Geoinformatics: Transforming data to knowledge for geosciences

A. Krishna Sinha*, *Dept. of Geosciences, Virginia Tech, Blacksburg, Virginia 24061, USA;* ***Zaki Malik,*** *Dept. of Computer Science, Wayne State Univ., Detroit, Michigan 48120, USA;* ***Abdelmounaam Rezgui,*** *School of Information Sciences, Univ. of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA;* ***Calvin G. Barnes,*** *Dept. of Geosciences, Texas Tech Univ., Lubbock, Texas 79409, USA;* ***Kai Lin,*** *San Diego Supercomputer Center, Univ. of California, San Diego, California 92093, USA;* ***Grant Heiken,*** *331 Windantide Place, Freeland, Washington 98249, USA;* ***William A. Thomas,*** *Dept. of Earth and Environmental Sciences, Univ. of Kentucky, Lexington, Kentucky 40506, USA;* ***Linda C. Gundersen,*** *U.S. Geological Survey, National Center, Reston, Virginia 20192, USA;* ***Robert Raskin,*** *NASA Jet Propulsion Laboratory, 300-320, Pasadena, California 91109, USA;* ***Ian Jackson,*** *British Geological Survey, Nottingham NG12 5GG, UK;* ***Peter Fox, Deborah McGuinness,*** *Dept. of Computer Sciences, RPI, Troy, New York 12180, USA;* ***Dogan Seber**,*** *San Diego Supercomputer Center, Univ. of California, San Diego, California 92093, USA; and* ***Herman Zimmerman,*** *National Science Foundation (ret.), 1337 NE Stanton Street, Portland, Oregon 97212, USA*

ABSTRACT

An integrative view of Earth as a system, based on multidisciplinary data, has become one of the most compelling reasons for research and education in the geosciences. It is now necessary to establish a modern infrastructure that can support the transformation of data to knowledge. Such an information infrastructure for geosciences is contained within the emerging science of geoinformatics, which seeks to promote the utilization and integration of complex, multidisciplinary data in seeking solutions to geoscience-based societal challenges.

INTRODUCTION

Over the centuries that humankind has been studying Earth, oceans, and sky, data were gathered toward explaining the physical phenomena of our surroundings. Understanding such events as eclipses, tides, volcanism, and earthquakes was challenging because of the difficulty of organizing observations within scientific frameworks that could provide an integrative understanding of these phenomena. Pioneers of the earth sciences, such as geologists Lyell (1797–1875) and Hutton (1726–1797), made multidisciplinary observations in stratigraphy, paleontology, and petrology, stored their observations in logbooks, and visualized them through interpretive products, such as maps and cross sections. We continue to conduct our science in similar ways. We make observations on the ground and through remote sensing techniques and store the information in computers, but we still find it difficult to achieve an integrative understanding of complex natural phenomena. The ability to find, access, integrate, and properly interpret data sets has been hampered by the expanding volumes and heterogeneity of the data. With the help of computer scientists, transformative advances in the geosciences are now possible through innovative approaches to interoperability, analysis, modeling, and integration of heterogeneous databases. This geoinformatics effort would require Web-based availability of data and applications, thereby removing geographic or political boundaries. Geoinformatics will give us the ability to encompass a variety of temporal and spatial scales, integrate heterogeneous data, and visualize data and analytical results.

WHAT IS GEOINFORMATICS?

Geoinformatics is an informatics framework for the discovery of new knowledge through integration and analysis of earth-science data and applications. Fostered by support from both national and international agencies, geoinformatics has emerged to address the growing recognition that problems with significant societal implications require integrative and innovative approaches for analysis, modeling, managing, and archiving of extensive and diverse data sets. In the United States, geoinformatics emerged as an initiative within the National Science Foundation (NSF) Division of Earth Sciences and other federal agencies, such as the U.S. Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). The impetus was the wide consensus that existing information management infrastructures were inadequate to cope with the complexities of earth processes.

