs of the Geological Society of America, pp. 177–203.

Wang, K. and He, J. (1999). Mechanics of low-stress forearcs: Nankai and Cascadia. *Journal of Geophysical Research*, **104**: 15 191–15 205.

Wang, K., Mulder, T., Rogers, G. C., and Hyndman, R. D. (1995). Case for very-low coupling stress on the Cascadia Subduction Fault. *Journal of Geophysical Research*, **100**: 12 907–12 918.

Wang, K., Wells, R., Mazzotti, S., Hyndman, R. D., and Sagiya, T. (2003). A revised dislocation model of interseismic deformation of the Cascadia subduction zone. *Journal of Geophysical Research – Solid Earth*, **108**: doi:10.1029/2001JB001227.

Ward, S. N. (1990). Pacific-North America plate motions: New results from very long baseline interferometry. *Journal of Geophysical Research – Solid Earth and Planets*, **95**: 21 965–21 981.

Wdowinski, S. and Bock, Y. (1994). The evolution of deformation and topography of high elevated plateaus: 2. Application to the Central Andes. *Journal of Geophysical Research*, **99**: 7121–7130.

Weldon, R., Scharer, K., Fumal, T., and Biasi, G. (2004). Wrightwood and the earthquake cycle: What a long recurrence record tells us about how faults work. *GSA Today*, **14**: 4–10.

Wells, R. E. and Simpson, R. W. (2001). Northward migration of the Cascadia forearc in the northwestern US and implications for subduction deformation. *Earth Planets and Space*, **53**: 275–283.

Wells, R. E., Weaver, C. S., and Blakely, R. J. (1998). Fore-arc migration in Cascadia and its neotectonic significance. *Geology (Boulder)*, **26**: 759–762.

Wernicke, B., Christiansen, R. L., England, P. C., and Sonder, L. J. (1987). Tectonomagmatic evolution of Cenozoic extension in the North American Cordillera. In *Continental Extensional Tectonics*, ed. M. P. Coward, J. F. Dewey and P. L. Hancock. Vol. **28**, Geological Society of London, Special Publication, pp. 203–221.

Wernicke, B., Clayton, R., Ducea, M. *et al*. (1996). Origin of high mountains in the continents: The southern Sierra Nevada. *Science*, **271**: 190–193.

Wernicke, B., Friedrich, A. M., Niemi, N. A., Bennett, R. A., and Davis, J. L. (2000). Dynamics of plate boundary fault systems from Basin and Range Geodetic Network (BARGEN) and geological data. *GSA Today*, **10**: 1–7.

Wernicke, B. P. (1992). Cenozoic extensional tectonics of the U.S. Cordillera. In *The Cordilleran Orogen: Conterminous U.S.*, ed. B. C. Burchfiel, M. L. Zoback and P. W. Lipman. Vol. **G-3**. Boulder, CO: Geological Societyof America, pp. 553–581.

Xing, H. L., Mora, P., and Makinouchi, A. (2004). Finite element analysis of fault bend influence on stick-slip instability along an intra-plate fault. *Pure and Applied Geophysics*, **161**: 2091–2102.

Yang, Y. and Liu, M. (2009). Algorithms for optimizing rheology and loading forces in finite element models of lithospheric deformation. In *Advances in Geocomputing*, ed. H. Xing. Lecture Notes in Earth Sciences, Vol. 119. Berlin: Springer-Verlag, pp. 119–138.

Yang, Y. and Liu, M. (2010). What drives short- and long-term crustal deformation in the southwestern United States? *Geophysical Research Letters*, **37**: doi:10.1029/2009GL041598.

Yang, Y., Liu, M., and Stein, S. (2003). A 3D geodynamic model of lateral crustal flow during Andean mountain building. *Geophysical Research Letters*, **30**: doi:10.1029/2003GL18.

Zhang, H., Liu, M., Shi, Y. *et al*. (2007). Toward an automated parallel computing environment for geosciences. *Physics of the Earth and Planetary Interiors*, **163**: 2–22.

Zoback, M. D., Anderson, R. E., and Thompson, G. A. (1981). Cenozoic evolution of the state of stress and style of tectonism of the Basin and Ranges Province of the western United States. *Royal Societyof London Philosophical Transactions*, Serial A 300: 407–434.

Zoback, M. D. and Zoback, M. L. (1981). State of stress and intraplate earthquakes in the United States. *Science*, **213**: 96–104.

# 7

# Next-generation plate-tectonic reconstructions using GPlates

JAMES A. BOYDEN, R. DIETMAR MÜLLER, MICHAEL GURNIS, TROND H. TORSVIK, JAMES A. CLARK, MARK TURNER, HAMISH IVEY-LAW, ROBIN J. WATSON, AND JOHN S. CANNON

## 7.1 Introduction

Plate tectonics is the kinematic theory that describes the large-scale motions and events of the outermost shell of the solid Earth in terms of the relative motions and interactions of large, rigid, interlocking fragments of lithosphere called *tectonic plates*. Plates form and disappear incrementally over time as a result of tectonic processes. There are currently about a dozen major plates on the surface of the Earth, and many minor ones. The present-day configuration of tectonic plates is illustrated in Figure 7.1. As the interlocking plates move relative to each other, they interact at plate boundaries, where adjacent plates collide, diverge, or slide past each other. The interactions of plates result in a variety of observable surface phenomena, including the occurrence of earthquakes and the formation of large-scale surface features such as mountains, sedimentary basins, volcanoes, island arcs, and deep ocean trenches. In turn, the appearance of these phenomena and surface features indicates the location of plate boundaries. For a detailed review of the theory of plate tectonics, consult Wessel and Müller (2007).

A plate-tectonic reconstruction is the calculation of positions and orientations of tectonic plates at an instant in the history of the Earth. The visualization of reconstructions is a valuable tool for understanding the evolution of the systems and processes of the Earth's surface and near subsurface. Geological and geophysical features may be "embedded" in the simulated plates, to be reconstructed along with the plates, enabling a researcher to trace the motions of these features through time. Even a single static reconstruction can reveal an illuminating configuration of surface or subsurface features. A time-sequence of reconstructions may be used to animate the motions of plates, producing kinematic and other time-derivative information, which in turn can provide insight into the geodynamic processes of the near subsurface and deeper Earth.

Plate-tectonic reconstructions are useful in a variety of contexts, such as research in geology, geophysics, and paleobiology; exploration for mineral and hydrocarbon

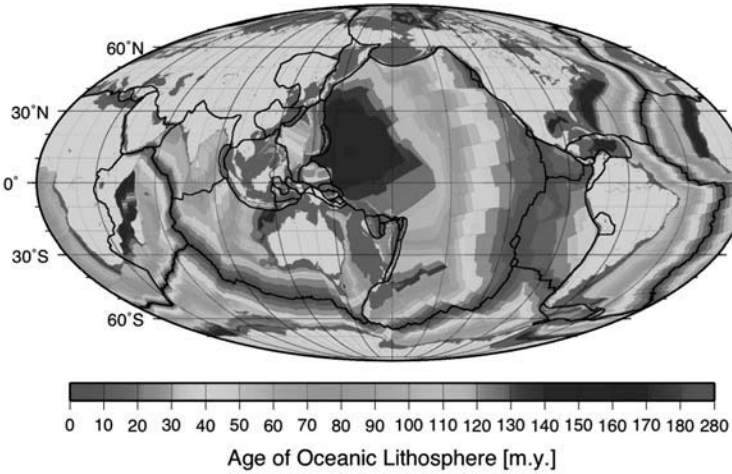



Figure 7.1. The age of the oceanic lithosphere, with plate boundaries shown in black and continents in light gray. See color plates section.

resources; modeling climate change and sea-level change over geological time-periods, including reconstructing paleotopography, paleovegetation patterns, and paleosurface temperatures to constrain these models (Herold *et al*., 2008; Herold *et al*., 2009); and providing kinematic boundary conditions for geodynamic models (e.g., Liu *et al*., 2008). The ability to reconstruct continents is crucial in any software that is intended to study and explore the distant geological past.

The capability of modern, open-source software has lagged behind the requirements for plate-tectonic reconstructions, which we have attempted to overcome with GPlates,[1] a desktop software application for the calculation and interactive visualization of plate-tectonic reconstructions. Building upon this foundation of reconstructions, GPlates offers a suite of integrated, interactive functionality for the study of the geological history of the Earth: the visualization of plate-tectonic processes and the dynamics of the upper mantle; the incorporation, reconstruction, and editing of geological, geophysical, and paleogeographic data within a plate-tectonic context; and the real-time graphical manipulation of the plate-motion models which dictate reconstructions. The main window of the GPlates user-interface is illustrated in Figure 7.2 and Figure 7.3.

GPlates can display a reconstruction for a single instant in geological time (in Figure 7.2 and Figure 7.3) or animate a sequence of reconstructions over a user-specified geological time-period. Time-derivative information such as plate-motion velocity fields and flow-lines may be calculated on-the-fly. A user may view reconstructions on either a "flattened-out" geographic map in a variety of 2-D

[1] GPlates web site: www.gplates.org.

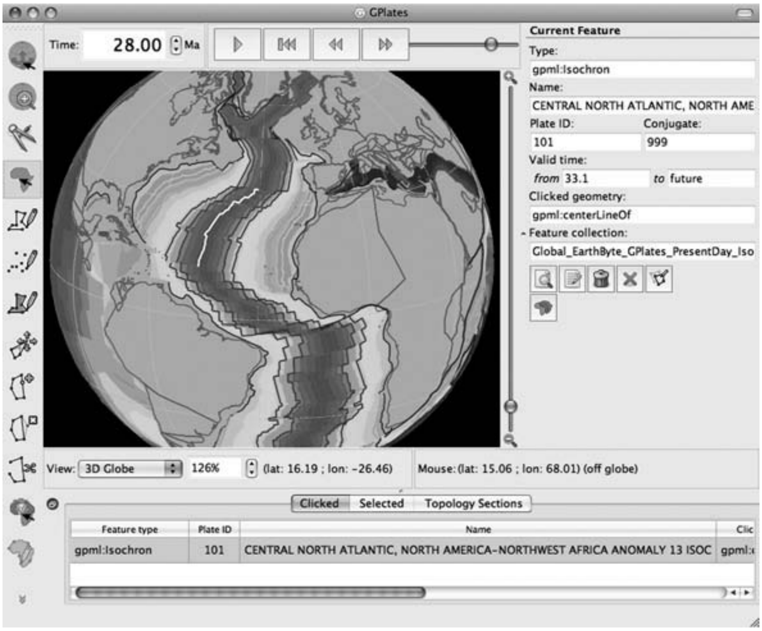



Figure 7.2. The GPlates main window on MacOS X, displaying a reconstruction at 28 Ma (million years ago). The user has clicked on an Isochron feature. See color plates section.

map projections, or an orthographic projection of a 3-D globe (as if observing the Earth from space). Reconstructions may be viewed and manipulated interactively, or they may be exported from GPlates as instantaneous data snapshots (for importation into other geospatial software programs) or as 2-D vector-graphics images in the Scalable Vector Graphics (SVG) format.

The goals motivating the development of GPlates include the need to visualize the reconstruction of geological, geophysical, and paleogeographic data in a variety of formats, including raster data;[2] to link plate kinematics to geodynamic models; to serve as an interactive client in a grid-computing network; to facilitate the production of high-quality paleogeographic maps; to serve as a freely shared desktop application and software platform, useful both in its own right (as a ready-to-use, stand-alone software product) and as the foundation for new features and specialized functionality to be built on top of the GPlates computational and visualization infrastructure.

GPlates is developed by an international team of scientists and software developers at the EarthByte Project[3] in the School of Geosciences at the University of

[2] A raster is a two-dimensional rectangular grid of values distributed over a spatial extent. Rasters will be discussed in greater detail below.

[3] EarthByte web site: www.earthbyte.org.

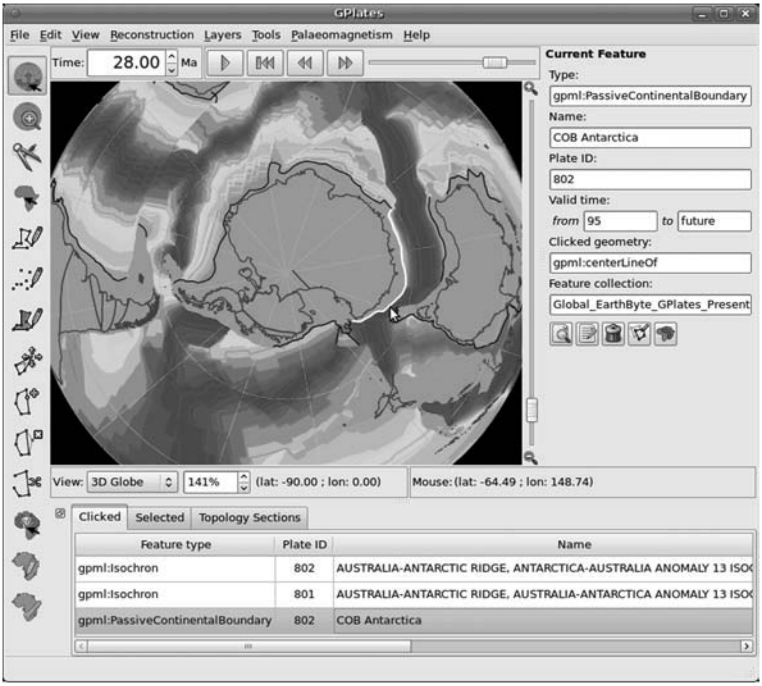



Figure 7.3. The GPlates main window on Linux, displaying a reconstruction at 28 Ma. The view of the globe may be re-oriented interactively by dragging on the globe with the mouse. See color plates section.

Sydney, the Division of Geological and Planetary Sciences[4] (GPS) at Caltech, and the Center for Geodynamics at the Norwegian Geological Survey[5] (NGU). Freely sharable, royalty-free data for use with GPlates are made available by collaborating scientists at these three institutions.

This chapter introduces GPlates in two parts: Section 7.2 is a survey of the advances in functionality and data-representation that GPlates brings to the domain of interactive plate-tectonic reconstruction software; Section 7.3 is a conceptual overview of the plate-motion model that GPlates uses to calculate reconstructions. Together, these two sections offer an introduction to the foundations and innovations of GPlates.

## 7.2 Advancing the state of plate-tectonics software

Purpose-built interactive visualization software for plate-tectonic reconstructions has existed since the 1980s. The functionality of such software has generally

[4] GPS web site: www.gps.caltech.edu. [5] NGU web site: www.ngu.no/no/.

included a number of key features: the ability to display a reconstruction for a user-specified geological time; the ability to calculate new reconstructions interactively, by "dragging" reconstructed plates to new locations using the mouse; the ability to control the visual presentation of the reconstructed data (generally a choice of color scheme and map projection, as well as fine-grained control over the visibility of different types of data); and the ability to interact with the data, to query or edit attributes, by "clicking" the visible data with the mouse.

GPlates offers all the familiar features of existing interactive plate-tectonics software, while incorporating a number of advances such as smooth, multiple-frame-per-second animation of reconstructions; interactive graphical manipulation of reconstructions, with continuous real-time feedback and automatic in-place modification of the loaded plate-motion model; seamless display of raster data on the surface of the globe; and automated post-processing of reconstructions, enabling new data to be computed on the fly (such as time-dependent, shape-changing plate-boundary polygons assembled from the geometries of other reconstructed data; and plate-motion velocity fields calculated within these plate-boundaries). These enhancements are presented within a modern, intuitive graphical user interface, which runs on Microsoft Windows, Apple MacOS X, and Linux.

### *7.2.1 Fast reconstructions and responsiveness*

GPlates offers unprecedented responsiveness as a result of its computationally efficient calculation of reconstructions and its effective use of hardware-accelerated computer graphics. Calculating a reconstruction is a computationally intensive operation. Due to the choice of programming language (C++) and judicious selection of the internal representations of geometries and geometric reconstruction operators (3-D unit-vectors (Schettino, 1999) and unit-quaternions (Altmann, 1986), respectively), GPlates is able to reconstruct even large datasets in sub-second times, enabling smooth, real-time animations of reconstructions.

Similarly, the use of the cross-platform, hardware-accelerated 3-D graphics library OpenGL[6] enables efficient graphics performance. The orthographic globe projection is particularly fast, since the internal representation of geometries (as 3-D unit-vectors) is also an efficient input format for OpenGL in this projection. When viewing the Earth as an orthographic globe, the OpenGL graphics updates are sufficiently fast that when the user "drags" the mouse to reorient the rendered globe, the globe is updated without a perceptible delay. This method of interaction – manipulating a displayed object like a real-world object, with rapid, incremental

[6] OpenGL web site: www.opengl.org.

feedback – is called *direct manipulation* (Shneiderman, 1983), and is one of the factors which makes a user interface seem more "natural" or "intuitive."

### 7.2.2 User-friendly graphical editing capability

GPlates extends the familiar graphical editing capability of GIS (Geographic Information System) software to the realm of plate-tectonic reconstructions. Operations such as tweaking geometry shapes and manually fine-tuning reconstructed plate positions are interactive graphical operations in the direct manipulation style. As the user manipulates an object displayed on-screen, GPlates provides continuous, immediate feedback about the result of the operation – for example, as the user drags a vertex of a geometry; a "ghost" of the resulting geometry is drawn and updated in "real-time." When the user releases the mouse button to end the drag operation, the ghost geometry replaces the previous version of the geometry, and the modifications are incorporated automatically into the loaded data. GPlates extends the direct manipulation metaphor to enable reconstructed plate positions to be modified in the same manner: A user can drag a plate into a new position to specify the new reconstruction of that plate at that geological time. GPlates will calculate new plate-motion parameters and incorporate them into the plate-motion model automatically.

New geometries may be digitized in GPlates just as they are in familiar interactive drawing software. However, due to plate-motion, a geometry on a plate will appear at different positions on the Earth at different geological times – so the position at which a geometry should be digitized is a function of geological time. Fortunately, a user of GPlates need not worry about this issue: GPlates allows a user to digitize a geometry at any geological time; the user should simply digitize the geometry at the appropriate position for the current reconstruction time. When the digitization is complete, GPlates will prompt the user for the plate on which the geometry should be located; the coordinates of the geometry will then be adjusted automatically, to compensate correctly for the reconstruction of that plate at that time.

### 7.2.3 Internal geometry representation

In addition to enabling fast reconstruction and graphics performance, the internal representation of geometries in GPlates (as 3-D unit-vectors) avoids the computational problem of "indeterminate" values at the geographic north or south poles: In a 2-D (latitude, longitude) space, the North Pole has a well-defined latitude of 90°, but no determinate longitude. The alternative is to nominate a particular longitude (for example, 0°) to be the longitude of the North Pole by convention; however, this results in a discontinuity between the North Pole and all other regions of high latitude that are not on the zero-longitude meridian. At best, either of these situations must be handled by special-case code, which slows the ubiquitous geometric calculations and makes the code more complex (and thus, more difficult to maintain); at worst, these situations are handled incorrectly or not at all during geospatial calculations, leaving the program unable to process data that pass over the poles.

When 3-D unit-vectors are used to encode positions on the surface of the Earth, with the dot product used as the measure of proximity, then the North Pole has a well-defined unit-vector of (0, 0, 1), without any discontinuities around it. This representation similarly liberates GPlates from any representational or computational discontinuities when geometries cross the International Date Line (and the longitude crosses from +180° to –180°, or vice versa) – or indeed, any other meridian. As a result, GPlates is able to represent data that pass over the poles or cross the Date Line (as illustrated in Figure 7.3), without requiring any special preprocessing of the data (such as cutting geometries into multiple pieces, or inserting extra vertices at –180° and +180°), making GPlates a truly "global" plate-tectonics application.

### *7.2.4 Effective representation of continuous fields*

Interactive plate-tectonics software has traditionally modeled geological entities as strictly *vector* geometries (points, lines, curves or polygons) with associated attributes. Vector geometries are well suited to the representation of location-based measurements (using points), boundaries and contour lines (using lines or curves), and the outlines of enclosed regions (using polygons). However, a continuous field of values distributed over an area – such as the interpolated age and calculated spreading-rate of the ocean floor (Müller *et al*., 2008), or the dynamic topography of the mantle (Steinberger, 2007) – is not well represented by a spatially discrete vector geometry. Such a field is better represented in raster form.

A *raster* is a two-dimensional rectangular grid of values distributed over a spatial extent. Each element of the raster is a rectangular cell that contains some scalar or vector values. The cells are contiguous (i.e., there are no gaps between them) and non-overlapping, which makes rasters well suited to the representation of continuous fields in an unambiguous fashion. Each cell contains a sampling of the field at the location of that cell. Since the grid structure of a raster is implied, a raster is a more space-efficient representation than the vector-geometric multi-point equivalent. There are also well-known compression methods for rasters, further increasing the efficiency.

GPlates is able to display rasters on the surface of the globe – both rasters of "raw" data (in which the value in each cell is a physical value, such as the measured seafloor depth or the calculated seafloor spreading-rate) and pre-rendered color images (in which the value of each cell is the color of a pixel). The first type of raster is generally the unprocessed result of some geospatial computation or measurement; the second type of raster is either a photographic image, or a computational result or measurement that has already been processed for visualization by some other software. A raster may cover either the entire surface of the globe or a user-specified rectangular surface extent.

GPlates is able to load color images stored in the widely used JPEG image format, as well as raw data rasters stored in the widely used NetCDF format. In the near future, GPlates will be able to display georeferenced rasters in the GeoTIFF and georeferenced JPEG formats.

GPlates is able to load rasters either as single, static images that do not change as the reconstruction time is varied or as frames in a time-indexed raster sequence. The latter are loaded collectively and treated as a single "time-dependent raster" which changes as a function of geological time. The former method of loading is used for a "normal" raster that is not part of a time-sequence and is not intended to change over time, while the latter is used for a pre-prepared raster sequence, which GPlates will animate.

### *7.2.5 Precise, consistent definition of data attributes*

Representing heterogeneous real-world geospatial information can be a complex issue. At one extreme, if we attempt to express all information uniformly in terms of a fixed set of attributes, we risk both missing important attributes that are relevant to only some types of information, as well as coercing our users' view of the world to fit into those attributes. At the other extreme, if we accept any combination of attributes and attribute types, there is no guidance about which attributes are recommended, expected, or required by the software. As a result, attributes will be chosen ad hoc from project to project, hindering attempts to share data between projects. Both of these extremes have been employed in widely used plate-tectonics software.

GPlates avoids these pitfalls through the definition of its own precise abstract *information model* for data to be handled by GPlates. Lee (1999) defines an information model as "a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse." Lee continues: "The advantage of using an information model is that it can provide sharable, stable, and organised structure of information requirements for the domain context."

An information model is used to obtain "a single, integrated definition of the data within an enterprise that is unbiased toward any single application of data and independent of how the data are physically stored or accessed," which "provides a consistent definition of the meanings and interrelationship of the data in order to share, integrate, and manage the data." (Lee, 1999)

The *GPlates Geological Information Model* (GPGIM) is the formal specification of all data in GPlates. It specifies the geologic, geophysical, and paleogeographic entities that GPlates simulates, the conceptual building blocks that GPlates defines, and the processing and simulation constraints to which GPlates adheres. It serves as the primary definition of the data that must be representable in GPlates, in order for GPlates to be able to present and process plate-tectonic reconstructions as the users of GPlates require. The current GPGIM specification is available[7] on the web site of the EarthByte e-research project.[8]

Each geological and tectonic entity is described precisely in formal modeling terms, without concern for how the entity will be represented within software or encoded in data files. This liberates the description of data from concerns of file-format representation; the model of each entity has exactly the attributes it needs, no more and no less. The precise definition avoids any ambiguity about which attributes are expected or required.

The GPlates software developers have worked closely with collaborating geologists to create, refine and evolve the GPGIM specification over time. The GPGIM is currently able to represent both the legacy geological data of the collaborating scientists, as well as various next-generation data for state-of-the-art functionality being developed in GPlates (such as time-dependent geometries defined by geospatial topological networks). The GPGIM continues to be extended to incorporate new types of data – recent additions include the paleomagnetic data-types from the PMAG portal of the Magnetics Information Consortium (MagIC) database[9] on the EarthRef web site.[10]

### *7.2.6 Ability to represent next-generation data*

Traditional plate-tectonics data have consisted of vector geometries and "simple type" attributes such as strings, integers, and floating-point numbers. On their own, these building blocks are insufficient to represent the next-generation data calculated and processed by GPlates, such as attribute values which vary as functions of geological time; geological timescales; structured edit-history metadata; evolving topological plate boundaries; and dynamic flowlines.

The last two examples are particularly illustrative of the paradigm shift that is being pioneered by GPlates: Rather than containing static, pre-calculated or prepopulated geometries and geospatial values, these two data types instead encode templates for geospatial calculations. To be specific, they contain parameters for calculations that are performed in GPlates as automated post-processing of a

[7] GPGIM specification on the EarthByte web site: www.earthbyte.org/Resources/GPGIM/public/.
[8] EarthByte e-research project: www.earthbyte.org. [9] MagIC database: http://earthref.org/MAGIC/.
[10] EarthRef: http://earthref.org/.

reconstruction, to generate results that are based upon the other data loaded in GPlates, including the plate-motion model. For example, a topological plate boundary lacks any static geometry of its own; at each reconstruction time, GPlates calculates a *topological geometry* by linking together the reconstructed geometries of other data. This topological geometry is a time-dependent, shape-changing polygon that encloses a region corresponding to a tectonic plate (Gurnis *et al*., 2010). As the tectonic plate changes shape over time, so too does the polygon.

A topological plate boundary *is* "data," because a user defines the topology by selecting geometries from other reconstructed data, to specify geometry sections from which a plate-boundary polygon should be assembled. This topology configuration is stored in a data file like any other data, and the assembly of the polygon occurs only when the topological boundary data are loaded, after the component sections have been reconstructed. After the polygon has been assembled, it is rendered on the globe just like any other geometry, enabling the topological plate boundary to be visualized and manipulated just like any other geospatial data.

Given a mesh of points, GPlates can use these plate-boundary polygons to "cookie cut" the mesh points and allocate them to plates. At each reconstruction time, GPlates can then calculate a velocity for each mesh point, by first calculating the angular velocity of the relevant plate motion; the result will be a plate-motion velocity field for that reconstruction time. In addition to being visualized in GPlates, such a plate-motion velocity field may be exported, to be used as a kinematic boundary condition to a geodynamic mantle-convection model such as *CitcomS* (Spasojevic *et al*., 2009).

The key to being able to represent these next-generation data in GPlates is the ability to describe and represent structured information in attributes – in particular, the ability to nest structured information inside other structured information to an arbitrary degree. Such nesting of structured information enables GPlates to represent arbitrarily complex information by composing structures like functions are composed in mathematics. The GPGIM provides a precise specification of this structured information, and GPlates implements this specification to represent these next-generation data internally – but this does not solve the problem of how to store these data on disk between GPlates sessions, or how to transport and exchange these data between systems. For this purpose, the developers of GPlates have specified a new file format for plate-tectonics data called GPML, the *GPlates Markup Language.*[11] GPML is XML-based, which enables it to define and contain arbitrarily nested structure. In fact, the structure of GPML mirrors the structure of the GPGIM exactly: A GPML document is an XML *serialization*[12] of GPGIM data.

[11] GPML: www.gplates.org/gpml.html.

[12] *Serialization* is the process of converting an in-memory object into a sequence of bytes so it can be stored in a file or transmitted across a network.

### 7.2.7 Superior interoperability of data

The GPGIM and GPML are both based upon GML, the *Geography Markup Language* (Cox *et al*., 2004; Lake *et al*., 2004). GML is a modeling language for geographic systems, as well as an encoding specification for the serialization of GML data to XML documents. GML is intended to provide a standardized, interoperable base for defining application-specific geospatial languages. The GML specification is defined and maintained by the Open Geospatial Consortium. GML was adopted as an ISO standard in 2007.

GML defines information model building blocks including geometries; temporal primitives (time-instants and time-periods); units of measure; and coordinate reference systems (Lake, 2005). These building blocks ease the process of interoperability both by providing recognizable, well-defined data components and by standardizing the XML representations of these components, to simplify the parsing of foreign data. GML also provides an approach to modeling, the "feature-oriented" approach, which adds additional consistency to the structure of data (Zhang *et al*., 2003). The GPGIM adopts the feature-oriented modeling approach and builds upon the GML primitives, to maximize interoperability with other GML-based languages such as GeoSciML (Sen and Duffy, 2005).

The *Web Feature Service* (WFS) specification (Vretanos, 2005) defines a web service interface to query, transmit, and manipulate geospatial data across the Internet or other networks. GML is the default encoding for the transmission of geospatial data to and from WFS servers; as an application schema of GML, the GPGIM can be used to define geospatial features in WFS communications, enabling GPlates data to be queried and transmitted across WFS networks. Future GPlates development efforts will implement WFS client operations in GPlates itself, to enable GPlates to query and transmit geospatial data as an interactive client in a WFS-compatible grid-computing network.

Finally, by building upon GML, the GPGIM will also benefit from future additions to GML. For example, Cox and Richard (2005) describe a GML-compatible model of geological timescales, which will be incorporated into GPlates in the future.

### 7.2.8 Availability of the software source code

Perhaps one of the most significant advantages of GPlates is that the GPlates source code can be freely shared as open-source software. As a result, anyone can enhance GPlates with new technology as it becomes available; anyone can extend GPlates with new functionality; anyone can fix any bugs in GPlates which might be discovered; and anyone can port GPlates to new computer systems if desired.

GPlates is distributed under the GNU General Public License (GPL), version 2,[13] which grants to every person who receives a copy of GPlates the permission to reproduce, adapt or distribute GPlates as long as the source code of any resulting copies or adaptations are also made available under the same "share alike" licensing scheme.

## 7.3 Reconstructions and plate motions in GPlates

### *7.3.1 Modeling tectonic plates*

In contrast to traditional GIS software, GPlates was designed and built from the ground up to be geotemporally aware and to handle plate-tectonic reconstructions. Thus, a core concept in GPlates is that geospatial entities move around through geological time due to plate tectonics or other deep-Earth geological processes. This motion is expressed in the plate-motion model, which is used to reconstruct geospatial entities to their past locations. The plate-motion model is itself composed of data that may be loaded, saved, visualized, and edited in GPlates.

The majority of geological, geophysical, and paleogeographic entities modeled by GPlates are embedded in tectonic plates, so will move around with these plates. However, a key aspect of the GPlates plate-motion model is that there is no fundamental "tectonic plate" entity with a well-defined boundary. This is because the boundaries of tectonic plates are not always well defined; their relative motions more clearly differentiate adjacent plates. Furthermore, the shapes and extents of plates change over geological time, as plate motion causes plates to diverge and converge, with new oceanic lithosphere being produced from the mantle or old oceanic lithosphere being consumed by the mantle. Plates can stretch, compress, and otherwise deform. Finally, the set of tectonic plates has not remained constant over time: Plates can break up and fuse due to tectonic processes.

Thus, the GPlates plate-motion model instead focuses on relative motion. The key concept is the *plate ID*: an integer identifier that associates all entities that move together as if they were embedded in a single rigid plate. The plate-motion model describes the relative motions of entities with different plate IDs; every reconstructable entity is given a plate ID to describe its motion. Many of the plate IDs *do* correspond to rigid subregions of plates, but other plate IDs correspond to non-plate entities that move relative to plates (such as mid-ocean ridges and hotspots). Plate IDs are not "hard coded" or otherwise pre-determined: The allocation of plate IDs is up to the user.

[13] GPL version 2: www.gnu.org/licenses/old-licenses/gpl-2.0.html.

### 7.3.2 Finite rotations and Euler poles

The hypothesis of plate tectonics states that tectonic plates are assumed to be internally rigid (Wessel and Müller, 2007). Euler's Displacement Theorem (also called Euler's Rotation Theorem) states that any displacement of a rigid body, such that one point of the rigid body remains fixed in space, is a rotation about some axis through that fixed point (Beatty, 1986). If the Earth is approximated as a sphere, then any motion of a rigid tectonic plate on the surface of the Earth may be described geometrically as a displacement of the plate about the fixed center of the Earth, and thus, as a rotation of the plate about an axis *through* the center of the Earth. Cox and Hart (1986) refer to such rotations as *finite rotations*. Finite rotations are commonly used in plate-tectonics literature to describe plate motions on a large scale.

By analogy with the Earth's geographic north pole (the point at which the Earth's own rotation axis intersects with the surface of the planet), a common means of specifying the axis of a finite rotation is to specify the point at which the axis intersects with the surface of the Earth. This intersection point is called the *Euler pole* of the finite rotation. The Euler pole and its spherical antipode are the only two points on the surface of the Earth that are not moved to a new location by the finite rotation.

The position of the Euler pole is described conveniently as a (latitude, longitude) pair, meaning that a finite rotation can be specified by a scalar triple: the latitude of the Euler pole, the longitude of the Euler pole, and the angle of rotation. Finite rotations are commonly published in this form in plate-tectonics literature; this is also the form used to enter finite rotations into GPlates, before they are converted automatically to the unit-quaternion representation used internally.

### 7.3.3 Relative motions and the rotation hierarchy

GPlates uses a plate-motion model representation in which plate motions are described in terms of relative rotations between pairs of plates. For example, in the GPlates-compatible 2008 EarthByte Global Rotation Model (Müller *et al.*, 2008), the South American plate (also known by the abbreviation "SAM," with plate ID 201) moves relative to the African plate ("AFR," 701), as does the Antarctic plate ("ANT," 802), while the Australian plate ("AUS," 801) moves relative to the Antarctic plate. This is illustrated in Figure 7.4.

Thus, the full plate-motion model is a directed graph of rotations between pairs of plates. Every plate in the model moves relative to some other plate or is the "fixed reference frame" relative to which some other plates move. In general, large plates are described relative to other large plates, and smaller plates are described relative to their nearest large plates. This tree-like structure is called the *rotation hierarchy.*

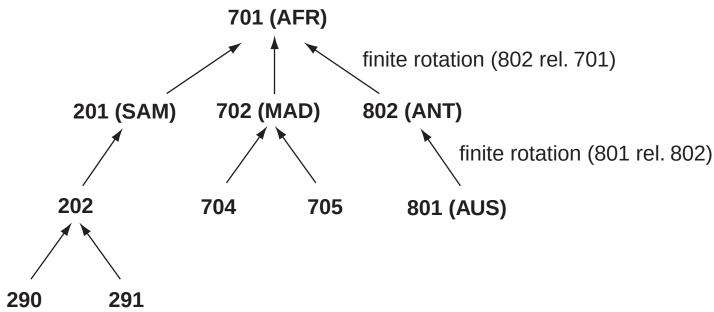



Figure 7.4. A small subset of the hierarchy of relative rotations used in GPlates.

Even if there is actually no relative motion between some pair of plates for some period of time, the plates may nevertheless be related by the identity rotation (a rotation of zero angle, which effects no change of position or orientation). The identity rotation does not determine a unique axis of rotation, and so is not strictly speaking a finite rotation, but since the identity rotation does not change the shape of the plate or move the plate relative to the center of the Earth, there is no geometric problem with allowing identity rotations in the rotation hierarchy.

### *7.3.4 Total reconstruction poles*

The relative displacement of adjacent plates may be determined quantitatively by a number of techniques, including the fitting of magnetic anomaly and fracture zone data on the ocean floor by matching conjugate isochrons (e.g., Hellinger, 1981) and the fitting of apparent polar wander curves for continents (e.g., Van der Voo, 1993). These changes may be described quantitatively using finite rotations, though Kearey and Vine emphasize (1990): "although the relative motion between two plates can be described in terms of an angular rotation about a pole, the plates do not necessarily follow the route predicted by a single rotation, and may have followed a much more complex path."

Thus, given a measurement of the relative displacement that occurred between two plates over some time-period, simply scaling the angle of the finite rotation (in order to interpolate the displacement over time) will not necessarily yield an accurate history of the relative positions of the plates over time. Measurements of intermediate displacements are necessary to build an accurate picture – the more measurements, the better.

The finite rotation that displaces a plate from its present-day position to its reconstructed position at some instant in the past is called a *total reconstruction pole* (Cox and Hart, 1986) (henceforth TRP). A TRP is associated with a fixed plate, a moving plate, and a particular instant in geological time: Such a TRP reconstructs the moving plate to its position relative to the fixed plate at that time. The TRP for the present day is the identity rotation. A sequence of TRPs for a particular moving plate is used to describe the plate's motion through time by sampling its relative displacement at key instants in the past; TRPs for intermediate times may be interpolated according to the mathematical rules described by Cox and Hart (1986).

### *7.3.5 Equivalent rotations and absolute reference frames*

Cox and Hart (1986) outline a simple algebra for composing TRPs of the same geological time. The key is to match up fixed plates and moving plates to form a continuous *plate circuit* through the graph of the rotation hierarchy. For example, the fixed plate and moving plate of a relative motion may be swapped if the direction of the finite rotation is reversed: The reverse of *B*-relative-to-*A* yields *A*-relative-to-*B*. Similarly, TRPs for the same geological time may be composed if the fixed plate of one TRP is matched to the moving plate of the other: *C*-relative-to-*B* may be composed with *B*-relative-to-*A* to give *C*-relative-to-*A*. Using this algebra, a circuit of relative motion may be traced between any two plates in the rotation hierarchy.

GPlates enables the user to specify an "anchored plate" which will be held fixed in the reconstruction. By composing and reversing the appropriate TRPs in the rotation hierarchy, GPlates can trace a plate circuit to any choice of anchored plate from every plate in the hierarchy, to calculate an *equivalent rotation* for each plate relative to the anchored plate (Ross and Scotese, 1988). These equivalent rotations may be used to reconstruct all the other plates relative to the anchored plate at that geological time. In an animation, the anchored plate would remain still, and all other plates would move relative to it. While this is not necessarily geologically realistic, it is nevertheless useful as a means to study relative motions, and is sufficient for many regional models.

Other uses of plate tectonics – such as using plate motions to provide kinematic boundary conditions to geodynamic mantle-convection models (O'Neill *et al*., 2005) – require information about the motions of plates relative to the interior of the Earth itself. For this purpose, it is necessary to identify a measurable geological frame of reference that is consistent over tens or hundreds of millions of years – this is termed an *absolute reference frame*. Examples of absolute reference frames include hotspot reference frames (relative to stationary or slow-moving mantle plumes that rise up through the mantle) and paleomagnetic reference frames (Torsvik and Smethurst, 1999) (relative to the average location of the Earth's magnetic north pole over millions of years).

The plate-motion predictions of different absolute reference frames do not always coincide. Part of this disagreement may be understood in terms of information limitations inherent to each particular type of absolute reference frame. For example, hotspot reference frames are disadvantaged by a limited number of

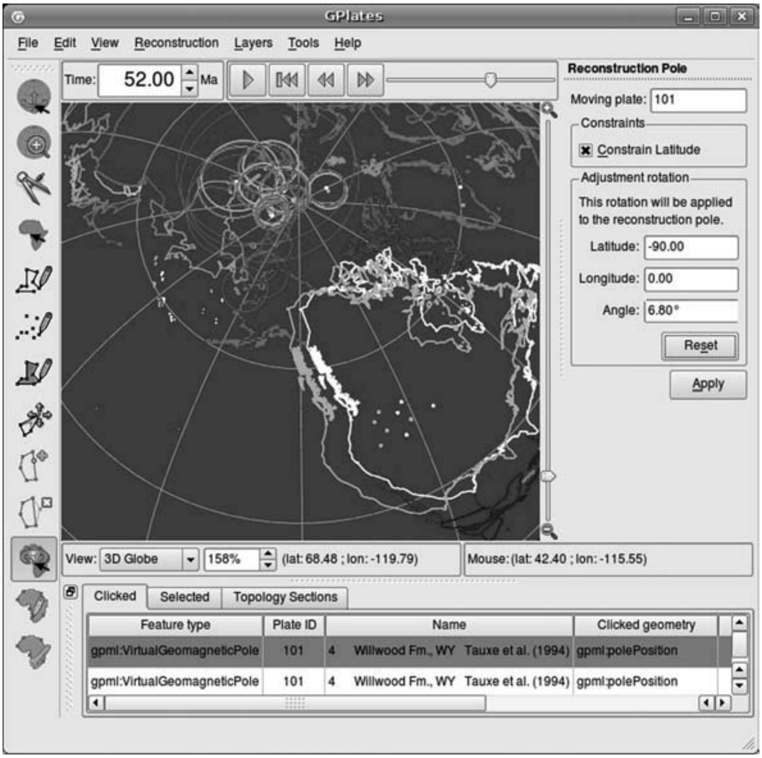



Figure 7.5. Modifying a plate reconstruction in GPlates using the interactive Modify Reconstruction Pole tool. Paleomagnetic data are being used to align the plates, so it is helpful to constrain the latitude of the plate; only the longitude of the plate will change. See color plates section.

hotspots; a lack of hotspots on many plates; and a lack of hotspot data older than 200 million years (Greiner, 1999). In contrast, the main disadvantage of paleomagnetic reference frames is that, due to the axial symmetry of the Earth's magnetic poles, paleomagnetism can only determine the latitude and orientation of a plate, not the longitude (Torsvik and Smethurst, 1999). As a result, a researcher must be able to test a reconstruction hypothesis using different absolute reference frames. Figure 7.5 illustrates how a GPlates user may experiment with an alternate plate reconstruction by varying only the longitude.

In the 2008 EarthByte Global Rotation Model, the relative motions between plates and absolute reference frames are also described using total reconstruction poles; the absolute reference frames have three-digit plate IDs which begin with a "0." A GPlates user may specify an absolute reference frame as the "anchored plate," to observe the motions of the plates relative to this geological frame of reference. This is illustrated in Figure 7.6.

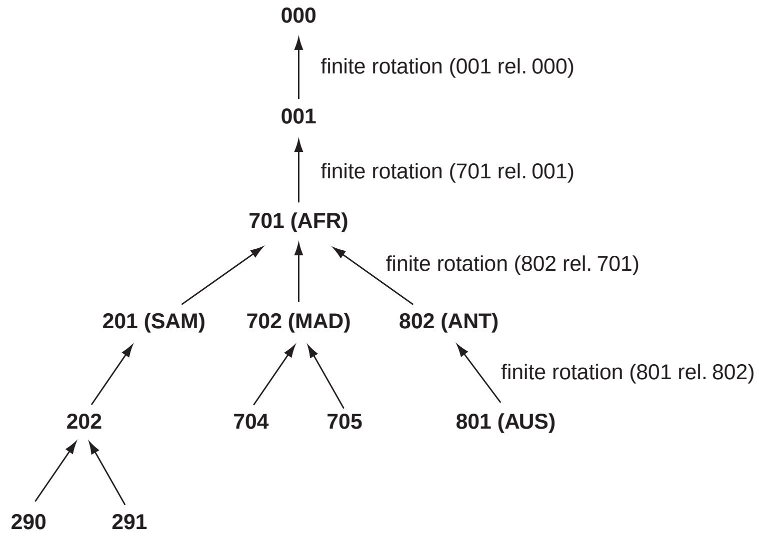



Figure 7.6. Relative rotations with an absolute reference frame, as used in GPlates.

## 7.4 Conclusions

Offering all the familiar functionality of traditional plate-tectonics reconstruction software, and based upon a solid theoretical foundation of finite rotations, relative motions and the rotation hierarchy, GPlates measures up as an interactive plate-tectonics application. In fact, more than simply meeting expectations, GPlates advances the state of the art along several dimensions of functionality. These advances are due to a combination of successful leveraging of technological advancements (OpenGL, Qt, XML, GML) and judicious selection of suitable techniques (such as the internal geometry representation based upon 3-D unit-vectors and unit-quaternions, and the careful modeling of the geological domain).

There are four significant differences in GPlates that set it apart from previous plate-tectonic reconstruction software: The rigorous modeling methodology used to model geological, geophysical and paleogeographic features in the GPGIM and enable GPlates to serve as a component in a larger GML-based infrastructure; the sophisticated information model building blocks which enable expression of advanced and abstract concepts; the paradigm shift of calculation templates for automated post-processing of reconstructions, to enable new data to be computed on the fly based upon plate motions; and finally, the fact that GPlates is released as open-source software, enabling anyone to build upon or modify this foundation, to extend GPlates to meet their needs, and then release their work back to the geology community, further enriching the community in a virtuous cycle.

Future GPlates development will include the ability to process paleomagnetic features to derive new total reconstruction poles automatically; the ability to read and write GeoSciML documents, and reconstruct GeoSciML features alongside GPGIM features; the ability to ready and write GML-based data, participating in a computing grid as a WFS client; and the ability to reconstruct rasters on plates. These innovations will enable users to investigate alternative fits of the continents interactively, simultaneously displaying and fitting data such as faults, sutures, orogenic belts, terranes, paleomagnetic poles, and raster data such as magnetic or gravity grids, to test hypotheses of supercontinent formation and breakup through time. Users will also be able to derive and test models for the evolution of tectonically complex areas such as the Arctic, the Caribbean and Southeast Asia. The diverse functionality of GPlates, paired with its free availability, will also allow teachers to explore creative ways of integrating plate tectonics into educational activities.

With a flexible, expressive, interoperable information model and the ability to represent calculation templates as data, GPlates leads the arrival of the next generation of interactive plate-tectonics software – a generation in which a plate-tectonics application does not operate in isolation, but as an integrated visualization and processing component within a computational grid; and a plate-tectonic reconstruction is no longer an isolated result, but a single stage in an adaptable workflow.

## References

Altmann, S. (1986). *Rotations, Quaternions and Double Groups*. Oxford: Oxford University Press.

Beatty, M. F. (1986). *Principles of Engineering Mechanics*. New York: Springer, 595pp.

Cox, A. and Hart, R. B. (1986). *Plate Tectonics: How It Works*. Oxford: Blackwell Scientific Publications.

Cox, S. J. D. and Richard, S. M. (2005). A formal model for the geologic time scale and global stratotype section and point, compatible with geospatial transfer standards. *Geosphere*, **1**: 119–137.

Cox, S. J. D., Daisey, P., Lake, R., Portele, C., and Whiteside, A., eds. (2004). OpenGIS Implementation Specification #03–105r1: OpenGIS Geography Markup Language (GML) Implementation Specification, version 3.1.1. February 2004.

Greiner, B. (1999). Euler rotations in plate-tectonic reconstructions. *Computers & Geosciences*, **25**(3): 209–216.

Gurnis, M., Turner, M., Zahirovic, S. *et al*. (2010). Plate reconstructions with continuously closing plates. *Computers and Geosciences*, in review.

Hellinger, S. J. (1981). The uncertainties of finite rotations in plate tectonics. *Journal of Geophysical Research*, **86**: 9312–9318.

Herold, N., Seton, M., Müller, R. D., You, Y., and Huber, M. (2008). Middle Miocene tectonic boundary conditions for use in climate models. *Geochemistry Geophysics Geosystems*, **9**, Q10009, doi:10.1029/2008GC002046.

Herold, N., You, Y., Müller, R. D., and Sdrolias, M. (2009). Climate model sensitivity to changes in Miocene paleo-topography. *Australian Journal of Earth Sciences*, **56**(8): 1049–1059.

Kearey, P. and Vine, F. J. (1990). *Global Tectonics*. Oxford: Blackwell Scientific Publications.

Lake, R., Burggraf, D. S., Trninić, M., and Rae, L. (2004). *Geography Markup Language: Foundation for the Geo-Web*. Chichester, UK: John Wiley & Sons.

Lake, R. (2005). The application of geography markup language (GML) to the geological sciences. *Computers & Geosciences*, **31**(9): 1081–1094.

Lee, Y. T. (1999). *Information Modeling: From Design to Implementation. National Institute of Standards and Technology*. Available at www.mel.nist.gov/msidlibrary/doc/tina99im.pdf, last accessed 2009–04–08.

Liu, L. J., Spasojevic, S., and Gurnis, M. (2008), Reconstructing Farallon plate subduction beneath North America back to the Late Cretaceous. *Science*, **322**: 934–938.

Müller, R. D., Sdrolias, M., Gaina, C., and Roest, W. R. (2008). Age, spreading rates and spreading asymmetry of the world's ocean crust. *Geochemistry Geophysics Geosystems*, **9**(4): 1–19, doi:10.1029/2007GC001743.

O'Neill, C. J., Müller, R. D., and Steinberger, B. (2005). On the uncertainties in hotspot reconstructions, and the significance of moving hotspot reference frames. *Geochemistry Geophysics Geosystems*, **6**: doi:10.1029/2004GC000784.

Ross, M. I. and Scotese, C. R. (1988). A hierarchical tectonic model for the Gulf of Mexico and Caribbean region. *Tectonophysics*, **155**: 139–168.

Schettino, A. (1999). Polygon intersections in spherical topology: Application to plate tectonics. *Computers & Geosciences*, **25**(1): 61–69.

Sen, M. and Duffy, T. (2005). GeoSciML: Development of a generic GeoScience Markup Language. *Computers & Geosciences*, **3**(9): 1095–1103.

Shneiderman, B. (1983). Direct manipulation: A step beyond programming languages. *IEEE Computer*, **16**(8): 57–69.

Spasojevic, S., Liu, L., and Gurnis, M. (2009). Adjoint models of mantle convection with seismic, plate motion and stratigraphic constraints: North America since the Late Cretaceous. *Geochemistry Geophysics Geosystems*, **10**, Q05W02, doi:10.1029/2008GC002345.

Steinberger, B. (2007). Effects of latent heat release at phase boundaries on flow in the Earth's mantle: Phase boundary topography and dynamic topography at the Earth's surface. *Physics of the Earth and Planetary Interiors*, **164**: 2–20.

Torsvik, T. H. and Smethurst, M. A. (1999). Plate tectonic modelling: virtual reality with GMAP. *Computers & Geosciences*, **25**(4): 395–402.

Van der Voo, R. (1993). *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*. Cambridge: Cambridge University Press.

Vretanos, A. P., ed. (2005). OpenGIS Implementation Specification #04-094r1: Web Feature Service Implementation Specification, version 1.1.0. May 2005.

Wessel, P. and Müller, R. D. (2007). Plate tectonics. In *Treatise on Geophysics*. Vol. 6. Amsterdam: Elsevier, pp. 49–98.

Zhang, C., Peng, Z. R., Li, W., and Day, M. J. (2003). GML-based interoperable geographical databases. Available at www.ucgis.org/summer03/studentpapers/chuanrongzhang.pdf.

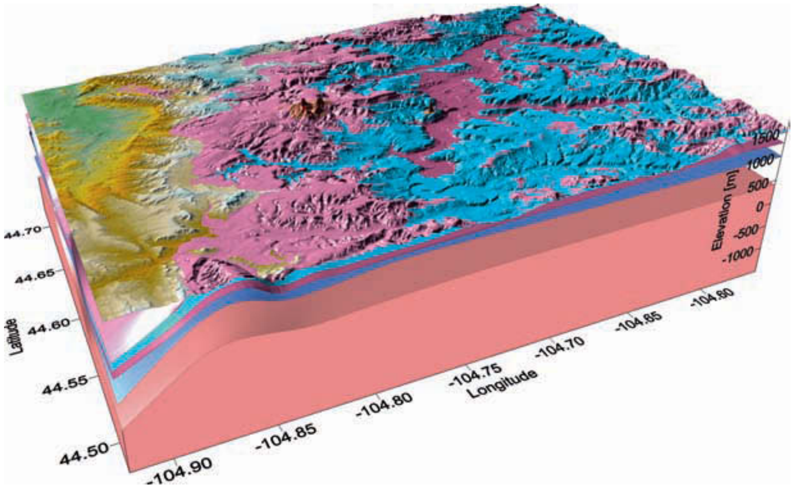



Plate 1. Example of a 3-D geological/geophysical model consisting of layers that are bounded by geologic interfaces that have been extracted from surface and subsurface geologic data. The lowest layer is the Precambrian basement. The interfaces are georeferenced and provide a framework for assigning physical properties to the layers between them. Image provided by Kevin Crain. (See Figure 1.1.)

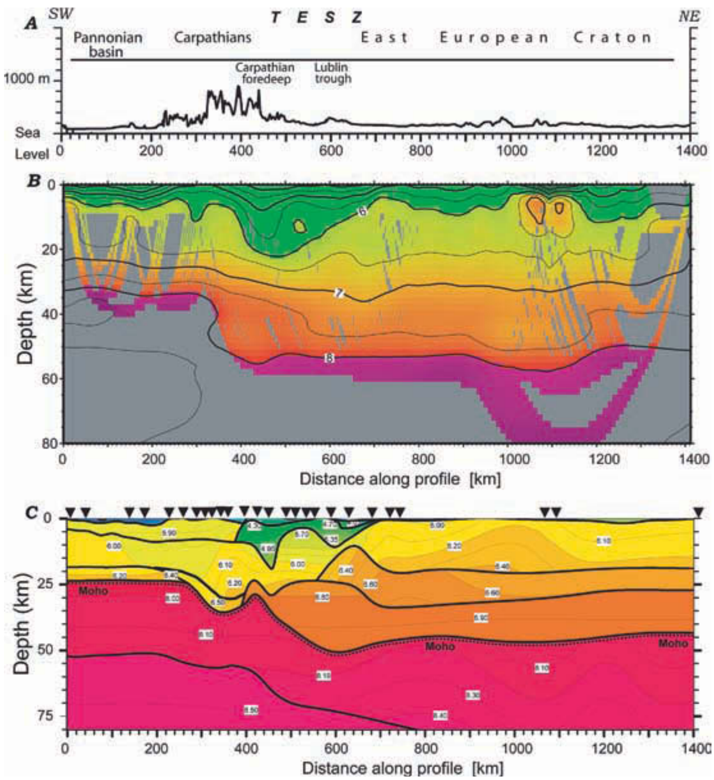



Plate 2. (*A*) Topographic profile showing the main geologic features present; (*B*) Preliminary seismic velocity model derived by tomographic inversion of the arrival times of the first seismic wave observed. The model is smooth and lacks the detail that is needed to make a suitable geological interpretation. The numbers in the model are P-wave velocities in km/s; (*C*) Seismic velocity model derived by ray trace forward (trail-and-error) modeling of all observed seismic arrivals. This approach has the advantage of providing more detail, but the formal analysis of certainty is difficult. The numbers in the model are P-wave velocities in km/s. Inverted triangles indicate the locations' shot points that produced the observed seismograms. (See Figure 1.3.)

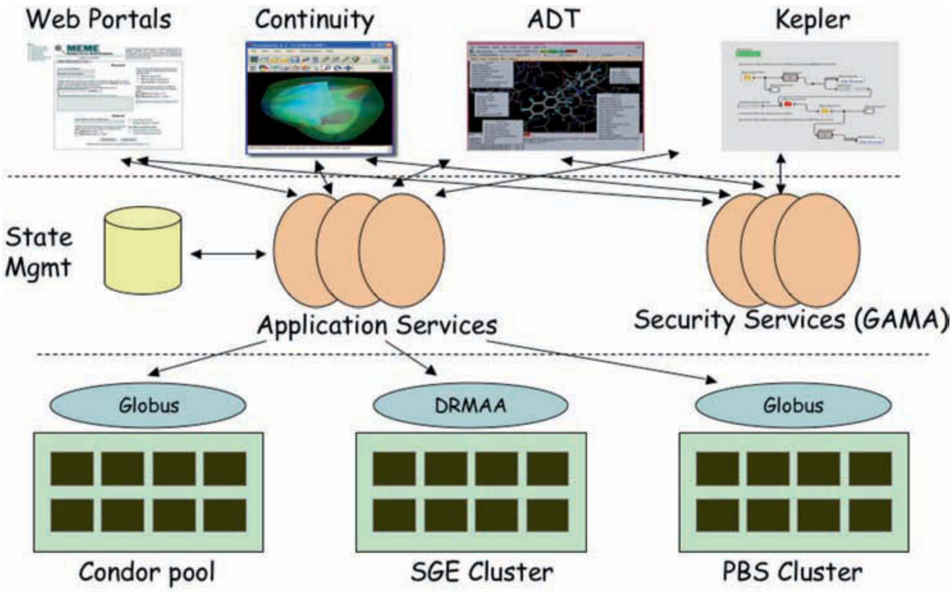



Plate 3. NBCR SOA architecture. (See Figure 4.1.)

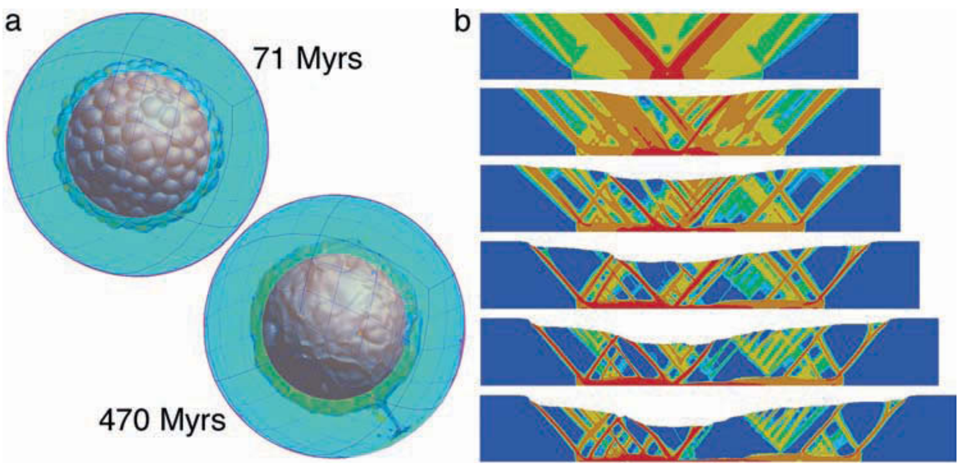



Plate 4. (a) Two time-instances in a calculation using *CitcomS* 3.0. The initial layer thickness is 250 km, and Ra = $1.4 \times 10^9$. The density of normal mantle increases 50% from the surface to the CMB due to compressibility. The anomalous mantle has 8% larger bulk modulus and is 3.8% denser at the surface and 1% denser at the CMB. The viscosity is temperature and depth dependent, varying 400× due to temperature and is 10× larger for the lithosphere and the lower mantle. 120M tracers and $64 \times 64 \times 64 \times 12$ = 3.1M elements and uses 48 processors. (b) Pictured is the strain-rate invariant for calculations of the GeoMod 2004 extension benchmark. The images range from 0 to 5 cm of extension, top to bottom. This model had a mesh of $512 \times 32$ and was computed on 32 processors using *Gale 1.3.0*. (See Figure 5.3.)

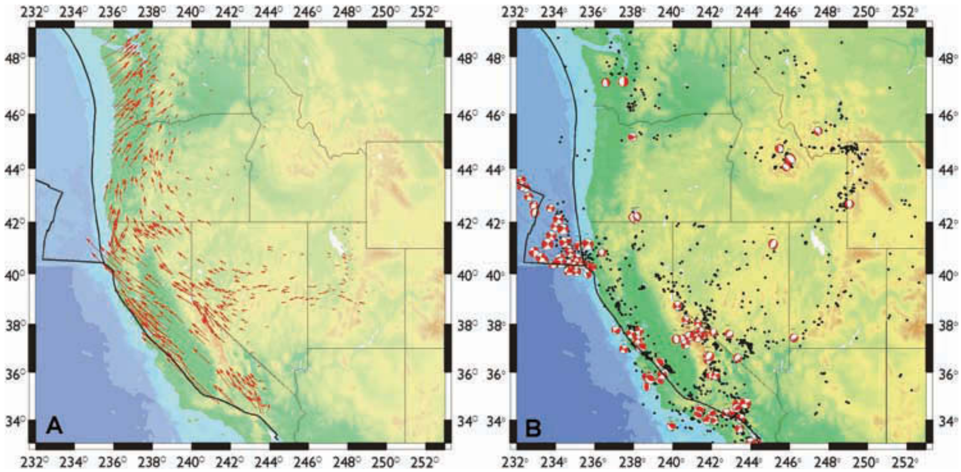



Plate 5. (A) Topography (background) and GPS velocity (vectors) relative to stable North America. GPS data are from USGS (http://quake.wr.usgs.gov/research/deformation/gps/qoca/index.html). (B) Seismicity (dots, from USGS) and focal mechanisms solutions (from Harvard catalog) in western United States. (See Figure 6.1.)

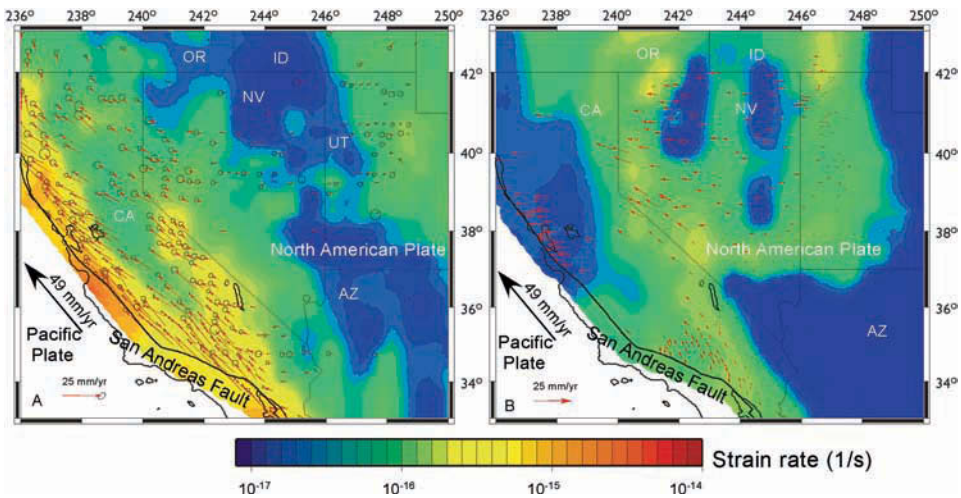



Plate 6. The GPS (A) and geologically (B) derived surface velocities (vectors) and strain rates (background color) in the southwestern United States. Source geologically determined surface velocity: McQuarrie and Wernicke (2005). (See Figure 6.3.)

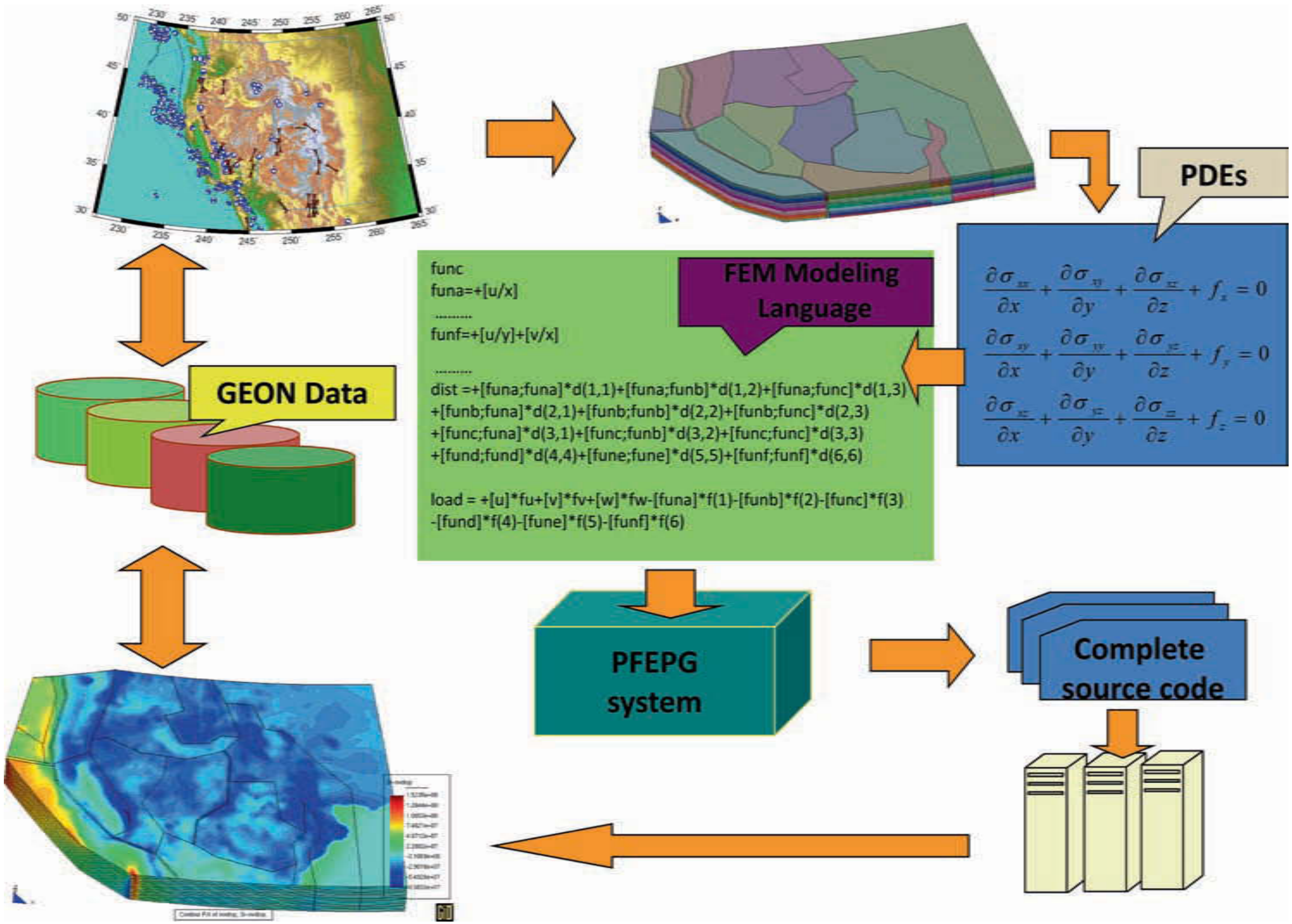



Plate 7. Workflow of the automated parallel FEM system. (See Figure 6.2.)

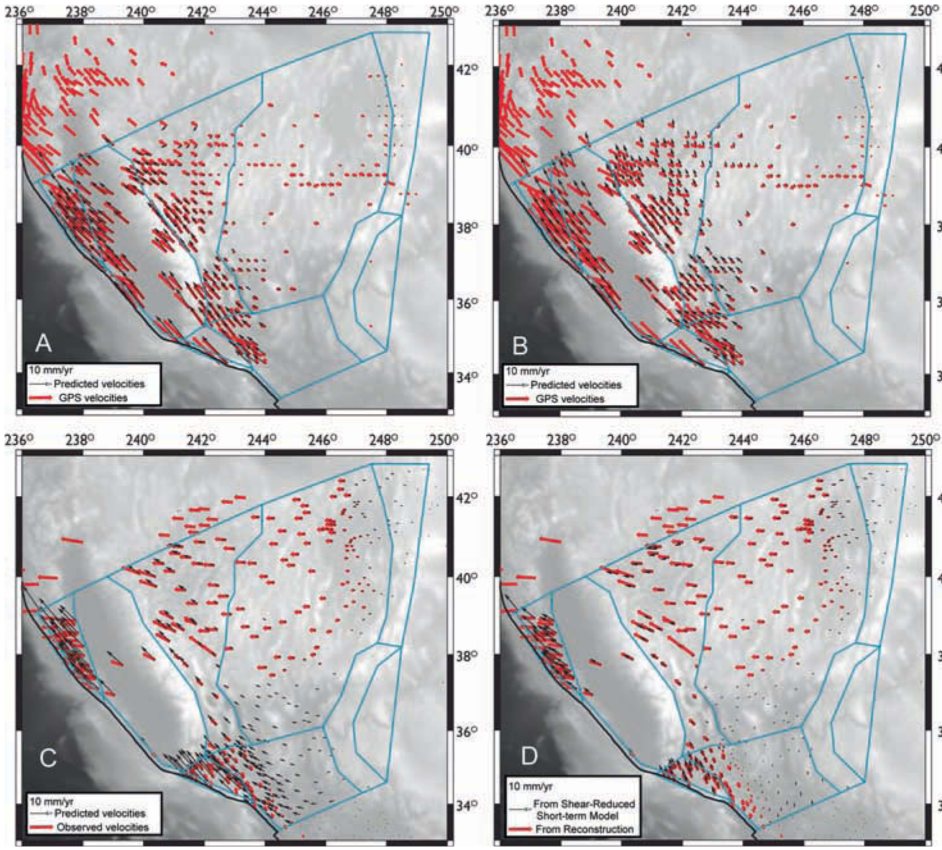



Plate 8. (A) Comparison of the GPS velocities with those predicted by a viscous model with plate boundary traction as the sole tectonic force. (B) Comparison of the GPS velocities with those predicted by an elastic model. (C) Comparison of the geologically derived crustal motion with the predictions of the short-term deformation model constrained by the GPS data. (D) A similar comparison but with reduced shear traction on the SAF in the model. (See Figure 6.5.)

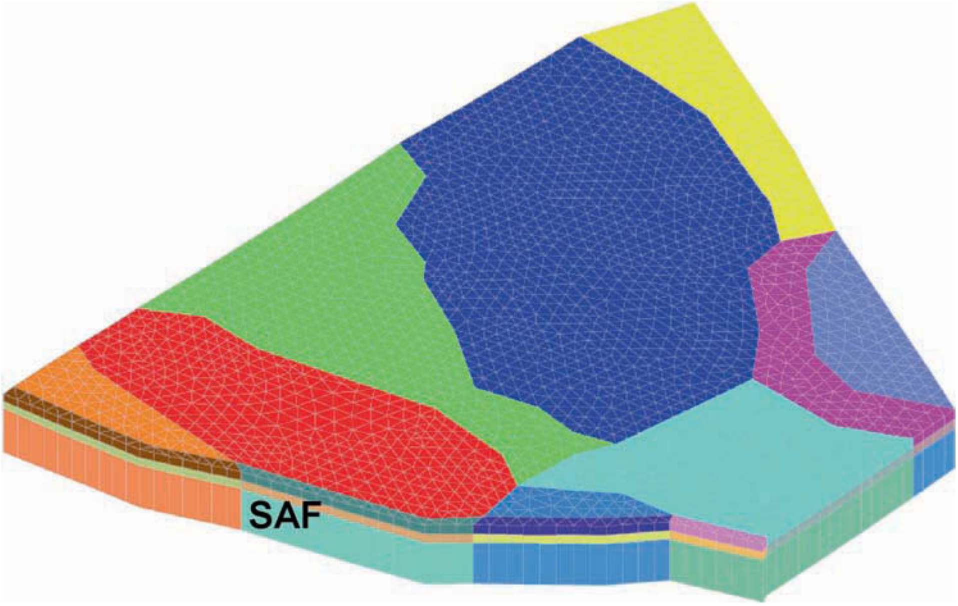



Plate 9. The finite element model. The blocks represent major tectonic units in southwestern USA. SAF: San Andreas Fault. (See Figure 6.4.)

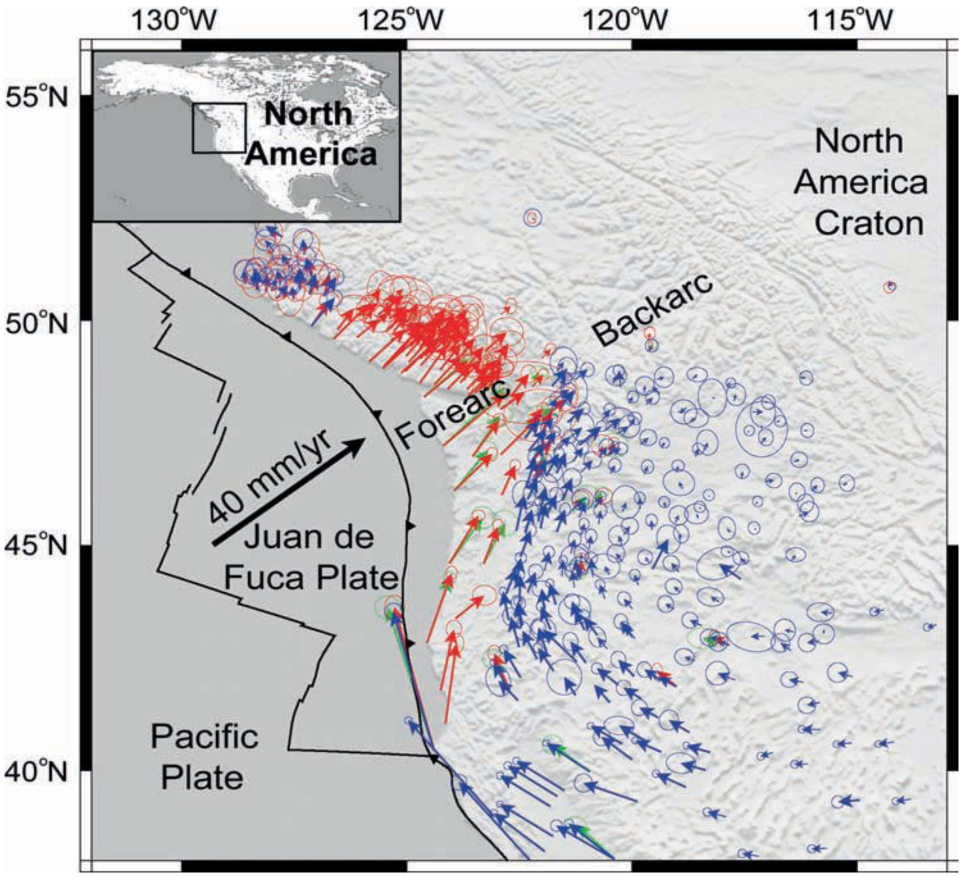



Plate 10. GPS velocities and topographic relief in the Cascadia subduction zone. Sources for the GPS data: Miller *et al*. (2001b) (green), Mazzotti *et al*.(2003) (red), and McCaffrey *et al*. (2007) (blue). (See Figure 6.6.)

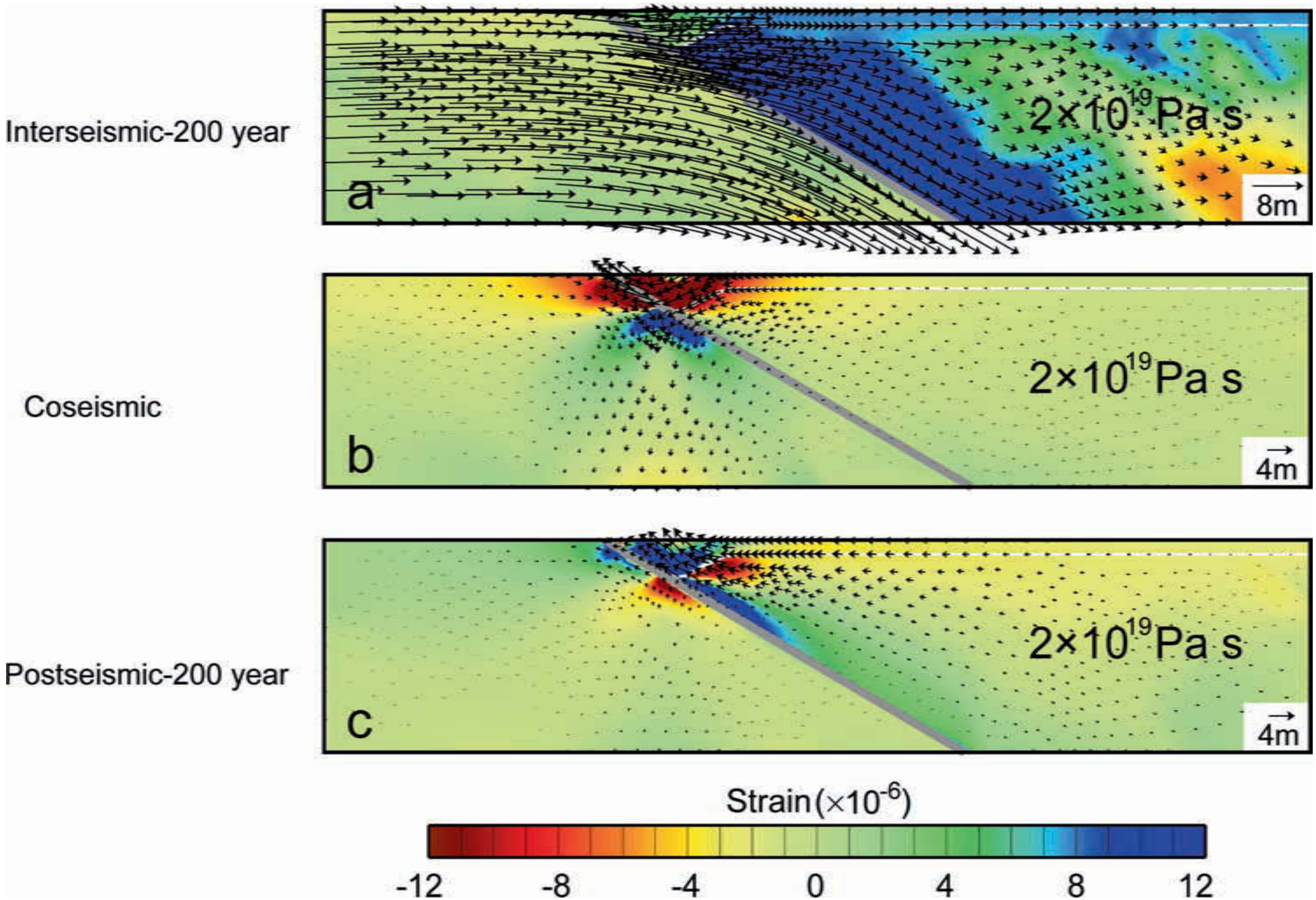



Plate 11. Predicted displacement (vectors) and horizontal strain (background color) across the plate boundary during one seismic cycle. (a) Accumulative inter-seismic displacement and strain for 200 years. (b) Co-seismic displacement and strain. (c) Accumulative post-seismic displacement and strain due to 200 years of viscous relaxation. Heavy gray lines show the subduction interface. White lines in the overriding plate show the bottom of the elastic or elastoplastic upper crust. Positive strain is compressive. (See Figure 6.8.)

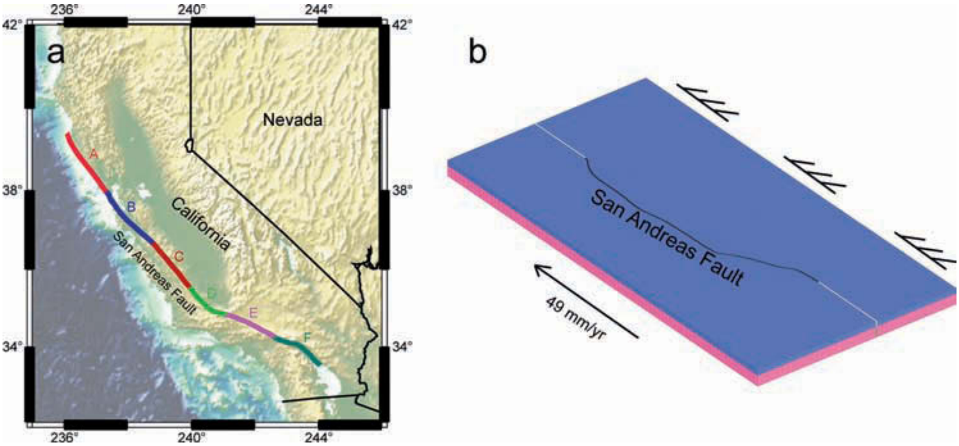



Plate 12. (a) Topographic relief and the San Andreas Fault (SAF). In the model, rupture patterns along each segment of the SAF are examined. (b) Numerical mesh and boundary conditions of the finite element model for the SAF system. The entire San Andreas Fault (black line) is explicitly included in the model. (See Figure 6.10.)

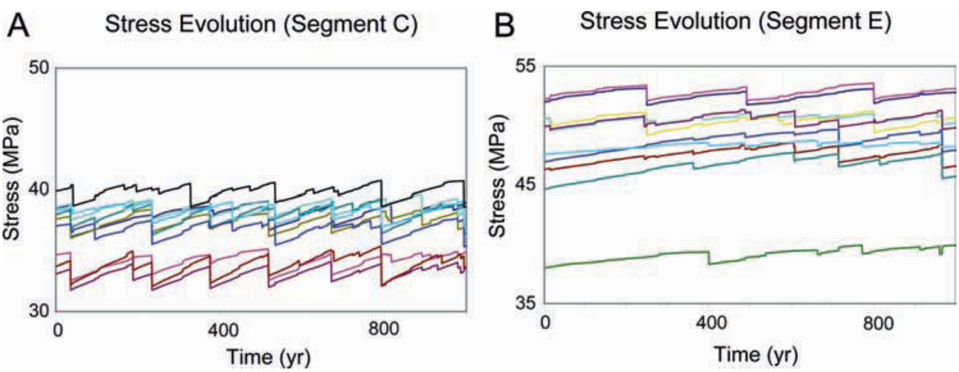



Plate 13. Predicted rupture patterns in two segments of the SAF, shown by the stress evolution of a group of sample points along each segment. (See Figure 6.12.)

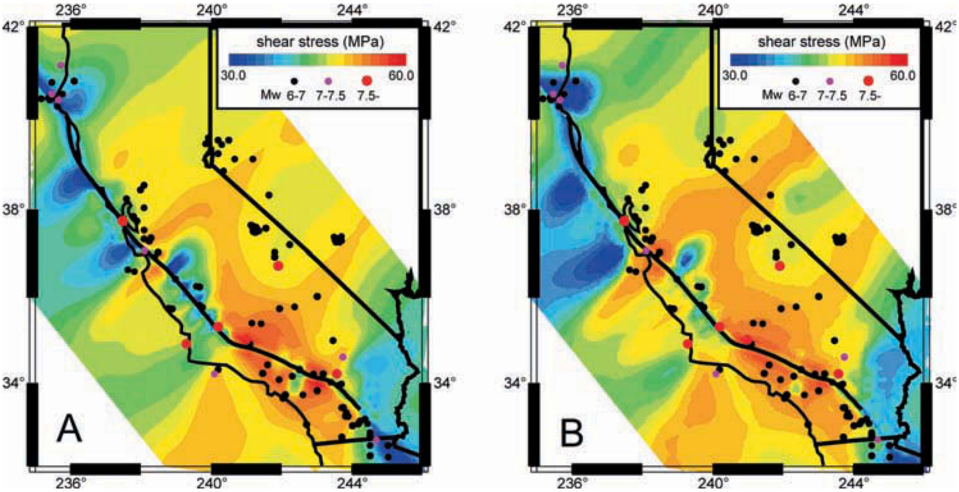



Plate 14. Predicted background shear stress. (A) Cumulative shear stress after 40 000 model years of repeated ruptures along various segments of the SAF. (B) Steady-state background shear stress predicted by the long-term faulting model where fault slip is simulated by continues plastic creep (Li and Liu, 2006). Dots are earthquakes. (See Figure 6.13.)

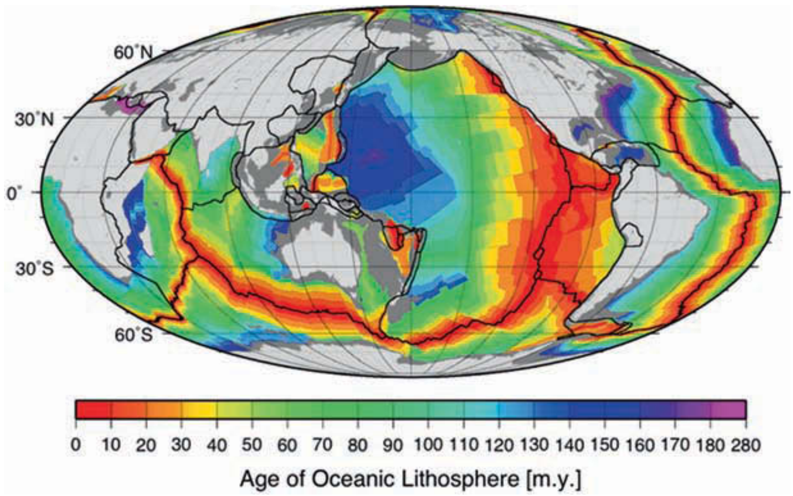



Plate 15. The age of the oceanic lithosphere, with plate boundaries shown in black and continents in light gray. (See Figure 7.1.)

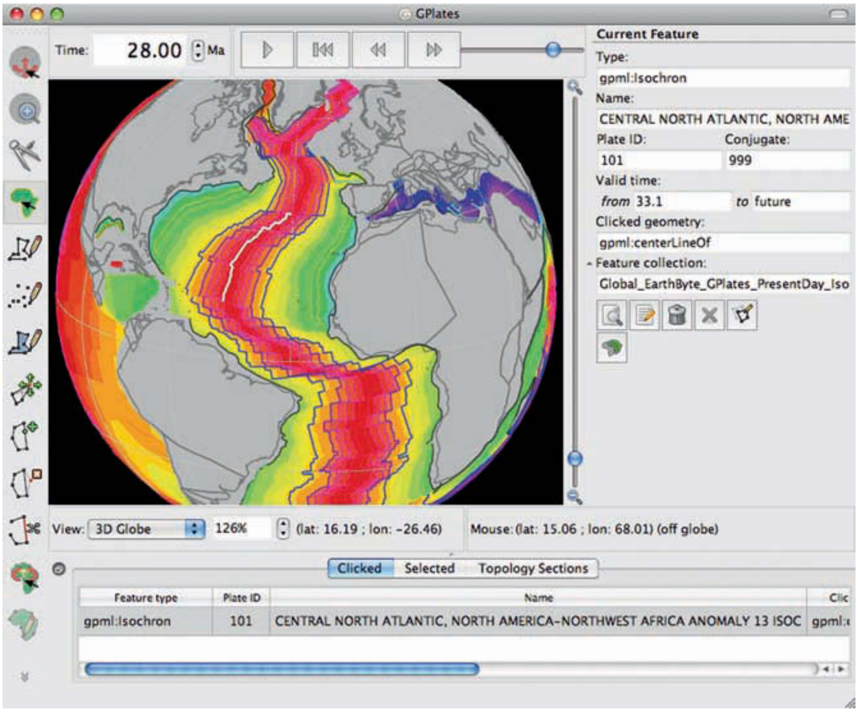



Plate 16. The GPlates main window on MacOS X, displaying a reconstruction at 28 Ma (million years ago). The user has clicked on an Isochron feature. (See Figure 7.2.)

