A WEB SERVICES-BASED UNIVERSAL APPROACH TO HETEROGENEOUS FAULT DATABASES

QuakeSim lets scientists study earthquake behavior over single or multiple seismic cycles. The system's semantics-based database component, QuakeTables, provides global real-time accessibility to a diverse set of earthquake and fault data.

In the past decade, the availability of space-derived crustal deformation data has transformed the solid earth geophysics field. Global Positioning System (GPS) networks deployed globally provide precise time-dependent information on how the Earth's crust responds to earthquakes and plate-tectonic processes. Interferometric Synthetic Aperture Radar (InSAR) data reveal spatially dense information on how the Earth's crust deforms and how faults interact with each other.

Deformation of the Earth's crust and the interaction between earthquake faults is a complex 3D process. Understanding these processes requires sophisticated models and the use of high-performance computers. Our simulation system, QuakeSim (http://quakesim.jpl.nasa.gov), aims to help members of the seismological, crustal deformation, and tectonic communities develop an understanding of active tectonic and earthquake processes. The project's major science goal is to create a virtual laboratory to probe earthquake behavior. Its computational goal is to produce a functional system that fully models earthquake-related data.[1]

QuakeSim is a Web browser-based problem-solving environment that provides a set of links between newly available resources from NASA's Earth-observing systems, high-performance simulations, automated data mining, and more traditional tools. It's the first Web services-based, interoperable environment for creating large-scale forward models of earthquake processes.[2] A Web services-based portal provides global access to geologic reference models of faults and fault data, simple analysis tools, new parallel forward models, and visualization support.

Effective use of large data sets in the solid earth sciences will soon require cyberinfrastructure tools (www.cise.nsf.gov/sci/reports/toc.cfm). QuakeTables, a database system for handling both real and simulated data, provides input for earthquake simulation tools using fault data. Later, it will in-

1521-9615/05/$20.00 © 2005 IEEE
Copublished by the IEEE CS and the AIP

LISA B. GRANT AND MIRYHA M. GOULD
University of California, Irvine
ANDREA DONNELLAN
Jet Propulsion Laboratory
DENNIS MCLEOD, ANNE YUN-AN CHEN, AND SANG-SOO SUNG
University of Southern California
MARLON PIERCE AND GEOFFREY C. FOX
Indiana University
PAUL RUNDLE
University of California, Davis

clude other types of earthquake data as well. This article describes our Web-based universal approach to heterogeneous earthquake databases using QuakeTables to demonstrate the design, development, and implementation challenges of incorporating solid earth science data sets in high-performance computing simulations.

Data Management Problems

Earthquakes are generated by sudden fault movements, which produce seismic waves and induce 3D deformation of the Earth's surface. Earthquake science data sets include geographical and temporal fault data, regional deformation, and descriptive characteristics of the earthquakes themselves. These heterogeneous data sets reside in various distributed databases constructed for specific data types. Earthquake science data is heterogeneous, and the interpretations of some data types differ from resource to resource and from scientist to scientist.[3] Because existing databases have different information, structural organizations, and data formats, it's difficult to compare the results of simulations with input from different databases. The QuakeTables database system manages various types of earthquake science data and information, providing for its definition, storage, query, and control. We use QuakeTables content to cross-validate different simulation methods, explore competing theories of plate-boundary development, perform case studies with widely accepted assumptions, and provide input for visualization software to display simulation results.

Scientific data management problems are, of course, not limited to earthquake science: digital libraries supporting biology,[4] for example, have investigated similar issues. These projects share several requirements:

- *Annotation* lets users make comments on data sets.
- *Pedigree*, or *provenance*, tracks data origins and ownership with mechanisms such as automatic time-stamping and associations with users or particular data sources (ranging from publications to simulation codes). All entries should be traceable to their origins to assist users in determining data quality and in isolating potential errors.
- *Data curation* provides services for authorized groups to "bless" certain entries (and to revoke such blessings as appropriate), as required when mixing validated and nonvalidated data.
- *Access controls* restrict access on nonvalidated data sets to particular groups of researchers to prevent inadvertent or improper use of nonvalidated results. They also protect a collaborative group's results until they're ready for broader publication.