Foundation technologies constitute the base infrastructure required to facilitate geoinformatics. These technologies include resources for communication, storage, and computation. Consequently, geoscientists are now better equipped (e.g., high-performance computing) to efficiently address complex questions. However, the true potential of these technologies can only be realized by enhancing our data- and application-management capabilities (shown as the geoinformatics components in Fig. 1). For instance, standards are needed for the exchange and understanding of data (e.g., shared data models, markup languages, ontologies, etc.), visualization, and computation. Data analysis

E-mails: Sinha: pitlab@vt.edu; Malik: zaki@wayne.edu; Rezgui: arezgui@sis.pitt.edu; Barnes: cal.barnes@ttu.edu; Lin: klin@sdsc.edu; Heiken: heiken@whidbey.com; Thomas: geowat@uky.edu; Gundersen: lgundersen@usgs.gov; Raskin: rob.raskin@jpl.nasa.gov; Jackson: ij@bgs.ac.uk; Fox: pfox@cs.rpi.edu; McGuinness: dlm@cs.rpi.edu; Seber: seber@nrc.gov; Zimmerman: hzimmerm@comcast.net.

*Adjunct, Dept. of Geological Sciences, San Diego State Univ., San Diego, California 92182, USA.

**Now at Nuclear Regulatory Commission, One White Flint North, Rockville, Maryland 20852, USA.

GSA Today, v. 20, no. 12, doi: 10.1130/GSATG85A.1

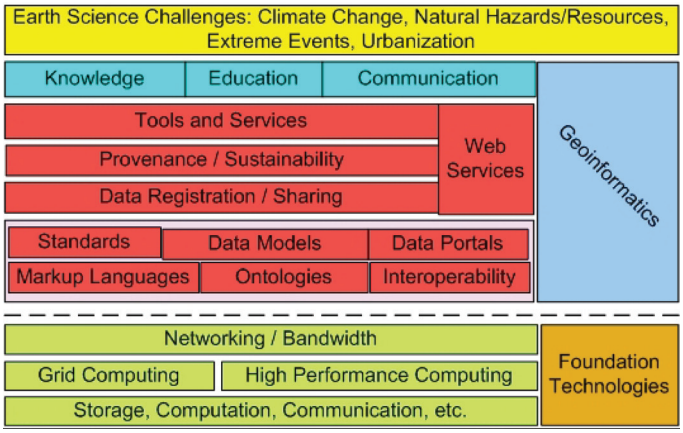

Figure 1. Representation of information technology intensive (green) and geoscience–computer science intensive (red) components for the emerging science of geoinformatics. Content adapted from Atkins et al., 2003. (See glossary, Appendix A.)

tools and services must be made Web-accessible; portals must enable easy location of registered data and services; data providers must retain ownership rights and credit through tracking of data sources and services (provenance); and, most importantly, these advances should be communicated and shared with the broader community (Simmhan et al., 2005).

Scientists facing the global challenges of climate change, natural hazards, and the discovery and management of natural resources will benefit greatly from an expanded integration of informatics into the geosciences. In this brief overview, we emphasize integration of data and services to meet such challenges. For example, management and discovery of natural resources requires many data types, such as geologic maps, geochronology, petrology, and geochemistry of fluids and solids, as well as access to ore deposit models. The ability to discover and incorporate these data into new, robust models for ore genesis would lead to an integrative view and make exploration much more efficient.

WHY DO WE NEED GEOINFORMATICS?

Communities of scientists around the world are working toward the goal of discovering new knowledge through a better understanding of the fundamental principles that underlie complex and heterogeneous data—a foundation for why the data values are what they are or an indication as to how the data would change over time through physical, chemical, and biological processes. Geoinformatics will support the next generation of knowledge discovery, markedly broaden our understanding of science and engineering, and allow us to solve challenging and complex problems previously unimagined.

There is common consensus that access to and integration of data are prerequisites for creating an information infrastructure. In addition, we argue that in order to fully exploit data in the pursuit of knowledge discovery and transformative science, new semantic models are needed to integrate scientific processes and methods within such an infrastructure. The semantic stages scientists follow on the pathway from data to knowledge and beyond involve seeking information as it relates to description, definition, or perspective (what, when, where) followed by derivation of knowledge, which comprises strategy, practice, method, or approach (how). These stages lead to new insight into fundamental principles (why).