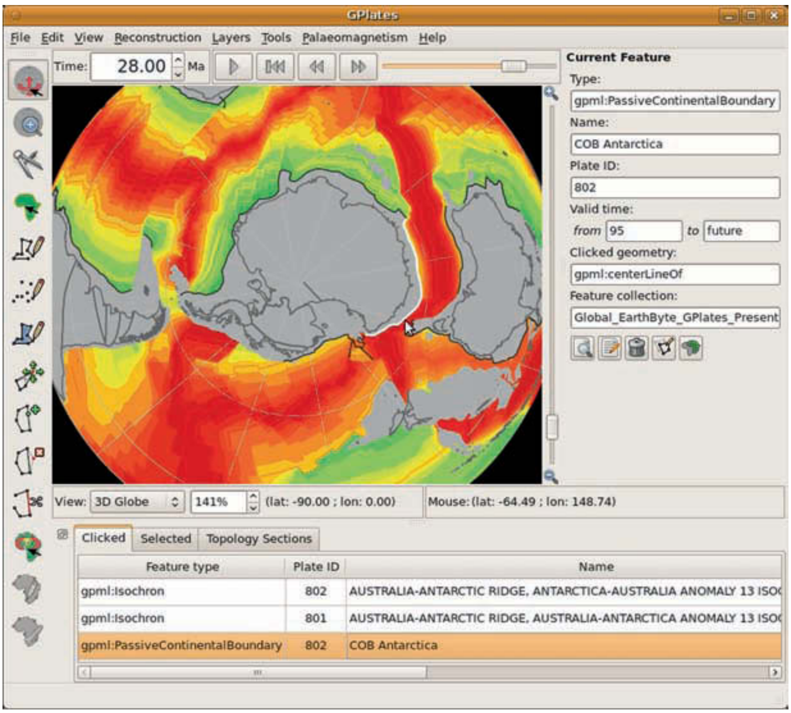



Plate 17. The GPlates main window on Linux, displaying a reconstruction at 28 Ma. The view of the globe may be re-oriented interactively by dragging on the globe with the mouse. (See Figure 7.3.)

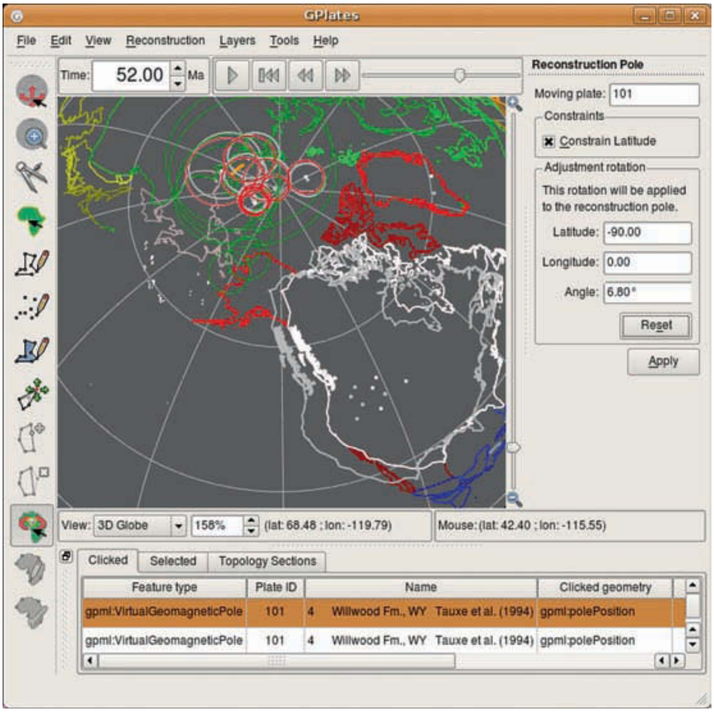



Plate 18. Modifying a plate reconstruction in GPlates using the interactive Modify Reconstruction Pole tool. Paleomagnetic data is being used to align the plates, so it is helpful to constrain the latitude of the plate; only the longitude of the plate will change. (See Figure 7.5.)

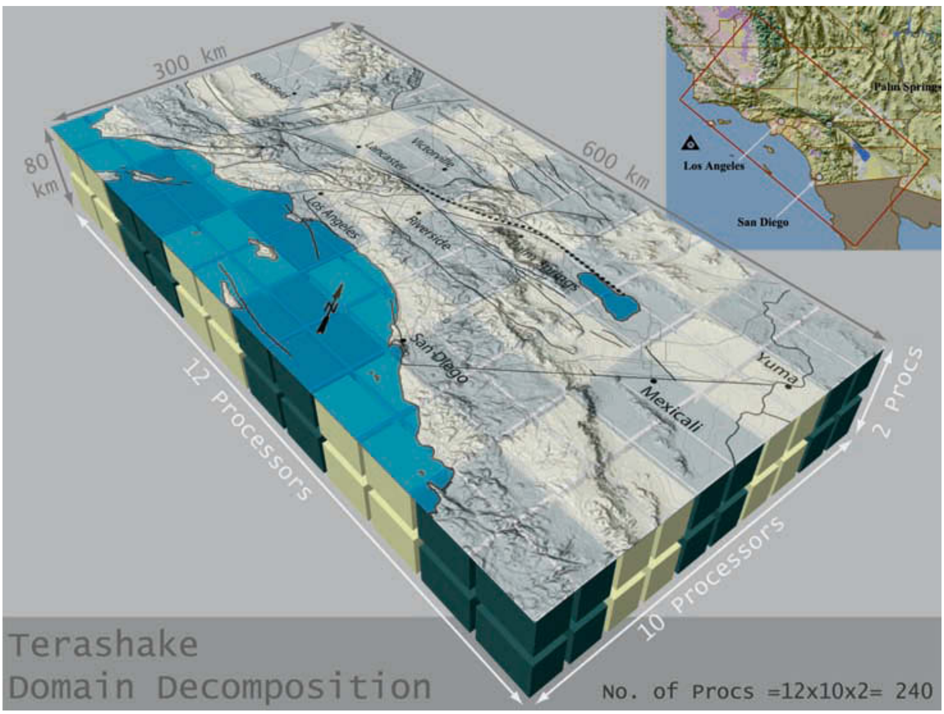



Plate 19. The top right inset shows the geographic location of the simulation region depicted by red rectangle. In the center, the topography, fault lines, and city locations are visible. This also shows the domain decomposition of this region into 240 processors. (See Figure 8.1.)

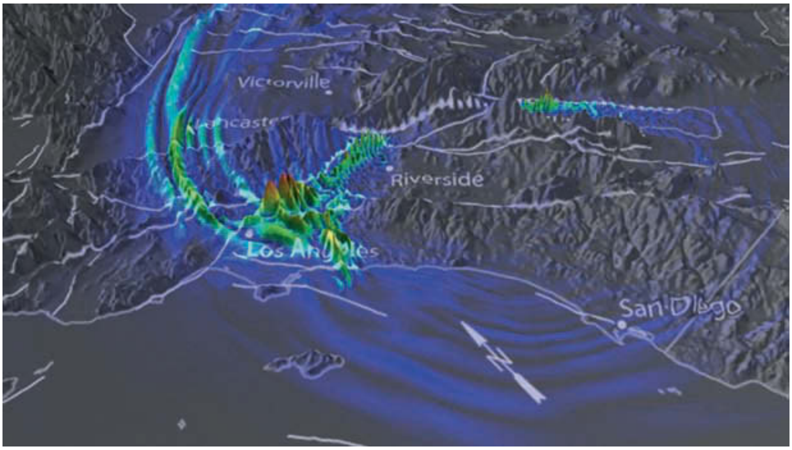



Plate 20. Snapshot showing deformation along the vertical axis of the terrain-based velocity magnitudes. (See Figure 8.2.)

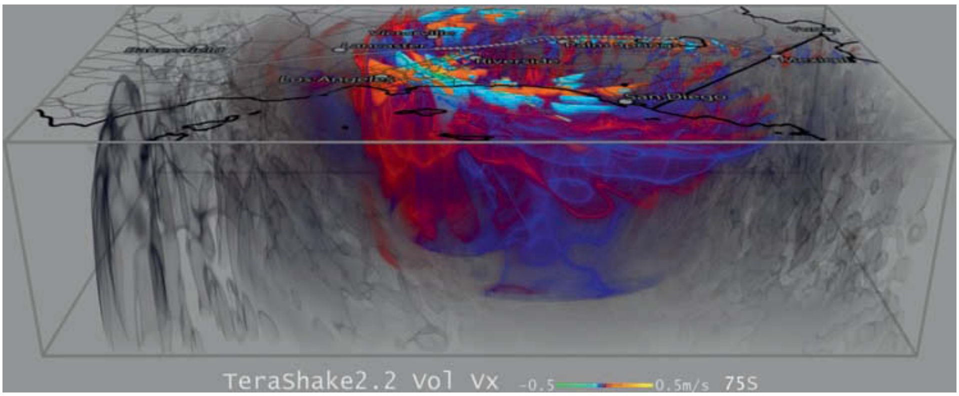



Plate 21. Snapshot showing volume-rendered velocity in *y* (shorter horizontal) direction. The *z*-axis (vertical) has been scaled twice. (See Figure 8.3.)

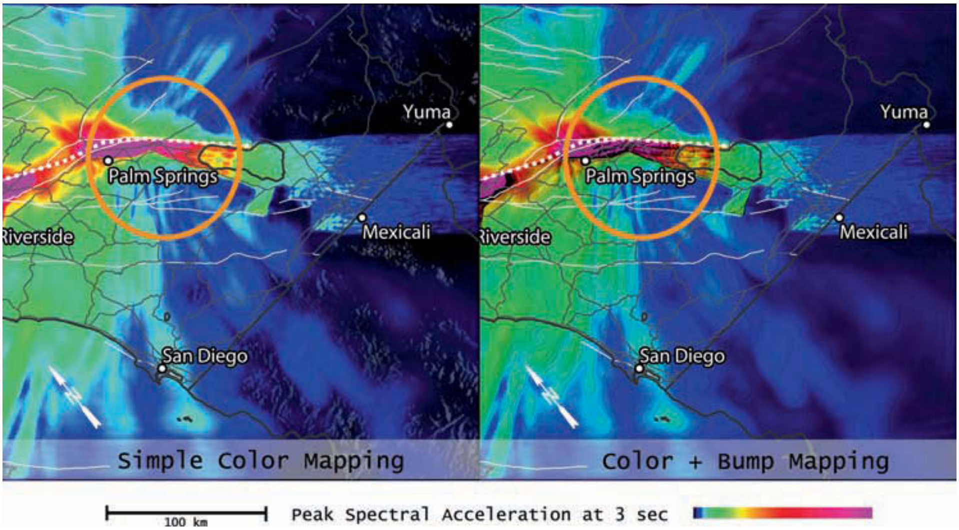



Plate 22. Map showing the advantage of the self-contouring technique in contrast to the simple color mapping. In the region highlighted by the orange circle, scientists identified a star-burst pattern. This indicates an unusual radiation of energy worthy of further investigation, which went unnoticed with simple color mapping. (See Figure 8.4.)

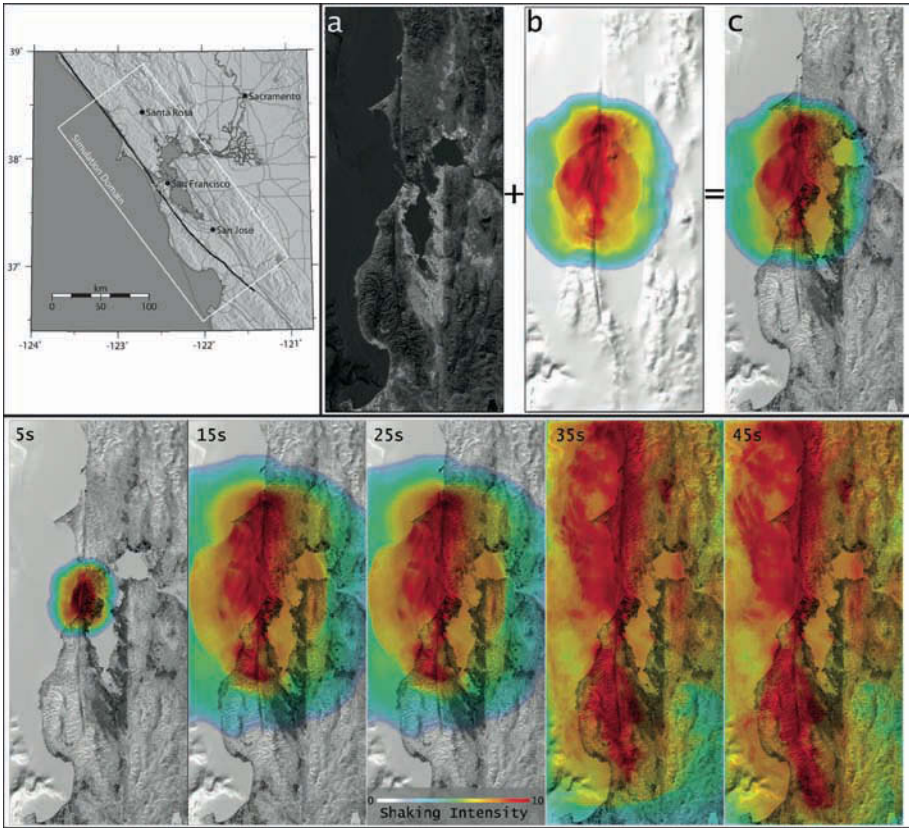



Plate 23. The top left inset shows a white rectangle indicating the simulation region 110 km long and 80 km wide. The top right inset shows the color blending of satellite imagery (a) with the coloring based on data (b) to produce the final image (c). The bottom image shows the earthquake's progression at 5, 15, 25, 35, and 45 seconds, respectively. (See Figure 8.5.)

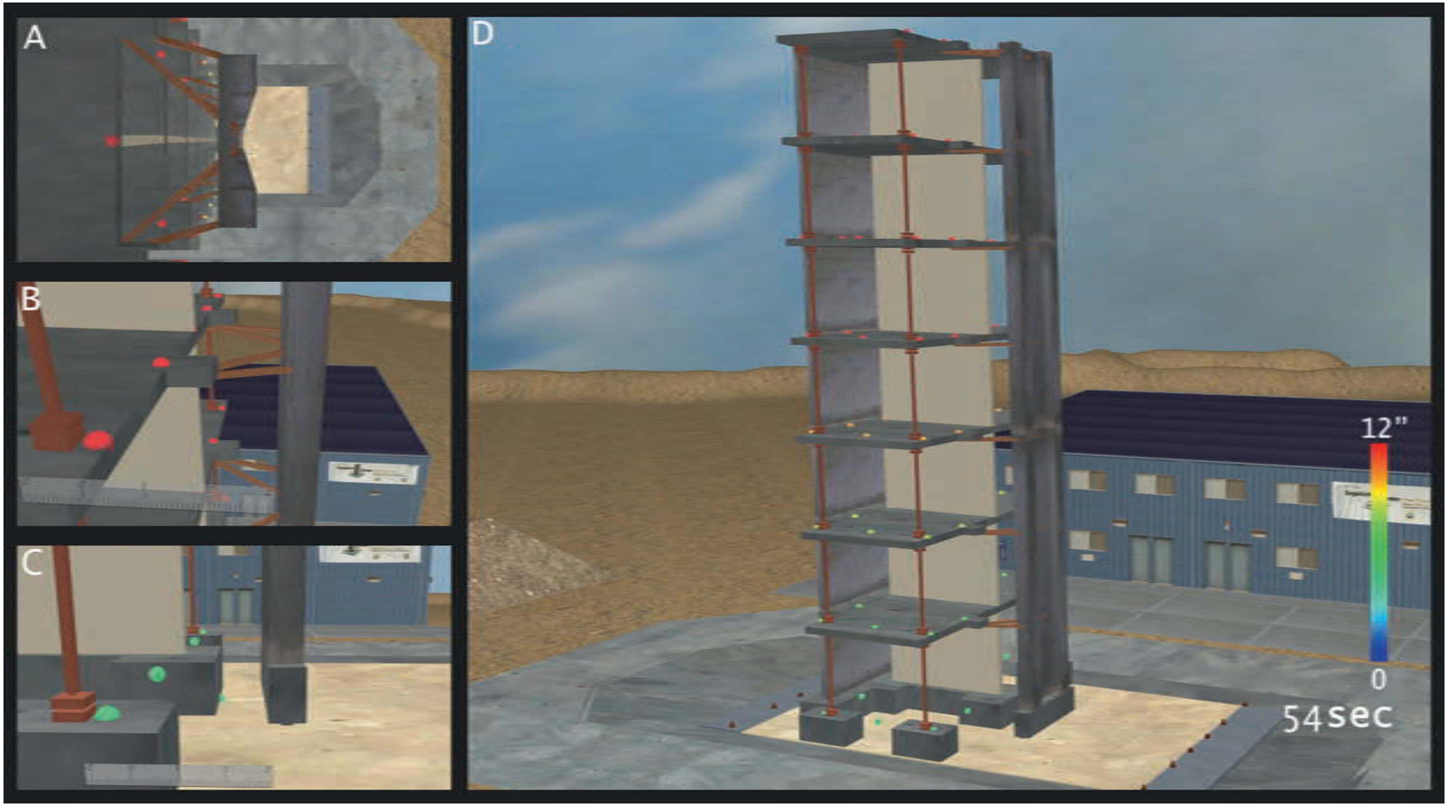



Plate 24. An illustration showing the particle system and a deformer skeleton. Each particle (sphere) corresponds to a node location on the deformation skeleton. The particle system is driven by PDC files translated from observed data. Each particle's color depicts the displacement at their location as shown in the color map. The position and color of the particles change with time based on observed data. An animation of skeleton deformation is available at the web site. (See Figure 8.6.)

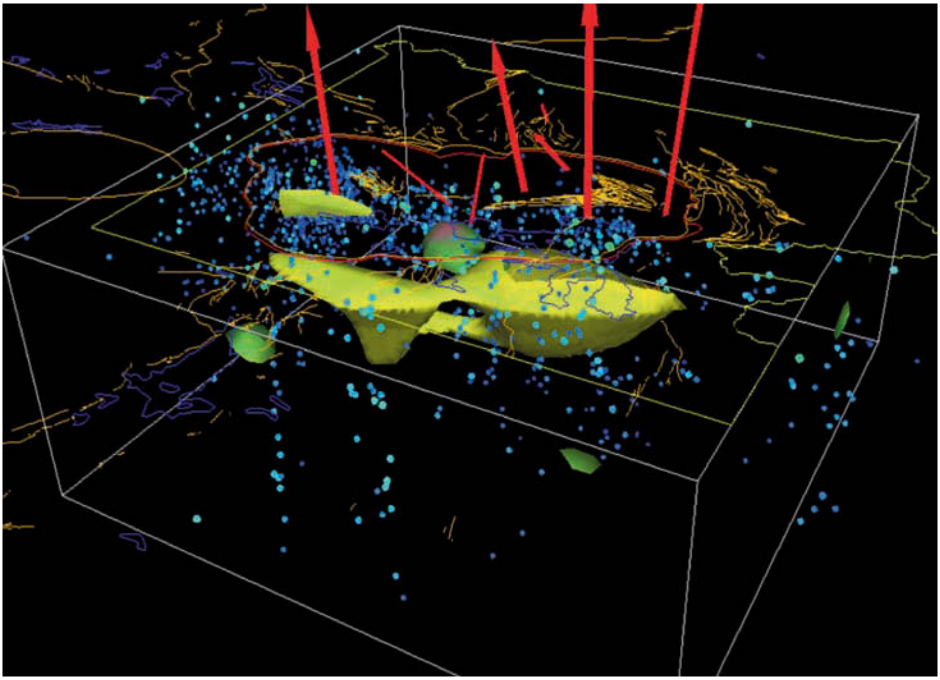


Plate 25. IDV display showing a 3-D oblique view from the southeast of the Yellowstone National Park area, with surface fault lines and caldera outline (orange and red, US Geological Survey *et al*., 2006), seismic tomography isosurfaces (yellow), earthquake hypocenters sized by magnitude (blue), GPS velocity vectors (red; Chang *et al*., 2007; Husen *et al*., 2002; Puskas *et al*., 2007), and isosurfaces of a fluid intrusion model (green; Vasco *et al*., 2007). (See Figure 9.1.)

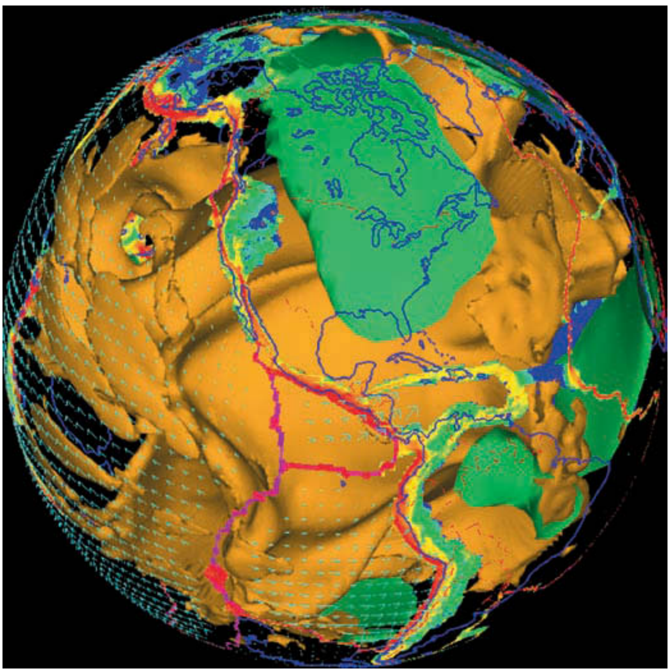


Plate 26. GEON IDV display of four diverse global datasets: a mantle convection isosurface (50% of core-mantle boundary temperature, orange; McNamara and Zhong, 2005), a seismic tomography isosurface (positive 3% anomaly surface, green; Gu and Dziewonski, 1999), and plate motion vector symbols with the associated image of the second invariant of global strain rate (Kreemer *et al*., 2003). (See Figure 9.3.)

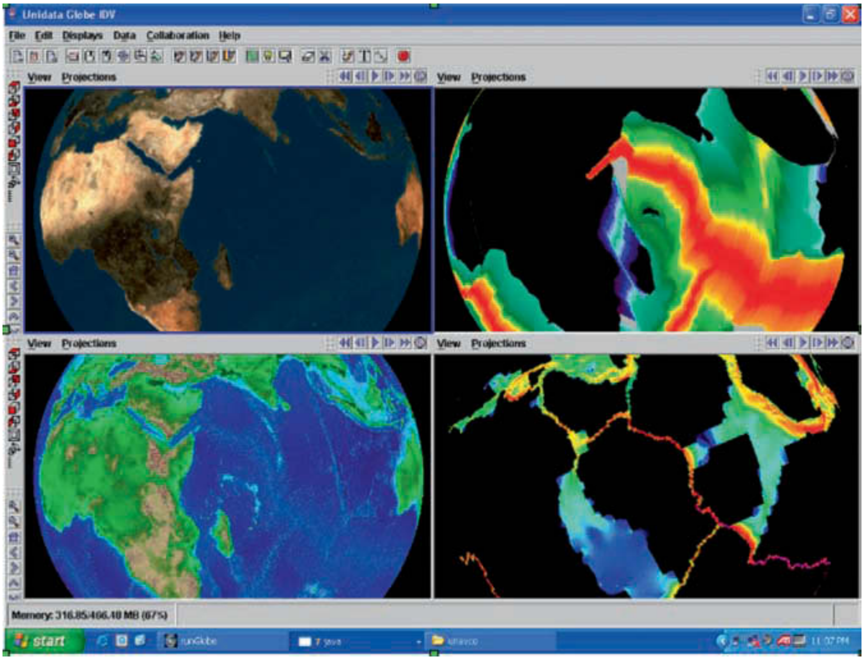



Plate 27. IDV display with four linked globe displays showing, from upper left, the Blue Marble image from the JPL OnEarth web map service, ages of ocean floor (Müller *et al*., 1997), topography colored by elevation (National Geophysical Data Center GLOBE Digital Elevation Model), and the second invariant of strain rate (Kreemer *et al*., 2003) colored by value. Interactive rotation and zoom links all four displays together. Multiple linked flat map displays may also be created in the IDV. (See Figure 9.4.)

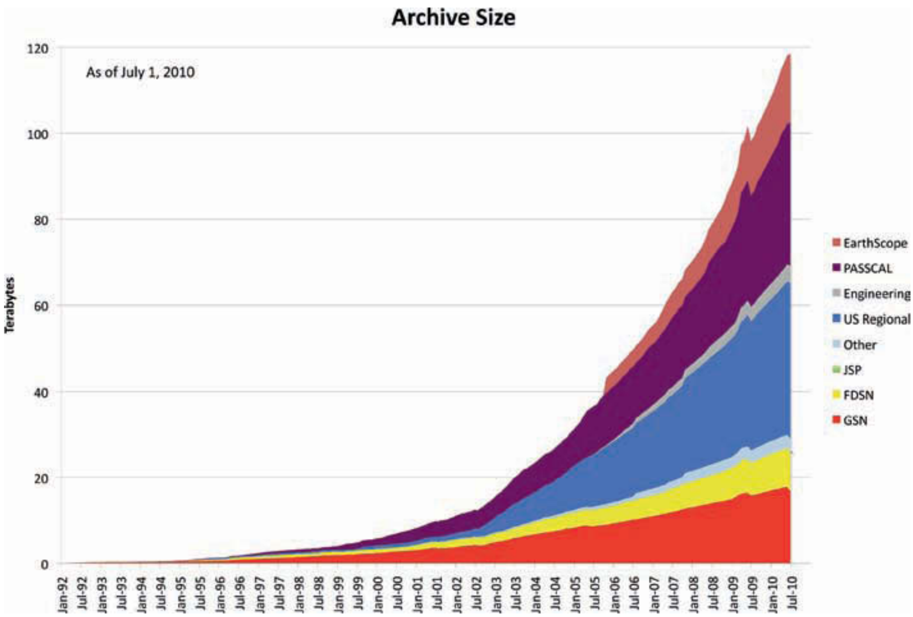



Plate 28. The growth of the IRIS DMC Archive. As of July 2010 the IRIS DMC manages approximately 120 terabytes of data. This figure shows the growth of the archive since 1992. Contributors from the bottom (red) to the top in the figure are (1) IRIS Global Seismic Network, (2) International Federation of Digital Seismographic Data, (3) Joint Seismic Program, (4) Other networks, (5) US regional Networks including the USGS ANSS network, (6) engineering data from structures, (7) data from the PASSCAL program of IRIS, and (8) data from the NSF EarthScope program that includes USArray and PBO seismic data. The archive is presently growing at about 20 terabytes per year. (See Figure 13.2.)

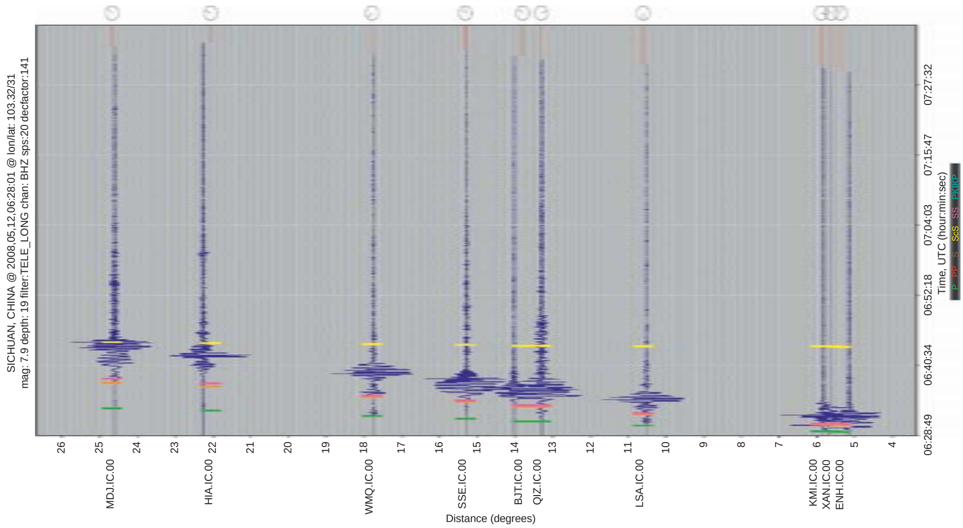



Plate 29. Seismic record section plot. This figure shows energy arriving at ten stations in China from the Sichuan earthquake of May 12, 2008. The various phases are marked with colored vertical bars. The earliest arriving phase (P) is marked in green, with later arriving phases in different colors. The secondary phase (S) is shown in orange and clearly shows the increased delay for signals arriving at more distant stations, even though these stations are fairly close (less than 25°) from the earthquake source. (See Figure 13.3.)

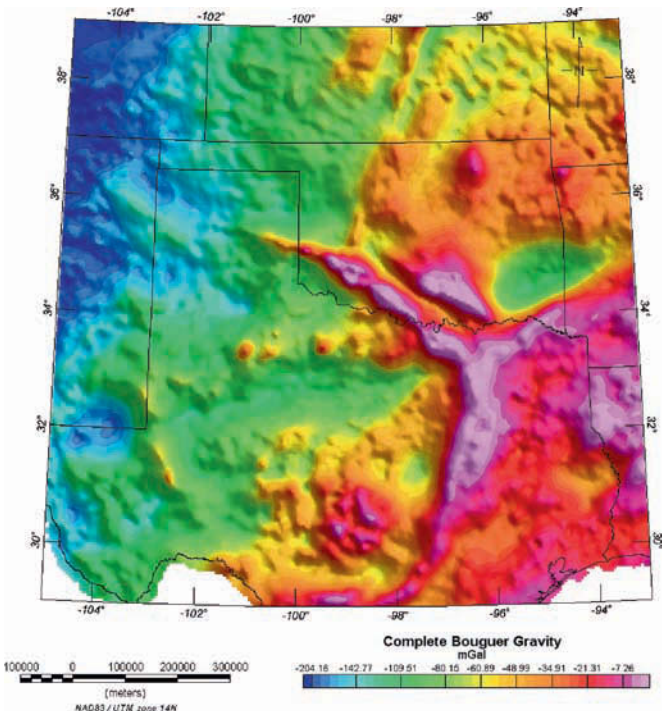



Plate 30. Complete Bouguer anomaly map of a portion of south-central USA. (See Figure 14.2.)

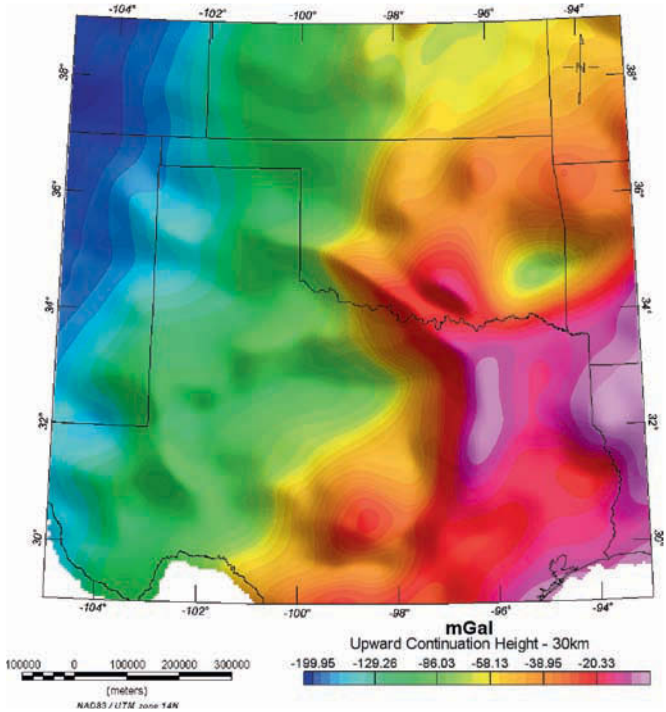



Plate 31. An upward continued version of the Complete Bouguer anomaly map shown in Figure 14.2. This physics-based procedure simulates the gravity anomaly map that would result if the readings had all been made 30 km above the Earth's surface. As is evident, this procedure enhances the long-wavelength features in the data. (See Figure 14.6.)

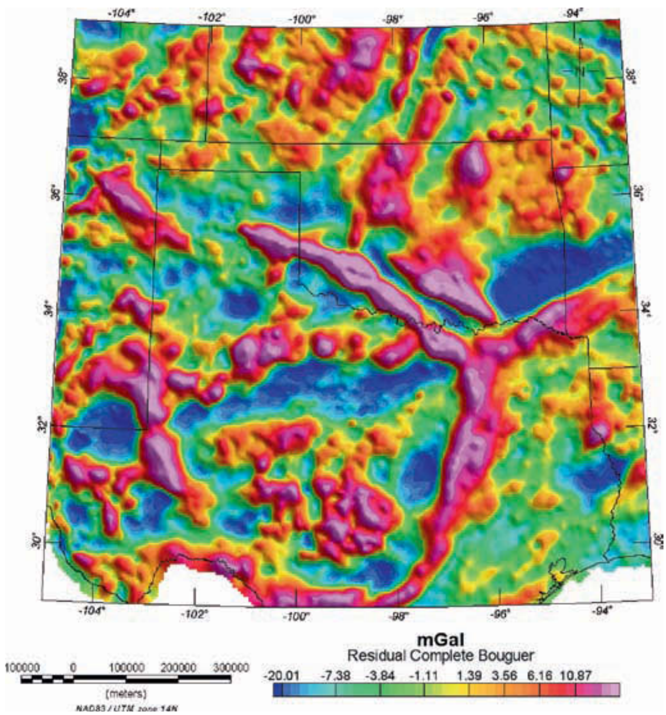



Plate 32. Residual gravity anomaly map obtained by subtracting the values in the grid that is visualized in Figure 14.6 from the original anomaly data grid that is visualized in Figure 14.2. The result is to enhance short-wavelength features. (See Figure 14.7.)

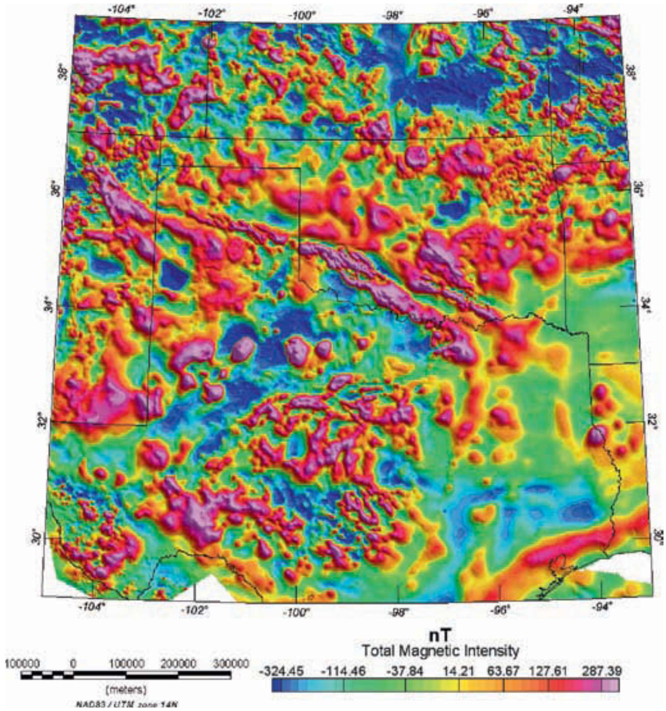



Plate 33. Total magnetic intensity map of a portion of south-central USA. The data were downloaded from our data system and visualized using the workflow in Figure 14.5. (See Figure 14.8.)

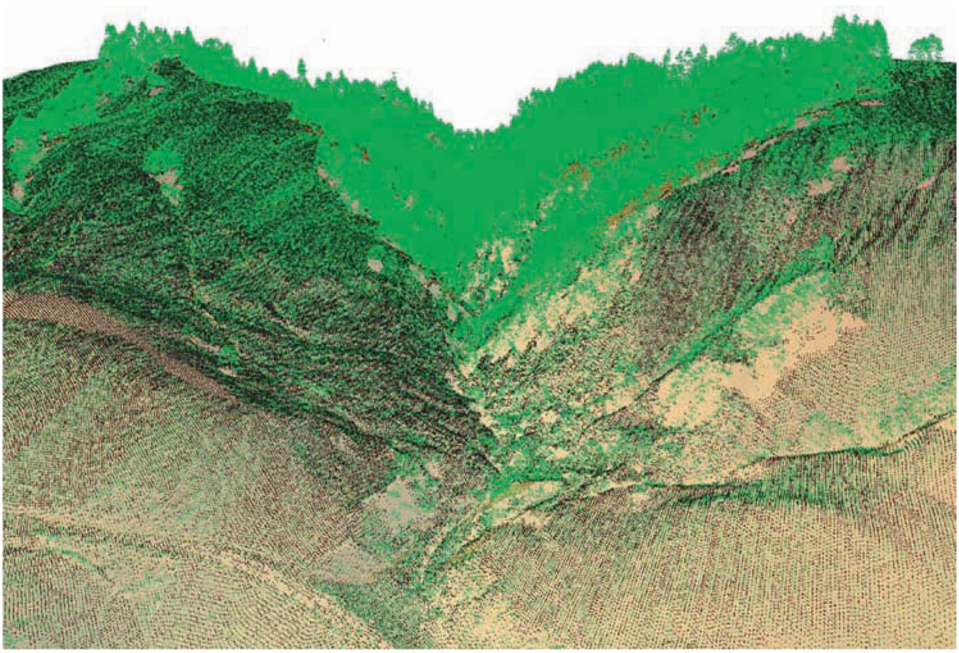


Plate 34. Three dimensional rendering of LiDAR ground returns (black) and vegetation returns (green) over the bare earth digital elevation model (DEM), shown in brown. (The Northern San Andreas Fault data were acquired by NASA, in collaboration with the United States Geologic Survey and the Puget Sound LiDAR Consortium, with funding provided by NASA's Earth Surface and Interior Focus Area. The data was collected in 2003 and processed by Terrapoint. The data are in the public domain with no restrictions on their use.) LiDAR instruments typically record multiple returns for each outgoing laser pulse. Individual points can be classified by their source by applying a filtering algorithm. (See Figure 16.1.)

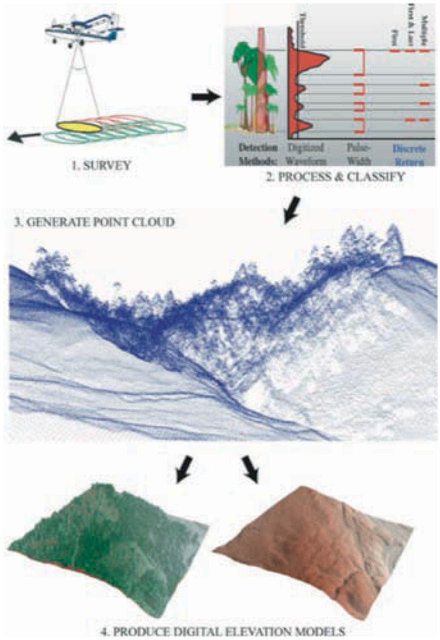



Plate 35. Generalized aerial LiDAR acquisition and processing workflow consisting of four steps: (1) Aerial data acquisition – scanner, IMU, and GPS (modified from R. Haugerud, USGS: http://duff.geology.washington.edu/data/raster/lidar/About_LIDAR.html), (2) processing of laser ranging, GPS and IMU data to generate LiDAR point cloud (modified from Harding, 2006), (3) generation of classified point cloud, and (4) generation of digital ground and vegetation models. Point cloud classification and generation of digital ground and vegetation models directly controls the specifications of the products upon which analyses will be performed. (See Figure 16.2.)

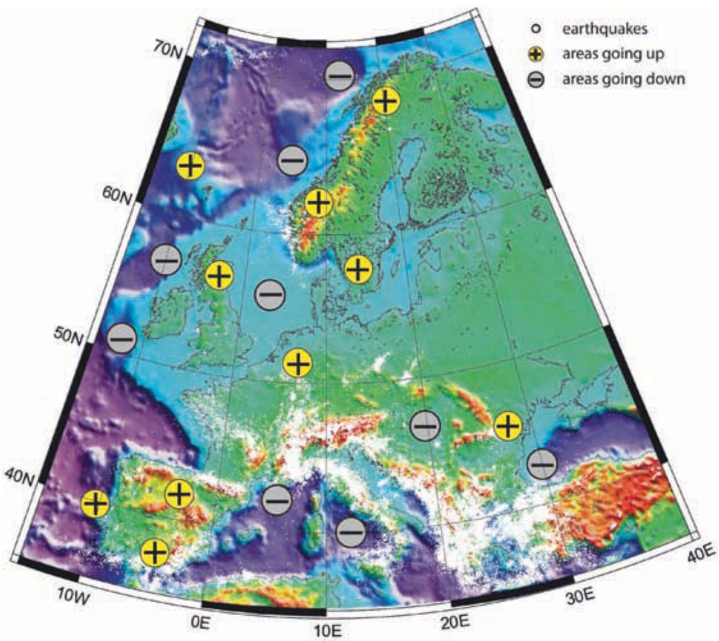



Plate 36. Present-day active intraplate deformation in Europe. White dots indicate location of earthquake epicenters. Also shown are intraplate areas of Late Neogene uplift (circles with plus symbols) and subsidence (circles with minus symbols). Background elevation images are extracted from the ETOPO2 dataset. (See Figure 19.2.)

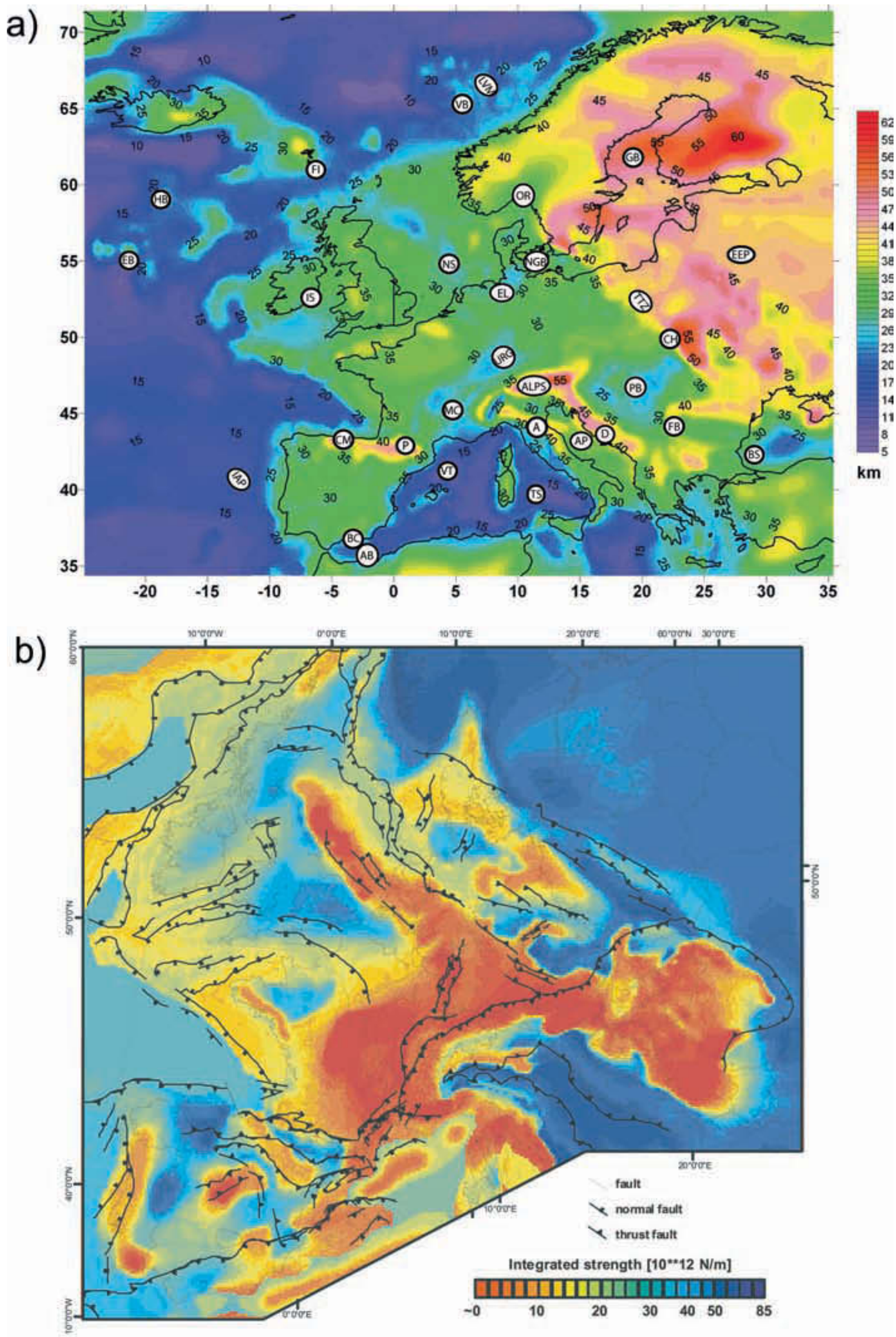



Plate 37. (a) European Moho depth (km). Abbreviations are as follows: A, Apennines; AB, Alboran Basin; AP, Adriatic Promontory; BC, Betic Cordillera; BS, Black Sea; CH, Carpathians; CM, Cantabrian Mountains; D, Dinarides; EB, Edoras Bank; EL, Elbe Lineament; EEP, East European Platform; FB, Focsani Basin, FI, Faeroe Islands; GB, Gulf of Bothnia; HB, Hatton Bank; IAP, Iberian Abyssal Plain; IS, Iapetus Suture; LVM, Lofoten-Vesterålen margin; MC, Massif Central; NGB, North German Basin; NS, North Sea; OR, Oslo Rift; P, Pyrenees; PB, Pannonian Basin; TS, Tyrrhenian Sea; TTZ, Tesseyre-Tornquist zone; URG, Upper Rhine Graben; VB, Vøring Basin; VT, Valencia Trough. From Tesauro *et al*. (2008).

(b) Integrated strength map for intraplate Europe (after Cloetingh *et al*., 2005). Colors represent the integrated compressional strength of the total lithosphere. Adopted composition for upper crust, lower crust, and mantle is based on a wet quartzite, diorite, and dry olivine composition, respectively. Rheological rock parameters are from Carter and Tsenn (1987). The adopted bulk strain-rate is $10^{-16}\,s^{-1}$. Main structural features are from Ziegler (1988) and Dèzes *et al*. (2004). (See Figure 19.3.)

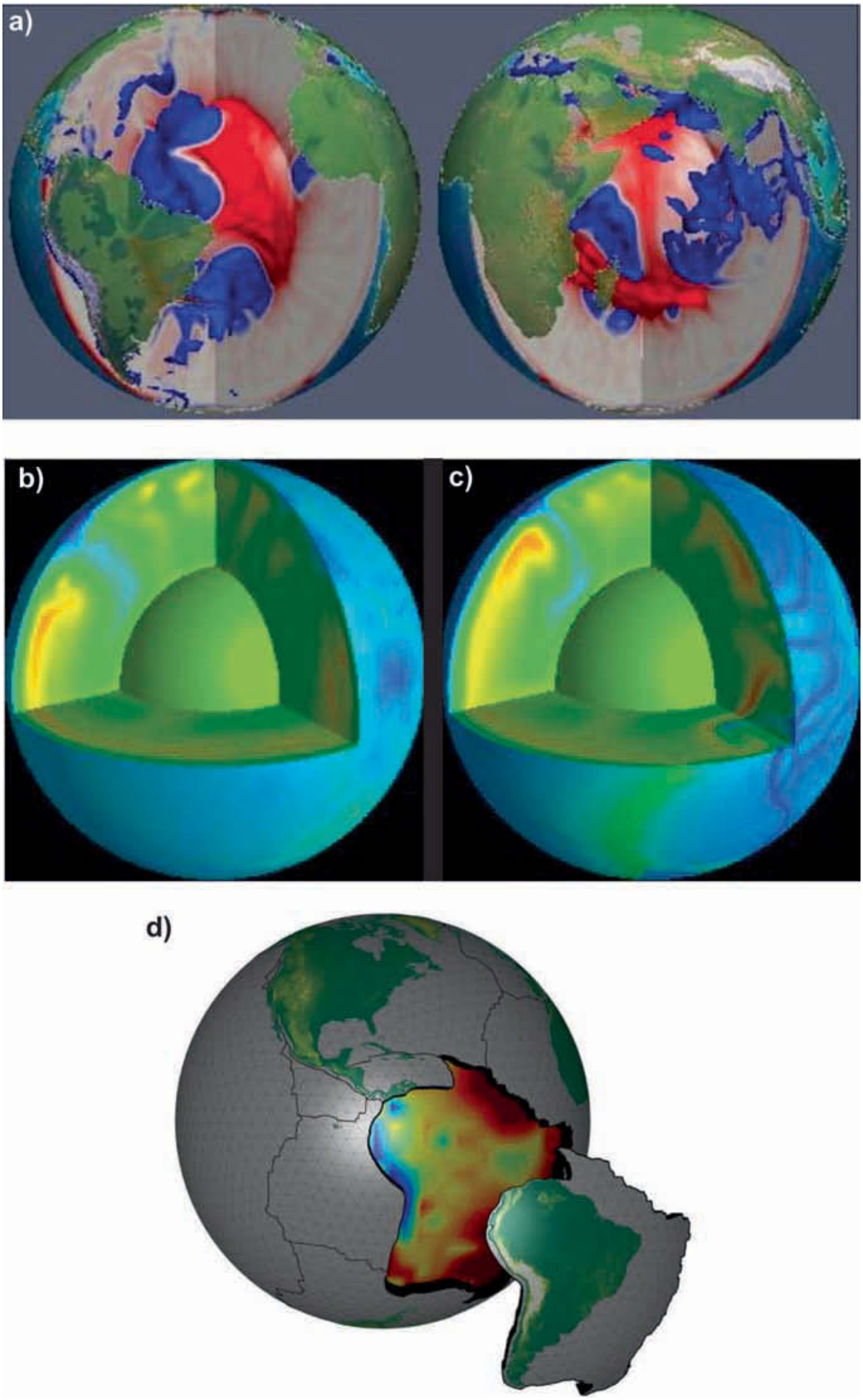



Plate 38. (a) Lateral temperature variations in two cross-sectional views from a high-resolution Mantle Circulation Model (MCM). Temperature iso-surfaces are taken to be at −600 and +400 kelvin. Continents are overlain for geographic reference. (b) Cutaway of the initial 3-D temperature field for a reconstruction model with a coupling between mantle convection and plate tectonics. Sample reconstructions using a perturbed mantle circulation model with a coupling of mantle convection and plate tectonics (after Bunge *et al*., 2003). Blue is cold, red is hot. The upper 100 km of the mantle is removed to the show the convective planform. (c) Same as (b) but after 100 Myr of present-day plate motion has been imposed. (d) Global neo-tectonic plate model incorporating realistic mantle-related buoyancy driving forces taken from a high-resolution MCM. After Iaffaldano *et al*. (2006, 2007) and Iaffaldano and Bunge (2008). (See Figure 19.4.)

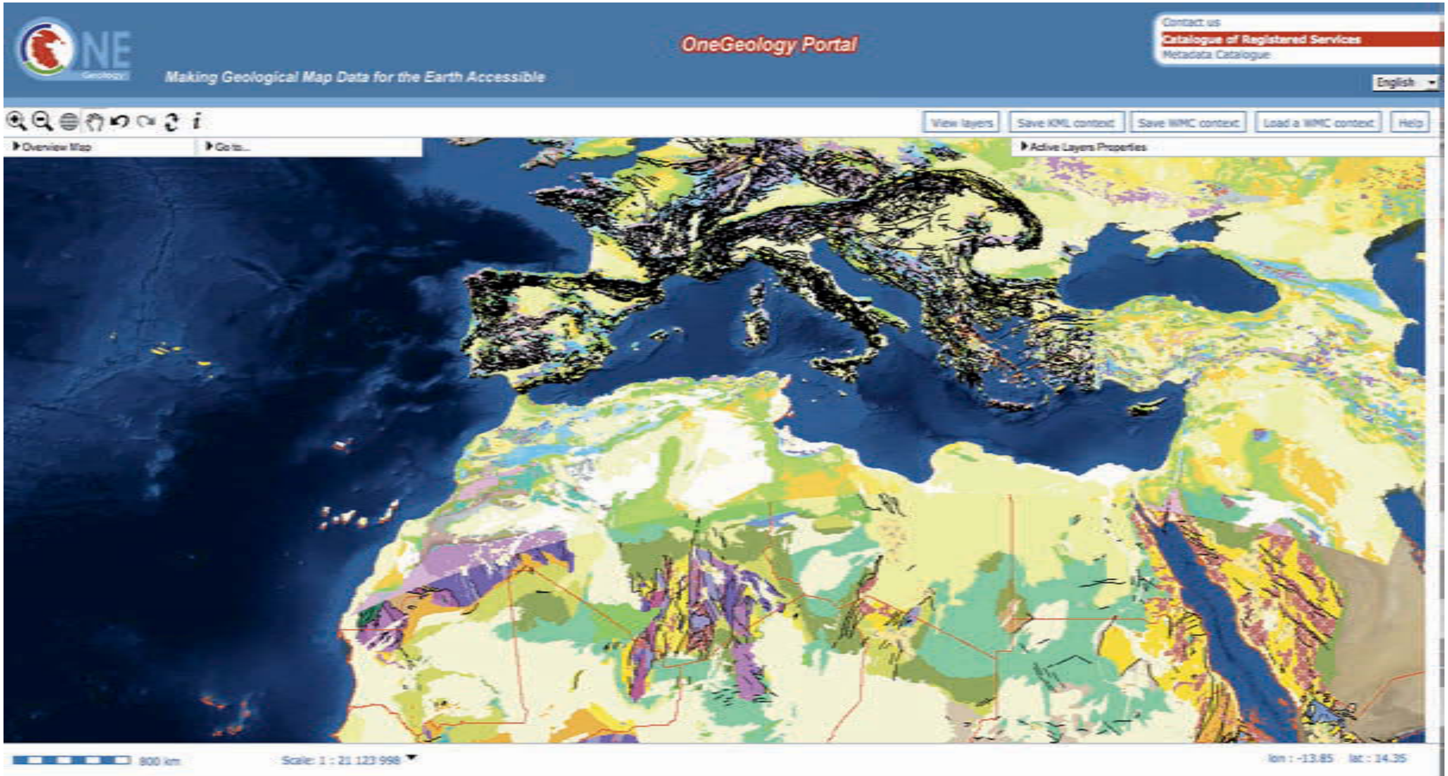



Plate 39. Screen capture of the OneGeology portal with selected geological layers displayed. (See Figure 20.2.)

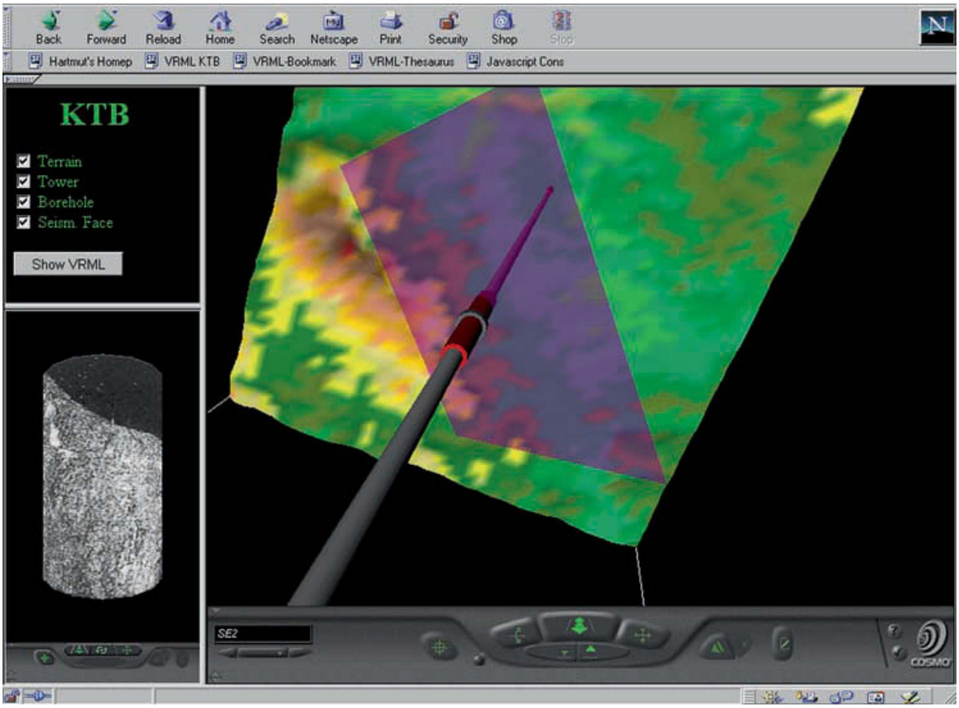



Plate 40. Screenshot of a borehole visualization from the German Continental Deep-Drilling Program (KTB). The visualization uses a VRML plug-in in a Netscape browser. (See Figure 21.1.)

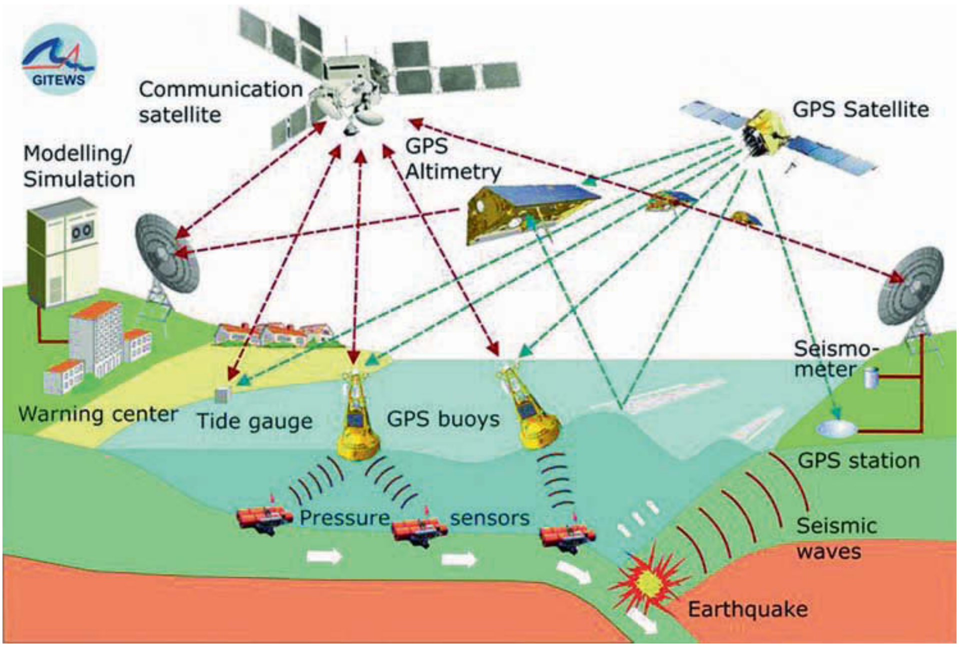



Plate 41. Schematic system diagram of the German-Indonesian Tsunami Early Warning System (GITEWS). The early warning system is based on the GITEWS sensor system platform (Tsunami Service Bus) to link sensor networks and models for a decision support system. (See Figure 21.2.)

# Part III

Visualization and data representation

# 8

# Visualization of seismic model data

AMIT CHOURASIA

## 8.1 Introduction

Computational modeling and simulations are becoming popular test beds and proving grounds in several areas of science. The seismology community is no exception to this trend, and the increased capability and availability of computational power has fueled further interest in performing such simulations using models. Over the last two decades, scientists have been making steady improvements to computational models, and we are just beginning to see the results of their efforts. The current models are capable of running on a variety of computational resources at multiple scales and sizes. One approach for studying the simulation models is by creating *probable* earthquake scenarios. These simulations require considerable computational and storage capacity for a successful run, and the amount of data produced by these simulations is enormous, requiring a non-trivial amount of processing for analysis of the results. Traditional analysis of model outputs, using statistical summarizations or simple examination of output data is impractical. Visualization offers an alternative approach for analyzing such huge stockpiles of data.

## 8.2 Visualization characteristics

It is important to identify a few key characteristics that any visualization should possess which can help us in designing and implementing a visualization solution for a given problem. We discuss the role of desirable visualization characteristics including *intuitiveness*, *trainability*, *focus*, *interactivity*, and *accessibility*.

- **Intuitiveness**
  Intuitiveness is fundamentally important for any visualization. This entails how the underlying data are represented or transformed to the visual domain. For example, consider some possible visual representations for showing ambient temperature in a room. One representation could be a smooth color-mapped image where each

colored pixel in the image represents the temperature at that point in the room. Another way to represent the same data would be to use an array of spheres, each colored from values in the data. Another method could depict the data as a height field surface. Or, perhaps a combination of color and height field could be employed. While all of these approaches aim to represent and convey the same information visually, some methods excel over others. Ideally, the visual representations should be clear, concise, simple, and unambiguous – in other words, they should be intuitive.

- **Trainability**
  While intuitiveness is an ideal goal for any visual representation, accomplishing this goal is a rather difficult task. Since synthetic representations are not always intuitive, one sometimes has to rely on training to enable people to interpret visualizations. This training effort could range from negligible in some cases to very significant in special cases. For example, the visual representations used for hurricane forecasts are relatively simple and may require little or no training to understand, but the representations used by neurosurgeons may be very complex and might require years of training experience to effectively interpret results.
- **Focus**
  Visualization generally has as its goal the understanding of some specific aspect of the data or simulation result. This is an essential motivation to embark on the visualization process. For instance, the goal in the above example was to show the temperature variation and distribution. This is a hard task to do just by interpreting raw data. The ability to focus quickly down to features of interest is where visualization excels.
- **Interactivity**
  Each visualization exercise has a specific focus that must be considered and accounted for in developing the interaction model. An ideal visualization will require the least amount of adjustments and provide mechanisms for interaction when needed.
- **Accessibility**
  Accessibility is the mode through which people access the visualization. Often this is an overlooked aspect of visualization. For example, a CAVE[1] environment offers a limited access, while geobrowsers like Google Earth™ are widely accessible.[2] Virtually all methods have some limitations, but a well-thought-out choice for accessibility could go a long way towards making visualization usable.

## 8.3 Visualization case studies

This chapter discusses three case studies on the visualization of seismic data. Two of the cases are for earthquake simulations: one based on the 1906 San Francisco

[1] CAVE information: www.evl.uic.edu/pape/CAVE/. [2] Google Earth information: www.earth.google.com.

earthquake and the other based on an earthquake prediction scenario in southern California. The third case study is based on structural engineering data recorded from a Shake Table experiment.

### *8.3.1 Case Study 1: TeraShake earthquake simulations*

TeraShake earthquake simulations are a set of seven different simulations (Cui *et al.*, 2007; Cui *et al.*, 2008; Cui *et al.*, 2009; Olsen *et al.*, 2006; Olsen *et al.*, 2008), four of which are based upon a Kinematic rupture called TeraShake1, and the remaining three named TeraShake2 are based on dynamic rupture of the fault. The simulations create a scenario for a magnitude 7.7 earthquake along the southern San Andreas Fault. The simulation's spatial domain covers a region that is 600 km long, 300 km wide, 80 km deep and is divided into a uniform mesh with 3000 × 1500 × 400 grid cells at 200 m in each direction.

*Computational schema* Before we delve into the specifics of the visualization process, it is important that we understand conceptually the simulation execution and the vital characteristics of the resulting data. Figure 8.1 illustrates the division of labor to carry out the simulation. Of the total 240 processors used for the simulation, each processor was responsible for solving the computational equations for a smaller subset region and for communicating with its neighboring zones. The simulation operated in discretized time increments and each processor computed the results based on input factors. Once the results are computed for a particular time step, all processors then write the chunk of data from their respective memory to data files in parallel. Typically, one file is created per time step per variable.

*Data attributes* It is also important for the visualization to understand the essential characteristics, or attributes, of the data. These data attributes may generate specific requirements and introduce limitations when undertaking the visualization process. For the simulations discussed here, the first data attribute is its temporal nature. The data comprised 22 728 time steps at an interval of 0.011 seconds, for a total time of 250.08 seconds. Next, the data are multivariate, implying that there are several variables represented in the datasets. Finally, the data are heterogeneous, consisting of 1-D seismograms, 2-D velocity fields, and 3-D velocity fields. The seismograms and 2-D velocity fields are recorded for every time step, while the volumetric data are recorded at either 100th or 1000th time step for efficient computation.

The data are stored in simple raw binary format, which is a brick of floating-point numbers. However, the coordinate system used by the data does not correspond with a left-handed coordinate system typically used by visualization systems, thus the data were translated by developing a customized data loader. One of the other big

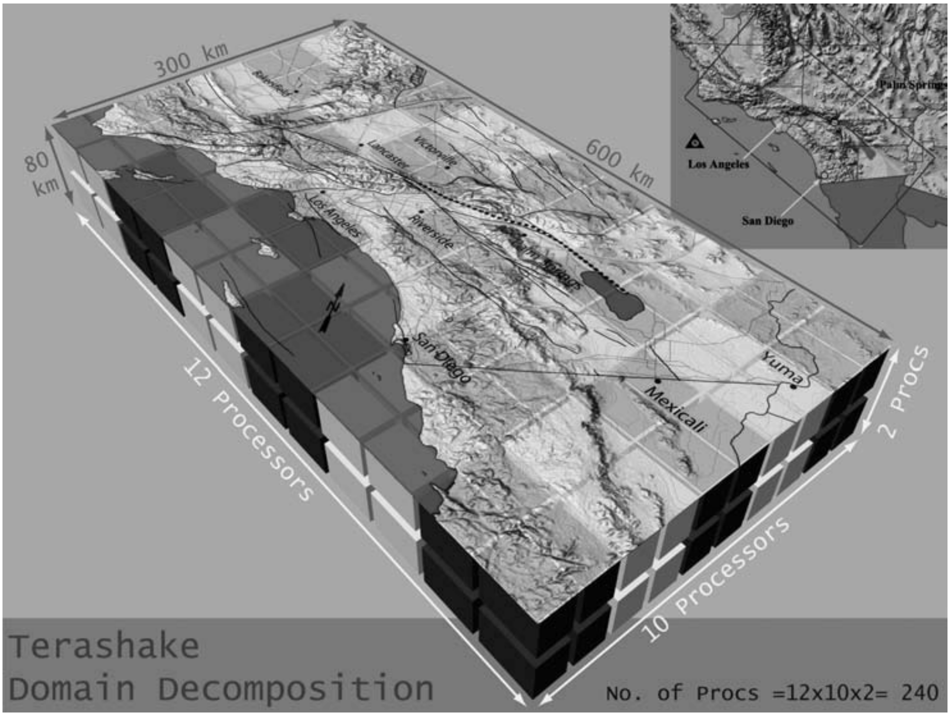



Figure 8.1. The top right inset shows the geographic location of the simulation region depicted by red rectangle. In the center, the topography, fault lines, and city locations are visible. This also shows the domain decomposition of this region into 240 processors. See color plates section.

challenges was data management and accommodating the very large datasets on available storage systems (Faerman *et al.*, 2007), which involved transferring data to different file systems, and from active storage to archival storage and vice versa, while ensuring no data corruption.

*Visualization motivation* The visualizations aimed to address a few key issues, one of which was to verify and validate the simulation. However, the most important aspect of the visualization was, of course, to gain scientific insight from the data. Finally, the visualization also helps in the ability to share information not only with experts but also with the wider audience.

*Visualization techniques* The TeraShake simulation produced several types of data products; thus different methods and techniques were applied to visualize these data. Here we categorize these techniques in five different forms.

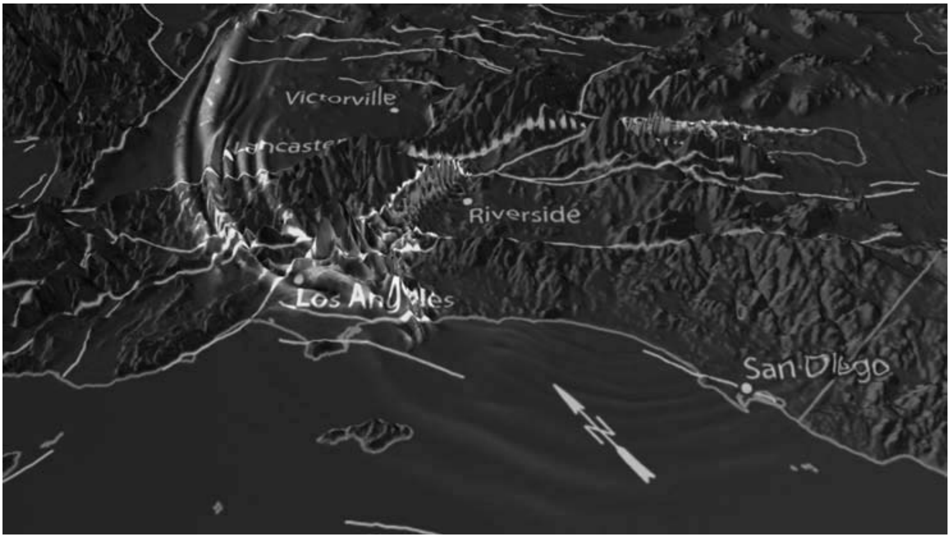



Figure 8.2. Snapshot showing deformation along the vertical axis of the terrain-based velocity magnitudes. See color plates section.

*Surface and topography visualization*

The key interest of seismologists is to analyze the impact of earthquakes on the Earth's surface. In this technique, we processed the 2-D surface data via direct 24-bit color maps overlaid with contextual geographic information. Annotations and captions provided additional explanations. Layering the geographic information with high-resolution simulation results provided precise, insightful, intuitive, and rapid access to complex information. Seismologists need this to identify clearly ground motion wave-field patterns and the regions most likely to be affected in San Andreas Fault earthquakes. We created surface visualizations for all 2-D data products using this method.[3] However, while the surface visualization was very useful for quickly interpreting seismic impact, the local features were harder to interpret. The topography visualization process used dual encoding of the surface velocity data as both color mapping and as displacement mapping. We used the surface velocity data to create a color-mapped image, and the displacement magnitude calculated from the surface velocity data to generate a gray-scale image. The system used the gray-scale image as a displacement map to create terrain deformation along the vertical axis (see Figure 8.2). The animation[4] corresponding to

[3] TeraShake visualizations: www.visservices.sdsc.edu/projects/scec/terashake/.

[4] Topography visualization movie for TeraShake 1.3 simulation: http://visservices.sdsc.edu/projects/scec/terashake/movies/TeraShakeNorthAsTopo.mov (10.4 MB);http://visservices.sdsc.edu/projects/scec/terashake/movies/Terashake2.1_narrated_720x405.mov (30 MB).

Figure 8.2 lets the scientist gain a better understanding of the kind of waves propagating through the model.

### *Volume visualization*

The surface and topography visualization are of central interest, but the bulk of the data and activity happens below the ground. The size of the data was a major hurdle for visualization since there were over 2270 time steps for three velocity components. Each volumetric snapshot was over 6.7 GB resulting in an aggregate of 44.6 TB of data.

Volume visualization is a method of creating images directly from 3-D volume data. There are several methods to achieve volume rendering, one of which is direct volume rendering using ray casting (Kaufman and Mueller, 2004). Our study used direct volume rendering for volumetric visualization and composited it with contextual information to provide a holistic view of the earthquake rupture and radiated waves (see Figure 8.3). Several transfer functions were developed to highlight features of interest. Our initial work helped the seismologists see the general depth extent of the waves. For example, depending on the wave field's component, waves propagating predominantly in the shallow layers can be identified as surface waves. Such waves typically contain large amplitude and long duration and can be particularly dangerous to certain structures. However, more research, in addition to existing work, needs to be done to represent multivariate data in a unified visual format.

Additional challenges in volumetric visualization include how to present visually to the user a global understanding of the seismic waves' behaviour, while at the same time allowing them to examine and focus on localized seismic activity. This challenge is important as, often, only reviewing the global behavior of the seismic

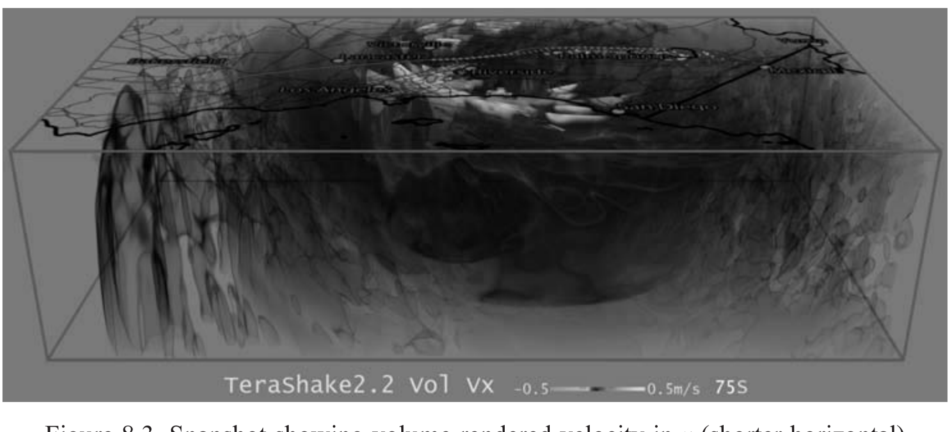



Figure 8.3. Snapshot showing volume-rendered velocity in $y$ (shorter horizontal) direction. The $z$-axis (vertical) has been scaled twice. See color plates section.

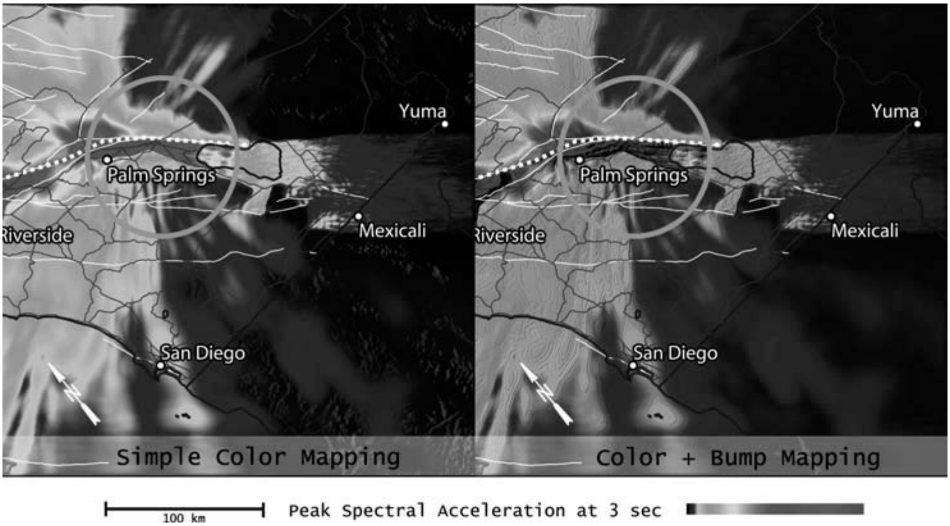



Figure 8.4. Map showing the advantage of the self-contouring technique in contrast to the simple color mapping. In the region highlighted by the orange circle, scientists identified a star-burst pattern. This indicates an unusual radiation of energy worthy of further investigation, which went unnoticed with simple color mapping. See color plates section.

wave hides important localized behaviors, while at the same time simply focusing on localized activity often hides how this localized motion impacts the waves' overall behavior.

### *Self-contouring technique*

We developed a technique to highlight features in 2-D by using bump mapping. A detailed analysis of this approach is provided in Wijk and Telea (2001). Encoding the spectral acceleration levels using both color maps and bump maps reveals subtle transitions between ground motion levels within localized regions with similar spectral acceleration properties. The color and bump encoding technique brought out variations in the data that were not previously visible (see Figure 8.4).

### *Web portal*

While the above visualizations have proved useful, there exists a need to recreate 2-D images with different color mapping and differing data range settings. We wanted to ensure that TeraShake could offer scientists hands-on analysis capabilities. This is important, because scientists need to be able to take this kind of approach to gain a better understanding of the output data. However, the size of TeraShake's data poses a significant problem for accessibility and analysis. We therefore developed a web front end where scientists can download the data and create custom visualizations over the Web directly from the surface data. The portal uses Linux, Apache, PHP, and Java technology for web middleware. On the back end it relies on specialized programs to fetch data from the archive, visualize, composite, annotate, and make them available to the client browser. Furthermore, there exists a need to view these maps in a richer geographic context in an interactive manner. The web portal fills this gap by creating outputs suitable to be viewed in modern and free geobrowsers like Google Earth$^{TM}$.

*Conclusion*

We visualized surface and volume data with different variables (velocities and displacements) and data ranges in multiple modes. The resulting animations have proven valuable not only to domain scientists but also to a broader audience by providing an intuitive way to understand the results. The visualizations played a significant role in the ability to easily detect several numerical errors and boundary condition failures. The visualization also proved very helpful in discussing various aspects of science and computation with images and movies as a reference. The use of multiple datasets with different visual representations (Figures 8.2–8.4) help the seismologists understand key earthquake concepts like seismic wave propagation, rupture directivity, peak ground motions, and the duration of shaking. The strong visual impact leads the viewer from the global context of earthquake hazards to the hazards in a specific region and then into the details about a specific earthquake simulation. Watching the earthquake simulation evolve over time helps viewers gain insight into both wave propagation and fault rupture processes. The simulation also illustrates the earthquake phenomena in an effective way to non-scientists. The visualization animations were also used for research dissemination to a wider audience. Some movies were used in television news coverage, and one of the animations was featured in a documentary entitled *LA's Future Quake* by the National Geographic Channel. An important result of this effort was that the research group became convinced about the value of visualization. They acknowledged the impact that visualization could have on verification and interpretation of results and have continued to incorporate visualization in their research endeavors. A subset of the data from one of the TeraShake simulations was made available to the visualization community as a challenge through the IEEE visualization conference.[5] A detailed visualization report for this study is described in the research article by Chourasia *et al.* (2007).

### *8.3.2 Case Study 2: 1906 San Francisco earthquake simulation*

In 2006, USGS Menlo Park conducted a detailed study of the magnitude 7.9 earthquake that occurred in 1906 in the San Francisco bay area. The major

[5] IEEE 2006 visualization design contest: http://vis.computer.org/Vis2006/session/contest.html.

motivation of this study was to understand the past scenario using modern computational tools, and to assess its potential impact in the current time-period. Here we discuss a visualization of one of the simulation datasets.

*Simulation and data* The simulation is based on a finite element model that uses tetrahedral topology to approximate the region having dimensions of 250 km long, 110 km wide, and 40 km deep. The tetrahedral elements on the ground surface are smaller to resolve the shorter wavelengths. The topography and bathymetry of the region are incorporated in the simulation model. The rupture is represented by propagating a dislocation on the tetrahedral mesh, and the ground motion is computed using the wave equation for elastic material.

The data are recorded only for the surface, which is an irregular triangular 3-D mesh consisting of about 65 000 triangles with more than 35 000 vertices. Velocity vectors and displacement vectors are recorded for 1200 time steps at an interval of 0.1 s for each of the vertices, yielding an aggregate data size of 1 GB that is stored in a custom binary format with floating-point precision.

*Visualization motivation* There are several possible methods to visualize these kind of data, which includes writing *ab initio* custom software or using existing visualization tools like Paraview,[6] Amira,[7] and EnSight.[8] While these standard methods have been used routinely, this study used the Maya software package for visualization. Maya offers a robust and flexible customization interface on top of excellent mesh, texture, and particle manipulation tools. Maya also helped shorten development time for the visualization. The goal of visualization was to create realistic visual representation of the data.

*Visualization process* The visualization for this study involved several steps. First, the surface topology information was used to create a 3-D triangular mesh, which was imported into Maya. Then the raw data comprising displacements and velocity components are translated into a Particle Disk Cache (PDC) format using a custom library. Then the 3-D mesh was coupled with a particle system, which has a particle for each vertex. This particle system was in turn driven by the PDC files, which contain displacement and color information for each vertex. The gray-scale satellite imagery was used on the 3-D mesh as a texture to provide contextual information, while the color at each vertex was blended with the satellite image. Light and cameras were added for viewing the terrain from various angles. Then the Mental Ray renderer for Maya was used for final rendering of the image sequence, which was encoded into a movie using Adobe After Effects.

[6] Paraview information: http://vtk.org/Wiki/ParaView.
[7] Amira information: www.amiravis.com. [8] EnSight information: www.ensight.com.

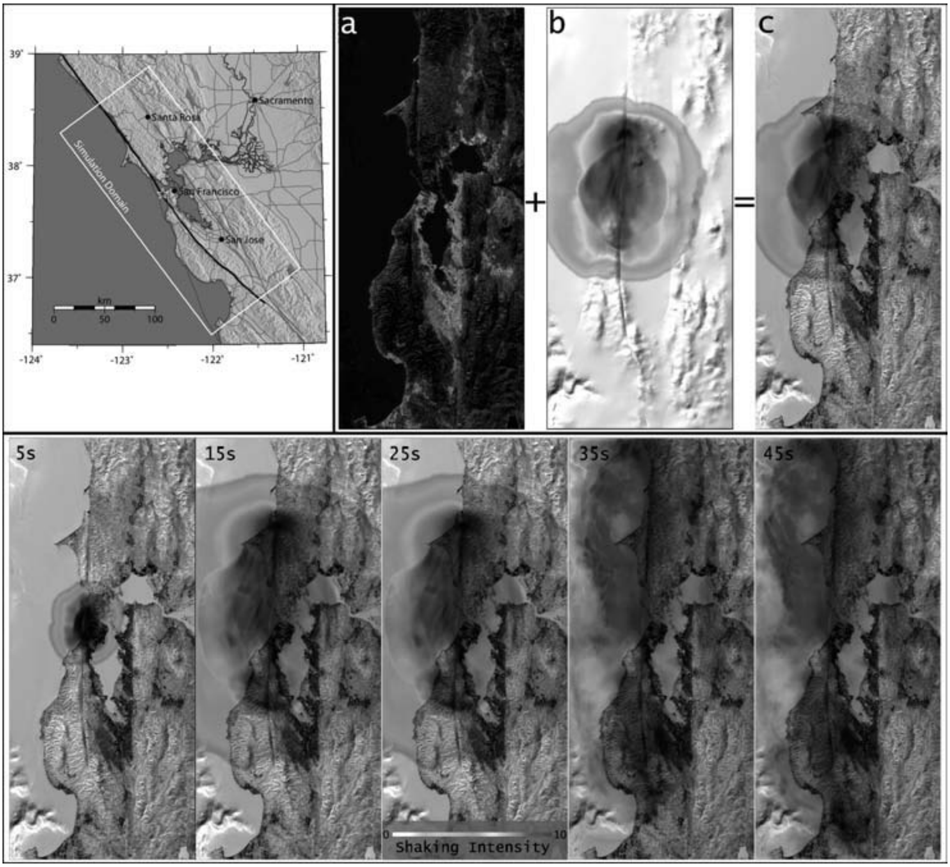



Figure 8.5. The top left inset shows a white rectangle indicating the simulation region 110 km long and 80 km wide. The top right inset shows the color blending of satellite imagery (a) with the coloring based on data (b) to produce the final image (c). The bottom image shows the earthquake's progression at 5, 15, 25, 35, and 45 seconds, respectively. See color plates section.

## *Conclusion*

The visualizations in Figure 8.5 portray the intensity of shaking using the same ground motion metric and color maps as ShakeMaps (Wald *et al.* 2005). The coloring of ground motion is based on modified mercalli intensity, which quantifies how the shaking is perceived by people and its potential effect on the manufactured structures on a scale of 1-X. The colors correspond to the peak intensity of shaking at each location up to the time on the animation frame. This visualization provides a vivid depiction of the earthquake, which includes topography and shaking and helps us to understand which areas undergo strongest shaking, as well as how the shaking is propagated by seismic waves. A high level of visual realism is achieved using the satellite imagery and state-of-the-art rendering without compromising accuracy.

This study demonstrates that animation tools could be used for scientific data with moderate data size and high-quality results. The results of this study were presented at the "100th Anniversary Earthquake Conference" in 2006 in San Francisco.[9, 10] A detailed visualization report for this study is described in the research article by Chourasia *et al*. (2008).

### 8.3.3 Case Study 3: Re-creation of seven-story Shake Table experiment

Our third case study for seismic visualization is based on a real experiment on a very large shake table. Structural engineers at University of California, San Diego, performed an experiment by shaking a 65 foot tall, 250 ton, seven-story tall building on the outdoor Shake Table based on the recorded data from the 1994 Northridge earthquake of magnitude 6.7. The building was fitted with several accelerometers and high-speed cameras to record several characteristics during the experiment. The goal of this study was to recreate this experiment in a virtual environment and present a concept of an integrated visual environment for data analysis.

*Visualization process* A 3-D model of the site that includes the seven-story building, the Shake Table, the control room, and surrounding elements was constructed using the digital blueprints. These models were textured using the photographs taken at the site to provide realism. The data of interest were chosen to be from the 115 sensors distributed across the entire building with 5250 time steps recorded at 50 Hz. Video footage was also available, which recorded the overall view of the experiment from a distant.

As in the previous case study, this study also chose Maya as the visualization platform. The seven-story building was fitted with a deformer skeleton based on the sensor locations. This skeleton was then driven by a particle system where each sensor corresponded to a particle and vertex on the skeleton. The particle system was in turn driven by the PDC files as in previous study, which had the information about displacements and colors for each particle. Each particle was rendered as a sphere that was colored based on its displacement from steady state. This system was interactively viewable from 360 viewpoints. A matching camera to the existing video footage was created which was used to render a sequence of images. This sequence of images was compared to the actual recorded video as a simple accuracy test. This posed several challenges since the camera characteristics, such as focal length and real-world position, were not available and there was no time synchronization of the data sources. Despite these difficulties, we were able to use

[9] 100th Anniversary Earthquake Conference commemorating the 1906 San Francisco earthquake: www.1906eqconf.org.
[10] 1906 San Francisco earthquake visualizations: http://visservices.sdsc.edu/projects/sfquake.

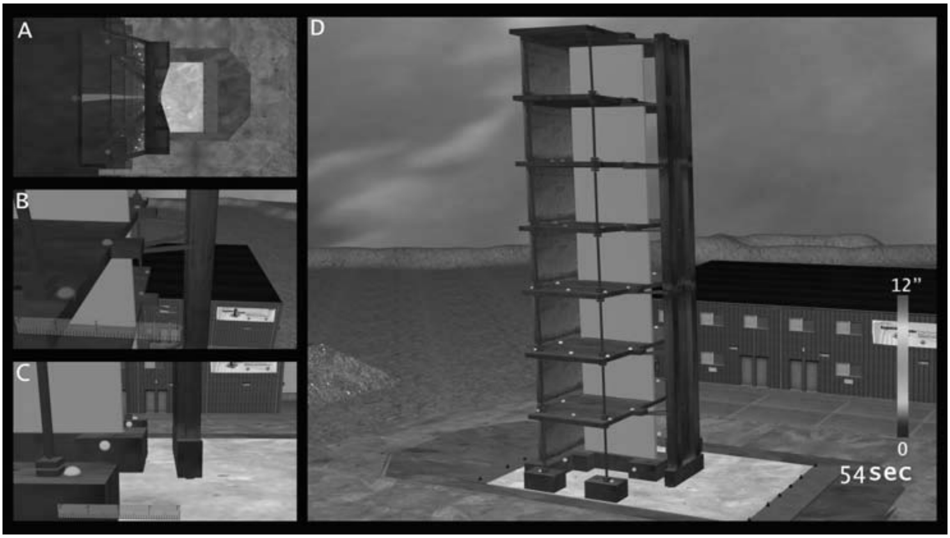



Figure 8.6. An illustration showing the particle system and a deformer skeleton. Each particle (sphere) corresponds to a node location on the deformation skeleton. The particle system is driven by PDC files translated from observed data. Each particle's color depicts the displacement at their location as shown in the color map. The position and color of the particles change with time based on observed data. An animation of skeleton deformation is available at the web site.[11] See color plates section.

well-judged approximations such that the recreated virtual environment was in good agreement with the real video footage (see Figure 8.6).