We designed QuakeTables specifically to handle these requirements. We've found, for example, that nonvalidated data sets containing simulated or unpublished data can be almost as useful in geophysical modeling as validated data sets. Geophysicists using modeling and simulation codes often want to compare simulation results after changing parameter settings and might wish to publish this data, along with associated results, to selected collaborators. QuakeTables allows this sort of limited publication (and sometimes retraction) of results, filling a gap in current geophysical database systems.

Web-Based Approach

In earthquake science, data sources can be observations, simulations, or hypotheses. Scientists can have their own interpretations and analyses of raw data, but data can be difficult to compile from distributed individual databases. Therefore, effective information retrieval and Web-based search for data of interest to a specific scientist requires a semantic metadata management system and Web service wrappers. Wrappers handle interface and data heterogeneity, whereas semantic metadata assists in information discovery and subsequent use (for example, scientists using their simulation model on fault data from another source).

Representing and extracting semantic meaning from information content is thus essential. Figure 1 shows an initial domain ontology (a description of key concepts and interrelationships in the domain) developed by computer scientists and earthquake science experts.

Our approach incorporates an object-based classified database model—the Classified Interrelated Object Model (CIOM)[5]—to structure a domain-dependent ontology for representing semantic information[6–8]—that is, information about a statement's or fact's meaning. To represent and understand the meaning of interrelationships, CIOM provides enriched semantics primitives (types of interrelationships the system understands) such as *subclass* (special kind of), *attribute* (property of), and *inverse* (inverse of a property), grouping classes that are second-order collections—namely, classes of classes and instances (specific fact occurrences).[5]

Ontronic[9] is our ontology-based metadata management system that supports CIOM in structuring semantic information to construct, refine, and expand ontologies. It supports analysis of the data sources and specifies the concepts and interrelationships among the concepts to establish a domain ontology.[5] Ontronic created the