The lack of a robust informatics infrastructure for sharing data and knowledge across all scientific disciplines has become a major hindrance to productivity, especially in multidisciplinary research (Atkins et al., 2003). Community-specific knowledge creation requires intra- and inter-community integrative capabilities. However, integrating and using data acquired by different investigators can be difficult. This is primarily because each data set uses heterogeneous schema and semantics. Such heterogeneities can be divided into three categories: syntactic, structural, and semantic (Sheth, 1998). Syntactic and structural transformation (e.g., database mediation) can be used to handle the first two kinds of heterogeneities but are not adequate for resolving semantic differences. The use of ontologies is considered a possible solution for the semantic heterogeneity problem (McGuinness, 2003).

We present two examples that demonstrate the current use of semantics for access and integration of an array of geologic data types and formats. Our purpose is to highlight the advantages of what may be considered elaborate semantics-based approaches to provide solutions for complex problems.

1. OneGeology (www.onegeology.org) is an international collaboration working to develop and serve a Web-accessible, worldwide geological map data set at a scale of 1:1,000,000. Its objective is to utilize community-endorsed standards for syntactic interoperability that enhance the use of existing data. To achieve this goal, the program has developed a data exchange model called GeoSciML (Commission for the Management and Application of Geoscience Information, 2008) that provides a controlled vocabulary within a common conceptual model. Such a model allows common description of geologic features leading to interoperability through a markup language for data interchange for the discovery and utilization of globally distributed geoscience data and information. GeoSciML is a critical first step in the use of informatics-based technologies (Simons et al., 2006).

2. Ontology-Enabled Map Integrator (OMI), developed at the San Diego Supercomputer Center (Lin and Ludäscher, 2003), utilizes ontologies for registering geologic data sets to assist in integrating and querying heterogeneous data. Although this system was implemented for integration of data associated with geologic maps, it is a geoscience breakthrough in regard to the use of ontologic capabilities for discovery and integration. Each data set is registered ("mapped") to an ontology-based association before it becomes available in a Web environment. The process of data registration semi-automatically generates mapping from data sets to existing ontologies; these mappings are then available to software applications that may be used to explore and extract information from diverse data arrays.

GEOSCIENCE-BASED SOCIETAL AND RESEARCH CHALLENGES

An Example of Cities at Risk and Volcanic Hazards

Sixty-three cities worldwide are situated near potentially active volcanoes and have populations of more than 100,000,

including two mega-cities with a combined population of more than 50 million. Thus, there is a great need to understand volcanic processes through pattern recognition and epidemiological forecasting. The need for informatics in hazard mitigation is evident in the data sets generated by disciplines represented at the International Association of Volcanology and Chemistry of the Earth's Interior's (IAVCEI) biannual conferences ("Cities on Volcanoes"). An informatics-based solution makes the integrative process across geoscience disciplines (and others) efficient, accurate, and cost-effective, thus making possible the discovery of new critical knowledge not accessible via manual analysis of data. For instance, (1) epidemiological data models enable comparisons with similar recorded events in real time, and (2) volcano visualizations and mining of data associated with volcano product characterizations facilitate efficient hypothesis formation and evaluation.

The example of cities at risk illustrates the need for integrative, multidisciplinary access to research-based data products. A host of other societally significant initiatives has similar needs; two examples are the joint USGS and Chinese Qingdao Institute for Marine Geology project on management of delta ecosystems (Delta Research and Global Observation Network) and the UK's Environment and Urban Regeneration Program for development of 3- and 4-dimensional (4-D), high-resolution shallow (first 200 m) subsurface models to aid assessment of urban risks associated with natural and anthropogenic ground instability, pollution, and flooding.