*Conclusion*

To summarize, this study was able to identify the basic requirement of a visualization visual system that could enable integration of disparate sources of data to be useful in analysis. It showed real-time interaction with realistic-looking models and realistic deformation could aid understanding. The study also demonstrated that animation tools can be customized for high-quality scientific visualizations. A detailed visualization report for this study is described in the research article by Chourasia (2007).

## 8.4 General discussion

The gap between analysis and ever-increasing data is widening at a rapid pace. The rate at which data produced by simulations and experiments are increasing makes it difficult for traditional methods of analysis to keep pace. This presents an

[11] Seven-story Shake visualization: http://visservices.sdsc.edu/projects/nees/article.php.

opportunity for novel data-analysis tools and techniques in data mining and visualization. Each project discussed here was a collaborative effort among computer and computational scientists and geoscientists (and earthquake engineers) who all contributed towards the final goals of the research. The visualizations have helped scientists easily identify numerical errors and other computational artifacts, demonstrating that visualization tools are effective in diagnosing such problems. In each case, the visualizations proved to play an important role, not only in discovery but also in the dissemination process.

Finally, while there are well-understood advantages of visualization, there are also several obstacles and potentially detrimental factors that need to be addressed in a responsible manner. Some of these issues include the understanding of domain knowledge when creating visualizations, the visual design used to represent multivariate data, techniques used to maintain temporal coherence and precision loss, interactivity of visualization versus batch visualization, perceptual issues in designing color maps, lighting, and displaying depth, and lastly the bias of the author and viewer, which are a big challenge to be addressed in an effective way.

## Acknowledgements

This work has been possible as a result of successful collaborations with several research groups and their contributions, as listed below.

**TeraShake**: I would like to thank Reagan Moore, Yifeng Cui, Steve Cutchin, Marcio Faerman, Yuanfang Hu, Jing Zhu, and Patrick Yau and other people at SDSC. I would also like to thank Phil Maechling, Kim Olsen, Steve Day, Bernard Minster, and Thomas Jordan and other scientists at SCEC. This work was partially funded by SDSC and SCEC. SCEC is funded by the US NSF Cooperative Agreement EAR-0106924 and US Geological Survey Cooperative Agreement 02HQAG0008. The SCEC contribution number for this article is 1085.

**1906 San Francisco earthquake**: I would like to thank Brad Aagaard, Steve Cutchin, Mike Rossmassler for their contributions to this project, and Mary Lou Zoback for her coordination on this project. Thanks to John Moreland and Peter J. Lu in helping with the PDC Utility.

**Seven-story Shake**: I would like to thank Anke Kamrath for envisioning and making this project possible and Steven M Cutchin and Lelli Van Den Einde for continuous feedback, support, and liaison with structural engineers. I would also like to thank Jose Restrepo and Marios Panagiotou for providing the data and explanation of the experiments in detail, Mike Rossmassler for creating the 3-D model of the site, and Ruben Soto-Alcauter and Andrew Collier for help in data gathering and translation. Also thanks to the folk at NEES who participated in discussions.

Last, but not least, I would like to thank John Moreland of the San Diego Supercomputer Center for his feedback on several of these projects and his review of this chapter.

## References

Chourasia, A. (2007). Digital re-creation of a seven-story building shake during an earthquake. *ACM Crossroads*, **13**(3): 6–8.

Chourasia, A., Cutchin, S. M., and Aagaard, B. (2008). Visualizing the ground motions of the 1906 San Francisco earthquake. *Computers & Geosciences*, **34**: 1798–1805, doi:10.1016/j.cageo.2008.01.012.

Chourasia, A., Cutchin, S. M., Olsen, K. B. *et al*. (2007). Visual insights into high resolution earthquake simulations. *IEEE Computer Graphics and Applications*, **27**(5): 28–34.

Cui, Y., Chourasia, A., Moore, R. *et al*. (2009). The TeraShake computational platform for large-scale earthquake simulations. In *Advances in Geocomputing.* Lecture Notes in Earth Sciences, Vol. 119. Berlin/Heidelberg: Springer-Verlag, pp. 229–278, doi:10.1007/978-3-540-85879-9_7.

Cui, Y., Moore, R., Olsen, K. *et al*. (2007). Enabling very-large scale earthquake simulations on parallel machines. In *Computational Science – ICCS 2007*. Vol. 4487. Berlin/Heidelberg: Springer Verlag, pp. 46–53.

Cui, Y., Moore, R., Olsen, K. B. *et al*. (2008). Toward petascale earthquakes simulations. In *Acta Geotechnica*. Berlin/Heidelberg: Springer Verlag, pp. 1861–1125, doi:10.1007/s11440-008-0055-2.

Faerman, M., Moore, R., Cui, Y. *et al*. (2007). Managing large scale data for earthquake simulations. *Journal of Grid Computing*, **5**(3): 295–302, doi:10.1007/s10723-007-9072-x.

Kaufman, A. and Mueller, K. (2004). Overview of volume rendering. In *The Visualization Handbook*, 1st edn.ed. C. D. Hansen and C. R. Johnson. Boston: Elsevier(Academic Press), pp. 133–138.

Olsen, K. B., Day, S. M., Minster, J. B. *et al*. (2006). Strong shaking in Los Angeles expected from southern San Andreas earthquake. *Geophysical Research Letters*, **33**, L07305, doi:10.1029/2005GL025472. 2006.

Olsen, K. B., Day, S. M., Minster, J. B. *et al*. (2008). TeraShake2: Simulation of Mw7.7 earthquakes on the southern San Andreas fault with spontaneous rupture description. *Bulletin of the Seismological Society of America*, **98**(3): 1162–1185, June 2008, doi:10.1785/0120070148.

Wald, D. J., Worden, B., Quitoriano, V., and Pankow, K. (2005). ShakeMap Manual: Technical Manual. *Users Guide, and Software Guide*, Version 1.0 ed., United States Geological Survey, 156pp.

Wijk, J. J. V. and Telea, A. C. (2001). Enridged contour maps. In *Proceedings of IEEE Visualization.* Washington, D.C.: IEEE CS Press, pp. 69–74.

# 9

# The UNAVCO GEON Integrated Data Viewer for exploration, analysis, and integration of geoscience data

STUART WIER AND CHARLES MEERTENS

## 9.1 Introduction

Modern geoscience data, like the Earth that they represent, often are 3-D and time-varying, with complexity on all length scales. Finding and studying significant features in such data is greatly improved with new approaches that use interactive 3-D visualizations.

As part of the Geosciences Network (GEON) development efforts, UNAVCO developed the GEON Integrated Data Viewer (IDV), a research-quality application for interactive 3-D display and analysis of geoscience data. The GEON IDV is designed to meet the challenge of investigating complex, multivariate, time-varying geoscience data (Meertens and Wier, 2007; Meertens *et al.*, 2005; Meertens *et al.*, 2006).

The Geosciences Network is a multi-institutional NSF ITR (Information Technology Research) project, and a scientist-centered cyberinfrastructure initiative that provides earth scientists with an array of tools, including data access and integration mechanisms, computational resources, and software for analysis, modeling, and visualization. Needing a powerful tool for visualization and integration of earth science data, one path the GEON team chose was to build on the UCAR (University Corporation for Atmospheric Research)/Unidata Integrated Data Viewer, an open-source, meteorologically oriented, platform-independent application for visualization and analysis. Unidata's IDV brings together the ability to display and work with satellite imagery, gridded data, surface observations, balloon soundings, NWS WSR-88D Level II and Level III RADAR data, and NOAA National Profiler Network data, all within a unified interface.

## 9.2 Software design

The Unidata IDV (Meertens *et al.*, 2004; Murray *et al.*, 2003) is a well-established package in meteorology, designed to provide true 3-D displays with time animation, using an interactive graphical user interface.

The Unidata IDV is designed for the conceptual setting of cyberinfrastructure – access to remote data sources using differing protocols – and it combines powerful visualizations with computational and analysis capabilities. The Unidata IDV software is designed to allow extension of the core software. It was not specially built for a limited set of particular data types. Unidata has invested more than 10 person-years in development of the IDV, and provides active online support for IDV use and software development. GEON, looking for ways to visualize geoscience data with full 3-D interactivity, time controls, and using the latest forms of distributed data access, chose to extend the UCAR/ Unidata IDV.

The IDV is Java code, with an object-oriented design suitable for adding new features. It can accept Java "plug-ins," additional code that extends or modifies IDV features. On the software level, plug-ins are subclasses to Java classes in the IDV core code. The IDV uses editable XML configuration files that control behavior. The GEON IDV software is a Java plug-in to the Unidata IDV that adds new features for solid earth geophysics and retains all the features of the original IDV.

The GEON IDV is open source, and runs on any system supporting Java and Java3D, including Windows, Linux, MacOS X, and Solaris. It can be quickly installed from instructions on UNAVCO's GEON IDV web site. The IDV usually is run as an interactive desktop application, but may be run as a scripted background process for automatic generation of images of IDV displays.

## 9.3 Data source types

The GEON IDV displays most earth-located numerical data, including 2-D and 3-D grids, observations at scattered locations on or inside the Earth, raster imagery, GIS shape files, and values along a geographic path such as a ship's track or vertical sounding. Observations and model output may have time variations, and may have a multi-parameter character, such as vectors with error estimates.

Raster images can come from photographic imagery, multi-spectral scanning, interferometric SAR, and derived products such as are available on web map servers (WMS). A significant barrier to combining imagery with other geoscience data in mapped displays is georeferencing the image. WMS sources also serve up this information; the IDV can connect to a WMS data source and display selected map imagery. The GeoTiff format allows for this in the raw image specification, but is not widely used yet. For raster files with no internal location metadata, such as jpeg or png files, simple IDV "ximg" files specify image corner locations in latitude, longitude, and depth.

## 9.4 Data formats

UNAVCO has a goal of facilitating data access and interoperability, with efficient search, exploration, and use of data from multiple data sources, by promoting web services and standards for geoscience data formats. Retention and propagation of semantics and metadata with observational and experimental values is essential for interoperability and understanding diverse data sources. Standard data formats are essential for this goal. Best practices for IDV data have formats and sources with metadata for earth locations, standard data unit names, and error indications.

The IDV relies on the power of the NetCDF file format. NetCDF is a software package with data format specifications, file creation and editing software, and online support. NetCDF files are self-describing. NetCDF files are binary and machine independent, and contain data, attributes, and georeferencing information. NetCDF is very suitable for earth-located data, such as grids or random locations, with one or more parameters and, optionally, time variations. NetCDF files should include georeferencing on the Earth, and should contain unit names and descriptions for every observed variable. NetCDF files shared with others may also, but need not, contain information such as data source attribution, revision levels and dates, authors, institutions, instrument types, processing methods, contact information, related publications, web sites, and sources of funding. NetCDF is well supported by the OPeNDAP data distribution system. Using OPeNDAP the online user can select and download a geographic subset of a large global grid. NetCDF is used by many applications, including the Generic Mapping Tool (GMT), Matlab, and GRASS 3D.

NetCDF data can be used for most numerical geoscience data, including 2-D and 3-D grids, scattered observations, tracks, and soundings. Multiple depths, time steps, and parameters can be stored in one file. The IDV can use NetCDF metadata from two or more files of differing sources for automatic unit conversions and remapping of grids when making computations and displays.

Recognizing the need for data reformatting to use the IDV, UNAVCO has put effort into explaining the NeTCDF formats and conventions, and also has created some conversion programs to convert some simple ASCII file formats into NetCDF. These materials are available from the UNAVCO web site.

For point data, observations at irregular or scattered locations, the IDV also recognizes a class of simple comma-separated value ASCII files. This format is useful for anything from GPS station names to earthquake epicenters to vertical soundings to aircraft tracks.

## 9.5 Data access methods

An essential ingredient in true data integration is removing the barriers to accessing, reading, and mapping the data. That is, ease of access to data from diverse and remote sources, the ability to read the data formats, and understanding of the meanings of the georeferenced coordinates and variables. The IDV has made considerable progress to achieving this end-to-end process (Figure 9.1).

The IDV is designed for remote and local access to data, including distributed data from several protocols and sources, including local files, URLs to files on FTP or HTML servers, OPeNDAP data servers, web catalog servers such as THREDDS (Domenico *et al.*, 2002; Nativi *et al.*, 2005), web map servers (OGC WMS, WFS, and WCS), and RSS feeds. The IDV has data readers for NetCDF files, KML and KMZ files, GIS shape files, DEM files, WMS server data, and for several specialized ASCII file formats for geoscience data such as earthquake focal mechanisms.

An example of an online data service that the IDV uses is the IRIS Earthquake Browser (IEB). The IRIS earthquake catalog has more than 1 900 000 worldwide earthquake locations with depths and magnitudes. On the IEB web site, the user specifies a search for earthquake epicenter data anywhere on Earth by location, depth, and magnitude, using an interactive map tool, and the service then makes a data file from the earthquake catalog. The IDV user copies the URL of the file at

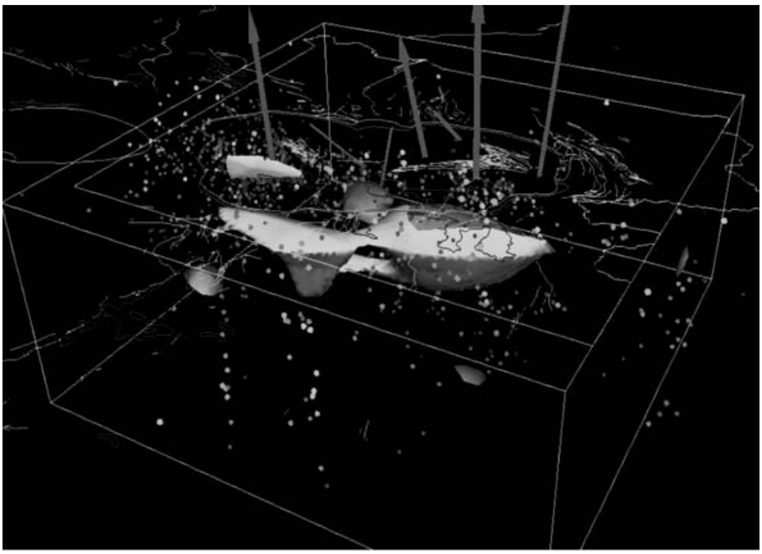


Figure 9.1. IDV display showing a 3-D oblique view from the southeast of the Yellowstone National Park area, with surface fault lines and caldera outline (orange and red, U.S. Geological Survey *et al.*, 2006), seismic tomography isosurfaces (yellow), earthquake hypocenters sized by magnitude (blue), GPS velocity vectors (red; Chang *et al.*, 2007; Husen *et al.*, 2002; Puskas *et al.*, 2007), and isosurfaces of a fluid intrusion model (green; Vasco *et al.*, 2007). See color plates section.

IRIS into the IDV's Data Chooser, without downloading the data file. The IDV reads the data directly from IRIS; no file is copied. The combination of the IRIS earthquake catalog, the IRIS IEB search service, and the IDV makes a powerful mapping system for new and historic earthquakes. The IDV can provide displays ranging from a simple map view of seismicity to a 3-D perspective movie of aftershock progression by time, including other geological, geophysical, and cultural data if desired.

## 9.6 GEON IDV visualizations

The GEON IDV is a true 3-D display system, ideal for exploring and displaying any data located on the Earth, inside the Earth, and above the Earth's surface. The GEON IDV shows otherwise hidden details, features, and relationships among diverse data sources. It has full interactive controls on point of view, zoom, pan, rotation and time, a large selection of map projections, and adjustable vertical scales from centimeters to planetary dimensions. The IDV has globe and flat map views, and standard and user-created color scales for coloring any display by data value.

Display types include earth-located raster imagery, 3-D topographic relief, point data symbols and labels, 2-D grids shown as contours or colored by value, vertical and horizontal cross sections and contours of gridded data, vertical profiles, *x*–*y* charting, time series of point observations (Figure 9.2), data transects, and data tracks or paths with time and parameter values along track such as aircraft measurements, drifting buoys, and seismic ray paths.

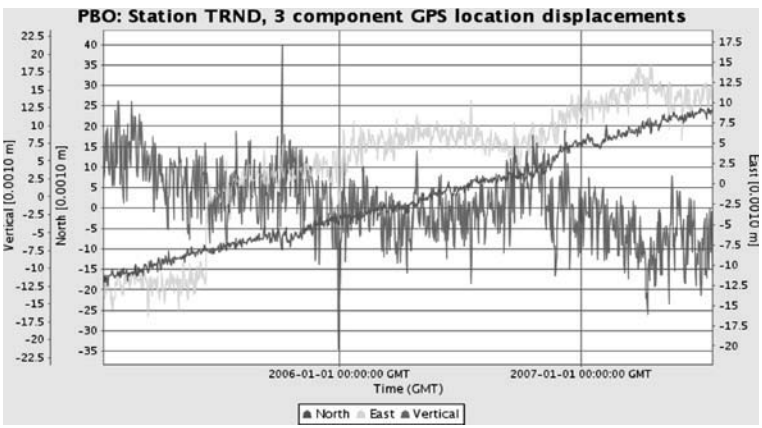



Figure 9.2. IDV time display of GPS displacements at the Plate Boundary Observatory station TRND.

Raster imagery may lie in horizontal or vertical planes, at any angle or depth, or be draped over a 3-D relief surface of any other parameter, including topography. An IDV display can drape a colored map of chemical species concentration on 3-D surface "topography" of the same data to emphasize high- and low-value regions, or drape the same concentration image on a 3-D relief surface of topography.

Volumes, data in 3-D grids, can be rendered with isosurfaces, 3-D surfaces of a single data value (Figures 9.1 and 9.3). Isosurfaces have shaded single colors, and IDV isosurfaces also can be colored by values from another gridded parameter, for example, to show uncertainty in the isosurface's data, varying by location.

Special GEON IDV symbols and displays for geoscience include, but are not limited to, GPS velocity vectors with error ellipses, earthquake focal mechanism "beachballs," earthquake locations colored by magnitude or depth, seismic ray paths in 3-D colored by a parameter such as seismic wave velocity, seismic

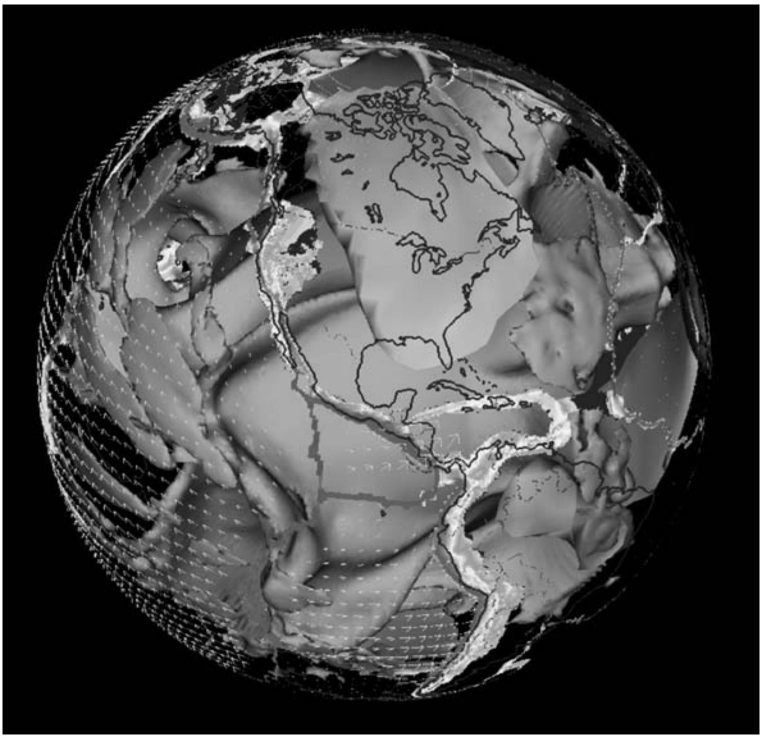


Figure 9.3. GEON IDV display of four diverse global datasets: a mantle convection isosurface (50% of core-mantle boundary temperature, orange; McNamara and Zhong, 2005), a seismic tomography isosurface (positive 3% anomaly surface, green; Gu and Dziewonski, 1999), and plate motion vector symbols with the associated image of the second invariant of global strain rate (Kreemer *et al.*, 2003). See color plates section.

anisotropy symbols, 3-D fluid flow vectors, earth strain axes and strain field imagery, geology maps and vertical cross sections (from imagery), and InSAR imagery draped on 3-D topographic relief.

True stereographic visualizations are easy with the IDV. The GEON IDV can drive a GeoWall or other stereo viewing system, such as any projector with circular polarization and "flicker glasses," or two plane-polarized projectors and plane-polarized glasses. The IDV includes a stereo mode with dual channel output for stereo viewing. All these methods preserve color in stereo displays.

## 9.7 Data integration in displays

The GEON IDV supports simultaneous display of datasets of differing types and from differing sources (Figures 9.1, 9.3, and 9.4). The IDV correctly maps all the data sources in a single display using each one's georeference information. The IDV also allows simultaneous linked displays of different data types (Figure 9.4) so that features of one data source are not hidden by another. Linked displays respond identically to interactive zoom and pan controls.

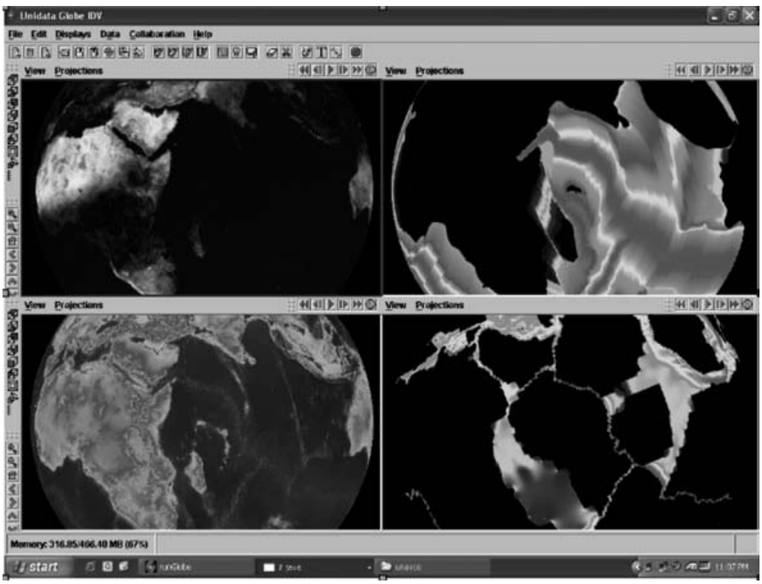



Figure 9.4. IDV display with four linked globe displays showing, from upper left, the Blue Marble image from the JPL OnEarth web map service, ages of ocean floor (Müller *et al.*, 1997), topography colored by elevation (National Geophysical Data Center GLOBE Digital Elevation Model), and the second invariant of strain rate (Kreemer *et al.*, 2003) colored by value. Interactive rotation and zoom links all four displays together. Multiple linked flat map displays may also be created in the IDV. See color plates section.

The IDV saves displays as imagery for publications and presentations, movies for presentations and web sites, and KML files for use in Google Earth$^{TM}$. New development is planned to allow the IDV to run as back-end support for a web site, making images of data displays in response to user requests on a web site.

## 9.8 IDV formulas and computations

The IDV has built-in data processing. For example, the IDV can compute and display relations between gravity anomalies, elevation, crustal thickness, and crustal densities, or compute and display the horizontal divergence of a grid of plate-motion vectors, showing where plates are deforming, using internal IDV vector calculations. Barnes analysis creates a regular data grid from scattered observational values.

If a computation calls on multiple gridded data sources with differing data units and data locations, the IDV automatically converts units and interpolates grid values between locations to create the resulting grid. A simple example is the difference between two seismic tomography models, in 3-D grids. In this case the difference is computed and displayed simply with the formula "a-b" and with a data source browser to indicate which two sources are grids "a" and "b." If the two 3-D grids' point locations differ, and if the units differ, the IDV automatically does interpolations and unit conversions before making the display. The user need not write code to loop over all the grid points in both grids; that is implicit in the IDV "a-b" formula.

Users may also enter their own one-line mathematical expression in the IDV as a new "formula" for functions not built into the IDV. For more complex analyses, users can write complete software programs for data processing, to include in the IDV computational library. The embedded mathematical capability uses Jython, a form of the Python language that runs inside Java. Python is a complete modern language with a very simple syntax. The IDV can be used for extensive and complex data analysis, and to simply describe computational concepts.

## 9.9 IDV bundle files and saving state

The state of the IDV includes information about all its data sources, displays, map projections, scales, color tables, display symbols, and points of view. The IDV saves its state in named "IDV bundle files." Later a bundle file may be used to recreate exactly a previous IDV configuration and display. Bundle files are small; they usually do not contain data values, though they may. They tell how to recreate the display from the data, rather than copying the displayed data or copying a static image of a display.

Bundle files have many uses, such as "snapshots" of incoming observations to archive progress of an experiment, or exchanging bundle files with colleagues to discuss some observed phenomena, data, or processing. Bundle files can be posted online for widespread duplication of IDV displays; for example, they can be auxiliary "figures" in online publications, allowing readers of a report to see and examine the data in 3-D, interactive, IDV displays. Bundle files can be used for classroom assignments and to submit homework.

## 9.10 IDV use in research geophysics

As an example of GEON IDV use, investigating a subduction zone could create displays including any or all of seismic tomography results (isosurfaces, vertical cross sections, and horizontal cross sections), seismic ray paths, seismicity, focal mechanisms, gravity anomalies, geoid heights, volcano locations, 3-D surface relief, GPS velocity vectors, seismic anisotropy splitting, geology maps, geology cross sections, imagery of seismic reflection profiles, refraction line models, and any other appropriate geoscience data. Recent work has demonstrated the IDV's value in seismic tomography (Wier *et al.*, 2006), mantle convective flow, the time-varying topography of California during the past 3 million years, and studies of the present earth crust using GPS displacements.

## 9.11 GEON IDV installation

The GEON IDV is available from UNAVCO's GEON web site (http://geon.unavco.org). Installation usually requires less than ten minutes. The UNAVCO GEON web site provides a tutorial, examples of IDV use in geophysics, standard geophysical datasets, a complete description of formatting data files for the IDV, and software tools for data file format conversion to NetCDF.

## 9.12 Current status

The GEON IDV is a very powerful tool to explore geophysical data using displays and computations. The IDV is a true 3-D interactive display and analysis system for earth science data. Strengths of the IDV include powerful display controls, time animation from seconds to millions of years, integration of very diverse datasets, direct access to remote data sources, and built-in computations. The IDV saves results as display images and animations, for research, presentation figures, publications, and educational materials. Bundle files allow other IDV users to recreate 3-D IDV representations of observations and results.

New features are being added to the Unidata IDV by Unidata, and to the GEON IDV by UNAVCO. Both institutions maintain extensive web sites with online help, and access to standard data sources ready for use in the IDV.

## Acknowledgements

This work was supported by NSF award EAR-0408228. Don Murray and Jeff McWhirter of UCAR/Unidata provided essential help in the development of GEON IDV software. Data for testing and validation were provided by the scientists named in the figure captions. GPS data services were provided by the Plate Boundary Observatory, operated by UNAVCO for EarthScope (www.earthscope.org) and supported by the National Science Foundation (No. EAR-0323309).

## References

Chang, W-L., Smith, R. B., Wicks, C., Farrell, J. M., and Puskas, C. M. (2007). Accelerated uplift and magmatic intrusion of the Yellowstone caldera, 2004 to 2006. *Science*, **318**(5852): 952–956, doi: 10.1126/science.1146842.

Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S. (2002). Thematic real-time environmental distributed data services (THREDDS): Incorporating real-time environmental data and interactive analysis tools into NSDL. *Journal of Digital Information*, **2**: 114.

Gu, Y. and Dziewonski, A. M. (1999). Mantle discontinuities and 3-D tomographic models. *Eos Transactions AGU*, **80**, F717.

Husen, S., Waite, G. P., and Smith, R. B. (2002). Seismicity and tomographic imaging of the Yellowstone caldera as a function of time, Abst. 14th Annual IRIS Workshop, Waikoloa, Hawaii, June 12–16, 2002.

Kreemer, C., Holt, W. E., and Haines, A. J. (2003). An integrated global model of present-day plate motions and plate boundary deformation. *Geophysical Journal International*, **154**: 8–34.

McNamara, A. K. and Zhong, S. (2005). Degree-one mantle convection: Dependence on internal heating and temperature-dependent rheology. *Geophysical Research Letters*, **32**, L01301, doi:10.1029/2004GL021082.

Meertens, C. and Wier, S. (2007). UNAVCO software and services for visualization and exploration of geoscience data. *Eos Transactions AGU*, **88**(52), Fall Meet. Suppl., Abstract IN33A-0872.

Meertens, C. M., Murray, D., and McWhirter, J. (2004). Collaborative visualization and analysis of multi-dimensional, time-dependent and distributed data in the geosciences using the Unidata Integrated Data Viewer. *Eos Transactions AGU*, **85**(47), Fall Meet. Suppl., Abstract SF13A-0710.

Meertens, C., Estey, L., Wier, S. *et al*. (2005). Visualization of earth science data. *Eos Transactions AGU*, **86**(52), Fall Meet. Suppl., Abstract IN23C-03.

Meertens, C. M., Wier, S., Murray, D., McWhirter, J., and Memon, A. (2006). The GEON IDV (Integrated Data Viewer) for data exploration and discovery in the geosciences. *Eos Transactions AGU*, **87**(52), Fall Meet. Suppl., Abstract IN31A-1306.

Müller, R. D., Roest, W. R., Royer, J.-Y., Gahagan, L. M., and Sclater, J. G. (1997). Digital isochrons of the world's ocean floor. *Journal of Geophysical Research*, **102**(B2): 3211–3214.

Murray, D., McWhirter, J., Wier, S., and Emmerson, S. (2003). The Integrated Data Viewer: a web-enabled application for scientific analysis and visual. Preprints. *19th International Conference on IIPS for Meteorology, Oceanography, and Hydrology*.

Nativi, S., Domenico, B., Caron, J., and Bigagli, L. (2005). Extending THREDDS middleware to serve the OGC community. *Geophysical Research Abstracts*, **7**, 06508, Sref-ID: 1607–7962/gra/EGU05-A-06508.

Puskas, C., Smith, R. B., Meertens, C. M., and Chang, W. L. (2007). Crustal deformation of the Yellowstone-Snake River plain volcanic system: Campaign and continuous GPS observations, 1987–2004. *Journal of Geophysical Research*, **112**, B03401, doi:10.1029/2006JB004325.

U.S. Geological Survey, Idaho Geological Survey, and Montana Bureau of Mines and Geology (2006). Quaternary fault and fold database for the United States, accessed 2006, from USGS web site: http://earthquakes.usgs.gov/regional/qfaults/.

Vasco, D. W., Puskas, C. M., Smith, R. B., and Meertens, C. M. (2007). Crustal deformation and source models of the Yellowstone volcanic field from geodetic data. *Journal of Geophysical Research*, **112**, B07402, doi:10.1029/2006JB004641.

Wier, S., Bensen, G. D., Meertens, C. M., Keller, G. R., and Seber, D. (2006). Integrated analysis of seismological models using GEON resources: An illustration from the western United States. *Eos Transactions AGU*, **87**(52), Fall Meet. Suppl., Abstract S51D-03.

# Part IV

## Knowledge management and data integration

# 10

# Data and tools for geologic timelines and timescales

PETER M. SADLER AND CINZIA CERVATO

## 10.1 Introduction

Geologic timescales provide the vocabulary for geologic correlation, bring organization to the discussion of diverse processes and events, and enable the assembly of geologic history with global scope. Since the early attempts to calibrate a geologic timescale (Holmes, 1947; 1960), new stratigraphic techniques and radioisotopic systems have been brought to bear on the task; the volume of relevant information has increased dramatically; and the most recently published calibrations attempt finer resolution and explicit estimates of uncertainty (Gradstein *et al*., 2004). These trends have encouraged the use of statistical tools and custom databases that assemble local timelines of events and numerical ages, using the stratal thickness scales of measured sections and drill cores. These are the initial field and laboratory products that precede correlation, calibration, and statistical manipulations. Combining the local timelines into regional and global composites is a hard intellectual challenge. The named divisions of the geologic timescale are a much simpler conventional overlay upon these temporally integrated datasets. The application of state-of-the-art information technology and the optimization tools of operations research to large time-stratigraphic datasets introduces the possibility that timelines may be custom-built 'on the fly' from specific data queries and the user's choice of correlation algorithms. The print media and the labor of calibration tend to limit the familiar conventional geologic timescales to infrequent update and limited resolving power. This apparently stable nature tends to obscure many of the underlying uncertainties and assumptions of calibration. Time scales could be transformed by information technology into dynamic and interactive, yet reproducible, products that would be nimble enough to explore the consequences of different data selections or changing geographic scope. The users' discomfort or inconvenience in seeing that timescales are not unique or stable solutions should be outweighed by the deeper understanding of the real scientific exercise involved.

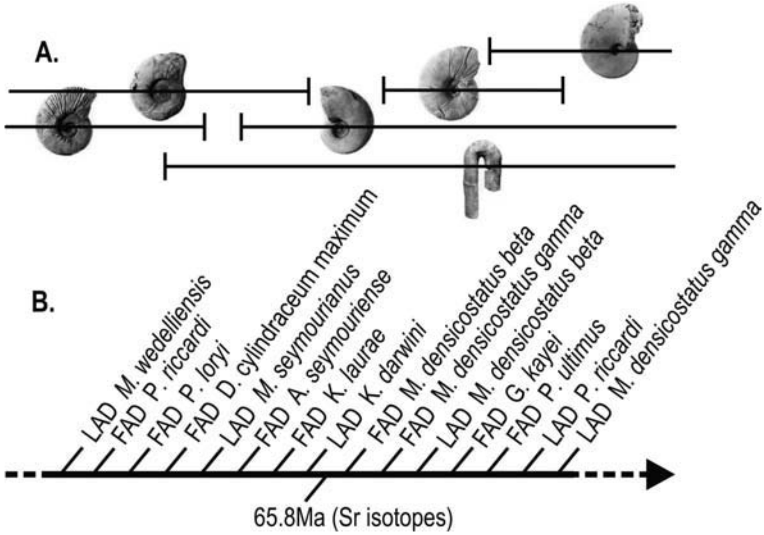



Figure 10.1. A fragment of a taxon range chart for Maastrichtian ammonites (A) and the same data embedded in a timeline (B). Time progresses from left to right.

We describe here the types of geologic events that can be assembled into timelines, the consequences of their different information content, and the demands they place on any integrated framework of databases and online tools that might be designed to support a fully dynamic product.

## 10.2 Building geologic timelines

Television detectives draw timelines to display events in order of occurrence. They insert numerical dates and times in the appropriate positions and use the diagrams to explore opportunity, cause, and effect. Stratigraphers' measured sections, well logs, and taxon range charts are timelines also (Figure 10.1). Using scales of rock thickness, they display the local order of events such as the appearance of new species, geomagnetic polarity reversals, geochemical anomalies, and dated ash falls (Figure 10.2). They contain the raw time-stratigraphic information that supports biodiversity histories, paleoclimate reconstructions and many other derivative timelines, plus the familiar geologic timescale. This timescale differs from the timelines in one crucial aspect, which we emphasize by distinguishing the two terms: timelines display a sequence of ancient events, each of which has one correct relative position on the line, however difficult this might be to infer correctly; the geologic timescale displays a human construct of named intervals that partition deep time. Both timelines and timescales need to be calibrated to numerical age scales.

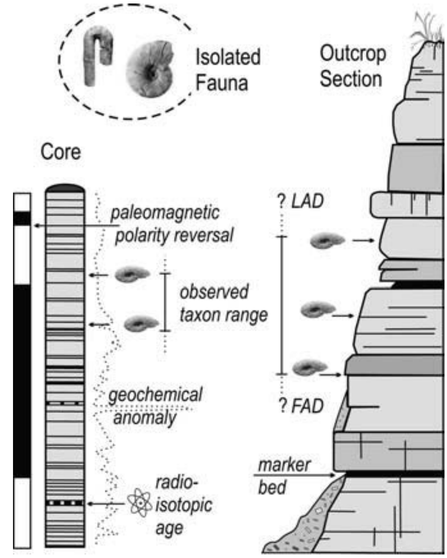



Figure 10.2. Summary of the three common sources of information about taxon ranges and coexistences, with examples of other stratigraphic data types suited to correlation and inclusion in timelines.

Building timelines entails several distinct inferential tasks. First comes the selection of the events that can be meaningfully ordered. This includes decisions about quality and correlation: deciding whether a fossil species has been identified consistently, for example, or whether the same turbidite bed has been recognized in different cores, or whether radioisotopically dated crystals determine the true age of an ash fall. The second task is seriation: inferring the order of occurrence of events that were observed at various localities and placing them into one ordered sequence that is as consistent as possible with all the local ordering information. For reasons that will soon become clear, difficulties in seriation arise from local differences in the preserved order of the most numerous events used for timelines and correlation – the first and last appearances of fossil. Optional third and fourth tasks estimate the relative sizes of the time intervals between events and interpolate numerical ages from radioisotopically dated events to other events. All these tasks place demands on the information content of publications and databases that describe the stratigraphic record. Some of the demands are discipline-specific matters of quality control; others are the minimum information requirements that determine whether or not a particular seriation tool can be applied.

Partitioning the geologic timescale entails more convention and convenience. The purpose is to facilitate discussion about Earth history; names and subdivisions remain in the timescale if they fill a practical need and enjoy consensus. When we seek an ordered series of diagnostic events to define the named intervals or attempt to interpolate the ages of interval boundaries, the standard moves from consensus to correctness and precision; answers must be derived from the underlying timelines. Just as timelines differ from timescales, two fundamentally different types of databases support time stratigraphy. Some databases store the vocabulary and history of geologic timescales and, perhaps, expert judgments about the age of events, expressed in terms of the named time intervals. They might draw from custom timescales to fit user-selected data or respond to passive inquiries about named intervals. What does the term Whiterockian mean, for example? How has the estimated age of the Permo–Triassic boundary changed over time? This storage and relaying of expert inferences is invaluable, but not our primary concern here; we return to it at the conclusion of the chapter where we examine the possibility of combining these functions with the second type of database – those that manage the raw information needed to support geologic inferences about the position of events in timelines. We shall see that the minimum requirements for these databases are determined by the need to sequence information across databases. The requirements are conceptually simple: either radioisotopic data sufficient to place an event with others on a numerical age scale, or some means to constrain the age of one event relative to at least one other, on a scale of stratal superposition. We start by considering the wide variety of events commonly built into geologic timelines. By contrasting the information and constraints each provides about their position in line, we can specify in more detail the minimum information that databases must provide in order to support timelines.

## 10.3 Geologic events

For correlation and seriation purposes, taxon range ends, paleomagnetic reversals, sequence boundaries, and other events bring different advantages and disadvantages (Table 10.1). None is ideal. For example, we shall see how uniquely identifiable events, which could support unambiguous correlation, tend to be preserved in a fashion that obscures their true stratigraphic levels; for events whose stratigraphic level is unequivocal, a lack of uniqueness may permit one-to-many correlations. Datasets that combine different event types lead to the best-constrained timelines.

### *10.3.1 Taxon range ends*

Every fossil species provides two global events for geologic timelines and correlation – the moment of its origination, or First Appearance Datum (FAD), and the moment of its extinction, or Last Appearance Datum (LAD). So many fossil species are known (Teichert, 1956; Valentine, 1970) that the time intervals between range-end events must average less than 100 years and might approach 10 years.

Table 10.1. *Properties of stratigraphic events*

| | *Recognition / correlation* | | | *Stratigraphic freedom to adjust* | | | | *Interval type* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Unique | Binary | Multiple | Up only | Up or down | Down only | None | Present-absent | Alternating | Repeating | Moment | At hiatus? | Reliable ordering information |
| **Taxon range ends** | | | | | | | | | | | | | |
| Reworked last app. | X | | | | X | | | X | | | | Y/N | above FAD |
| Last appearance | X | | | X | | | | | X | | | Y/N | superposition only[2] |
| First appearance | X | | | | | X | | X | | | | Y/N | subposition only[1] |
| Caved first app. | X | | | | X | | | X | | | | Y/N | below LAD |
| **Reversals** | | | | | | | | | | | | | |
| Paleomagnetic reversal | | X | | | | | X | | X | | | Y/N | position[3] |
| Isotope cycle boundary | | X | | | | | X | | X | | | Y/N | position[3] |
| **Resettings** | | | | | | | | | | | | | |
| Sequence boundary | | | X | | | | X | | | X | | Y | position[3] |
| Flooding surface | | | X | | | | X | | | X | | Y | position[3] |
| **Spikes** | | | | | | | | | | | | | |
| Marker bed | ? | | ? | | | | X | | | | X | N | position[3] |
| Isotopic transient | ? | | ? | | | | X | | | | X | N | position[3] |

Table 10.1. (*cont.*)

| | *Recognition / correlation* | | | *Stratigraphic freedom to adjust* | | | | *Interval type* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Unique | Binary | Multiple | Up only | Up or down | Down only | None | Present-absent | Alternating | Repeating | Moment | At hiatus? | Reliable ordering information |
| Taxon acme | | | X | | X | | | | | | X | N | between FAD and LAD |
| **Dated ash fall** | | | | | | | X | | | | X | N | position[3] and age value[4] |

[1] must be older than type [2]and [3]events higher in section, but might be shifted with respect to lower events.
[2] must be younger than type [1]and [3]events lower in section, but might be shifted with respect to higher events.
[3] must be younger than type [1]and [3]events lower in section and older than type [1]and [3]events higher in section.
[4] age relative to other type [4]events determined by value in numerical age scale; does not require occurrence in same section.

This richness and the unique nature of species make FADs and LADs the most practical events for generating high-resolution timelines for the geologic past. But these advantages are offset by much less desirable properties (Table 10.1). The geographic range of a species is limited by ecological considerations. Furthermore, the time of first and last appearance of a species differs from place to place and few, if any, of these locally observed range ends will correspond with the global first- and last-appearance datums. The discrepancy arises naturally from the patchy and dynamic nature of species distributions. Incomplete preservation and collection add to the patchiness. If these were the only reasons that local first and last appearances departed from the global origination and extinction times, we could be sure that the observed taxon ranges fit within the longer true ranges. Unfortunately, there are situations in which this simplicity may be violated.

Last appearances may extend beyond the true range if fossils are reworked as detrital particles into younger sediments. Reworking is not limited to cases of large-scale erosion and redeposition into younger sediment, such as can be observed at sea cliffs and stream banks. Microfossils, in particular, are susceptible to stratigraphic displacement by burrowing organisms in the benthic mixing zone. Although the mixing zone may be only a few centimeters thick, in slowly accumulating abyssal oozes this "averaging window" may mix the record of thousands of years. Although the mixing may be almost diffusive in scale, the action is persistent and can allow some particles to remain near the surface for geologically significant intervals. The fragility of fossils ultimately limits the effect. In some well-drilling operations, fossils may fall into older levels with cave-ins from the well walls and mix with older sediment at the drill bit. It is common practice to assume that these anomalously early and late occurrences can be recognized and eliminated from the data. The validity of the assumption varies considerably with the size and fragility of the fossils. It must be considered when choosing among the algorithms that purport to sequence stratigraphic events.

Quality control issues for taxon range-end events arise from taxonomic instability (Alroy, 2002). If the same species is recognized by different names, potential seriation constraints and resolving power are lost. If different species are assigned to the same name, the consequences may be far more serious: timelines are incorrectly ordered. The best data sources for taxon range-end events will be associated with rigorously maintained taxonomic dictionaries. Nonetheless, human taxonomists will disagree and the species concepts that emerge from combinations of sources will inevitably tend to be looser than those of individual publications.

### *10.3.2 Paleomagnetic reversals and stable isotope ratios*

Paleomagnetic intervals lack the unique recognition characteristics of fossil species. Two states, normal and reversed, generate an alternating binary signal. Relative to taxon range ends, however, paleomagnetic reversal events have the advantage that they are global in scope and are recorded at the appropriate levels in sections and cores. Concerns of quality control include possible secondary magnetization, for example. Quasi-periodic time series of stable isotope ratios may be subdivided at arbitrary threshold values to distinguish alternating intervals of relatively high and low values. The boundaries between intervals have time-stratigraphic properties similar to paleomagnetic polarity intervals, but their scope is not necessarily global.

### *10.3.3 Sequence boundaries*

Some intervals may be considered repeating rather than unique or alternating. Fluctuating accommodation rates in peritidal environments, for example, lead to repetitions of hiatus-bound sediment packages, each of which shallows upward. The bounding hiatuses or flooding surfaces are multiple successive events of one type; they are correlated by differences in the intervening strata and by patterns in successive differences. Unfortunately, the stratal character may change laterally with depositional environment and accommodation space. All the events discussed so far are changes of state. A fossil species appears or disappears; the geomagnetic field reverses; low isotopic values are succeeded by high values; a set of upward-shallowing strata is succeeded by a flooding surface that resets the deposits to deeper conditions. Matched pairs of such events mark the ends of time intervals, such as a taxon range or a polarity subchron. An unfortunate characteristic of these events is that the true time of the change of state may lie within a hiatus; i.e., there does not need to be sediment deposited at the moment of change; deposition before and after suffices to record the event. Some geologic intervals are short enough that the interval itself is treated as an event. In these cases there is always a physical deposit recording the time of the event, not a hiatus.

### *10.3.4 Marker beds*

Rapidly deposited ash fall or turbidite beds are typically treated as instantaneous events. It may be a very difficult quality control issue to ensure that the same bed is identified everywhere. A marker bed traced correctly from place to place may tie together many local sequences and resolve numerous contradictions in order, but false correlation of a marker bed will force serious flaws into a timeline. Decisions about the correlation of marker beds might be made prior to data entry, or databases might contain sufficient information to allow the correlation to be evaluated later. In the construction of timelines, a third option arises – a sensitivity analysis can be performed which examines the consequences of including or excluding the correlation. Short-lived isotopic excursions may be treated as marker beds. More problematic is the practice of using taxon acme horizons as correlative levels; there is no underlying theoretical support for the assumption that peaks in abundance are synchronous and the validity likely falls as the geographic scope widens.

### *10.3.5 Radioisotopically dated events*

Dated events from two locations constrain timelines in powerful ways and the power increases rapidly with the number of these events in a timeline problem. Their position with respect to one another is revealed without any evidence of superposition and the size of the intervening interval is also evident; i.e., these events can order and space themselves on internal information alone. They help tie together faunal provinces that lack shared taxa and they limit the degrees of freedom to correlate nonunique events that alternate or repeat. Of course, the estimate of age is provided as a mean and standard deviation. If these estimates of age overlap for two events, then their order in the timeline is not constrained.

## 10.4 Time-stratigraphic properties

Radioisotopically dated events possess the powerful property that they reveal their ages relative to one another whether or not they occur in the same stratigraphic section. For any one taxon, the FAD must occur before the LAD and abundance spikes cannot lie beyond these extremes. All other event pairs must be ordered relative to one another by superpositional relationships observed where two or more events are recorded in the same sequence of strata. For taxon range-end events, this line of reasoning is compromised by the fact that the observed range ends almost invariably depart from the horizons deposited at the time of the true FAD and LAD. This property may be termed the *fidelity* of stratigraphic position. Stratigraphic infidelity makes the seriation task seriously hard (Dell *et al.*, 1992). It may be overcome by considering numerous measured sections, but the computation time escalates with the number of taxa and sections.

### *10.4.1 Fidelity of stratigraphic position*

Events like ash falls that can be recognized from the properties of a single stratigraphic horizon record their stratigraphic position faithfully in the sense that there is no uncertainty about their true position in a section. But many more event types are revealed by changes of state between successive stratigraphic samples (Table 10.1). The separation of these samples amounts to an uncertainty interval on the precise position of an event. If uncertainty intervals overlap, then the relative age of two events recorded in the same section may be ambiguous. Imagine, for example, that successive sampled horizons record the change from normal to reversed remanent magnetization and that the down-section sample contains the highest occurrence of a diatom species; both events might be arbitrarily drawn midway between the samples, but the order of the last-appearance event and the reversal is undetermined. If different horizons were sampled for the magnetic polarity and the diatoms, and the events placed midway between their respective samples, the apparent order of events can be false. For the diatom species in the previous example, stratigraphic infidelity has another component: the local last occurrence is more likely an emigration or preservation event and, thus, older than the true evolutionary LAD. The negative evidence of barren samples up-section is not trustworthy. The LAD is not necessarily constrained by the first blank sample and it is permissible to infer the position of the LAD anywhere up-section. In other words, to match the sequence of events from section to section, we have stratigraphic freedom to shift the top of a taxon range up-section from its observed limit. Consequently, the local last appearance of a species may establish that the LAD event is younger than events preserved below it but provides no ordering constraints for events preserved above it. In the terms of Table 10.1, last appearances may prove superposition, but not subposition. For a reworked last occurrence, the true LAD may lie up- or down-section. Similar logic applies to FADs, but with the directions reversed.

### *10.4.2 Range length*

It is a special property of taxon range-end events that they are uniquely paired. The range length is the separation of the FAD and LAD event for the same species. Range lengths can be considered as the uncertainty on the age significance and correlation value of a single find of a taxon. The total length of the range is the maximum discrepancy between an observed range end and the corresponding FAD or LAD. Consider the likelihood that two local range ends will be observed in the reverse of their true global order. It increases with the taxon range lengths, but decreases with the separation between the true ages of the events. Relatively small displacements between a find and its true level can cause two closely spaced events to be preserved in the wrong order. Larger failures of preservation or collection are required to reverse the apparent order of events that differ more widely in age. Unless reworking or caving has occurred, range-end events cannot be preserved in the wrong order unless the true stratigraphic ranges of the two taxa overlap in time;

i.e., contradiction in the order of two range-end events from one locality to another is evidence that the corresponding taxa coexisted at some time.

### *10.4.3 Coexistence*

If there has been no reworking or caving, true taxon ranges can only be equal to or longer than the observed ranges. Thus, if two taxa can be shown to coexist anywhere, this is a powerful partial constraint on the true order of events: both FADs must be older than both LADs in the timeline. For strict proof of coexistence, however, the two taxa should be found together in successive samples. Co-occurrence at a single horizon is a looser criterion for coexistence; it leaves a greater chance that the fauna is artificially condensed (Kowalewski and Bambach, 2003). Longer-ranging taxa tend to overlap with larger numbers of other taxa, increasing both the number of constraints and the number of potential contradictions of event order. Although finds of short-ranged taxa offer greater resolving power, they generate fewer coexistence constraints and are more likely than long-ranged taxa to be entirely missing from a local stratigraphic section.

### *10.4.4 One-to-many correlations*

We have seen how the presence–absence nature of taxon range data undermines the ordering of FADs and LADs. Changes of state that generate alternating (binary) and repeating (multiple) event types have a different problem that results from their nonuniqueness. If a stratigraphic hiatus spans two (or more) sequence boundaries or an odd number of successive paleomagnetic reversals, there still appears to be a single simple change of state. Thus an apparently single event of this type in one section may correlate with many events in another section.

## 10.5 Sequencing tools

In order to determine the information needed for seriation, we may examine a few of the algorithms written for the task. Refer to Tipper (1988) for a more complete catalog of correlation methods or Sadler (2004) for a more detailed discussion of seriation strategies. We discuss the methods here only to the extent necessary to reveal different constraints on the seriation task and how the nature of the information sources limits the available methods.

Shaw's (1964) graphic method of correlation likened geologic correlation to linear regression. If pairs of stratigraphic sections are plotted as the axis of an orthogonal "$x$–$y$" graph, then their shared events have $(x, y)$ coordinates and a line of correlation may be fit to these points. To account for the stratigraphic infidelity of range ends, Shaw established a parsimony principle that he called "economy of fit." It states that the preferred estimate of the true sequence of events would be the one that requires the smallest net adjustment of all the locally observed range ends; i.e., it minimizes the implied stratigraphic infidelity that must be attributed to failures of preservation and collection. Shaw allowed range ends to be weighted differentially. His economy of fit sought the smallest adjustment of the "best established" local range ends. The relative reliability of taxon range ends can be estimated from the number of separate finds within the range (Marshall, 1990; Spencer-Cervato *et al*., 1994; Strauss and Sadler, 1989). In effect, the size distribution of gaps due to failures of preservation and collection within the observed range is used to estimate the size of gaps between the observed and true range ends. Of course, Shaw's method requires that fossil finds are arrayed in stratigraphic sections.

A very different approach minimizes the number of coexistences of pairs of taxa that are implied by a hypothetical sequence of FADs and LADs but not seen in any of the observed range charts. This minimization is used in the Unitary Association (Guex, 1991), and Conjunct (Alroy, 1992) methods. They can be applied to datasets in which stratigraphic superposition is not known or stratigraphic distance cannot be reliably measured; i.e., these methods can be used to sequence isolated faunas or collections from highly deformed beds. Alroy's (1994) appearance event ordination extends the logic to stratigraphic sections. Here, any observation of a first appearance below a last appearance constrains the corresponding FAD and LAD to the same order in the timeline; appeal to stratigraphic infidelity allows the distance between the two observed events to be increased not decreased, assuming there has been no reworking or caving.

Methods that minimize range lengths or implied coexistences assume that reworking and caving are not significant – a risky assumption in some settings. The RASC method (Agterberg and Gradstein, 1999) takes the opposite position – that observed ranges are equally likely to be too long or too short. RASC considers events pair-wise and attempts to build a complete sequence out of the most commonly observed order for all event pairs. Notice that RASC compares the properties of the observed and hypothetical sequences. The solution is likely to resemble the most common local observations.

Graphical correlation, by contrast, seeks extreme events – the earliest first occurrence and the latest last occurrence. This attempts to find the true evolutionary sequence of range-end events but leaves the result very sensitive to reworked occurrences. Also, graphical correlation examines the cost of making the observed data fit a single hypothetical sequence. This is a second cause of differences in the resulting timelines; if many pair-wise contradictions can be eliminated by a single range extension, graphical correlation assesses a small misfit penalty, whereas

RASC sees a numerically large difference. Graphic correlation builds a composite sequence one stratigraphic section at a time. The order in which sections are added has a profound influence on the results and must be counterbalanced by an iterative correction process. CONOP9 (Sadler, 2007; Sadler *et al*., 2003) uses the simulated annealing search heuristic to consider all sections at once. The trial-and-error process is sufficiently straightforward and assumption-free to allow a wide range of event types to be included and to mimic or combine the minimization strategies of all the other methods. Thus, for example, it can combine stratigraphic sections and isolated faunas. Each event type has its own rules for stratigraphic freedom (adjustment of position) and recognizing its age relative to other events. When minimizing range extensions, CONOP9 allows users to measure the range extensions in terms of stratigraphic thickness, or event levels, or events leapfrogged. The first option most resembles Shaw's method and favors the sequences observed in the thickest sections. This is often inappropriate. The other two options favor sequences preserved in the most intensely sampled and taxon-rich sections; they are less sensitive to differences in local accumulation rates and to be preferred as default settings.

Clearly, the choice among sequencing tools must be driven by three considerations: the available data, the necessary assumptions, and the questions to be answered. Whether the available data are limited to isolated samples or include a preponderance of measured sections determines whether range extensions can be used as a measure of the fit between observations and hypothetical timelines. Unless reworking may be assumed to be absent, for example, RASC (or a comparable setting in CONOP9) may be the safest approach. If measured sections have similar accumulation histories, range extensions can be measured in rock thickness; otherwise, it is best to use a scale of event horizons. RASC answers questions like: "based on wells drilled so far, what are we most likely to find in the next well" (Cooper *et al*., 2001). Graphic correlation and its CONOP9 variants answer a significantly different question: "based on the wells drilled so far, what is the most likely basin-wide sequence of events."

## 10.6 Integrating tools and databases

If seriation tools are to be coupled with stratigraphic databases, at least two more considerations arise. The sequencing tools were developed by different academics, using different programming languages and input/output formats. The output of queries to the database must be remapped to the input format of the individual tools, but this should be a relatively simple task. More difficult may be the challenge to serve the database and the tool on the same platform, because most of the tools were developed before the databases. Rewriting the tool to a new language should be avoided; it opens the possibility that the tool developer will be unable to supply upgrades directly into the new language. The first fundamental decision is likely to be whether the end user hosts the tool and imports data to it or whether the database server runs the tool and provides the user with output. For the second option another decision may follow: whether the output is restricted to the minimum set of generic files that will enable the user to implement any and all analytical options provided by the tool, or whether the database server offers the full suite of analyses.

It is most troublesome for the database service to offer the full capacity of those tools that provide runtime animations that illustrate the progress of the search for best-fit solutions (e.g., CONOP9). More than mere progress bars, these animations allow an experienced user to assess the appropriateness of the runtime settings and, if necessary, abort a misguided run. In one experiment, we coupled CONOP9 with the CHRONOS database, without implementing runtime graphics. CHRONOS could generate CONOP input files directly from database queries or users could upload their own input files. The CHRONOS web service ran the sequencing algorithms (in a different FORTRAN dialect) and returned a small set of files from which the end user could reproduce all the post-run analysis. In addition to the inconvenience of maintaining two FORTRAN dialects, one for the server, and one for the desktop, we discovered a problem with user experience. It was so easy to run the tool and so much information was available through the database that a novice user could run a huge dataset without the necessary learning experience of watching the timeline grow piecemeal through small incremental steps while building a database manually. The result was hopelessly inappropriate runtime settings coupled with an inability to diagnose the problems they caused or track down quality control issues. For many sequencing algorithms, the run time grows rapidly with the size of the dataset. This is another consideration in the decision whether or not to host the tool. Users will likely appreciate access to the power of the central servers, but the service will need to develop a measure to apportion time. The task becomes much more manageable if the central service can send the user both a dataset and a kernel of the code needed to run the sequencing tool on the user's desktop.

Not surprisingly, CHRONOS had better outcomes for two web-based tools developed together with the databases that served them. Built on Shaw's (1964) graphic correlation principles, ADP (age/depth plotting) is a Java-based intuitive visual tool that is seamlessly connected to the Neptune database of marine plankton records from deep-sea drilling sites (Bohling, 2005; Lazarus, 1992, 1994). Given its optimal suitability for one-dimensional geologic archives that extend over considerable thickness of sediments like deep-sea cores, this tool has been extensively used to construct age models for Deep Sea Drilling Project (DSDP) and Ocean Drilling Program (ODP) cores (e.g., Spencer-Cervato, 1999). Users can construct age models (a line of correlation) for stratigraphic sections based on age-calibrated biostratigraphic datums, magnetostratigraphy, and geochemical proxies. They can save new models in an Age Model database with attribution or on their desktop.

The construction of the line of correlation is guided by the available datums but is a subjective interpretation, involving weighing individual datums differently, based on known reliability. The subjectivity can be eliminated by computing certain criteria and automatically constructing a best-fit line (best practices from experienced users, for example, prefer a paleomagnetic reversal over a local last occurrence that is represented by a single specimen and requiring that every sequence of samples must have increasing age with depth). The quality of the automated outputs varies widely, however, and an expert is needed to manually adjust the output and provide information on overall quality. Individual databases like Neptune require a consistent approach to the age vs. depth correlation to ensure quality outputs and a unique age/depth relationship for each sample.

Practical issues arise when experts disagree. This can be overcome, dictatorially, by having a single expert or restricted group of experts construct age models that become the vetted versions for a particular instance of the database. A modified age model would replace the existing one. However, this precludes other expert users from providing their own interpretation to the age model. This was addressed by creating an Age Model database that stores multiple versions of age models for a single hole together with minimal metadata to indicate its pedigree (author, date, time, a quality attribution, and comments in addition to the description of the hole and its coordinates). Implementation of XML can then change the age interpretations of samples from this site and output data that reflect the version of the age model preferred by this user. Alternatively, users can rely on the interpretation of a defined expert or on the qualitative attributions, selecting the best available interpretation.

Another web-based graphical visualization tool, Age Range Chart (ARC), uses the age models stored in the Neptune database to allow users to plot individual occurrences of one or more taxa and their abundance in multiple drill holes, showing sampling density and hiati present in the age model. Sites can be arranged geographically by latitude, ocean basin, or longitude, allowing the user to visualize time-transgressive occurrences and extinctions and the ranges of local first or last occurrences for each taxon within a single drill hole or a region. Because such a tool is more illustrative than analytical, it gives rise to far fewer issues than those that include correlation and sequencing algorithms.

## 10.7 "Steno Core"

In some ways, databases resemble human scientists as stores of information. In order to work well together in multidisciplinary projects, both need some shared language for common concepts in addition to their larger specialized vocabularies. In this section we consider what shared vocabulary might be needed to enable timelines to be populated with information harvested from a variety of specialized databases, each with its own data structure and quality controls.

Consider, first, a challenge that has been more widely explored: compiling biogeographic maps from a range of zoological and botanical databases, museum catalogs, etc. This task has been considerably eased by encouraging databases to share a common minimum storage format for geographic position and taxonomic names. Several standard protocols exist for sharing location information: ISO 19115 and the Federal Geographic Data Committee (FGDC) standards for geospatial data (www.fgdc.gov/metadata) and Dublin Core (www.dublincore.org) for resource description. They are more or less cumbersome and detailed, in some cases including hundreds of elements. The shared vocabulary for storing biological names is a better role model for our needs. "Darwin Core" (Wieczorek, 2007) specifies a small set of data fields designed to be incorporated in all biological data structures. The Darwin Core protocol specifies simple, stable, minimum data requirements that simplify access to common kinds of records of the occurrence of organisms across numerous databases with different specialized purposes and data structures.

We suggest that stratigraphers develop a similar protocol to support stratigraphic seriation. It might be named "Steno Core," in honor of the superposition rule attributed to Niels Steensen under the familiar Latin name Steno. Borrowing from Darwin Core, it would meet the following principles:

- Facilitate the assembly of information from many sources into new structures
- Demand a minimum additional effort to support analysis of the data
- Avoid restricting data content and quality at the source
- Accommodate data that are common to more than one discipline, and found frequently enough to be broadly useful
- Achieve stability by including in the core only those concepts that are already broadly agreed upon and relatively likely to remain unchanged over time

As with Darwin Core, there would be no pretense that Steno Core is adequate to support a primary database. The goal is to be simple enough to minimize disincentives to primary data suppliers and avoid barriers to data import into a variety of user platforms. Assuming that geographic locations and names of events follow conventions established elsewhere, the basic Steno Core requirements would be of only three kinds:

- *Relative Position*: Each usable event must be linked unambiguously, by relative stratigraphic position, to *at least* one other event, either directly or via a common datum. With enough such links, it is conceivable that all linked pairs could be assembled into a single ordered chain. The linkages need to function within and between databases; that is, the superpositional order may be specified relative to another event in the same database or a different database. For the most diversely populated timelines, the cross-database links should be maximized. Relative age may be established on a numerical age scale, a stratigraphic thickness scale, or within an ordered series of samples.

- *Temporal Scope*: There should be some means to assess the temporal scope, or resolving power of the information. For paleontological faunal lists, this might be the thickness of the sampled interval, or a statement that it was a single bedding surface. For paleomagnetic reversals and first- or last-appearance datums, the sample spacing matters.
- *Correlation Uncertainties*: For taxon ranges, it is widely recognized that the name of the taxonomist must be provided. For marker beds, some level of assurance might be provided that the same bed has indeed been located in different sections. The bed might have been physically mapped out between sections or there may be reference to supporting analyses of composition.

To see how Steno Core might function, consider the three most common sources of stratigraphic information: sediment cores drilled down from the surface of the crust, stratigraphic sections measured at outcrop, and isolated samples. Cores typically present the best ordering information and the fewest ambiguities; outcrop sections present more ambiguities; isolated samples provide the least ordering information.

Drill cores, such as those extracted by petroleum companies or the Ocean Drilling Program, provide the most readily referenced information. Cores have unique well identifiers and core repositories ensure that all investigators sample the same physical record. The obvious datum for relative position is depth below the seabed, or the land surface, or the driller's Kelly bar. All events from a single core can be referenced to the same datum. Cores that preserve the same taxa or marker horizons can readily be correlated.

Stratigraphic sections that are measured at outcrops allow access to much more rock than a drill core. They can provide more richly sampled taxon ranges for macrofossils and better assessment of hiati, faults, and discontinuous bed forms. For the same reason, however, it is much more difficult to replicate the position of previously reported horizons on successive visits to outcrop sections. Anyone who has measured the same section twice understands the difficulty of replicating measurements of stratigraphic distance in irregular surface outcrops. Datum levels are not always easy to establish and reoccupy. Where different authors report independent measurements of the "same" section, it is prudent practice to treat repeat samplings as separate sections and to allow the seriation tools to correlate them. As many distinctive levels as possible may be used as marker horizons in the construction of timelines. Any sufficiently distinctive stratigraphic level may serve as a reference for repeat measurements – generally it is better to insert new information into a previously measured section by distance from the nearest distinctive bed, rather than a datum at the top or base of the section. An author who claims to have reoccupied the same measurement scale should provide some assurances, perhaps by checking all favorable conditions that apply; e.g., "same measurer," "accompanied by original measurer," "found original painted scale," "recalibrated at unmistakable horizons," etc.

Isolated samples provide only one type of information – co-occurrence. Depending upon the scope of the sample, the evidence of co-occurrence may be more or less reliable. Observations from the same thin bed or bedding plane are best. Samples of thick intervals or whole formations may generate lists of taxa that did not actually coexist. For this reason, some seriation tools include the option of requiring overlapping ranges as evidence of co-occurrence, not merely appearance in the same sample.

## 10.8 Concluding thoughts about managing timeline data

The potential resolving power of a timeline increases with the number of events that it includes. The likely robustness of a timeline depends upon the volume, quality, and variety of supporting field and laboratory observations. As the number of events and volume of data increase, however, the compilation, quality control, and sequencing tasks become more daunting and the likelihood decreases that the total dataset can be published with the timeline. We close by addressing two issues: how best to ensure data quality and how to store the derivative products – the timelines themselves.

The information necessary for geologic timelines might be managed centrally in one large database and realize some advantages of a single structure. Alternatively, a federation of specialized databases is formed in which one umbrella component handles information integration and provides users the appearance of searching a single structure. Quality control issues, which differ significantly with subdiscipline, would seem to tip the scale heavily towards a federal model. The CHRONOS experiment (www.chronos.org) demonstrated the feasibility of supporting a virtual stratigraphic record by federating databases and tools for timelines. The look and feel of a single database can be achieved by managing the web-based interoperability between many individual database formats and the input requirements of various analytical tools. Each database remains matched to its own data-donor community and provides its own expert quality control (Fils *et al.*, 2008; Greer *et al.*, 2005). Subdisciplinary communities retain their own representation of concepts and the data behind them, individual best practices, and platform preferences. In order to describe the relationship of data elements in different databases, as well as overarching concepts like timescale, CHRONOS uses the Web Ontology Language (OWL), Simple Knowledge Organization System (SKOS), and Resource Description

Framework (RDF). It is critical that the complete ontology captures the full scientific underpinnings of the databases – that is, the science for which the databases are designed. An evident weakness of the federal model is that in fiscal crises the umbrella component may appear to be a luxury.

The derived timelines are valuable information products that need to be retrievable together with the data from which they were generated. One option stores the coupled data-and-timeline objects with metadata that include the sequencing tool settings. Another stores only a set of instructions sufficient to reproduce the data query and the seriation exercise, assuming that the timeline can be rebuilt when needed. The former requires storage space, the latter computation time and careful versioning of the growing database so that former contents can be reproduced. Because heavily populated timelines based on large data volumes may require many days to optimize, the former option would seem to be the best choice. Each data-timeline object would need a unique identifier for inclusion in publications and to permit rapid retrieval. The long computation times, likely to grow as the stratigraphic information accumulate, lead us to wonder whether some proven sequencing tools and timeline projects should be running continually in a fashion that automatically incorporates new information as the database grows. If feasible, this unites the timeline and timescale database concepts we introduced at the outset. The lay inquirer could see the current results of automated timescale calibration at any time. Clearly, this goal would face several hurdles. A more practical first step might use automated sequencing tools to examine the nature of new data as they are added. Very soon after each entry of a new core or fauna such tools should be able to report whether the additional information corroborates current solutions or generates severe contradictions. These two conditions do not correspond in any simple fashion with good and bad data but might be useful flags to inform the quality control process.

## Acknowledgement

Partial support for the CHRONOS project was provided by National Science Foundation awards EAR-0315216 and EAR-0524285.

## References

Agterberg, F. P. and Gradstein, F. M. (1999). The RASC method for ranking and scaling of biostratigraphic events. *Earth-Science Reviews*, **46**: 1–25.

Alroy, J. (1992). Conjunction among taxonomic distributions and the Miocene mammalian biochronology of the Great Plains. *Paleobiology*, **18**: 326–343.

Alroy, J. (1994). Appearance event ordination: A new biochronological method. *Paleobiology*, **20**: 191–217.

Alroy, J. (2002). How many species names are valid? *Proceedings of the National Academy of Sciences*, **99**(6): 3706–3711.
Bohling, G. (2005). CHRONOS Age Depth plot: A Java application for stratigraphic data analysis. *Geosphere*, **1**: 78–84.
Cooper, R. A., Crampton, J. S., Raine, J. I. *et al*. (2001). Quantitative biostratigraphy of the Taranaki Basin, New Zealand: A deterministic and probabilistic approach. *American Association of Petroleum Geologists Bulletin*, **85**: 1569–1498.
Dell, R. F., Kemple, W. G., and Tovey, C. A. (1992). Heuristically solving the stratigraphic correlation problem. *Proceedings of the First Industrial Engineering Research Conference*, **1**: 293–297.
Fils, D., Cervato, C., and Reed, J. (2008). CHRONOS architecture: Experiences with an open source services-oriented architecture for geoinformatics. *Computers & Geosciences*, **35**(4): 774–782.
Gradstein, F. M., Ogg, J. G., and Smith, A. G. (2004). *A Geologic Time Scale 2004*. Cambridge: Cambridge University Press, 589pp.
Greer, D. S., Bohling, G., Diver, P. *et al*. (2005). A data infrastructure to support Earth history research. *Geological Society of America Annual Meeting*, **37**: 416.
Guex, J. (1991). *Biochronological Correlations*. Berlin: Springer Verlag, 252pp.
Holmes, A. (1947). The construction of a geological time-scale. *Transactions of the Geological Society of Glasgow*, **21**: 117–152.
Holmes, A. (1960). A revised geological time-scale. *Transactions of the Edinburgh Geological Society*, **17**: 183–216.
Kowalewski, M. and Bambach, R. K. (2003). The limits of paleontological resolution. In *High-Resolution Approaches in Stratigraphic Paleontology*, ed. P. J. Harries. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 1–48.
Lazarus, D. B. (1992). Age Depth Plot and Age Maker: Age depth modeling on the Macintosh series of computers. *Geobyte*, **71**: 7–13.
Lazarus, D. B. (1994). The Neptune Project: Developing a large relational database of marine microfossil data on a personal computer. *Mathematical Geology*, **26**: 817–832.
Marshall, C. R. (1990). Confidence intervals on stratigraphic ranges. *Paleobiology*, **16**: 1–10.
Sadler, P. M. (2004). Quantitative biostratigraphy: Achieving finer resolution in global correlation. *Annual Reviews of Earth and Planetary Science*, **32**: 187–213.
Sadler, P. M. (2007). *CONOP9 version DEC 7.451*. Riverside, CA: University of California.
Sadler, P. M., Kemple, W. G., and Kooser, M. A. (2003). Contents of the compact disc: CONOP9 programs for solving the stratigraphic correlation and seriation problems as constrained optimization. In *High-Resolution Approaches in Stratigraphic Paleontology*, ed. P. J. Harries. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 461–465.
Shaw, A. (1964). *Time in Stratigraphy*. New York: McGraw-Hill Book Co., 365pp.
Spencer-Cervato, C. (1999). The Cenozoic deep sea microfossil record: Explorations of the DSDP/ODP sample set using the Neptune database. *Palaeontologia Electronica*, **2**(2), art. 4, 268pp., 2.4 MB. http://palaeoelectronica.org/1999_2/neptune/issue2_99.htm.
Spencer-Cervato, C., Thierstein, H. R., Lazarus, D. B., and Beckmann, J. P. (1994). How synchronous are Neogene marine plankton events? *Paleoceanography*, **9**: 739–763.
Strauss, D. J. and Sadler, P. M. (1989). Classical confidence intervals and Bayesian posterior probabilities for the ends of local taxon ranges. *Mathematical Geology*, **21**: 411–427.
Teichert, C. (1956). How many fossil species? *Journal of Paleontology*, **30**: 967–969.

Tipper, J. C. (1988). Techniques for quantitative stratigraphic correlations: A review and annotated bibliography. *Geological Magazine*, **125**: 475–494.
Valentine, J. W. (1970). How many marine invertebrate fossil species? *A New Approximation. Journal of Paleontology*, **44**: 410–415.
Wieczorek, J. (2007). The design and purpose of Darwin core, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DesignAndPurpose.

# 11

# Modeling geodynamic processes with ontologies

HASSAN A. BABAIE

## 11.1 Introduction

Ontologies define the hierarchical structure and types of object and process entities and their properties, and reflect the nature of the spatial objects and regions involved in processes, relations among processes and spatial entities, and the temporal characteristics of the processes (e.g., Noy, 2004; Smith, 2003). Earth scientists study naturally, experimentally, or simulationally induced processes, and develop conceptual models to help them understand and simulate these processes in the laboratory. The metadata, i.e., information required to understand data, inherent in ontologies, can help the integration, reuse, and interoperability of these models, and enhancement of their functionality.

Despite the fact that object, state, process, and event constitute the main ingredients of an ontology (Galton and Worboys, 2005), most ontologies in the earth sciences only focus on the static part of reality, i.e., on objects (e.g., fault, subduction zone) and their properties and relations, leaving the processes (e.g., faulting, subduction), which constitute the dynamic part of reality, out of the picture. In other words, these ontologies ignore change through time and the processes that materialize these changes, despite the fact that earth scientists continuously collect data about individual spatial objects and processes in their research. With the advent of the Web and sophisticated digital data acquisition equipments, which produce an immense volume of data in short periods of time, and cover spatial regions of variable scale, there is an emerging and urgent need for data and information storage and interchange through ontology-based knowledge bases.

Although existing top-level earth science ontologies such as SWEET (Raskin and Pan, 2005) provide sets of terms (i.e., classes) related to processes, they do not define an exhaustive set of properties for these processes or relate them to spatial objects. The main reason for avoiding the formal inclusion of processes in ontologies is the inherent difficulty in depicting the Earth's dynamics compared to its static part. However, there has been some significant recent progress in the depiction and formalization of spatiotemporal reality in ontologies. These formalizations, axiomatized with first-order logic, are applied mainly in biomedical sciences (Smith and Rosse, 2004; Smith *et al.*, 2005) and, to a lesser extent, geosciences (e.g., Feng *et al.*, 2004). In this chapter, I review the application of the so-called SNAP and SPAN meta-model (Grenon, 2003a, b, c; Grenon and Smith, 2004) for developing ontologies in earth sciences, and provide examples from a set of entities in the Nankai Trough in southwestern Japan as a case study to model the processes and objects in this subduction zone.

This chapter focuses on the design of dynamic ontologies, without considering the issues of implementation in ontology languages such as OWL, their deployment, or production of the web interface for their presentation and query. I will describe the two meta-models that define orthogonal perspectives of reality; one views the static aspects of spatial objects and their qualities (SNAP), and the other views the processes that act to bring change in these objects (SPAN). The main objectives of this paper are: (1) to introduce the ontological notions of process, event, and state, and their relation to state change in geological objects, (2) to show to the geoscientists how ontology can be used to depict their knowledge about both the static and dynamic aspects of the earth systems, and (3) to introduce the main ontological relations that associate the spatial entities to processes. In this chapter, class names are given with the first letter in the upper case (e.g., Mylonite, Fault), and relations (properties) are written in the lower case (e.g., realizes, involves). Compound class names are separated with an underscore (e.g., Subduction_Zone; Accretionary_Prism).

## 11.2 Ontological views of the world

Each subsystem of the earth system (e.g., lithosphere, hydrosphere) includes material, concrete things, or objects that exist, at specific times independent and outside of our minds, as substantial individuals (e.g., the San Andreas Fault in California or Nankai accretionary prism in southwestern Japan). Each object (individual), including the Earth itself, *denoted* by a special sign or symbol (e.g., a term in natural languages such as "fault"), is *referenced by* a concept that *instantiates* a universal *type* (e.g., Fault) (Dietz, 2006). While each object exists in space and time (e.g., Idaho batholith; Nankai Trough), the *universal* is a construct (e.g., batholith, trough) that exists only in our mind (Bunge, 1977). For example, "Kumano Forearc Basin" (a symbol), which *refers* to a specific, real basin (an object) in the Nankai Trough, *instantiates* the abstract concept of "Forearc_Basin" (a universal type). The real, *individual* Kumano Forearc Basin is one of many objects in the Forearc_Basin class (concept), which *conform* to this universal type by possessing the set of properties that characterize the Forearc_Basin type. The type supplies global identity conditions to the individuals (Guarino and Welty, 2002).

Every part of the earth system, as a substantial (material) object, is endowed with a series of properties. Tectonic movements lead to lawful (allowed) physical, chemical, and biological processes that occur in the so-called lawful *event space*, and transform the state of things (Bunge, 1977). Through systematic analyses, scientists have discovered that properties in a given type are related to each other through physical (natural) *laws* that are defined for the universal type, and that are themselves complex properties. For example, the linear law of elasticity, that relates stress and strain, applies to rocks deforming at shallow depths, under a restricted range of mechanical and environmental conditions (e.g., low strain, stress, and temperature). At any instant of time, the state of any object (e.g., rock) that participates in a process (e.g., deformation), defined by the collection of values for its properties, is in constant flux. The change of state, a subject of study by earth scientists, also occurs in a lawful *state space*, which is constrained by the properties and the physical laws (Bunge, 1977). For example, the state space for rock deformation is within the range of pressure and temperature in which rocks remain solid. Science progresses by discovering new properties and laws that restrict the relationship among properties in a given type, and the processes that change the state of objects that conform to the type (Bunge, 1977). The individuals constitute the subject matter for ontologies (Smith, 2004).

Entities that occur on the surface of the Earth, below its surface, and in the atmosphere around it, can be grouped, based on their mode of existence and persistence, into two general, non-overlapping (i.e., disjoint) categories: (1) enduring entities, (2) perduring entities (e.g., Bittner *et al.*, 2004; Heller and Herre, 2003; Lowe, 2005). Enduring entities, known as endurants or continuants, include substantial objects (both material and immaterial) such as fault, water, gas, subduction zone, mylonite, and cave (an immaterial entity), that occur in the lithosphere, hydrosphere, biosphere, atmosphere, and cryosphere components of the earth system. These entities obtain different properties at different times, e.g., an accretionary prism growing in size through offscraping and underplating. The earth systems science views the entities in these components as interacting through different sorts of processes such as weathering, hydration, compression, flow, and erosion. The perduring entities, known as perdurants, include the processes and events that involve the endurants that participate in these processes. The perdurants include such things as displacement (process), deformation (process aggregate), and the instantaneous parts (i.e., the beginning and ending events).

While the endurants are part of reality that occupy spatial regions at different or similar times, i.e., have spatial parts, processes occupy spatiotemporal regions, and have temporal parts (i.e., phases, stages) that unfold over a time interval (e.g., Hawley, 2001; Koslicki, 2008). For example, a river (an endurant whole) has spatial parts such as meander, cutoff bank, ox-bow lake, and delta that occupy different spatial regions at a given time. On the other hand, a process such as slip along a plate boundary fault in a subduction zone may have the following temporal parts: an earlier strain hardening phase that may be followed in time by discrete intervals of frictional sliding, cataclasis, or crystal–plastic deformation at different depths (under different pressures and temperatures), followed possibly by seismic rupture. Each of these simple or complex (i.e., aggregate) earth processes has instantaneous beginning and ending temporal parts, which are referred to as events. For example, the first instant of an earthquake when the rupture occurs is an event.

Geologic reality, like other realities, can be viewed from different perspectives (e.g., map, cross section) and granularities (e.g., microscopic, lithospheric, global). Philosophers have depicted the world, and the persistence of objects in it, mainly through two competing accounts of reality: three-dimensionalism or endurantism (e.g., Stell and West, 2004) and four-dimensionalism or perdurantism (e.g., Hawley, 2001; Sider, 2001). While some philosophers believe that the 3-D and 4-D views are metaphysically equivalent (Miller, 2005), others believe that effective ontologies must have both views (e.g., Simons, 1987) and that the four-dimensionalist view provides a truthful perspective on reality (Smith and Grenon, 2004).

The 3-D and 4-D paradigms have direct implications to the work of earth scientists in their efforts to depict geologic realities. The following are the important distinctions between the two paradigms (Bittner *et al.*, 2004; Koslicki, 2008; Sider, 2001; Stell and West, 2004): (i) individuals (e.g., specific processes) in the 4-D paradigm exist in four dimensions, in the past, present, and future. These entities can have one, two, or three spatial dimensions in addition to a time dimension. On the other hand, the 3-D ontology assumes that individuals are three-dimensional and are wholly present at every moment in time. (ii) In the 4-D ontology, the world is viewed from outside of time in contrast with the 3-D ontology that views objects from the present. (iii) In the 4-D ontology, individuals extend in space and time and have spatial and temporal parts, whereas those in the 3-D ontology do not have any temporal part. (iv) Individuals (particulars) occupying the same spatiotemporal extent are assumed to be the same in the 4-D paradigm, in contrast with the 3-D ontology which allows different physical objects to coincide (e.g., water or oil in a pore space). (v) The 4-D paradigm considers the whole life of an object to be of primary interest, vs. the 3-D paradigm, which considers the objects at instants of time to be of primary interest.

## 11.3 Spatiotemporal reality

Effective ontologies must capture persisting endurant (continuant) objects through time and space, and account for processes that lead to changes in these individuals. The continuant earth objects such as rock, water, accretionary prism, and fault, endure by preserving their (same but not necessarily exact) identity through time despite their continuous qualitative changes, for example, in their length, density, shape, and temperature. These objects occur, as complete wholes, in spatial regions (i.e., portions of space). This means that, at each instant of time, an object occurs in its entirety, i.e., it persists as a self-connected mereological whole (Casati and Varzi, 1999; Lowe, 2002; Pribbenow, 2002; Simons, 1987). For example, the San Andreas Fault zone in California is undergoing a continuum of changes that includes brittle and ductile deformation at different spatial regions (within the spatial region occupied by the fault zone) through which the fault zone changes shape (length and width), spatial location (fault blocks displacing by translation and/or rotation), and spatial parts (e.g., formation of new bends, steps, sag ponds, sets of fracture, mylonite zones). At each instant of time when earth scientists make observations about the fault, the structure's spatial parts are wholly present, although some have changed, or are new. Because scientists' perspective toward endurant entities is from the present moment of time (i.e., now), the zero- to three-dimensional objects are measured in space ($x$, $y$, $z$), with a time index ($t$) (e.g., a satellite image of the trace of the San Andreas Fault taken today at 10:30 am).

The perdurant (occurrent) earth objects such as processes, events, on the other hand, occur in intervals of time ($t_{i+n}$-$t_i$) through a succession of temporal parts (phases, stages) (e.g., Needham, 2003). An example is a video of a hurricane or tsunami, taken over a period of time by a satellite, showing different phases of the same hurricane or tsunami. The processes (e.g., fluid flow, deformation) unfold themselves over time and, at each time slice ($t$) they present an incomplete part of the whole, meaning that they are mereologically incomplete (Casati and Varzi, 1999; Simons, 1987). We can simplify the depiction of space-time with a 2-D graph of time ($t$) vs. space ($s$), on which spatiotemporal objects occupy different ($t$, $s$) coordinates over their lifetime (Bunge 1977; Lowe, 2002;). For example, the Hurricane Katrina, a 4-D perdurants individual that started in summer 2005, occupied different spatial regions at different times. The hurricane (as a process) had different temporal parts, characterized by different spatial qualities such as category, intensity, velocity, size, and location, at different times and places. A rupture that forms an earthquake process that starts in a spatial region along a brittle fault (e.g., the seismogenic zone along the plate-boundary fault in the Nankai Trough) propagates in space-time, and occupies different spatial regions at different times. This spatiotemporal earthquake process entity, like the hurricane, assumes different temporal parts during successive phases of its life. It propagates from the hypocentral region across concentric spheroidal (isotropic media) or ellipsoidal (anisotropic media) time surfaces in space, and manifests itself by different properties, which are represented by attributes such as the propagation direction, slip surface area, and velocity, at different times and in different spatial regions. For example, a minute after rupture along a fault, P-waves lead to displacement of rock particles in the direction of the wave propagation (compression process; a temporal part), and at a later time (phase), particles in the same spatial region undergo (i.e., participate in a) displacement transverse to the wave propagation (shearing) when the S-waves pass through the same spatial region. Scientists capture different temporal parts of the same seismic process (that started with the same instantaneous rupture event) at different times and spaces (i.e., seismic stations) on seismograms. The same earthquake may cause a tsunami, which may inflict damage in one or more spatial regions, involving continuants at different or same times. The earthquake or the tsunami is said to be a spatiotemporal entity, meaning that its life is made of a succession of temporal parts (Miller, 2005).