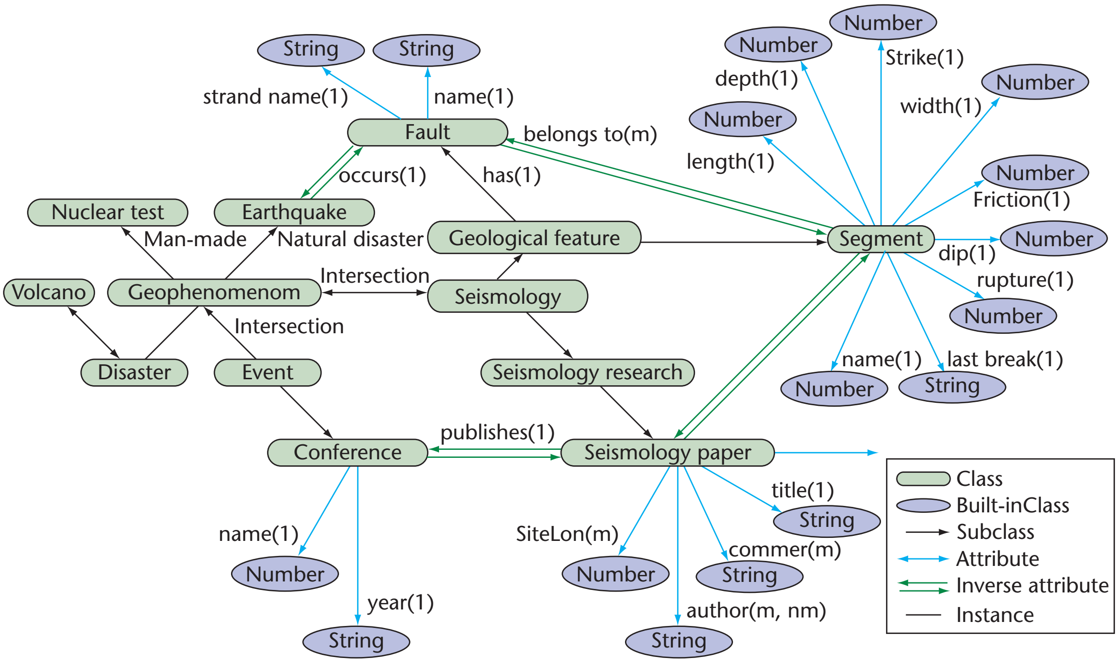


Figure 1. A seismology domain ontology (created by Ontronic).[5] The two roots in the example—event and seismology—can be further categorized. Seismology, for example, can be specialized as Geophenomenon, Seismology research, or Geological feature, and Event can be subcategorized as Disaster, Conference, or Geophenomenon. Each subclass can be further specialized, as shown.

seismology-domain ontology in Figure 1 using QuakeTables parameters and seismology-domain knowledge.

The limitation of accessibility is a problem in current heterogeneous fault databases. The properties of observatory earthquake and fault data are enormous and temporal-based. Sharing the abundant and valuable data is difficult because of the diversity of data formats, storage, and organization. For example, data can be saved as plain text, with user-defined file types, or in various database management systems.

To break the barrier of accessibility due to heterogeneity, we use a universal approach based on Semantic Web and Web services technologies. This lets scientists access the abundant data in heterogeneous databases with minimal delay and without having to deal directly with formats or system platforms. A system based on this universal approach must, of course, provide integration portability to manage interoperability for heterogeneous data.

Tools for Global Accessibility

To achieve real-time global accessibility, data transmission via the Internet must be based on a lightweight protocol—a transmission agreement between the server and client machines with minimal overhead to reduce transmission time. Web services technologies minimize the data transmission overhead by using an XML-based protocol and schema of interface definitions to invoke the applications among servers and clients. We implement the service interface using a Web-friendly programming language such as Java or Python. Web services let platforms and applications exchange information and make remote application invocation possible.[10]

XML is a key Web services technology. We use XML schemas to describe different data sources' metadata. XML schemas specify metadata's structure (such as the elements or concepts within the structure and the relations among these elements or concepts) and define each element's or attribute's data type.

Users have different requirements for retrieving data. Using Web services technologies, client stubs developed according to user requirements let users request information on literature references, retrieve data for use in graphical simulations with virtual reality tools, and collect data from several resources for experiments. The user-friendly, easily accessible, and browser-based authorized interfaces manage the databases and provide access at different levels of abstraction. The interface designs are based on the concurrent efforts of scientists and engineers.

For support to SOAP, we use the Web services

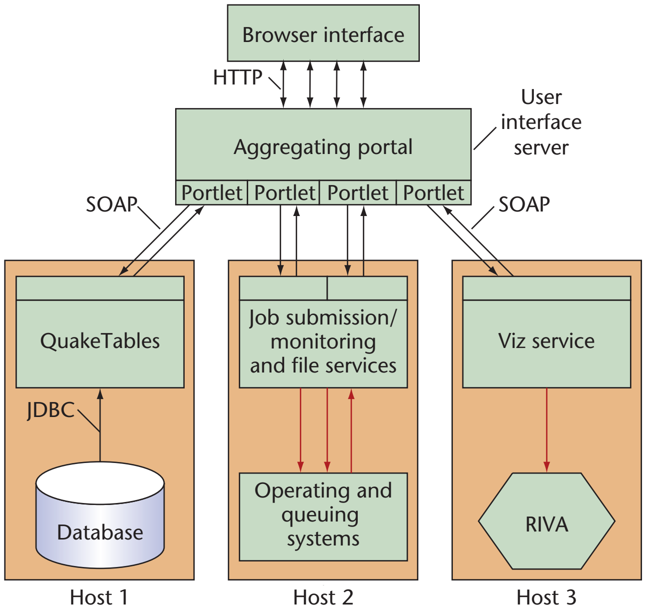


Figure 2. QuakeSim portal and service architecture. The database is the QuakeTables database; RIVA is the Remote Interactive Visualization and Analysis System. RIVA can be used as an interactive system to explore and visualize large terrain data sets in 3D perspective views, or as an animation tool to generate fly-by movies using high-resolution images and digital elevation.