Basic research in geoscience also benefits from semantics-based geoinformatics. For example, construction of a 4-D, kinematically balanced, palinspastic restoration of a continental margin orogenic belt and foreland also requires geoinformatics-based solutions to gain a more robust understanding of geologic processes. The necessary first step in interdisciplinary integrative research is data discovery. The current method of Web-based data discovery (mainly through search engines) requires sifting through a large number of Web pages. Also, because human interaction is required, integration normally results in the "layering of data" through a GIS system to retrieve new information (e.g., Takarada et al., 2007). Alternatively, the user must create a data integration layer to capture the location, format, and structure of the underlying data leading to a logical view. This activity requires the adoption of a common data model (e.g., North American Data Model [Boisvert et al., 2003]). Such techniques are effective but laborious and not the most rational and efficient way to analyze complex information (Doan and Halevy, 2005).

The main impediment to data discovery and integration is the lack of semantics to enable machines to "understand" and "automatically" process the data that they now merely display (Cardoso and Sheth, 2006). Figure 2 shows the different types and levels of interoperability leading to integration through semantics-based techniques. For example, taxonomy can classify information hierarchically without defining the nature of connections, while a thesaurus contains associations with semantic constraints. Both levels of semantic models are for standard classification schemes in a single discipline (e.g., rock classification [one-dimensional]) and are unable to represent and interoperate across multiple dimensions and/or varied conceptual models (Obrst, 2003; McGuinness, 2003). The more

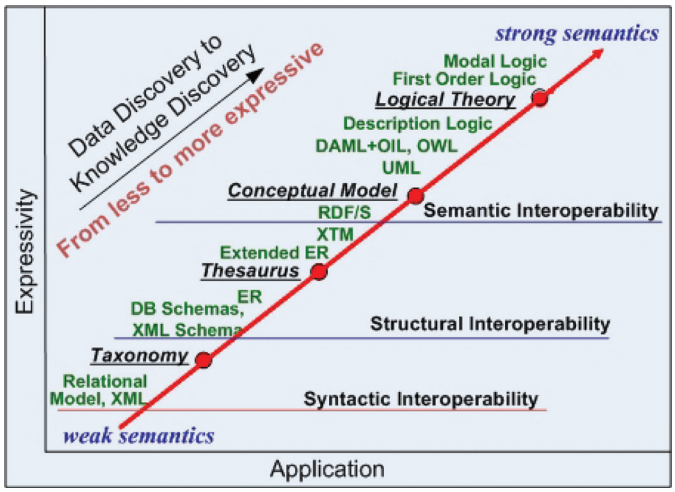

Figure 2. Multiple levels of semantics and associated interoperability capabilities (from Obrst, 2003). Increasing interoperability services requires increasing community agreement on conceptual relationships across participating geoscience disciplines. Strong semantics allow inferences from dataset contents. Terms defined in Appendix A.

expressive semantics, in the form of ontologies, are underpinned by logical theories and provide increased capabilities for deductions and inferences based on known associations and rules (Baader et al., 2004; Sinha et al., 2006). Enabling software tools and languages, such as XML (W3C, 2003), RDF (W3C, 2004a), and OWL (W3C, 2004b; McGuinness and Harmelen, 2004), allow interoperability at increasing levels of semantics (i.e., from weak to strong), resulting in a transition from data to knowledge. We endorse the definition of knowledge discovery as a nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al., 1992).

To enable strong semantic interoperability, current research emphasizes ontology-based data registration, discovery, and integration (Obrst, 2003; Noy, 2004; Raskin, 2006; Malik et al., 2007a; Fox et al., 2008). The primary purpose of ontologies (e.g., Noy and McGuinness, 2001) is to provide an organizational structure for automated data discovery and automated inferencing capabilities (Baader et al., 2004). For example, a relationship between the occurrence of ignimbrites and hazardous volcanic eruptions can be inferred by an automated reasoning system even though this fact is not contained in the database, but only if the ontologic framework effectively captures such a relationship (Fig. 3). The conceptual relationships are based on the ontologic relationships: (1) ignimbrite *is a* pyroclastic rock *is a* volcanic rock *is a* rock; (2) a hazardous eruption *is an* explosive eruption *is an* eruption; and (3) an explosive eruption *has material* pyroclastic rocks; therefore, ignimbrites are a product of hazardous volcanic eruptions.