The possibility of the presence of more than one correct and complementary perspective is known as perspectivalism in philosophy (e.g., Grenon *et al.*, 2004), which is constrained by realism, meaning that the reality is independent of our perspective (Smith and Grenon, 2004). Earth scientists habitually apply a perspectivalist view toward geological objects by applying a map, cross section, or other views to the same reality. Because the two veridical (formal) 3-D and 4-D perspectives described above are mutually disjoint, objects belonging to one cannot belong to the other (e.g., Miller, 2005). For example, a fold or a fault, as endurant entities, cannot belong to the folding or faulting process, respectively, because the fold and fault entities have spatial parts but no temporal parts, whereas folding and faulting have temporal parts but no spatial part. Thus, we must develop two non-overlapping ontologies to take care of our earth objects and processes. Earth science ontologies can only capture the reality if they consider both of these orthogonal perspectives (i.e., endurant and perdurant) on the persistence of real objects. Obviously, geological processes (e.g., faulting, melting, crystallization) “involve” geologic objects (fault, rock, magma). This is how the two perspectives intersect, and how we relate the two ontologies. The two endurant and perdurant perspectives are represented by the SNAP and SPAN meta-models, respectively, which are described below.

## 11.4 The SNAP ontology

The Basic Formal Ontology (BFO) (Grenon *et al.*, 2004) is said to be a formal, upper-level ontology because it is developed at a highest and most domain-neutral level of generality (Grenon, 2003a, b, c), like the SWEET ontologies in earth sciences (Raskin and Pan, 2005), or DOLCE (Gangemi *et al.*, 2002; www.loa-cnr.it/DOLCE.html). This means that it applies equally well to these sciences as it does to other domains such as biology or business. The bi-categorical BFO ontologies of particulars (individuals) include the two endurant and perdurant perspectives described above, which hereon are referred to as SNAP and SPAN ontologies, respectively (Grenon, 2003a, b, c; Grenon and Smith, 2004). These SNAP and SPAN components of the BFO characterize the static and dynamic views of the world, respectively. The SNAP ontologies represent a snapshot (i.e., 3-D) perspective on reality, whereas SPAN ontologies characterize a 4-D view (over a span of time) and hence a temporal aspect of existence.

Endurants, which are constituents of the SNAP ontology, are captured at times of measurements and observations. The SNAP ontology is therefore an inventory of time-indexed measurements of individuals, such as a specific fault zone or river observed at some specific instants of time while moving along the linear time continuum. Philosophically, the SNAP ontology is a 'presentist' view of the world, where individual objects are measured at moving 'present' instants of time (Lowe, 2002). The observed endurant object of course may have existed for a long time before the present moment (i.e., now), when it is captured, and may exist after now, and be captured in the future time slices. Change in a SNAP ontology may be measured by comparing the discrepancies (from snapshots) among the qualities (e.g., shape, location) of the continuants measured at different time indexes.

SNAP entities have spatial parts, boundaries, and aggregates which are themselves SNAP entities (Grenon, 2003a, b, c; Grenon and Smith, 2004). For example, discontinuities (e.g., bedding, foliation) that define the limbs of a fold, and the fold itself, are both SNAP entities. The 2-D (i.e., planar) boundaries of the fold limb are SNAP entities. The 1-D fold axis along the intersection of the two limbs is a SNAP entity. However, no SNAP entity can be a part of a SPAN entity. For example, a subducting plate is not a part of the subduction process. We say that a subducting plate is "involved_in" or "participates_in" the subduction process, and becomes subducted when the subduction process occurs over a finite time interval. The plate, which perhaps was unmetamorphosed before subduction, changes its state by modifying, for example, its composition (a "quality" or property, and a SNAP dependent entity, see below).

The SNAP entities include the following types (Grenon, 2003a, b, c; Grenon and Smith, 2004): (1) Independent SNAP entities, which include substantial entities (e.g., mineral, river, mylonite); aggregates of substances, such as a sequence of rock layers at all scales (e.g., lamination, bed, member, formation); boundaries (e.g., unconformity, intrusive contact, fault, grain boundary); fiat (arbitrary) parts (e.g., the seismogenic fault zone, the equator, 30 °N latitude, border between Nevada and California); and sites, which include empty spaces (holes) such as cavity, pore, cave, and conduit. (2) Dependent SNAP entities, which include SNAP entities' specific qualities or "tropes" (e.g., density, color, temperature, composition); functions (e.g., of a river to transport sediment load and of solutions to carry ions); conditions (e.g., of a rock; altered vs. unaltered, of a volcano: dormant vs. active); shape (e.g., sigmoidal shape of a megacryst in a mylonite); and role (e.g., of water in hydraulic fracturing or hydrolytic weakening). Dependent SNAP entities can be monadic belonging to an object (e.g., mineral's specific gravity, hardness, planarity of a discontinuity), or relational, between two objects (e.g., sutured contact, unconformable contact, parallel lamination).

## 11.5 The SPAN ontology

Since SPAN entities have temporal parts, time is a main part of a SPAN ontology. These entities are of three types (Grenon, 2003a, b, c; Grenon and Smith, 2004): (1) Processuals entities, which include processes (e.g., alteration of rocks near a river, mylonitization in a shear zone, accretion in a subduction zone), and their instantaneous temporal boundaries (i.e., events such as the instant a meteorite hits the Earth, the last instant of an interval of the flow of a gusty wind). Processuals also include temporal aggregates, i.e., a successive series of processes that lead into one another (e.g., the arrival of series of body and surface waves to a location at different times, and the consequent compressional or shear particle motions). (2) Temporal regions that are part of time and could be scattered or connected (both intervals and instants). (3) Spatiotemporal regions which are part of space-time and which can also be scattered or connected.

SPAN entities can also be of monadic dependent type (e.g., dilation of rock above a magma chamber), or relational dependent type (e.g., water fracturing rock, strain shortening a layer of rock). Some SNAP entities owe their existence to the SPAN entities, e.g., a deformation band that could not exist without the occurrence of a deformation process, or a slip surface or a cataclastic band that depend for their existence on slip and cataclasis processes, respectively. Processes have fiat temporal parts (as opposed to bona fide temporal parts (Smith, 2001; Smith and Varzi, 2000) arbitrarily partitioned along the time dimension, e.g., the first half of a folding process, or the last phase of crystallization out of a magma. The 4-D counterparts of the 3-D SNAP "site" entities are called "setting" in the SPAN ontology; these are space-time regions in which processes occur, e.g., brecciation in the upper 2 km of a fault zone during the past 10 my. Thus, change and dynamic reality can better be captured through SPAN entities (e.g., events, processes), which unfold themselves over an interval of time (Smith and Grenon, 2004). The following section focuses on SPAN entities such as events and processes.

## 11.6 Events, processes, and states

In a spatiotemporal framework, objects, and sets of events and processes that involve these objects, populate the space and time. Time is perceived by us through the occurrence of these events and processes that bring change within the period between these events. In this framework, time is an ordered set of instants or interval entities. There are infinitely many points along the time dimension, assuming a linear notion of time, analogous to the unending set of the real numbers (e.g., Galton, 2000). Processes "occur" within time intervals between two instants $t_i$ and $t_{i+n}$ ($t_i$ is earlier than $t_{i+n}$) that are defined by the "happening" of event $e_i$ that "precedes" event $e_{i+n}$ (Allen, 1983, 1984). These ordered instants of time and the events define the beginning and end of processes that occur in the time interval. For example, one can argue that the first foreshock of a seismic event ($e_1$) defines the beginning of a seismic slip (a process) along a fault, at a specific instant ($t_1$), and the last aftershock, which occurs at a later instant ($t_2$), defines the ending "event" ($e_2$) of the seismic slip (a process). Notice that in contrast to events that occur at discrete time instants, processes occur over a series of instants (i.e., within a time interval).

There are some questions regarding the relationship between processes and their spatial location. Processes do not occupy space because of their mereological incompleteness at instants of time. For any object $x$ to move (by translation or rotation), at least some parts of $x$ should be wholly present (located) at different times and different places (cf. Casati and Varzi, 1999). For a process to move, its temporal parts must be located at different places at different times. For example, the motion of a tsunami or a hurricane (as a process) involves motion of its temporal parts from one spatiotemporal region to another. A process may move if a part of it, which occurs in one spatiotemporal region (e.g., melting of rock in one region, sliding along a fault), moves to another spatiotemporal region (melting in another region, sliding in another spot along the fault). Another interesting thing about a process is that *same* individual process (i.e., occurrence), such as the 1980 eruption of Mount St. Helens, cannot occur at the *same* place at different times, unless its temporal span is greater than the interval between the times (Mayo, 1961). Notice, however, that the same universal type (extended by a class) of a process (e.g., eruption) can occur in the same region at different time intervals (i.e., different occurrences). For example, the universal process of seismic slip (but not the specific individual slip) can occur along the same segment of a fault at different times.

Events can be perceived to define temporal discontinuities or transitions, triggering change in the "state" of the continuants (endurants) that participate in the process. Spatial change in endurants can occur in their dimension (e.g., linear objects becoming broad and planar), connectivity (making or breaking connections), location (translation, rotation), orientation (through rotation), size (through strain), shape (through strain), and other qualities such as density and composition (Galton, 2000).

The series of change as the world goes through time is called history; events are boundaries between states in the history of the world (e.g., Bunge, 1977). State is characterized by the presence of a constant situation (temporal homogeneity) and the absence of change; it is the way the world is at a particular time interval. This does not mean that there is no change at all during a state; perception of change is scale and perspective dependent. It just means that the processes that are happening (at some scales) are not enough to trigger a change. For example, the "seismic gap" segments of an active fault zone (e.g., San Andreas), although not sliding, are undergoing microscopic ductile deformation such as creep. Only after the resolved shear stress along a fault plane overcomes the friction between the two fault blocks, an event can initiate a frictional sliding (slip) process along the fault, which then stops (sticks) (at another instantaneous event) when friction exceeds the resolved shear stress. The world is continuously changing its state, i.e., it is changing the structure that defines the way its constituents are put together in space, at some instant (i.e., time slice). An event, such as the onset of down-slope movement of a landslide, disrupts the temporal homogeneity (i.e., state) that exists before the slide, and triggers a new non-steady state in the rocks, which gives way through another event to a later state when the landslide comes to rest. There are of course a lot of small-scale processes in the rocks that occur before the landslide in the original state of the rock. So the rock as a whole is perceived to be in a homogeneous, steady state of rest before and after the slide. However, the moving body enters a new state which is defined by translation, rotation, and strain, and which is disrupted again by an event that brings it to rest.

Real-world processes are instances of a structured collection of activities (i.e., things that occur) that change the world from one state to another in a predictable fashion (Menzel and Gruninger, 2001). These are chemical, mechanical, thermal, and other types of transformation that result in changes in the participating substantial entities. Understanding events and processes requires the formalization of frameworks that define the time and processes. Here, I follow Galton's (2000) absolute time, which assumes a fixed, absolute temporal framework, independent of anything that happens, and which is populated with processes, events, and states.

Consider a seismic event along the plate boundary fault in the Nankai Trough. Before the onset of the seismic event, the rocks along the seismogenic fault are in a "homogeneous state" characterized by a series of microscopic and mesoscopic processes in space-time (i.e., within different regions of the fault zone at different times), which despite possibly being inhomogeneous at those scales (granularities) can be perceived as homogeneous at a larger, macroscopic scale, and hence a state. The instant of the first foreshock marks the time when the state of the fault zone changes, and the rocks embark on a new state (defined by sorts of processes such as frictional sliding), which later switches to a final state of rest, perhaps marked by the last aftershock. It is actually the processes that bring about states, and then events that disrupt the states, and then more events that disrupt the new state, and so on. If there are no processes, no change would happen, no state would change, and no notion of time would make sense (Bunge, 1977).

## 11.7 Structure of processes

In addition to describing the objects, ontologies should characterize the structure of the world by defining the inter-relationships between the objects occupying space-time, emphasizing the way the constituent objects are put together, and the constraints, characteristics, and facts about these relationships (Menzel and Gruninger, 2001; Needham, 2003).

Space-time is interrupted by many events that define the beginning and end of processes that lead to qualitative change in endurants. Some of these processes coincide or overlap in time and space. Processes can occur synchronically (i.e., within the same time intervals) or polychronically, involving the same or different objects, in the same or different temporal regions. Complex processes are aggregates of one or more processes. This means that the temporal region of an aggregate process (e.g., deformation) may be divided into several sub-intervals within which unique, but possibly (causally) related, subprocesses occurred. Whether an occurrent is a process or subprocess depends on granularity (scale) and one's perspective. For example, a volcanic eruption, viewed at a large scale, may be characterized by extrusion of lava, emanation of gases, and ejection of pyroclastic material. At a finer granularity each of these processes may include other subprocesses, for example, the ejection of pyroclastic material may trigger a mudslide. The subduction may be viewed at a large scale as a process through which lithospheric plates are forced back into the mantle. At a finer scale, this process includes many other related processes that occur at the same time in different regions within the subduction zone, such as accretion, which itself includes subprocesses of offscraping and underplating, each of which includes subprocesses of faulting, rotation, and fracturing among others. This clearly shows that granularity should be a major part of any process ontology, and the description of processes should be based on scale.

One major aspect of processes is that they unfold themselves over a time interval. If we measure a process at an instant, for example, a snapshot of a single seismic event, flow of a river, stress buildup along a fault, bulging of a surface of a volcano, we are only capturing a temporal part of each of these processes. We cannot capture the mereological whole of any of these processes unless we continuously observe and measure (as in a video or a continuous time series) all the temporal parts of each process over its life span.

## 11.8 Process ontology

Philosophers have always pondered on whether or not processes such as diffusion or sliding are as real as endurants, which persist through time and space. Presentists (Lowe, 2002) believe that different parts of space coexist in the present moment, and the present is the only part of time that exists (cf., Kennedy, 2003). Followers of the 4-D paradigm (e.g., Sider, 2001), on the other hand, believe that past and future are as real as present, and coexist with it. They believe that the universe is like a 4-D block, which when cut by the present (time slice), produces a 3-D world at time instants (like the world we see around us). Because of the continuity of time, a series of these time slices gives the illusion of motion for the objects that occupy space-time, like a movie revealed by flipping a deck of cards on which sequential images have been drawn. To appreciate this notion, imagine the universe is like an elongate, rectangular cube of ice, with its long axis along the time dimension, that contains frozen three- and lower-dimensional objects (Kennedy, 2003; Sider, 2001). Also imagine a laser beam outside of this cube is melting the cube at infinitesimally, closely packed instants of time. The melted front of the ice block is the "present moment." The melted cube before the present is the "past," the future is yet to be melted, and the present is the instantaneous temporal boundary between past and future. The unfolding of time at a perceptible pace, unleashes objects, from the front edge of the frozen block (universe), gradually to "move" out along time and space.

The following example may help us better understand this 4-D model. A clay-rich body of rock, which only few moments ago was in a resting, stable state starts to slide by going through a sudden instant ($t_1$) of instability (first event, $e_1$), perhaps caused by intense raining or an earthquake (causal processes). The event dislodges the rock, under the influence of gravity, which initiates the process of sliding ($P_1$). An observer, with a video camera located on the other side of the canyon, records a great number of pictures as the landslide moves down slope, and the "present moment" moves forward with time (as the "4-D ice block melts"). As the time ($t$) reaches a second state-transition instant ($t_2$), the rock stops moving and a state change occurs (second event, $e_2$). The two events and the intervening process of sliding ($P_1$) now only exist in the past. From this moment, the space and time of the events and processes cannot change. Meanwhile, the infinitesimally small "present moments" proceed along the continuum of time dimension toward the future, and the rock remains in a stable state, until another instability event ($e_3$) triggers a new slide, starting a new process ($P_2$), which will end with a later event ($e_4$). Similar scenarios can be envisioned for all geoprocesses such as the eruption of a volcano, movement along a fault, or flow of water.

## 11.9 Spatial location of processes

Events, processes, and time are the main components of the 4-D model. The enduring entities provide the spatial location for the events by participating in these processes. However, the span of processes, i.e., the spatiotemporal region in which the processes occur (Simons, 1987), depends on the spatial and temporal distribution of the qualitative features of rock (e.g., variation of the water content, shear strength) and the external fields (e.g., gravity, stress/strain field). For example, in the case of a landslide, the transition from stable to unstable states depends on the angle of slope (which relates to the gravitational field), amount of clay, pore fluid pressure along the potential plane of failure, and distribution of discontinuities (e.g., fractures) among other factors. In the case of a volcanic eruption, the process depends on such things as presence and type of magma, gas content, temperature, and viscosity of the magma. In the case of a fault zone, the processes of sliding and earthquake generation are controlled by the far field as well as local stresses, and the conditions along the fault, such as pore fluid pressure, and rock's mechanical and chemical properties. These imply that the "span" of a process always includes a "spell" (i.e., exact time interval that the process occupies) around the "present moment," and a "spread" (the exact space the process occupies) where potential participants are located. For example, the far field stresses, that are distributed systematically due to lithospheric-scale plate interactions along the San Andreas Fault zone, lead to inhomogeneously distributed, local spans in which specific processes occur at each "present moment" (i.e., time instant) because certain required criteria are met in those spans. In the 4-D paradigm, this means that the 4-D block is only "melting" selectively those endurants that can participate in the process. Rocks away from the fault zone (i.e., outside of the process zone), which are strong or out of the influence of the far-field stress, will remain stationary and stable, and will not become part of the process. Areas along the fault zone, with rocks that are potentially weak, or in which resolved shear stresses are high enough for sliding to occur, "come into play," and "participate" in the process, thereby defining the spatial location of the sliding process (by their spatial location). Processes always happen in the present moment, but may continue into the future, which later become part of the past.

## 11.10 Hierarchy of processes

Processes, like SNAP entities, can be organized in hierarchical structures using the "is-a" and "part-of" relations, reflecting specialization and part-whole (mereological) relations, respectively. In this chapter, we only deal with the hierarchical structure of processes. If a process $P$ subsumes another process $P_1$ (i.e., $P_1$ is-a $P$), then for all $x$, if $x$ is an occurrence of $P_1$, $x$ is also an occurrence of $P$ (Galton, 2000; Smith *et al.*, 2005). This assertion is stated in first-order logic notation as: $\forall x\ P_1(x) \rightarrow P(x)$. For example, the assertions: Oxidation is-a Weathering or Folding is-a Deformation, state that instances of Oxidation or Folding are also instances of the Weathering or Deformation processes, respectively. The assertion that the processes of recrystallization and recovery are subtypes of the crystal–plastic deformation means that an actual occurrence of grain-boundary migration recrystallization or sub-grain rotation, that occur say during mylonitization in a shear zone, are also mechanisms of crystal plasticity. This means that the actual mylonite that participated in the two subprocesses also participated in the superprocess (i.e., crystal–plastic deformation).

An individual process $p_1$ is "*part-of" p* if and only if an instance of $p_1$ is also an instance-level part-of $p$ (Smith *et al.*, 2005). For example, Rotation part-of Cataclasis or Shearing part-of Frictional_Sliding. The mereological (i.e., part-whole) structure of spatiotemporal processes is defined by temporal parts. For example, the process of flow, diffusion, or subduction may have parts (i.e., phases or stages) that occur say faster than other parts. These parts are assumed to be contiguous, and without temporal gaps. Presence of gaps will lead to the presence of a subprocess or a new process. This means that complex mechanical, chemical, or biological processes have temporal parts that may not be of the same kind as the whole.

Entities such as processes and events include the following parts (Sider, 2001; Stell and West, 2004): (1) 4-D spatiotemporal part, for example, the accretionary prism in the Nankai Trough over the lifespan of the prism that extends from the time since the accretionary process started to form to the present. To see this we need a video that shows the prism over this whole period. To comprehend this, picture a video of a hurricane or tsunami over its lifetime as it changes its spatial location, dimension, and other characteristics. (2) Temporal part, which is the time slice of the 4-D entity, which gives the whole object at a given time instant or interval. For example, a seismic reflection section of an accretionary prism showing a temporal 2-D section of the prism at the present time. This is equivalent to taking a snapshot of the hurricane or tsunami at an instant of time. (3) Historical part, which is a part of the spatiotemporal whole over space and time.

Given that a process leads to changes in the continuants involved in that process, the problem of identity of the spatial entities that participate in the process becomes an issue. Welty and Guarino (2001) relate identity to the problem of distinguishing a specific instance of a certain class from other instances. For example, a granite undergoing weathering changes its composition, color, density, and other properties through time, during the weathering process, but its identify may not change (e.g., weathering of the Idaho batholiths).

Process ontologies provide general descriptions of universal process- and event-types that can occur more than once in many time intervals and spatial regions. For example, deformation, erosion, and melting are process-type. Melting (a process) of a specific rock at a depth of 20 km, in the Nankai Trough subduction zone in southwest Japan, is an occurrence of the universal Melting process type. These occurrences, which are analogous to the instances of endurant entities, occupy specific regions of space-time. In contrast to the continuants that can exist in different time intervals in different spatial regions, occurrences only exist in unique temporal interval and spatial region (i.e., a unique spatiotemporal region). For example, an occurrence of the Crystallization (a universal process type) of a specific magma (an endurant) occurred in Stone Mountain Georgia at a depth of about 10 km, around 300 Ma, forming the Stone Mountain granite. The spatiotemporal region of the occurrence of the crystallization process that formed the Stone Mountain granite cannot be changed. Subsequent processes of uplift and erosion of overlying rocks exposed the granite at the surface in the Atlanta area, modifying its properties (qualities). Notice that time is attached to the notion of process, and cannot be removed from it. This is in contradistinction with the endurants, which are only "observed" through snapshot at the time of observation. For example, while we cannot change the spatial and temporal regions of the 2004 Indonesian tsunami (a perdurant), the meander and the drainage pattern in a river (a continuant) can change with time. Geoscientists are lucky that records of time become preserved in the rock. For example, the time interval of the crystallization process in the Stone Mountain granite can be revealed by isotopic dating of its minerals. It is clear that the crystallization process is not unique to the Stone Mountain area, and can happen anywhere else on Earth.

## 11.11 Formal perdurant relations

The way the entities are configured in the world, through their properties, structure, and relations they stand in, is the main subject of scientific endeavors and hence ontologies (Menzel and Gruninger, 2001). The entities that are part of reality stand in special relations to each other. The main goal of scientific research in every domain is to unravel the fundamental patterns and structures behind the inherent relations among the entities in that domain.

Querying knowledge bases that use the two diverse SNAP and SPAN perspectives requires trans-ontological relations that relate endurants to the processes and events. These formal relations should traverse across: (1) the border between the two perspectives, connecting the endurants and processes together, (2) the granularity boundaries (e.g., microscopic to lithospheric scale), and (3) the temporal divide, e.g., between now and later times (Smith and Grenon, 2004). These relations are between the entities of the two types of ontologies, for example, between the independent, dependent, and spatial entities of the SNAP ontology, and the processual, space-time regions, and time regions, entities of the SPAN ontology. These relations represent such things as dependence, parthood, and identity, and are characterized by signature, arity, and directionality (Smith and

Grenon, 2004). The signature reflects the kind of ontologies that the relation links, such as: <SNAP$_i$, SNAP$_j$> of distinct time indices $i$ and $j$, <SPAN, SPAN>, <SNAP, SPAN>, < SPAN, SNAP>. Geological examples for the main signatures are given below.

An example of a relation between a SNAP entity with another SNAP entity is when a proper part of a SNAP entity is related to its whole, e.g., a fault bend (SNAP) or step (SNAP) is related to a strike-slip fault (SNAP). This is a mereological part-whole relation, and works in this case only if the part and whole have the same time index, that is, the two ontologies have temporal overlap. A SNAP object can also be related to itself at two different times, e.g., snapshots of the waves of a tsunami at two instants of time. An example for relations linking a SPAN to another SPAN entity reflects successive causality; i.e., one earlier process leading into another that follows it. For example, shortening that causes shearing along a fault. An example of mereological relation between two SPAN entities (i.e., processes) is when we relate a subprocess part to its whole, such as that which obtains between comminution and cataclastic flow. The relation between SNAP independent entities and SPAN entities is one of participation. Examples include relations between rock (participant) and deformation (process), or fluid and flow. The relation between a SPAN and SNAP entities is one of involvement. An example is folding that involves bedding plane.

**Case Study: Nankai Trough Seismogenic Zone Experiment**

The IODP Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE) involves drilling, sampling, and instrumenting the Kumano forearc basin, its underlying accretionary prism, subducting sediments seaward of the deformation front, the seismogenic zone of the plate-boundary fault, and a large-scale splay fault within the prism (Tobin and Kinoshita, 2006). The trough, where the Philippine Sea oceanic plate is subducting beneath Honshu and Shikoku islands, is located in the Nankai region of southwestern Japan (Figure 11.1). The NanTroSEIZE project leads to the continuous collection of a large volume of diverse, potentially related data, over a long period of time, that help geoscientists to better understand the geodynamic processes in the seismogenic zone of the plate-boundary fault under the accretionary prism. These data, collected directly during and after drilling in several boreholes, under different P–T and fluid conditions, include rock composition and structure in cores, as well as normal stress magnitude, strain rate, and pore fluid pressure. The data will also include long-term borehole monitoring and measurements (for example, seismicity, hydrology, and stress) in several drill sites that will include an ultra-deep borehole to be drilled during phase 3 into the seismogenic zone and potential underplated sediment. A tremendous amount of spot and time series data will be collected during drilling and post-drilling monitoring of the sites through instruments installed in the boreholes.

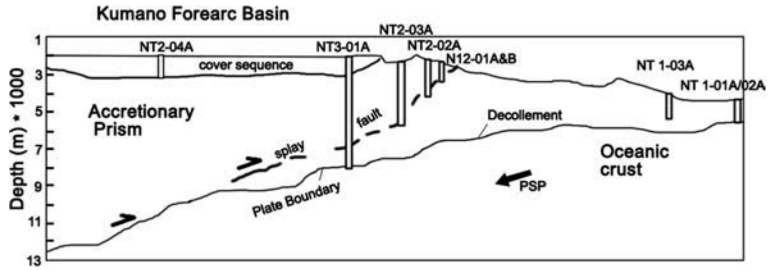



Figure 11.1. Subduction of the Philippine Sea oceanic plate (PSP) beneath Honshu and Shikoku islands, in the Nankai region of southwestern Japan, and position of the Nankai Trough accretionary prism. After Tobin and Kinoshita (2006).

The diverse, large volume of data, collected at different times, in several spatial regions above the subduction zone, under different conditions of stress, strain, temperature, and pore fluid pressure, creates a challenging opportunity to the NanTroSEIZE geoscientists for broadening their understanding of the complex geological and seismic processes in the accretionary prism system. Processing and analyses of these heterogeneous data will lead to difficulties in the computational, query, reasoning, and inferential aspects of scientific research. Unraveling the inherent complexities in such a prism system requires building an integrated set of knowledge bases whose designs are based on the domain knowledge represented by ontologies. Currently, the knowledge about the trough resides in hundreds of published papers and several unintegrated public databases.

Developing the NanTroSEIZE knowledge-based, integrated information system requires designing ontologies that correctly (1) depict our interpretations and theories of accretion above the subduction zone, (2) represent material entities, such as plate-boundary fault, seismogenic zone, and splay fault, and process entities such as accretion and offscraping, as they exist in reality (i.e., in and around an accretionary prism), and (3) establish explicit formal vocabulary (e.g., core, fault, fluid pressure) to share between applications (Noy, 2004). The ontologies will enable efficient handling and management of the onslaught of information expected during the NanTroSEIZE project. Extraction and distillation of information and knowledge from different sets of geological, geophysical, core, and other types of NanTroSEIZE data should focus mainly on unraveling and uncovering the fundamental patterns and structures behind these inherently related data in the Nankai Trough system.

Ontologies in the NanTroSEIZE project may depict and formalize the relation among, for example, the accretionary prism and the forearc basin, offscraping, underplating, and subduction erosion, stable/unstable sliding and seismic/aseismic behavior, friction and mechanical properties of fault rock, and qualities such as pressure, temperature, and strain rate. Once the exact meanings of these terms are defined in ontologies, and the collected NanTroSEIZE data are stored in a database or knowledge base, the data from diverse sources (e.g., core, logging, stress measurement, seismicity, heat flow) can be transformed (with XML) into other forms as input for analysis, visualization, simulation, and modeling applications. The interoperability and data interchange provided by ontologies are central to knowledge acquisition and discovery, and building new models and hypotheses (Babaie and Babaei, 2005a, b).

The seismogenic zone can include properties such as thickness, extent, depth, dip, rock type, structure, fluid, and seismicity. The zone also has mereotopological features (i.e., part-whole and connection relations) that can be defined in the ontology. For example, the boundaries between rocks below and above the seismogenic zone in the plate-boundary fault zone, or between the seismogenic and aseismic parts of the plate-boundary fault, or forearc sediments and the prism, constitute the mereotopology of the system. In other words, the seismic and aseismic segments (i.e., parts) in the plate-boundary fault, or accreted and underplated thrust sheets, constitute the mereology (Simons, 1987) of a prism system, and their boundaries and connections are topological features. Other potential parts in the seismogenic zone (which is a mereological whole) include slivers of undeformed rock, bands of cataclasite or mylonite, sets of riedel shear bands, and other spatial heterogeneities.

The seismogenic zone <SNAP, SPAN> ontology can define fault and faulting, rock and mylonitization, earthquake and wave propagation, and strain and folding/fracturing, some of which are shown in Figures 11.2 and 11.3. It can depict subprocesses that cause other subprocesses, such as dilatation and microfracturing, comminution and cataclasis, heating and recrystallization, or stable sliding and seismicity. In this type of ontology we can formalize which spatial region of the plate-boundary fault is deforming brittlely, and which mereological part is ductile; which thrust sheet (a mereological part of the prism) is older (i.e., accreted earlier), and which is younger. We can depict the sequence of seismicity in both space and time within the seismogenic zone. We can answer questions such as: Which way is the seismic activity propagating through time – up-dip or down-dip? Is the seismic activity moving into the splay thrust with time, or does it migrate along the decollement? Is any part of the plate-boundary fault locking? Is there any spatial-temporal pattern in the stick-slip processes along the fault? What are the mechanical properties of the underplated sediments compared to those in the offscraped or subducting sediments? What are the distributions of the fracture systems in the seismic and aseismic parts of the plate-boundary fault? What are the states of stress and hydrology in different parts of the fault at different times?

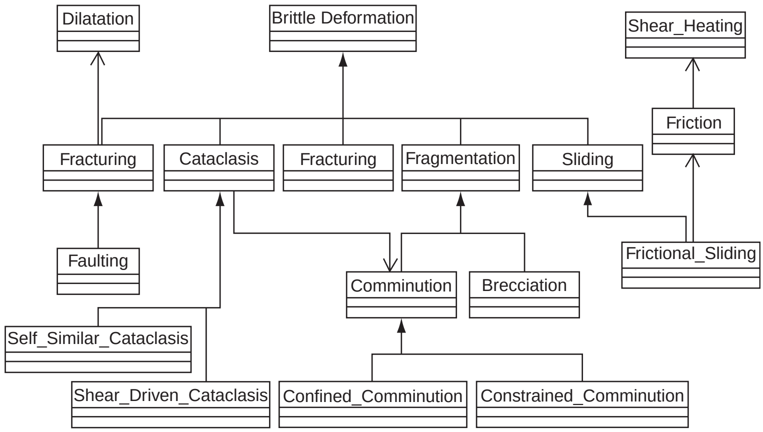



Figure 11.2. UML class diagram showing the hierarchy of part of the processes that may occur in a brittle environment. Arrows with open head show specialization (is-a) relationship (e.g., fracturing is-a brittle deformation). Regular arrows show association (e.g., fracturing leads-to dilatation).

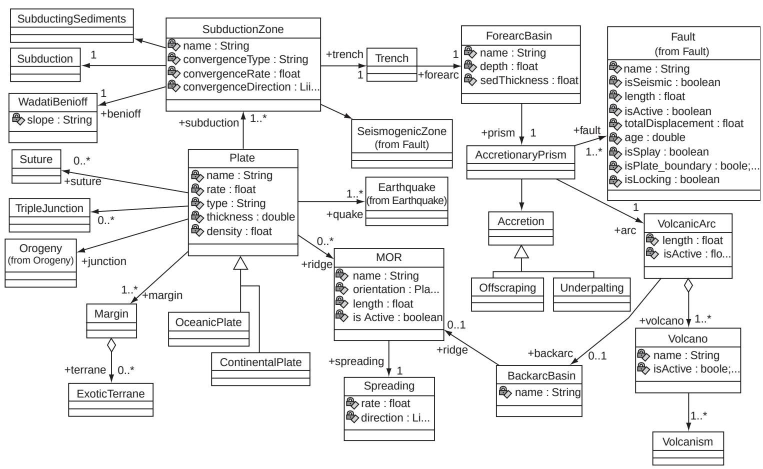



Figure 11.3. UML class diagram showing the relationships among some objects and processes in a subduction zone. The attributes of only some of the classes are shown for brevity.

A seismic event in the seismogenic zone, which is triggered at a time instant t, in the spatial region r of this zone, may initiate concurrently with, or partially overlap, different stages (foreshocks, main shock, and aftershocks) and processes of earthquake generation and fracturing. A down-hole monitoring instrument, recording this seismic activity over time, registers the beginning and end instants (events) of each temporal sub-interval for each subprocess, as well as the location (hypocenter) of each rupture, in addition to other data. Processing of all of these inherently related data can best be handled with software that already ‘knows’ the semantics (i.e., formal meaning) of the data through the underlying ontology. The ontology would “know,” for example, the following transitive temporal rule: “for all instants t (time of fracture nucleation in a damage zone), $t_1$ (time of dilatation), $t_2$ (time of rupture), if $t$ precedes $t_1$, and $t_1$ precedes $t_2$, then $t$ precedes $t_2$” (Allen 1983, 1984). Or, it would “know” the meaning of the temporal statement: if “*i*” (e.g., introduction of high pore fluid pressure) is an initial sub-interval of “*j*” (e.g., tensile fracturing), that is, if the beginning of “*i*” coincides with the beginning of “*j*,” but “*i*” ends before “*j*” ends, or more formally: $\text{beg}(i) = \text{beg}(j) \wedge \text{end}(i) < \text{end}(j)$ (the “^” and “$<$” symbols mean “and” and “precedes,” respectively) (Galton, 2000). We can formalize statements such as: “the beginning of the subinterval of pseudotachylite injection coincides with that of a seismic event,” or “the beginning of dilatancy and uplift start at the same time.” The time series data collected by down-hole instruments in the NanTroSEIZE project require formalization of this kind of temporal, logical statement. The software based on formal ontologies, would “know” from the inference rules embedded in the ontology, as domain scientists know, that high pore fluid pressure reduces the effective shear stress required for fracturing of a given rock, or that increasing the ambient temperature, or reducing the strain rate, enhances ductility. The software would “know” the meaning (semantics) of all the terms defined in the ontology. It would “know,” just like the domain experts, the meaning of such things as ductility, seismic moment, fracture propagation, fracture energy, and strain rate.

## 11.12 Architecture of the NanTroSEIZE ontology

It is expected that several ontologies, and corresponding knowledge base and databases, are needed to represent all the knowledge about the accretionary prism above the plate-boundary fault in the subduction zone. The NanTroSEIZE ontologies can be divided into the SNAP and SPAN groups to reflect the static and dynamic aspects (Grenon, 2003a, b, c; Grenon and Smith, 2004) of the geological entities in the NanTroSEIZE domain as described above. The static (SNAP) group of sub-ontologies will cover things such as the forearc basin, accretionary prism, decollement, splay fault, plate-boundary fault, geophysical well log, drilling profile, drill mud, boundary of forearc basin, and oceanic crust. The dynamic (SPAN) group will include ontologies depicting processes such as deposition, volcanic eruption, metamorphism, deformation, offscraping, underplating, drilling, coring, logging, sampling, and earthquakes. The ontology of the two main groups will be integrated (i.e., related with the <SNAP, SPAN> and <SPAN, SNAP> signatures, see above) to each other based on reality in the domain. For example, the following static and dynamic concepts will be related: fault rock and shearing, heat flow and metamorphism, damage zone and sliding, and seismogenic fault and earthquakes.

For efficient maintenance of the ontologies and knowledge bases, minimization or elimination of redundancies, and simplification of the problem, ontologies will be developed in an autonomous (i.e., modular) yet integrated fashion. This means that the core and drilling ontologies will be developed and maintained separately from, say, the seismicity or stratigraphy ontology. The drill core ontology, for example, will "know" the lithostratigraphic concepts of "formation" or "member" through its communication with the stratigraphy ontology. The communication is made possible by the OWL's and XML's import and namespace facilities. This means that, if required, the drill core ontology will "import" or "include" the stratigraphy or rock ontology, which will make it possible to have an integrated and distributed, yet autonomous, system of ontologies and knowledge bases. The ontology will take advantage of the mereotopology and time-event-process ontologies (Casati and Varzi, 1999; Galton, 2000; Simons, 1987). An advantage of having several autonomous ontologies, as opposed to one large ontology and database, is that each OWL ontology will map into a Java package, relational database, and XML XSD schema. The concepts (i.e., classes) in each ontology will map into different Java classes, XML elements, and relational database tables (Babaie and Babaei, 2005b). A big advantage of this modular architecture and design is that applications that are to be designed to process, analyze, and visualize and query the NanTroSEIZE data and knowledge base, can use the corresponding Java and XML files as well as databases that are based on the ontology. This design will guarantee easier maintenance, integration, and upgrade of the ontologies and databases over time.

## 11.13 Conclusions

Geological reality, like any other, includes both spatial and spatiotemporal entities. Individual (i.e., instances of) spatial geological entities and their spatial parts are constantly changing their properties and state due to the occurrence of individual spatiotemporal entities (e.g., processes). Ontologies are designed to define subsumptive and/or mereological taxonomies by giving a hierarchical structure to the way the spatial and spatiotemporal universal types are related to each other. Ontologies define the type for these two kinds of entities in our domain by defining their properties and special relations. In this chapter, two ontological views of the world were described, and the application of two meta-models to build ontologies was discussed. It was shown that the two views and the related models, which represent the static and dynamic parts of geological reality, could provide a veridical account of the geological environments. Ontologies that include both the spatial objects and the processes that involve these objects can more successfully depict our domain knowledge and theories compared to those that only include spatial objects. The application of the bi-categorical approach in developing ontologies, to the IODP Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE), shows that complex geological settings, such as a subduction zone, can completely be depicted in ontologies that are based on the two meta-models described in this chapter.

## References


Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, **26**: 832–843.

Allen, J. (1984). Towards a general theory of action and time. *Artificial Intelligence*, **23**: 123–154.

Babaie, H. A. and Babaei, A. (2005a). Modeling geological objects with the XML Schema. In *Applications of XML in Geosciences*, ed. H. A. Babaie and R. Ramachandran. *Computers & Geosciences*, Special Issue, **31**: 1135–1150.

Babaie, H. A. and Babaei, A. (2005b). Developing the earthquake markup language and database with UML and XML Schema. In *Applications of XML in Geosciences*, ed. H. A. Babaie and R. Ramachandran. *Computers & Geosciences*, Special Issue, **31**: 1175–1200.

Bittner, T., Donnelley, M., and Smith, B. (2004). Endurants and perdurants in directly depicting ontologies. *AI Communications*, **17**(4), Special Issue on *Spatial and Temporal Reasoning*, pp. 247–258.

Bunge, M. (1977). *Treatise on Basic Philosophy, Volume 3: Ontology I. The Furniture of the World*. Dordrecht, The Netherlands: D. Reidel Publishing Co., 352pp.

Casati, R. and Varzi, A. C. (1999). *Parts and Places: The Structures of Spatial Representation*. Cambridge, MA: The MIT Press, 238pp.

Dietz, J. L. G. (2006). *Enterprise Ontology: Theory and Methodology.* Berlin: Springer, 239pp.

Feng, C.-C., Bittner, T., and Flewelling, D. M. (2004). *Modeling Surface Hydrology Concepts with Endurance and Perdurance*. LNCS 3234. Berlin/Heidelberg: Springer, pp. 67–80.

Galton, A. (2000). *Qualitative Spatial Change*. Oxford/New York: Oxford University Press, 409pp.

Galton, A. and Worboys, M. F. (2005). Processes and events in dynamic geospatial networks. In *Proceedings of First International Conference on Geospatial Semantics*, ed. M. A. Rodrıguez, I. F. Cruz, M. J. Egenhofer and S. Levashkin. *LNCS* 3799, Berlin/Heidelberg: Springer, pp. 45–59.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). *Sweetening ontologies with DOLCE. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. LNCS 2473. Berlin/Heidelberg: Springer, pp. 223–233.

Grenon, P. (2003a). Spatiotemporality in basic formal ontology: SNAP and SPAN, upper-level ontology, and framework for formalization. IFOMIS Technical Report 05, 89pp.

Grenon, P. (2003b). The formal ontology of spatio-temporal reality and its formalization. In *Foundations and Applications of Spatio-Temporal Reasoning (FASTR)*, ed. H. W. Guesguen, D. Mitra and J. Renz. AAAI Spring Symposium Technical Report Series. Menlo Park, CA: AAAI Press, pp. 27–34.

Grenon, P. (2003c). Spatio-temporality in basic formal ontology: SNAP and SPAN, upper-level ontology, and framework for formalization. IFOMIS Technical Reports Series 05/03.

Grenon, P. and Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, **4**(1): 69–104.

Grenon, P., Smith, B., and Goldberg, L. (2004). Biodynamic ontology: Applying BFO in biomedical domain. In *Ontologies in Medicine*, ed. D. M. Pisanelli. Amsterdam: IOS Press, pp. 20–38.

Guarino, N. and Welty, C. (2002). Identity and Subsumption. In *The Semantics of Relationships: An Interdisciplinary Perspective*, ed. R. Green, C. A. Bean and M. Sung Hyon. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 111–125.

Hawley, K. (2001). *How Things Persist*. Oxford: Clarendon Press, 221pp.

Heller, B. and Herre, H. (2003). Ontological categories in GOL. In *Process Theories: Cross-Disciplinary Studies in Dynamic Categories*, ed. J. Seibt. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 57–76.

Kennedy, J. B. (2003). *Space, Time, and Einstein: An Introduction*. Montreal: McGill-Queen's University Press, 192pp.

Koslicki, K. (2008). *The Structure of Objects*. New York: Oxford University Press, 288pp.

Lowe, E. J. (2002). *A Survey of Metaphysics*. New York: Oxford University Press, 402pp.

Lowe, E. J. (2005). Vagueness and endurance. *Analysis*, **65**(2): 104–112.

Mayo, B. (1961). Objects, events, and complementarity. *The Philosophical Review*, **70**(3): 340–361.

Menzel, C. and Gruninger, M. (2001). A formal foundation for process modeling. *Proceedings of the International Conference on Formal Ontology in Information Systems*, pp. 256–269.

Miller, K. (2005). The metaphysical equivalence of three and four dimensionalism. *Erkenntnis*, **62**: 91–117.

Needham, P. (2003). Continuants and processes in macroscopic chemistry. In *Process Theories: Cross-Disciplinary Studies in Dynamic Categories*, ed. J. Seibt. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 237–265.

Noy, N. F. (2004). Semantic integration: A survey of ontology-based approaches. *SIGMOD Record, Special Issue on Semantic Integration*, **33**(4): 65–70. Available at: www.sigmod.org/sigmod/record/issues/0412/.

Pribbenow, S. (2002). Meronomic relationships: From classical mereology to complex part-whole relations. In *The Semantics of Relationships: An Interdisciplinary Perspective*, ed.R. Green, C. Bean and M. Sung Hyon. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 35–50.

Raskin, R. and Pan, M. J. (2005). Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). In Applications of XML in Geosciences, ed. H. A. Babaie and R. Ramachandran. *Computers & Geosciences*, **31**: 1119–1125.

Sider, T. (2001). *Four-Dimensionalism: An Ontology of Persistence and Time*. Oxford: Clarendon Press.

Simons, P. (1987). *Parts: A Study in Ontology*. Oxford, New York: Oxford University Press, 390pp.

Smith, B. (2001). Fiat objects. *Topoi*, **20**(2): 131–148.

Smith, B. (2003). Ontology. In *Blackwell Guide to the Philosophy of Computing and Information*, ed. L. Floridi. Oxford: Blackwell, pp. 155–166.

Smith, B. (2004). Beyond Concepts: Ontology as reality representation. In *Formal Ontology and Information Systems 2004*. Amsterdam: IOS Press, pp. 73–84.

Smith, B. and Grenon, P. (2004). The cornucopia of formal-ontological relations. *Dialectica*, **58**(3): 279–296.

Smith, B. and Rosse, C. (2004). The role of foundational relations in the alignment of biomedical ontologies. In *Proceedings MEDINFO 2004*, ed. M. Fiechi *et al*. Amsterdam: IOS Press, pp. 444–448.

Smith, B. and Varzi, A. C. (2000). Fiat and bona fide boundaries. *Philosophy and Phenomenological Research*, LX, **2**: 401–420.

Smith, B., Ceuster, W., Kellages, B. *et al*. (2005). Relations in biomedical ontologies. *Genome Biology*, **6**, R46. Open Access.

Stell, J. G. and West, M. (2004). A four-dimentionalist mereotopology. In *Formal Ontology in Information Systems*, ed. A. C. Varzi and L. Vieu. Amsterdam: IOS Press, pp. 261–272.

Tobin, H. J. and Kinoshita, M. (2006). NanTroSEIZE: The IODP Nankai Trough Seismogenic Zone Experiment. *Scientific Drilling*, **2**: 23–27.

Welty, C. and Guarino, N. (2001). Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering*, **39**: 51–74.

# Part V

## Web services and scientific workflows

# 12

# Service orientation in the design of a community hydrologic information system

ILYA ZASLAVSKY AND DAVID R. MAIDMENT

## 12.1 Introduction

Hydrology is an observational science concerned with the motion of the water through the hydrologic cycle. The Committee on Opportunities in the Hydrologic Sciences (1991, p. 43) defined hydrology in this way: "Hydrology is the science that treats the waters of the Earth, their occurrence, circulation, and distribution, their chemical and physical properties, and their reaction with their environment, including their relation with living things." Bearing in mind the vast extent of the waters of the Earth, their ceaseless motion, the existence of water as ice, liquid, and water vapor, and the global distribution of water throughout the atmosphere, land surface, and subsurface, hydrologic science is complex. Constructing a comprehensive picture of its phenomena involves the assembly of data from many sources, both organizations and individual researchers. The Committee on Opportunities in the Hydrologic Sciences (1991, p. 265) stated further that "Advances in the hydrologic sciences depend on how well investigators can integrate reliable, large-scale, long-term datasets."

Observation of hydrologic phenomena can be by direct sensing, or by remote sensing, depending on whether the measurement of water properties is made by a gage or water sample directly from a water body, or whether the water property is inferred from a camera or sensor measuring reflected or emitted radiation at a distance from the water body. Directly sensed observations from rainfall or streamflow gages are expressed as time series, in which a regular sequence of data values is indexed by the measurement time, and georeferenced by the latitude and longitude of the sampling site. Water quality samples are taken in the field and returned to a laboratory for analysis of the chemical and microbiological content of water. Such sampling is also irregular in time but many constituents are measured with each sample, and over a period of time, if a sequence of samples is obtained, a corresponding set of time series of measured results is produced. The water level in groundwater wells changes slowly, and at many locations is sampled infrequently and irregularly in time. Within the United States, time series of water properties are measured at more than 2 million locations by hundreds of water agencies and researchers. Much of these data are accessible through web pages, but each organization has its own way of publishing their data on web sites, both as to the structure of the web site, the format of its data output, and the manner in which they name the water properties being observed.

The concept of a service-oriented architecture is defined by Josuttis (2007) as “a concept that applies to large, distributed information systems that have many owners, are complex and heterogeneous, and have considerable legacies from the way their various components have developed in the past.” This concept fits very well with the situation faced with water observations data. Compared to other data sources, such as remote sensing or climate grids, point water observations data are syntactically and semantically complex; that is, there are many data formats and variable definitions for the same basic data types. However, water observations data are relatively “light-weight” when compared to remote sensing or climate grids – each time series, describing one variable measured at one location by one organization over a period of time, is a small compact data package that is well suited to transmission through the Internet in XML format. Larger data collections comprising many such series can be sent as a sequence of elementary data parcels of the same type. Within the United States, some of the key water observations data sources are the USGS’s National Water Information System (NWIS) for surface and groundwater quantity and quality, the EPA’s Storage and Retrieval (STORET) repository for water quality, USDA’s Snowpack Telemetry (SNOTEL) for snow conditions, and the Integrated Surface Database of climatic data from the National Climate Data Center. Many comparable data collections from state, regional and local water agencies, and from academic investigators are also included in the service-oriented architecture described here.

The ability to easily discover, retrieve, and interpret hydrologic time series regardless of physical location or peculiarities of individual data sources is one of the key requirements for a hydrologic information system (Bandaragoda *et al.*, 2005; Zaslavsky, 2006). Integrating hydrologic data across agency boundaries and combining them with observation time series collected by multiple academic research projects has been challenging due to the extreme heterogeneity in how the different repositories are organized, described, accessed, and maintained. To support advanced data-intensive hydrology research, it should be possible to access such disparate information sources in a uniform standards-based manner; easily publish locally collected observational data; and easily interface with a variety of community models and analysis and visualization codes. A higher level of interoperability of hydrologic information will enable applications and research designs that were either prohibitively expensive or just not possible before. Detailed analysis of water balance at the local, regional, and continental scales would enable a number of tasks, including creation of a comprehensive national portrait of the state of water resources; modeling of the system at finer spatial and temporal scales; and discovery of regions with comparable patterns of hydrologic cycle at the global scale.

This chapter describes a community cyberinfrastructure project, named CUAHSI Hydrologic Information System (http://his.cuahsi.org), conducted under the aegis of the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI: www.cuahsi.org). The goal of the project is to implement the above vision and advance hydrologic science and education by enabling uniform access to hydrologic data and metadata, analysis and visualization tools and models for a wide audience of academic hydrologists (Maidment, 2009a; Maidment *et al.*, 2009b). The project has developed now to a point when the initial infrastructure components have been deployed at universities and hydrologic observatories across the United States, and its web service methodology and specifications are adopted or replicated by a range of government agencies and academic projects. In Section 12.2, we outline the design principles of a hydrologic information system as derived from the needs and specifics of the hydrologic research community and as influenced by parallel technological advances. Section 12.3 describes the main components of CUAHSI HIS as a system of distributed water data services and applications. It is followed by a discussion of community standards for hydrologic data exchange developed or adopted by the HIS project. We conclude with an outline of new challenges and prospects of CUAHSI HIS development as a standards-based community information system for hydrology.

## 12.2 SOA design for hydrologic research community

A design strategy for large information systems, *service-oriented architecture* (SOA) relies on a collection of loosely coupled self-contained functional modules, referred to as services, which communicate with each other and can be called from multiple clients or chained in processing workflows. Service-oriented architecture is the theme of the chapters in this section of the book. Service is a metaphor for functionality accessible over the Web or another network, where the communication follows a standard protocol and remote functions can be invoked irrespective of their underlying implementation language or computing platform, as long as the service provider and the service consumer follow a communication "contract." The services can generally represent any packaged functionality that is sufficiently granular and independent to encourage reuse of data and other resources for different user scenarios. The goal of SOA is to allow such services, which are likely to be developed and maintained at different institutions, to be combined into a multitude of applications.

Not surprisingly, this development paradigm has been central to several large projects with the mission to enable collaborating research teams to share and integrate geographically distributed heterogeneous scientific data and tools. As mentioned in the introductory section, these projects have also been referred to as cyberinfrastructure projects, after the term was articulated by NSF (NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, 2003) as the cornerstone of NSF's strategy for connecting high-performance computing resources, information resources, and researchers to support scientific discovery. Such projects have emerged in many science disciplines, including in the solid earth sciences (The Geosciences Network, or GEON: www.geongrid.org), ecology (The National Ecological Observatory Network, or NEON: www.neoninc.org), oceanography (The Ocean Observatories Initiative, or OOI: http://oceanobservatories.org), and atmospheric sciences (Linked Environments for Atmospheric Discovery, or LEAD: http://lead.ou.edu), to name just a few of NSF-supported initiatives. The projects have different scope and research foci, addressing challenges that range from efficient management of real-time multimedia data streams from multiple sensors to heterogeneous data integration to visualization of large distributed datasets, metadata and data publication and discovery services, long-term data preservation, secure access to and pooling high-performance computing resources, and distributed data analysis and modeling. Service-oriented architecture is the backbone of such large projects, focusing developers' attention on encapsulating system functionality as a collection of service modules, and developing applications that are service consumers.

In creating a hydrologic information system, leveraging cyberinfrastructure developed in neighboring earth science projects mentioned above, as well as generic data exchange specifications, is an attractive option, especially where it concerns standard infrastructure components and technologies for managing common data types. This approach makes cross-disciplinary data sharing easier, and lets the HIS design team focus on the core services needed by hydrologists. However, the adopted components, specifications, and design ideas shall be tuned to specific characteristics of the hydrologic research community. These characteristics include community organization, research paradigms, user preferences, the adopted semantics and vocabularies, expressed or (more likely) implied information models used by different data providers, and common research tools and data resources. Our experience in developing the hydrologic information system architecture, and surveys of hydrology researchers conducted throughout the project, brought the following conclusions about the specifics of hydrological cyberinfrastructure:

(1) Researchers in hydrology rely to a large extent on federally organized, observational data repositories, including the above mentioned USGS NWIS, EPA STORET, USDA SNOTEL, and NCDC Integrated Surface observation network. In addition, national-scale gridded datasets are available from MODIS, DAYMET, and similar observation and modeling projects. The data are in public domain, and repositories are freely accessible via respective web portals. This has at least

two consequences for the cyberinfrastructure design: (1) Making access to such repositories simpler, more uniform, and model-driven would directly support research efforts for a large group of hydrologists, and (2) the emphasis on data ownership and restricted access is relatively weaker in hydrologic analysis as compared with other communities where cyberinfrastructure projects are being developed, e.g., in geology or neuroscience. This justifies the focus on common web service interfaces and a hydrologic data discovery and retrieval portal easing access to federal observation network archives, without the necessary service authentication as is customary in other portal environments. At the same time, there is a strong desire to control access to measurements generated in the course of academic projects, at least until the research is published.

(2) The hydrologic research community appears to be organized, to a larger extent than other geoscience communities, by natural boundaries that are regional in extent, specifically by river watersheds and catchments. This suggests a "natural" network of relatively autonomous hydrologic data nodes that provide access to locally collected and curated data resources and applications. Therefore, development, deployment, and technology support for such nodes is an important component of creating a networked environment for hydrologic data sharing. At the same time, data and model outputs published at the nodes are federated and made accessible through a common gateway, to enable meta-analysis and hydrologic modeling at the regional and continental scales.

(3) From the data perspective, there are sub-groups in the community focused on analyzing point time series (and incidentally relying largely on the Windows platform), and focused on analyzing remote sensing data and time series (and using Linux/Unix platforms to a larger extent than the first group). Supporting different groups of researchers requires that HIS relies on cross-platform data management services and portals that can be deployed and accessed in both environments.

(4) Due to existing water use and water management regulations, which are local in nature, a community hydrologic information system creates a lot of opportunities for partnership with public sector entities (such as state and local water authorities and related small engineering firms). Such partnerships can leverage data publication and information integration components of the system, providing access to local hydrologic data for government departments, researchers, and citizen groups.

(5) As revealed by CUAHSI user surveys, and our subsequent deployment experience, the community relies on several common COTS (commercial off-the-shelf) software packages, most importantly Excel, ArcGIS, and Matlab. Enabling access to time series repositories from these clients, as well as from such popular coding environments as FORTRAN and VisualBasic, is an important consideration for HIS architecture. Yet, there is a lack of a unifying desktop hydrologic analysis and modeling platform, which creates a software development opportunity.

(6) Hydrology is an integrative science, and hydrologic models rely on data inputs from several neighboring disciplines: atmospheric and ocean observations, soil surveys, vegetation and land use, geomorphology and geology, social and demographic datasets, etc. Consequently, the HIS infrastructure supports interoperation with data and processing services being developed in other earth science disciplines, and ideally follows similar standard formats for handling spatiotemporal information. In the same manner, HIS maintains a single and easy-to-use gateway to its resources, including a catalog of available hydrologic data accessible via standard services from other domain systems.

(7) While the organizing model for hydrologic measurements used in HIS is the cube with spatial, temporal, and attribute dimensions (Horsburgh *et al.*, 2008), treatment of each dimension is complex. Different hydrologic analyses may require different representations of space and time. For example, a hydrologic information system may need to expose observations in both local and UTC time: the former is common in large-scale watershed-level studies, while the latter may be needed for compatibility with weather and climate data services. Additional complexity is added by varying time support for differently recorded and aggregated data, especially for derived data with often incomplete provenance information. The same applies to handling spatial locations of hydrologic observations, where multiple types of offsets from hydrologic landmarks are commonly recorded, and data are collected and organized by a variety of spatial sampling features (time series at a site, measurements along a transect, trajectory, or a vertical profile, etc.).

(8) Perhaps the most complex dimension of the cube is the attribute one. Federal data repositories often contain large numbers of measured parameters, in particular on water quality. For example, USGS NWIS lists 17 736 unique parameter codes (USGS NWIS, 2010; as of June 2010), while a complete list of EPA STORET parameter codes includes 16 521 entries (EPA STORET, 2010; as of June 2010). A growing number of new measurement techniques and sensor instruments, coupled with often inconsistent semantics of variable and measurement unit descriptions across observation networks, make the development of observation data catalogs and knowledge bases, and reconciling semantics of parameter measurements, indispensable. Therefore, scalability in terms of measured parameters, and their efficient management to support search across observation networks, is one of the main performance characteristics of a comprehensive hydrologic information system.

(9) For a national-level system, scalability in terms of measurement station counts and observation counts is also critical, and remains a challenging issue, as existing standard services for observational data, e.g., the OGC Sensor Observation Service (OGC SOS, 2007), are usually ill-equipped to carry massive volumes of data. The size of the metadata catalog for US observations assembled by the HIS project highlights the scale of this problem. This requires a careful scalability

design throughout the system and reliance on special techniques such as OLAP (Rodriguez *et al.*, 2010), which have not been used widely in hydrology yet.

(10) Given the huge and growing volumes of hydrologic observations available in an integrated information system, the data need to be managed by certain agreed levels of aggregation. As we determined in the course of the CUAHSI HIS project, time series provide such a natural grouping of observations which is common across multiple repositories and applications. Efficient management of time series became the central data management issue in HIS, which maintains a comprehensive series catalog, and a series-centric service API. Further, time series may be grouped into time series collections, one type of which, representing a named thematic collection of time series, we termed a hydrologic theme. In our experience, hydrologic time series and themes are the core aggregate objects that hydrologists manipulate, and hence require standard means for describing, publishing, discovering, retrieving, integrating, and visualizing them as part of a hydrologic CI. One of the key challenges here is the disparity in time series definitions across major hydrologic data providers.

This list is not exhaustive, and is presented here to illustrate the key features and challenges in building an information system for hydrologic data. While necessarily generalized and simplified, the features listed above let us conceptualize the architecture components and deployment strategies of a hydrologic information system. From this list we omitted issues that are common across different CI efforts, such as supporting single sign-on authentication, data quality control and maintaining data provenance, long-term data preservation, virtualization of storage and processing capabilities, ability to tailor outputs to audiences with different levels of expertise, data replication and sustainability, establishing trustworthy services, etc., not to mention a range of economic, societal and organizational challenges brought in by the data and service sharing environment. Instead, here we focus on the characteristics that are germane to hydrologic research and the hydrologic research community, while referring to the growing CI literature for the treatment of general issues. We pose that SOA, being an innovative yet mature enough approach to distributed information system development, is a proper strategy for HIS design at this time. The next section describes the main components of HIS SOA, and illustrates how service-oriented computing supports HIS flexibility, incremental and scalable development, and ease of configuration for new research scenarios.

## 12.3 Main components of HIS service-oriented architecture

The initial focus of the CUAHSI HIS CI development is limited to observational data collected at stationary points: gages, measurement stations, wells, etc. The

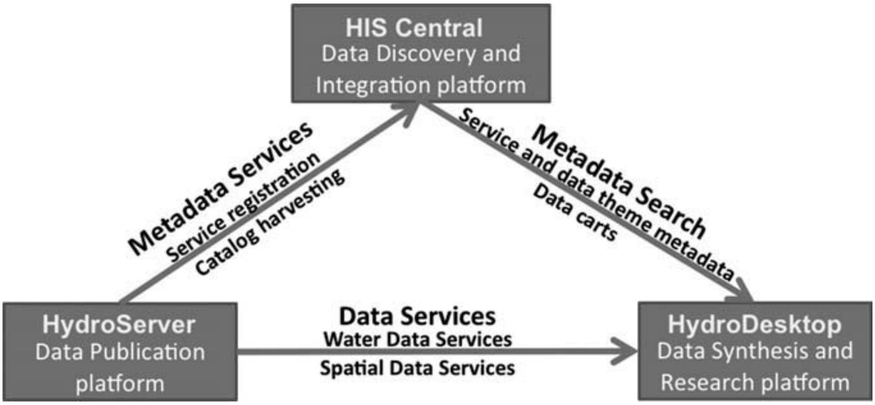



Figure 12.1. The key components of the CUAHSI HIS service-oriented architecture.

semantics of such data is captured in a point observations information model, which is implemented as a relational schema called the Observations Data Model, or ODM (Horsburgh *et al.*, 2008) and as an XML schema called the Water Markup Language, or WaterML (Zaslavsky *et al.*, 2007). WaterML-based services, called *WaterOneFlow* services, are the core of the CUAHSI HIS cyberinfrastructure. These services support discovery and retrieval of observational data from multiple distributed sources, including federal and state repositories and academic projects. A high-level view of the CUAHSI HIS service-oriented architecture is shown in Figure 12.1. The key components include:

- **The CUAHSI HIS data publication platform**: a software stack for ingesting both static and streaming data into an ODM-compliant database, and exposing them via web services compliant with WaterML. This software stack, also referred to as **HydroServer**, has enabled individual scientific investigators and research groups to publish their hydrologic observations and make them available within CUAHSI HIS (Horsburgh *et al.*, 2010; Horsburgh *et al.*, 2009; Whitenack *et al.*, 2007). The HydroServer suite of applications can be downloaded from the project web site. It is available in two versions: besides the complete HydroServer suite, which relies on commercially available components, there is an HIS Server Lite, with limited functionality and a software stack composed of freely available components. HydroServers are installed at various universities and public agencies. To publish hydrologic time series, data managers can use various data loaders to ingest observational data into instances of ODM, perform quality control of the data, configure web services over the ODM databases, and customize an online mapping application for accessing the new data.

Table 12.1. *Federal agency water data services at HIS Central catalog (as of November 2010)*

| Network Name | Site Count | Value Count (thousands) | Earliest Observation | Notes |
|---|---|---|---|---|
| NWISDV | 31,800 | 304,000 | 1/1/1861 | WaterML-compliant GetValues service from NWIS, catalog ingested |
| EPA | 236,000 | 78,000 | 1/11/1900 | SOAP wrapper over WQX services, catalog ingested |
| NWISUV | 11,800 | 169,000 | 120 DAYS | WaterML-compliant GetValues Service, catalog ingested |
| NCDC ISH | 11,600 | 3,000* | 1/1/2005 | WaterML-compliant GetValues service from NCDC |
| NCDC ISD | 24,800 | 18,200 | 1/1/1892 | WaterML-compliant GetValues service from NCDC |
| NWISIID | 376,000 | 86,500 | 9/1/1867 | SOAP wrapper over NWIS web site, catalog ingested |
| NWISGW | 834,000 | 8,490 | 1/1/1800 | SOAP wrapper over NWIS web site, catalog ingested |
| RIVERGAGES | 1,300 | 264,000 | 1/1/2000 | WaterML compliant REST services from the Army Corps of Engineers |

* Estimated

- **The central data discovery and integration platform** comprises a water metadata catalog and a service registration, harvesting, and discovery application called **HIS Central** (Whitenack *et al.*, 2008). It enables users and applications to search across a comprehensive catalog of hydrologic time series available from academic and government data sources. As of November 2010, the HIS Central Catalog contained metadata for about 23.3 million observational time series, for over 18,000 variables measured at 1.96+ million sites, providing web service access to over 5.2 billion data values. This includes several federal repositories that are critical for hydrologic research (USGS NWIS, EPA STORET, NCDC Integrated Surface Hourly (ISH) and Daily (ISD) datasets, data from the Army Corps of Engineers, USDA, etc.), many state agencies, and over 50 academic research sites. Table 12.1 shows the content of selected federal agency services included in the catalog.
- The system currently brokers data requests averaging over 5000 per day, which is an increase of nearly 200% from the previous year, and 1200% increase over 2008 (see Figure 12.2). Once data managers configure a HydroServer, the water data web services are registered at the central HIS service registry. As a new water

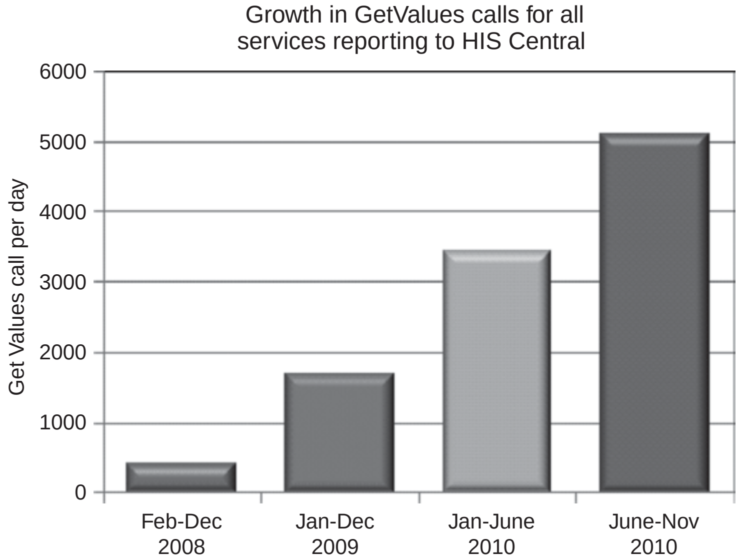



Figure 12.2. *GetValues* requests per day brokered by HIS Central.

data service is registered, the HIS Central application harvests observation metadata from the service (sites, variables, and periods of record that are accessed via the service) and appends it to the central metadata catalog. In addition, HIS Central supports semantic tagging of the registered data, by associating the harvested variables with concepts from a community ontology of hydrologic parameters.

- **The data synthesis and research platform** includes a range of applications that consume web services from HIS Central, HydroServers, and other partner sites that comply with the WaterML standard or otherwise publish their data as web services. The main component in this software suite is **HydroDesktop**, a .NET-based open source desktop application that supports processing of both geospatial data and observational time series. It is designed as a single comprehensive environment for analysis and synthesis of hydrologic data, development of local analytical databases, management of time series collections, and model and workflow integration based on local and remote services. In particular, HydroDesktop relies on the notion of a *hydrologic theme*, which is defined as a thematically organized collection of hydrologic time series expressed as a spatial database (e.g., Texas salinity theme, Chesapeake Bay nitrogen theme). Hydrologic themes allow users to query and subset hydrologic time series based on the numeric values of observations stored in a theme (e.g., select all data values greater than a threshold value for a particular variable) – a key HIS functionality that otherwise cannot be efficiently supported by the physically separate data servers within the current HIS implementation. In addition to HydroDesktop, several client applications over water data services have been developed within and outside the project, including extensions of Microsoft Excel and ESRI's ArcGIS, which support discovery and retrieval of hydrologic data via HIS Central, and a Google Earth$^{TM}$-based interface to the water data services developed at CSIRO, Australia.

The three components reflect the key functionality expected of a hydrologic cyberinfrastructure, as described in Section 12.2. At the same time, they follow an SOA design strategy, including reliance on autonomously maintained services, ability to publish hydrologic data as services, discover and chain services to address a variety of research tasks, and extend HIS as needed by bringing in additional services or service clients without re-engineering the entire system.

## 12.4 Development of community standards for hydrologic data

Reliance on community information models and independently governed open standards for hydrologic data exchange is a central requirement for building hydrologic cyberinfrastructure. Common CI designs represent a blend of commercial, academic and community-developed software, governed by different licensing models and use constraints, residing on different computing platforms, and managed by different geographically distributed teams. Creating a community consensus about data exchange protocols does not only "glue" multiple loosely coupled components into an interoperable working system, but also ensures that the system is sustainable and extensible, can accommodate third-party components, encourages competition and innovation, and prevents a vendor lock-in. The design of CUAHSI Water Markup Language, an XML exchange schema used by the current water data services within CUAHSI HIS, was motivated by our attempt to reconcile structural differences across key hydrologic data repositories in the USA (NWIS, STORET, and NCDC) and create a uniform and scalable backbone for the HIS infrastructure. Another starting place for the WaterML schema was the CUAHSI ODM, which outlined recommended metadata for hydrologic observations. An additional goal of WaterML design was to create a fairly small and rigid foundation for the hydrologic data infrastructure, to ensure that compatible services and clients can be created quickly and information exchanges are unambiguous, thus creating the lowest barrier for adoption by hydrologists.

As an XML schema, WaterML defines valid elements, their attributes and relationships, and provides simple and complex type definitions (including *timeSeriesResponseType*, *siteInfoType*, *variableInfoType* – see Zaslavsky *et al.*, 2007). In addition to the syntactic declarations, WaterML contains a permissible set of terms (vocabulary) that can be slotted into the open placeholders that the grammar framework provides. The HIS hosts several controlled vocabularies (CV) of various degrees of complexity (including quality control levels, units, censor codes, sample mediums, value types), and maintains an online system for managing the CVs shared between ODM and WaterML (http://his.cuahsi.org/mastercvreg.html). Enforcing the CVs in ODM and water data services eases interpretation of hydrologic time series conveyed over WaterML. This model, however, has not worked for variable names, in particular as used in large independently managed data repositories such as NWIS and STORET. In order to accommodate the semantic differences in variable naming and support parameter-based search across such sources, a CUAHSI HIS variable ontology is used. Once variables in each registered hydrologic dataset are associated with ontology concepts in the HISCentral application, the respective time series can be discovered in HydroDesktop and other discovery and integration clients. Currently, over 96% of measurements referenced in the HISCentral metadata catalog are available for ontology-based discovery. Establishing community consensus mechanisms for curating and evolving the parameter ontology is an important component of HIS sustainability and wider adoption, and is a focus of the ongoing ontology development (see Chapter 3 for more discussion on the use of ontologies for science applications).

WaterML 1.x, which is our initial attempt to capture the semantics of hydrologic data discovery and retrieval, has been validated in multiple applications in the community of academic hydrologists in the USA. However, additional requirements necessitated its further development towards a better alignment with other information exchange standards, in particular those being developed within the Open Geospatial Consortium. The enhanced WaterML, referred to as WaterML 2.0, is being developed as a profile of the OGC Observations & Measurements specification (OGC O&M, 2007). Compliance with O&M and other OGC specifications developed within the OGC Sensor Web Enablement (SWE) framework will support interoperability with earth science data across domains. The WaterML 2.0 work is ongoing as a collaboration between the Land and Water Division of CSIRO, Australia, CUAHSI HIS, and several partner agencies and companies within the recently created Hydrology Domain Working Group (OGC HDWG, 2009). The mission of the Hydrology DWG, which is a joint activity of OGC and the World Meteorological Organization's Commission for Hydrology, is to orchestrate a community process supporting development of international standards for hydrologic data exchange, and to promote best practices in standards-based hydrologic data management.

## 12.5 Key community requirements and technical challenges of CUAHSI HIS

Several federally funded projects in the USA are now using HIS software (e.g., Ball *et al.*, 2008; WATERS Network, 2008). The CUAHSI ODM database schema has been ported to MySQL and PostgreSQL, while WaterML-based web services have been adopted by the United States Geological Survey, the National Climatic Data Center, and the Army Corps of Engineers to publish their hydrologic data (see Table 12.1). At the same time, the exponential growth of the system over the last two years and ample feedback from the hydrologic research community on initial deployments has helped us clarify the limitations of the current system and additional challenges. Below we summarize the most important challenges following the above outline of the key HIS components.

### *12.5.1 Challenges of the data publication platform*

Contributing hydrologic data to the current HIS infrastructure requires that individual researchers or research groups establish a computer server, install the HydroServer software stack, obtain commercial software licenses as needed, load their data and configure web services, and maintain the hardware and software to ensure the data remain available. It has been our experience that the hardware, software, and personnel resources required to establish and maintain a HydroServer can be a barrier for many individual researchers who would like to publish hydrologic data but lack resources for establishing or maintaining their own HydroServer. Allowing potential data contributors to join the federation of shared hydrologic data without the need to set up and maintain their own hydrologic data servers is one of the key conditions for a wider adoption of the system that has been expressed by the hydrologic research community.

We have also observed that data publishers need more granular control over access to their data, in particular as time series pass quality assurance workflows. The identification of entire services as public or private should be complemented by the ability to control access at the data series level. Additional requirements often voiced by the community include the ability to publish additional types of data using standard formats (in particular, gridded datasets and spatial data), and republishing hydrologic analysis and modeling results.

### *12.5.2 Central metadata catalog and service registry challenges*

The shared metadata catalog at HIS Central, where observation series metadata is collected, associated with a hydrologic community ontology, and made available for searching via web services, is a unique feature of the CUAHSI HIS that enables unprecedented access to the United States' water data across many independently managed observation networks and datasets. However, maintaining the central metadata catalog in synchronization with multiple sources of hydrologic data has been one of the main bottlenecks to HIS. While automatic metadata harvesting has been conducted on a weekly schedule for HydroServer-based academic data sources, harvesting metadata from large federal and state data repositories has followed a different model. CUAHSI HIS collaborators at USGS, EPA, and several other agencies periodically place their catalogs or entire repositories, in either ASCII or database dump format, on FTP sites for the CUAHSI HIS project to ingest. The information is imported into Microsoft SQL Server, and database views are created to present the data in a format that can be ingested by the HIS Central harvesting application. The sheer size of the catalogs, difference in information models, and incompatibility of database platforms make such updates a lengthy process. This underscores the need for a new method for maintaining cross-agency catalog information, where data contributors and partners independently commit to maintaining their own sections of the catalog and agree on a standard format for exchanging catalog fragments.

Once time series catalog information is harvested into HIS Central, catalog maintainers or HydroServer data managers proceed to associating measured variables with concepts from a hydrologic domain ontology. For large segments of the catalog, which reflect the content of federal and state agency data repositories, this process is time consuming given the number of measured variables available in the catalog. Exposing the catalog tagging procedure to our collaborators at federal and other agencies will lead to a higher degree of transparency, accuracy in ontology mappings, availability of catalog records, and accuracy in search results.

### *12.5.3 Hydrologic data synthesis and research platform challenges*

Development of hydrologic themes with HydroDesktop enables queries of time series information based on observation values, once users have downloaded theme data to their own machine (e.g., "find all nitrogen measurements in a given theme that exceed a given threshold"). There are, however, limitations to the number and size of themes that can feasibly be assembled and stored on individual workstations due to the volume of available data and processing time required to gather all of the actual time series data values for large themes from distributed data services. Therefore, the ability to query over arbitrary hydrologic themes or theme collections is currently missing from the CUAHSI HIS.

Within the current HydroDesktop implementation, downloaded themes may contain potentially incompatible time series (e.g., time series of values for the same variable but reported in different units, or available at different time intervals). Such incompatibilities present a serious challenge for research. They must be reconciled before the series can be integrated for analysis or joint visualization – which is currently done manually after the data are downloaded to a user's workstation. With the already large and growing amount of data that are available online for assembly into themes, a consistent solution for reconciling incompatibilities prior to download is desirable.

## 12.6 Conclusion

We present the design and main components of the CUAHSI Hydrologic Information System, a large cyberinfrastructure project built around the services model and driven by hydrologic community requirements. Close alignment between patterns and needs of the hydrologic research community, and SOA principles, has been a strong argument in favor of using the latter as the basis of HIS design. The technical challenges discussed in the previous section can be addressed within SOA and related technologies, in particular with HIS moving to cloud and data grid platforms. Yet the critical issue of HIS development as an interoperable, sustainable, and extensible information system remains its adoption by the community and engagement of a wide audience of hydrologic researchers in governing service standards for data exchange and community semantics, and in development of hydrologic analysis and modeling applications compliant with such standards.

## Acknowledgements

CUAHSI and NSF support for the hydrologic information system development (awards EAR 0413182, 0413265, 0412904, 0622374) is gratefully acknowledged. In addition, we specifically thank the project development team and CUAHSI HIS partners at federal and state agencies and research projects.

## References

Ball, B., Burns, R., Cuker, B. *et al*. (2008). A Prototype System for Multi-Disciplinary Shared CyberInfrastructure: Chesapeake Bay Environmental Observatory (CBEO). *Journal of Hydrologic Engineering, ASCE*, **13**(10): 960–970.

Bandaragoda, C. J., Tarboton, D. G., and Maidment, D. R. (2005). User needs assessment (Chapter 4). In Hydrologic Information System Status Report, Version 1, ed. D. R. Maidment, pp. 48–87, www.cuahsi.org/docs/HISStatusSept15.pdf.

Committee on Opportunities in Hydrologic Science (1991). *Opportunities in the Hydrologic Sciences*. Washington, D.C.: National Academies Press.

EPA STORET (2010). STORET Legacy Data Center, Reference Tables, www.epa.gov/storet/legacy/ref_tables.htm, last accessed June 6, 2010.

Horsburgh, J. S., Tarboton, D. G., Maidment, D. R., and Zaslavsky, I. (2008). A relational model for environmental and water resources data. *Water Resources Research*, **44**, W05406, http://dx.doi.org/10.1029/2007WR006392.

Horsburgh, J. S., Tarboton, D. G., Piasecki, M. *et al*. (2009). An integrated system for publishing environmental observations data. *Environmental Modelling & Software*, **24**(8): 879–888, http://dx.doi.org/10.1016/j.envsoft.2009.01.002.

Horsburgh, J. S., Tarboton, D. G., Schreuders, K. A. T. *et al*. (2010). HydroServer: A platform for publishing space-time hydrologic datasets. In *Proceedings of the AWRA Spring Specialty Conference on GIS and Water Resources*, Orlando, FL, March 29–31.

Josuttis, N. M. (2007). *SOA in Practice: The Art of Distributed System Design*. Sebastopol, CA: O'Reilly Press, 324pp.

Maidment, D. R., ed. (2009a). CUAHSI Hydrologic Information System: 2009 Status Report, Consortium of Universities for the Advancement of Hydrologic Science, Inc., 79pp. http://his.cuahsi.org/documents/HISOverview_2009.pdf, last accessed June 6, 2010.

Maidment, D. R., Hooper, R. P., Tarboton, D. G., and Zaslavsky, I. (2009b). Accessing and sharing data using CUAHSI Water Data Services. In *Hydroinformatics in Hydrology, Hydrogeology and Water Resources*, ed. I. Cluckie, Y. Chen, V. Babovic *et al*. Proceedings of Symposium JS4 held in Hyderabad, India, September 2009. Publication 331. Hyderabad, India: IAHS, , pp. 213–223, http://iahs.info/redbooks/331.htm.

NSF Blue Ribbon Advisory Panel on Cyberinfrastructure (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure*. Washingron, D.C.: National Science Foundation.

OGC O&M (2007). Open Geospatial Consortium, Observations and Measurements: Part 1. Observation Schema; Part 2. Sampling Features. OGC documents 07-002r1 and 07-002r3, Wayland, MA, USA.

OGC SOS (2007). Open Geospatial Consortium, Sensor Observation Service. OGC Implementation Standard. Document OGC 06-009r6, www.opengeospatial.org/standards/sos.

OGC HDWG (2009). Open Geospatial Consortium – Hydrology Domain Working Group, http://external.opengis.org/twiki_public/bin/view/HydrologyDWG/.

Rodriguez, M., Valentine, D., Whitenack, T., and Zaslavsky, I. (2010). OLAP-based analysis and visualization of large volumes of hydrologic observation and model data. AWRA 2010 Spring Specialty Conference, Orlando, FL, USA, March 29–31, 2010, www.awra.org/orlando2010/.

USGS NWIS (2010). National Water Information System, Parameter Codes, http://nwis.waterdata.usgs.gov/usa/nwis/pmcodes, last accessed June 6, 2010.

WATERS Network (2008). WATERS Test Bed Sites, www.watersnet.org/wtbs/, last accessed June 3, 2010.

Whitenack, T., Valentine, D., Zaslavsky, I., and Djokic, D. (2007). Hydrologic information system server: The software stack and the initial deployment experience. Geoinformatics 2007 Conference, May 17–18, 2007, http://gsa.confex.com/gsa/2007GE/finalprogram/abstract_122332.htm, last accessed May 19, 2010.

Whitenack, T., Zaslavsky, I., and Valentine, D. (2008). HIS Central and the Hydrologic Metadata Catalog. *Eos Transactions AGU*, **89**(53), Fall Meet. Suppl., Abstract IN51A-1142.

Zaslavsky, I. (2006). Information management needs of hydrologic observatory test beds. Survey at the CUAHSI HIS – test bed projects joint meeting, Austin, TX, November 15–17. Retrieved July 1, 2009, from www.cuahsi.org/his/austin/Austin-survey-results-200611.pdf.
Zaslavsky, I., Valentine, D., and Whiteaker, T. (2007). CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0. www.opengeospatial.org/standards/dp.

# 13

# Web services for seismic data archives

ROBERT CASEY AND TIMOTHY K. AHERN

## 13.1 Introduction

Seismology has a long tradition of sharing data across boundaries, be they international, institutional, or scientific domain. This sharing began decades before the era of digital data but modern methods of digital access to seismological data have revolutionized the ways in which these data are now available. Normally, seismological information is accessible on a global basis within seconds of real time. Since scientific studies also demand access to data spanning decades, access to voluminous archives of time series data is also necessary. This paper attempts to highlight some of the more recent applications of web services to provide access to seismological data of several types.

There are thousands of seismic stations operating in the world today; most are used for monitoring seismicity on a local, national, regional, or global scale. Broadband seismometers are now more and more common and greatly enhance the scientific usefulness of these data. The IRIS Data Management Center (DMC) is a major center that incorporates data from several seismic networks, including those funded by the US National Science Foundation, US Geological Survey, and networks funded by other international and national sources. As an international data center, the IRIS DMC actively works with international groups in a manner where data from many networks around the world, funded by international sources, are available through the IRIS DMC, as well as through distributed data centers. Figure 13.1 shows the locations of the stations that have contributed data to the IRIS DMC.

While aspects of what we present here are specific applications applied at the IRIS Data Management Center, the general characteristics of web services for seismological data centers are universal.

## 13.2 A description of seismic data

There are four categories of seismic data for which web services have been developed. These include (1) the time series of ground motion, (2) metadata describing

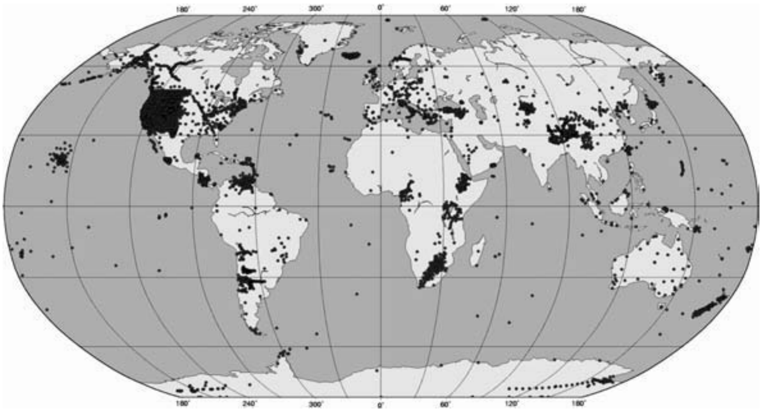



Figure 13.1. Stations providing data to the IRIS DMC. As of July 2010 the IRIS DMC has data from more than 14 264 seismic stations. These include stations that operate as part of 267 permanent networks, and stations that were part of 469 temporary deployments. The IRIS DMC has data from IRIS networks (GSN and PASSCAL), networks of the International Federation of Seismograph Networks (FDSN), the USGS Advanced National Seismic System (ANSS), and data from temporary deployments from the NSF-funded Ocean Bottom Seismographic Instrumentation Pool (OBSIP), the SEIS-UK project in the United Kingdom, the SEISMOB-FR project in France, and some deployments funded by Canada.

the time series data, (3) catalogs of earthquakes (locations, times, and magnitudes), and (4) seismological products.

### *13.2.1 Time series data*

Most modern ground motion data are recorded by sensors that are sensitive to changes in the velocity of the ground. Ground motions as small as one nanometer per second can be detected and these ground motions are measured along three axes to describe the three-dimensional motion of the ground at that sensor's location on the Earth. The ground motion is typically sampled at a variety of rates, from hundreds of samples per second to periods of 10 or 100 seconds per sample. The useful bandwidth of the recorded signals is from a few tens of hertz to several thousands of seconds, capturing the full spectrum of interesting seismic signals. Data are recorded in a self-describing format, usually in the FDSN SEED format (USGS, 2006) or in the CD1.x format used by the UN's Comprehensive Test Ban Treaty Organization (SAIC, 1997, 2001). Most modern recording instruments write the data in a compressed format to minimize the storage capacity needed to store the individual samples. Data are normally converted to other formats when detailed analysis of the seismograms is carried out. Time series data can be voluminous, and web service methods are sometimes not well suited for direct transmission of seismic data.

Figure 13.2 shows the amount of data that was archived at the IRIS Data Management Center through mid 2010. While seismological data have not yet reached petabyte scale in size, 120 terabytes are available, and the volume is growing at a rate of about 20 terabytes per year at the IRIS DMC.