client stubs for earthquake simulations. For support to HTTP, we've implemented a browser-based user interface. We developed a basic search using an HTTP-based wildcard search engine as well as an intelligent search engine for fault data: users can enter an author name, a fault name, or a title keyword to search for fault data. The search engine's intelligence lets it accept a partial string of one author or one partial fault name as input. It displays the query results as a list of data entries that includes the attribute values of the (partial) author name or fault name requested by the user.

QuakeTables Database

The need for compilations of fault data for seismic hazard analysis has existed for a long time, and scientists have constructed several databases for this purpose. Most existing databases[11–14] provide input for probabilistic seismic hazard assessment (PSHA). Because they're text-based, these databases are generally neither accessible nor structured for numerical simulations and modeling of earthquake processes. Although the recently released US Geological Survey (USGS) Quaternary Fault and Fold Database[14] contains a wealth of validated fault data, for example, much of it is descriptive text that can't be input to simulation codes without labor-intensive effort. In addition, fault attributes or parameters that are useful for seismic hazard assessment might differ from parameters required for tectonic modeling and understanding active deformation processes at varying temporal and spatial scales. Most faults in existing databases, for example, are divided into characteristic segments that are expected to rupture as a unit. Geologic slip rates refer to entire faults or segments rather than to the specific locations (geographic coordinates) where they were measured. These simplifications are useful for seismic hazard analysis, but they introduce subjective interpretations that could bias the results of fault behavior simulations over different time scales.

To reduce this problem, the QuakeTables database includes both primary and interpreted or subjective "nonprimary" fault parameters.[3] Geologic fault parameters include fault location and geometry, such as dip angle. The database also contains paleoseismic data about active faults' activity and earthquake history. Paleoseismic data describe fault activity over time scales of tens to thousands of years. Primary paleoseismic data parameters include measurements of fault slip rate, dates and locations of previous ruptures, earthquake recurrence intervals, and the amount of fault displacement per earthquake. Nonprimary fault parameters include characteristic segment definitions and characteristic or prehistoric rupture magnitudes.

System Architecture

Figure 2 shows how QuakeTables relates to the QuakeSim system architecture. Figure 3 shows a simplified extended entity relationship (EER) schema for the initial QuakeTables database. We developed the parameters to provide input to QuakeSim model codes. Documentation describing a relational implementation of the database is available on the QuakeSim Web page (http://quakesim.jpl.nasa.gov/).

We use MySQL, a commercially available general-purpose database management system, to support QuakeTables. The system runs on a PC under Linux and supports the definition, storage, access, and control of collections of structured data. We implement the HTTP-based application program interfaces (APIs) using HTML and JavaScript, and format the user interfaces as forms to provide simple but sufficient functions. For the SOAP-based API, we use Java-based technologies to implement the client stubs. Users familiar with SQL, the database query language, have more power to manage the data with client stubs.