### *13.2.2 Metadata describing the recording instruments*

The metadata describing the seismic instrumentation, sensors, and other information necessary to use the time series data are well documented. The metadata include descriptions of the sensor locations, manufacturer, and other descriptive information. The most complex part of the metadata is that describing the processes by which seismic time series of differing sample rates are derived from the seismic sensors. The metadata includes descriptions of various filter stages used to generate

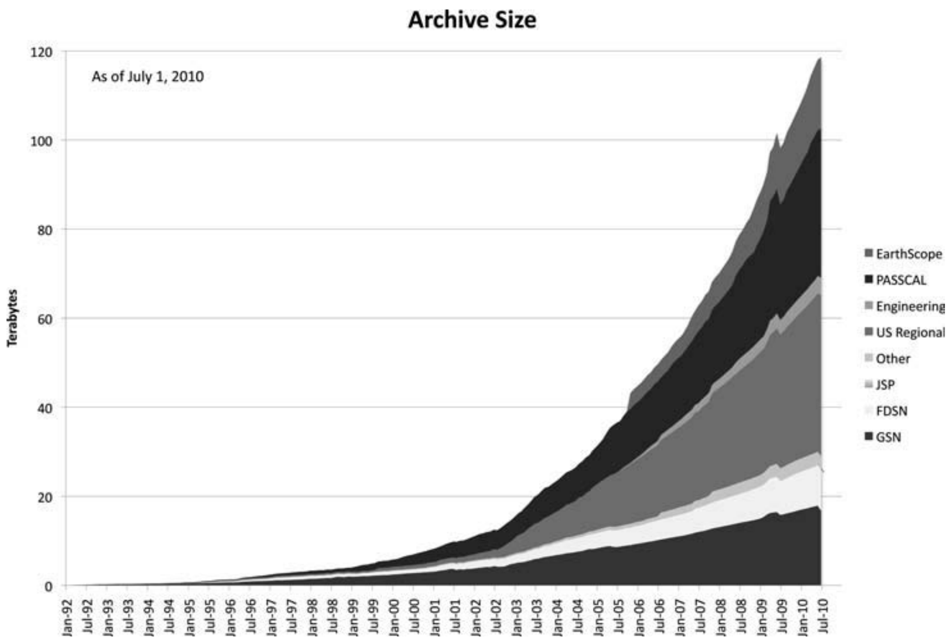



Figure 13.2. The growth of the IRIS DMC Archive. As of July 2010 the IRIS DMC manages approximately 120 terabytes of data. This figure shows the growth of the archive since 1992. Contributors from the bottom (red) to the top in the figure are (1) IRIS Global Seismic Network, (2) International Federation of Digital Seismographic Data, (3) Joint Seismic Program, (4) Other networks, (5) US regional Networks including the USGS ANSS network, (6) engineering data from structures, (7) data from the PASSCAL program of IRIS, and (8) data from the NSF EarthScope program that includes USArray and PBO seismic data. The archive is presently growing at about 20 terabytes per year. See color plates section.

the data streams, gains of the instruments, and the response of the seismometers to ground motion. The information is complete enough such that the actual ground motion can be extracted from the recorded signals and ground motion can be reproduced with high fidelity. The methodology is completely described in the FDSN SEED Reference Manual (USGS, 2006).

Metadata about the availability of specific time series is derived as the data are archived at the DMC. While it is the goal to have continuously recording channels, in fact, telemetry and instrumentation problems can prevent meeting this goal, so time markers are placed in the database to indicate where the data start and stop as a normal part of the data management process.

### *13.2.3 Earthquake location data*

One of the most fundamental products produced by seismologists are lists of earthquakes that are "located" using time series data from the globally distributed seismic stations. While it is beyond the scope of this article to describe the methodology of the earthquake location process, fairly complete "catalogs" of earthquakes exist as a result of institutions that dedicate their expertise to the task. These catalogs contain the time, location, and magnitude of detected seismic events. The National Earthquake Information Center (NEIC) of the U.S. Geological Survey (USGS) in the United States and the International Seismological Centre (ISC) in the United Kingdom are two centers that generate earthquake catalogs and make them widely available. Earthquake bulletins also provide additional information about specific events and normally include source mechanism information, as well as the arrival times of specific types of energy observed at seismic stations. Waveform data centers like IRIS DMC ingest these catalogs and bulletins and make them available through a variety of techniques, including web services.

### *13.2.4 Seismic data products*

In addition to time series data, metadata, and earthquake catalogs, value-added products are routinely generated by many seismological data centers. These products consist of such things as collections of waveforms containing energy from specific events, animations showing ground motion or particle motion of derived parameters, or static images of waveform record sections and amount of ground shaking. Product management systems now exist that manage these data entirely within a web services framework, facilitating computer to computer as well as human to computer interactions.

While seismic data are presented in a form that is usable to seismologists, the process required to start working with it on a user's workstation can be time consuming and overly technical. It is even more difficult for those outside the seismology domain to use the raw data for their studies. For this reason we see a series of services made available through a service-oriented architecture (SOA) as a promising method to allow a more approachable and flexible form of access to seismic data for all scientists and educators. For instance, services that convert the compressed time series data to measured ground motion and present it as a simple text listing for import to a spreadsheet serves as just one example of reaching a wider audience.

## 13.3 Traditional data access methods

Access to digital seismic data via the Internet preceded the age of the World Wide Web. The processes that were in place in the early days of networked data center access involved users asking for data they thought they needed, without being certain that the data had been recorded. Users had little or no interactivity to supplement the transaction.

In this early scenario, a request form would be submitted electronically (typically via email) and would be queued into a batch-oriented processing system for data retrieval. If the data a user wanted existed at the data center, they would get the data they asked for in the form of a latent, asynchronous response from the data center. If there were entries in the request where no data were present, then there would be no data to return. In any case, the user could only know of the data's existence (or lack thereof) in hindsight.

This kind of *proactive* approach is still in heavy use today due to its simplicity to implement, deploy, and use. The downside is that users cannot gather information on what they can ask for and just what a data center offers that they might not be aware of. This is akin to mailing an order for a book without benefit of a catalog to confirm the title or its availability.

Instead, a more *interactive* process is needed to discover, first, what is available, and secondly, how to get it. This process is known as a *data discovery* search.

## 13.4 Discovery of seismic data

When a scientist is interested in seismic waveforms, they invariably focus their search on a set of sensing stations that recorded the ground motion of interest. The *station identifier* names must be found, collated, and referenced by the participating data center to facilitate seismic data discovery. The question then becomes one of determining the techniques for selecting this set of stations.

There are three primary components to seismic data discovery: station, recording channel, and time. In finding the station or set of stations of interest, we satisfy the *geographic* dimension of our data search in that we identify the point locale and the instrument that recorded the seismic waveform. A scientist typically selects stations in one of three ways:

(1) Specific stations that are located in an geographic selection space;
(2) Stations that are owned and operated by a particular institution, or are members of a particular organization or initiative;
(3) Stations that are within a certain travel-time distance and/or direction from one or more earthquake locations.

Identifying the stations of interest is key to gaining access to the seismometer recordings. It is not enough to simply ask for data from a point on the globe. Once an instrument has been referenced, there are two other properties that must be explored and specified.

One of these properties is *time*. Stations do not remain in the field indefinitely, and sometimes are either not operational or do not exist during the time-period of interest. For this reason, users specify a *window* of time that they are interested in gathering data within, normally times associated with earthquakes.

This time selection not only addresses stations that are operational, but is also used to query for the availability of data at the data center. The condition of seismic data being available changes over time and cannot be guaranteed at the time of data discovery.

There are two primary techniques used for specifying a time window in the search for data:

(1) Specifying a single arbitrary window of time, consisting of a start time and an end time, which applies to all sensing stations;
(2) Specifying a start time and an end time that pertain to earthquake signal phase arrivals at each station. Earthquake phases are associated with different energy propagation paths through or around the surface of the Earth.

The first case of time selection is used for gathering continuous data, irrespective of one or more seismic anomalies. The span of time is on the order of minutes, hours, or days. This is the simplest form of time specification since only a single start and end time set need be specified.

The second, earthquake-oriented approach focuses on a specific seismic event and computes the data windows at *each* individual station to provide a snapshot of the earthquake signal at the moment the earthquake waves arrive at the station. This is the more complicated of the two processes because it requires additional information to be provided beforehand:

(1) The time and location of the earthquake in question;
(2) A travel-time model that will estimate the timing of phase arrivals based on a station's distance from the event.

Using an appropriate algorithm that accepts the above two pieces of information, an iteration over a set of stations results in time selections that are later at stations that are more distant from the event. When viewing the waveforms displayed with time on the horizontal axis and distance between station and event on the vertical axis, in what is known as a *record-section plot*, the scientist will see the energy arriving for several phases at each station and these phases arrive later on seismograms from more distant stations (Figure 13.3).

Whereas global location and time are essential aspects to discovery of seismic data, a less essential but important filter for selecting data is the sample rate. Today's broadband seismometers can record anything from one sample every 100 seconds to tens of samples every second. Many stations offer a selection of data rates, but a user could still be seeking one that a station cannot supply.

Each of these frequencies has a special use and purpose for the scientist. Therefore, most data-discovery queries will involve interest in a particular data rate. This, coupled with the bi-directional orientation of the sensor (North/South, East/West, Up/Down), forms the basis of an identifying tag commonly referred to as a *channel*. While station identifiers are unique to a location on the globe, channel

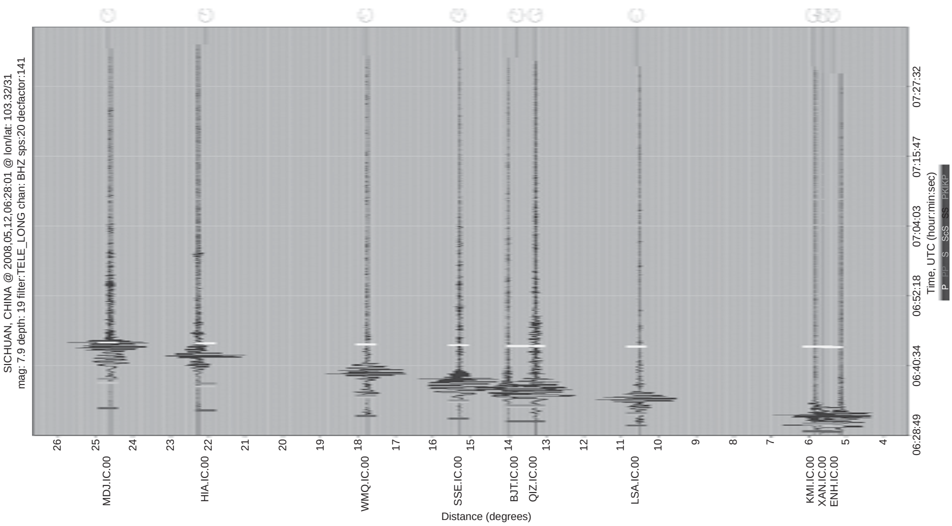



Figure 13.3. Seismic record section plot. This figure shows energy arriving at ten stations in China from the Sichuan earthquake of May 12, 2008. The various phases are marked with colored vertical bars. The earliest arriving phase (P) is marked in green, with later arriving phases in different colors. The secondary phase (S) is shown in orange and clearly shows the increased delay for signals arriving at more distant stations, even though these stations are fairly close (less than 25°) from the earthquake source. See color plates section.

identifiers are *universal* among sensing stations, the naming scheme being specific to an agreed-upon convention. For this reason, it is possible to search across an entire set of stations, seeking a specific channel, in order to ensure that the desired seismic waveforms are selected for retrieval.

For traditional access mechanisms, a user must know in advance which stations they are interested in, which channels they want, and the time-period of the recording before submitting a request for batch processing.

More modern systems provide interactive facilities that allow a user to explore, select, and gather their query parameters before they submit a request. A query and feedback interface lets users form their data queries in a true discovery-oriented fashion, not committing them to any data gathering and delivery until they have put together a request that meets their needs. An example of such a client is the jWeed application available from IRIS (www.iris.edu/dhi/jweed_ request.htm).

While the latter procedure sounds superior to the a-priori approach of batch requests, there is something to be said about offering both interactive and proactive approaches when it comes to *programmatic web services*.

## 13.5 Data discovery with web services

Web technologies, high-speed networks, and advanced marshalling frameworks have come a long way in enabling the scientific data center to go from a virtual drop box to a medium of negotiation and quality of service. In changing the way we think of what a data center provides, designers of new data-discovery systems are frequently tempted to create tools that act strictly in a choreographed chain of events, requiring clients to *drill down* through many discovery layers, regardless of what they may already know up front.

Part of this tendency stems from the nature of seismic data in that it is identified and organized in a hierarchical fashion, starting at the station site identifier all the way down to the specific clock time of seismic samples. This natural arrangement lends itself to access interfaces that pull users through a *tree-oriented* access pattern, which can result in multiple query and response steps to arrive at a desired dataset.

While drill-down procedures are in many cases needed in order to *find* data, interactive negotiation can exhibit the negative characteristics of increased network traffic, extensive state persistence, and a tendency to cause expensive client-server iteration activities. The end result is a slow, cumbersome means of access that familiar clients quickly abandon for more direct, traditional methods to get the data they need.

There is a tradeoff to consider in supplying the naive user (or automated system) with the answers they need to find seismic data versus giving more proactive and knowledgeable clients access to the data without multiple steps. You do not want to alienate either audience, since both are highly valued by the scientific data center. It becomes clear that accommodation should be provided for both camps in a web service architecture.

A flexible data-discovery system should have interfaces exposed all along the data-discovery chain, each of which can help to narrow the overall search. This is easily realized by creating a public toolkit of *decoupled* service interfaces that can allow data discovery with as much or as little interaction as needed to arrive at the final parameter set for discovering and requesting data. In addition, the separation of such services into independent modules results in a more robust, easily tested, and maintained data-service system.

## 13.6 Data-service components

In creating a componentized system for web services, we need to perform an analysis of how users interact with a seismic data center, both in the information provided and the patterns of exploration that they tend to follow.

We already know that there are three principal query filters used for data search (location, channel, time), so this sets the basis for many of the component queries that will be passed to data service modules. In fact, since our principal goal is to point the user to seismic data, we should probably have the first module we create be one that accepts these three basic parameters and returns a listing of available waveforms. These results are typically *chunked* into station, channel, and time range elements, the reference of which will extract and deliver the seismic data in some well-defined format.

What if the user does not know what stations are available in a certain part of the globe, but they know the time range and channel of interest? This means they know two out of three parameters of interest, but require some level of interaction to get the station list. Our next web service module could accept a latitude and longitude range and return a list of station identifiers. The user could review and select from this list, providing the necessary set of parameters to feed our first example module, which in turn finds us our data.

This simple example is just the tip of the iceberg in terms of branching features away from the main query track in order to allow the user to find what they are looking for. Instead of forcing them on a long critical path to the data they want, a set of detours and tangents should be offered when users do not quite know what question to ask. In the end, however, they still come back to the main topic at hand: locating the seismic data that suit their interest.

## 13.7 Delivery of seismic data

Data recorded by seismic instruments are currently a regular time series of signed binary values, normally compressed, always annotated, and self-identifying. The formatting of these data can vary widely depending on the supporting organization and its intended use by the audience. Sometimes the format is a simple columnar representation of ASCII characters, but oftentimes it is a tightly formatted self-describing package of metadata descriptions, file markers, and author annotations to provide maximal scientific value to the data recordings.

The traditional means of delivering seismic data to the user is via file transport protocols such as FTP and http. In this case, there are no supporting annotations other than the file name and the contents of the file itself. This means of access is primitive but fairly quick with current network bandwidth.

While file transfer has been the predominant means of transferring seismic data to others, options that are more flexible have been desired by the seismic community to allow it to overcome the limitations of forcing the user to contend with large data packages in an offline capacity once interaction with the data center has long finished.

At times, users simply want to get a quick view of the data. Server side processing can create images for presentation, but what if the user wants to interact with that image? It is much better if the user can receive the data in a form that allows them to view, explore, measure, and decide if they want some form of permanent copy. This leads us into a new kind of seismic data delivery that is more transparent, accessible, and easily processed by programs.

The issue, then, becomes one of transport and representation. The above-mentioned file transport protocols might suffice, but they generally lack an easy-to-parse representation that can then be unbundled, processed, and displayed in client code. The files themselves tend to be tightly bundled and lacking in helpful annotations.

A means of delivering data that is more programmatically accessible is to treat the data as objects and serve them up using an *object request broker* (ORB), a software service that delivers copies of objects created on the server and mirroring them on the client side. This mirroring operation of generating objects from a network transmission is called *unmarshalling*. Objects, which are a data structure that combines data with operations to act on those data, provide a much more flexible and portable data representation to the traditional separation of code and data. Mirrored on the client, the objects are accessed just as if they were created from local files and can provide annotated information as well as operations to act directly on the data. All the while, the client program is really interacting with a server at great distance through mediation protocols. The best known standard for this process is CORBA.

Seismic data can and have been delivered in this object-oriented fashion for users, but there are some known limitations to the CORBA approach. First is the barrier to entry, where developers must leverage vendor-specific code on both sides of the network and follow a prescribed set of communication steps before data transfer can take place (Chappell, 1998). Second, network communication processes occur in an opaque fashion through non-ubiquitous port connections, which both limits the exposure and means of carrying out such transactions and makes troubleshooting issues difficult (Henning, 2006). Finally, the ORB model has a heavy dependence on use of an object-oriented language to carry out client–server communication, which rules out applications written in procedural languages and scripts (Kaye, 2003).

In the mid nineties, we saw the rise of the Web as a ubiquitous interfacing tool for all kinds of data access. It is a tool that everyone in the general public has become comfortable with. On its heels was the creation of the XML markup language, which could be transmitted across web space between computers and processed by programs. This was then followed by protocols such as SOAP that formalized the process of XML data exchange and opened the door to extensible data descriptions and registered services.

Because communication is initiated over the network port connection used by web browsers around the world, web services offers a universally understood point of contact for outside users to gain access. In addition, web services, just like web pages, refer to software and data via Universal Resource Locators, or URLs, which provide an easy-to-reference hierarchical address scheme to access services and data.

While it is easy to format text information in XML and send it to a scientist, seismic data present more of a challenge. Seismic recordings are generally stored in raw binary, not text, format, which necessitates some form of envelope or wrapper for such data using well-defined encoding schemes. In addition, seismic data tend to be very large in size, which can result in slow transmission and translation of the data when encoded in an XML document. It is simply not the most efficient means of transfer.

For this reason, there is no one particular way of sending seismic data that are preferred overall. For smaller datasets, the data can be converted directly to text value representations and sent in the XML document. Another method is to encode the binary values as an *attachment*, using methods such as MTOM and MIME to embed the data inside of an XML message.

When it comes to large datasets, a preferred approach is the use of an outside channel of data delivery, where an XML document provides a separate reference to the data file or raw binary stream, which is then accessed using traditional file transfer methods. This presents a compromise where the data are only partly represented in XML annotation, leaving the majority of the specific information for further processing as an auxiliary step.

## 13.8 Issues with service-oriented data access

When first setting out to create services for seismic data, it becomes abundantly clear how much depth there is to cover in a dataset, as well as the various angles of access that one can take. When a user sets out to explore a well-populated seismic data center's holdings, they can be faced with referencing thousands of data references on the first query pass. This not only makes browsing the data holdings a non-trivial task for the user, but it presents great challenges to those that design the data services to begin with.

As an example, the EarthScope Transportable Array program currently has more than 400 operational broadband stations in the continental United States; each of these stations has approximately 6 channels of constantly streaming data, which already presents more than 2400 separate data streams to choose from. Couple this with daily packaging of data over the span of even five years and you now present, in the simplest case, 876 thousand separate waveform products for the user to choose from.

Simple browsing capabilities simply will not suffice when dealing with this scope, so our services have to be ready to accept search parameters of various types to narrow down the list. We discussed these search patterns in an earlier section, but an effective service architecture will implement these filter views in whole or in part in order to bring to the user a small, manageable list of data to work with. The filters not only have to be present on the server side, but the client programs that implement them have to be effective and easy to use. It is very tempting to create user interfaces that multiply a simple click-and-select model over an entire return set. The end result is an interface that is simply not usable when you speak of data returns in the thousands.

What can work effectively in this realm is to let multiplication work for access as well. Data returns should not only list the name of a data product, but it needs to *categorize* itself as a member of larger group entities. For instance, a channel's seismic recording can be grouped with other channels in the same station, or they can be grouped with seismic recordings from all stations that have the same channel name. It could also be grouped by instrument type, what region of the globe it is located, or the organization that maintains it. Such groupings are generally represented as a tree structure, and allowing users to access data by higher nodes in the tree greatly simplifies the user experience for grabbing large sets of related data at once.

It is also necessary to remember that not all forms of access to seismic web services is via a graphical user interface or web browser. Automated programs are emerging to take advantage of these open standards to provide seamless access to data. This means rapid, repeated access to services that can easily overwhelm software and hardware if its design does not take high utilization into account. Services must be quick, accommodating, and robust. It must also have interfaces that are easy to tap into via direct, automated means.

## 13.9 Strategies for advanced seismic services

The opportunities for data discovery and access using web services only scratches the surface for what is truly possible with an effective, service-oriented architecture.

The real promise comes from providing computational and transformational services that act on behalf of the user to produce derivative results and datasets without the user needing to download a single iota of raw data.

Computational services can take data from local and remote resources to produce ground motion data representations, create waveforms with noise filtered out, and provide cross-sectional wave slowness images all based on user input, taking advantage of the vast resources of processing speed and storage capacity that can already be found at many data centers, but are perhaps not being utilized to their furthest potential and are out of reach of so many researchers.

Having a well-defined set of service components that are exposed, ready for use, and easily found on a registry, will provide new possibilities in scientific computation and data representation.

Web services promise to be the gateway to shared data and shared processing and what is required to make this truly effective in the community as a whole is to get the data and processing centers to work together and share their resources on the global scale.

## 13.10 Integration into a geoscience cyberinfrastructure

As the complexity of scientific problems increases, the need for a scientific study to integrate data across multiple scientific domains has also increased. The earth sciences typically have data that have a geospatial aspect and this has been leveraged to assist in the data-discovery part of the challenge. GeoWS was a collaboration between IRIS, UNAVCO, and the Marine Geophysical Data System of Columbia University to enable discovery of geophysical datasets across these three centers. This effort leveraged web-mapping standards from the Open Geospatial Consortium (www.opengeospatial.org/). Additionally certain amounts of metadata about the datasets could be exposed. This however does not address issues such as unique data formats, calibrations, and a variety of things that remain esoteric within a domain.

Another challenge in data integration lies with semantic interoperability. One scientific domain may use the same words as another but the underlying meaning can be very different. The goal is to develop systems that will allow semantic interoperability across domains using ontologies, but at this time this only remains a goal. The potential of web services using SOAP and WSDL appear to be more viable. While it will not completely divorce the researcher from understanding the vocabulary used within a discipline, the use of Web Services Definition Language (WSDL) also appears to be a viable approach that will help in the data integration problem. More recently RESTful web services and their corresponding service definitions in WADL (Web Application Description Language) are becoming more and more widely used and appear to be viable.

Most current developments in web services show promise in providing well-documented services that can be invoked by a variety of clients. At the present time, viable systems remain those that require a certain amount of domain knowledge in order to perform the required integration into a broader geosciences cyberinfrastructure.

## 13.11 Closing thoughts

Mechanisms to deliver seismic data to the scientific community are mature. However, we believe that the need to broaden the community that can make use of seismic data warrants an expansion into delivery of our data via web services. This will ease the integration of seismic data and products into other scientific studies and improve the repurposing of seismic data. Conversely, by moving seismology into web services we believe that it will ease access to other types of data for seismologists, supporting more integrative studies. To reach the ultimate goal of total interoperability will take many years, but providing access to the most fundamental data, metadata, and products in our domain can be done using existing technologies found in web services.

## References


Chappell, D. (1998). The trouble with CORBA.. David Chappell & Associates, www.davidchappell.com/articles/article_Trouble_CORBA.html. accessed August 2008.

Henning, M. (2006). The rise and fall of CORBA. *ACM Queue*, **4**(5).

Kaye, D. (2003). *Loosely Coupled: The Missing Pieces of Web Services*. Marin County, CA: RDS Press, p. 88.

SAIC (1997). CD1.0 Manual: Science Application International Corporation, Formats and Protocols for Continuous Data, SAIC-97/3022.

SAIC (2001). CD1.1 Manual: Science Application International Corporation, Formats and Protocols for Continuous Data CD-1.1, Revision 0.2 SAIC-01/3027Rev02.

USGS (2006). Reference Manual: International Federation of Digital Seismograph Networks: Incorporated Research Institutions for Seismology: United States Geological Survey. SEED Format, Version 2.4, March 2006.

# 14

# Development of robust data system for gravity and magnetic anomaly data: A case study of a community-based effort for point data

RAED ALDOURI AND G. RANDY KELLER

## 14.1 Introduction

Key elements in various descriptions of geoinformatics as a field of research development (e.g., Sinha, 2006; www.geoinformatics.info/; Keller, this volume) include open access to data as well as a software toolbox for processing and analyzing the data. However, if a particular type of data is to be useful to a broader community for integrated analysis, raw data must be organized, edited, and formatted into a database that meets a minimal set of community standards for content, processing, and metadata (e.g., FGDC, 1998). In order to make the database useful to non-experts, it should be the foundation of a data system that includes features such as tutorials on how to use the data and software tools for data processing, analysis, and mapping.

Measurements of Earth's gravity and magnetic fields are examples of point measurements that are relatively straightforward to make and qualitatively and quantitatively interpret, and considerable amounts of these data reside in the public domain. These data are useful in a wide variety of applications ranging from groundwater, environmental and natural hazard investigations, to exploration for natural resources, and to studies of deep earth structure. Thus, they are of broad utility to the geoscience community. This chapter presents a case study of the development of a data system that is based on new databases for both gravity and magnetic field point measurements that are downloadable, and contains tutorials and related software tools that have been developed to support data analysis and integrated geologic/geophysical investigations.

The initial database development effort was focused on gravity data for the lower 48 states of the USA, but the goal was to expand to related datasets and to all of North America and considerable progress has been made in this regard. In 2002, a new magnetic database for North America was released as a grid by Canadian, U. S. and Mexican national geoscience organizations (http://crustal.usgs.gov/projects/namad/). It is easily downloaded, but it is not searchable. Thus for use in our data system, the data values in this grid were converted to point values (latitude, longitude, and value) so they could be obtained only for the area of interest to the user. In addition, point measurements of gravity anomalies are now publicly available for Canada and a matching effort is underway in Mexico. The first version of the gravity database for the lower 48 states was posted online in 2003 (Hildenbrand *et al*., 2003) and involved cooperation between several large government agencies (US Geological Survey, National Geospatial-Intelligence Agency, National Geodetic Survey, and National Geophysical Data Center), academic groups, and industry. This cooperation has been maintained, but keeping the team together does take a steady concerted effort.

As described by Keller *et al*. (2006), the philosophy behind the construction of the gravity and magnetic data system was to build a community consensus via a series of steps:

- Seek input from the international scientific community
- Hold a series of workshops to seek the input and participation of interested parties
- Publish widely disseminated reports that present the results of these workshops (e.g., Hildenbrand *et al*., 2002, 2003; Keller *et al*., 2002)
- Devise new standards with community input, and make professional presentations and publish peer reviewed articles about them (e.g., Hinze *et al*., 2005)
- Provide open access to the resulting database, initially for the lower 48 states via a web portal (http://research.utep.edu/paces). This approach has now also been adopted by Canada (http://gsc.nrcan.gc.ca/gravity/index_e.php)
- Develop tools and tutorials to increase the utility of the data to a broad community from specialists to high-school students and teachers. See both URLs above
- Continue to update and improve the databases and other elements of the data system

The data system concept is intended to be a dynamic construct, which is accessed through a web portal that is designed to provide opportunities for the scientific community to provide input, tools, and contribute new data. Thus, it is designed to evolve as new data, procedures, and tools become available and are contributed by the community. In the case of this data system, a spreadsheet for the processing of gravity data collected in the field was designed by community members outside the organizers of the data system, carefully checked by the peer review system, and published in a major journal (Holom and Oldow, 2007). In addition, new information requested by users (e.g., point values in new coordinate systems) has been added to the data structure. As of the writing of this paper, the new database has 1 238 881 points (Figure 14.1) after removing duplicate and bad points, and provides geographic coordinates of data points referenced to multiple geodetic data.

One goal of this effort is that users will start a study using this database, which can be thought of as legacy data, but will often find that they need to collect additional

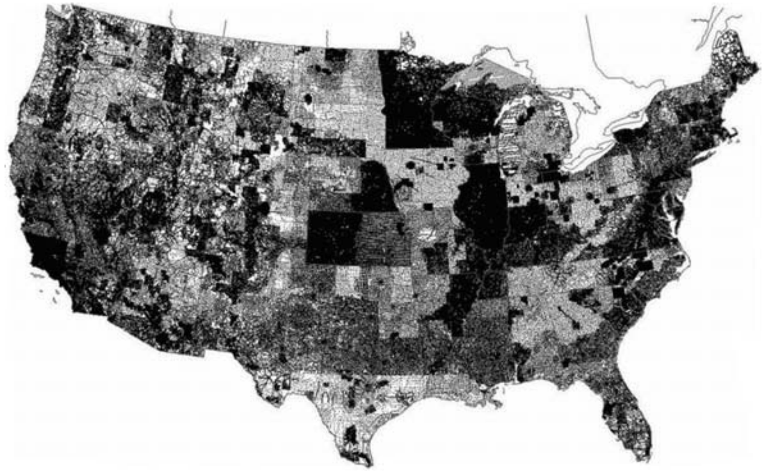


Figure 14.1. Map of the distribution of data points in the gravity database.

data in their study area and eventually contribute them to the database. Data contributions are presently handled on a case-by-case basis, but this process should be automated. Any new gravity data-collection effort requires a tie to an established base station where a highly accurate value of the Earth's gravity field is available. These base stations can be thought of as being like benchmarks for traditional land surveys. Therefore, our community felt that it was imperative for users to have access to gravity base-station information for their area of research, and the National Geodetic Survey agreed to provide this information. In addition, we maintain a listing of National Geospatial-Intelligence Agency gravity base stations. However, the user must contact the NGA directly for this information.

## 14.2 Technical development of the data system

The following discussion describes the various steps in the process of developing the existing data system.

### *14.2.1 Database*

An important goal in geoinformatics is making datasets that are important to the scientific community easily available for research and teaching. Keller *et al*. (2006) described the creation of the gravity and magnetic database being discussed here in considerable detail. However, the steps listed above concerning community involvement and standards were followed. The most important steps were getting all the major stakeholders who possessed major datasets committed to a consensus about the content and standards for the database. The various datasets were combined so as to retain unique source identification numbers that are assigned by the National Geospatial-Intelligence Agency. As expected, this process produced duplicate points because of undocumented coordinate system issues that could make the same data point appear to have horizontal positions that differed by tens of meters. Obvious duplicate points were deleted from the online version of the database using the technique of Torres *et al*. (2004) that was developed for this purpose. The first version of the database was also cleaned to delete points that had obviously erroneous gravity anomaly values. In both cases, no data were actually deleted from the master database in case a user wanted to examine these data and possibly correct them. The latitude and longitude coordinates for the data point locations were based on the North American Datum 1927 (NAD27) for horizontal location and NGVD29 for vertical datum in the initial version of the database.

In a second version of the database, a major advancement was that the dataset was terrain corrected (e.g., Hammer, 1939; Plouff, 1977) by the US Geological Survey team members. Advance in geodesy based largely on the availability of GPS (global positioning system) made it advisable that users employ the NAD83 (WGS84) and NGVD88 horizontal and vertical datums, respectively. Therefore, we converted the horizontal and vertical positions in the original database to these reference systems and added columns for these coordinates to the data structure in the current (3rd) version of the database. Additional data have been identified, and their addition to the database is pending as of July 2010.

The database was initially housed in an Oracle relational database management system with carefully crafted tables, including every possible attribute required to maintain information necessary to have an accurate and trusted database. For example, the instrumentation tables have the necessary information about the instrument used in the field to collect the data, model of the instrument, the calibration constants related to that instrument, the accuracy of the instrument, and how location data were collected. The gravity table includes the principal facts that are latitude, longitude, elevation, and the observed gravity value required to produce the traditional free air and Bouguer gravity anomaly values at each station.

Early in the process, the database was linked to an online user interface to empower users with the ability to search and retrieve the data from their specific geographic regions of interest. The gravity data file from a search is provided in a simple text format that is saved to the user's local desktop for further processing. However, advances in both technology and geospatial sciences are steering data presentation and publishing toward web services that are more robust and easier for communication. The US gravity database and North American magnetic database were moved to a PostgriSQL database since this is an open source tool (www.postgresql.org). A simple but elegant web service interface was created to make it easier for many users to access the data directly (http://irpsrvgis00.utep.edu/repositorywebsite) or via the data system web site (http://research.utep.edu/paces). The user interactively specifies a rectangular region covering their area of interest, and a selection button for either gravity or magnetic data uses a different coding within the web service to extract the records requested as specified by the search parameters. A preview of the search results displays ten records and the total number of points found within the search region. This is an intermediate step where the user could discard the search and start a new search, or click the download button to obtain a text file that can be saved on the user's local drive for additional processing and analysis.

### *14.2.2 Data reduction standards and tools*

The magnetic database entries are simply coordinates (latitude and longitude) and a value for the magnetic field at that position so no further processing is required to make a traditional magnetic anomaly map. In the case of gravity data, elevation is very important, and its effects are traditionally calculated in what are called free air and Bouguer corrections (e.g., Blakely, 1995). However, other second-order corrections are desirable, so the database team decided on an advanced set of corrections and reported these values to the community via publication in a major journal (Hinze *et al*., 2005). Members of the user community coordinated with the data system team to devise a spreadsheet that could apply these corrections, and a peer-reviewed article has been published about this spreadsheet that is available online from the publisher (Holom and Oldow, 2007). Terrain corrections are an important aspect of the database, but they cannot be done online due to the complexity of this processing at the level of the community standards adopted.

### *14.2.3 Data processing and analysis tools*

Once the gravity data have been reduced, gravity and corresponding magnetic anomaly maps can be created using many geospatial analysis tools. Generic Mapping Tools (Wessel and Smith, 1995; http://gmt.soest.hawaii.edu) is a versatile open-source package that is used by many geoscientists. Many algorithms are available to grid point data, but for gravity and magnetic data, most users prefer the minimum curvature gridding approach (e.g., Swain, 1976) because these potential fields are relatively smoothly varying and this approach avoids the introduction of short-wavelength gridding artifacts. The gridded values can be visualized as a traditional contour map. However, most users employ color, with red representing high anomaly values and blue representing low values (Figure 14.2). Geophysicists have developed many filters over the years that can be applied to enhance features that are of specific interest to the user (e.g., Blakely, 1995). Operations such as the calculation of

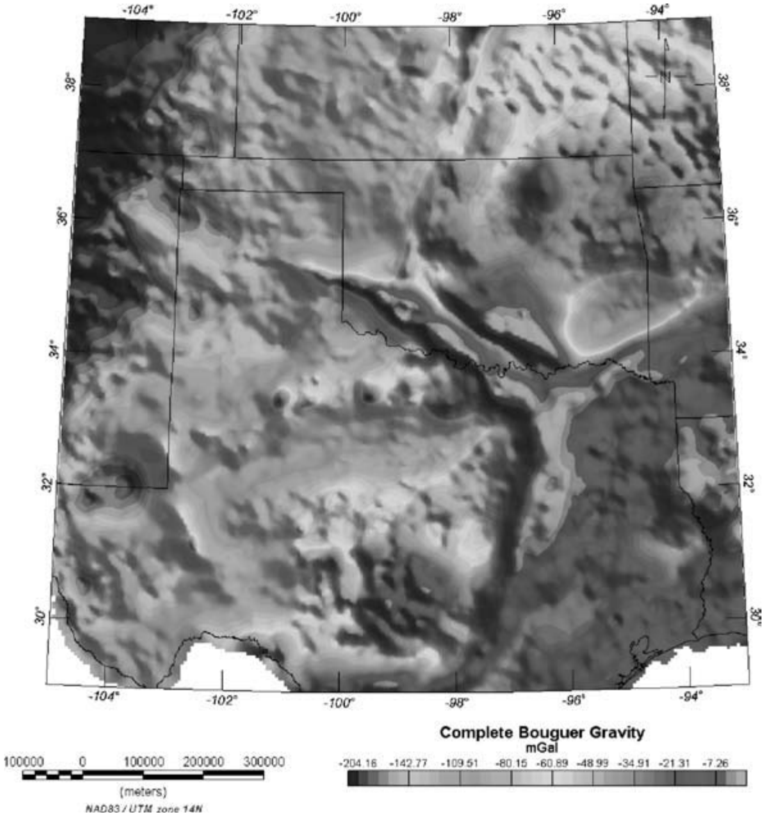



Figure 14.2. Complete Bouguer anomaly map of a portion of south-central USA. See color plates section.

horizontal gradients in the data help locate the edges of anomalous bodies. An open-source package of many filtering and other useful data analysis routines is available for downloading from the U.S. Geological Survey (Jeffrey Phillips, http://www.pubs.usgs.gov/fs/fs-0076-95/FS076-95.html).

The anomalies on gravity and magnetic maps reveal useful information about the subsurface. Thus, it is a common practice to construct models of the earth structure and composition in 2-D (distance along a profile vs. depth) or in 3-D by employing forward (trial-and-error iteration) or inverse (automatic iteration) methods in which the subsurface geometry and the density and/or magnetic susceptibility are varied. Modeling in 3-D remains a challenge in many respects even though the mathematical basis for the calculations is well established (e.g., Bott, 1960; Martins *et al.*, 2010).

Modeling in 2-D is very straightforward and is well suited for integrated analysis (e.g., Mickus and Keller, 1992). Thus, we developed a 2-D forward modeling software package for gravity profiles called *Talwani* (Cady, 1980; Talwani *et al.*, 1959) that can be downloaded as part of our data system. This modeling approach

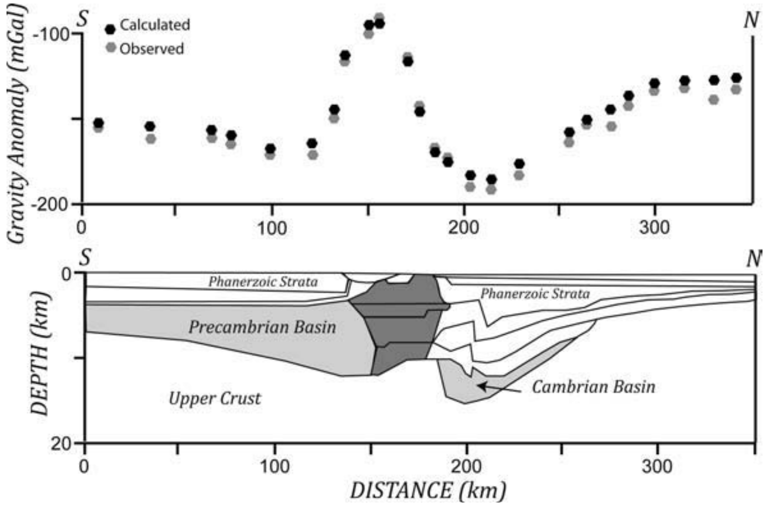



Figure 14.3. Schematic example of a 2-D gravity model that represents the output of the *Talwani* modeling software. The light gray bodies are not known from geological data and must have relatively low densities in order to match the observed data. In addition, the dark gray body must have a relatively high density to match the observed data. Matching the observed and calculated data involves modifying the geometries and/or densities of the polygonal bodies that constitute the model.

employs interlocking polygons to build the model (Figure 14.3). In order to use this program, GIS techniques should be used to extract gravity data points near a line between two reference points. All the data points that are within a specified buffer zone defined by the user should be selected. The output of this search should include the total distance (offset) from the starting point to each data point chosen. The output file should contain the offset and gravity anomaly value at the selected points. This file is required as input to the forward modeling program.

As with any geophysical modeling procedure, a starting model is used based on some knowledge of the earth structure present. The Talwani software is interactive and is designed so the user can select a vertex of a polygonal body in the model and move it to change the geometry of the model using a mouse. This action automatically changes the calculated gravity values from these bodies as the user manipulates the model to achieve a match between the observed and calculated gravity values (Figure 14.3) while honoring other types of data as constraints. Similarly, density values within the polygons that constitute the model can be changed to achieve a data match.

### *14.2.4 Workflows/ontologies*

The earth sciences community has begun the complex task of creating ontologies through creating detailed workflows. Capturing knowledge in the earth sciences

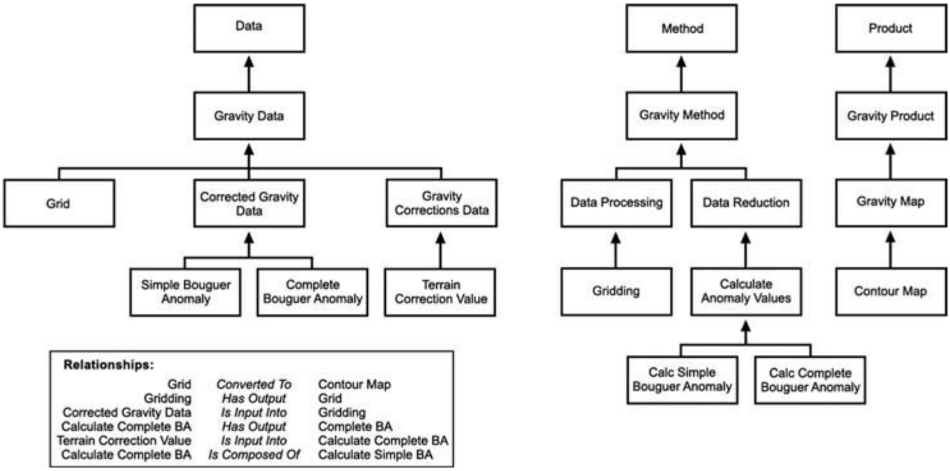



Figure 14.4. A diagram of a portion of the gravity ontology of Salayandia *et al.* (2006) and Gates *et al.* (2007). This diagram illustrates steps and relationships that are involved in the making of a gravity map such as in Figure 14.2.

through ontologies is intended to support technology that will facilitate computation, for example, by helping the composition of services and discovery of related data and concepts (e.g., Babai, this volume). Salayandia *et al.* (2006) and Gates *et al.* (2007) presented a workflow driven ontology for gravity data. Their approach specified multiple relationships between classes across three hierarchies (data, methods, products), and an example from their ontology is shown in Figure 14.4. When implemented, this particular ontology would drive the automation of the composition of services needed to create a gravity contour map.

## 14.3 Discussion and an example

This data system has became an invaluable tool for many who did not have access to such data either at all or only in their local area. It is being utilized by researchers and students around the world, as well as by government agencies and industry. We believe that this system can serve as an example for many types of point data, especially those that lend themselves to being analyzed via construction of contour maps. A nice example of another type of point data that is of broad interest to geoscientists is the World Stress Map project (www.world-stress-map.org/). In this case, vectors that show the principal stress directions can be generated online.

The simple workflow shown in Figure 14.5 illustrates the process that was used to provide an example of the utility of our data system by producing the sequence of maps of the south-central portion of the USA shown as Figures 14.2, 14.6, and 14.7. In this case, the first value chosen to be contoured (Figure 14.2) was the Complete

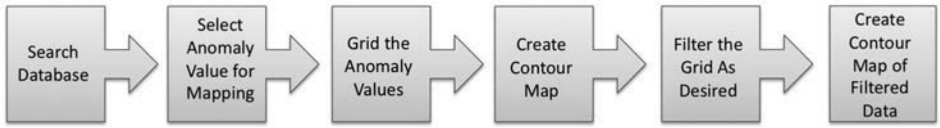



Figure 14.5. Simple workflow diagram that illustrates the steps needed to produce the maps shown in Figures 14.2, 14.6, 14.7, and 14.8 using the gravity and magnetic data system.

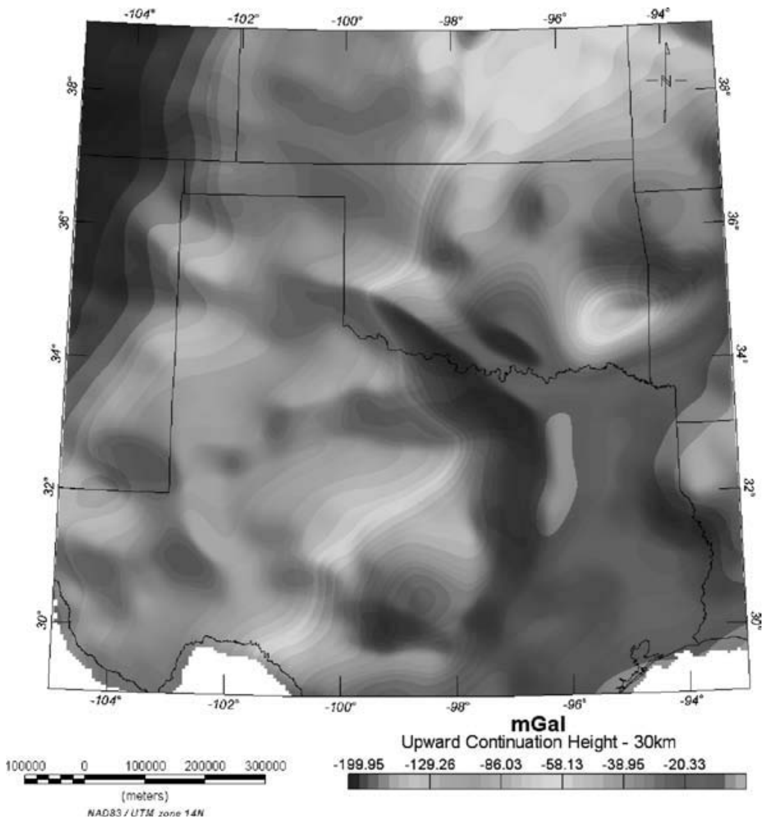



Figure 14.6. An upward continued version of the Complete Bouguer anomaly map shown in Figure 14.2. This physics-based procedure simulates the gravity anomaly map that would result if the readings had all been made 30 km above the Earth's surface. As is evident, this procedure enhances the long-wavelength features in the data. See color plates section.

Bouguer Anomaly, which is the initial type of gravity anomaly used in most geophysical studies (e.g., Blakely, 1995; Hinze *et al*., 2005). In this map, there is an obvious long-wavelength component in the data that causes an increase in anomaly values to the east and obscures the shorter wavelength features that are

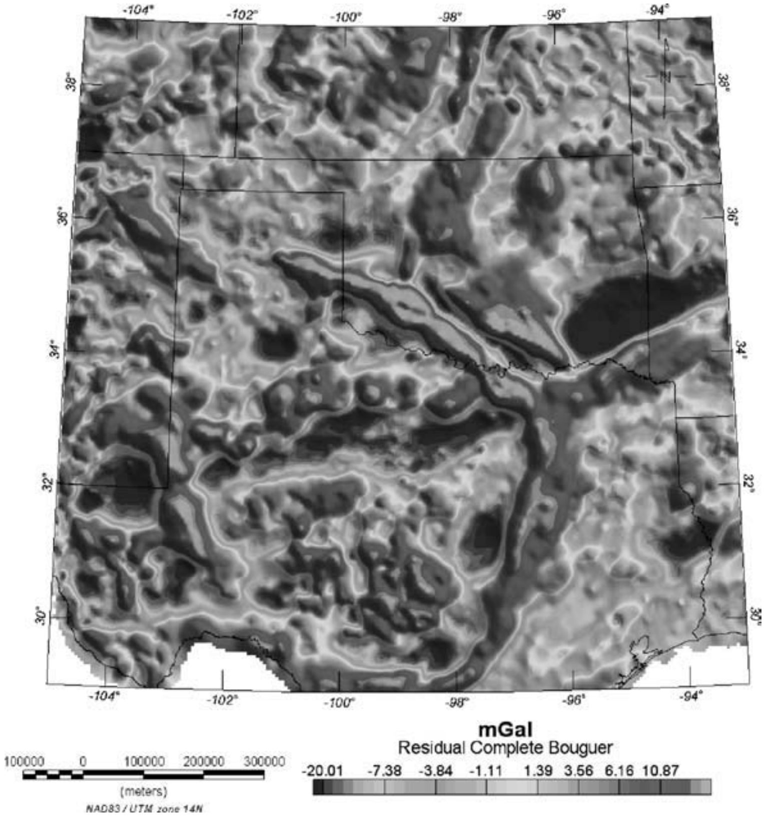



Figure 14.7. Residual gravity anomaly map obtained by subtracting the values in the grid that is visualized in Figure 14.6 from the original anomaly data grid that is visualized in Figure 14.2. The result is to enhance short-wavelength features. See color plates section.

of interest in most studies. Thus, we applied a filter, in the general sense of the term, to the data by first calculating the gravity field predicted by the data at an elevation of 30 km. A contour map of the result of this calculation is shown in Figure 14.6 and clearly demonstrates the west-to-east increase in gravity values. A similar result could have been obtained by applying a high-pass filter that would have removed long-wavelength features in the data or by fitting a low-order polynomial surface to the data. In our approach, by simply using grid math to subtract the gridded values on this smooth surface from the original grid, we produced the residual anomaly map shown in Figure 14.7. This map obviously shows a lot more interpretable texture than the map in Figure 14.2.

As mentioned above, this data system also provides access to the entire North American gridded magnetic database. The workflow to create a map of these data is the same as shown in Figure 14.5, but there is only one anomaly value, the total

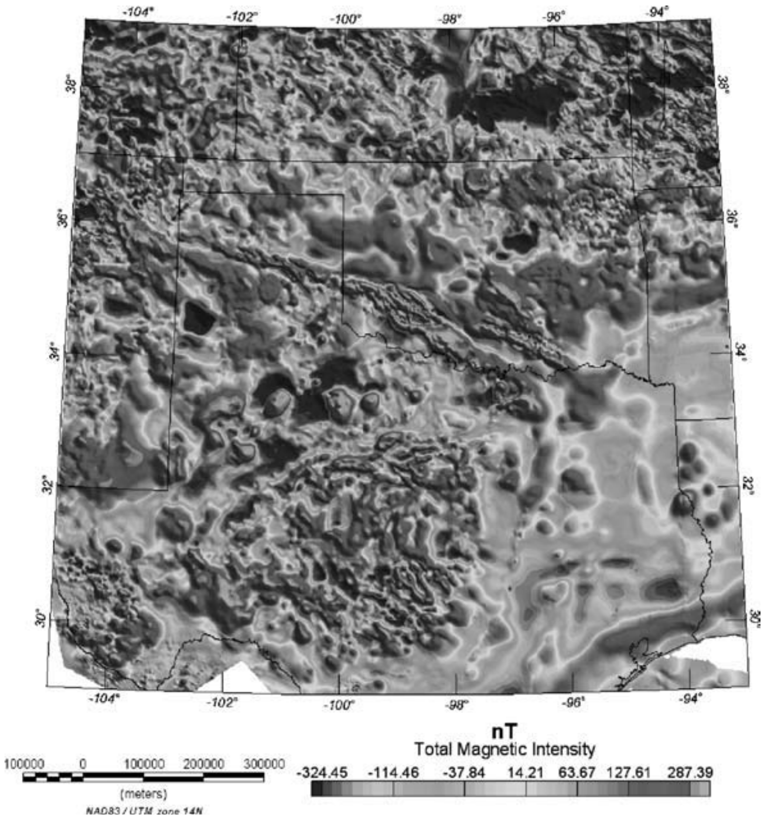



Figure 14.8. Total magnetic intensity map of a portion of the south-central USA. The data were downloaded from our data system and visualized using the workflow in Figure 14.5. See color plates section.

magnetic intensity. A magnetic anomaly map of the same area as in Figure 14.7 is shown in Figure 14.8. Because the magnetic susceptibility of rocks varies over a much wider range than density, and because magnetic minerals lose their magnetization at the Curie depth, which is about 20 km in most continental areas, this map is dominated by short-wavelength anomalies.

One does not have to be a geophysicist to appreciate that maps such as Figures 14.7 and 14.8 would provide considerable information for geological studies of many types. Many case histories of gravity and magnetic studies can be found in Hinze (1985) and Gibson and Millegan (1998). The data system described here provides two types of geophysical data, the software to process field measurements, 2-D gravity modeling software, and directions for using GIS approaches to construct anomaly maps. This is all the result of an ongoing community effort that will continue to improve and expand the system for years to come.

## Acknowledgement

The work described in this paper was partially funded by the NSF GEON project EAR-0225670.


## References


Blakely, R. J. (1995). *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press, 441pp.

Bott, M. H. P. (1960). The use of rapid digital computing methods for direct gravity interpretation of sedimentary basins. *Geophysical Journal of the Royal Astronomical Society*, **3**: 63–67.

Cady, W. J. (1980). Calculation of gravity and magnetic anomalies of finite-length right polygonal prisms. *Geophysics*, **45**: 1507–1512.

FGDC: Federal Geographic Data Committee (1998). FGDC-STD-001-1998: Content Standard for Digital Geospatial Metadata (Revised June 1998). Washington, D.C.: Federal Geographic Data Committee, www.fgdc.gov/metadata/contstan.html.

Gates, A. Q., Keller, G. R., Salayandia, L., Pinheiro da Silva, P., and Salcedo, F. (2007). The gravity data ontology: Laying the foundation for workflow-driven ontologies. In *Proceedings of the Second International Conference on Geospatial Semantics* (GeoS 2007), ed. F. Fonseca, M. A. Rodríguez and S. Levashkin. Heidelberg: Springer-Verlag, pp. 278–287.

Gibson, R. I. and Millegan, P. S. (1998). *Geologic Applications of Gravity & Magnetics: Case Histories*. Tulsa, OK: Society of Exploration Geophysicists, 162pp.

Hammer, S. (1939). Terrain corrections for gravimeter stations. *Geophysics*, **4**: 184–194.

Hildenbrand, T. G., Briesacher, A., Flanagan, G. *et al*. (2002). Rationale and Operational Plan to Upgrade the U.S. Gravity Database. U.S. Geological Survey, Open File Report 02–463, 12pp.

Hildenbrand, T. G., Briesacher, A., Hinze, W. J. *et al*. (2003). Web-based U.S. gravity data system planned. *Eos Transactions AGU*, **83**: 613.

Hinze, W. J., ed. (1985). *Utility of Regional Gravity and Magnetic Anomaly Maps*. Tulsa, OK: *Society of Exploration Geophysicists*, 468pp.

Hinze, W. J., Aiken, C., Brozena, J. *et al*. (2005). New standards for reducing gravity data: The North American gravity database. *Geophysics*, **70**: 325–332.

Holom, D. I. and Oldow, J. S. (2007). Gravity reduction spreadsheet to calculate the Bouguer anomaly using standardized methods and constants. *Geosphere*, **3**: 86–90, doi:10.1130/GES00060.1.

Keller, G. R., Hildenbrand, T. G., Kucks, R., Roman, D., and Hittleman, A. M. (2002). Upgraded gravity anomaly base of the United States. *The Leading Edge*, **21**: 366–367, 387.

Keller, G. R., Hildenbrand, T. G., Kucks, R. *et al*. (2006). A community effort to construct a gravity database for the United States and an associated web portal. In *Geoinformatics: Data to Knowledge*, ed. A. K. Sinha. Special Paper 397. Boulder, CO: Geological Society of America, pp. 21–34.

Martins, C. M., Barbos, V. C. F., and Silva, J. B. C. (2010). Simultaneous 3D depth-to-basement and density-contrast estimates using gravity data and depth control at few points. *Geophysics*, **75**: 121–128, doi:10.1190/1.3380225.

Mickus, K. L. and Keller, G. R. (1992). Lithospheric structure of the south-central United States. *Geology*, **20**: 335–338.

Plouff, D. (1977). Preliminary Documentation for a Fortran Program to Compute Gravity Terrain Corrections Based on Topography Digitized on a Geographic Grid. U.S. Geological Survey, Open-File Report 77–535.

Salayandia, L., Pinheiro da Silva, P., Gates, A. Q., and Salcedo, F. (2006). Workflow-driven ontologies: An earth sciences case study. *Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing*.

Sinha, A. K., ed. (2006). Geoinformatics: Data to Knowledge. Special Paper 397. Boulder, CO: Geological Society of America, 283pp.

Swain, C. J. (1976). A FORTRAN IV program for interpolating irregularly spaced data using the difference equations for minimum curvature. *Computers & Geosciences*, **2**: 231–240.

Talwani, M., Worzel, J. L., and Landisman, M. (1959). Rapid computations for two-dimensional bodies with application to the Mendicino submarine fracture zone. *Journal of Geophysical Research*, **64**: 49–59.

Torres, R., Keller, G. R., Kreinovich, V., Longpre, L., and Starks, S. A. (2004). Eliminating duplicates under interval and fuzzy uncertainty: An asymptotically optimal algorithm and its geospatial applications. *Reliable Computing*, **10**: 401–422.

Wessel, P. and Smith, W. H. F. (1995). New version of Generic Mapping Tools released. *Eos Transactions AGU*, **76**: 329.

# 15

# Scientific workflows for the geosciences: An emerging approach to building integrated data analysis systems

ILKAY ALTINTAS, DANIEL CRAWL, CHRISTOPHER J. CROSBY, AND PETER CORNILLON

## 15.1 Scientific method and the influence of technology

Due to the increasing number and sophistication of data acquisition technologies, the amount of raw data acquired has vastly increased over the last couple of decades (Berman, 2008). This explosion of scientific data, growth in scientific knowledge, and the increase in the number of studies that require access to knowledge from multiple scientific disciplines amplify the complexity of scientific problems. In order to answer these "grand challenge" scientific questions, scientists use computational methods that are evolving almost daily. The basic scientific method, however, remains the same for the individual scientist. Scientists still start with a set of questions, then observe phenomena, gather data, develop hypotheses, perform tests, negate or modify hypotheses, reiterate the process with various data, and finally come up with a new set of questions, theories, or laws (http://en.wikipedia.org/wiki/Scientific_method). A recent change in this scientific method is that it is continuously being transformed with the advances in computer science and technology. The simplest examples of this transformation are use of personal computers to record scientific activity and the way scientists publish and search for publications online. More advanced technologies within the scientific process include sensor-based observatories to collect data in real time, supercomputers to run simulations, domain-specific data archives that give access to heterogeneous data, and online interfaces to distribute computational experiments and monitor resources. The complexity of today's scientific problems and the many state-of-the-art computer science enabled technologies has resulted in a group of tools, known as "scientific workflows," a.k.a., "problem solving environments," developed for making the automation of scientific processes more efficient and faster, with the goal being to help scientists better utilize emerging technologies and resources.

Scientific workflow systems typically allow scientists to develop formal, customizable, reusable, and extensible definitions of all or part of a scientific process and execute them efficiently. Thus, they are becoming critical to the way modern scientists conduct their work, by making technological advances more approachable through integrative user interfaces. These workflow systems promote "scientific discovery" by providing tools and methods to generate workflows via extensible and customizable graphical user interfaces, and support computational experiment creation, execution, sharing, reuse, and provenance tracking. The scientific workflow approach offers a number of advantages over traditional scripting-based approaches, including:

- the ability to formalize the scientific process;
- easier sharing of software processes, deployment under different platforms, customization and rerunning of processes, and extensibility and reuse;
- helping with management of complexity and usability of scientific processes;
- creation of a unified interface to different technologies, tools, and resources;
- annotation of processes with domain knowledge;
- ability to track provenance of the data and the processes, thereby making reproducibility of scientific experiments an easier possibility; and,
- assisting with system monitoring, fault tolerance, and "smart" re-running of a computational experiment.

Various surveys of existing workflow management systems have been performed to investigate the usage and challenges facing such systems in the e-Science context. A larger study on the taxonomy of workflow systems is published in Yu and Buyya (2005) and a recent survey including an analysis of research directions was presented in Barker and van Hemert (2008). As indicated in these studies, there are different types of scientific workflows with different technological and scientific domain foci, e.g., rapid application development and experimental design; distributed execution on the Web, Grid, or Cloud; and, support for efficient validation and dissemination of workflow run results. Several well-known workflow systems include Kepler (Ludäscher *et al.*, 2006), Taverna (Turi *et al.*, 2007), Pegasus (Deelman *et al.*, 2004), Triana (Taylor, 2006) and Vistrails (Freire *et al.*, 2006).

In the geosciences, usage of scientific workflows includes analysis of existing data using scientific algorithms (Jaegar-Frank *et al.*, 2006), combining well-known GIS tools with codes developed in-house (Pennington *et al.*, 2007), and management of streaming data (Barseghian *et al.*, 2008; Barseghian *et al.*, 2010; Fricke *et al.*, 2004). In this chapter, we provide a brief overview of the Kepler scientific workflow environment as a tool for developing and managing geoscience workflows, and describe two workflows that were developed in Kepler as examples. More information on the Kepler project and software system can be obtained from http://kepler-project.org.

## 15.2 Kepler scientific workflow environment

The Kepler scientific workflow system has been developed as a cross-project collaboration to serve scientists from different disciplines, currently coordinated by the San Diego Supercomputer Center. Since 2003, several large-scale projects as well as individual researcher projects across multiple disciplines have utilized Kepler to manage, process, and analyze their scientific data. An important example is the CAMERA project (Altintas *et al.*, 2010) where Kepler scientific workflows are used to create, deploy, and execute applications for the analysis of metagenomics data using community-provided tools. Inherited from the Ptolemy II system (http://ptolemy.berkeley.edu/ptolemyII/), Kepler adopts the *actor-oriented* modeling paradigm for scientific workflow design and execution.

Kepler provides a graphical user interface for designing workflows composed of a linked set of components called actors that may execute under different Models of Computations (MoCs) (Goderis *et al.*, 2007). Actors are the implementation of specific functions that need to be performed, and communication between actors takes place via tokens that contain both data and messages. MoCs specify what information flows as tokens between the actors; how the communication between the actors is achieved; when actors execute (or, *fire*); and when the overall workflow can stop execution. The support for multiple MoCs in Kepler is provided by components called *Directors*. The designed workflows can then be executed through the same user interface or in batch mode from other applications.

Kepler also provides support for users to edit and manage scientific workflows, collect provenance information related to the developed workflows, and generate monitoring reports showing the execution status of actors and progress of execution of the overall workflow. The execution engine can be separated from its graphical user interface thus enabling the execution of workflows in batch mode, in a centralized or distributed fashion (Sudholt *et al.*, 2006; Wang *et al.*, 2008). Using built-in tools, customized workflow components (i.e., actors) can be exported for reuse either locally or publicly, through the Kepler actor repository (see: http://library.kepler-project.org/kepler/).

### 15.2.1 Provenance tracking in Kepler and related system modules

Kepler also provides a provenance framework (Altintas *et al.*, 2006; Crawl and Altintas, 2008; Ludäscher *et al.*, 2008) that keeps a record of the "chain of custody" for data and derived products within the workflow design and execution contexts. Provenance recording is a very important feature of scientific workflow systems as it facilitates tracking the origin of scientific end products, and validating and repeating the experimental processes that were used to derive those products. The *Kepler Provenance Recorder* (KPR) collects information about workflow structure and executions to enable users to track data generated by domain-specific programs. As seen in Figure 15.1a, KPR has plug-in interfaces for new data models, metadata formats, and storage destinations, designed to serve the multidisciplinary requirements

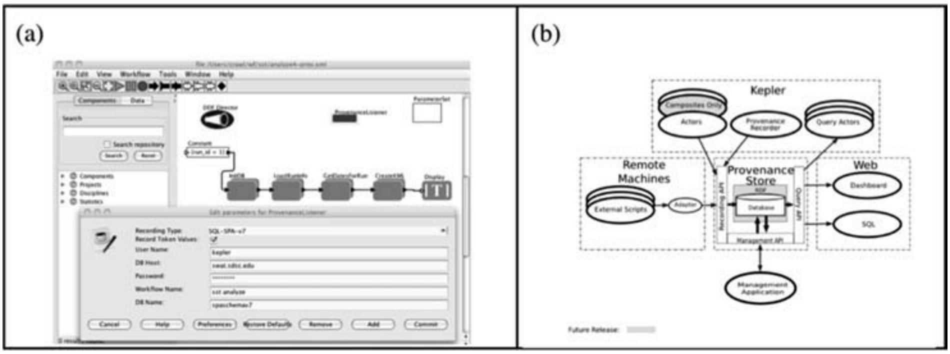



Figure 15.1. The configurable (a) Kepler Provenance Recorder acts as a recording component of the (b) Kepler Provenance Architecture.

of a broad user community. An architecture has been created that includes the KPR (see Figure 15.1b), which allows for binding different data models, collecting application-specific provenance data, and using the results via a dashboard (Vouk *et al.*, 2007). The central aspect of this architecture is the *Provenance Store*: a database providing physical storage and an API to access the database. The API has three components: (1) Kepler, its actors, and external scripts use a Recording API to collect and save provenance information; (2) a Query API provides different query capabilities for dashboards, and query actors in Kepler; and (3) a Management API. Three types of provenance information are collected: the workflow structure, e.g., actor types and parameter values; workflow evolution, e.g., how parameter values change over time; and workflow execution, e.g., the data products read and written by actors during execution. The provenance information collected by the recorder can be stored to multiple data models, including an SQL schema. Additionally, a Query API has been implemented to retrieve provenance information from this schema and is used by the *Kepler Reporting System*, as seen in Figure 15.2a, (Leinfelder *et al.*, 2009) and the dashboard, as seen in Figure 15.2b (Vouk *et al.*, 2007).

In addition to KPR, Kepler also provides a fault-tolerance framework (Mouallem *et al.*, 2010) to deal with system failures that may be encountered during data- and compute-intensive workflow runs, such as parameter-sweep studies. This framework is divided into three major components: (i) a *general contingency* Kepler actor that provides a recovery block functionality at the workflow level, (ii) an external monitoring module that tracks the underlying workflow components, and monitors the overall health of the workflow execution, and (iii) a checkpoint mechanism that provides smart resume capabilities for cases in which an unrecoverable error occurs. This framework takes advantage of the provenance data collected by the KPR to detect failures and help in fault-tolerance decision-making.

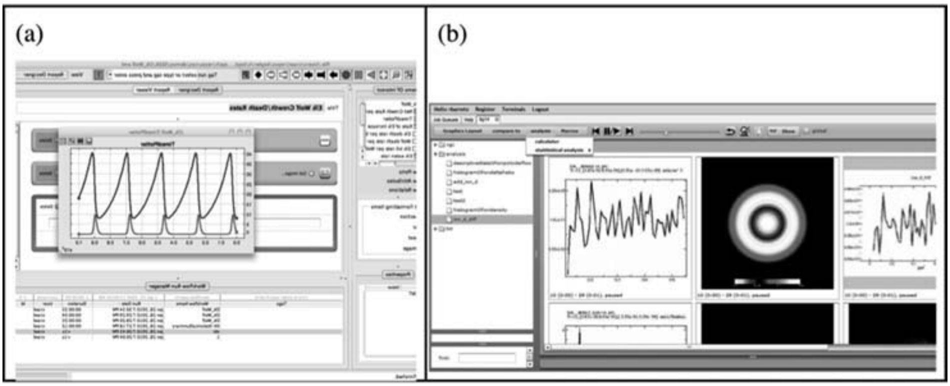


Figure 15.2. Screenshots that illustrate (a) reporting and workflow run manager modules in Kepler, and (b) the dashboard.

### *15.2.2 Distributed computing in Kepler*

Kepler provides a variety of approaches to distributed computing, including web-based workflow execution, web-service-based workflow composition (Altintas *et al.*, 2004), Master-Slave distributed execution of workflow instances (Wang *et al.*, 2008; Wang *et al.*, 2009a), MapReduce-enabled data-intensive execution (Wang *et al.*, 2009b), peer-to-peer execution (Cuadrado, 2008), high-throughput computing using Nimrod/K (Abramson *et al.*, 2008), Globus (Sudholt *et al.*, 2006; Wang *et al.*, 2010), and specialized actors in Kepler for this purpose (Podhorszki *et al.*, 2009).

Figure 15.3 shows a set of parameter-sweep (a.k.a. UQ) execution workflows, which were designed in Kepler: (a) MapReduce (Wang *et al.*, 2009) used in the Ecology domain for House Finch spatial stochastic birth-death process simulation; (b) Master-Slave distributed framework (Wang *et al.*, 2008; Wang *et al.*, 2009a) used in the Ecology domain for House Finch spatial stochastic birth-death process simulation; (c) Nimrod/K, also called TDA, director (Abramson *et al.*, 2008; Abramson *et al.*, 2009; Smanchat *et al.*, 2009) used in the Chemistry domain for quantum chemical calculations; (d) Ad-hoc SSH tunneling controlled by Kepler workflows composed of generic Kepler actors and dataflow directors (Podhorszki *et al.*, 2007; Podhorszki *et al.*, 2009) used in the Physics domain for plasma fusion simulation.

## 15.3 Scientific workflows for geosciences in Kepler

In this section, we present two geoscience-specific workflows constructed using the Kepler workflow system.

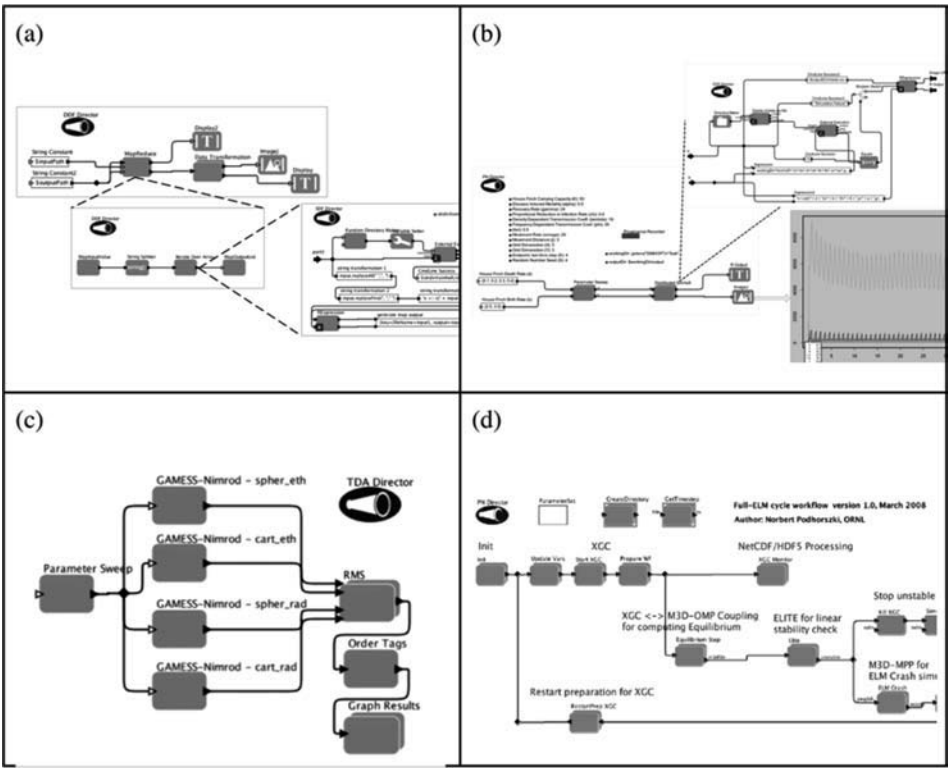



Figure 15.3. Different parameter sweep execution techniques in Kepler are applied to numerous computational science challenges.

### *15.3.1 The GEON LiDAR workflow: A workflow-based architecture for high-resolution topography data access and processing*

As described in Chapter 16, high-resolution topography data acquired with Light Distance and Ranging (LiDAR) technology have emerged as a fundamental tool for earth science research. These data provide an unprecedented three-dimensional representation of the Earth's surface and allow scientists to study active processes at resolutions not previously possible yet essential for their appropriate representation (Carter *et al.*, 2001). Over the past decade, LiDAR data have enabled a range of scientific discoveries, ranging from earthquake science to coastal change (e.g., Prentice *et al.*, 2009; Sallenger *et al.*, 2003).

LiDAR is a remote sensing technology that combines a high-pulse-rate scanning laser with a differential global positioning system (GPS) and a high-precision inertial measurement instrument on an airborne platform to record very dense measurements of the position of the ground, overlying vegetation, and built features (Carter *et al.*, 2007). Operating at tens to hundreds of thousands of pulses per second, LiDAR instruments acquire multiple measurements of the Earth's surface per square meter over large land areas. The resulting dataset, a collection of measurements in georeferenced coordinate space known as a "point cloud." may number hundreds of millions or billions of returns and occupy terabytes of space on disk. The volume of data generated by LiDAR technology has historically impeded wider dissemination and utilization of the data. Many organizations that acquire LiDAR topography struggle to manage the datasets due to their large size, and the lack of disk capacity, bandwidth, compute resources, and in-house expertise necessary to fully process and utilize the data. A more in-depth discussion of LiDAR data and data management issues is provided in Chapter 16. Here, we describe the *GEON LiDAR Workflow* (GLW), a Kepler-based workflow system that was developed for processing LiDAR point cloud data and generating customized digital elevation models (DEMs) based on parameter settings specified by a user (Crosby *et al.*, this volume; Jaeger-Frank *et al.*, 2006).

The GLW project was undertaken as part of the Geosciences Network (GEON) Project (www.geongrid.org). The GLW's goal was to harness emerging cyberinfrastructure technologies to enable internet-based access and processing of such data, thereby mitigating many of the challenges faced by scientists working with large-scale datasets, such as LiDAR topography. The GLW consists of three main computational steps deployed on distributed resources: (i) accessing and subsetting point cloud data, (ii) processing the data to a raster called a digital elevation model (DEM), and (iii) visualizing the results. Coordinating these steps in a Kepler scientific workflow provided built-in sequencing and monitoring of each task, including communication among the various resources. As mentioned before, the scientific workflow environment also provides modularity and extensibility through reusable actors. The LiDAR processing workflow, depicted in Figure 15.4, provides a *conceptual* workflow where each of the components can be dynamically customized by the availability of the data and processing algorithms and the specifics are set "on the fly," prior to the execution of that specific instance of the workflow. This modularized approach can be captured as a workflow pattern of *Subset*, *Analyze*, and *Visualize*, and is defined using the Kepler workflow description language, MoML.

- ***Subset***: LiDAR data are stored in an IBM DB2 spatial database (Nandigam *et al.*, 2010). An interactive, map-based, subset query by the user returns all points (X, Y, Z + attributes) that reside within the selected bounding box. The database connection information along with the query are specified as workflow parameters and are set on the fly prior to the workflow execution. The query is performed on the database and results are stored on an NFS-mounted disk.
- ***Analysis***: The analysis step consists of gridding the data. As shown in Figure 15.4, the query response is shipped to the *analysis* cluster and is then gridded into a raster.

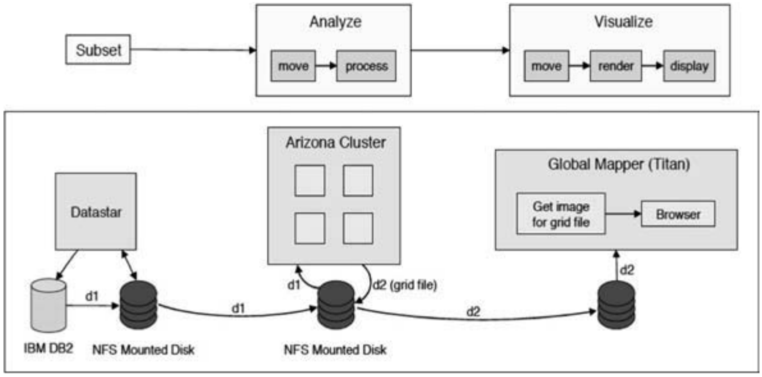



Figure 15.4. Conceptual LiDAR workflow of *subset*, *analyze*, and *visualize* steps. Distributed heterogeneous resources are coordinated in a single scientific workflow environment to execute this workflow.

Currently we use a high-performance local binning algorithm (Kim *et al.*, 2006) deployed on a 4-node cluster.

- ***Visualize***: Finally, the gridded results may be visualized and/or downloaded. A Global Mapper-based web service (www.globalmapper.com) is used to create browser images and Google Earth™ visualizations from the gridded products.

For this workflow, GLW was implemented as a three-tier architecture consisting of a *portal layer*, which is the interface through which the user interacts with the system; the *workflow layer* or control layer, provided by the Kepler workflow system, and the *Grid layer*, which is the processing/computation layer.

***Portal layer***. A Gridsphere portlet (www.gridsphere.org) serves as a front-end user interface. The portlet provides a Google Map™ interface to enable users to define an area of interest, as well as select processing algorithms, processing parameters, and the type of the resulting derivative products. A predefined, parameterized Kepler workflow template with placeholder stubs is then populated with the actual values for this particular workflow instance, prior to the workflow execution. Thus, the workflow instance is created on the fly based on the user's selections, and then scheduled for execution by the workflow layer.

***Workflow layer***. This layer provides the linkage between the *portal* and the *Grid layers*. Controlled by the Kepler workflow manager, it coordinates the multiple distributed Grid components as a single data analysis pipeline. Kepler is capable of submitting and monitoring jobs on the Grid, and handle transfer of derived intermediate products among compute clusters, according to the workflow specification. In addition, it sends control information to the portal client about the execution of the process. The Kepler engine executes the workflow in batch mode, and sends an email notification to the user when the process has completed.

The LiDAR processing workflow involves long-running processes on distributed computational resources under diverse controlling authorities. There is a possibility of system component failures in such a scenario. To manage workflow execution under these circumstances, Kepler provides a data provenance and failure recovery capability using a job database and smart reruns. The job database logs the workflow execution trace and stores intermediate results along with associated metadata. The workflow engine maintains information about the status of each intermediate step and can be used to initiate a smart rerun from a failure point, thus eliminating re-execution of computationally intensive processes.

***Grid layer***. The actual processing implementations are deployed on a distributed computational Grid. A simple submission and queuing algorithm is used for mapping jobs between various resources based on the number of allocated tasks and the size of the data to be processed.

By using a workflow system such as Kepler, the GLW has served to democratize access to the very large and computationally challenging LiDAR datasets by providing simple, web-based access to these data and the corresponding processing services.

Online access to LiDAR topography data is discussed in much more detail in Chapter 16.

### *15.3.2 Sea surface temperature MatchUp workflow*

The Real-time Environment for Analytical Processing (REAP) project is a cyberinfrastructure development effort focused on creating a technology platform in which scientific workflow tools can be used to access, monitor, analyze, and present information from field-deployed sensor networks, for both the oceanic and terrestrial environments, and across multiple spatiotemporal scales (see http://reap.ecoinformatics.org/). The goal is to provide an open-source, extensible, and customizable framework for designing and executing scientific models that consume data streams from sensor networks. The project is combining real-time data with the Kepler scientific workflow system for analysis of such data. One of the sub-projects in REAP is the development of workflows to compute and integrate Sea Surface Temperature (SST) data. SSTs constitute important information for many oceanographic, biological, and ecological analyses. Many types of SSTs datasets are available, from buoys and shipboard sensors to satellite-borne instruments. Two scientific workflows have been developed to facilitate the integration and comparison of these heterogeneous data sources and to perform statistical analyses on the SSTs (Barseghian *et al.*, 2010).

Since SST datasets can be very large, and are available from many different, widely dispersed organizations, only small subsets of these data can be reasonably compared. To facilitate the analysis of these subsets, a *MatchUp* database can be created to store the individual measurements from differet datasets with corresponding temporal and spatial locations. The workflow in Figure 15.5a builds such a *MatchUp* database, based on user-supplied parameters including which SST datasets to use, the number of samples (i.e., size of the subset), and the timespan and region of ocean. For each sample, the workflow randomly chooses a time and location within the user-supplied ranges and downloads the corresponding SST measurements from each dataset into the MatchUp database. Since many SST datasets are stored using OPeNDAP servers, we have to create a Kepler OPeNDAP actor to access these data (OPeNDAP, 2010). Additionally, the Kepler Relational Database actors write data into the MatchUp database.

Once a MatchUp database has been created, the Analysis workflow (Figure 15.5b) creates a KML file for Google Earth™ that displays each sample's location and time along with the average and standard deviation of the temperatures between the datasets (Figure 15.5c). The calculations are performed as part of the workflow using Kepler actors for the statistical package R, as well as other Python-based processing actors.

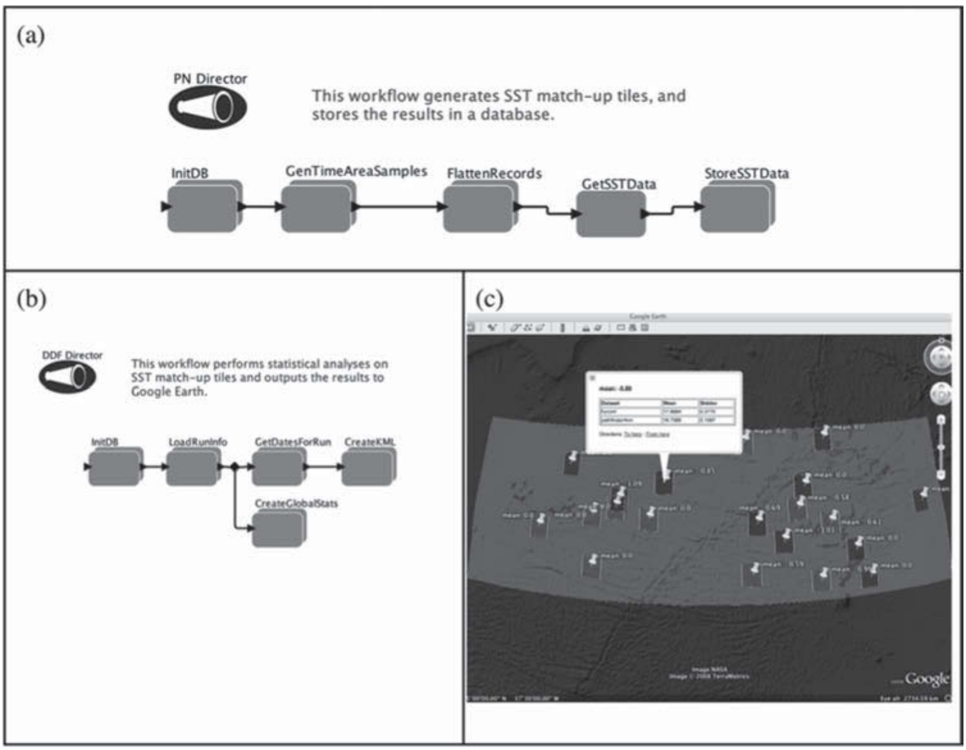



Figure 15.5. (a) Build MatchUp workflow, (b) Analysis workflow, and (c) viewing the results in Google Earth™.

## 15.4 Some challenges in using scientific workflow systems

Scientific workflow systems have been used to streamline scientific data analysis, integration, visualization, and distributed execution. There have been significant advances in the functionality and robustness of such scientific workflow systems, thereby making them an increasingly useful component of the cyberinfrastructure for scientific research. However, some challenges still remain in making scientific workflows easy to use and efficient in the context of a scientist's day-to-day research activity. There are still challenges in providing access to heterogeneous data and computational resources, building linkages to domain knowledge, the ability to interface to multiple analysis tools and workflow systems (Crawl and Altintas, 2008; Ludäscher *et al.*, 2008; Prentice *et al.*, 2009), and comprehensive support for computational experiment creation, execution, sharing, reuse, and provenance. These systems must fit the methodologies of scientists while managing complexity, and user and process interactivity. In order for workflow-based systems to be most efficient, they should minimally provide extensions for optimized failure recovery, smart re-runs, and more importantly, conceptually consolidated file-and-process transport mechanisms that are technology independent from a user's perspective.

Scientific workflows have the potential to be useful in all steps of the scientific process, from managing observational data to publication of scientific results. However, this requires that scientific workflow management systems have support for not only standardized data management and analysis tasks, but that connectivity to common domain tool sets, such as GIS software, and scientific computing environments, such as Matlab and R, are available. Further, the workflow systems must also interact with networked scientific instruments and data from observatories and community data archives. Finally, provenance tracking within and across experiments modeled as workflows can help preserve the data analysis process beyond the active lifetime of a workflow, and could potentially serve as a future citation for the actual analysis process.

## Acknowledgements

The authors would like to thank Bertram Ludaescher and the rest of the Kepler, GEON and REAP teams for their collaboration in the development of the Kepler system and the workflow examples mentioned here.

This work was supported by NSF SDCI Award OCI-0722079 for Kepler/CORE, DOE grant DE-FC02-01ER25486 for SciDAC/SDM, NSF CEO:P Award No. DBI 0619060 for REAP and Gordon and Betty Moore Foundation award to Calit2 at UCSD for CAMERA. The GEON LiDAR Workflow project was supported by NSF grants EAR-0225673 and EAR-0744229.

## References

Abramson, D., Bethwaite, B., Enticott, C., Garic, S., and Peachey, T. (2009). Parameter space exploration using scientific workflows. *ICCS 2009*, Baton Rouge, LA, USA, May 2009.

Abramson, D., Enticott, C., and Altintas, I. (2008). Nimrod/K: Towards massively parallel dynamic Grid workflows. In *Proceedings of Supercomputing 2008 (SC 2008)*: p. 24.

Altintas, I., Barney, O., and Jaeger-Frank, E. (2006). Provenance collection support in the Kepler scientific workflow system. In *Proceedings of International Provenance and Annotation Workshop (IPAW2006)*, pp. 118–132.

Altintas, I., Jaeger, E., Lin, K., Ludaescher, B., and Memon, A. (2004). A web service composition and deployment framework for scientific workflows. In *2nd International Conference on Web Services (ICWS)*, San Diego, California, July 2004.

Altintas, I., Lin, A. W., Chen, J. *et al*. (2010). CAMERA 2.0: A Data-Centric Metagenomics Community Infrastructure Driven by Scientific Workflows. IEEE 2010 Fourth International Workshop on Scientific Workflows, Miami, FL, USA.

Barker, A. and van Hemert, J. (2008). *Information Scientific Workflow: A Survey and Research Directions*. LNCS 4967. Berlin: Springer, pp. 746–753.

Barseghian, D., Altintas, I., and Jones, M. B. (2008). Accessing and using sensor data within the Kepler scientific workflow system. In *Proceedings of Environmental Information Management Conference*, ed. C. Gries and M. B. Jones. pp. 26–32.

Barseghian, D., Altintas, I., Jones, M. B. *et al*. (2010). Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*, **5**: 42–50.

Berman, F. (2008). Got data? A guide to data preservation in the information age. *Communications of the ACM*, **51**: 12, 50–56.

Carter, W. E., Shrestha, R., and Slatton, K. C. (2007). Geodetic laser scanning. *Physics Today*, **60**(12): 41–47.

Carter, W. E., Shrestha, R. L., Tuell, G., Bloomquist, D., and Sartori, M. (2001). Airborne laser swath mapping shines new light on Earth's topography. *Eos Transactions AGU*, **82**: 549–550, 555.

Crawl, D. and Altintas, I. (2008). A provenance-based fault tolerance mechanism for scientific workflows. In *Proceedings of International Provenance and Annotation Workshop (IPAW 2008)* , Salt Lake City, UT, USA, pp. 152–159.

Cuadrado, D. L. (2008). Automated distribution simulation in Ptolemy II. Ph.D. thesis, Aalborg University.

Deelman, E., Blythe, J., Gil, Y. *et al*. (2004). Pegasus: Mapping scientific workflows onto the Grid. In *European Across Grids Conference*, pp. 11–20.

Freire, C. T., Silva, S. P., Callahan, E. *et al*. (2006). *Managing Rapidly-Evolving Scientific Workflows*. In International Provenance and Annotation Workshop (IPAW), LNCS 4145. Berlin: Springer, pp. 10–18.

Fricke, T. T., Ludaescher, B., Altintas, I. *et al*. (2004). Integration of Kepler with ROADNet: Visual dataflow design with real-time geophysical data. AGU Fall Meeting, San Francisco, CA, USA, December, 13–17, 2004.

Goderis, A., Brooks, C., Altintas, I., Lee, E. A., and Goble, C. A. (2007). Composing different models of computation in Kepler and Ptolemy II. *International Conference on Computational Science*, **3**: 182–190.

Jaeger-Frank, E., Crosby, C. J., Memon, A. *et al*. (2006). *A Three Tier Architecture for LiDAR Interpolation and Analysis, LNCS* 3993. Berlin: Springer, pp. 920–927.

Kim, H., Arrowsmith, J. R., Crosby, C. J. *et al*. (2006). An efficient implementation of a local binning algorithm for digital elevation model generation of LiDAR/ALSM dataset. *Eos Transactions AGU*, **87**(52), Fall Meet. Suppl., Abstract G53C-0921.

Leinfelder, B., Altintas, I., Barseghian, D. *et al*. (2009). An integrated approach to managing workflow runs and generating reports in Kepler. In *Eighth Biennial Ptolemy Miniconference*, April 2009.

Ludäscher, B., Altintas, I., Berkley, C. *et al*. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice & Experience*, **18**(10): 1039–1065.

Ludäscher, B., Podhorszki, N., Altintas, I., Bowers, S., and McPhillips, T. M. (2008). From computation models to models of provenance: The RWS approach. *Concurrency and Computation: Practice & Experience*, **20**(5): 507–518.

Mouallem, P., Crawl, D., Altintas, I., Vouk, M., and Yildiz, U. (2010). *A Fault-Tolerance Architecture for Kepler-Based Distributed Scientific Workflows*. In SSDBM 2010, ed. M. Gertz and B. Ludascher. LNCS 6187. Berlin: Springer, pp. 452–460.

Nandigam, V., Baru, C., and Crosby, C. J. (2010). Database design for high-resolution LIDAR topography data. In *SSDBM 2010*, ed. M. Gertz and B. Ludascher. LNCS 6187. Berlin: Springer, pp. 151–159.

OPeNDAP: Open-source Project for a Network Data Access Protocol, http://opendap.org/, 2010.

Pennington, D. D., Higgins, D., Peterson, A. T. *et al*. (2007). Ecological niche modeling using the Kepler workflow system. In *Workflows for e-Science Scientific Workflows for Grids*, ed. I. J. Taylor, E. Deelman, D. B. Gannon and M. Shields. New York: Springer, pp. 91–108.

Podhorszki, N., Klasky, S., Liu, Q. *et al*. (2009). Plasma fusion code coupling using scalable I/O services and scientific workflows. In *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science (WORKS09) at Supercomputing 2009 (SC2009) Conference*. Portland, OR: ACM.

Podhorszki, N., Ludaescher, B., and Klasky, S. (2007). Workflow automation for processing plasma fusion simulation data. In *2nd Workshop on Workflows in Support of Large-Scale Science (WORKS07) at the 16th International Symposium on High-Performance Distributed Computing (HPDC-16 2007)*, Monterey, CA, USA, 2007.

Prentice, C. S., Crosby, C. J., Whitehill, C. S. *et al*. (2009). Illuminating northern California's active faults. *Eos Transactions AGU*, **90**(7): 55–56.

Sallenger, A. H., Krabill, W., Swift, R. *et al*. (2003). Evaluation of airborne scanning lidar for coastal change applications. *Journal of Coastal Research*, **19**(1): 125–133.

Smanchat, S., Indrawan, M., Ling, S., Enticott, C., and Abramson, D. (2009). Scheduling multiple parameter sweep workflow instances on the Grid. *IEEE e-Science 2009*.

Sudholt, W., Altintas, I., and Baldridge, K. (2006). Scientific workflow infrastructure for computational chemistry on the Grid. *International Conference on Computational Science*, **3**: 69–76.

Taylor, I. (2006). Triana generations. In *2nd International Conference on e-Science and Grid Technologies (e-Science)*. New York: IEEE Computer Society, p. 143.

Turi, D., Missier, P., Goble, C., Roure, D. D., and Oinn, T. (2007). Taverna workflows: Syntax and semantics. *In eScience*, Bangalore, India, pp. 441–448.

Vouk, M. A., Altintas, I., Barreto, R. *et al*. (2007). Automation of network-based scientific workflows. *Proceedings of the IFIP WoCo 9 on Grid-based Problem Solving Environments: Implications for Development and Deployment of Numerical Software, IFIP WG 2.5 on Numerical Software*, Prescott, AZ, USA, In *Grid-Based*

*Problem Solving Environments*, ed. P. W. Gaffney and J. C. T. Pool. IFIP, Vol. 239. Boston: Springer, pp. 35–61.

Wang, J., Altintas, I., Berkley, C., Gilbert, L., and Jones, M. B. (2008). A high-level distributed execution framework for scientific workflows. In *Proceedings of Workshop SWBES08: Challenging Issues in Workflow Applications, 4th IEEE International Conference on e-Science (e-Science 2008)*, New York, pp. 634–639.

Wang, J., Altintas, I., Hosseini, P. R. *et al*. (2009a). Accelerating parameter sweep workflows by utilizing ad-hoc network computing resources: An ecological example. In *Proceedings of IEEE 2009 Third International Workshop on Scientific Workflows (SWF 2009), 2009 Congress on Services (Services 2009)*, pp. 267–274.

Wang, J., Crawl, D., and Altintas, I. (2009b). Kepler + Hadoop : A general architecture facilitating data-intensive applications in scientific workflow systems. In *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science (WORKS09) at Supercomputing 2009 (SC2009) Conference*. Portland, OR: ACM.

Wang, J., Korambath, P., Kim, S. *et al*. (2010). *Theoretical enzyme design using the Kepler scientific workflows on the Grid. Accepted by 5th Workshop on Computational Chemistry and Its Applications (5th CCA) at International Conference on Computational Science (ICCS 2010)*, Amsterdam, The Netherlands, 2010.

Yu, J. and Buyya, R. (2005). A taxonomy of scientific workflow systems for Grid computing. In *ACM SIGMOD Record*, ed. B. Ludaescher and C. Goble. *Special Issue on Scientific Workflows*, **34**(3).

# 16

# Online access and processing of LiDAR topography data

CHRISTOPHER J. CROSBY, J RAMÓN ARROWSMITH,
VISWANATH NANDIGAM, AND CHAITANYA BARU

## 16. 1 Introduction

Real-time sensor networks, space and airborne-based remote sensing, real-time geodesy and seismology, massive geospatial databases, and large computational models are all enabling new and exciting research on the forefront of the earth sciences. However, with these technologies comes a prodigious increase in the volume and complexity of scientific data that must be efficiently managed, archived, distributed, processed, and integrated in order for it to be of use to the scientific community. Data volume, processing expertise, or computing resource requirements may be a barrier to the scientific community's access to and effective use of these datasets. An emerging solution is a *shared cyberinfrastructure* that provides access to data, tools, and computing resources. A key objective of geoinformatics initiatives (e.g., Sinha, 2000) is to build such cyberinfrastructure for the geosciences through collaboration between earth scientists and computer scientists.

Airborne LiDAR (Light Distance And Ranging) data have emerged as one of the most powerful tools available for documenting the Earth's topography and its masking vegetation at high resolution (defined here as pixel dimensions less than 2 meters). LiDAR-derived digital elevation models (DEMs) are typically of a resolution more than an order of magnitude better than the best-available 10-meter DEMs. The ability to use these data to construct 2.5-D and 3-D models of the Earth's topography and vegetation is rapidly making them an indispensable tool for earth science research (e.g., Carter *et al.*, 2001). The data enable research into surface processes at the fine scales and broad extents not previously possible, yet essential for their appropriate representation. However, LiDAR technology also represents an excellent example of the massive volumes of data associated with emerging technologies. In many cases, accessing and processing these datasets can be challenging for even the most sophisticated users.

We present an approach to the internet-based distribution and processing of LiDAR data, with particular emphasis on research-grade, community datasets. This approach is a model for the utilization of cyberinfrastructure to tackle the data access and processing challenges presented by the next generation of research-grade earth science data. As discussed in several other chapters in this book, the development of community-oriented data portals as well as internet-based processing and visualization tools is applicable to many types of scientific data, and approaches such as the one illustrated here are likely to revolutionize the utilization of massive and complex datasets.

## 16.2 Airborne LiDAR

Acquisition of airborne LiDAR data utilizes a pulsed laser ranging system operating at 10s to 100s of kilohertz mounted in an aircraft equipped with a kinematic Global Positioning System (GPS) to provide precise positioning information. An inertial measurement unit (IMU) monitors the orientation (roll, yaw, and pitch) of the aircraft. By combining the laser ranging system, GPS, and IMU on an aircraft, it is possible to quickly and economically acquire hundreds of millions to billions of individual point measurements of the absolute $x$, $y$, and $z$ coordinates of the ground surface, vegetation, and buildings in a survey area. LiDAR instruments typically sample the ground surface multiple times per square meter and provide an absolute vertical and horizontal accuracy of 5–10 cm (Shrestha *et al.*, 1999). Collectively, these LiDAR measurements are referred to as the "point cloud" ($x$, $y$, $z$ plus attributes) and typically consist of 100s of millions or billions of returns depending upon the size of the survey area and resolution of the data being acquired. Modern LiDAR instruments are capable of recording multiple returns from each outgoing laser pulse, making it possible to classify the individual laser returns by applying a filtering algorithm to differentiate ground returns from vegetation returns (e.g., Haugerud and Harding, 2001; Sithole and Vosselman, 2004) (Figure 16.1). The ability to segregate the point cloud data based upon the origin of the return significantly enhances the utility of these data.

A generalized airborne LiDAR acquisition and processing workflow consists of the following four steps: (1) Data acquisition, (2) processing of laser ranging, GPS, and IMU data to generate point cloud, (3) point cloud classification, and (4) generation, manipulation, and delivery of ground and vegetation models (Figure 16.2). Typically the first half of this workflow, data acquisition, and point cloud generation (and often the point cloud classification) is handled by the data provider and is usually not a concern to the earth scientist using LiDAR products; although a small percentage of users may have interest in the data acquisition, GPS positioning, and point cloud generation steps in the workflow for geodetically

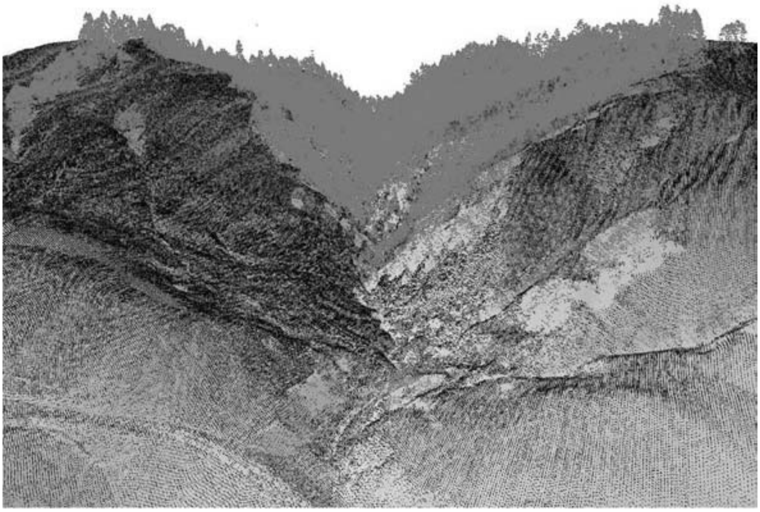


Figure 16.1. Three dimensional rendering of LiDAR ground returns (black) and vegetation returns (green) over the bare earth digital elevation model (DEM), shown in brown. (The Northern San Andreas Fault data were acquired by NASA, in collaboration with the United States Geologic Survey and the Puget Sound LiDAR Consortium, with funding provided by NASA's Earth Surface and Interior Focus Area. The data was collected in 2003 and processed by Terrapoint. The data are in the public domain with no restrictions on their use.) LiDAR instruments typically record multiple returns for each outgoing laser pulse. Individual points can be classified by their source by applying a filtering algorithm. See color plates section.

oriented studies (e.g., Shan *et al.*, 2007). Therefore, for most scientific applications, it is appropriate to consider the LiDAR point cloud the "raw" data product. The later portion of this workflow is of great interest to scientific users because classification and generation of DEMs directly controls the character of the products upon which analyses are performed.

Earth science LiDAR users typically analyze "bare-earth" DEMs derived from LiDAR ground returns because of their science applications and the requirement of computing resources and knowledge not readily available to handle and process LiDAR point cloud data. DEMs are created by gridding the LiDAR point cloud to a raster with elevation data sampled at uniformly spaced intervals ("resolution"; e.g., Mitas and Mitasova, 1999). Gridding millions of densely spaced points is computationally intensive, so LiDAR DEMs – bare-earth and full-feature (unfiltered) – are typically generated once by the data provider and the raw point cloud data are then discarded or ignored in future analysis. In some cases the user may not even take delivery of the point cloud. Users who work only with the vendor-generated DEM products are not taking full advantage of the richness of information held in the point cloud.

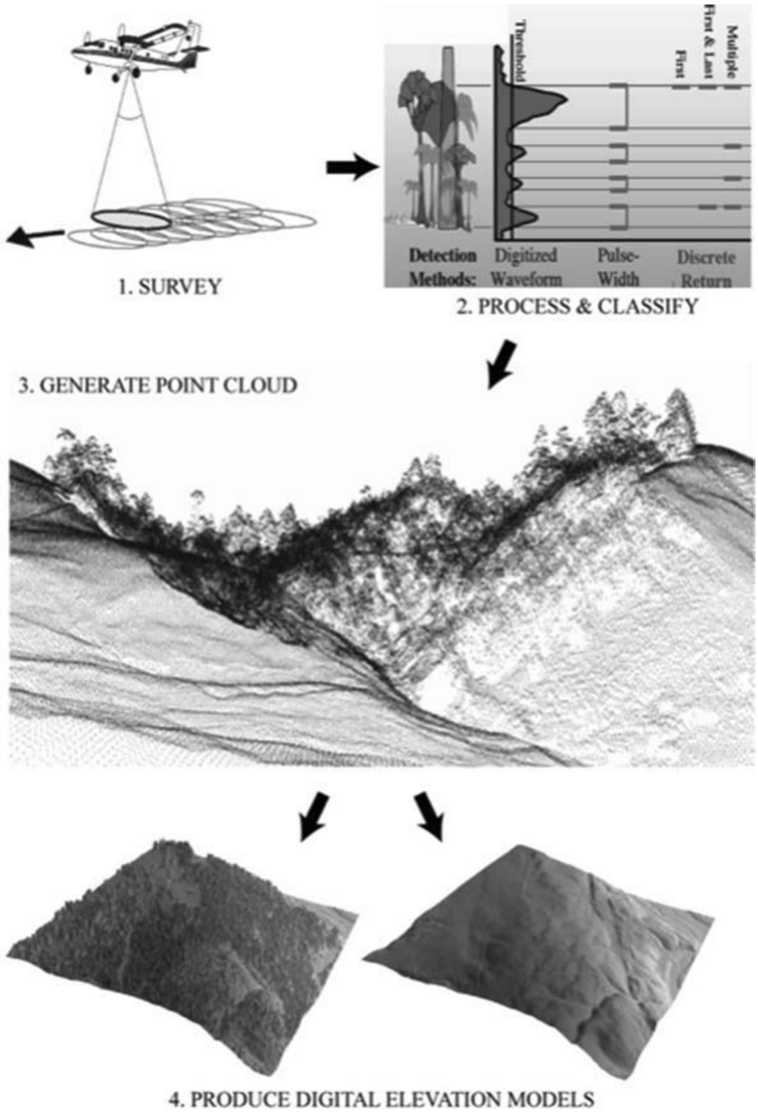



Figure 16.2. Generalized aerial LiDAR acquisition and processing workflow consisting of four steps: (1) Aerial data acquisition – scanner, IMU, and GPS (modified from R. Haugerud, USGS: http://duff.geology.washington.edu/data/raster/lidar/About_LIDAR.html), (2) processing of laser ranging, GPS and IMU data to generate LiDAR point cloud (modified from Harding, 2006), (3) generation of classified point cloud, and (4) generation of digital ground and vegetation models. Point cloud classification and generation of digital ground and vegetation models directly controls the specifications of the products upon which analyses will be performed. See color plates section.

By beginning their analyses with the “raw” point cloud data, users can obtain a better understanding of the data and also gain some control over how those data are processed to characterize the landscape. Details such as the selection of the particular gridding algorithm and the grid resolution employed can significantly affect how well the resulting DEM represents the actual landscape. In addition, beginning with the LiDAR point cloud data allows users to assess the density and heterogeneity of returns, which can vary due to topography, canopy characteristics, and acquisition parameters, and evaluate potential artifacts in the data caused by errors in data acquisition and processing (e.g., point misclassification).

The availability of topographic data of the resolution and accuracy provided by airborne LiDAR has profound implications for research in earth surface processes, natural hazards, ecology, and engineering (Carter *et al*., 2001; Committee on Challenges and Opportunities in Earth Surface Processes, 2010; Stoker *et al*., 2006). For example, in the geosciences, airborne LiDAR data have been used to re-evaluate fault offset and earthquake recurrence for the 1857 Fort Tejon, CA earthquake (Zielke *et al*., 2010); measure topographic change during the 2004 eruption of Mount St. Helens, WA (Haugerud *et al*., 2004); calibrate and test hillslope transport laws (Roering *et al*., 1999); characterize fundamental topographic metrics (valley and ridge spacing; Perron *et al*., 2009); quantify beach topography changes (e.g., Sallenger *et al*., 2003); and, improve flood hazard and floodplain maps under the FEMA National Flood Insurance Program (e.g., North Carolina Floodplain Mapping Program, 2000). In addition to the wide spectrum of geoscience applications for LiDAR data, the ecology community also uses airborne LiDAR data for its ability to provide measurements of vegetation height, biomass, and canopy structure and function (Lefsky *et al*., 2002). Finally, airborne LiDAR data are being utilized in urban environments for planning, feature extraction, and 3-D visualization (e.g., Haala and Brenner, 1999; Maas and Vosselman, 1999).

Due to this range of applications, LiDAR data acquisition has exploded in the past 10 years. While it is difficult to count the actual number of publicly available LiDAR datasets, it is reasonable to assume that the datasets available represent 100s of billions to trillions of individual LiDAR points. Many of these datasets were gathered for a specific scientific or management purpose, yet they are extremely valuable as community datasets if they can be repurposed and made easily accessible. In addition, the simultaneous acquisition of visible and hyperspectral imagery with LiDAR is becoming increasingly common. These datasets should be hosted together and would increase the total size of a dataset. Although this expanding volume of airborne LiDAR data represents a significant opportunity for new scientific endeavors, it also presents a massive challenge in terms of management, distribution, and processing. The current demand for, and interest in, aerial LiDAR data far outpaces the resources available within the earth science community for the distribution and processing of these data.

## 16.3 The challenge of community LiDAR

LiDAR data providers typically deliver the point cloud data in a generic ASCII format or in the binary LAS format (ASPRS, 2009) on a hard drive or DVDs. The system for distribution of these massive public-domain datasets to a user community is often haphazard. Mailing DVDs or hard drives to users who request the data is inefficient and time-consuming. A more elegant approach is to utilize high-performance cyberinfrastructure resources to distribute the data via an interactive portal. The National Oceanic and Atmospheric Administration (NOAA) Digital Coast LiDAR data access site (www.csc.noaa.gov/digitalcoast/data/coastallidar/index.html) and the U.S. Geological Survey's Center for LiDAR Information Coordination and Knowledge, CLICK (Stoker *et al.*, 2006), are examples of internet-based access to LiDAR point cloud data. The NOAA system is devoted to coastal data and is therefore limited in its scope. The USGS CLICK effort is much broader as it seeks to be an archive and distribution resource for raw, unfiltered, data from across the United States.

Although a centralized resource for locating and accessing LiDAR point cloud data such as CLICK is an excellent first step towards allowing earth scientists to utilize LiDAR data in their research, it does not fully address the challenges associated with LiDAR data distribution and processing. It hosts and serves LiDAR point cloud data organized as USGS quarter quadrangles in either ASCII or LAS binary formats. Because of the massive number of returns in a single LiDAR dataset, a single quarter quadrangle's worth of data may contain 100s of millions of returns and result in file sizes of 100s of megabytes. Additionally, by serving data only in quarter quadrangles, users are forced to download multiple data tiles and merge the data by themselves on their own computer systems, if they are interested in a region that falls on a quadrangle boundary.

Although nearly ubiquitous high-speed internet access now makes it possible to download massive LiDAR point cloud data files such as those provided by CLICK, visualization and DEM generation for these data typically require complex and typically expensive software packages (see U.S. Army Topographic Engineering Center, 2006, for software examples), which may be beyond the reach of the average earth scientist due to constraints on finances, computational capabilities, and expertise. This barrier is one of the current limiting factors for earth science users who wish to incorporate LiDAR point cloud data analysis and processing into their research.

## 16.4 A cyberinfrastructure-based approach

The massive data volumes and processing difficulty associated with LiDAR point cloud data make it unrealistic to expect all earth science users who wish to work with these data to acquire independently the necessary tools and resources. An enticing alternative is to employ an approach that utilizes cyberinfrastructure resources to build a community-oriented data distribution and processing gateway for these data. The National Science Foundation Information and Technology Research program-funded Geoscience Network (GEON) project (http://geongrid.org/) (Owens and Keller, 2003) provided the resources and expertise necessary to undertake the development of such a community LiDAR resource.

GEON was designed as an equal collaboration between Information Technology (IT) and geoscience researchers, with the goal of developing an IT platform to facilitate the next generation of geoscience research and education. GEON is based on a "service-oriented architecture (SOA)." This architecture takes advantage of a distributed network of datasets, tools, and computing resources to provide access to high-performance computing platforms for data analysis and model execution. The GEON portal provides a web-based interface to access the various resources.

In addition to the computing resources provided by the project, GEON is designed to bridge cultural and disciplinary boundaries to bring together earth and computer science experts for the common goal of developing the next generation of earth science tools. This close collaboration enables GEON researchers to identify opportunities and tackle problems that they might not otherwise be equipped to handle within their own discipline. The GEON project provides a unique suite of resources and capabilities that can be applied to the LiDAR distribution and processing challenge.

## 16.5 A conceptual workflow for LiDAR data distribution and processing

Given the distribution and processing challenges presented by LiDAR point cloud data, we have created a conceptual workflow solution for an internet-based LiDAR distribution and processing system (Figure 16.3). This workflow capitalizes on cyberinfrastructure technologies, and resources available via the GEON project to offer interactive point cloud data distribution, generation of DEMs, and visualization and analysis of products. The GEON conceptual LiDAR workflow (GCLW) is an end-to-end approach beginning with data distribution and ending with download and visualization of products in two and three dimensions. The workflow is designed to utilize a modular, service-based architecture that is scalable and dynamic. The goal of the modular approach is to allow new tools and resources to be easily added and to allow the workflow to be customized on the fly based upon the processing selections made by the user.

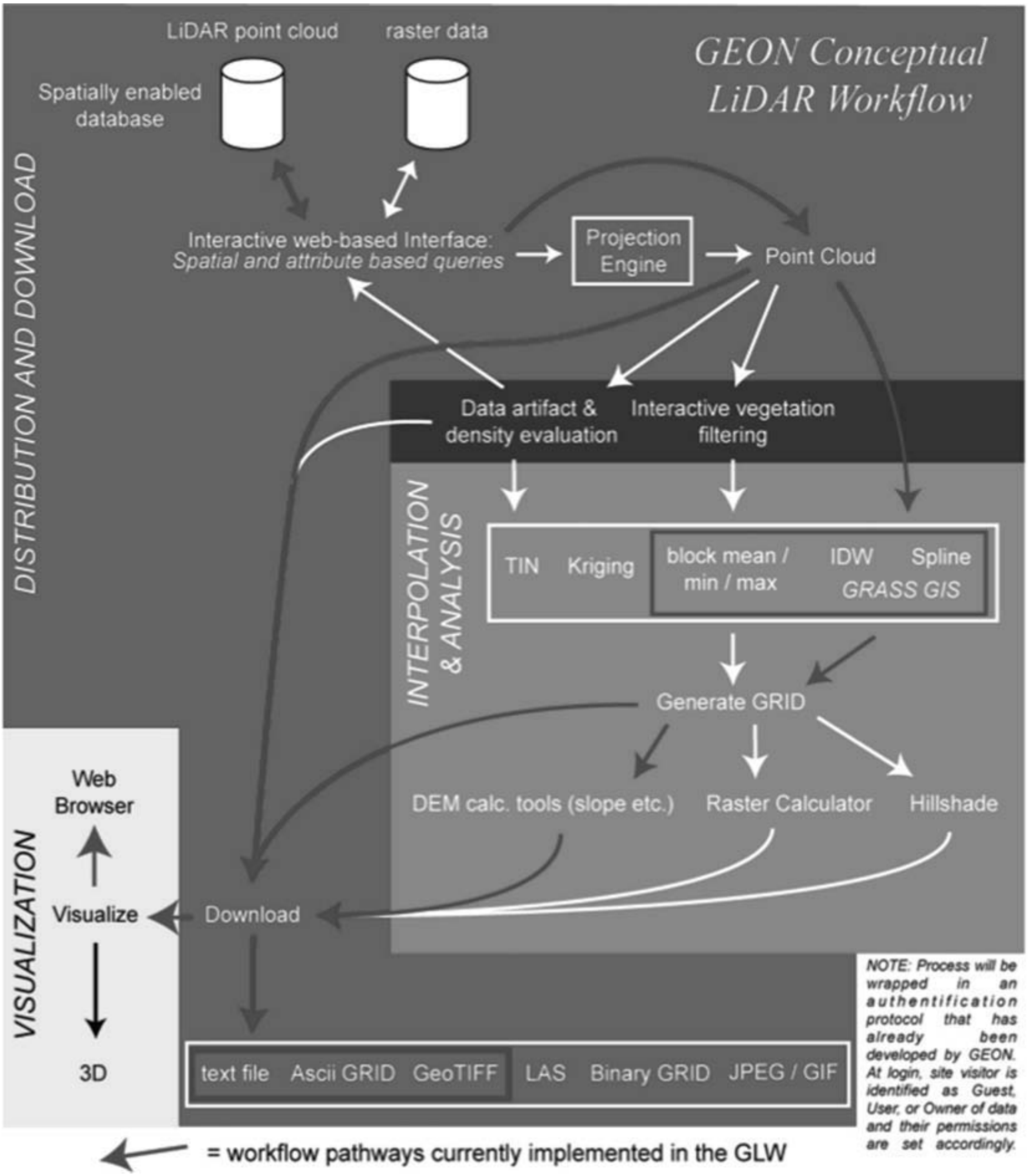



Figure 16.3. GEON Conceptual LiDAR Workflow. The workflow capitalizes on cyberinfrastructure available via the GEON project to offer an end-to-end LiDAR point cloud distribution and processing toolset. The workflow begins with point cloud data (and raster imagery) selection and includes DEM generation and analysis, as well as product download functionality. Pathways shown in gray are currently implemented in the OpenTopography version of the GEON LiDAR Workflow. Pathways shown in white or black are under development.

In the GCLW, LiDAR point cloud data are hosted in a spatially enabled database, which is accessed through a web portal. By storing the point cloud data in a database, users are able to perform spatial and attribute-based queries on the data. For example, the user can select all of the ground returns (attribute) within a given polygon (spatial) drawn on an interactive map in the portal. This approach allows users to select just the data in which they are interested and avoids the problems incurred by hosting data in tiles. We also propose hosting raster imagery that was acquired in tandem with the LiDAR data so that users can also interactively access these data as well.

Once the area of interest and the working geographic projection are selected, users may decide to exit the workflow by downloading the point cloud in either ASCII or LAS formats to their local computer to perform their own processing and analysis. For these users, the workflow acts as a LiDAR point cloud data distribution resource. Other users, however, may choose to take advantage of a gridding and analysis toolset with which they can perform tasks such as point cloud classification, evaluation of return density and classification errors, and DEM generation and analysis. Finally, users can download their data products in a variety of common file formats for easy import into their software package of choice for additional analysis. In certain situations, users may wish to view their data products before, or in place of, downloading to a local computer. In this case, the workflow offers the option of either 2-D or 3-D visualization of data products within the web browser or in free software such as Google Earth™. This functionality could be employed for verifying data products before beginning large downloads, for users who lack visualization and mapping resources locally, or for educational purposes where students use the workflow to explore landscapes digitally.

For earth science users, the DEMs derived from LiDAR are the most frequently used LiDAR product for scientific analysis and 2.5-D visualization. DEM generation can be performed with a number of different interpolation and gridding algorithms, and the workflow provides the user with a suite of algorithms (e.g., spline, TIN, IDW, Kriging, block mean) along with control over algorithm parameters and grid resolution to allow fine-tuning of the resultant surface representation based upon the user's scientific applications. The goal is to provide an interactive processing environment for iterative exploration of various DEM generation algorithms and processing options.

By utilizing the distributed computing resources available through the GEON project, the GCLW allows the user to run multiple jobs with different algorithms and/or parameter settings and compare results. Similar iteration may take days or weeks if done locally on a single computer. The GCLW is designed to allow computationally intensive data processing to be handled by the cyberinfrastructure resources available through GEON and to offer the user manageable download products in common file formats relatively quickly. This approach removes many of the management and computational barriers to working with LiDAR data, thereby democratizing access to these datasets.

## 16.6 Implementation

The GEON LiDAR Workflow (GLW) offers a subset of the GCLW functionality discussed above (gray pathways in Figure 16.3) and provides users access to LiDAR point cloud data, DEM generation and analysis algorithms, and download and visualization of data products all through an internet-based portal. In its most recent incarnation, the GLW is running as a production component in the OpenTopography Facility (www.opentopography.org/pointcloud), an NSF-funded data facility devoted to high-resolution topography data and tools.

The GLW uses distributed computing resources to complete the three main processing tasks within the workflow: point cloud data selection and download, DEM generation and analysis, and download and visualization of derived products. Coordination of these resources is handled through a workflow system developed by the GEON project (Jaeger-Frank *et al.*, 2006) utilizing the Kepler scientific workflow system (Ludäscher *et al.*, 2005) to link the various databases, processing tools, and computing resources (Figure 16.4). Because each tool in the processing workflow is designed to be a modular service, Kepler can dynamically generate a custom workflow – by linking the appropriate modules – based upon the processing options and parameters selected by the user at the portal. The modularity of the GLW architecture also allows us to easily add new datasets, processing tools, and computing resources, making the architecture adaptable and extensible.

The GLW, via OpenTopography, provides access to approximately 37 billion LiDAR returns covering several thousand square kilometers of the western United States. Currently, the available data are primarily focused on active faults and other tectonic landforms. Each of these datasets is hosted in a highly customized IBM DB2 database to allow rapid retrieval of data. Initial implementation of the LiDAR DB2 database was done on DataStar, a TeraGrid data-intensive resource located at the San Diego Supercomputer Center (www.sdsc.edu/user_services/datastar/). The powerful hardware, 64-bit OS, and large real memory and large capacity disk space made this an ideal system for hosting the LIDAR database. Following the decommissioning of DataStar by SDSC, the GLW was updated to use a dedicated 8-node DB2 cluster. This dedicated system uses data partitioning to provide better query performance and scalability at lower cost (Nandigam *et al.*, 2010).

Users interact with the GLW via an internet-based portal built using the Gridsphere Portal Framework (www.gridsphere.org) that enables user authentication. A Google Map™ shows the extent of available data and allows the user to

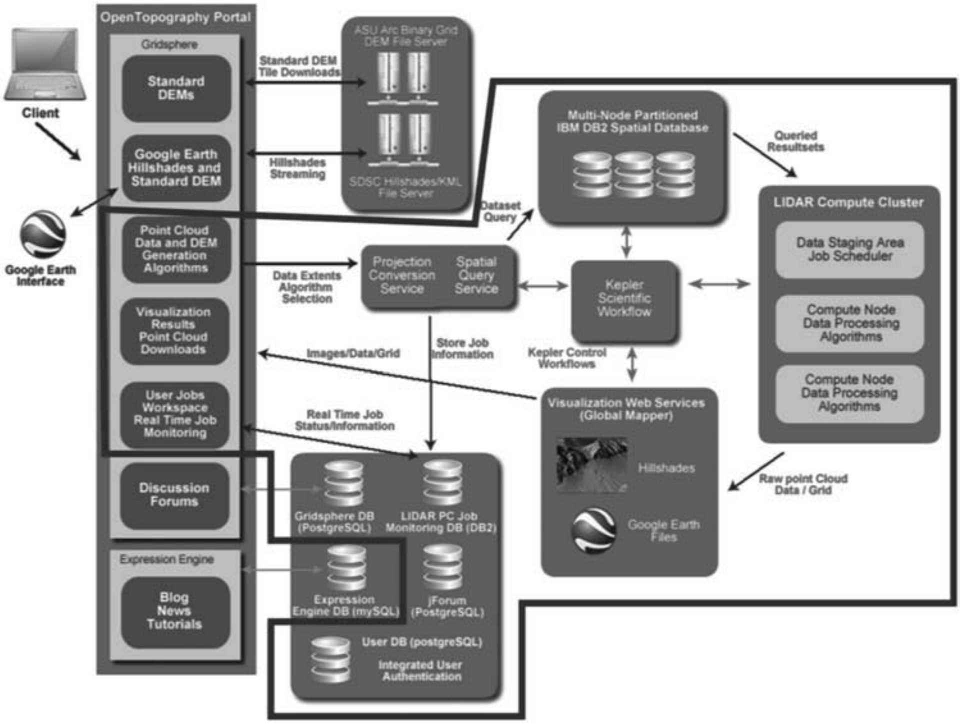



Figure 16.4. OpenTopography system architecture. The GEON LiDAR Workflow system has been incorporated into the overall OpenTopography Portal infrastructure to provide access and processing of LiDAR point cloud data. Components of the OpenTopography system that are relevant to the GLW system are outlined in black. Users interact with the portal interface but are accessing databases and processing resources hosted at the San Diego Supercomputer Center. Processes and distributed resources are coordinated via the Kepler workflow system.

browse and draw a box on the map that defines the spatial query to be submitted to the database. In addition to the spatial selection, users are also able to query by the return classification attribute of the point data (vegetation, ground, structure, etc.) if available.

Once the user has defined their GLW query they are able to make choices about the download products and file formats they would like to receive. The GLW currently offers generation of DEMs via two methods: A spline with regularized smoothing and tension algorithm (Mitasova and Mitas, 1993) available in the GRASS Open Source Geographic Information System (GIS) package (Neteler and Mitasova, 2004); and a local binning algorithm developed by the GEON team specifically for the GLW (Kim *et al*., 2006). At the web portal, users are given control over the DEM resolution as well as algorithm parameters, such as spline tension and smoothing and bin radius, to allow them to produce a DEM best suited to their scientific application. Also through the GRASS GIS service, the GLW will calculate the geomorphic metrics of slope, aspect, and profile curvature as derivatives of the custom DEM (Mitasova and Hofierka, 1993). All data products available via the GLW can be downloaded. Point cloud data are provided in compressed ASCII format, while DEMs and derivatives are available in two ASCII formats and GeoTIFF. Users receive email notification at the completion of their job with links to download results.

In the current OpenTopography implementation, all GLW hardware is located at the San Diego Supercomputer Center. Upon submission of a job, the user's spatial and attribute query information is sent to the LiDAR database, where the query returns all points with the given attribute value that reside within the selected bounding box. Next, the "clipped" subset of data, along with the processing parameters, is sent to a GEON computing cluster to perform the data processing. The GRASS GIS and local binning services deployed on the cluster accept the data and processing parameters and execute the request. Once the DEM and derived product generation are complete, they are passed to a visualization web service that invokes Global Mapper (www.globalmapper.com) to generate hillshade browse images and Google Earth™ KMZ files. Finally, a results page with the hillshade images and links to the user's products is generated and an email is sent to notify the user that their job is complete. The system also generates metadata on the fly that captures information about the user's processing selections and this information is delivered along with the data products at download.

In the current implementation, an end-to-end run of a GLW job will typically complete in tens of minutes depending upon size of query, resolution of DEMs, and system load. Role-based authentication permits us to grant users variable levels of data access. At present, we limit guest jobs (jobs run without logging into OpenTopography) to a maximum of 50 million points, while top-tier "power users" have access to 150 million points per job.

Because the GLW architecture is designed to access and utilize distributed databases and computing resources (Figure 16.4), the system can easily be adapted to expose new resources as they become available. Although the LiDAR data distributed via the GLW are currently hosted by OpenTopography at the San Diego Supercomputer Center, the architecture allows us to expose additional datasets in the GLW portal if they are hosted in a manner that is accessible to the GLW's service-based architecture. Likewise, the GLW architecture allows us to harness additional distributed computing resources if they become necessary.

The flexibility and extensibility of the GLW architecture lends itself to a community-oriented approach to distribution and processing of LiDAR point cloud datasets. This model allows funding agencies or data hosts to maintain control over archiving and management of their datasets, yet still allows their users to seamlessly take advantage of processing tools and resources hosted elsewhere. Through the OpenTopography Facility project, we are exploring this approach to providing centralized data access and processing for remotely hosted data.

## 16.7 Future work

The GLW as presented here represents one step in the ongoing evolution of the system. As Figure 16.3 illustrates, there are still several pathways within the GEON Conceptual LiDAR Workflow that need to be implemented. Efforts are underway, as part of the OpenTopography Facility project, to significantly enhance the GLW's functionality through the addition of additional data processing and analysis tools.

We recognize a need for tools that can be used to manipulate the point data before the user generates their DEMs. A "projection engine" that would allow the user to convert their data selection into a variety of geospatial coordinate systems is essential to allow users to integrate LiDAR data with their pre-existing geospatial datasets. Also necessary are a suite of tools that would allow users to interactively filter the raw point cloud to extract vegetation, buildings, and bare-earth products. DEM generation is a central component to the GLW and therefore we plan to expand upon the current high-performance gridding and interpolation tools currently available. Because each algorithm generates a terrain model from the point data differently, having a variety of choices available is advantageous. Triangular Interpolation (TIN) and Kriging are obvious additions to the GLW.

Although DEM generation from point cloud data is fundamental, we are also working to provide tools that enable higher-level analysis of the DEMs. Beyond the basic geomorphic metrics of slope, aspect, and profile curvature already offered by the GLW, hydrologic routing on high-resolution DEMs is a logical next step. The computational demands of flow routing routines on meter-scale terrain data are large, and the application of cyberinfrastructure resources (including parallel processing) to this problem is likely to be fruitful.

## 16.8 Conclusions

High-resolution topography from airborne LiDAR data is a powerful new tool for the study of the Earth's surface, vegetation, and built environment with utility for a range of scientific, engineering, and planning applications. Access to raw LiDAR point cloud data allows users to take full advantage of the richness of these massive datasets. However, due to the volume and computational challenge of these data, easy internet-based access is currently limited.

Using cyberinfrastructure available through the GEON project, we have developed a geoinformatics-based approach to the distribution and processing of airborne LiDAR point cloud data. The GEON LiDAR Workflow is an end-to-end data distribution and processing system that leaves the computationally challenging data processing to cyberinfrastructure resources. As a result, the GLW has democratized access to LiDAR point cloud for the greater scientific community. The recent initiation of the OpenTopography Facility, which harnesses the GLW, illustrates that online access to LiDAR topography data and related processing tools is a powerful and potentially transformative step for earth science research.

## References

ASPRS: American Society for Photogrammetry and Remote Sensing (2009). LAS 1.3 Format Specification, July 14, 2009, www.asprs.org/society/committees/standards/lidar_exchange_format.html.

Carter, W. E., Shrestha, R. L., Tuell, G., Bloomquist, D., and Sartori, M. (2001). Airborne laser swath mapping shines new light on Earth's topography. *Eos TransactionsAGU*, **82**: 549–550, 555.

Committee on Challenges and Opportunities in Earth Surface Processes (2010). *Landscapes on the Edge: New Horizons for Research on Earth's Surface*. Washington, D.C.: National Research Council, 180pp.

Haala, N. and Brenner, C. (1999). Extraction of buildings and trees in urban environments. *ISPRS Journal of Photogrammetry & Remote Sensing*, **54**: 130–137.

Harding, D. J. (2006). Overview of NASA Airborne and Satellite Laser Altimeters and Commercial Analysis Software, 2006 UNAVCO Science Workshop, Boulder, CO, USA. Original presentation at http://lidar.asu.edu/06unavco_lidarsig.html.

Haugerud, R. A. and Harding, D. J. (2001). Some algorithms for virtual deforestation (VDF) of LIDAR topographic survey data. *International Archives of Photogrammetry and Remote Sensing*, Vol. **XXXIV**-3/W4, pp. 211–217.

Haugerud, R. A., Harding, D. J., Mark, L. E. *et al.* (2004). Lidar measurement of topographic change during the 2004 eruption of Mount St. Helens, WA. *Eos Transactions AGU*, **85**(47), Fall Meet. Suppl., Abstract V53D-01A.

Jaeger-Frank, E., Crosby, C. J., Memon, A. *et al*. (2006). A three tier architecture for LiDAR interpolation and analysis. *LNCS* **3993**. Berlin: Springer, pp. 920–927, doi:10.1007/11758532_123

Kim, H., Arrowsmith, J. R., Crosby, C. J. *et al*. (2006). An efficient implementation of a local binning algorithm for digital elevation model generation of LiDAR/ALSM dataset. *Eos Transactions AGU*, **87**(52), Fall Meet. Suppl., Abstract G53C-0921.

Lefsky, M. A., Cohen, W. B., Parker, G. G., and Harding, D. J. (2002). Lidar remote sensing for ecosystem studies. *Bioscience*, **52**: 19–30.

Ludäscher, B., Altintas, I., Berkley, C. *et al.* (2005). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice & Experience*, Special Issue on Scientific Workflows, **18**(10): 1039–1065.

Maas, H.-G. and Vosselman, G. (1999). Two algorithms for extracting building models from raw laser altimetry data. *ISPRS Journal of Photogrammetry & Remote Sensing*, **54**: 153–163.

Mitas, L. and Mitasova, H. (1999). Spatial interpolation, In *Geographical Information Systems: Principles, Techniques, Management and Applications*, ed. P. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind. New York: Wiley, pp. 481–492.

Mitasova, H. and Hofierka, J. (1993). Interpolation by regularized spline with tension II: Application to terrain modeling and surface geometry analysis. *Mathematical Geology*, **25**: 657–667.

Mitasova, H. and Mitas, L. (1993). Interpolation by regularized spline with tension: I. Theory and implementation. *Mathematical Geology*, **25**: 641–655.

Nandigam, V., Baru, C., and Crosby, C. J. (2010). Database design for high-resolution LIDAR topography data. In *SSDBM 2010*, ed. M. Gertz and B. Ludascher. LNCS 6187. Berlin: Springer, pp. 151–159.

Neteler, M. and Mitasova, H. (2004). *Open Source GIS: A GRASS GIS Approach*, 2nd edn.. Kluwer International Series in Engineering and Computer Science, 773.Dordrecht, The Netherlands: Kluwer Academic Press/Springer, 424pp.

North Carolina Floodplain Mapping Program (2000). About the North Carolina Floodplain Mapping Program, www.ncfloodmaps.com/pubdocs/NCFPMPHndOut.htm.

Owens, T. J. and Keller, G. R. (2003). GEON (GEOscience Network): A first step in creating cyberinfrastructure for the geosciences. *Electronic Seismologist*, **74**(4).

Perron, J. T., Kirchner, J. W., and Dietrich, W. E. (2009). Formation of evenly spaced ridges and valleys. *Nature*, **460**: doi:10.1038/nature08174.

Roering, J. J., Kirchner, J. W., and Dietrich, W. E. (1999) Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology. *Water Resources Research*, **35**: 853–580.

Sallenger, A. H., Krabill, W., Swift, R. *et al*. (2003). Evaluation of airborne scanning LiDAR for coastal change applications. *Journal of Coastal Research*, **19**(1): 125–133.

Shan, S., Bevis, M., Kendrick, E. *et al*. (2007). Kinematic GPS solutions for aircraft trajectories: Identifying and minimizing systematic height errors associated with atmospheric propagation delays. *Geophysical Research Letters*, **34**, L23S07, doi:10.1029/2007GL030889.

Shrestha, R. L., Carter, W. E., Lee, M., Finer, P., and Sartori, M. (1999). Airborne laser swath mapping: Accuracy assessment for surveying and mapping applications. *Journal of American Congress on Surveying and Mapping*, **59**: 83–94.

Sinha, A. K. (2000,). Geoinformatics: A Defining Opportunity for Earth Science Research. White Paper Submitted to the National Science Foundation, www.geoinformaticsnetwork.org/whitepaper.html.

Sithole, G. and Vosselman, G. (2004). Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds. *ISPRS Journal of Photogrammetry & Remote Sensing*, **59**: 85–101.

Stoker, J. M., Greenlee, S. K., Gesch, D. B., and Menig, J. C. (2006). CLICK: The new USGS Center for LiDAR Information and Knowledge. *Photogrammetric Engineering & Remote Sensing*, **27**: 613–616.

U.S. Army Topographic Engineering Center Topography, Imagery and Geospatial Research Division Data Representation Branch (2006). Survey of Terrain Visualization Software, www.tec.army.mil/TD/tvd/survey/index.html.

Zielke, O., Arrowsmith, J. R., Grant Ludwig, L., and Akciz, S. O. (2010). Slip in the 1857 and earlier large earthquakes along the Carrizo Plain, San Andreas Fault. *Science*, doi:10.1126/science.1182781.

# 17

# Use of abstraction to support geoscientists' understanding and production of scientific artifacts

ANN Q. GATES, PAULO PINHEIRO DA SILVA,
LEONARDO SALAYANDIA, OMAR OCHOA, AIDA GANDARA,
AND NICHOLAS DEL RIO

## 17.1 Introduction

With advances in cyberinfrastructure (CI), scientists are able to solve more complex problems; in particular, CI assists in solutions that require collaboration, use of advanced instrumentation, discovery of knowledge resources, and distribution of work across a network of machines. The promise of CI is that it can impact scientific discovery and problem solving by providing the ability to access data collected from shared instruments, evaluate hypothesis by computationally intensive analysis, and retrieve results compiled by other scientists from around the world (Gil *et al.*, 2007). Furthermore, CI-based techniques and methods support the integration, coordination, transformation, and visualization of data and other resources.

Because of the growing complexity of problems being solved by scientists with the use of CI, the use of abstraction is becoming more relevant in helping scientists work with complex processes. Abstraction is the notion of removing unnecessary details to support understanding. How information is viewed, i.e., the perspective of information, guides how abstraction can be used. For example, consider the description of an automobile. From the perspective of a valet, the color, make, and year of the automobile are sufficient to describe a car, and other details can be ignored; from the perspective of a mechanic, details about the engine are necessary. In addition, one can define different levels of abstraction by expanding the amount of detail presented, albeit the amount of detail included at any one level could vary depending on the need. For example, from the perspective of a mechanic, changing the coolant in an automobile does not require the schematics of the engine, while changing the timing belt would require expanded information about the engine.

Specific CI technologies that leverage the use of abstraction and the management of knowledge for working with complex processes include scientific workflows, ontologies, and provenance. Scientific workflows are graphical representations of processes that capture how scientific artifacts are produced. A scientific workflow may be presented visually for human consumption, e.g., to use as a communication artifact among colleagues to discuss a scientific process, or it may be presented as a specification that a computer can execute to carry out a scientific process. There are numerous scientific workflow tools that use varying notations to represent scientific workflows (Ludäscher *et al.*, 2006; Oinn *et al.*, 2005). In general, however, the graphical representation of processes in scientific workflows consists of nodes that denote structured activities that transform data (e.g., activities that are carried out by a program executing in a computer), arrows that denote the data that the activities receive and transform, and additional rules or heuristics that indicate how the graph should be traversed to emulate the scientific process. The value of scientific workflows centers on its promise to capture complex processes that support reuse of resources (e.g., data and processes), reproduction of scientific results, and dissemination of knowledge. One specific challenge about scientific workflows is the ability to present appropriate abstraction levels to different people who may be involved in the specification of a scientific process and who may come from different fields of expertise, i.e., the presentation of workflows from different perspectives and at different levels of abstraction.

Ontologies, accessible through distributed environments such as the Web, encode knowledge about a particular domain and support the discovery, access, integration, and dissemination of knowledge. Indeed, a key function of an ontology is to establish a body of knowledge (including a vocabulary) that can be shared across workflows. Ontologies can be categorized according to their level of dependence on a particular task or point of view (Guarino, 1997). For example, top-level ontologies describe general concepts, e.g., space, time, and matter. Domain ontologies and task ontologies specialize terms used in top-level ontologies. Domain ontologies describe the vocabulary of a particular domain, e.g., P-wave in the domain of seismology, and task ontologies describe the tasks or activities of a particular domain, e.g., computation of P-wave arrival times (Gates *et al.*, 2007). From this categorization, one can say that ontologies leverage abstractions and that concepts in top-level ontologies tend to be more abstract than concepts in domain and task ontologies.

Artifact provenance complements the process knowledge captured in workflows by identifying specific sources of data and data used in performing a task to derive scientific artifacts, e.g., specific instrumentation used to collect data and a gravity dataset, respectively. Provenance can play multiple roles in support of workflows, including that of explaining to scientists how products were created. These explanations may include the identification of sources used, activities followed to create the product, parameters used (if appropriate), and creators of structured activities, among others. Provenance may be overwhelmingly detailed nd hard to understand; thus, abstraction is an essential mechanism to explain provenance (Furtado *et al.*, 2007). For example, provenance visualization tools such as ProbeIt! (Del Rio and Pinheiro da Silva, 2007) can retrieve workflow specifications and use their abstractions to present related provenance. With the help of these abstractions, the tools can present information in a more sensible way to scientists who can then better understand and accept artifacts as valid scientific products, or better understand artifact imperfections.

Broadly, the objective of this chapter is to illustrate how abstraction can be used to aid the scientist in understanding and generating solutions to complex problems. Abstraction is viewed from data and process perspectives, as well as scientist and technologist perspectives. The focus of the chapter is on the scientist perspective. The rest of this chapter is organized as follows. Scenarios from geosciences are used to illustrate the need for different levels of abstraction. After summarizing how other tools and environments support abstraction, the chapter ends with a summary and discussion of future work.

## 17.2 Abstraction perspectives

Figure 17.1 presents four partitions that categorize the gamut of scientific workflows across the dimensions of: (1) level of detail (or abstraction level) specified in the workflow, and (2) classification of the user of the workflow. With respect to the first dimension, workflows range from specifications that offer high-level views of a scientific process, to specifications that describe the finer points of a scientific process. The second dimension ranges from workflows that are geared for use by scientists, and hence capture a scientific process as a scientist would explain to a colleague, to workflows that are geared for use by technologists, and hence capture a scientific process for a technologist (i.e., expressing execution details). While there are four quadrants visibly delineated, the transition from one quadrant to another may be gradual (i.e., a user may expand a workflow that uses abstraction to add more detail when moving horizontally, or gradually introduce technical terminology and notations when moving vertically, or a combination of both when moving diagonally).

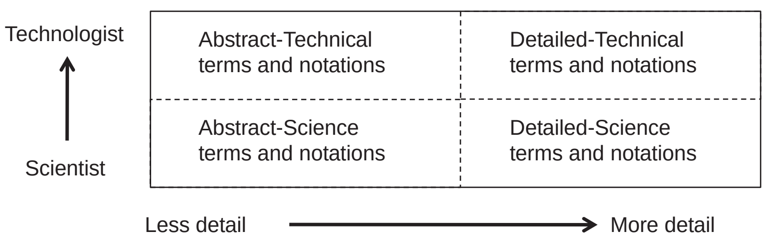



Figure 17.1. Perspectives of scientific workflow.

While there may be numerous uses of workflows, this chapter concentrates on four representative scenarios, where the primary user of each scenario is a scientist. The scenarios elucidate the use of abstraction and separation of technical versus scientific terms and notations. The scenarios are as follows:

- Scenario 1: Preservation of scientific workflows. A scientific result has been generated from a sequence of activities, and the workflow that was used to create it has been preserved. A scientist, probably in a new environment or location, wishes to reproduce the scientific result that was created elsewhere, and he or she re-executes the workflow. Various levels of abstraction are provided for the workflow to support understanding, especially when the workflow involves complex processes.
- Scenario 2: Preservation of provenance. The emphasis of this scenario is on recording how processes and data were used to generate artifacts. Provenance provides resources that help the scientists evaluate: (1) data sources, (2) data, (3) activities, and (4) artifacts produced for use in their scientific endeavors.
- Scenario 3: Experimentation with input. In this scenario, the scientist's interest is on the input and output values of the workflow. The scientist may expand the level of detail of the input data in order to understand what is entered into the process represented at the highest level of abstraction. The aim of the scientist is to experiment with different input values to determine the impact of the values on the output, e.g., a scientist experimenting with the initial model entered into a 3-D seismic tomography modeling program described below.
- Scenario 4: Experimentation with process (components). A scientist may want to change an activity in the workflow, e.g., replacing an algorithm in the workflow, to determine the impact of these changes on an established workflow. In this case, the scientist would need to be at a detailed level using the perspective of a technologist in order to replace the appropriate module.

It is important to note that there are numerous variations of the general scenarios discussed above. The main purpose of the scenarios is to relate how abstractions can be used and provide a common view when reviewing workflow tools.

## 17.3 Abstraction in workflows

Figures 17.2 through 17.4 illustrate various levels of abstraction for a workflow from a scientist's perspective, i.e., the bottom two quadrants of Figure 17.1. The workflow depicts the data and processes for creating a seismic velocity model of the Earth's crust using a nonlinear tomography inversion procedure (Hole, 1992; Hole and Zelt, 1995; Hole *et al.*, 2000) and a finite difference calculation (Vidale, 1988, 1990) using observed travel times of seismic waves through structures in the Earth's crust. This computation, referred to as Hole's Code here, uses seismic waves that are

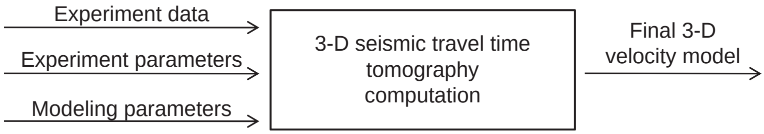



Figure 17.2. Highest level of abstraction for Hole's Code (Level 0).

generated by explosions at shot points and measurements of the arrival times of the waves at seismographs. The velocity models of the Earth are created from a preliminary velocity model that is refined iteratively using data collected from field experiments and application of the aforementioned calculations. The computation uses travel-time picks at specific seismograph locations and shot point times and locations to derive calculated travel times that are used in the iterations of the velocity model until the observed and calculated travel times agree within a tolerance specified by the user. Such models are useful in a variety of situations such as studies of deep earth structure and in oil exploration.

The highest level of abstraction of Hole's Code can be seen in Figure 17.2. Data and activities are described in general scientific terms. Data are represented as labeled arrows; the arrow's direction denotes the flow of data being entered as input to an activity, or emanating as output from an activity. Boxes denote activities or abstracted activities that can be further expanded to finer levels of detail. Notice that specialized notations for iteration and selection of information are not used in the scientist's perspective at this level. It is suggested that these are more applicable in a technical perspective.

The workflow depicted in Figure 17.2 presents a box labeled "3-D seismic travel time tomography computation," and does not present any additional particulars about the activity other than the required input data elements, and the expected output data element. The abstraction represented in Figure 17.2 allows the scientist to focus on exploring the input requirements of the workflow in order to recreate some scientific result, as described in Scenario 1: Preservation of scientific workflows. In addition, the abstraction of Figure 17.2 is beneficial to a scientist that may want to fine-tune the input "Modeling parameters" in order to improve the resulting "Final 3-D velocity model", i.e., Scenario 3: Experimentation with input.

Figure 17.3 presents the sub-activities that comprise the "3-D seismic travel time tomography computation" activity depicted in Figure 17.2, as well as the data interaction involved in each sub-activity. Notice that the notation used to describe data and activities is also refined to include more detail. For example, the input data line labeled "Experiment data" of Figure 17.2 is shown in Figure 17.3 as being translated to "Travel times." An additional notation introduced in Figure 17.3 is the dashed-line box. The dashed-line box represents an activity that is intended to be

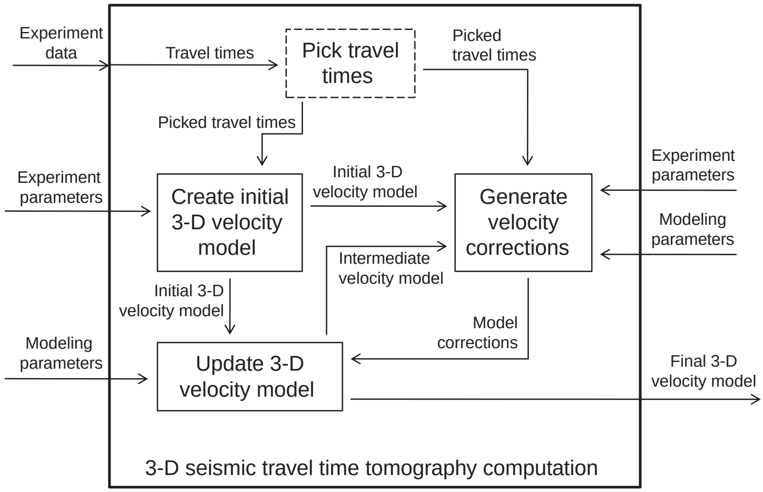



Figure 17.3. A more detailed view of the 3-D seismic travel time tomography computation activity (Level 1).

carried out by a human. In the case of Figure 17.3, the activity labeled “Pick travel times” denotes the user action of picking travel times.

The abstraction level shown in Figure 17.3 supports understanding the activities used to generate new results and experimentation to test new ideas for generating the results, as described in Scenario 1: Preservation of scientific workflows, and Scenario 4: Experimentation with process, respectively. The view captures the main concepts of the workflow without showing the details of each activity, allowing scientists to use this representation of the workflow as a communication tool. This view provides a clear separation between different parts of the workflow, allowing a scientist to identify the activities depicted in the workflow that can be replaced or modified to test new ideas. The scientist could then drill down as far as necessary to target a particular activity. For example, Figure 17.4 shows even more details about the “Create initial 3-D velocity model” activity depicted in Figure 17.3.

## 17.4 Abstraction in provenance

In the context of producing scientific results through the use of scientific workflows, provenance refers to the identification of data and information sources used as input to a workflow, as well as the sequence of structured activities that were performed to produce a scientific artifact (e.g., a gravity contour map).

Although there are various tools and technologies available to encode provenance, the discussion of provenance in this chapter is centered around the Proof

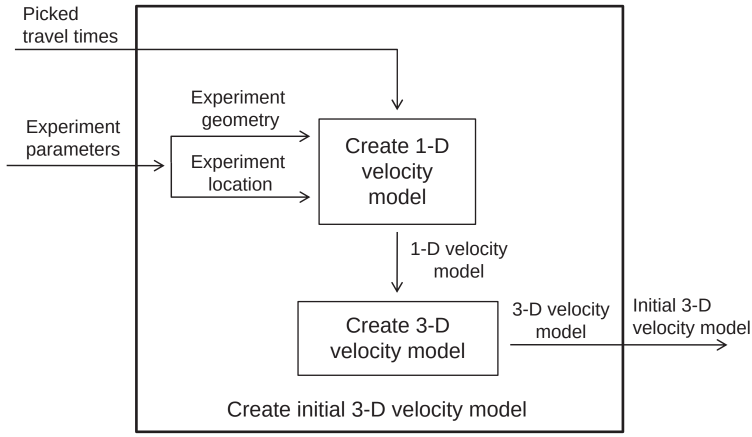



Figure 17.4. A more detailed view of the create initial 3-D velocity model activity (Level 2).

Markup Language (PML) (Pinheiro da Silva *et al.*, 2006) because (1) PML is a technology that is independent of any specific workflow tool, and hence, it is adequate to separate the discussion of abstraction with respect to workflows and provenance, and (2) PML was designed to accommodate a vast variety of scenarios (McGuinness and Pinheiro da Silva, 2004) that range from capturing provenance from undemanding processes to capturing provenance from rigorous mathematical processes, e.g., theorem provers.

The main building block defined in PML is the node set, which contains both a conclusion (i.e., a logical expression) and a collection of inference steps, each of which provide a different justification for the conclusion; in its simplest composition, a single PML node set represents a single proof step. Inference steps contain the following elements: antecedent, rule, and inference engine, which correspond to the antecedent node sets providing input information, the rule applied to the antecedents, and the inference engine responsible for the node set conclusion, respectively. From the point of view of recording what occurred during a workflow execution, antecedents describe data dependencies between the activities of the workflow, rules represent the types of activities that were executed, and inference engines are the platforms or mediums on which the rules executed, e.g., the workflow engine.

In a PML node set, antecedents are references to other node sets comprising the rest of a justification, effectively forming graphs with node sets as graph nodes and antecedents acting as edges. More specifically, this justification graph is directed and acyclic with the edges always pointing towards the direction of the root node set that has the conclusion of the entire justification. In this sense, node sets always contribute to the final conclusion of a justification graph.

### *17.4.1 Provenance logging*

There are different techniques for incorporating provenance logging capabilities into workflow environments. Some workflow management systems rely on a workflow engine to both execute the workflow and manage the related provenance. An alternative is a proxy-based approach where the workflow engine executes each activity in the workflow through a wrapper or intermediate functional module. This wrapper does not disrupt the original workflow execution outcome; however, it intercepts the execution interaction to construct the provenance log.

One characteristic shared by many provenance logging techniques are that they are restricted to recording information specified at a very detailed and technical level, where the preference is to capture all possible details to minimize the risk of missing relevant provenance recordings. For example, Figure 17.2 specifies the term "Experiment Parameters" as data that is input to "Create initial 3-D velocity model." Capturing provenance at a finer level of abstraction, however, may result in a provenance trace that references the terms "Experiment geometry" and "Experiment location" as described in Figure 17.4. This may not seem limiting until one thinks about how to present provenance information to users. In some cases, it may be more useful to present provenance in terms of a higher level of abstraction.

### *17.4.2 Presentation of provenance*

Just as abstraction can be used to present workflows to support understanding of complex processes, abstraction can be used to present provenance to support understanding of how results were generated. Currently, provenance is presented to users at the level of detail that it is recorded, i.e., the workflow execution level. This may not be the best way of presenting information to users considering that too much information may hide key aspects of provenance that can facilitate understanding of how scientific artifacts were derived. To maintain the same level of abstraction with the workflow and provenance, it would be necessary to annotate provenance information with abstract workflow specifications. Thus, provenance visualization tools could leverage the abstractions provided by the workflow specifications (such as the abstract workflows in Figures 17.2 and 17.3) to present the provenance related to more concrete workflows, i.e., the provenance related to the execution of the workflow specification in Figure 17.4.

## 17.5 Workflow tools

Scientific workflow tools provide environments that support varying degrees of abstraction to assist users in designing, managing, and executing processes that model their scientific tasks. Scientific workflow tools often provide features that allow scientists to work at a conceptual level of problem solving that avoids implementation-specific details required to execute a workflow, e.g., choosing a data type or making decisions with respect to execution platforms. Many times, however, in order to allow for the automation of scientific tasks, technologists must facilitate access to data and hardware resources to allow scientists to reuse and combine software components to create executable applications. The next sections discuss some of the abstractions used in five scientific workflow tools: Kepler Scientific Workflow System, Taverna Workbench, VisTrails, WDO-It!, and Wings. The descriptions of these workflow tools in this section are aligned according to their support for the scenarios presented in Section 17.2.

### *17.5.1 The Kepler scientific workflow system*

The Kepler scientific workflow system (Ludäscher *et al.*, 2006) is a tool that has been created to support science in general; thus, the abstractions within the tool are not specific to any domain and are meant to facilitate the documentation of scientific analysis in varied scientific areas. Kepler's main abstraction in describing workflows is the "actor." Actors are executable components such as services or local executables and are configured with ports to specify the input and output connections with data. Kepler provides a library of actors and data sources for use within workflows, and users can create and register more actors for use within the Kepler interface. Finding or implementing executable components often requires detailed knowledge about technical terms and notations, e.g., data types; therefore, the ability to register new actors within the actor library would align more with the tasks of a technologist. A scientific workflow is represented as an interconnection of actors and data sources on a canvas, a graphical window used to display the workflow. Abstraction is supported further in Kepler with "composite actors" that represent sub-workflows containing more process detail. Composite actors coexist with regular actors on a canvas and are interconnected via ports defined when the composite actors are created. In this way, scientists can define workflows using stepwise refinement, adding details as needed, and can import composite actors to interact with the design of the workflow. Via the visual representation of scientific analysis displayed on the canvas, Kepler workflows promote the dissemination of knowledge where actors and composite actors within a scientific workflow can be collapsed and expanded to different levels of detail supporting the understanding of complex processes at various levels of abstraction.

With Kepler, scientific workflows can be created to execute a scientific process and reproduce scientific results, as discussed in Scenario 1. All Kepler workflows are evaluated for dataflow execution. A "director," another abstraction within

Kepler, is used to hide workflow evaluation details of how actors are executed, and how they pass information to each other within a workflow. There are several types of directors in Kepler; therefore, in order to execute a workflow and achieve the desired result, a detailed understanding of how actors work (including inputs, port types, and how they execute within a given director) must be understood. Due to the level of understanding needed to execute a workflow, a Kepler user would function more like a technologist, ironing out technical terms and notations to achieve the needed artifact. As a result, experimentation of input values and executable components, as discussed in Scenarios 3 and 4, is a challenge from the scientist's perspective.

Kepler provides a "provenance recorder" (Altintas *et al.*, 2006), a mechanism for collecting provenance as described in Scenario 2. Adding a provenance recorder to a workflow allows for the collection of workflow building steps and workflow evolution, a method of capturing important instances or versions of a workflow as it is being modified. Furthermore, a workflow's provenance recorder can capture process provenance, i.e., data related to the execution of the workflow. Some of the information captured during workflow execution is embedded in existing components within Kepler; however, for a more complete provenance capture, in particular with details within actor execution, Kepler requires that actors be extended with an API that is used to forward important information to the provenance recorder, e.g., the creation of new artifacts or execution status values. Kepler provides an interface for scientists to configure the collection of provenance within a workflow; the extension of actors to utilize the API and forward provenance would be a technologist-oriented task.

### *17.5.2 The Taverna workbench*

The Taverna workbench (Oinn *et al.*, 2005) is a scientific workflow editing system that was created by the myGrid project (www.mygrid.org.uk) to support experimental biology. In Taverna, scientists are able to link data and third-party applications in a graphical environment highly focused on the user being a life scientist, not a technologist. An important component to providing abstractions in Taverna is its dependence on the configuration of a myGrid middleware created to facilitate the interaction between the scientist and the specification of scientific analysis. The main abstraction supported in Taverna is the "service," usually implemented as a web service. Services are selected from a pre-configured library of available services that have been registered in the middleware by technologists. Moreover, if no service is available for a needed executable, the application is installed by a technologist on the middleware and deployed as a web service. In this way, Taverna tries not to limit the functionality that is available to scientists, but it does require more detailed technical knowledge to facilitate the use of additional technology. Data are also registered by technologists in the middleware and are made available from a repository list.

A workflow in Taverna is an interconnection of data and services displayed as a graph on a workspace. The visual representation is automatically generated; components are automatically placed on the user interface as the scientist makes data and service selections. To the scientist, the allowable connections between data and services are seamless. As the scientist interconnects data with services, Taverna makes suggestions based on knowledge encoded in the middleware by technologists. Technologists are tasked with understanding alignment issues concerning semantic requirements of data when actually connecting data to services. Furthermore, the ability to interchange services and data within a workflow is dependent on how well input and output data align with the semantics of the service. Therefore, support for experimentation of input values or process components, as discussed in Scenarios 3 and 4, is available to the scientist, if a technologist has tailored the needed components within the middleware. Since Taverna is primarily focused on the life sciences community, the myGrid middleware has already been configured with the infrastructure needed to support a broad scope of problems for the community. Workflows in Taverna can be used to support the understanding of complex processes as discussed in Scenario 1, but only at a single level of abstraction. Although the user can configure the amount of information shown on a workflow diagram, neither the data nor the services can be expanded or collapsed to provide further levels of detail in understanding the scientific analysis that is defined within the workflow.

Taverna workflows are executed in a dataflow model of execution. Services are closely tied to executable resources when configured by the technologist in the middleware. This means that the scientist is not responsible for identifying platform details as it is already part of the configuration of the service. Furthermore, once an artifact has been created, workflows can be saved and reused to recreate the artifact.

Taverna captures provenance as metadata that describes different aspects of workflow execution (Zhao *et al.*, 2008), as discussed in Scenario 2. In addition to raw data, either as intermediate data or final artifacts, data lineage is captured for all data. Audit trails are kept with the services invoked as well as information from third-party providers. Finally, Taverna is able to overlay logs and data lineage with provenance details. To leverage the provenance, a suite of APIs is provided as a set of services allowing for the analysis of results through provenance records. A technologist could leverage these APIs to facilitate provenance visualization for a scientist, but these visualization tools are not a default part of the Taverna workbench.

### *17.5.3 VisTrails*

VisTrails (Freire *et al.*, 2006) is a scientific workflow tool that was created to support data exploration through visualization. VisTrails workflows are specified by an iterative process in which a scientist transforms, evaluates, and explores the data under study. With VisTrails, a scientist can create a scientific workflow by choosing from a library of functions, interconnecting the functions with selected data and defining workflow related parameters. All functions that are used within VisTrails must be registered in the library of functions; therefore, initial work must be invested by technologists to make new functions usable within the VisTrails framework. VisTrails supports understanding of complex processes in a workflow, as discussed in Scenario 1, at a single level of abstraction. There is no composition of workflow components allowing for collapsing and expanding to hide or see more details of the analysis.

VisTrails differs from traditional graph-based workflow design tools, such as Kepler or Taverna, because the main abstraction within VisTrails is a “vistrail,” a concept of workflow evolution where all the trial-and-error steps are followed to construct a set of related data products. With VisTrails, a scientist can modify an existing workflow by transforming data or modifying parameters to create a new instance of a workflow and result artifact. Changes to workflows and workflow parameters are systematically stored in a VisTrails abstraction called a “provenance tree” in support of Scenario 2. The provenance tree is used to manage the many paths within the evolution of a workflow. The provenance tree is rooted at the initial version of a workflow, and each node represents changes to the data either from applying a function from the library or by modifying the parameter space. Nodes in a provenance version tree are never deleted or modified; changes result in a new node.

VisTrails workflows are executable workflows. As the scientist is modifying parameters or adding transformations, these are being applied to the data to create new data artifacts directly related to a version of the workflow. VisTrails has a repository for saving workflows for future reuse, supporting the preservation of scientific workflows discussed in Scenario 1. In addition, VisTrails provides mechanisms, e.g., a visual spreadsheet, to allow scientists to select representations of artifacts to make exploratory changes. Scientists can explore input parameters and transformations on data in an exploratory fashion to create different paths of scientific analysis. This aligns with the experimentation of input values and process components as discussed in Scenarios 3 and 4, respectively. Furthermore, through the provenance tree, a user has easy access to each version of the workflow facilitating comparisons of workflows and their artifacts.

### 17.5.4 WDO-It!

WDO-It! (Salayandia *et al.*, 2006) is a workflow tool that allows scientists to build Workflow-Driven Ontologies (WDOs) and create non-executable workflows, i.e., conceptual workflows, that use concepts and relations from these ontologies. Conceptual workflows provide features that allow the scientist to work at a level of problem solving that avoids implementation-specific distractions required to execute the workflow, e.g., choosing a data type or making decisions with respect to execution platforms. Scientists capture domain-specific concepts either by creating concept entries in the WDO or by harvesting them from existing domain ontologies. Concepts captured in WDOs are intended to be used in the specification of processes; hence, concepts in a WDO are classified into two main groups: Data and Method. Data concepts include such things as experiment parameters, datasets, and artifacts, e.g., maps and graphs. Method concepts include such things as software programs, scripts, algorithms, tools, techniques, calculations, and other processes that transform data and yield an artifact of scientific relevance. In these terms, a workflow is considered to be the application of one or more Method concepts in some predetermined order, where each Method concept requires Data concepts as input and yields a Data concept as output. Scientists capture input and output relations among Data and Method concepts to document their intended use in the specification of conceptual workflows. As a result, WDOs can be developed and shared among scientific communities to create common vocabularies, i.e., through the elicitation of concepts and relations, that can be leveraged towards the creation of conceptual workflows.

In WDO-It!, a workflow is created by dragging data and method concepts from a WDO into the graphical canvas. Data concepts are represented by arrow lines, and Method concepts are represented by rectangles. An oval is used to represent a source or entity from which data originate. When a concept is introduced into the canvas, the tool suggests connection points according to the input and output relations captured in the WDO. The scientist at this point has the option to follow the guidance offered by the tool or to define alternate behavior in the workflow. The workflow specified in the canvas is the model that the scientist is endorsing as the intended process. At a later point, the scientist may choose to introduce alternate behavior into the WDO by including alternate concepts and input and output relations.

The notion of information source is adopted from the Proof Markup Language (PML) ontology, and it is used to determine provenance of data. The intention is to use the combination of provenance of data and the process specification captured in conceptual workflows to determine provenance of the artifacts produced through workflow execution. PML and related tools provide the complementary mechanisms to enable such a scenario, which is consistent with Scenario 2 of section 17. 2.

Conceptual workflows in WDO-It! can be modeled at varying levels of abstraction. Method concepts can be broken down into finer conceptual sub-workflows and Data concepts can be extended to constituent Data concepts or aliased to match terminology of a different abstraction level. For example, suppose there exists the Data concept "Principal facts." At a finer level of detail, "Principal facts" can be extended to the Data concepts "Latitude," "Longitude," and "Elevation." In addition, "Elevation" can be renamed to "Decimal number" at a more technical level of abstraction.

Creating conceptual workflows at varying abstraction levels with WDO-It! aligns with the description of the scenario about understanding complex processes in Scenario 1. Scenarios 3 and 4 from Section 17.2, however, are not supported by WDO-It! because they are relevant mainly for executable workflows, not conceptual workflows.

### *17.5.5 Wings*

Another executable workflow tool is Wings (Gil *et al.*, 2006), which was created to support scalable scientific applications consisting of possibly thousands over distributed grid-computing environments. Wings makes extensive use of ontologies to support its workflow building and execution framework that can be used to create large scientific workflows. For the scientist, the main abstraction is the "workflow template." Workflow templates are used to specify sequences of complex scientific analysis at a conceptual level. This is done by connecting data files to nodes, executable components via links, and descriptions of connection constraints. The goal of using workflow templates is to allow for a high level of specification of process steps without concern for data or resource requirements. Nodes and links within a workflow template are semantic objects written as OWL-DL ontologies. The annotation of links and workflow components can be quite detailed and focused on scientific terminology as well as semantic descriptions as well. As a result, setting up the environment to support workflow templates would move more in the direction of a technologist. Nevertheless, using ontologies allows Wings to suggest valid workflow components to scientists while constructing workflows and to validate workflows based on ontology constraints. Since many details of the workflow are specified via ontologies, Wings does not support composition of workflows. There is no drilling down to different levels of abstraction, although Wings does allow for workflow components to be defined in terms of collections of workflow components, to support iteration over data files.

Another abstraction within Wings is the "workflow instance," which is a representation of the workflow template with associated data files. This mechanism allows the scientist to replicate the workflow template across data collections to emulate large-scale scientific analysis. Wings' support for the understanding of complex processes, Scenario 1, is primarily at a single level of abstraction, where a scientist can see the top-level interactions between data and executable components through the workflow template or workflow instance. Workflow instances are specified independent of executable resources.

Wings workflows become executable once workflow instances are mapped to hardware resources. The mapping of workflow instances to hardware resources is a more detailed process, requiring expertise in the management of processes and distribution of data over distributed grid-computing environments. For this, Wings utilizes a workflow mapping system called Pegasus (Deelman *et al.*, 2005), which requires knowledge of the computational infrastructure needed to support large-scale scientific analysis.

Wings supports the preservation of workflows as described in Scenario 1, in that scientists can load and rerun workflows. The number of nodes for a higher-scaled workflow, e.g., thousands of workflow elements, is highly dependent on the availability of the hardware resources utilized to complete the computations. At this level, the verification of workflow execution is a more detailed task and requires support from a technologist. At an abstract level, and assuming that appropriate constraints have been defined for workflow components and links by technologists, a scientist could potentially experiment with input values and workflow components, as discussed in Scenarios 3 and 4. The scientist would still have to instantiate the workflow and create an executable version before an execution could be run, in order to evaluate the impact of the changes on the output.

Within the Wings/Pegasus framework (Kim *et al.*, 2008) provenance is captured at two levels as described in Scenario 2. Wings, through the semantic representations of the interconnections between data and executable components, captures application-level provenance, provenance related to the construction of the workflow, and stores it within the libraries and catalogs of the workflow framework. Pegasus then refines the provenance collection when the workflow is mapped for execution by logging workflow details into the workflow framework, e.g., file locations and hardware selections, and adding provenance-capturing scripts to key phases within the workflow steps. These scripts collect execution provenance and details about the workflow execution, and log them within the workflow framework. With the extensive use of semantic annotations within the workflow components and the logging of provenance in libraries and catalogs during workflow creation and execution, provenance can be queried within the workflow framework concerning workflow specification details. For example, provenance could describe the steps of a process or data lineage, as well as more dynamic information like specific optimizations used during a workflow execution. The tools available to perform such queries are SQL and SPARQL, an OWL-based query language, which are more appropriate for a technologist to leverage.

## 17.6 Summary

To make workflow tools more accessible to scientists, it is critical that they allow one to view a workflow at different levels of detail and use notations and terminology that are geared to the scientist. By removing unnecessary details depending on the immediate task, abstraction can support a scientist who wishes to experiment with scientific processes by changing data sources or high-level processes. Abstraction can aid in understanding complex processes by expanding or collapsing workflow elements. In addition, it can be helpful for teaching and learning about new scientific processes.

From a technologist's perspective, several issues should be considered regarding workflow tools, as noted by Taylor *et al.* (2006). Because of the difficulty of debugging executable workflow components when a workflow is provided at a high abstraction level, workflow tools must also be able to present technical details when needed. Furthermore, details are essential when workflow components are distributed and implemented over web or grid services available through cyberinfrastructure. It is important not to limit the ability to create workflow representations that result in executable implementations. Another consideration is the potential ambiguous nature of abstract workflow notations. Often, the usability challenges of workflow tools can present distractions for the scientist.

The relationship between data and sources is one that needs to be explored. As different levels of abstraction exist between data, it is logical to have different levels of abstraction for data sources. At a conceptual level, data sources must be equally abstract as the data with which they associate. This allows users of the workflow to experiment with combining different outputs into inputs that match the conceptual source description. Because many data sources may generate the same data, concrete workflows must have the ability to mediate the semantics, depending on the source of the data, or multiple data sources must be able to feed into one data type.

Analyzing provenance that was captured at a very fine level of detail has the same limitation as analyzing proofs generated by theorem provers; the amount of information tends to be overwhelming for human consumption and provides no clear indication of which parts of the proofs are relevant to support understanding. The idea presented in this chapter of leveraging workflow specification and, in particular, to leverage abstractions embedded in workflow specifications to abstract workflow specifications is a promising approach for provenance abstraction. Abstractions defined at design time can help explain provenance generated at runtime. In addition, abstractions defined at design time serve as a scope of possible execution paths under which provenance is restricted, providing a verification mechanism of compliance to intended design.

This chapter presents a variety of scientific workflow tools that assist scientists through the use of abstraction with respect to data, processes, and provenance. The tools vary in the amount of support that they provide to scientists and technologists. All tools focus on preservation of information in support of documenting scientific processes and recreation of scientific results. With the trend toward globalized scientific research, technologies such as ontologies and semantic annotations need to be incorporated more into workflow tools in order to allow the sharing and reuse of information, i.e., the sharing of knowledge and processes from different domains through discovery via cyberinfrastructure.

## Acknowledgements

This work was partially funded by the NSF under grants EAR-0225670, HRD-0734825, and CNS-0837556. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the NSF.

## References

Altintas, I., Barney, O., and Jaeger-Frank, E. (2006). *Provenance Collection Support in the Kepler Scientific Workflow System*. L. Moreau and I. T. Foster, eds. *LNCS* 4145. Berlin: Springer, pp. 118–132.

Deelman, E., Singh, G., Su, M.-H. *et al*. (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming*, **13**(3): 219–237.

Del Rio, N. and Pinheiro da Silva, P. (2007). Probe-It! Visualization support for provenance. In *ISVC (2)*, pp. 732–741.

Freire, J., Silva, C. T., Callahan, S. P. *et al*. (2006). Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop, IPAW 2006*. LNCS 4145, Berlin: Springer, pp. 10–18.

Furtado, V., Pinheiro da Silva, P., McGuinness, D. L. *et al*. (2007). Abstracting web agent proofs into human-level justifications. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference, May 7–9, 2007, Key West, FL, USA*, ed. D. Wilson and G. Sutcliffe. Menlo Park, CA: AAAI Press, pp. 80–85.

Gates, A. Q., Keller, G. R., Salayandia, L., Pinheiro da Silva, P., and Salcedo, F. (2007). The gravity data ontology: Laying the foundation for workflow-driven ontologies. In *Proceedings of the Second International Conference on Geospatial Semantics (GeoS)*. LNCS 4853. Berlin: Springer, pp. 278–287.

Gil, Y., Ratnakar, V., Deelman, E., Spraragen, M., and Kim, J. (2006). *Wings for Pegasus: A semantic approach to creating very large scientific workflows*. *Proceedings of the OWL: Experiences and Directions (OWLED-06)*, Athens, GA, USA, November 10–11, 2006.

Gil, Y., Deelman, E., Ellisman, M. *et al*. (2007). Examining the challenges of scientific workflows. *Computer*, **40**(12): 24–32, doi:10.1109/MC.2007.421.

Guarino, N. (1997). Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In *Proceedings of SCIE'97 International Summer School on Information Extraction*, pp. 139–170.

Hole, J. A. (1992). Nonlinear high-resolution three-dimensional seismic travel time tomography. *Journal of Geophysical Research*, **97**(B5): 6553–6562.

Hole, J. A. and Zelt, B. C. (1995). 3-D finite difference reflection traveltimes. *Geophysics*, **121**: 427–434.

Hole, J. A., Brocher, T. M., Klemperer, S. L. *et al*. (2000). Three-dimensional seismic velocity structure of the San Francisco Bay area. *Journal of Geophysical Research*, **105**(B6): 13 859–13 874.

Kim, J., Deelman, E., Gil, Y., Mehta, G., and Ratnakar, V. (2008). Provenance trails in the Wings/Pegasus system. *Concurrency and Computation: Practice & Experience*, **20**(5): 587–597.

Ludäscher, B., Altintas, I., Berkley, C. *et al*. (2006). Scientific workflow management and the Kepler system: Research articles. *Concurrency and Computation: Practice & Experience*, **18**(10): 1039–1065.

McGuinness, D. L. and Pinheiro da Silva P., (2004). Explaining answers from the semantic web. *Journal of Web Semantics*, **1**(4): 397–413.

Oinn, T., Greenwood, M., Addis, M. J. *et al*. (2005). Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience*, **18**: 1067–1100.

Pinheiro da Silva P., McGuinness, D. L., and Fikes, R. (2006). A proof markup language for semantic web services. *Information Systems*, **31**(4–5): 381–395.

Salayandia, L., Pinheiro da Silva, P., Gates, A. Q., and Salcedo, F. (2006). Workflow-driven ontologies: An earth sciences case study. In *Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing*. New York: IEEE, pp. 212–220.

Taylor, I. J., Deelman, E., and Gannon, D. B. (2006). *Workflows for e-Science: Scientific Workflows for Grids*. New York: Springer, 514pp.

Vidale, J. (1988). Finite-difference calculation of travel times. *Bulletin of the Seismological Society of America*, **78**(6): 2062–2076.

Vidale, J. (1990). Finite-difference calculation of travel times in three dimensions. *Geophysics*, **55**(5): 521–526.

Zhao, J., Goble, C., Stevens, R., and Turi, D. (2008). Mining Taverna's semantic web of provenance. *Concurrency and Computation: Practice & Experience*, **20**(5): 463–472.

# Part VI

Emerging international and other efforts

# 18

# It's not your data, it's everyone's: The benefits of a corporate approach to scientific information

IAN JACKSON AND RICHARD HUGHES

## 18.1 Background

The British Geological Survey (BGS) is the national geological survey of the United Kingdom with national responsibility for the acquisition, analysis, management, and delivery of geoscience data, information, and knowledge. BGS employs approximately 750 people and its budget is ~£60 million, only 50% of which is funded directly by government.

In 1992, a UK Treasury agency study of BGS and its data and information management and concluded that:

> The fragmentation of data, and the questions about its accuracy, mean that BGS information systems cannot support BGS's claims to be the National Archive for Geological Data or a National Geoscience Information System. They also call into doubt the value of the unique data holdings as a major competitive strength.

Perhaps reflecting the internal view of the relative importance of data prevailing in BGS at that time, this indictment did not result in immediate remedial action. Data storage in non-interoperable information silos and fragmented management of the data (and arguably the organization) continued. However, the criticism was to be repeated in a further external review in 1996 and, following that, action was taken. In 2000, after 2+ years of pilot studies and developing a new strategy, BGS restructured radically. It moved from a hierarchically managed organization to a matrix managed one and also created a dedicated Information Directorate. This new Directorate was assigned approximately one third of the BGS budget and corporate priority was given to work on data and information. In particular, work on metadata, standards, data product development, and delivery was accelerated. BGS had recognized that the commodity it dealt in was data and information and that its role was, in a digital era, an information organization. Corporate policies for data management and standards were developed and implemented.

The investment of ~33% of the allocatable BGS budget in data and information – equivalent to approximately £5 million per annum for the last 10 years, coupled with continued priority for information activity – has delivered very significant internal and external benefits, perhaps best illustrated by the availability of up-to-date, quality assured and interoperable national versions of all prime geoscience datasets, and internal and external access to an extensive range of core and value-added internet information services. But those gains were not made easily.