QuakeTables accommodates several types of fault

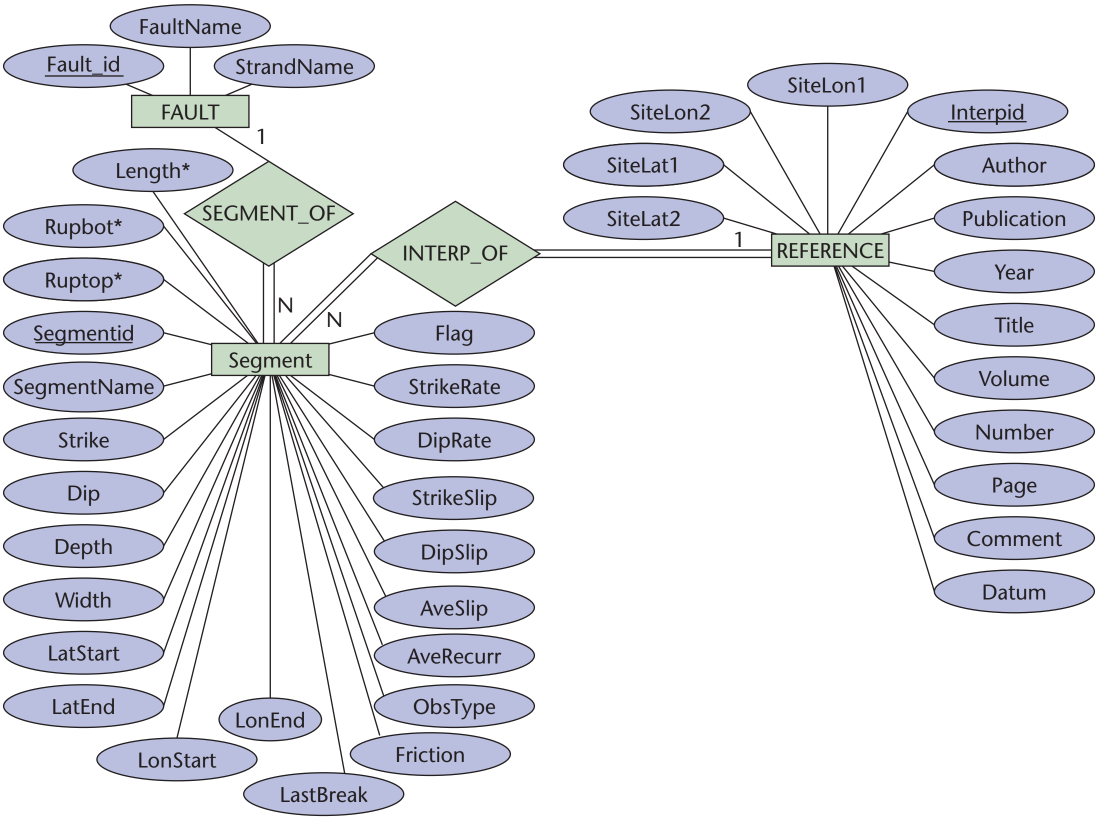


Figure 3. A simplified extended entity relationship model specification for the QuakeTables fault database shows relationships among attributes and faults.

data and data sets, as well as simulated or hypothetical data. There are pre-existing collections with Web-based access interfaces; there are also some structured collections managed by general-purpose database management systems. QuakeTables lets users characterize dynamically defined earthquake faults and includes material rectilinear-layer parameters for 3D tectonic deformation modeling.

The QuakeSim system is operational, is geographically extensible, and has proven useful for earthquake research. QuakeTables contains data from California faults, but no geographic restriction exists for future data entries. We extracted the data in QuakeTables from refereed journal articles, professional papers, professional reports, and conference abstracts.