## 18.2 Challenges

The challenges faced by BGS in addressing the poor state of its data and information policy and practice in the 1990s are likely to be similar to those faced by every science-led organization. They reflected the dominant (understandable, but not excusable) attitudes and priorities of the scientists and science managers prevalent at the time. These challenges included:

- Fieldwork and research being accorded much more formal and informal priority; data management was seen as an inherently tedious and unproductive task.
- Scientists were protective of their data; they saw it as "their" property and worried that others may misuse it. Citation and publication do not sit happily with data dissemination.
- Individual (idiosyncratic) approaches meant that standards, either technical or semantic were not complied with.
- Current research objectives were prime; the post-project (long-term) value of data was accorded far less priority.
- A cultural divide between information system and technology experts and scientists.
- A focus on localized independent research projects and a failure to give precedence to the unique national remit and the need for systematic consistent coverage.
- The difficulty of getting the survey scientist to truly focus on the needs of society and their stakeholders and to effectively engage and communicate with them to establish and meet their real needs.

## 18.3 Progress

Over the last decade, BGS and the majority (but not yet all) of its staff have recognized that its Unique Selling Proposition (USP) is that it is "National, Long-term and Strategic" and both its expertise *and* data that make up its core competence are central to its USP. The BGS took a major and radical decision to change, and invested heavily in corporate best practice information management and delivery. That investment is the basis for more comprehensive data that are more easily accessed by the entire diverse range of its stakeholders, for state-of-the-art geoscience information systems, and for knowledge that is enabling BGS to play a full role in global geoinformatics.

One of the fundamental objectives of the BGS approach was – and is – to deliver high-quality national digital baseline datasets. These provide the basis for value-added datasets that have also been developed in-house, some of which are of significant commercial value. For example, development of the BGS DiGMapGB digital geological map dataset (an attributed, vector data-set, not merely a digitally drawn map) enabled BGS to develop a suite of national ground stability datasets (GeoSure) that was first launched in 2003 and is now used extensively by the property and insurance sector. More recently BGS – in partnership with the Health Protection Agency – has released a natural radon dataset for England and Wales by combining geological data with empirical natural radon measurements.

Unless geologists can share digital data and information between each other and with other sciences, the value and contribution of geoscience will never be fully understood or appreciated. BGS's investment in geoscience information and informatics, and the resultant expertise, which it developed, is allowing it to play a leading part in initiatives aimed at achieving global interoperability in the geosciences. The organization plays a full role in the development of GeoSciML and instigated OneGeology. Within the organization, it is operationally deploying three-dimensional modeling systems and workflows and is moving rapidly on to four-dimensional systems. To support and also exploit these developments, the BGS continues to develop innovative intuitive modeling, as well as virtual field reconnaissance and visualization applications. Digital data capture has been implemented corporately and is now the dominant method of fieldwork for BGS field geologists. The field map case most of us are familiar with has been largely replaced by an A5 ruggedized graphics pad and GPS.

But the most fundamental point is that these and all other deployment of web, modeling, and GIS applications is that they could not be contemplated and would not deliver the efficiencies or new scientific opportunities without the information management infrastructure in place to support them.

Investments in Information Management and the creation of national datasets have been matched by investments in web-based delivery systems for digital data and information. These systems service the internal demands of BGS scientists and the external demands of the wider user community. Internally a very high proportion of data and information is now delivered in digital format to the scientist's desktop. The demand for BGS data and information from the wider community is considerable and grows every year. Users include government (local, regional, devolved, and central), commerce, industry, utilities companies, academia, and the education sector. Enquirers can use the online GeoIndex geospatial search interface (www.bgs.ac.uk/geoindex) to find what data are held on their area of interest. To help those who need it, we have created BGS's dedicated enquiries telephone and web helplines.

A range of delivery systems is able to service varying demands for data and information. At one end of the spectrum are commercial customers who want multi-seat, multi-year licence agreements to use entire national datasets. These data can be supplied in various ways, including secure extranet delivery. At the other end of the spectrum are private individuals who want BGS data and information on a site-specific basis. These needs are serviced by a number of delivery streams, including the BGS's own GeoReports service (cited by the UK government's Department for Trade and Industry as an example of best practise in public sector knowledge transfer). Other delivery streams have been developed in partnership with other public sector bodies such as ground stability reporting with the UK Coal Authority (which has statutory responsibilities for coal-mining-related hazard data in the UK). BGS data and information are also packaged and sold on a site-specific basis by property and environmental report companies, themselves part of the burgeoning knowledge economy.

BGS's expertise in web delivery of digital data has resulted in a number of commissions from the private sector. For example, BGS developed a portal (Digital Energy Atlas and Library – DEAL) that provides web access to hydrocarbon exploration and infrastructure data. These data help the UK government and industry to maximize the exploitation of dwindling hydrocarbon reserves on the UK continental shelf, to improve environmental management and regulation of oil- and gas-related infrastructure, and to assist other industries such as fisheries. Significant effort and informatics innovation is also invested in public outreach products and services, and we are trying to make sure we are providing the UK university and education sectors with the data and information they need for teaching and research.

## 18.4 Fee and free

The exploitation of BGS science through investment in information delivery yields considerable economic and wider benefits to both its users and the BGS. In 2006, the BGS's parent body, the Natural Environment Research Council (NERC) commissioned the economists PricewaterhouseCoopers (PwC) to carry out a study into the economic benefits of its research portfolio. PwC chose a small number of case studies for close analysis, including the BGS national ground stability dataset. PwC concluded that this dataset could save the UK insurance industry up to £270 million between 2006 and 2030 in reduced insurance claims alone. This figure excludes a range of wider "invisible" benefits to the individual, and is therefore very likely to be an underestimate of true economic benefit. Other BGS products such as the natural radon dataset for England and Wales (developed jointly with the Health Protection Agency) will have major positive public health impacts through the avoidance and mitigation of illness and death due to the cancer-causing effects of natural radon. Benefits to BGS include support of our business strategy through recovery of costs (in line with current UK government policy on charging for public sector information), maintaining a strong focus for our research programme on meeting real user needs, and a strong external reputation in delivering knowledge transfer that yields real economic and social benefits.

However, over the last few years, the BGS has also progressively made more and richer data freely available via the Internet. In December 2009, it launched its new web site – OpenGeoscience – a free service where users can view maps and download photographs and other information. OpenGeoscience material is free-of-charge for non-commercial private study, research, and educational activities. Users can explore the six OpenGeoscience sections – Data, Education, Maps, Pictures, Reports and Software on an easy-to-use web interface. Notably the site provides the world's first open-access supply of street-level geological mapping for a whole country – better than 50 m resolution – with on-the-fly viewing of bedrock and superficial geology overlaid on street maps and aerial photography. The web map service (WMS) provides access to 1:50 000 scale geological mapping for mash-up use (allowing access via OneGeology) and has open access to 50 000 photos, databases, educational resources, reports, and software.

Encouraging innovation, research, teaching, and non-commercial use of BGS materials, OpenGeoscience complements established commercial supply routes for BGS resources and provides the platform for further open-access releases of BGS materials, e.g., scanned records information. Illustrating just how popular OpenGeoscience is with the ordinary citizen, on the day of its release, it was the most popular (shared) story on BBC News Online, with over 70 000 visitors to the BGS web site from the time it appeared at 10.30 a.m. – the number of visitors the BGS web site normally gets in an average month. There were more than 10 million image hits (1000 a second!) and 300 000 page views.

## 18.5 In summary

The challenge for geological surveys is to make sure that the centrality of information and informatics is not forgotten, and that we share and spread best practice as widely as we can. Data, information, and knowledge on the geology of our territories are the commodities that national and state geological surveys deal in and, more to the point, are what society expects us to provide. In a digital era, that expectation has moved with the times – geological surveys must do likewise, and to do that they must take their information management responsibilities corporately and seriously.

# 19

# TOPO-EUROPE and cyberinfrastructure: Quantifying coupled deep earth – surface processes in 4-D

S. A. P. L. CLOETINGH, H. P. BUNGE, AND THE TOPO-EUROPE WORKING GROUP[1]

## 19.1 Introduction

The earth sciences differ from other scientific disciplines in their focus on processes one cannot easily repeat or control. Examples are the nucleation of earthquakes as brittle failure along faults, or the creeping flow processes in the Earth's interior driving plate tectonics and the geologic activity of our planet. The inherent experimental limitations and the indirect nature of our observations explain the need for sophisticated modeling approaches. And with continued growth of computer hardware performance, the crossing of some long-standing thresholds in capacity and capability computing is finally underway. For instance, it is now feasible to implement earth models having in excess of 1–10 billion grid points, a number that matters because it allows us to overcome in three-dimensional (3-D) models the disparate length scales characteristic of key geologic phenomena: an earthquake rupturing a fault segment over a distance of some 100 km while emanating seismic energy throughout the planet (10 000 km), or the peculiar nature of plate tectonics with deformation concentrated along narrow plate boundaries of 10–100 km width separated by plates of dimension 1 000–10 000 km. But capable hardware and raw compute power are not sufficient by themselves to advance demanding earth system simulations. Equally critical are sophisticated software, visualization tools, data

[1] TOPO-EUROPE Working Group: U. Achauer, A. Adam, P. A. M. Andriessen, I. M. Artemieva, G. Bada, R. T. van Balen, A. Barabas, F. Beekman, J. Behrmann, Z. Ben-Avraham, A. Beylich, J. M. Bonow, P. J. F. Bogaard, J. Braun, J.-P. Brun, S. Buiter, E. B. Burov, R. Carbonell, L. Csontos, J. Dehls, P. Dèzes, C. Dinu, E. Dombradi, J. Ebbing, E. Eide, C. Facenna, T. Fancsik, L. Fodor, O. Fredin, A. Friedrich, W. Frisch, J. Gallart, L. Gemmer, J. Genser, A. G. Green, G. Grenerczy, J. M. Hagedoorn, S. Harangi, E. Hegedus, O. Heidbach, A. Helge, B. Hendriks, F. Horváth, E. Horvath, G. A. Houseman, H. Igel, P. Japsen, A. G. Jones, J. Kiss, V. Klemann, J. Kuhlemann, A. Lankreijer, E. Larsen, J. Lauterjung, O. Longva, J. Ludden, E. Lundin, A. Madarasi, P. Maguire, L. Matenco, G. Molnar, J. Mosar, A. Nador, J. Negendank, F. Neubauer, A. Novak, O. Olesen, O. Oncken, P. T. Osmundsen, C. Pascal, S. Sliaupa, G. Peters, C. Puigdefabregas, L. Rise, Z. Ruszkiczay, M. Sacchi, S. Schmid, M. Smelror, A. Soesoo, W. Spakman, W. Stackebrandt, K. Stalsberg, R. Steel, B. Steinberger, R. A. Stephenson, C. Szabo, P. Szafian, L. Szarka, B. Szekely, O. Sztano, M. Tesauro, A. Thieken, G. Timar, H. Thybo, T. Torsvik, L. Toth, J. van Enst, J. D. van Wees, G. Varga, G. de Vicente, Z. Weber, F. Wenzel, M. Wilson, D. Wolf, M. J. R. Wortel, P. A. Ziegler.

portals, and shared middleware and software libraries, collectively known as *cyberinfrastructure*.

Cyberinfrastructure forms a comprehensive modeling environment to integrate computing hardware, data, networks, digitally enabled sensors, observatories, and experimental facilities. Its sophistication presents new challenges and opportunities, which we group here along two trends. One arises from the growing complexity of computational hardware, which now includes a variety of computing systems with complementary performance capabilities. For example, the petascale computing environment is likely to consist of: (1) a significant number of systems with peak performance in the 1–50 teraflops range, deployed and supported at the local level by individual research groups; (2) a few mid-tier systems with peak performance of 100+ teraflops, and (3) at least one system in the 1–10 petaflops range at the European level to support a limited number of projects at the highest level of computing performance. To operate effectively within this environment means that one must pool computer resources across heterogeneous networks, as envisioned by the *Distributed European Infrastructure for Supercomputing Applications* (DEISA), which necessitates the adaptation, benchmarking, verification, and standardization of central modeling tools to run in a stable fashion across multiple platforms. Such central tools hold the promise to greatly reduce the time to solution of computationally intensive research problems by permitting local development of research codes that one then transfers quickly to production environments.

Cyberinfrastructure is also a key component of TOPO-EUROPE, which is a multidisciplinary European collaborative research program that addresses the interaction of processes inherent to the deep Earth (lithosphere, mantle) with surface processes (erosion, climate, sea level), which together shaped the topography of Europe (Cloetingh *et al.*, 2007). Topography influences various aspects of society, not only in terms of the slow process of landscape evolution, but also through climate change, for instance, due to mountain building and changes in sea-water levels and ocean currents. TOPO-EUROPE integrates monitoring, imaging, reconstruction, and modeling of the interplay between processes controlling continental topography and related natural geohazards. TOPO-EUROPE will make a strong effort to develop robust common software libraries, so that the European cyberinfrastructure can be used with high efficiency.

The other trend to shape earth system modeling profoundly in the coming years is the explosive growth in remote sensing capabilities with digitally enabled sensors. This trend is already transforming the earth sciences from a state of data starvation to a state rich in data. Prominent examples are the ambitious new data-acquisition schemes now underway, such as *Euro-Array* or the *CHAMP*, *GRACE* and *ØRSTEDT* as well as the future *GOCE* and *SWARM* geophysical satellite missions. Together they will provide crucial insight into the European seismic, magnetic, gravimetric, and geodetic fields. But each day they will also generate terabytes of observational data that need to be visualized, processed, distributed, and archived in an efficient manner. The net effect of this new data abundance will be the promotion of assimilation and 4-D integrated modeling schemes in the form of sophisticated sequential and variational techniques. These will enable essential fore- and hind-cast capabilities, allowing us to far better understand the temporal (4-D) evolution of complex geologic systems. But before we address challenges of earth system modeling, let us take a brief look at our planet.

## 19.2 Deep Earth: Key for understanding the surface record

The structure and processes of the deep Earth may sound remote from everyday concerns, but both have strong relevance for humanity's basic needs, such as supply of *water* and *resources*, protection against *natural hazards*, and control of the *environmental degradation* of the Earth. The answer to these long-standing questions depends in major ways upon a better knowledge of the mass transfer at Earth's surface, and its feedback with deep earth recycling, and they form the core of the TOPO-EUROPE research effort. The Earth's interior is complex, consisting of three distinct regions. Starting from the outside there is first the cold lithosphere, which is dominated by brittle behavior. Then follows the solid mantle, which deforms slowly over geologic time by a mechanism known as ductile creep. Finally, near the Earth's center there is the (mostly) liquid core. As a result of convective and other forcings, all three regions are in motion, albeit on different timescales. On the longest timescale, solid-state convection (creep) overturns the mantle once in about every 100–200 million years (Bunge *et al.*, 1998). This overturn is the primary means by which our planet rids itself of primordial and radioactive heat. Tectonic processes operate on shorter timescales, up to a few million years or so. They include rapid variations in plate motions, which are revealed by the recent arrival in the earth sciences of highly accurate space geodesy techniques, such as the global positioning system (GPS) (Dixon, 1991). On still shorter timescales of perhaps 1–1000 years convection of the liquid iron core generates the Earth's magnetic field through a complicated dynamo process that will not concern us here. On a timescale of hours to seconds the core, the mantle, and the lithosphere are traversed by seismic waves, and seismologists are now turning to computer models to study seismic wave propagation through our planet. In recent years we have come to understand the solid Earth in more measurable ("quantitative") ways. Better seismic techniques have brought us to a better understanding of the 3-D structure of the Earth's mantle and lithosphere.

During the past decade, for example, the analysis and understanding of dynamic crust–mantle processes has greatly progressed owing to major advances in the field of seismic tomography at global and regional scales (e.g., Bijwaard *et al.*, 1998).

Tomographic imaging techniques are applied to observations of body and surface waves and provide spectacular images of mantle structures. These images can readily be linked to global plate-tectonic processes, such as past and active subduction of lithosperic plates (Bunge and Grand, 2000; Fukao *et al.*, 2001; Sigloch *et al.*, 2008; Wortel and Spakman, 2000;). Thus, we can describe, in numerical terms, how the deep earth system works. However, at the same time, quantitative analysis of the basins in which sediments accumulate has allowed us to connect the *deep earth system* with the record of those changes written in the sediments that build up over geological time.

## 19.3 Continental topography

Continental topography is at the interface of deep earth, surface, and atmospheric processes (Figure 19.1). Topography influences society, not only as a result of slow landscape changes but also in terms of how it impacts geohazards and the environment. When sea-, lake-, or ground-water levels rise, or land subsides, the risk of flooding increases, directly affecting the sustainability of local ecosystems and human habitats. On the other hand, declining water levels and uplifting land may lead to higher risk of erosion and desertification. In the recent past, catastrophic landslides and rock falls have caused heavy damage and numerous fatalities in Europe. Rapid population growth in river basins, coastal lowlands and mountainous regions and global warming, associated with increasingly frequent exceptional weather events, are likely to exacerbate the risk of flooding and devastating rock failures. Along active deformation zones, earthquakes and volcanic eruptions cause short-term and localized topography changes. These changes may present additional hazards, but at the same time permit quantification of stress and strain

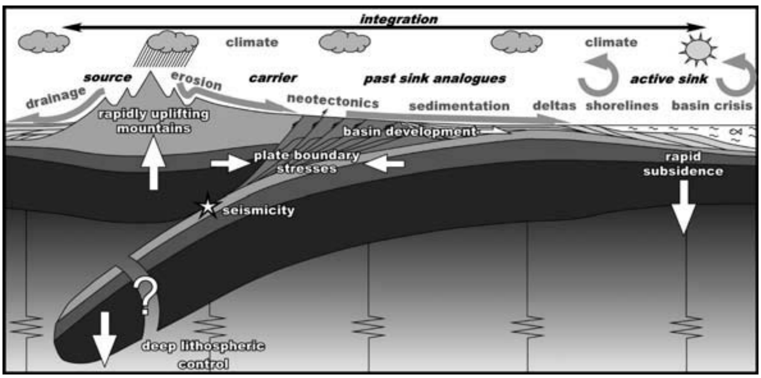



Figure 19.1. Shaping of continental topography as a result of the interaction of processes inherent to the deep Earth (lithosphere, mantle) with surface processes (erosion, climate, sea level). After Cloetingh *et al.* (2007).

accumulation, a key control for seismic and volcanic hazard assessment. Although natural processes and human activities cause geohazards and environmental changes, the relative contribution of the respective components is still poorly understood. That topography influences climate is known since the beginning of civilization, but it is only recently that we are able to model its effects in regions where good (paleo-) topographic and climatologic data are available.

The present state and behavior of the shallow earth system is a consequence of processes operating on a wide range of timescales. These include the long-term effects of tectonic uplift and subsidence, seismicity and the development of river systems, residual effects of the ice ages on crustal movement, natural climate and environmental changes over the last millennia and up to the present, and the powerful anthropogenic impacts of the last century. If we are to understand the present state of the earth system, to predict its future and to engineer our use of it, this spectrum of processes, operating concurrently but on different timescales, needs to be better understood. The challenge to geosciences is to describe the state of the system, to monitor its changes, to forecast its evolution and, in collaboration with others, to evaluate modes of its sustainable use by human society.

Using space satellites to survey the Earth has allowed us to obtain ever-higher resolution when monitoring the vertical motions of Earth's surface. Modeling the way topography changes with time has now reached the stage where it is possible to couple studies of sediment supply and erosion in time and space. At a much smaller scale, we face problems of sedimentary architecture (the way different sediments are structured), and of imaging this architecture using *remote sensing techniques* that use *seismic or electromagnetic waves* to see inside them, like a "body scanner."

Despite enormous progress in the last 15 years, such remote imaging barely keeps pace with the great demands society places upon it, with urgent needs for water supplies, mineral resources, protection against natural hazards, and control of the environment.

## 19.4 Topography and natural hazards

To gain a better understanding of the interrelation between topography, geohazards, and the environment, the temporal evolution of topography needs to be assessed, not only during the recent past but also during the last 10 or so million years. There are, however, some complex problems inherent to paleotopography analysis. Apart from dealing with topography that no longer exists, the dimensions and timing of events and the underlying dynamic processes that controlled its development, as well as the topographic life cycle, pose major challenges, the complexity of which cannot be solved by a single subdiscipline but requires support by other disciplines. The geographic scope of the TOPO-EUROPE programme demands cooperation on a European scale to avoid a fragmented approach. Mountain ranges (increasing surface topography) and adjacent sedimentary basins (decreasing surface topography) record signals and proxies that tell the story of the topographic life cycle. In this, the source-to-sink relationship is of key importance. However, signals and proxies are still poorly understood and we have only started to decipher the few of which we are aware. A major challenge is to extract all available information contained in the system and to interpret it in terms of processes. Innovative analytical techniques, improvement of methodologies, back to back with innovative conceptual and quantitative modeling, are required to resolve these problems.

The main challenge in topography-related geological hazard research is to create and verify physical models of hazardous earth systems that integrate all relevant data, describe hazards as a function of time, and understand them as resulting from a nonlinear system evolution under which processes acting on various temporal and spatial scales can become catastrophic. In this context it must be understood that topography plays a prominent role, as it results from the interaction of shallow and deep earth processes, and as such permits us, in combination with other parameters, to assess the state of stress and its change through time.

There are obvious relations between geological hazards and topography. Topography is a major factor controlling slope instabilities, which can lead to the development of landslides, both on- and offshore. For example, uplift of Fennoscandia and the Romanian Vrancea area (Figure 19.2) has caused increased landslide and rock-fall hazards. The second important parameter for catastrophic earth movements is the internal friction of soil, which in turn depends largely on the hydrological conditions and water input by precipitation. Regional climate changes when associated with a precipitation increase tend to cause increased slope instability and corresponding landslide activity.

Earthquakes result from crustal-scale fault-related deformation and occur in various parts of Europe (Figure 19.2). Although areas with a high frequency of large magnitude earthquakes are mostly bound to the Mediterranean domain, the strong concentration of people and high-value infrastructure in densely populated areas in Europe can turn, in these regions, moderate hazards into large risks. The currently used "3rd-generation" hazard assessment method can be coined "seismotectonic probabilism." This method largely relies on historical and paleoseismological earthquake records, and results in maps giving an annual exceedance probability of a certain damage parameter. The challenge to solid earth science researchers lies in developing 4th-generation hazard assessment methods, relying much more on a physical understanding of processes leading to earthquakes and on assessment of the actual state of stress on faults. The state of stress is strongly influenced by surface topography, but also by the topography of lithospheric boundaries (Moho, lithosphere–asthenosphere boundary). Highly sophisticated

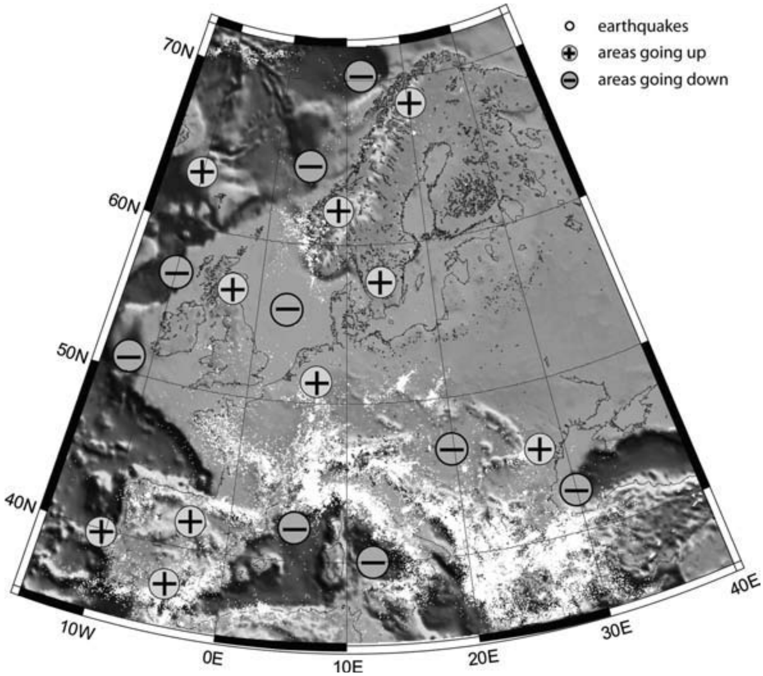



Figure 19.2. Present-day active intraplate deformation in Europe. White dots indicate location of earthquake epicenters. Also shown are intraplate areas of Late Neogene uplift (circles with plus symbols) and subsidence (circles with minus symbols). Background elevation images are extracted from the ETOPO2 dataset. See color plates section.

models for time-dependent hazard assessment that link several processes, such as mantle dynamics, structure and rheology of the crust and mantle, change in topography, mass redistribution by erosion and sedimentation, and post-seismic relaxation, can be established today. Verification of these models requires data on recent deformation (both from GPS and geological reconstruction for the Holocene/Pleistocene) and tectonic stress (e.g., through the World Stress Map project). This can lead yield substantial new insights into the stress and strain evolution of the key seismogenic areas of Europe. This type of modeling may prove a particularly valuable approach to constrain extreme events with their high societal impact.

Intraplate seismicity is still poorly understood and tends to follow episodic intermittency patterns rather than quasi-periodic earthquake activity more characteristic for plate boundaries. TOPO-EUROPE will establish a database that allows for a systematic combination of lithospheric data (e.g., geometry of boundaries, temperature, stress, structure) and recent movements, including topography changes over an area that covers all levels of seismicity, such as highly active plate boundary domains, moderate intraplate activity, and seismic quiescence.

Europe is exposed to recurrent flooding events that pose major hazards to population and industrial agglomerations. The damaging potential of floods is intrinsically linked to even minor topographic changes that control the depth of inundation. Thus, it will be a challenging task for TOPO-EUROPE to combine regional climate predictions with changes in sea and river level and topography during the Holocene to fully assess Europe's future flood hazards.

The main risk-generating factor for human society is the increased exposure and vulnerability of its assets (buildings, infrastructure, and social systems). However, during the past 150 years, anthropogenic modification of the planetary environment has caused changes in the hazard potential itself. For instance, extraction of large amounts of ground water beneath and near cities modifies surface elevations and thus their inundation potential during floods. At the same time, this impacts on the stability of the subsurface with consequences for ground motion during future earthquakes, associated liquefaction potential, and landslide hazard. Again, TOPO-EUROPE opens avenues to address systematically these issues on a European scale.

## 19.5 The TOPO-EUROPE network

The TOPO-EUROPE network was officially launched during a symposium held in October 2005 in Heidelberg, Germany with the objective to tackle the challenges in continental topography research. TOPO-EUROPE aims at integrating European communities that hitherto have been active under discrete research and implementation schemes in the field of continental topography research. Subjects that have been addressed include land subsidence and uplift, and fluctuations of the erosional base level in response to sea-level changes and local tectonics.

With the establishment of a strong network of collaborating institutes, the international TOPO-EUROPE project will be able to tackle a set of outstanding questions pertaining to lithospheric, surface, and climate-related processes controlling the ongoing topography evolution and related natural hazards of the Alps/Carpathians–Pannonian Basin System, the West and Central European Platform, the Apennines–Aegean–Anatolian region, the Iberian Peninsula, the Scandinavian Atlantic Margin, the East-European Platform and the Caucasus–Levant area. These natural laboratories comprise some of the best-documented orogens, sedimentary basins, and continental margins worldwide. As such, they offer key study areas for the development of a new generation of models for ongoing lithospheric deformation and its effect on continental topography development, both on regional and local scales.

Research will focus on the interplay between active tectonics, topography evolution, and related sea-level changes and drainage pattern development (e.g., Cloetingh and Cornu, 2005; Cloetingh *et al.*, 2006). This includes the development of an integrated observation and analysis strategy, focusing on large-scale changes in vulnerable areas of Europe. Geoprediction in poly-phase deformed and tectonically active orogenic systems requires multidisciplinary efforts and, therefore, the interaction and collaboration of researchers covering a broad field of expertise. Among other eminent scientific disciplines, geology, geophysics, geodesy, hydrology, climatology, as well as various fields of geotechnology will be integrated. TOPO-EUROPE will address several scientific issues of key relevance, such as (1) the 4-D development of Europe's topography, (2) the quantification of source-to-sink relations to quantify sediment budgets, (3) the quantification of land subsidence in the basins and deltas of Europe, (4) the quantification of land uplift in orogenic and intraplate domains, (5) the quantification of tectonically controlled river evolution, and (6) the effects of climate changes.

Below we give an overview on continental topography research, its future challenges, and expected breakthroughs. We first summarize old and new methods and techniques available to TOPO-EUROPE researchers. We also discuss the state of research in some of the "natural laboratories" that have been selected as targets for TOPO-EUROPE research during the next 10 years. Subsequently, the TOPO-EUROPE science plan is presented, including specific targets and expected deliverables of the programme in the years to come.

## 19.6 The natural laboratory concept: From orogen through platform to continental margin

The TOPO-EUROPE network provides a discussion forum for a multidisciplinary research programme, which functions in a feedback mode between advancement of new numerical modeling concepts and their validation by an array of geological and geophysical datasets from a number of natural laboratories in Europe. To this purpose the network concentrates on well-documented regions, each of which is optimally suited to address the coupling between tectonic (endogenic) and surface (exogenic) processes and the related effects on topography development and inherent geohazards.

TOPO-EUROPE upscales the expertise acquired during the ESF EUROPROBE programme through the integration of the above-described components as the fundamental approach to provide conditions for closing the loop between observation, reconstruction, and process-oriented modeling. TOPO-EUROPE integrates geology, geophysics, geodesy and geotechnology and provides the frame for intense cross-fertilization between these disciplines. By working together in a concerted effort on common datasets, an optimal dissemination of results will be achieved. An example of an integration of such datasets is provided by the new crustal thickness map of Europe recently published (Tesauro *et al.*, 2008; open source ftp://ftp.agu.org/apend/gl/2007gl 0322 44). This map (Figure 19.3a) is based on all the presently

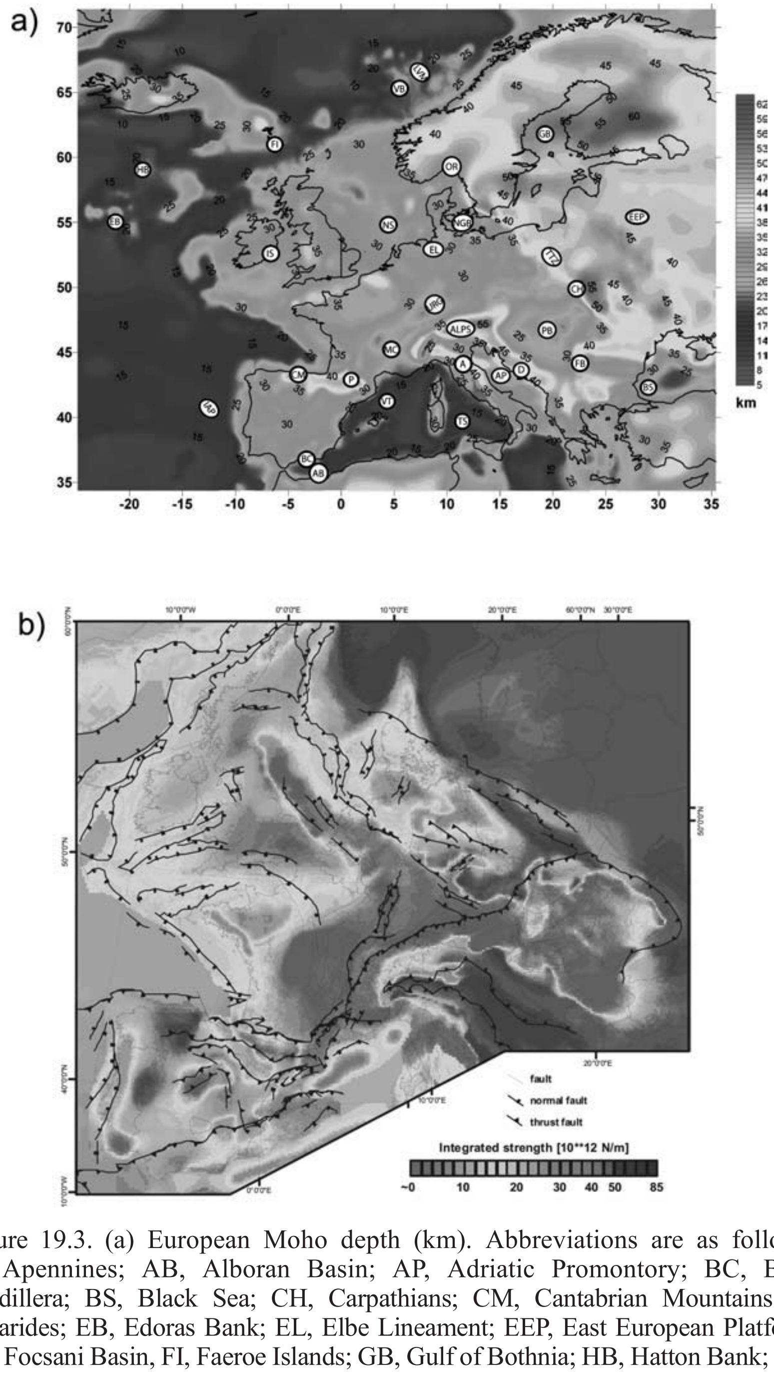



Figure 19.3. (a) European Moho depth (km). Abbreviations are as follows: A, Apennines; AB, Alboran Basin; AP, Adriatic Promontory; BC, Betic Cordillera; BS, Black Sea; CH, Carpathians; CM, Cantabrian Mountains; D, Dinarides; EB, Edoras Bank; EL, Elbe Lineament; EEP, East European Platform; FB, Focsani Basin, FI, Faeroe Islands; GB, Gulf of Bothnia; HB, Hatton Bank; IAP, Iberian Abyssal Plain; IS, Iapetus Suture; LVM, Lofoten-Vesterålen margin; MC, Massif Central; NGB, North German Basin; NS, North Sea; OR, Oslo Rift; P, Pyrenees; PB, Pannonian Basin; TS, Tyrrhenian Sea; TTZ, Tesseyre-Tornquist

available seismic reflection, refraction and receiver function data, providing an update of the earlier compilation of Dèzes and Ziegler (2004). With this information, residual mantle anomalies of the gravity field measured by satellite missions can be observed after removal of the crustal effect from the observed field (Tesauro *et al.*, 2007). The new crustal model also served in the construction of the first integrated strength map for intraplate Europe (Figure 19.3b) (Cloetingh *et al.*, 2005).

Integration of datasets and data handling is vital to the efficient transmission of findings through the above-mentioned chain of components. In TOPO-EUROPE this can be achieved via a number of connected implementation steps centered on three key cells, namely: (1) The creation of new think tanks for the development and implementation of new conceptual approaches and testing of their viability against geological and geophysical data from selected natural laboratories; (2) the creation of new earth system teams working jointly on unexplored interfaces between existing research activities; and (3) building of information technology cells to optimize integrated data handling, interdisciplinary modeling, and software integration.

The TOPO-EUROPE integrative research program is centred on critical regional and continental-scale earth science problems in carefully selected natural laboratories that cover a wide range of geodynamic settings and geohazard provinces, for each of which extensive databases are available. Together these natural laboratories provide a set of world-class opportunities to probe and quantify the entire range of plate interaction processes affecting topography in the context of presently active geological processes. In these areas we can obtain the highest-possible resolution required to discriminate between endogenic and exogenic earth processes required to quantify the coupling between solid earth and surface processes. TOPO-EUROPE's natural laboratories discussed below offer unique key study areas for developing a new generation of models explaining ongoing deformation of the lithosphere and its repercussions on continental topography and the human habitat. Other natural laboratories may be selected as the project advances. These will be chosen based on their merits. TOPO-EUROPE operates in an iterative manner with

---

Caption for Figure 19.3. (*cont.*)
zone; URG, Upper Rhine Graben; VB, Vøring Basin; VT, Valencia Trough. From Tesauro *et al.* (2008).

(b) Integrated strength map for intraplate Europe (after Cloetingh *et al.*, 2005). Colors represent the integrated compressional strength of the total lithosphere. Adopted composition for upper crust, lower crust, and mantle is based on a wet quartzite, diorite, and dry olivine composition, respectively. Rheological rock parameters are from Carter and Tsenn (1987). The adopted bulk strain-rate is $10^{-16}\,s^{-1}$. Main structural features are from Ziegler (1988) and Dèzes *et al.* (2004). See color plates section.

initial models being developed to explain existing datasets and concepts. In parallel, new higher-resolution data will be acquired in a number of carefully selected European natural laboratories. The derived numerical models will then be tested and refined on the basis of the new data.

An important step has been the selection of TOPO-EUROPE in early 2008 by the European Science Foundation (ESF) as one of its large-scale European collaborative research initiatives (EUROCORES). In response to the ESF call for proposals, 42 outline proposals were submitted, resulting in 22 full proposals submitted for international peer review. Out of these, the following ten collaborative research projects (CRPs) were selected for funding with a total amount of 13 million euros for the ESF EUROCORES TOPO-EUROPE, opening up research positions for more than 50 Ph.D. students and postdocs:

- **VAMP**: "Vertical Anatolia Movements Project" (coordinator: Manfred Strecker – University of Potsdam, Germany)
- **TOPO-4D**: "Mantle forcing of Earth surface evolution in Europe and the Mediterranean: From Past to Present" (coordinator: Wim Spakman – Utrecht University, The Netherlands)
- **TOPOALPS**: "The Topographic History of the Alps and its Tectonic and Climatic Drivers" (coordinator: Sean Willett – ETH Zürich, Switzerland)
- **Thermo-Europe**: "Coupled climatic/tectonic forcing of European topography revealed through thermochronometry" (coordinator: Peter van der Beek – University of Grenoble, France)
- **RESEL-GRACE**: "Refined European sea-level estimations by combining altimetry, tide gages, hydrographic and other datasets with improved regional GIA modeling and tailored regional GRACE gravity field models" (coordinator: Juergen Kusche – University of Bonn, Germany)
- **TopoMed**: "Plate reorganization in the western Mediterranean: Lithospheric causes and topographic consequences" (coordinator: Marinus Wortel – Utrecht University, The Netherlands)
- **PYRTEC**: "Spatial and temporal coupling between tectonics and surface processes during lithosphere inversion of the Pyrenean–Cantabrian mountain belt" (coordinator: Ritske Huismans – University of Bergen, Norway)
- **TopoScandiaDeep**: "Searching for the origin of the topography of the Scandes" (coordinator: Valerie Maupin – University of Oslo, Norway)
- **SedyMONT**: "Time-scales of sediment dynamics and topographic change in mountain landscapes" (coordinator: Fritz Schlunegger – University of Bern, Switzerland)
- **SourceSink**: "Integrated natural hazard assessment through the quantification of mass transfer from mountain ranges to active sedimentary basins" (coordinator: Paul Andriessen – VU University Amsterdam, The Netherlands)

## 19.7 High-performance computing and virtual earth models

The speed of microprocessors increased dramatically the past two decades – from about 1 Mflops on SUN Microsystem's venerable SPARC 1 in the early 1990s to about 10 Gflops on current INTEL or AMD processors – with concurrent dramatic gains in price performance. And with computer hardware performance growing exponentially (gate density doubling every 18 months, storage capacity every 12 months, and network capability every 9 months) we are now entering an era with heterogeneous "multi-core" architectures that promises further gains in computer speed and cost. This progress has eroded the distinction between computers made from low-cost microprocessors and more conventional supercomputers operated at national centers.

It is also promoting a new balance in high-performance computing (HPC), whereby powerful local computer facilities, exploiting an approach to parallel computation known as clustered computing, add a viable mid-tier distributed level to more traditional computing centers. The mid-tier systems deliver superior price/performance and are well suited to supply the compute capacity required for many earth simulations. Moreover, their user communities tend to have well-defined computing needs centered around key applications, such as climate modeling, tectonic and seismic calculations, or *ab initio* material simulations. In other words, their computing needs are topical. A tradition of topical computing exists in the earth sciences in the form of NCAR (National Center for Atmospheric Research) or the DKRZ (Deutsches Klimarechenzentrum). But the approach can now be brought to models of the solid Earth, where computing needs rival those of climate models. Compute power is now, in principle, sufficient to model the full plate tectonic/mantle convection system, promising breakthroughs in understanding the coupled solid earth system. Further computing growth should allow us also to model seismic frequencies of one Hz (or less) globally in numerical wave propagation codes in the coming years.

Topical computing can realize unprecedented gains in capacity computing, but requires a paradigm shift in the approach to "CPU-rich" applications and a qualitatively a new level of cooperation with computational science. In TOPO-EUROPE, we envision a number of mid-tier distributed facilities. These systems will be on fast networks, configured with grid services, and will provide the processing, data-archiving and visualization backbone for the expected explosive growth in high-end computing as sophisticated simulation codes and remote sensors transition to web services. They will also serve as general-purpose mid-range compute platforms for jobs too large for local resources, and we anticipate that national and European compute centers will host some of these systems.

Software development is another major issue, as the age of undocumented academic codes specific to one or two research groups is passing. This includes advanced visualization, which is indispensable in computational earth sciences.

Datasets are enormous, and one must manipulate, display, and explore them efficiently. We envision powerful distributed visualization servers to produce bitmaps from complex data and ship these on the Web to the end user interactively. The development of documented and well-structured codes, libraries and visualization tools requires long-term funding. The recently funded NSF program in Computational Infrastructure for Geodynamics (CIG; www.cig.org) is a start in this direction and will serve as a template for TOPO-EUROPE.

Several classes of advanced computational earth models are available to TOPO-EUROPE. One involves the process of mantle convection recognized now as the driving mechanism of plate tectonics. Mantle convection acts over multiple scales because of the many nonlinearities in the system arising from rheology, or from compositional and thermal gradients (Tackley, 2000). Grand-challenge problems include (1) spherical 3-D convection with earthlike Rayleigh number, (2) the incorporation of constraints from mineral physics models, and (3) nonlinear rheologies and thermochemical convection. Progress in mineral physics computer simulations (e.g., *ab initio* calculations, molecular dynamics, and thermodynamics) is essential for understanding material properties for the exotic conditions of the deeper Earth and the relation to geodynamic processes (Piazzoni *et al.*, 2007). Coupling the mineral physics models to large-scale mantle flow models will facilitate their explicit comparison to tomograpy. Several high-resolution 3-D spherical mantle convection codes that run efficiently on clusters are now available (TERRA: Bunge *et al.*, 1996; *CitcomS*: Zhong *et al.*, 2000). Combining these models with reconstructions of past plate motion (e.g., Lithgow-Bertelloni and Richards, 1998) makes it possible to infer the large-scale circulation and the associated buoyancy structure of the mantle (Bunge *et al.,* 1998; 2002). This is essential when trying to assess the plate tectonic force balance and the large-scale tectonic stress field responsible for horizontal and vertical movement. A representative MCM at high numerical resolution (about 100 million grid points) is shown in Figure 19.4a. While the current MCM generation compares reasonably well with mantle tomograms (e.g., Schuberth *et al.*, 2009), they suffer in a fundamental way from lack of initial condition information, a problem shared with oceanic and atmospheric circulation modeling, as we will detail below.

Computational seismology is also at the forefront of HPC. The grand-challenge problem here is to enhance the quality of tomograms of the Earth's interior in conjunction with improving models of the rupture process during an earthquake. Although, in principle, one cancalculate the dependence of seismograms to earth structure numerically, the computational cost of doing so is prohibitive. Drawing connections among waveform tomography and adjoint methods is finally opening the way to solving the full 3-D seismic inverse problem (e.g., Fichtner *et al.*, 2006a, b; Tromp *et al.*, 2005). Moreover, it is possible to use high-resolution, 3-D models of

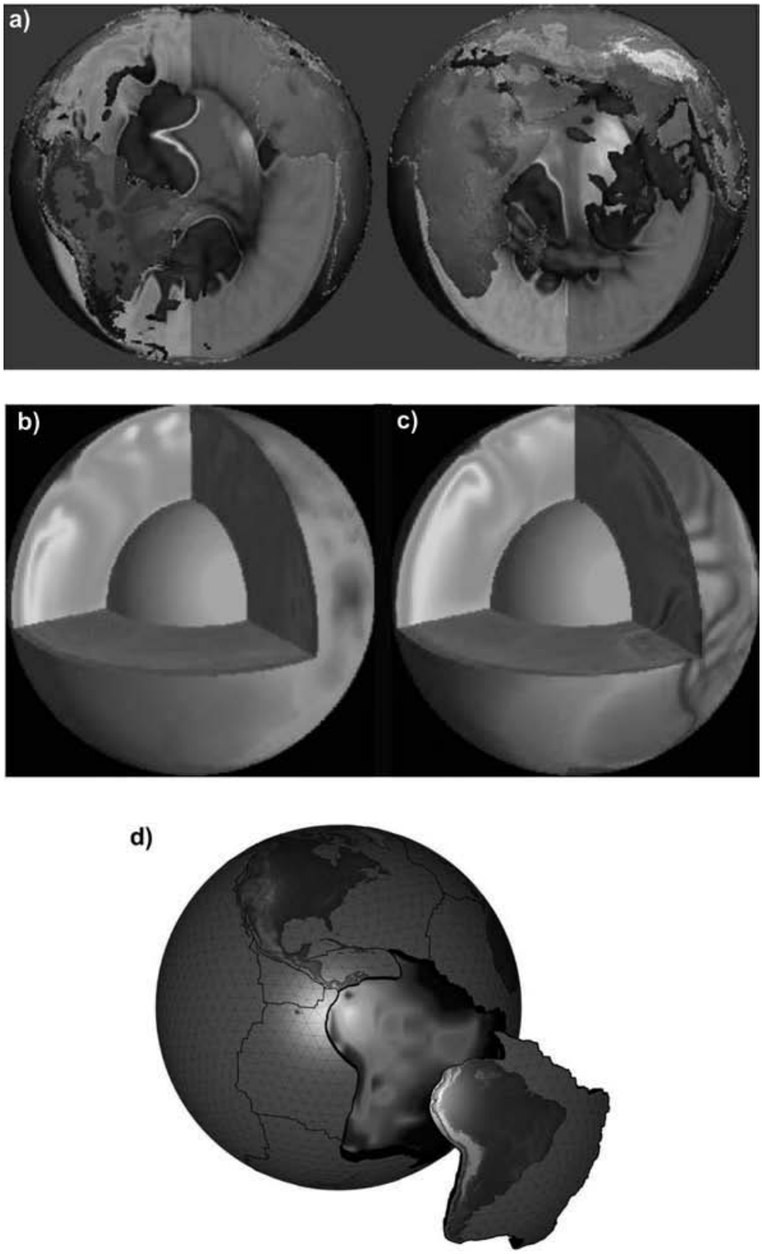



Figure 19.4. (a) Lateral temperature variations in two cross-sectional views from a high-resolution Mantle Circulation Model (MCM). Temperature iso-surfaces are taken to be at −600 and +400 kelvin. Continents are overlain for geographic reference. (b) Cutaway of the initial 3-D temperature field for a reconstruction model with a coupling between mantle convection and plate tectonics. Sample reconstructions using a perturbed mantle circulation model with a coupling of mantle convection and plate tectonics (after Bunge *et al.*, 2003). Blue is cold, red is

sedimentary basins with their associated fault structure (e.g., Plesch *et al.*, 2007) in conjunction with modern numerical methods to predict ground motions that could cause major damage to buildings (e.g., Komatitsch *et al.*, 2004; Olsen *et al.*, 1997). Such scenario simulations won the Gordon Bell Special Achievement prize in 2003 (Akcelik *et al.*, 2003), and are based on the location and size of potential earthquakes and the regional structure of particular seismically active regions. Source processes are also important in connection with peak ground motion. Parameter estimation through theoretical (numerical) simulation of rupture processes combined with wave propagation models promises significant progress (e.g., Brietzke *et al.*, 2004; Igel, 1999; Igel *et al.*, 2002). The feedback between models, experiments and observations is essential to "ground-truth" these complex geosimulations.

A persistent challenge in earth modeling is the realistic treatment of plate tectonics, because it is difficult to simulate shear failure along plate boundaries. One strategy combines strongly temperature-dependent viscosity combined with a plastic yield stress in global mantle convection models (Richards *et al.*, 2001; Trompert and Hansen, 1998). The former causes the cold upper boundary layer (lithosphere) to be strong, while the latter allows the boundary layer to fail locally in regions of high stress. Another approach includes plate boundaries directly into the computational grid in global neo-tectonic models. These models are now at a high level of maturity, allowing them to account for surface topography, regional variations of lithosphere density and thickness, realistic plate configurations and ductile as well as brittle deformation mechanisms (Bird, 1998). When combined with realistic plate driving forces from mantle circulation models it is possible to explicitly assess the force balance in plate tectonics. We show a coupled tectonic mantle convection model in Figures 19.4b and 19.4c. A first-order result of this innovative approach is the recent explanation of short-term plate motion variations offshore of South America (Figure 19.4d). By explicitly coupling MCMs to neo-tectonic models, Iaffaldano *et al.* (2006; 2007) show that the recent topographic growth of the Andes is a key factor controlling the evolution of plate motion and stress orientation (Heidbach *et al.*, 2008) in this region.

From the discussion and examples above, it is clear that simulation is becoming a ubiquitous tool in geosciences. To project the future needs one must consider where computational power is going. Moore's law still works. But new problems, such as

Caption for Figure 19.4. (*cont.*)
hot. The upper 100 km of the mantle is removed to the show the convective planform. (c) Same as (b) but after 100 Myr of present-day plate motion has been imposed. (d) Global neo-tectonic plate model incorporating realistic mantle-related buoyancy driving forces taken from a high-resolution MCM. After Iaffaldano *et al.* (2006; 2007) and Iaffaldano and Bunge (2008). See color plates section.

electrical power usage increasing exponentially, emerge. Greater computing speed can be obtained by increasing the size of clusters and transitioning to multi-core architectures. But power use and heat production still remain a growing concern. An efficient approach will involve funding of mid-tier compute and visualization centers that act as servers for web services to perform computations and analyze and serve results to the users. For example, when an earthquake occurs and appears on a data site, one could request 3-D synthetic seismograms to study local or global structure. In principle, these services require no applications for computer accounts and have little entry barrier. Such grid and visualization services will form a part of the coming earth science computing environment.

Software is as important as hardware. An efficient software environment will include community-based standards, modular-based libraries, cross-platform portability, documentation and system upgrades, with easy access to data storage systems. Within the TOPO-EUROPE effort, we will share, develop, maintain, and document software to form an integrated software infrastructure, which will allow developers and end users to integrate algorithms, scientific expertise, data, and visualization from a wide range of disciplines. Model archiving and metadata management is another key issue for maintaining, comparing, and understanding the output of models and to capture the essential information by which they can be reconstructed (e.g., a full description of the model/modules, input variables, compilers, and machine it was run on). Automatically recording the metadata produced by models is closely related to efforts for cataloging "real" data and metadata. Comparing models to real data requires that the modeling framework interoperates with databases and produces output that shares the same metadata structures as real data. To develop useful standards and interfaces requires significant cooperation between model developers and data groups, such as NERIES, the Network of Research Infrastructures for European Seismology (www.neries-eu.org), *GPlates* (www.gplates.org) and the Southern California Earthquake Center (www.scec.org). As the years have seen unprecedented growth in the amount and quality of geophysical data, grids will provide a convenient service-based interface to these data. For example, already the Southern California Earthquake Center is developing a modern database for use in simulations and modern forecasting technologies. Sensors, computers, data repositories, networks, software, and visualization must be integrated and added in the future to data curation efforts. We expect that the "data flood" will require new algorithms, such as data assimilation schemes, and will drive essential synergies between data mining and the simulation.

The ultimate goal of forward modeling is, of course, to solve the inverse problem and to use available data to constrain solid earth properties that cannot directly be measured. For full-waveform tomography, this problem is nearly solved given new adjoint methods (e.g., Tromp *et al.*, 2008); however, building this functionality into tectonic and geodynamic models will require significant efforts. Important aspects will include generating adjoint operators for tectonic simulations automatically (Iaffaldano *et al.*, 2007) through automatic differentiation methods (www.autodiff.org), ensemble Kalman filtering, and error covariance analysis that are commonly used in data assimilation research. As more-sophisticated geodynamical models become available, we envision that data assimilation will become an important research area in geosciences whereby surface observations are assimilated into model solutions to predict changes in various earth components.

## 19.8 4-D integrated approaches

One of the primary and most innovative objectives of TOPO-EUROPE is to promote 4-D approaches that will lead to integrated interpretations of existing and newly acquired geomorphologic, geologic, geophysical, geodetic, remote sensing, and geotechnologic datasets. A major challenge is the incorporation of different temporal and spatial scales in the analyses of solid earth and surface processes. By combining sophisticated adjoint techniques for data assimilation, the observational constraints of plate motion or deep mantle structure can be exploited to construct time-dependent (4-D) earth models. For MCMs this approach is already underway and suggests that it is possible to overcome the problem of unknown initial conditions. The approach is equivalent to formulating a large-scale fluid dynamic inverse problem. Essentially one seeks optimal initial conditions that minimize, in a weighted least-squares sense, the difference between what an MCM predicts as mantle heterogeneity structure and the heterogeneity one actually infers from tomography. This class of problems is known as *history matching* and is also often referred to as *variational data assimilation*, meaning that model parameters are inferred from a variational principle through the minimization of a cost function. For an example in meteorology, see the review by Talagrand (1997). The necessary condition for this minimum leads to the usual mantle convection equations coupled to a corresponding set of so-called adjoint equations. The adjoint equations, which have been derived recently (Bunge *et al.*, 2003), together with large-scale simulations show that flow can be inferred back in time for at least 100 million years and are nearly identical to the forward model except for forcing terms. Unfortunately, adjoint MCM modeling at realistic convective vigor comes at a heavy computational price. Weeks to months of dedicated integration time are needed to solve this class of problems even on some of the more powerful parallel machines currently in use. Such resources, however, are coming within reach of topical clusters dedicated to capacity computing (Oeser *et al.*, 2006). Examination of the role played by climate, erosion, and tectonics on landscape evolution should provide key constraints for quantifying feedback mechanisms and teleconnections that link the solid earth, active tectonics, and surface processes. Monitoring horizontal and vertical motions of the surface and mapping the subsurface, using modern geophysical, geodetic, remote sensing, and geotechnical techniques, will provide new constraints on present-day deformation patterns and related topographic changes. Analog and numerical modeling, based on these new constraints as input parameters, can be used to test integrated interpretations and to provide information on dynamic processes controlling topography development in intraplate settings and adjacent orogens.

## 19.9 Innovative modeling of mantle-to-lithosphere-to-surface processes

The evolution of surface topography and morphology depends strongly on the interplay of subsurface and surface processes (Figure 19.1). Erosion unloads growing topography whereas sedimentation accelerates basin subsidence (Figure 19.5). This is clearly demonstrated by the strong correlation between denudation and tectonic uplift rates in zones of active deformation. During collision, surface processes contribute towards the localization and growth of mountain belts and fault zones, and ensure stable growth of topography (Avouac and Burov, 1996). During crustal extension, syn-rift erosion contributes towards widening of the rifted basin, so that apparent extension coefficients can increase by a factor of 1.5–2 (Burov and Cloetingh, 1997; Burov and Poliakov, 2001). Poly-phase subsidence and other deviations from thermal subsidence models can be also controlled by feedback between surface and subsurface deformation. Similarly, recent work

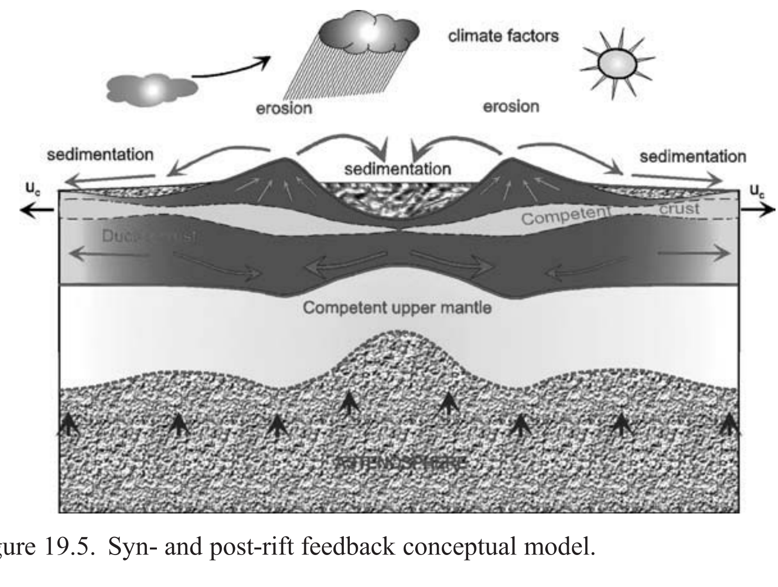



Figure 19.5. Syn- and post-rift feedback conceptual model.

(Burov, 2010; Burov and Cloetingh, 2009) has demonstrated that the surface expression of lithospheric folding (Cloetingh *et al.*, 1999) and mantle plumes (Burov *et al.*, 2007) is strongly controlled by the stratified rheology of the continental lithosphere and by the interplay with surface processes.

A new generation of 3-D and 4-D tectonically realistic models is required for an understanding of dynamic feedbacks between tectonic and surface processes (e.g., Cloetingh and Negendank, 2010), providing new insights into the evolution of tectonically active systems and related surface topography:

- Morphologically and tectonically consistent collision and exhumation models
- Basin modeling, synthetic stratigraphy
- Climate-coupled modeling

The topographic reaction to surface loading and unloading depends on the mechanical strength of the lithosphere, as well as on the strength partitioning between the crust and lithospheric mantle. Consequently, testing different rheological profiles in areas where the data on denudation/sedimentation rates are well constrained may provide new possibilities for constraining the long-term rheology of the lithosphere (Figure 19.3b) (e.g., Burov and Watts, 2006).

Reliable information on (de)coupling processes at the crust–mantle and lithosphere–asthenosphere boundaries and at the two principal phase transitions within the deeper mantle (at about 410 and 660 km depth) will be of fundamental importance for modeling surface topography. The quantification of dynamic depth-to-surface relationships is a major challenge that requires innovative approaches to 4-D modeling. The principles of available conventional fluid-dynamic modeling are robust, but they require greatly increased computer power to provide adequate resolution of a convection system characterized by thermal boundary layers, slabs, and plumes of complex structure that may evolve rapidly. New approaches need to incorporate (1) elastoplastic rheologies for crustal and mantle materials, (2) an integrated modeling approach of material flow and elastic deformation (also crucial for predicting realistic topography evolution), and (3) crustal and lithospheric weakness zones and/or faults. Accounting for elastic and plastic deformation may actually require modifying available large-scale mantle dynamics models to solve, at least for the lithospheric part, full stress equations with free upper surface boundary conditions instead of flow approximations. Mantle models need to be constrained by mantle tomography, geodetic, and electromagnetic data. The latest geomodeling tools are able to consistently treat homogeneous and inhomogeneous deformation with realistic faults, so that the magnitude of uplift, subsidence, fluid flow, and other types of deformation (derived from geological markers or GPS, stress in boreholes and earthquakes) can be linked and interpreted quantitatively. The goal of 4-D modeling is to quantify the dynamic evolution of solid earth boundaries and phase transitions and associated surface deformation, and to define the present state of surface deformation, including its space-time gradient (a prerequisite for geoprediction). To achieve this goal, very high resolution at temporal and spatial scales (e.g., 50–100 years, 5–10 km) is required.

## 19.10 Synergy between analog and numerical modeling

Novel tectonic modeling concepts and their implementation in numerical modeling software provide new opportunities for quantifying the interplay between stresses and rheology during deformation of the lithosphere. Computer simulations will focus on the links between mountain-forming and basin-forming processes, basin geometries, and vertical motions in space and time. Furthermore, thermomechanical numerical modeling schemes, accounting for the physics of strain localization in the lithosphere and its consequence for poly-phase deformation and associated vertical motions, can be designed and implemented.

Analog modeling (e.g., Smit *et al.*, 2008; Sokoutis *et al.*, 2005; Sokoutis *et al.*, 2007) will provide independent validation of numerical models and will be particularly useful in complex settings, such as those with pronounced 3-D geometries (e.g., strike-slip systems and compressional mountain belts). Various scales can be handled: shallow to deep, local to regional with advantages of analog modeling in terms of complexity and proximity to geological observations and advantages of numerical modeling in terms of physical clarity and higher potential for sensitivity studies and parameter variation (Figure 19.6) (Garcia-Castellanos *et al.*, 2003). In analog experiments, geomechanical boundary conditions and material properties will be dynamically scaled to simulate lithospheric conditions.

With respect to modeling techniques, ever-faster computer systems and ever-larger datasets have resulted in vast improvements in modeling capabilities. A transition to true 4-D modeling has been achieved during the last years. However, modeling approaches to different problems are still developed on an ad hoc basis.

## 19.11 Conclusions

The modern earth system approach requires a comprehensive integration of existing databases, with the capacity to expand to allow for storage and exchange of new data collected during the growth of the TOPO-EUROPE programme. Unification and coupling of existing modeling techniques is needed to achieve full integration of what are currently discipline-oriented approaches and to expand fully the next generation of 4-D numerical applications. Furthermore, flexible

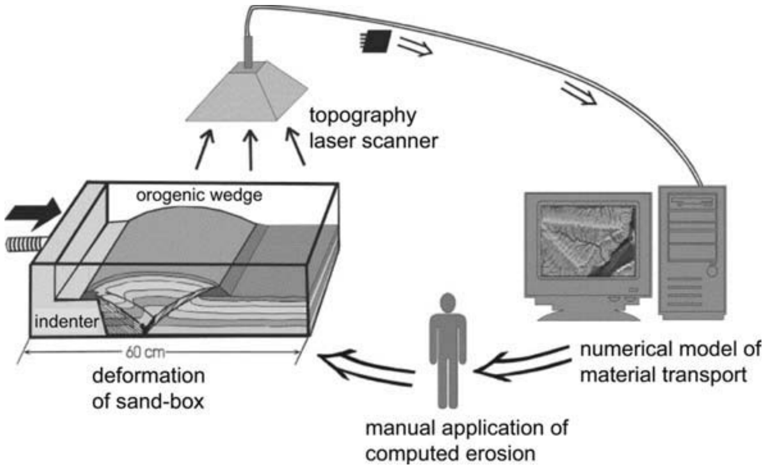



Figure 19.6. Schematic diagram of a coupled analog-numerical modeling system (Persson *et al.*, 2004). The analog modeling facilities are used to simulate upper crustal deformation and its dynamic response to surface erosion and sedimentation predicted by the developed numerical models. The scanner is used to transmit the surface topography of the analog model to the numerical model. The displayed image shows the drainage system of the Ebro River (NE Spain) (Garcia-Castellanos *et al.*, 2003). Subsequently, the calculated erosion/deposition is manually applied to the analog model.

exchange for feedback loops in the quantification of earth processes is needed between databases and modeling tools. Consequently, further implementation of state-of-the-art cyberinfrastructure in modern integrated earth science is of major importance.

## References

Akcelik, V., Bielak, J., Biros, G. *et al*. (2003). High-resolution forward and inverse earthquake modelling on terascale computers. *Proceedings of the Supercomputing SC2003 Conference*, www.sc-conference.org/sc2003.

Avouac, J. P. and Burov, E. B. (1996). Erosion as a driving mechanism of intracontinental mountain growth. *Journal of Geophysical Research – Solid Earth*, **101**(B8): 17 747–17 769.

Bijwaard, H., Spakman, W., and Engdahl, E. R. (1998). Closing the gap between regional and global travel time tomography. *Journal of Geophysical Research*, **103**: 30 055–30 078.

Bird, P. (1998). Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. *Journal of Geophysical Research*, **103**: 10 115–10 129.

Brietzke, G., Igel, H., Jahnke, G. *et al*. (2004). Computational elastic wave propagation: Advances in global and regional seismology. In *High Performance Computing in Science and Engineering*. Munich 2004 Berlin: Springer, pp. 445–458.

Bunge, H.-P. and Grand, S. P. (2000). Mesozoic plate-motion history below the northeast Pacific Ocean from seismic images of the subducted Farallon slab. *Nature*, **405**: 337–340.

Bunge, H.-P., Hagelberg, C. R., and Travis, B. J. (2003). Mantle circulation models with variational data assimilation: Inferring past mantle flow and structure from plate motion histories and seismic tomography. *Geophysical Journal International*, **152**: 280–301.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R. (1996). The effect of depth dependent viscosity on the planform of mantle convection. *Nature*, **379**: 436–438.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R. (2002). Mantle-circulation models with sequential data assimilation: Inferring present-day mantle structure from plate-motion histories. *Royal Society of London Philosophical Transactions*, **A 360**: 2545–2567.

Bunge, H.-P., Richards, M. A., Lithgow-Bertelloni, C. *et al*. (1998). Time scales and heterogeneous structure in geodynamic earth models. *Science*, **280**(5360): 91–95.

Burov, E. (2010). Thermo-mechanical models for coupled lithosphere-surface processes: Applications to continental convergence and mountain building processes. In *New Frontiers in Integrated Solid Earth Sciences*, ed. S. Cloetingh and J. Negendank. New York: Springer Verlag, pp. 103–143.

Burov, E. and Cloetingh, S. (1997). Erosion and rift dynamics: New thermo-mechanical aspects of post-rift evolution of extensional basins. *Earth and Planetary Science Letters*, **150**: 7–26.

Burov, E. and Cloetingh, S. (2009). Interactions of mantle plumes and lithospheric folding: Impact on intra-plate continental tectonics. *Geophysical Journal International*, **178**: 1691–1722.

Burov, E. and Poliakov, A. (2001). Erosion and rheology controls on syn- and post-rift evolution: Verifying old and new ideas using a fully coupled numerical model. *Journal of Geophysical Research*, **106**: 16 461–16 481.

Burov, E. and Watts, A. B. (2006). The long-term strength of continental lithosphere: "jelly-sandwich" or "crème-brûlée"? *GSA Today*, **16**: 4–10.

Burov, E., Guillou-Frottier, L., d'Acremont, E., Le Pourhiet, L., and Cloetingh, S. (2007). Plume Head–lithosphere interactions near intra-continental plate boundaries. *Tectonophysics*, **434**: 15–38.

Carter, N. L. and Tsenn, M. C. (1987). Flow properties of continental lithosphere. *Tectonophysics,* **136**, 27–63.

Cloetingh, S. A. P. L. and Cornu, T. G. M. eds. (2005). Neotectonics and Quaternary fault-reactivation in Europe's intraplate lithosphere. *Quaternary Science Reviews*, **24**(3–4): 235–508.

Cloetingh, S. and Negendank, J. eds. (2010). *Frontiers in Integrated Solid Earth Sciences*. New York: Springer Verlag, pp. 414.

Cloetingh, S. A. P. L., Burov, E., and Poliakov, A. (1999). Lithosphere folding: Primary response to compression? (from central Asia to Paris basin). *Tectonics*, **18**: 1064–1083.

Cloetingh, S., Cornu, T., Ziegler, P. A., Beekman, F., and Environmental Tectonics (ENTEC) Working Group (2006). Neotectonics and intraplate continental topography of the northern Alpine Foreland. *Earth Science Reviews*, **74**: 127–196.

Cloetingh, S. A. P. L., Ziegler, P. A., Beekman, F. *et al*. (2005). Lithospheric memory, state of stress and rheology: Neotectonic controls on Europe's intraplate continental topography. *Quaternary Science Reviews*, **24**: 241–304.

Cloetingh, S., Ziegler, P., Bogaard, P. *et al*. (2007). TOPO-EUROPE: The geoscience of coupled deep Earth-surface processes. *Global and Planetary Change*, **58**: 1–118.

Dèzes, P. and Ziegler, P. A. (2004). Moho depth map of Western and Central Europe (2004). EUCOR-URGENT home page, www.unibas.ch/eucor-urgent.
Dèzes, P., Schmid, S. M., and Ziegler, P. A. (2004). Evolution of the European Cenozoic Rift System: Interaction of the Alpine and Pyrenean orogens with their foreland lithosphere. *Tectonophysics*, **389**: 1–33.
Dixon, T. H. (1991). An introduction to the Global Positioning System and some geological applications. *Reviews of Geophysics*, **29**: 249–276.
Fichtner, A., Bunge, H.-P., and Igel, H. (2006a). The adjoint method in seismology: I. Theory. *Physics of the Earth Planetary Interiors*, **157**(1–2): 6–104.
Fichtner, A., Bunge, H.-P., and Igel, H. (2006b). The adjoint method in seismology: II. Applications: Traveltimes and sensitivity functionals. *Physics of the Earth and Planetary Interiors*, **157**(1–2): 105–123.
Fukao, Y., Widiyantoro, S., and Obayashi, M. (2001). Stagnant slabs in the upper and lower mantle transition region. *Reviews of Geophysics*, **39**: 291–323.
Garcia-Castellanos, D., Verges, J., Gaspar-Escribano, J., and Cloetingh, S. (2003). Interplay between the tectonics, climate, and fluvial transport during the Cenozoic evolution of the Erbo Basin (NE Iberia). *Journal of Geophysical Research*, **108**: 2347, doi:10.1029/2002JB002073.
Heidbach, O., Iaffaldano, G., and Bunge, H.-P. (2008). Topography growth drives stress rotations in the central Andes: Observations and models. *Geophysical Research Letters*, **35**(8): doi:10.1029/2007GL032782.
Iaffaldano, G. and Bunge, H.-P. (2008). Strong plate coupling along the Nazca/South America convergent margin. *Geology*, **36**(6): 443–446.
Iaffaldano, G., Bunge, H.-P., and Buecker, M. (2007). Mountain belt growth inferred from histories of past plate convergence: A new tectonic inverse problem. *Earth and Planetary Science Letters*, **260**: 516–523.
Iaffaldano, G., Bunge, H.-P., and Dixon, T. H. (2006). Feedback between mountain belt growth and plate convergence. *Geology*, **34**: 893–896.
Igel, H. (1999). Wave propagation in three-dimensional spherical sections by the Chebyshev spectral method. *Geophysical Journal International*, **136**(3): 559–566.
Igel, H., Jahnke, G., and Ben-Zion, Y. (2002). Numerical simulation of fault zone trapped waves: Accuracy and 3-D effects. *Pure and Applied Geophysics*, **159**: 2067–2083.
Komatitsch, D., Liu, Q., Tromp, J. *et al*. (2004). Simulations of ground motion in the Los Angeles basin based upon the spectral-element method. *Bulletin of the Seismological Society of America*, **94**: 187–206.
Lithgow-Bertelloni, C. and Richards, M. A. (1998). Dynamics of Cenozoic and Mesozoic plate motions. *Reviews of Geophysics*, **36**: 27–78.
Oeser, J., Bunge, H.-P., and Mohr, M. (2006). Cluster design in the earth sciences: TETHYS. In *High Performance Computing and Communications*. Second International Conference, HPCC 2006, Munich, Germany. LNCS **4208**, Berlin: Springer, pp. 31–40.
Olsen, K., Madariaga, R., and Archuleta, R. (1997). Three-dimensional dynamic simulation of the 1992 Landers earthquake. *Science*, **278**: 834–839.
Persson, K. S., Garcia-Castellanos, D., and Sokoutis, D. (2004). River transport effects on compressional belts: First results from an integrated analogue-numerical model. *Journal of Geophysical Research*, **109**, B01409, doi:10.1029/2002JB002274.
Piazzoni, A. S., Steinle-Neumann, G., Bunge, H.-P., and Dolejš, D. (2007). A mineralogical model for density and elasticity of the Earth's mantle. *Geochemistry Geophysics Geosystems*, **8**: 1–23.
Plesch, A., Shaw, J. H., and Kronman, D. (2007). Mechanics of low-relief detachment folding in the Bajiaochang field, Sichuan Basin, China. *AAPG Bulletin*, **91**: 1559–1575.

Richards, M. A., Yang, W. S., Baumgardner, J. R., and Bunge, H. P. (2001). Role of a low-viscosity zone in stabilizing plate tectonics: Implications for comparative terrestrial planetology. *Geochemistry Geophysics Geosystems*, **2**, doi:10.1029/2000GC000115.

Schuberth, B., Bunge, H.-P., Steinle-Neumann, G., Moder, C., and Oeser, J. (2009). Thermal vs. elastic heterogeneity in high-resolution mantle circulation models with pyrolite composition: high plume excess temperatures in the lowermost mantle. *Geochemistry Geophysics Geosystems*, **10**: doi:10.1029/2008GC002235.

Sigloch, K., McQuarrie, N., and Nolet, G. (2008). Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nature Geoscience*, **1**(7): 458–462.

Smit, J., Brun, J.-P., Fort, X., Cloetingh, S., and Ben-Avraham, Z. (2008). Salt tectonics in pull-apart basins with application to the Dead Sea Basin. *Tectonophysics*, **449**: 1–16.

Sokoutis, D., Burg, J.-P., Bonini, M., Corti, G., and Cloetingh, S. (2005). Lithospheric-scale structures from the perspective of analogue continental collision. *Tectonophysics*, **406**: 1–15.

Sokoutis, D., Corti, G., Bonini, M. *et al*. (2007). Modelling the extension of heterogeneous hot lithosphere. *Tectonophysics*, **444**: 63–79.

Tackley, P. J. (2000). Mantle convection and plate tectonics: Toward an integrated physical and chemical theory. *Science*, **288**: 2002–2007.

Talagrand, O. (1997). Assimilation of observations: An introduction. *Journal of Meteorological Society of Japan*, **75**(1B): 191–209.

Tesauro, M., Kaban, M. K., and Cloetingh, S. A. P. L. (2008). EuCRUST-07: A new reference model for the European crust. *Geophysical Research Letters*, **35**, L05313, doi:10.1029/2007GL032244.

Tesauro, M., Kaban, M. K., Cloetingh, S. A. P. L., Hardebol, N. J., and Beekman, F. (2007). 3D strength and gravity anomalies of the European lithosphere. *Earth and Planetary Science Letters*, **263**: 56–73.

Tromp, J., Komatitsch, D., and Liu, Q. (2008). Spectral-element and adjoint methods in seismology. *Communications in Computational Physics*, **3**(1): 1–32.

Tromp, J., Tape, C., and Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, **160**: 195–216.

Trompert, R. and Hansen, U. (1998). Mantle convection simulations with rheologies that generate plate-like behaviour. *Nature*, **395**: 686–689.

Wortel, M. J. R. and Spakman, W. (2000). Subduction and slab detachment in the Mediterranean-Carpathian region. *Science*, **290**: 1910–1917.

Zhong, S., Zuber, M. T., Moresi, L. N., and Gurnis, M. (2000). Role of temperature dependent viscosity and surface plates in spherical shell models of mantle convection. *Journal of Geophysical Research*, **105**: 11 063–11 082.

Ziegler, P. A. (1988). Evolution of the Arctic-North Atlantic and the Western Tethys. *American Association of Petroleum Geologists Bulletin*, **46**, 198pp.

# 20

# OneGeology – from concept to global project

IAN JACKSON

## 20.1 Background

The OneGeology concept originated in the UK in early 2006. With the potential stimulus of the UN International Year of Planet Earth (IYPE) very much in mind, an idea was opportunistically presented a few weeks later to the General Assembly of the Commission for the Geological Map of the World (CGMW) in Paris. The challenge laid down was: could we use IYPE to begin the creation of an interoperable digital geological dataset of the planet? Fourteen months later, the concept was unanimously endorsed by the international geoscience community at a meeting in Brighton, UK, and goals were set to achieve a global launch at the 33rd IGC in Oslo in August 2008. The goals that the Brighton meeting agreed for OneGeology were deceptively simple. They were to:

- improve the accessibility of geological map data;
- exchange know-how and skills so that all nations could participate; and
- accelerate interoperability in the geosciences and the take up of a new "standard" (GeoSciML).

In Oslo 18 months later, the author, Simon Winchester, launched a fully operational project with 94 national participants, and 25 of those nations were already serving national map datasets through a state-of-the-art web map portal.

## 20.2 The current situation

Today there are 114 countries participating in OneGeology (Figure 20.1), more than 40 of which are serving data using a web map portal and protocols, registries, and technology to "harvest" and serve data from around the world (www.onegeology.org/). An essential part of the development of OneGeology has been the exchange of know-how and provision of guidance and support so that any geological survey can

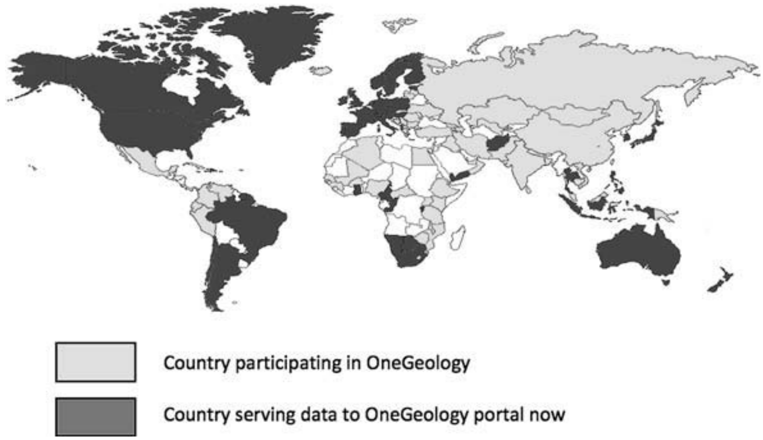



Figure 20.1. Index map of countries participating in OneGeology.

participate and serve their data. The team has also moved forward and raised the profile of a crucial data model and interoperability standard – GeoSciML – a Geography Markup Language for the geosciences developed by the International Union for the Geological Sciences Commission for Geoscience Information (www.geosciml.org/).

OneGeology is coordinated through a two-part "hub" – a Secretariat based at the British Geological Survey (BGS), and the portal technology and servers provided by the French geological survey, BRGM. The "hub" is guided and supported by two international groups – the Operational Management Group (OMG) and the Technical Working Group (TWG). A Steering Group to provide strategic guidance for OneGeology and comprising geological survey directors representing the six continents was formed at the end of 2008. The Steering Group has recently approved a proposal that OneGeology should move to not-for-profit incorporated status to better ensure its sustainability.

A set of success criteria for the next 3 years are providing new goals for the OneGeology work programme. Within these goals, major aims are to increase the number of participants, increase the number of those participants serving data, and increase the number of participants moving from a web map service to a web feature service offering significantly improved functionality. An important aim will be to establish an effective and sustainable governance structure.

Two regional initiatives that are strongly linked to OneGeology are OneGeology-Europe and the US project Geoscience Information Network (GIN: http://usgin.org/). The former is a 2 year, €2.6 million project funded by the European Commission and involves 19 European countries; GIN involves all the US states and the USGS and is funded by the National Science Foundation. Both these initiatives heavily involve OneGeology (global) people and each reinforces the other.

Communication and outreach have always been a priority for OneGeology; nonetheless, the volume of global media coverage the project has received has been astounding. A dynamic web site with rich and regularly updated content is a strong factor in that outreach. Equally important has been a readiness by the team to give presentations across the globe; presentations which echo common messages and themes. The audiences for these presentations range from geoscientists, to informatics and spatial data specialists, to environmental scientists, politicians, and not least the general public. OneGeology has proved to be a project that has much broader appeal (and thus more opportunity to communicate the relevance of geology to society) than was ever envisaged.

## 20.3 Technical aspects

Today over 200 map datasets (local, national, continental, or global) are registered and documented with standardized metadata. Using the web service address available in the register, each dataset can be displayed in various GIS packages or portals. The OneGeology portal (Figure 20.2) has been designed and optimized to search and display multiple layers coming from distributed servers. It provides the usual visualization tools (e.g., zoom, pan, transparency control) as well as the functionality to save a combination of datasets into a "web map context" that can be shared with other users. To support the portal and the registry access, the BRGM has put in place an infrastructure of 15 virtual servers.

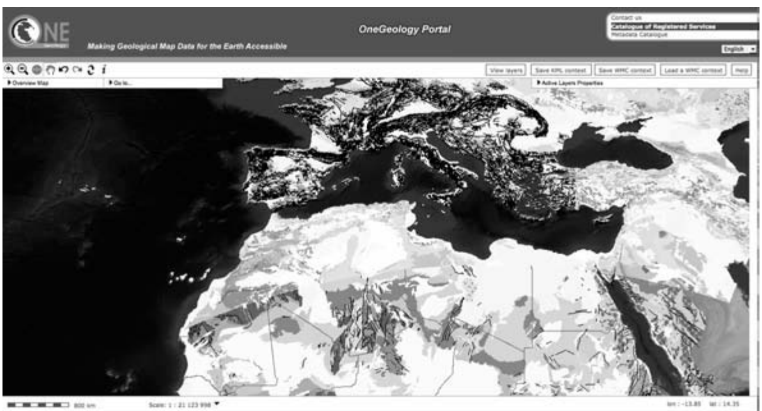



Figure 20.2. Screen capture of the OneGeology portal with selected geological layers displayed. See color plates section.

To help the participants to register their datasets, the Technical Working Group has developed standards for naming the datasets. It also provides cookbooks for preparing web services to deliver maps or features according to GeoSciML standard. The cookbooks are available for download on the OneGeology web site. A new version of the portal was released in June 2009 and is based on a new version of the OpenLayer technology. It provides better performance, an improved user interface, and new functions for searching the registry dataset. The services that deliver features (WFS) will be more visible for the users looking for data access and download. A technical evolution from the current "register" to an Open Geospatial Consortium (OGC) catalog was achieved before end of 2009. This makes possible the connection between the OneGeology catalog and external catalogs supporting the OGC/ISO standards.

## 20.4 Future challenges

While the progress that OneGeology has made is impressive, some serious challenges remain. For example, getting in touch with countries that are not participating in OneGeology and reaching the right people in these organizations has not been easy. Some participating nations have said they are unable to serve data because their national laws/organization business models say they must charge for geological map data. Many participating nations are happy to have their data served by others, but the long-term goal is to have them serve their own data – that will require money and training.

The incorporation of OneGeology is a radical step for the project; will it be supported by the community and succeed in its goals? Where will long-term funding come from and will OneGeology get agreement on a policy on sponsorship? What is the optimum relationship between OneGeology and component/linked regional initiatives such as those in Europe and the USA? OneGeology's global profile has resulted in enquiries from universities, individuals, and companies asking if it is possible for them to participate and serve geological map data – how should it deal with these requests without overloading an already stretched coordination team? The longer-term plans for OneGeology – after 2012 – are not yet well articulated: what is the collective vision of the OneGeology community?

## 20.5 Lessons learned

The question of why OneGeology has been successful has often been raised. There is no single reason; it owes the success it has had to a combination of factors. A short simple mission and vision, coupled with three straightforward objectives was paramount given the diversity of nations and cultures involved. Within that was an uncomplicated initial target of making existing data accessible using basic technology (Web Map Service), and accepting that geological surveys would move progressively to more sophisticated functionality and higher data quality. The project is inclusive: all geological surveys are welcome to participate regardless of their development status. The technical model adopted has meant minimal intrusion into local systems. Coordination and governance have been pragmatic – those prepared to do the work, take the lead and drive OneGeology forward. The underlying ethos has been "let's do it, not excessively strategize about it and discuss it" and this is coupled with a strong "the perfect is the enemy of the good" philosophy. The pre-existence of an international network of geoscientists and geological surveys is a fundamental factor, with at their core, a small group of highly motivated and passionate people who share a common vision. The web site is genuinely dynamic, which is not always something that geological projects and surveys achieve. It provides up-to-date information on all aspects of the project and includes much downloadable and digestible technical and general help and information. The project adopted a "buddy" system where countries with the expertise pair up with those nations who need help to get started. OneGeology took a decision early on to put significant effort into outreach and its media profile, and this has been a major factor in encouraging participation and ensuring a more coherent approach to geoscience data delivery and standards.

OneGeology is a basic global Spatial Data Infrastructure (SDI) for the geosciences. It is a fact that most SDI (and Information System) strategies, no matter how transparent to their architects, are all too often difficult for wider stakeholders, and in particular those who have to implement SDIs to comprehend "What does this really mean for me and my organization?" Taking a practical approach, as OneGeology has, means that those directly involved have been able to understand in their terms, exactly what implementing an SDI means. Through practical deployment, they appreciate the importance of producing metadata, building discovery and view services, and establishing a shared data model and standards to enable this. A practical approach also quickly exposes problems and difficulties of a strategy in a way that a theoretical examination cannot. These invaluable lessons fed back into OneGeology and into the individual geological surveys themselves, and the way geological surveys will implement regional generic SDIs like the INSPIRE Directive in Europe. One of the most important benefits of the OneGeology project itself has been raising competences through exchanging know-how across geological surveys globally (and that also meant helping the more-able surveys to understand what assistance the less well able surveys require). The development of standards and interoperability – both semantic and structural – has been accelerated. The prime objective of improving accessibility has been realized with over 200 datasets now being served by more than 40 nations, and more being added every month. Data are now moving up to higher resolutions as the understanding of the benefits and capabilities of web mapping grows. OneGeology's success globally has meant that geology has been picked up as a case study by the GEO/GEOSS community (www.earthobservations.org/) and is thus ensuring a profile for geoscience in a geoinformatics world where geology used to be a peripheral player. Finally, the fact that OneGeology seems to have been the right project at the right time cannot be underestimated – geological surveys were ready and eager to engage in web mapping, standards, and interoperability; they had an appetite for collaboration and a recognition that operating alone in a shrinking digital world was not the way forward. Without the enthusiastic support of those geological surveys and their staff, OneGeology would not have succeeded.

# 21

# Geoinformatics developments in Germany

JENS KLUMP, JOACHIM WÄCHTER, PETER LÖWE, RALF BILL, AND MATTHIAS LENDHOLT

## 21.1 Geoinformatics history in Germany

With computers becoming available for geosciences in the 1970s, the German research community in the earth sciences realized the potential of an informatics approach to scientific questions in the earth sciences. As early as 1979, groups started investigating how these "new media" could be used in earth science research (Vinken, 1983). The development of applications in the mid 1980s was driven mainly by the needs of land surveying and those of utility companies. By the end of the 1980s, the application of Geographic Information Systems (GIS) began to establish itself as a methodology in academic research, predominantly in the fields of geodesy and geography. However, progress in the field of geoinformatics was hampered by missing standards and immature technology. The term *geoinformatik* for the application of computer science in the earth sciences has been in use in Germany since the mid 1990s. However, its definition remains vague and other terms, such as "geoinformation science" and "geomatics" are still in use.

In the early 1990s, the German National Science Foundation (Deutsche Forschungsgemeinschaft, DFG) established a working group for interoperable GIS (AG IOGIS). The concept of IOGIS can be seen as a precursor to an interoperable geospatial infrastructure. This was also the time when geoinformatics began to establish itself as a subdiscipline of computer science and to reach beyond GIS as its application. However, despite the activities of IOGIS, geoinformatics applications were still isolated developments for subdisciplines of the earth sciences, like geodesy, geography, geophysics, geochemistry, and seismology. Integration of these approaches was attempted in the German Continental Deep-Drilling Program (Kontinentales Tiefbohrprogramm, KTB, Figure 21.1), but progress was slowed by the still immature technology.