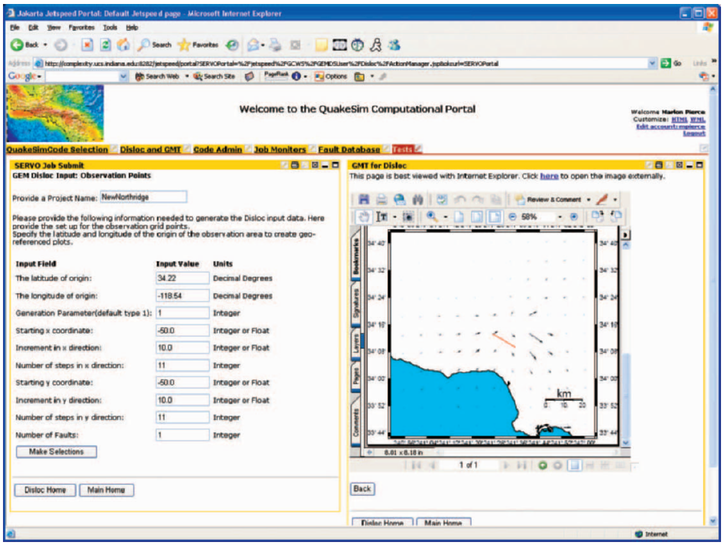
QuakeTables also contains paleoseismic data from major faults as well as three structured data sets: a recent version of Virtual California[15] and two fault databases[12,13] published by the California Geological Survey (CGS) and the USGS for seismic hazard analysis. These structured data sets provide geographic coordinates, geometry, and summary attributes for many active faults and fault segments in California. We'll add more paleoseismic data from research publications in the future.

Web Portal Access

QuakeSim offers several high-end computing simulation tools and application programs (available at http://complexity.ucs.indiana.edu:8282/jetspeed/index.jsp). You can access the QuakeSim portal from the QuakeSim homepage (http://quakesim.jpl.nasa.gov/).

Figure 2 is a diagram of the portal and its three-tiered architecture. Web portals are a common approach for providing user-friendly interfaces for launching and controlling complex chains of codes and data or for simply downloading query results. These portals are sophisticated user environments in their own right, providing many services for managing the user experience.

Users interact with QuakeSim through the Web browser interface. The browser connects to an aggregating portal[2,9,10,16] running on the user interface server. This portal collects and manages dynamically generated Web pages that can be developed independently of the portal and run on



Jakarta Jetspeed Portal: Default Jetspeed page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
http://complexity.ucs.indiana.edu:8282/jetspeed/portal?SERVOPortal=%2Fjetspeed%2FGCWS%2FGEMDSUser%2FDisloc%2FActionManager.jsp&okurl=SERVOPortal
Welcome to the QuakeSim Computational Portal
Welcome Marlon Pierce
Customize: HTML WML
Edit account: mpierce
Logout
QuakeSimCode Selection
Disloc and GMT
Code Admin
Job Monitors
Fault Database
Tests
SERVO Job Submit
GMT for Disloc
Review & Comment
58%
km
0 10 20
8.01 x 6.18 in
1 of 1
Back
Internet

GEM Disloc Input: Observation Points

Provide a Project Name: NewNorthridge

Please provide the following information needed to generate the Disloc input data. Here provide the set up for the observation grid points.
Specify the latitude and longitude of the origin of the observation area to create geo-referenced plots.

Input Field	Input Value	Units
The latitude of origin:	34.22	Decimal Degrees
The longitude of origin:	-118.54	Decimal Degrees
Generation Parameter(default type 1):	1	Integer
Starting x coordinate:	-50.0	Integer or Float
Increment in x direction:	10.0	Integer or Float
Number of steps in x direction:	11	Integer
Starting y coordinate:	-50.0	Integer or Float
Increment in y direction:	10.0	Integer or Float
Number of steps in y direction:	11	Integer
Number of Faults:	1	Integer

Make Selections

Disloc Home | Main Home

This page is best viewed with Internet Explorer. Click here to open the image externally.

Figure 4. Browser interface for an example QuakeSim portal component. The screenshot shows two user interfaces, a visualization of Disloc output on the right and the code input interface on the left. The interfaces are independent Web pages pulled into the aggregating portal. Users can navigate to other portlet component displays using the tabs across the top.

separate servers. Portlets manage particular Web site connections.

Figure 4 is an example user interface featuring output from Disloc, a program that models dislocations resulting from movement of a fault. Disloc also handles multiple arbitrarily dipping dislocations (faults) in an elastic half-space to produce surface displacements.

The QuakeTables architecture, both in its current form and with planned future enhancements, provides programmatic and human access to fault data. By adopting an OpenGIS-based Web service architecture, we separate the logic and implementation of data representation and access from the application layer that's used to build client programs. We can also use these client applications as application interfaces or as remote procedure calls embedded in geophysical application codes. Or, the applications can be human interfaces in which the system delivers the data product directly to the end user, who can then incorporate the results in offline or desktop applications.

In the project's next phase, we'll use Web (Grid) service technology to demonstrate the assimilation of multiple distributed data sources into a major parallel high-performance computing earthquake-forecasting code. A key design requirement is environmental support for defining and using interfaces between components that address different scales of the earthquake process, which can range from continental to a grain of sand. Support for such processes might include special data operations such as filtering and parameterization.

We'll also add GPS, InSAR, and other geophysical data types, as well as additional fault data. Current work involves developing a metaontology: a federation of semantic specifications for various types of geophysical data. Future work involves developing user-friendly interfaces to find, browse, extract, and use data from various sources. With these improvements, the QuakeSim project and QuakeTables database component will continue to provide cyberinfrastructure tools for modeling regional deformation and earthquake processes in a Web services environment. CiSE

Acknowledgments

We thank Gabriela Noriega-Carlos and Maggie Ta for their contributions. The Computational Technologies Program of NASA's Earth-Sun System Technology Office funds the QuakeSim project. We performed portions of this work at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA, and additional work at our respective facilities. The QuakeSim project includes collaboration with Brown University, Indiana University, the University of California, Davis, the University of California, Irvine, and the University of Southern California.

References

1. G.C. Fox et al., "Introducing a New Paradigm for Computational Earth Science—A Web-Object-Based Approach to Earthquake Simulations," *GeoComplexity and the Physics of Earthquakes*, J. Rundle, D. Turcotte, and W. Klein, eds., Am. Geophysical Union, 2000, pp. 219–245.
2. M.E. Pierce, C. Youn, and G. Fox, "Interacting Data Services for Distributed Earthquake Modeling," *Proc. Int'l Conf. Computational Science*, LNCS 2659, Springer-Verlag, 2003, pp. 863–872.
3. L.B. Grant and M.M. Gould, "Assimilation of Paleoseismic Data for Earthquake Simulation," *Pure and Applied Geophysics*, vol. 161, no. 11/12, 2004, pp. 2295–2306.
4. C.A. Goble et al., "Knowledge Integration: In Silico Experiments in Bioinformatics," *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufman, 2003, pp. 121–134.
5. S. Sung and D. McLeod, *Ontology-Based Semantic Information Management for Seismology and Geoscience*, tech. report imsc-05-002, Integrated Media Systems Center, Univ. of Southern Calif., 2005.
6. A.Y. Chen et al., "Interoperability and Semantics for Heterogeneous Earthquake Science Data," *Proc. Semantic Web Technologies for Searching and Retrieving Scientific Data Conf.*, CEUR, 2003; http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-83/sia_5.pdf.
7. L. Khan, D. McLeod, and E. Hovy, "Retrieval Effectiveness of an

Ontology-Based Model for Information Selection," *VLDB J.*, vol. 13, no. 1, 2004, pp. 71–85.