In the late 1990s, a second impulse for geoinformatics had gained momentum at the universities. This resulted from two factors: (1) Since 1994 the Internet had

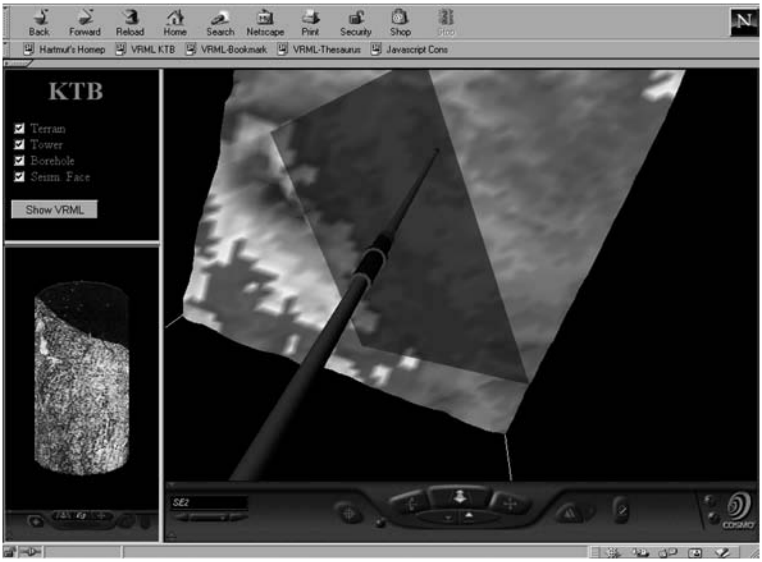



Figure 21.1. Screenshot of a borehole visualization from the German Continental Deep-Drilling Program (KTB). The visualization uses a VRML plug-in in a Netscape browser. See color plates section.

become available as a communication platform for sharing software and data, and (2) the concept of "Free Open Source Software" (FOSS) licenses, which was started by Richard Stallman (Free Software Foundation, www.fsf.org) for the free distribution of program code, had gained popularity, enabling non-commercial collaborative development efforts. This new approach to software licensing allowed the geoinformatics community to revive abandoned software projects ("orphaned code") and continue development in groups of dedicated individuals at other locations.

This needs-driven, grass-roots approach was originally started by early adopters motivated to transcend the limited options in university education for geoinformatics in the geosciences. The first notable project of this kind was GRASS GIS that was originally developed by the US Corps of Engineers and has been in the public domain since the early 1980s. While the code repository was maintained in the USA, the development activities moved on to the University of Hannover, Germany. A modernized code base was released under the General Public Licence (GPL) in 1999. However, the potential of this new wave of GI development was not recognized by the majority of academic institutions where it occurred. As a result, the early adopters left academia for the geoinformatics industry. In the past years this expertise has started to be reintroduced to academia in a "teach the teachers" approach. In the course of the evolution of the German FOSS geoinformatics community, the focus of activity shifted from applications to independent software development, service-based projects and mash-up projects. By now, several of the original German "early adopters" serve as active members of the international umbrella organization for open-source geoinformatics development (OSGeo: www.osgeo.org/), coordinating worldwide activities in this field.

The internet boom around the year 2000 brought a paradigm shift in the development of geoinformatics in Germany, extending the field beyond stand-alone applications towards interoperable services. This development of a geospatial data infrastructure (GSDI) was primarily driven by industry and by federal agencies. Academic involvement in geoinformatics did not keep pace and GSDI was introduced to university curricula only much later. A major development programme for geoinformatics was the launch of the funding programme "GEOTECHNOLOGIEN" in 2001 by the German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung, BMBF) in conjunction with the German National Science Foundation (DFG). This funding programme, which is now extended until 2015, aims at the development of technology related to geoscience research and transfer of these new developments into industrial applications.

By 2003, a number of geoinformatics infrastructures and services had evolved, and the focus of activities had shifted from stand-alone analytical tools towards GSDI. This shift towards service-oriented architectures also prepared geoinformatics for the introduction of Grid technology for high-volume computing. Consequently, representatives from the German geoinformatics community were involved in the D-GRID initiative that prepared the framework for the D-GRID funding programme. In order to realign its funding strategies, the DFG hosted a round table discussion in 2004 to discuss future developments in geoinformatics and geoscience information portals.

Other important drivers that have contributed to the development of geoinformatics in Germany are the German Umbrella Organization for Geoinformation (Deutscher Dachverband für Geoinformation, DDGI: www.ddgi.de) and the German Geoinformatics Socienty (Gesellschaft für Geoinformatik, GfGI: www.gfgi.de). Of these two organizations, DDGI is more oriented towards geoinformatics in government and business applications, while GfGI addresses the academic community.

## 21.2 Spatial data initiatives and spatial data availability in Germany

Technical networking of data resources in geosciences research has been seen as a desirable feature since the early 1990s, as documented by the IOGIS group. However, it was mainly outside of academia that the goal of interoperable geospatial data infrastructures (GSDI) was actively pursued and where the political dimensions of GSDI and of the availability of data were first recognized. As an example, the first major attempt at setting up a nationwide GSDI (GDI-DE: www.gdi-de.org) was initiated by a consortium of agencies at several levels of government, ranging from federal to local governments. To transfer the results of GSDI activities into industrial application, the Federal Ministry of Economics initiated a Commission on Geospatial Industries (GeoBusiness) with the aim to better utilize geospatial information held by government agencies.

In the field of geoscience research data, the main drivers for the development of GSDI are the geoscience research centers under the umbrella of the Helmholtz Association and of the Max-Planck Society. As part of their commitment to international data exchange and to the enhancement of GSDI interoperability, three of these centers host World Data Centers (WDC) accredited with the International Council of Scientific Unions (ICSU), a branch of UNESCO. These four research centers work closely together, and with international partners, on mutual interoperability and the promotion of geospatial infrastructure technologies (Lautenschlager *et al.*, 2005). The German WDCs now serve as prototypes for the realignment and restructuring of the WDCs into a new structure called the World Data System (WDS).

The **World Data Center for Marine Environmental Sciences** (WDC-MARE: www.wdc-mare.org) is maintained by the Alfred Wegener Institute for Polar and Marine Research (AWI), a research center of the Helmholtz Association, and the Center for Marine Environmental Sciences (MARUM), University of Bremen. WDC MARE is collecting and disseminating data related to global change and earth system research in the fields of environmental oceanography, marine geosciences, and marine biology. The PANGAEA information system is working as a long-term archive and publication unit.

The **World Data Center for Climate** (WDCC: www.mad.zmaw.de/wdc-for-climate/) is collecting and disseminating data for climate research. Together with the Max-Planck-Institute for Meteorology in Hamburg with its computing center MPI-DKRZ they are storing and delivering around 5 petabytes of data. While a specific WDCC portal is under development, the access is available through the "Climate and Environmental Retrieving and Archiving" (CERA) database system.

The **World Data Center for Remote Sensing of the Atmosphere** (WDC-RSAT: http://wdc.dlr.de) is hosted by the Cluster of Applied Remote Sensing (CAF) of the German Aerospace Centre (Deutsches Zentrum für Luft- und Raumfahrt, DLR). WDC-RSAT provides data and information on atmospheric trace gases, clouds, and the Earth's surface. The data are primarily gathered by satellite-based sensors. Higher-level data and information products are also generated from the data through assimilation into numerical models of the atmosphere and of its interaction with the biosphere.

As a national research center the **German Research Centre for Geosciences** (Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum GFZ: www.wdc-terra.org) is home to a wealth of data on geology and geophysics of the solid Earth and for geodesy. The data are sourced from geodetic satellite missions, geophysical observatories, geophysical measurements of lithospheric structure and properties, seismology, and deep-drilling projects. The GFZ is in the process of making its data and data catalogs available through web services.

Beside the scientific community the federal administrative offices related to earth sciences started to set up a common spatial data infrastructure named GDI.de. This covers topographical data from the Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodäsie, BKG: www.bkg.bund.de), geological data from the Federal Institute for Geosciences and Natural Resources (Bundesanstalt für Geowissenschaften und Rohstoffe, BGR: www.bgr.bund.de), hydrographical data from the Federal Maritime and Hydrographic Agency (Bundesamt für Seeschifffahrt und Hydrographie, BSH: www.bsh.de), environmental data from the Federal Agency for Nature Conservation (Bundesamt für Naturschutz, BfN: www.bfn.de) and many other spatial information providers. This portal solution is based on international standards defined by ISO with the ISO family 191xx, specifications and implementations given by the Open Geospatial Consortium (such as WMS and WFS), and running in the Internet based on W3C standards. This GSDI is currently extended to the federal states and the municipality level.

## 21.3 Current issues and emerging technologies

The current setup of spatial data infrastructures is based on national and international standards. In addition to implementing these standards, development guides are published that define the rules for interoperable implementations at state and municipal levels, and also for private enterprises. At present, however, access to data is not limited by technological barriers or by lack of availability, but due to licensing aspects and costs.

Even though the value of free data exchange was recognized already in the late 1990s, the policies on access to data showed little effect for many years. Following the adoption of the Berlin Declaration (Berlin Declaration, 2003), the German science organizations established working groups on access to data. The publication of the OECD "Principles and Guidelines for Access to Research Data from Public Funding" (OECD, 2007) initiated a discussion and consultation process in the scientific community that will eventually result in national legislation concerning access to data from publicly funded research. This will, of course, also affect the availability of geospatial data.

To make data centers and scientific web portals effective ways of data sharing, scientists need to be convinced that preparing their data for online publication is a worthwhile effort. An incentive to scientists would be, if data publications were citable publications, adding to their reputation and ranking among their peers

(*Nature* editorial, 2009). To achieve the rank of a publication, a data publication needs to meet the two main criteria, persistence and quality (Klump *et al.*, 2006).

For data to be citable it is necessary that they can be referred to in a persistent way. The location of internet resources, and thus their URLs, may easily change, which in most cases means to the user that the data are lost. Therefore, a prerequisite for data access via the Internet is the use of persistent identifiers, such as DOIs or URNs, to have an address to locate the desired dataset that is reliable and available over a long time (Brase, 2004). Another aspect of persistence is that the data are stored in trusted repositories that guarantee long-term operation. This condition is met by modern data centers, some of which are part of the ICSU system of World Data Centers (Lautenschlager *et al.*, 2005).

Whereas persistence is mainly a technological question, data quality is a far more difficult concept. In ISO 9000:2000, the term “quality” is defined as the “degree to which a set of inherent characteristics fulfills requirements” (ISO, 2000). In terms of data, these could be credibility, usability, and interpretability (Hinrichs and Aden, 2001). Work on metrics for the characterization of data quality in geoinformatics is conducted by the project “Publication and Citation of Scientific Primary Data” (STD-DOI: www.std-doi.de) and by a working group of the DDGI. The DDGI draft has been published as a Publicly Available Specification (PAS) by the German Bureau of Standards (DIN, 2007).

In Germany efforts to create interoperable geoinformatics applications and services have highlighted the need for standardized formats for data and metadata and for their transmission. Efforts towards standardized formats and services have been driven by stakeholders in academia, industry, and government bodies. A significant impulse for this effort was generated by the processes leading up to the European Union directive towards a common “Infrastructure for Spatial Information in Europe” (INSPIRE). Its implementation in European Union member states continues to act as a driving force for government agencies dealing with geospatial data and services towards whom INSPIRE is primarily directed. In the broader context of geoinformatics research and applications, INSPIRE is only one of many approaches and little experience has been gained in its application to geoinformatics research and research-oriented services.

Following the example of the National Science Foundation (NSF) Cyberinfrastructure Initiative in the United States and the development of an e-Science programme in the United Kingdom, stakeholders in research and industry developed a similar programme in Germany, the D-GRID Initiative. As a result of this initiative, several projects have been funded to develop applications of Grid technology and e-Science for research and development. Until now, two projects related to geoinformatics have been funded. In the project “Collaborative Climate Community Data and Processing Grid” (C3GRID: www.c3grid.de) the emphasis

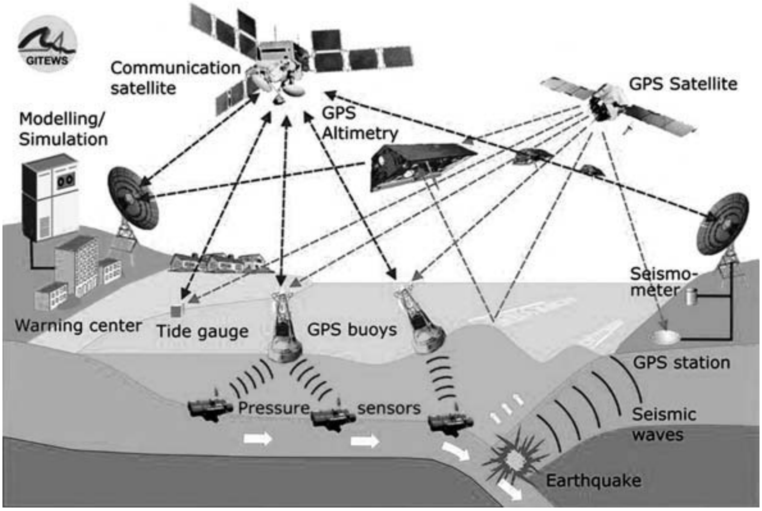



Figure 21.2. Schematic system diagram of the German-Indonesian Tsunami Early Warning System (GITEWS). The early warning system is based on the GITEWS sensor system platform (Tsunami Service Bus) to link sensor networks and models for a decision support system. See color plates section.

is on making Grid resources available for earth system modeling, while in GDI-GRID (www.gdi-grid.de) the focus is on integrating GSDI services with Grid resources, in particular services that require large computing resources, such as the Web Processing Service (WPS). A state-of-the-art report is given in *GIS. Science*, Vol. 3/2009 (www.abcverlag.eu/GISS).

In the wake of the Sumatra earthquake of December 26, 2004, Germany and Indonesia started work on a tsunami early warning system for the Indian Ocean region of Indonesia. This project coincided with the emergence of Sensor Web Enablement (SWE) in the context of the Open Geospatial Consortium and conceptual models and system elements of SWE were introduced into the "German-Indonesian Tsunami Early Warning System" (GITEWS) and the EU-funded project "Distant Early Warning System" (DEWS, Figure 21.2).

A thematically related project is the "Sensor based Landslide Early Warning System" (SLEWS), which covers the whole functional chain of a modular service-orientated early warning system for landslides, based on modern web technologies from geoinformatics and real-time data from ad hoc wireless sensor networks using digital micro sensors (Fernández-Steeger *et al.*, 2009).

Further development of SWE is carried out not only in the context of this project, but also in small and medium enterprises spun out of academic departments, notably the 52°North Initiative for Geospatial Open Source Software GmbH (http://52north.

org/) and lat/lon GmbH (www.lat-lon.de), and others. The advancement of SWE technology is also a component of GDI-GRID.

## 21.4 Conclusions

Comparing the situation in Germany to the cyberinfrastructure programs in the USA, or to the e-Science programme in the UK, the question has been asked whether Germany is still catching up or whether it may benefit from a late-comer advantage (Schroeder *et al.*, 2007). In this context, geoinformatics in Germany may certainly profit from advances in technology that have been gained since the inception of the aforementioned programmes. Research and development in geoinformatics in Germany over the past decades has produced a number of noteworthy results, both in software applications and in conceptual advancements. However, many are still insular efforts. These excellent results need to be coordinated into one systematic approach to address the grand challenges faced by society (climate, energy, mega-cities, etc.). Geoinformatics in Germany will only succeed and grow if it can contribute significantly to addressing these challenges.

## References

Berlin Declaration (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.

Brase, J. (2004). Using digital library techniques: Registration of scientific primary data. In *Research and Advanced Technology for Digital Libraries*, ed. M. Jones *et al*. LNCS. Heidelberg: Springer-Verlag, pp. 488–494.

DIN (2007). *Qualitätsmodell für die Beschreibung von Geodaten*, DIN PAS 1071:2004–10. Berlin: Deutsches Institut für Normung.

Fernández-Steeger, T., Hass, S., Walter, K. *et al*. (2009). SLEWS: Aa prototype system for flexible real time monitoring of landslides using an open spatial data infrastructure and wireless sensor networks. 17. Tagung für Ingenieurgeologie, Freiberg (Sachsen), Germany.

Hinrichs, H. and Aden, T. (2001). An ISO 9000:2000 compliant quality management system for data integration in data warehouse systems. In *International Workshop on Design and Management of Data Warehouses*, Interlaken, Switzerland, June 4, 2001, ed. D. Theodoratus *et al*., pp. 1.1–1.12, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-39/.

ISO (2000). Quality Management Systems: Fundamentals and Vocabulary. ISO 9000: 2000.

Klump, J., Bertelmann, R., Brase, J. *et al*. (2006). Data publication in the Open Access Initiative. *Data Science Journal*, **5**: 79–83.

Lautenschlager, M., Diepenbroek, M., Grobe, H. *et al*. (2005). World Data Center cluster "Earth System Research": An Approach for a common data infrastructure in geosciences. *Eos Transactions AGU*, **86**(52), Fall Meeting Suppl., IN43C-02.

*Nature* editorial (2009). Credit where credit is due. *Nature*. **462**(7275): 825.

OECD (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: Organisation for Economic Co-operation and Development.

Schroeder, R., den Besten, M., and Fry, J. (2007). Catching up or latecomer advantage? Lessons learned from e-research strategies in Germany, in the UK and beyond. In *German eScience Conference*, ed. W. Bühler. Berlin: Max Planck Digital Library.

Vinken, R. (1983). Die automatische Datenverarbeitung als Hilfsmittel bei der Aufnahme und Konstruktion geowissenschaftlicher Karten. In *Geologisches Jahrbuch – Reihe A*, ed. D. Pfeiffer. Hannover: Bundesanstalt für Geowissenschaften und Rohstoffe (BGR).

# 22

# iGEON: Networking the Indian geosciences community through GEON

K. V. SUBBARAO, ARUN AGARWAL, RAJEEV WANKAR, AND CHAITANYA BARU

We must be ever prepared to accept and experiment with new materials from all sources and determine what can be useful to us and take the fullest possible advantage of such materials.

*Sir C. V. Raman (1958)*

## 22.1 Geoinformatics: An Indian perspective

We provide an overview of geoinformatics activities that are being coordinated by the University of Hyderabad, India via two research institutes at the University, viz. the *University Center for Earth and Space Sciences* (UCESS) and the *Center for Modeling, Simulation, and Design* (CMSD) (Agarwal, 2006). We also describe cyberinfrastructure-related activities undertaken as part of the *iGEON-India* project, which has been a collaboration with the GEON project in the USA, supported by a grant from the *Indo-US Science and Technology Forum*, IUSSTF (Agarwal *et al.*, 2009).

During the past 5 years or so, India has made great strides in mineral and other natural resource surveys using both direct and indirect (satellite-based remotely sensed data) techniques, leading to exploration and exploitation of both mineral and forest wealth, and improved data bases for weather prediction and disaster management. We recognize that wise utilization of our natural resources including, minerals, water, and forests, will be possible only through free and unfettered dissemination of accurate data and timely sharing of such information. Data relating to earth sciences are accumulating at an astonishing rate all over the world. Indeed, it has become impossible to keep pace with the growing tempo of new information from research publications. Fortunately, we are also concurrently witnessing spectacular advances in communication technology and in instrumental data acquisition and retrieval. Information gathering and its dispersal have acted as catalysts for the further promotion of research, and information is the key to natural resource management.

Here, we enumerate available datasets and their applications that are being generated in India by various organizations. We will also describe our collaboration, since 2005, with the GEON project – initiated in the USA by the National Science Foundation and coordinated by the San Diego Supercomputer Center (SDSC), University of California, San Diego – in helping promote cyberinfrastructure for the earth sciences in India. Indeed, this “IT head start,” so to speak, has helped us in data fusion and visualization efforts related to a variety of earth science datasets. We believe that this initiative undertaken by the University of Hyderabad is helping create synergy among earth, oceanic and atmospheric realms, as well as space and information sciences, thereby helping bridge a “digital divide,” and promote knowledge-driven development not only in India but in the Southeast Asia region. The initiative has also helped us make significant progress in innovative manpower development by creating a new “hybrid” programme to “cross-train” students in information technology from different science streams.

### *22.1.1 Priority application areas*

Key areas in which the application of cyberinfrastructure techniques is critical and can provide an immediate benefit in the Indian context include the following:

(1) Assembly of recent geophysical, i.e., gravity, magnetic and aeromagnetic, and geochemical surveys over large sections of Indian territory. These offer a challenge in developing innovative methods for integration and fusion of different databases to identify potential mineralized zones, and also to develop 2-D, 3-D, and 4-D models to obtain insight into subsurface structures of the Earth’s crust and mantle.
(2) Accumulation of ocean data through the Tsunami Data Center, Indian National Center for Ocean Information Services, Hyderabad (INCOIS; INTEWS), which has a vast potential for developing improved weather models and early warning systems for disaster management for cyclones, floods, etc.
(3) Management of agriculture through land use, geomorphologic, and ground water maps.
(4) Development of domain data processing algorithms using such data and for the types of applications indicated above, e.g., using expert systems for automated image processing and other automated schemes for interpretation, in order to deal with the data deluge.

### *22.1.2 Geoinformatics research topics*

Several geoinformatics research and development themes arise directly from the data management, integration, analysis, visualization, and interpretation tasks implied by research in the aforementioned priority areas, which we are interested in promoting, such as:

(i) The role of information technology in the development, curation, and use of large, common databases, as well as geological, geophysical and land use maps focused on availability and stewardship of natural resources.
(ii) Understanding climate variability and its impact on agriculture, fishing, water quality and quantity, ecosystems, air quality and human health; and linking this integrative science with the needs of the decision makers.
(iii) Deployment of cyberinfrastructure with the appropriate hardware and software capabilities at different centers around the country to enable quick and proper dissemination of data and knowledge.

## 22.2 Availability of Indian datasets

A wide variety of geoscience datasets are collected by different agencies and institutions in India. Our challenge is to first assemble a list of available resources and, next, to establish a framework within which these data may be shared. We provide a listing of available data in this section. In the next section, we describe how we are attempting to utilize the GEON cyberinfrastructure to encourage sharing of metadata and data among stakeholders and interested users, by making it relatively easy to publish, including via access-controlled interfaces, so that the level of access to different datasets may be controlled by the original dataset providers, based on the privileges they choose to provide to each user.

Following is a listing of available sources of geoscience data in India:

(a) **Indian Satellite Programme and its Roadmap: An Eye in the Sky**: Indian Space Research Organization, Department of Space, Government of India, has conducted earth observation satellite missions since 1993, including for ocean data (OCEANSAT), cartographic data (CARTOSAT), meteorology (INSAT), and land and water data at different resolutions (ISRO). These include:
   (i) Natural resources database:
      - land use/land cover; wastelands
      - geomorphology; soils
      - glaciers, snow melt run-off
      - forest cover
   (ii) Agricultural crop inventory
   (iii) Watershed development planning and monitoring
   (iv) Irrigated command area monitoring
   (v) Irrigation infrastructure monitoring
   (vi) Potential ground-water zones

(vii) Potential fishing zones
(viii) Urban/*I*nfrastructure planning
(ix) Biodiversity at landscape level
(x) Spatial database; space-based information
(xi) Decision support systems for disaster management support

(b) **Data Repositories: Custodian Institutes in India**: In addition to the ISRO, a number of national organizations contribute to the preparation of topographic, geological, and geophysical maps on various scales. Table 22.1 shows the major sources of data and the contributing organization.

Table 22.1. *Major sources of geoscience data in India and the contributing organizations*

| Institution | Resource |
|---|---|
| NGRI, GSI, Oil Companies | Seismic data |
| Mining Companies, NGRI, GSI | Ground geophysical data |
| Ministry of Surface Transport (MST) | Surface transport information |
| Data from State Governments | Census, local administrative network, cadastral information |
| National Atlas and Thematic Mapping Organisation (NATMO) | Thematic data in the form of maps and atlases |
| Ministry of Environment & Forests (MoEF) | Coastal land use maps |
| Indian Meteorological Department (IMD) | Meteorological and seismic data |
| National Hydrographic Department (NHD) | Naval charts |
| Ministry of Ocean Development (MOD) | Oceanic data |
| Survey of India (SOI) | Topographical maps, geodetic trig. And levelling data, gravity & geomagnetic data, GPS data, tidal data, repetitive geodetic & geophysical data |
| Geological Survey of India (GSI) | Geological maps on various scales, geological and seismic data |
| National Remote Sensing Agency (NRSA) | Satellite imageries, land use and waste land maps on small scales |
| Forest Survey of India (FSI) | Forest maps and data of Indian forests |
| Central Ground Water Board (CGWB) | Hydrology maps |
| National Bureau of Soil Survey and Land Use Planning (NBSS&LUP) | Soil maps and land use data |
| Central Water Commission (CWC) | Command area maps |
| National Geophysical Research Institute (NGRI) | Seismological data |
| Indian Institute of Geomagnetism (IIG), NGRI | Magnetic observatory data |
| Geological Survey of India (GSI), Atomic Minerals Directorate (AMD), NGRI, NRSA | Aero-geophysical data |

A few important additions to the Indian geophysical databases are:

(i) Gravity map series of India, produced by NGRI, GSI, SOI, ONGC, OIL, 2006

(ii) Aeromagnetic map of Southern Peninsular India

(iii) Geophysical atlas of Rajasthan, produced by GSI, 1998

(c) **Availability of data for research and collaboration**: Despite vast amounts of databases created by the Indian agencies, only limited data are currently available for public use. For example, most of the topographic maps produced by the Survey of India (SOI) are restricted. However, some of the maps at 1:50 000 scale are now ready to be released in digital form. Similarly, geophysical data, including gravity, aeromagnetic, seismic, seismological, and magnetic observatory data for parts of the Indian terrain are available on request from the respective organizations (see Table 22.1). Satellite imagery at different scales and resolutions are available for sale from the National Remote Sensing Agency (NRSA). As part of an international programme (IGCP) and an NSF-sponsored project, a large geochemical dataset of nearly 10 000 rock analyses for the Deccan basalt province is also available at the University of Hyderabad (Subbarao, 2010).

## 22.3 The cyberinfrastructure framework

The CMSD and UCESS centers at the University of Hyderabad started a collaboration with the GEON project (www.geongrid.org) in 2005, with the goal of leveraging and replicating some of the GEON geoinformatics activities in India, with the University of Hyderabad serving as the hub (Chandra *et al.*, 2008). Corresponding to the structure of GEON in the USA (which was a collaboration between computer scientists and earth scientists, with the San Diego Supercomputer Center taking the lead on the computer science activities and principal investigators from about 12 different universities representing the earth sciences), at the University of Hyderabad, CMSD, a supercomputer center, took the lead on the computer science activities, while UCESS took the lead in the earth sciences.

The Indo-US Science and Technology Forum (IUSSTF) funded this collaborative effort called, iGEON-India, under their *Knowledge R&D Networked Centers* initiative. Hyderabad is an ideal location to establish a knowledge networked center in geoinformatics in India since, not only is the University of Hyderabad one of the leading science research universities in India, Hyderabad is also home to a number of leading research labs and educational institutions with special strengths in this area, including the National Geophysical Research Institute (NGRI), National Remote Sensing Center (NRSC), India National Center for Ocean Information Science (INCOIS), Geological Survey of India (GSI), Central Ground Water Board (CGWB), Indian Institute of Chemical Technology (IICT), and Atomic Minerals Directorate (AMD), and Hyderabad is a key information technology hub in India. The formal partners of the Knowledge R&D Networked Center were the San Diego Supercomputer Center, University of California, San Diego, University of Oklahoma, and Penn State University, in the A, and the University of Hyderabad, Pune University, and the University of Jammu in India.

### *22.3.1 Cyberinfrastructure workshops for geoscientists*

A key early activity that was established as part of the iGEON Center was a week-long workshop on the topic of "Cyberinfrastructure for Geoscientists" held at the University of Hyderabad, with faculty attending from the USA as well as India (Subbarao *et al.*, 2006). Modeled along the lines of the "Cyberinfrastructure Summer Institute for Geoscientists," which has been held every year at SDSC since 2004 (CSIG), the first workshop was held in October 2005 and subsequent workshops have been held in August 2007 and January 2009. The goal was twofold: one, to introduce relevant, state-of-the-art IT technologies to earth scientists and, two, to inform the audience about how advanced geoinformatics tools can be used in research and education. The workshops have covered a number of topics ranging from Grid computing, the role of knowledge representation and use of ontology in data integration, data and information visualization techniques, web services, and portal-based online environments for access to data and tools. They also emphasized how research and education in geophysics and geology can exploit these technologies.

Following are some of the important observations from these workshops:

- **Heterogeneity is the hallmark of the workshop**: The workshops bring together computer scientists and geoscientists and create a forum for discussion among the participants.
- **To be most effective, the workshop should address specific application-related and field-related problems**: The developments in information technology must be driven by the requirements of well-defined applications.
- **Datasets must be made available**: While still difficult, there is willingness among scientists and organizations to share data, especially when they realize that other scientists could perform value-based processing of these data and, in turn, register and share their published work via a community portal.
- **Agencies should upload their applications, tools and algorithms**: We believe this will be possible, as long as agency concerns are addressed by providing appropriate access controls for restricted datasets and tools.
- **A forum is needed for the development of standards in a number of areas**: For example, metadata, data formats, and digitization of paper maps, with the ultimate goal being to collect and register the resulting data.

- **The University of Hyderabad (UofH) is in a good position to support the creation of a well-knit community in this area**: UofH is in a position to develop linkages and partnership with various stakeholders in this area.

There was a clear need and opportunity to conduct a survey of the kinds of datasets and analysis services that can be made available from the iGEON-India portal. Also, the need to continue these workshops in future was recognized, possibly by colocating them with the annual meetings of the Geological Society of India.

### *22.3.2 iGEON-India grid*

The goal of iGEON-India is to promote the use of cyberinfrastructure at participating institutions in India to facilitate the sharing of geosciences data and tools via GEON middleware (Youn *et al*., 2005). The University of Hyderabad has deployed the GEON portal server and a data server at their site in Hyderabad to enable local scientists to register their datasets. GEON PoP nodes were deployed at the University of Pune and Adikavi Nanayya University, Rajahmundry. The PoP nodes serve multiple purposes, including providing capabilities for (a) development and deployment of services, including data access services, by local users to the broader iGEON community, (b) integration of local department or campus resources into iGEON, and (c) provision of resources for system functions such as data caching and monitoring. In the process of establishing these nodes, the staff and students at UofH were trained on various grid middleware software components.

## 22.4 Development of application tools for geosciences research

The cyberinfrastructure workshops identified several requirements for iGEON-India. For example, there is a need for a *geoinformatics toolbox* that would bundle together several functional components to create a complete geoscience application, so that the end user does not have to deal with using each function as a web service. Examples of functionality that could be combined together include:

(1) Conversion of topographic maps (topomaps) to shape files (Sandhya *et al.*, 2008), which involves:
   (a) Reading appropriate images
   (b) Applying filters to the data
   (c) Applying thinning algorithms on the data
   (d) “Cleaning” small segments of results for correctness
   (e) Development and application of contour tracking algorithms
   (f) Conversion of the final results into shapefiles

(2) Change detection:
  (a) Reading and processing of temporal images
  (b) Detecting and measuring change over time specified
  (c) Georeferencing images with topomaps
  Other functions of interest to users include:
  (a) Uploading of datasets
  (b) Formatting and transformations of data
  (c) Georeferencing of data
  (d) Registration/layering of different datasets
  (e) Image enhancement
  (f) Color image processing
  (g) Image restoration
  (h) Classification of object in an image
  (i) Image segmentation
  (j) Analysis of image for qualities such as tone, shape, pattern, texture
  (k) Identification of objects

## 22.5 Future directions

The Indo-US iGEON-India project, as well as funds from the US NSF, have enabled several visits by faculty members and students from one side to the other during the last 2 years for conducting the GEON Workshops at the UofH, participating in AGU meetings, training graduate students in the use of GEON tools, and setting up iGEON nodes in India. As part of this programme, faculty members have traveled from the USA to UofH for the workshops; students from UofH have visited SDSC in the USA; a four-month visit by two UofH women graduate students and a one-month visit by a UofH faculty member have been successfully completed with Professor Randy Keller's lab at the University of Oklahoma. Finally, perhaps the single most important result of this effort has been in developing ties and synergy between earth science and computer science communities in India.

The plan of action for iGEON-India is to (i) complete the ongoing database creation, such as for hydrology, geophysics (i.e., gravity, magnetic, aeromagnetic), geomorphology, geochemistry, paleontology and dinosaur data, geological maps, topographic maps, followed by an integration of all of these various datasets into a common scheme, (ii) make these databases available via the iGEON portal, and (iii) make new applications available tools to the iGEON community. Finally, an important goal is to train UofH and other students and professionals in India in the areas of cyberinfrastructure and geoinformatics through short courses and Masters Programmes.

## Acknowledgements

The success so far of the iGEON-India activities is due to collaborations among teams of scientists and students between India and the USA. Several US researchers and prominent India scientists have participated in the iGEON-India workshops, and researchers and students have exchanged visits between India and the USA, under this programme. Some of the key individuals who have participated in this collaboration are listed below.

## iGEON-India Workshops (2005, 2007, 2009)

Individuals who have contributed their time to participate in the iGEON-India workshops (in 2005, 2007, and 2009) include: Professor Randy Keller, Oklahoma University, Norman, Oklahoma; Dr. Peter Arzberger, Dr. Dogan Seber, Sandeep Chandra, Ashraf Memon, Christopher Crosby, Viswanath Nandigam, University of California, San Diego, California; Dr. Boyan Brodaric, Natural Resources Council, Ottawa, Canada; Professor Mark Gahegan, Penn State University, University Park, Pennsylvania; Dr. K. Radhakrishnan, Director, INCOIS/NRSA; S. K. Subramanian, NRSA; Dr. V. P. Dimri, Dr. Harsh Kumar Gupta, and Dr. H. V. Rambabu, NGRI; Dr. Shailesh Nayak and Dr. M. Ravi Chandran, INCOIS; Dr. J. Saibaba and Mrs Lakshmi, Advanced Data Research Institute (ADRIN), *Hyderabad*; Dr. K. Mruthyunjaya Reddy, Dr. V Raghu and Mehar Baba, *Andhra Pradesh State* Remote Sensing Application Centre (APSRSAC), Dr. Manish Gupta, IBM India; Mr. Rajesh Chabbra, Altair, Bangalore; Sri L. P. Singh, Sri Ramamurthy, Geological Survey of India, GSI.

## Workshop on Geophysical Approaches for Integration/Fusion of Heterogeneous Datasets and Joint Interpretation (2009)

This workshop was jointly conducted by Professor Randy Keller, Oklahoma University and Dr. M. V. Rama Krishna Rao, Datacode, Nagpur, at the UofH on January 9–11, 2009.

## References

Agarwal, A. (2006). Grid Initiative at University of Hyderabad. An invited article for the Anniversary Issue of C-DAC CONNECT, March 2006, pp. 7.

Agarwal, A., Baru, C., Crosby, C. *et al*. (2009). iGEON-India: An International Collaborative Activity of the GEON Project, accepted in Earth and Space Science Informatics-1 session, Geophysical Research Abstracts, Vol. 11, EGU2009–11792–1, 2009, European Geosciences Union General Assembly, Vienna, Austria, April 19–24, 2009.

AMD: Atomic Minerals Directorate, Hyderabad, India, www.amd.gov.in/.
CARTOSAT-2B, ISRO, www.isro.gov.in/satellites/cartosat-2b.aspx.
CGWB: Central Ground Water Board, Southern Region, Hyderabad, India, http://cgwb.gov.in/.
Chandra, S., Baru, C., Subbarao, K. V., and Agarwal, A. (2008). Building a Portal Middleware Grid for Geoscience Communities, European Geosciences Union General Assembly, Vienna, Austria, April 13–18, 2008.
CSIG: Cyberinfrastructure Summer Institute for Geoscientists, www.geongrid.org/index.php/education/summer_institute/.
CMSD: Center for Modeling, Simulation, and Design (Professor-in-Charge: Arun Agarwal), University of Hyderabad, www.uohyd.ernet.in/academic/specialized_centres/CMSD/
GSI: Geological Survey of India, Regional Headquarters, Hyderabad, India, www.portal.gsi.gov.in/.
IICT: Indian Institute of Chemical Technology, www.iictindia.org/.
INCOIS: Indian National Center for Ocean Information Services, www.incois.gov.in.
INTEWS: Indian National Tsunami Early Warning System, www.tsunami.incois.gov.in/ITEWS/HomePage.do.
ISRO: Indian Space Research Organization satellites, www.isro.org/satellites/satelliteshome.aspx.
IUSSTF: Indo-US Science and Technology Forum, www.indousstf.org/.
NGRI: National Geophysical Research Institute, Hyderabad, India, www.ngri.org.in/.
NRSC: National Remote Sensing Center, Hyderabad, India, www.nrsc.gov.in/.
OCEANSAT-2, ISRO, www.isro.org/satellites/oceansat-2.aspx.
Sandhya, B., Wankar, R., Agarwal, A., and Rao, C. R. (2008). Processing of Topomaps on GEON, ESE-OP22, Emerging Trends in Science and Technology, 1st Andhra Pradesh Science Congress 2008, November 14–16, 2008, Hyderabad, India.
Subbarao, K. V. (2010). Geochemical dataset of nearly 10,000 rock analyses for the Deccan basalt province, private collection. Contact: Professor K. V. Subbarao, kvsubbarao_iitb@yahoo.co.in.
Subbarao, K. V., Agarwal, A., and Baru, C. (2006). Workshop on cyberinfrastructure for the geosciences. *Journal of Geological Society of India*, **67**: 700–701.
UCESS: University Center for Earth and Space Sciences (Professor-in-Charge: Professor K. V. Subbarao), University of Hyderabad, India, www.uohyd.ernet.in/academic/specialized_centres/UCESS/.
Youn, C., Baru, C., Bhatia, K. *et al*. (2005). GCE 05 Workshop on Grid Computing based on SC 05, Seattle, WA, USA, November 2005, www.geongrid.org/pubs/local/choonhan_youn_paper16.pdf.

# 23

# Geoinformatics in the public service: Building a cyberinfrastructure across the geological surveys

M. LEE ALLISON, LINDA C. GUNDERSEN, AND STEPHEN M. RICHARD

## 23.1 The need for a national geoscience information network

Geologic data and information are required to find effective scientific solutions to the challenges that society faces today, whether it is developing new sources of energy, sustaining mineral and water resources, mitigating natural hazards, or anticipating climate change. Even when environmental data for a given area are readily available, they are often exceptionally difficult to discover and may exist in different formats and via different services, with different access conditions. The cost and time needed to manually find, assemble, and reformat data are considerable, and many times these data are needed to respond quickly to an issue or emergency.

In early 2007, the Federal and State geological surveys in the United States agreed to the development of the U.S. Geoscience Information Network (http://usgin.org, http://lab.usgin.org) as a data integration framework that is distributed, interoperable, uses open-source standards and common protocols, respects and acknowledges data ownership, fosters communities of practice, and is based on web services and clients (Allison *et al.*, 2008a). This common approach enhances the geoscience community's ability to coordinate within and across scientific domains and, as service-oriented architecture (SOA) designs become more common, it can be fully integrated with the growing global cyberinstructure. The "USGIN" as the network has come to be known, has attracted a large number of collaborators across government, industry, and academic institutes and working groups, including such organizations as the U.S. Department of Energy, Energistics, Inc., Microsoft Research, and the San Diego Supercomputer Center.

Geological surveys have unique resources and mission-specific requirements that include the gathering, archiving, and dissemination of unique long-term earth science data. These data and information represent one of the largest, most extensive collections on the geology and natural resources of the United States. Historically, these data and information have only been available in paper format or in disparate digital systems, which require significant time and resources to explore, extract, and reformat. By using modern information technology and a virtual SOA design that provides common discovery tools and standards, the geological surveys and the general science community will benefit in multiple ways. First, online data and other informational products from each survey will be more readily available to the world audience and will be more valuable because they will be interoperable. Second, data and applications from external sources (databases, catalogs, and inventories) will be readily accessed and integrated with each participant's own data system. Third, a large, federated data network will create immeasurable opportunities for the broader community, including academia and the private sector, to build applications utilizing this huge data resource, and to integrate it with other data. The breadth and depth of survey-based data are so large that they constitute one of the largest data resources in the geosciences, in essence, a national data "backbone."

## 23.2 Vision for the US Geoscience Information Network

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

- Develop a coordinated, national geoscience framework to access and integrate state survey and USGS-information resources (databases, maps, publications, methods, applications, and data services).
- Function as a "community of practice" in developing geoinformatics and the geoscience network.
- Develop prototypes (pilots, test beds) to show proof of concept, to determine realistic levels of effort, and to compare costs and benefits while providing immediate benefits in the form of user services.
- Build the network through an iterative and evolutionary process.
- The basic architecture of the network should be distributed and leverage existing systems, map services, and data with local autonomy, by using standards to enable interoperability.
- Review, test, and adopt standards and protocols for developing the system, including metadata and Open Geospatial Consortium (OGC) protocols and standards (www.opengeospatial.org/).
- Help develop and adopt GeoSciML (geoscience mark-up language) as a protocol and consider proposing it as a standard to the Federal Geographic Data Committee.
- Recognize that there are priority data for which we have mission requirements and inherent partnerships amongst the geological surveys, including data and

information on bedrock and surficial geology, geochemistry, geophysics, mineral and energy resources, geologic hazards, water resources, and subsurface information such as borehole and well data.

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

## 23.3 Conceptual elements for data integration and system design

When USGIN is implemented, we envision a scenario where any user may go to a geological survey, or USGIN partner, to search and view all USGIN catalogs through a simple piece of software served on each geological survey's web site. Other applications being developed and tested now will allow viewing of available data geospatially for a specific State and adjacent States, if desired, and then accessing available web services for download. Because all these data will use common vocabularies and interchange formats, the user can immediately select or download the needed data and load them into any number of applications, including in-house, freeware, and proprietary commercial products that conform to OGC standards. It is intended that the original data source would be credited with the download. This type of "decoupled" system, where the data providers need not know details about the clients or applications and vice versa, provides ease of use and contrasts sharply with centralized systems where data can only be accessed by a dedicated client that is custom built for that application. This latter design restricts or prohibits interoperability and hinders open integration of data and services.

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

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

### *23.3.1 Data integration at the USGS*

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

USGS is using SOA design principles in constructing a new architecture for all USGS data and science applications – a complex challenge for a 131-year-old institution that has been collecting earth science data since its inception (Gallagher *et al.*, 2007). This effort requires operating on many aspects of architecture creation simultaneously, while dealing with extensive legacy analog and digital data. The approach is to create tools and services that assist with the scientists' workflow process while addressing all aspects of the data management life cycle. Projects are underway that include building a federated database network, a master metadata catalog called ScienceBase, creating new web services for discovery of data, creating community-specific data models and vocabularies, creating easy-to-use registry and data upload applications, providing tools for modelers to integrate modeling outputs, and building integrated earth system science applications (Gundersen, 2008). The USGS is partnering across the spectrum of the informatics community to leverage and contribute to data integration efforts. The USGS supports the use of Open Geospatial Consortium standards and is working with Unidata (www.unidata.ucar.edu/) to implement their data access and management tools: THREDDS (Thematic Realtime Environmental Distributed Data Services) and NetCDF (Network Common Data Form), as well as their data access protocol OPeNDAP (Open-source Project for a Network Data Access Protocol).

The USGS is engaged in evaluating the agency's data holdings and creating a searchable, spatially enabled digital catalog that provides discovery tools, metadata, and knowledge of where the scientific data are held and what they pertain to. The USGS has developed a master metadata catalog that it calls "ScienceBase" that provides harvesting and registration services to help users find the best available data sources. The development of ScienceBase is leveraging the technology being used in USGIN to employ an open-standards cataloging method (OGC-CSW). This specification will provide, among other things, a way for ArcGIS users to query directly for all available map-type services that can be incorporated directly into ArcGIS projects. It is also designed so that other USGS catalogs such as the National Digital Catalog of Data and Materials (http://datapreservation.usgs.gov/index.shtml) can be readily accessed from a single search.

The USGS is in the ongoing process of creating, national, regional, and topical map and web services in a wide range of scales and resolution for a broad variety of natural resources data. The National Map (http://nationalmap.gov/) provides imagery, elevation, hydrography, geographic names, and land cover in a new viewer that allows geospatial analyses and downloads. The Mineral Resources On-Line Spatial Data service (http://mrdata.usgs.gov/) provides national databases and maps of mines, historical mining, mineral occurrences, geochemistry of rocks and sediments, lithology, geology, and geophysics of the United States. The National Water Information System (http://waterdata.usgs.gov/nwis) provides access to water-resources data collected at approximately 1.5 million sites in all 50 States, the District of Columbia, Puerto Rico, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands. All three of these data sources provide map browsers, downloads in multiple formats, and web services.

### *23.3.2 National geothermal data system*

A coalition of state geological surveys (via AASG) is expanding and enhancing the National Geothermal Data System (NGDS: www.geothermaldata.org) by creating a national, sustainable, distributed, interoperable network of data providers representing all 50 US states that will develop, collect, serve, and maintain geothermal-relevant data that operates as an integral compliant component of NGDS (www.stategeothermaldata.org). The data exchange mechanism is built on the Geoscience Information Network (GIN) protocols and standards. The NGDS system is described further in Chapter 24.

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

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

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

### *23.3.3 OneGeology*

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

WFSs utilizing GeoSciML as the data transport schema. The OneGeology effort is described in Chapter 20.

## 23.4 Sustainability

One of the challenges facing the field of data integration and geoinformatics, in general, is sustainability. The National Science Foundation now requires that new projects address sustainability in proposals and incorporate planning for sustenance after initial funding ends. Many worthwhile projects have disappeared at the end of the grant funding cycle because of the lack of long-term cyberinfrastructure. The USGIN is a more sustainable enterprise given that geological surveys are government entities that will likely continue with their core missions. A USGIN sustainability path is emerging as groups and companies adopt the framework creating a broad user and contributor base, with growing demand for its services. A broadly deployed system also means that the cost of maintenance can be spread among a larger community so that no one group or organization is burdened with it. The initial validation of USGIN by USGS and AASG set the stage for national deployment and continuity right from the start. Subsequently, the use of USGIN in NGDS, by the petroleum industry as a prototype for metadata standards, and participation by a growing cadre of companies, State and Federal agencies, and data integration and networking projects in related sciences, augurs well for the evolving approach.

## Acknowledgements

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

The ideas behind GIN developed over many years as part of community and project-level discussions with a great many of our colleagues whose contributions have been instrumental in helping us get to this point. We need to highlight the leadership of NSF Acting Division Director Art Goldstein, who pushed for the community workshop in 2007 that prompted the creation of GIN. The system design and architecture described in this paper benefited greatly from extensive discussions with Sky Bristol and his colleagues at the USGS, Kristine Asch at the Bundesanstalt fur Geowissenschaften und Rohstoffe in Germany, and with the International Union of Geological Sciences Commission for the Management and Application of Geoscience Information, GeoSciML Working Group. We also thank Peter Fox for insights on web services and organizations.

## References

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

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

Gallagher, K. T., Bristol, S. R., and Gundersen, L. C. (2007). A data integration and interoperability blueprint for the U.S. Geological Survey (expanded abstract). In *Geoinformatics 2007 – Data to Knowledge, Proceedings*, ed. S. R. Brady, A. K. Sinha and L. C. Gundersen. U.S. Geological Survey Scientific Investigations Report 2007–5199, pp. 58–60.

Gundersen, L. C. (2008). Answers to earth systems science questions: The evolution of geoinformatics at the U.S. Geological Survey. In *Geoinformatics 2008 – Data to Knowledge, Proceedings*, ed. S. R. Brady, A. K. Sinhaand L. C. Gundersen. U.S. Geological Survey Scientific Investigations Report 2008–5172. pp. 2–3.

Keller, G. R., Maidment, D., Walker, J. D. *et al*. (2007). A community workshop and emerging organization to support a national geoinformatics system in the United States (expanded abstract). In *Geoinformatics 2007 – Data to Knowledge, Proceedings*, ed. S. R. Brady, A. K. Sinha and L. C. Gundersen. U.S. Geological Survey Scientific Investigations Report 2007–5199, pp. 75–76.

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

Richard, S. M., Allison, M. L., Clark, R. J., and Grunberg, W. (2009). US GIN: Interoperable geoscience data services on the Web: How do we get there? *Geological Society of America Abstracts with Programs*, **41**(7): 99.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. New York: Cambridge University Press.

# 24

# Application of the US Geoscience Information Network to deploying a national geothermal data system

STEPHEN M. RICHARD, RYAN CLARK, AND WOLFGANG GRUNBERG

## 24.1 Introduction

The U.S. Department of Energy calls the National Geothermal Data System the "cornerstone" of the Geothermal Technologies Program for the next several years in meeting the leading demand of industry for better and easier access to data. As described in the original Department of Energy Funding Opportunity Announcement (FOA):

> The National Geothermal Database will store critical geothermal site attribute information such as temperature at depth, seismicity/microseismicity, fracture maps, drilling data, permeability data, well logs, geophysical surveys, etc. The database should be inclusive of all types of geothermal resources such as hydrothermal, geopressured, Enhanced Geothermal Systems, geothermal fluids co-produced with oil and/or gas, etc. It should also utilize information from existing USGS geothermal resource assessments and DOE funded R&D projects. This standardized set of geothermal resource data will be made available to the public and serve to focus geothermal exploration activities, thereby mitigating investment risks.

DOE further elaborated (http://apps1.eere.energy.gov/geothermal/projects/projects.cfm/ProjectID=27) that:

> The NGDS will be able to handle the full range of geoscience and engineering data pertinent to geothermal resources as well as incorporate data from the full suite of geothermal resource types. It will be able to handle data on geothermal site attributes, power plants, environmental factors, policy and procedure data, and institutional barriers. It will provide resource classification and financial risk assessment tools to help encourage the development of more geothermal resources by industry. It will be an easy to use system that meets the needs of the professional and the public for information on geothermal resources.

Given the breadth of the vision, mission, and resources as the data integration mechanism for NGDS, it is also serving as the flagship effort for design, implementation, deployment (nationwide), and population of USGIN. The system design under development for NGDS since mid 2009 is scalable and transportable so it serves as the base for USGIN, and in turn provides a two-way path for other resources to come in through USGIN in support of the NGDS (Clark *et al.*, in press). The preliminary concepts better inform the implementation.

The U.S. Geoscience Information Network, a joint project of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS), is defined by a collection of service specifications, best practices, registered resources, and reference implementations to enable interoperable web services for geoscience information (see Chapter 22 in this book; Allison *et al.*, 2008). Our goal is to provide a framework that will enable applications constructed to accept data input using standard web service interfaces to work with any server that provides data using a known interface. Data providers (e.g., Geological Surveys) can then publish information by implementing web services without having to simultaneously devote resources to developing client applications to utilize the information. Once a service implementation is in place, addition of content provided by the service is straightforward. Key components of the network are the service specifications, servers and clients implementing the specifications to provide and utilize content, and metadata catalog services that enable users to locate and utilize network resources.

The USGIN project is working on implementing reference server implementations for Open GeoSpatial Consortium (OGC) Catalog Service for the Web (CSW), georeferenced map-image delivery using OGC Web Map Service (WMS), and GeoSciML-encoded geologic map data using Web Feature Service (WFS) (Richard *et al.*, 2009). Wherever possible, we are using existing free open-source projects to avoid duplicating development effort, and to keep the cost of implementation as low as possible. The idea is that the reference implementations may be used as templates for other data providers.

We describe here the data integration framework, based on USGIN concepts, and developed to accommodate State Geological Survey contributions to the NGDS. The results will effectively deploy USGIN nationwide, populate it with geoscience and other data required for geothermal exploration and development that also have wider relevance and application.

## 24.2 Conceptual elements for data integration

Critical system components include standardized Catalog Service for the Web (CSW – ISO 19115 profile) to register and discover resources, web map services to display georeferenced images, and feature services to transport data, using OGC service components to obtain results in a standard format (Richard and Grunberg, 2010).

The shift to a web services (service-oriented architecture – SOA) has followed demands to handle increasingly complex data structures on the Web, while reducing the long-term care and maintenance of providing and updating those resources. Static web pages are relatively easy to create and sustain. The burden is on the web site owner, however, to constantly update the site with new static data as they are updated, added to, or modified. This requires substantial resource commitment or the site is obsolete quickly. It also requires the user to invest significant resources in seeking out non-federated static web sites and manually accessing and integrating the data.

In the past decade, the move to dynamic web pages that capture data on the fly improved the situation, but that was a transitional technology that is now being quickly subsumed by web services. Creation of web services is more difficult and takes more resources than static web pages, but in the long run, the overall cost to maintain and even expand a web services operation is much lower and offers the potential for effectively real-time data integration. As Peter Fox describes it (personal communication, 2010), the burden is shifted from the user to the provider.

## 24.3 System design

***System technical design principles***: The National Geothermal Data System must provide online resources to make it easy for users to extract, assess, and synthesize data according to criteria they select. Data will be provided by a community of data providers, many of whom maintain their own data management systems. There are also numerous kinds of existing "legacy" data in various tables, spreadsheets, and databases that need to be made accessible through the system, as well as many documents that are or could be in digital form and accessible through the system. Some of these legacy data are "orphaned" in that the original producer of the data is no longer involved, and there is no acting steward for the data.

Resources (e.g., data, metadata, catalogs, services, tools) are made accessible through the system by creating metadata conforming to a shared content model and inserting them into the metadata catalog system. The metadata provide description of the resources that can be indexed for discovery by search engines, information about provenance and quality of the resource so users can evaluate the resource for their application, and information describing how to access the resource. The access instructions should be in a format that can be utilized by software clients to automate the access process and minimize the amount of user interaction required to bring the resource to their desktop.

Users should be able to search all resources in the system through a single, but not sole, search client. This means that there can be multiple portals and client applications for accessing system resources; it requires that a single client can search different metadata catalogs in the system without the user having to reconfigure the software.

Providing quality information to evaluate system resources requires criteria that can be used to filter data and categorize them according to established and user-defined quality levels. These quality filters will vary depending on the type of data and their targeted use.

Structured data are provided through NGDS services that have published protocol and documented interchange formats. The idea is that this allows multiple data providers to readily present the same kind of information in the same way, and a client that implements an NGDS service can access that service from any server in the system that offers that service and get data that integrate with minimum operator intervention.

The following points are extracted from the original project proposal, led by Walter Snyder at Boise State University, and subsequent Statement of Project Objectives to help explain the scope of the project:

- Design must be expansive; capture the full physical, geologic, geophysical, and geochemical context of geothermal systems on scales ranging from regional to the individual well bore to the thin section and microscopic scales
- Information in the system must be supported by metadata to document authority and to provide people and projects that compile data the appropriate level of recognition and support:
  - All data will credit the original intellectual source and host server of record for that data
  - Standard measures of "quality" should be available (e.g., variability, bias, systematic error, imprecision, accuracy, precision, reproducibility, etc.)
- Able to adapt to evolving requirements, new technologies and standards, and expanded scope as necessary
- Use existing or emerging standards and technology whenever possible rather than developing new ones
- Open source and open accessibility is preferred to encourage third parties to independently develop software applications that can use the content and services provided by the system
- People who produce data can integrate those data into the data system
- Provide a means of capturing legacy data
- Distributed data system, connected by the principle of data sharing and interoperability among linked sites
- Two-way system of both data-in and data-out
- Provide the users with the base data behind data products
- Assign Digital Object Identifiers (DOI) to datasets
- Accessible through multiple browsers
- Easily maintained

***Data access***: The USGIN framework includes the following preliminary design attributes for data access:

- Provide open access to public data
- Contributors can require user consent to license conditions on data (e.g., noncommercial use only)

- Implementation of access controls and security to limit access to datasets is at discretion of the data provider
- Data owners retain control of access to all data regardless of where they are stored

***Catalog service for the Web, development of interoperable service for discovering resources***: Formal metadata was invented to allow managing and searching for information resources based on domain and content specific criteria, to more efficiently locate and evaluate scientific or technical content that does not index effectively using automated text and http-link-based algorithms. There are compelling business reasons for standardized metadata content, format, and services: Improve efficiency searching for data and information; avoid use of incomplete, inaccurate, or superseded versions of content; avoid duplication of effort searching multiple catalogs or developing software to interact with different catalogs.

To achieve interoperable metadata catalog services, a community of practice must agree on a metadata content model, vocabularies to use keywords to index content, and the protocol for searching catalogs and obtaining metadata. The ISO 19115 and 19119 specifications are becoming widely accepted as content models for metadata, but for application in a particular community, more specific profiles of these specifications are necessary. The US GIN project is collaborating with our partners in the geoscience community to develop metadata profile for indexing geoscience resources. Various keyword thesauri are in use for indexing geoscience resources (e.g., GeoRef), but these have mostly been used in traditional library-type systems, and must be adopted for operation in a distributed system linked by web services; considerable work is also necessary to harmonize the various vocabularies and develop standard encodings, service interfaces, semantic relationships. The USGIN project is implementing metadata catalog services using the Open GeoSpatial Consortium (OGC) Catalog Service for the Web (CSW), testing free, open-source implementations by Deegree and GeoNetwork. Currently, variations in interpretation or ambiguity in the specifications interfere with actual metadata catalog interoperability, but problems are being identified, and the project will document "best practices" to utilize the existing software to link multiple clients and servers.

### *24.3.1 Efficient searching*

A search should return results that are actually relevant. Existing web search tools are very good at indexing relevance based on association of words in text, and using links and user navigation history for those links. This kind of indexing does not work for datasets, in which the information may be encoded in binary format, and proximity of strings may be a function of the data serialization algorithm, not the semantics. Semantic technology is advancing rapidly, and there is significant effort devoted to increasing search efficiency using background information (common sense) encoded in ontologies. The use of controlled vocabularies (ideally linked to an ontology) to index structured data will enable the system to take advantage of semantic technology to increase search efficiency. Determining the elements requiring such vocabularies must be based on specific use cases.

### *24.3.2 Identifiers*

A widely used identifier scheme is important to reduce duplication, and determine associations between resources. Globally unique identifiers are essential for the described resource, and for the metadata record.

The current thinking in the WWW community appears to be converging on a consensus to use http URIs that are expected to de-reference to some useful resource representation. A widely used and understood identifier scheme also enables semantic web functionality. The "anyone can say anything about anything" paradigm requires being able to identify the things.

### *24.3.3 Query complexity*

The complex search examples in the use-cases section involve associations between resources, or resource-specific properties. The following table is a decomposition of some complex query examples. Careful consideration of such decomposition is necessary to determine the boundary between metadata catalog services with metadata search, and data services that allow filtering of data elements based on their properties.

Consideration of these queries indicates a requirement to distinguish metadata service from a data service. When the request involves properties of specific instances of a particular resource type, a data service for that resource should be accessed. The metadata for that service should describe the properties offered for resource instances in that service.

Cases 1–3 can be handled in a general way by a service chaining process, in which the metadata catalog is searched for services offering the feature of interest with the property of interest that will be used as a selection criteria. This approach keeps the top-level resource metadata catalog simpler, but makes discovery operations significantly more complex. Cases 1–3 can also be handled with scoped keyword terms, where the scope includes things like "analysis type," "geologic unit," "related resource type." In this usage, the scope specifies a controlled vocabulary of categories related to some concept. Addition of new querying capabilities requires adding additional scoped keywords in the metadata. The second approach is viewed as more appropriate in a "keep it simple" design framework for minimum metadata requirements.

Table 24.1. *Analysis of complex queries*

| Case | Plain language query | Decomposition | Simplified solution |
|---|---|---|---|
| 1 | Boreholes that have core in a particular depth interval in a given area. | Borehole-centric approach – geographic search for borehole resources (assume collar location), filter for those that have a related resource "core," filter again for property of related resource "core interval = min, max depth meters." Alternatively, view search as actually for a "core" resource, so search should be for "core" with some given vertical extent. The core resource must provide an ID "xxxx" for the borehole from which it was obtained. To obtain more details about the borehole, search for metadata on borehole with resource ID = "xxxx." | Include keywords for other resources associated with borehole. Put information about these in the abstract. User searches metadata catalog for borehole with keyword (thesaurus=related resource) = "core," reads abstract to see if it is what they want. The keywords would have to be a controlled vocabulary. |
| 2 | Boreholes that penetrate the Escabrosa formation in a given area. | Geographic search for borehole resources (assume collar location), filter for property "intersects Escabrosa formation." Alternatively, search for borehole service that includes property = "formation tops," then query that service. Service properties would have to be from controlled vocabulary. | Include names of penetrated formations as keywords on a borehole. Formation names ideally from a geologic unit lexicon. |
| 3 | Locations for samples with uranium–lead geochronologic data in a given area. | Search metadata catalog for geochronology data service with property = "analysis type" and backtrack to location point through sample metadata, or search metadata catalog for U– | Include keywords for kinds of analytical data associated with a sample in the sample metadata record. Search for samples with keyword (thesaurus=analysis type) = "U–Pb geochronology." |

Table 24.1. (*cont.*)

| Case | Plain language query | Decomposition | Simplified solution |
|---|---|---|---|
| | | Pb Geochronology Data Service and backtrack to location point through sample metadata, or search for "sample service" with property = "analysis type." In the second case, there would still need to be some metadata property to indicate the analysis type for the service. Approach via the analytical data service requires chaining to the sample feature service, analogous to case 1 for borehole service. | |
| 4 | Find links to pdfs of publications by Harold Drewes on southeast Arizona. | Search for document resource with author = "Harold Drewes" and geographic extent = "SE Arizona," and online distribution format = "pdf." | Is search by representation format high enough priority to support? |
| 5 | Find geologic maps at scale <100 000 in the Iron Mountains. | Search for geologic map resource with geographic extent = "Iron Mountains, and resolution scale denominator <100 000." | Is search by resolution high enough priority to support? |
| 6 | Who has a physical copy of USGS I-427? | Search for document publisher = USGS, Series ID = I-427, offline distribution format = "paper copy." | Include the document ID in the resource description. |

Cases 4–6 are related to document-oriented searches, for which distribution format and online access are important, and a number of bibliographic properties (scale, publisher, series, series ID, media, file format) come into play.

### *24.3.4 Accessing resources*

In order for software to utilize URLs in metadata without operator intervention, strong conventions are necessary to guide what URLs are in the metadata and where they are placed. Links in metadata to access resources should in general be complete URLs that can be invoked with a simple http GET, without having to add additional request parameters. Formal elements (with controlled vocabulary content) should provide machine-processable information to distinguish links that will return a document from links that invoke a service or access an online interactive application. The idea is that sufficient information should be provided so that client software can parse the metadata record and provide useful functionality on the resource with minimal user interaction.

For many resources, different representations may be available. These might be different file formats for the same document for information resources. For non-information resources, a variety of representations that have different uses might be available. For example, a physical sample may be represented by a text description of the sample, a GeoSciML XML description, visible light photograph, or images of the sample using other sensors. A geologic map may be available as a paper copy, a scanned image, a georeferenced scanned image, a vector dataset in one of several formats (gml, shape file, file geodatabase, MIF, DWG), through a web map service, or through a web feature service. Metadata for a resource should be able to describe all of these different representations that the resource provider wishes to make available, in such a way that automated clients can seek representations useful to that client, or search clients can present users with links to access different formats or representations.

### *24.3.5 Citation and contact information*

Citation information specifies the source of some content. Citations for the described resource specify the origin for the resource intellectual content. The cited agent may have played various roles relative to the resource – author, compiler, editor, collector, etc., and a controlled vocabulary is necessary to specify these. Citation for a metadata record specifies the agent responsible for producing the record, typically thought of as the metadata record creator. Metadata production involves elements of authoring, compiling, and editing. Minimally, citations must identify an individual person, an organization, or a role in an organization that is the agent filling a specified role relative to the cited resource. In most cases an organization will be specified, either as the employer or sponsor of a person, an institutional actor, or the host for some role (web master, metadata editor). In addition, information required to contact the cited actor is necessary to enable metadata users to contact a person with some knowledge of the cited resource. For long-lived metadata, contact for an agency role is most likely to persist. The minimum metadata contact information recommended is either an email address or telephone number.

### 24.3.6 Fitness for purpose

The metadata should provide sufficient description of the resource for a user to determine if the resource is likely to meet their needs, and to determine what representation to access. The simplest approach is to provide such information as text in the metadata abstract, including why the resource was produced, what sort of observation procedures were used, assessment of data completeness, accuracy, and precision, and comparison with other known similar resources. The data quality section of ISO 19115 provides a data structure to formally describe this information, but the cost of using this is high (complex data entry), and there do not currently appear to be clients that utilize the information. The guiding principle should be that if users need to search on some particular quality criteria, specific guidance on how to encode that information in the metadata is necessary (e.g., which ISO 19139 elements, what controlled vocabulary to use if terminology is involved).

### 24.3.7 Branding

In a distributed, federated metadata catalog system with harvesting, metadata records are expected to propagate far beyond their original point of introduction into the system. If an organization producing metadata wishes to be recognized, and in order for users to be able to contact the metadata originator, contact information for the metadata originator must be considered part of the metadata record, and maintained in harvest processes. For presentation to users, it is desirable to provide a link to an icon that can be displayed with records to brand the origination of the metadata.

The same considerations hold for the resource itself.

### 24.3.8 Access constraints, legal limitations

Metadata records that are not for public consumption should never be exposed to a harvesting request. Implementation of security and access control must occur at a lower layer in the network stack than the metadata catalog service is operating, such that authorization/user authentication information is handled by the environment containing the metadata catalog client and server. Metadata for commercially licensed resources may be publicly accessible, but should clearly indicate the licensing requirements and procedure to access the resource.

### 24.3.9 Low cost of entry

Metadata producers should be able to reuse and build on existing structured metadata. Minimum requirements should be limited to information that is commonly available. Resource-specific details should be provided in text elements in the metadata. Special information necessary to utilize web links (e.g., web service operation) in metadata should be provided by text in the metadata or through linked documents.

*Discussion*

One of the basic concepts of USGIN and requirements of the NGDS is to make access to data simpler. A major time-consuming aspect of bringing disparate datasets together is data integration. This process involves matching field or element names in the schema for various datasets, selecting those that contain the information of interest, and then merging content into a single dataset with consistent usage of vocabulary and units of measure in a standardized collection of fields or elements. Data integration in our current system of scientific information interchange is mostly left to the data consumer. One obvious path to simplify data access is to develop standard formats for integrating common datasets (e.g., borehole temperature data, heat flow measurements) that is used to deliver content to data consumers.

A major decision for data delivery in a federated system is where the data integration occurs. Until recently, the most common approach was for a database compiler to collect various datasets and integrate them into a single database that was then made available. This approach works fine while the data compiler has the resources to continue integrating new data, and while the data compiler is still professionally active.

A second approach is to do the data integration as close as possible to the original data provider. By documenting data schema, encoding formats and practices for vocabulary usage, data can be put into the "data integration" format when it is made available on the Web. This requires education of the data providers/publishers on the use of the integration formats, but results in a larger community of information technology personnel who know how to get data into and out of the integration format. The originator of the data is likely to be a better judge when it comes to making decisions on how to map their content into an interchange format (assuming they are comfortable with the interchange format). In the case of geological surveys, policies can be developed to always present data in the data integration format (along with any other formats that the data publisher wants to use). The net effect is a greater likelihood that the federated information system using the documented interchange formats will outlast any particular researcher, data provider, project, or agency. HTML on http and NetCDF are examples of data integration formats that have achieved wide usage and long-term usefulness. If a widespread core user/provider constituency such as the geological surveys, adopts certain formats, these could become de facto standards across the wider community.

The use of schema and encoding specifically designed for data integration and interchange means data producers and consumers can continue to use internal data formats that are optimized for their business requirements. Use of the community interchange formats reduces the amount of work required because only one

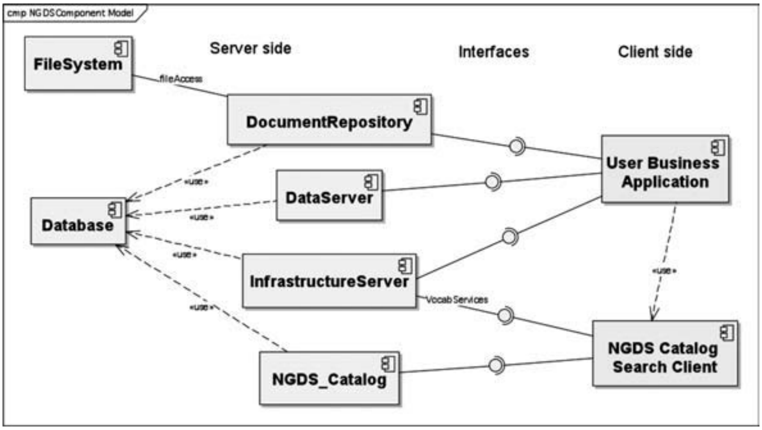



Figure 24.1. Functional components of a data integration framework as proposed for the National Geothermal Data System. Server side components are on the left, and client components are on the right. Client components interact with server components through a variety of interfaces.

transformation from internal to interchange format has to be engineered for each interchange format in use.

## 24.4 System architecture

The framework for implementing data handling requirements is a community of data providers exposing information through standardized internet-accessible interfaces (services), a community of software developers building applications that will utilize the information resources available to the community, and a community of users taking advantage of the software and information to develop geothermal resources. The service inventory would be focused on entity data services that provide information resources. A key component is the metadata catalog service through which data providers register the availability of resources, and users discover, evaluate, and access resources.

The system architecture will be described in terms of the functional components shown in Figure 24.1. These are discussed in the following sections.

## 24.5 Functional components

### *Metadata catalog*

An NGDS metadata catalog component implements one or more protocols for searching a metadata store and returning metadata. At least one of the implemented protocols and interchange formats used for delivering metadata must conform to an NGDS specification. Initial metadata catalog testing and prototypes are using the OpenGeospatial Consortium Catalog Service for the Web (CSW), but other protocols such as the Open Archive Initiative Protocol for Metadata Harvest (OAI-PMH) or the Z39.50 protocol may also prove to be useful. The CSW was selected for initial development work because it operates in the same framework as the other Open Geospatial Consortium services being tested for data delivery (the Web Map Service and Web Feature Service), is designed for geospatial data, and has a variety of free, open-source software projects developing clients and servers for the protocol, as well as a variety of commercial products (including ESRI ArcGIS) that are implementing the protocol.

The CSW service requires all conformant implementations to return metadata using a simple XML encoding of the Dublin Core elements and terms, and defines a subset of these that are core queryable and returnable elements (see OGC 07–006r1). The base CSW specification adds a bounding box as a core queryable requirement for any CSW metadata catalog. The CSW service can operate with any XML schema for metadata content, and in the geospatial community, the most widely used profile is for the ISO 19115/19115 metadata. Use of this metadata schema allows richer metadata content that enables greater automation of access to resources.

An NGDS metadata catalog may be implemented with various software and hardware configurations on any node in the system. To be an NGDS compatible/compliant metadata catalog, the only requirement is that it implements an NGDS metadata catalog service profile, and provides metadata in at least one output format schema and profile that conforms to an NGDS metadata interchange specification. Our recommendation is to use the OGC CSW with its base metadata schema (Dublin Core elements and terms), and for more in-depth metadata, the ISO 19139 encoding of the ISO19115/119 metadata content model, following recommendations proposed in the USGIN ISO metadata profile.

### *Document repository*

Data in documents will be accessed via URL from document repositories, which are basically web-accessible file systems. In this context, “document” is used in a very general way as a packaged body of intellectual work with an author (or editor, compiler, or similar originating role), a title, and some status with respect to review/authority/quality. Documents can be packaged in a single file or a group of related, linked digital files. Documents provide a straightforward path to get data online quickly and easily for the data provider, but if this approach is used for datasets (e.g., Excel spreadsheets, Microsoft Access databases), it requires the data consumer to do all data integration work themselves.

Many options are available for implementing document repositories, including DSpace (FOSS), OCLC ContentDM (commercial), and the Drupal-based document repository developed in collaboration with the USGIN project. In order to integrate holdings in system document repositories, metadata for contained resources in a documented NGDS metadata interchange format containing the required metadata content must be made available for insertion into the NGDS metadata catalog system.

### *Data servers*

A data server is any component that implements a service providing data using at least one protocol and interchange format conforming to an NGDS specification. Data service delivery of content differs from the simpler document-based delivery because it requires that the format and content delivered will conform to some known set of rules, allowing software to interact directly with the data server to facilitate user acquisition and integration of data into their work environment.

Data delivery through a service requires the service provider to perform any necessary data integration operations to get content into the schema conforming to the service profile. This requires more work for the data provider than the simpler document-deliver approach, and thus will have to be implemented incrementally based on the quantity and significance of various data items. Data types that are deemed suitable for service delivery will have NGDS protocols, interchange formats, and vocabularies defined to enable automated access to those data.

Since many of the data types are associated with geographically located features, the OpenGeospatial Consortium Web Feature Service (WFS) is proposed as the starting point for implementation of feature services. This protocol uses GML geometry for location description, and allows feature types to be defined that are characterized by feature-specific XML schema.

A number of international efforts are under way to develop specifications for data interchange of geoscience information (GeoSciML), and basic observation and measurement data (ISO19156). These XML schema are very flexible to allow representation of a wide range of content, but are thus correspondingly complex. Currently there are no client applications that can do more than transform complex XML to HTML for display.

Thus, in the initial phase of the project, services will be defined using simple XML schema with string and numeric-valued elements (Richard and Grunberg, 2010). These services can be consumed by existing clients like ArcMap and Quantum GIS. These simple schema will be compatible with the ISO specifications to the degree that is practical. As clients are developed for richer-content, complex feature services, the NGDS will migrate towards use of the more complex schema. There are also a number of other data formats in use in related communities for geoscience information interchange, including WaterML in use by the CUAHSI project, NetCDF, which is widely used for large numeric datasets in the atmospheric and remote sensing communities, and an XML mark-up developed for geochemical data by the EarthChem project. Whereever possible, NGDS data providers should reuse existing schema to take advantage of tools developed to consume data in these formats.

### *Infrastructure server*

The extensive requirements for the NGDS laid out in the requirements section prescribe a collection of functions that must be available on a system wide basis. These functions will be provided by infrastructure servers, prime among which is the NGDS Core at Boise State. The most important infrastructure services that have been identified at this point include caching, mirroring, and backing-up system data; providing a home for orphaned data or legacy data; user authentication for access control, vocabulary services for provision of community vocabularies for semantic interoperability, and identifier registration services that will provide URI de-referencing and mapping between identifier schemes to avoid unrecognized duplication of resources. Other infrastructure functionality that would be useful includes validation of information interchange documents to determine if and to what degree they conform to system specifications; and social networking functions such as resource rating, comment, feedback; and usage monitoring and reporting. Development of such infrastructure services should be prioritized to support data services that are actually being implemented.

### *Database and file system*

Various databases and file systems accessed by server applications will house the actual system resources. For security and simplicity, these will probably not be directly accessible to system users, but will be accessed through NGDS servers or clients such as the Geothermal Desktop. Many user applications may also have a local data store and file system used to cache resources obtained from the system for offline usage, better performance, and reliability (not dependent on operation of the Internet).

### *Clients*

The client applications implement most of the desktop analytical and search functionality required by the system. These are outside the scope of this data access system architecture, except for the provision that they operate with the NGDS metadata catalog for resource discovery and evaluation, and utilize NGDS services and repositories for data access.

## 24.6 System deployment

*Nodes*

Any server that is internet accessible and implements one or more NGDS services, including a document repository containing files indexed by metadata in NGDS catalogs, is effectively a node in the system (Figure 24.2). Each node will implement one or more of the abstract components, and will need to register public resources available at that node in the metadata catalog system.

Some nodes will implement special functions, including archives, system specification repositories, and registries of identifiers.

The deployment diagram indicates a key aspect of the system – the user client software interacts with components on the server side through a pipe labeled "NGDS services." This connection represents any and all service protocols used to link clients and data servers in the system. These services define interfaces that decouple the

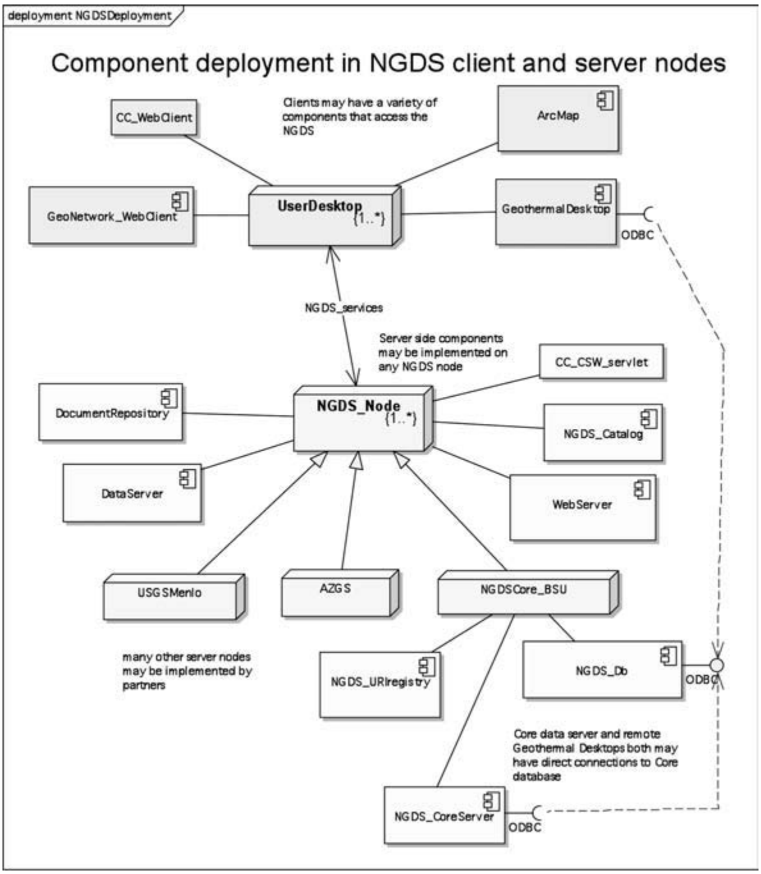



Figure 24.2. Deployment of components to nodes in the system.

clients and servers such that as long as the operations and behavior of the service do not change, any upgrades or modifications that occur in the client or server software do not break the system. This loose coupling is a key design feature necessary to allow the system to evolve as technology and user requirements change.

Direct connections using proprietary technology may exist between clients and servers managed by some participants (e.g., ODBC to ODBC connection indicated between client Geothermal Desktop and server NGDS_Db). Such connections may be necessary for expediency, security, or special performance requirements, but should be considered interim solutions because they violate the premise of an open system in which services offered are publicly documented and available to any client in the system.

## 24.7 Data acquisition plan

The central idea of the data access architecture proposed here is the idea that data providers and client applications should be linked through open source interfaces that decouple clients and servers such that they can evolve independently without breaking the system. The hypertext transfer protocol (http) and hypertext mark-up language (HTML) are far and away the best established protocols and interchange formats in use on the Internet, and in the near term these will probably continue to be the mainstay of most interaction in the NGDS.

We are developing a road map for bringing data into the information infrastructure that is the foundation of USGIN and the NGDS as quickly as possible by using existing, tested OpenGeospatial Consortium services, particularly Web Map Service (WMS) and Web Feature Service (WFS). In a nutshell, the steps in this plan are:

(1) Identify the kinds of information to be made available through the system.
(2) Prioritize acquisition according to availability, importance for geothermal resource evaluation and development, and difficulty of acquisition.
(3) Make data resources accessible:
   (a) For document-based resources and datasets that do not have specifications for interchange protocols, data schema, and file format: create metadata for resource and make resource available in a web-accessible location linked to from the metadata.
   (b) For high-value datasets with sufficient volume, design and implement XML schema based on any applicable standards to use as an interchange format in WFS service response documents, and make the data available through WFS service. Metadata describing service function and content go in metadata catalog.

(c) For map-based portrayals of information made available as documents, and as WMS service layers. Metadata describing map content and distribution points go in metadata catalog.

Data types for which NGDS data acquisition services and interchange formats have not been specified will be made available in user-defined data files that will be described by metadata in the system catalog and placed in web-accessible servers. Standardization of automated, interoperable data acquisition via services and community interchange formats will be developed incrementally, starting with highest-priority data types. Priority will be determined by data availability and requirements from application developers working on client software useful for geothermal resource development.

For interoperable data to be presented to the system using standardized protocols, interchange formats, and vocabularies, the development team will need to work with the user community (data providers and consumers) to determine a useful starting collection of attributes for entities or features that will be delivered, including units of measure and required controlled vocabularies. Interoperability means in practice that software will use the same access protocol for a given kind of information from any NGDS data provider, without any provider-specific customization. Some important requirements include:

(1) Ensure interoperability among datasets with members adopting common standards and protocols.
(2) The data schema must be vetted with stakeholders.
(3) Data schema for interchange formats must be versioned, such that expanded or modified versions can be introduced without disrupting working systems.

### *File-based data*

File-based data access will be the option of choice for text documents, but will also be used for datasets that do not have a standard interchange protocol and file formats defined. Some tabular file formats may already be in use, or be specified by groups of users to simplify exchange of some kinds of information, and if widely used these would be obvious candidates for system interchange formats. The recommended metadata for file-based (document) resources is designed to allow discovery, evaluation of the resource based on text description, and access to the resource via a web link (URL).

### *Data to be scanned*

Reports, logs, maps, and other documents pertinent to geothermal energy exploration, evaluation, development, and production that exist in hard copy but are not available online may be converted to digital form by scanning to create digital image files. If the resource is a map, it should be georeferenced (geoTiff or world file) if possible. Preferred document formats are pdf, tif, jpg, or png. File formats that are specific to particular (especially proprietary) software are undesirable and their use will need to be justified and approved by the project management. OCR processing of text to make Adobe Acrobat files searchable is highly desirable. Georeferenced map images ideally will be published through a Web Map Service (WMS) as well as accessed from document repositories. Deliverable digital documents must be publicly available online, and registered in the NGDS metadata catalog. A prototype document repository, implemented using Drupal software is available for deployment by data providers that do not currently have such an online repository (http://repository.usgin.org/). This application also supports production of metadata meeting NGDS requirements. Instructions for deployment are available at http://lab.usgin.org/groups/drupal-development/creating-document-repository-drupal.

### *Online digital data*

Implementation of online data services will involve several steps. First, an application profile for the service or services to deliver a particular kind of data will have to be developed. In most cases, we anticipate that existing standard services like the OpenGeospatial Consortium (OGC) Web Feature, Map, or Coverage services (WFS, WMS, and WCS), will provide the necessary framework for services we require. The NGDS technical team will need to develop profile documents specifying the details of how a particular type of information (e.g., borehole temperatures, water chemistry analytical data) will be encoded. Once a profile is in place for a particular data resource, the next step is working with the data-providing organization to implement the service with their data.

The actual mechanics of bringing particular datasets online will be dependent of the format of existing data, and the IT resources of the data owner. Some organizations may choose to implement web services on their own servers to expose datasets, others may choose to work with a partner that has better IT support to host services.

The second part of the online service implementation and deployment is registering the new data service with the metadata catalog system. This will require creating a metadata record for the service, and loading it into a metadata catalog server that is harvested by the NGDS metadata catalog system, such that the fact of the service's existence, and information to evaluate and access the service becomes available to the community. The data acquisition process will thus need to include guidance on what kind of metadata will be required to register resources with the metadata catalog system to make them available.

For online data services, registration of a dataset in the metadata catalog, and its availability online will constitute “data acquisition.” Thus, implementation of the metadata catalog as an operational service will need to be one of the first steps in system implementation. AZGS has developed a prototype metadata catalog, implementing the CSW 2.0.2 catalog service using Geonetwork OpenSource, currently at v. 2.6, but in active development with new versions coming out 2–3 times per year.

### *Data delivery options*

Participants have two options on how to make their data available:

- register files in an NGDS-compliant document repository; submit metadata to an NGDS-compliant catalog. If the files contain datasets, then the structure of the data (entities, attributes, vocabulary) should be described in the metadata such that someone using the file dataset can figure out what they have got; or
- implement a web service for direct online access to the data. Submit metadata to an NGDS-compliant catalog.

Data will be considered part of the NGDS when they are locatable using the NGDS core metadata catalog, and accessible via the Web according to procedures described in the metadata record obtained from the NGDS core catalog.

### *Metadata*

*Metadata* should be created and submitted for any resource that is meant to be accessible individually via the Web.

*Individual documents* require one metadata record per document. Some document types may consist of a bundle of files, e.g., ESRI shape file. In general, these should be bundled into a single file like a zip archive or UNIX tar file. The metadata must include the URL at which the document can be accessed. These documents might be scans of well logs, scanned reports or publications, or data in a spreadsheet, such as an Excel file.

*Datasets* include internal record-level source information, documenting details of observation or measurement procedure and other information specific to a particular data type. This includes information such as location, data and time of observations, and the source of the data. These metadata are delivered with the data, and only summarized in the dataset metadata that are published to the NGDS-compliant catalog.

## Acknowledgements

This report is based in part on work done at the Arizona Geological Survey under the auspices of DOE award DE-EE0001120, in conjunction with related work on development of the Geoscience Information Network (US GIN) supported by NSF grant EAR-0753154, and a parallel DOE award DE-EE1002850 to compile and publish geothermal data from state geological surveys to integrate with the NGDS.

The ideas behind USGIN developed over many years as part of community- and project-level discussions with a great many of our colleagues, whose contributions have been instrumental in help us get to this point.

The system design and architecture described in this paper benefited greatly from extensive discussions with Sky Bristol and his colleagues at the USGS, Kristine Asch at the Bundesanstalt fur Geowissenschaften und Rohstoffe in Germany, and with the International Union of Geological Sciences Commission for the Management and Application of Geoscience Information, GeoSciML Working Group.

## References

Allison, M. L., Dickinson, T. L., and Gundersen, L. C. (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, submitted to National Science Foundation for Award 0723437 to the Association of American State Geologists, March 1, 2007 to February 29, 2008. Arizona Geological Survey Open-file Report 08-01, 23pp.

Clark, R., Richard, S., and Grunberg, W. (in press). The national Geothermal Data System: The Geoscience Information Network in action. In *Digital Mapping Techniques '10 – Workshop Proceedings*, ed. D. Soller. U.S. Geological Survey Open-file Report.

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

Richard, S. M., Allison, M. L., Clark, R. J, and Grunberg, W. (2009). US GIN: Interoperable geoscience data services on the Web: How do we get there? *Geological Society of America Abstracts with Programs*, **41**(7): 99.

# Index