8. G. Aslan and D. McLeod, "Semantic Heterogeneity Resolution in Federated Database by Metadata Implantation and Stepwise Evolution," *VLDB J.*, vol. 18, no. 2, 1999, pp. 120–132.
9. C. Youn, M.E. Pierce, and G. Fox, "Building Problem Solving Environments with Application Web Service Toolkits," *Proc. Int'l Conf. Computational Science*, LNCS 2660, Springer-Verlag, 2003, pp. 403–412.
10. M.E. Pierce et al., "Interoperable Web Services for Computational Portals," *Proc. Supercomputing*, ACM Press, 2002, pp. 1–12.
11. Working Group on California Earthquake Probabilities, "Seismic Hazards in Southern California: Probable Earthquakes, 1994–2024," *Bulletin Seismological Soc. of Am.*, vol. 85, no. 2, 2001, pp. 379–439.
12. M.D. Petersen et al., *Probabilistic Seismic Hazard Assessment for the State of California*, open-file report, US geological survey no. OF 96-0706, 1996.
13. A.D. Frankel et al., *Documentation for the 2002 Update of the National Seismic Hazard Maps*, open-file report, US geological survey no. OF 02-0420, 2002.
14. K.M. Haller et al., "US Quaternary Fault and Fold Database Released," *EOS,* vol. 85, no. 22, June 2004; www.agu.org/eos_elec/000655e.html.
15. J.B. Rundle et al., "GEM Plate Boundary Simulations for the Plate Boundary Observatory: A Program for Understanding the Physics of Earthquakes on Complex Fault Networks via Observations, Theory, and Numerical Simulation," *Pure and Applied Geophysics*, vol. 159, no. 10, 2002, pp. 2357–2381.
16. D. Gannon et al., "Building Grid Portal Applications from a Web-Service Component Architecture," *Proc. IEEE*, vol. 93, no. 3, Mar. 2005.

Lisa B. Grant is an assistant professor in the Department of Environmental Health, Science, and Policy at the University of California, Irvine. Her research interests include earthquake geology and active faults, environmental significance of earthquakes, and computational earthquake science. Grant has a PhD in geology and geophysics from the California Institute of Technology. She is a member of the American Geophysical Union, the Seismological Society of America, and the Geological Society of America. Contact her at lgrant@uci.edu.

Andrea Donnellan is deputy manager of the Science Division at the Jet Propulsion Laboratory. Her research interests include computational earthquake science, GPS, and InSAR measurement of crustal deformation. Donnellan has a PhD in geophysics from the California Institute of Technology. She is a member of the American Geophysical Union. Contact her at donnellan@jpl.nasa.gov.

Dennis McLeod is a professor of computer science at the University of Southern California and a strategic scientist at the USC Integrated Media Systems Center. His research interests include database system interoperation and networking, multimedia information modeling and sharing, database system user interfaces, and applied machine learning. McLeod has a PhD in computer science from the Massachusetts Institute of Technology. He is a member of the ACM and the IEEE Computer Society. Contact him at mcloed@pollux.usc.edu.

Marlon Pierce is a senior research associate in the Community Grids Lab at Indiana University. His research interests include developing tools for computational science based on emerging Internet and computational grid technologies. Pierce has a PhD in physics from Florida State University. Contact him at mpierce@cs.indiana.edu.

Geoffrey C. Fox is a professor of computer science, informatics, and physics at Indiana University. His research interests include basic technology for Grid computing and its application to earthquake science, distance education, complex systems, and particle physics. Fox has a PhD in theoretical physics from Cambridge University. He is a member of the IEEE and the ACM. Contact him at gcf@indiana.edu.

Anne Yun-An Chen is a PhD candidate of computer science and a member of the Semantic Information Representation Group at the University of Southern California. Her research interests include federated database systems, Web services, and information recommender systems. Chen has an MS in computer science from the University of Southern California. She is a member of the IEEE and the ACM. Contact her at yunanche@usc.edu.

Miryha M. Gould is a PhD student in the School of Social Ecology at the University of California, Irvine. Her research interests include earthquake geology and environmental psychology. Gould has an MA in social ecology from the University of California, Irvine. Contact her at miryha@uci.edu.

Sang-Soo Sung is a PhD student of computer science at the University of Southern California and a member of the Semantic Information Representation Group. His current research focuses on ontology-based information fusion and management. Sung has a BS in computer science from the Korea University in the Republic of Korea. Contact him at sangsung@usc.edu.

Paul Rundle is an undergraduate at the University of California, Davis. His research interests include methods of data assimilation and numerical simulation as they pertain to large-scale fault system models. Contact him at paul_rundle@hmc.edu.