Recognizing the significance of semantics, we see the future as a virtual environment that allows science communities to go beyond data discovery toward modeling and understanding processes through shared data and services. We recognize the need to establish a tripartite semantic infrastructure for automated discovery, analysis, utilization, and understanding of data (through both inverse and forward modeling capabilities), leading to new knowledge. This infrastructure will consist of

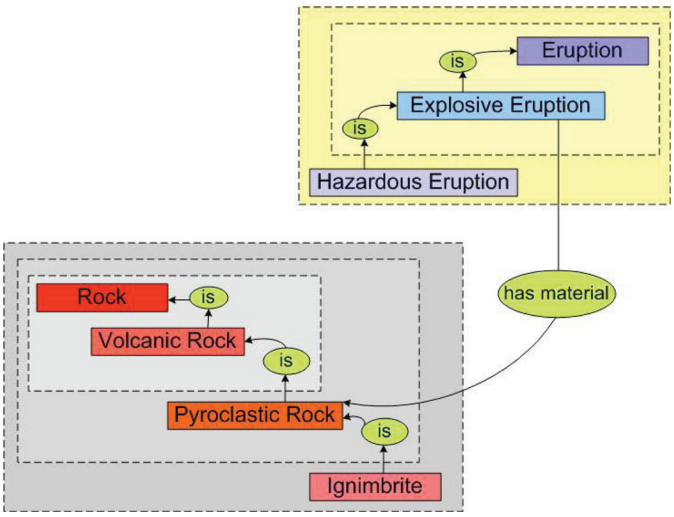

Figure 3. Graphical representation of an ontology leading to automated capability of logical deduction through defined taxonomies and inference rules.

three categories of ontologies: objects (e.g., materials), processes (e.g., chemical reactions), and services (e.g., simulation models). Objects represent our understanding of the state of a system when the data were acquired, whereas processes capture the forcings on the objects that may lead to changes in state over time (Sinha et al., 2006). Service ontologies would enable appropriate tools for computation and visualization to be discovered as Web services. Such a semantic model would provide crucial machine-interpretable information to the knowledge discovery process.

Object ontologies exist at many levels of abstraction and are often related to a tiered structure composed of upper-level, mid-level, and domain ontologies (Semy et al., 2004). Upper-level ontologies, such as Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001) and Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Masolo et al., 2003), provide a conceptual framework for developing domain ontologies, leading to interoperability, automated inference, and natural language processing. For example, a geoscience ontology being developed as a mid-level ontology (Malik et al., 2007a) could eventually contain all possible geoscience terms and their associations, similar to the well-developed semantic capabilities in bioinformatics (Stevens et al., 2004).

The use of existing ontologies (e.g., SWEET ontology library [Raskin and Pan, 2005], which contains numeric, time, and units ontologies) will accelerate the development of additional subject-specific ontologies in the geosciences (e.g., Ramachandran et al., 2006; Sinha et al., 2007; Tripathi and Babaie, 2008). Thus, we envision community-supported ontologies that would enable automated discovery, analysis, utilization, and understanding of data through both induction and deduction along the pathway from data to knowledge and ultimately to insight of scientific principles. We emphasize that through technologies such as ontology mappings (Fensel, 2004) it is possible to share ontologic frameworks within and across scientific communities, regardless of consensus level. For example, rock classification schemes used by the British Geological Survey and the Geological Survey of Canada are dissimilar, but a user can still map the concepts of one to the other based on either classification scheme.

The semantic interoperability problems of data discovery and integration are similar to those associated with the use of geoscientific services (e.g., visualization or modeling codes), which have experienced limited re-use because of differences in operating systems, formats, etc. The Web Services Initiative undertaken by the World Wide Web Consortium (W3C) is a step toward resolving the problem of service-sharing across computing environments (Alonso et al., 2003). A Web service user need not be concerned with the operating systems, development language environments, or component models used to create or access the service. Therefore, tools and services developed by geoscientists can be wrapped as Web services registered to a service and process ontologies and made accessible to the scientific community at large.

Figure 4 shows a software system architecture for organizing geoscientific data and tools through ontologies. Registration to ontologies of these data and tools as Web services would

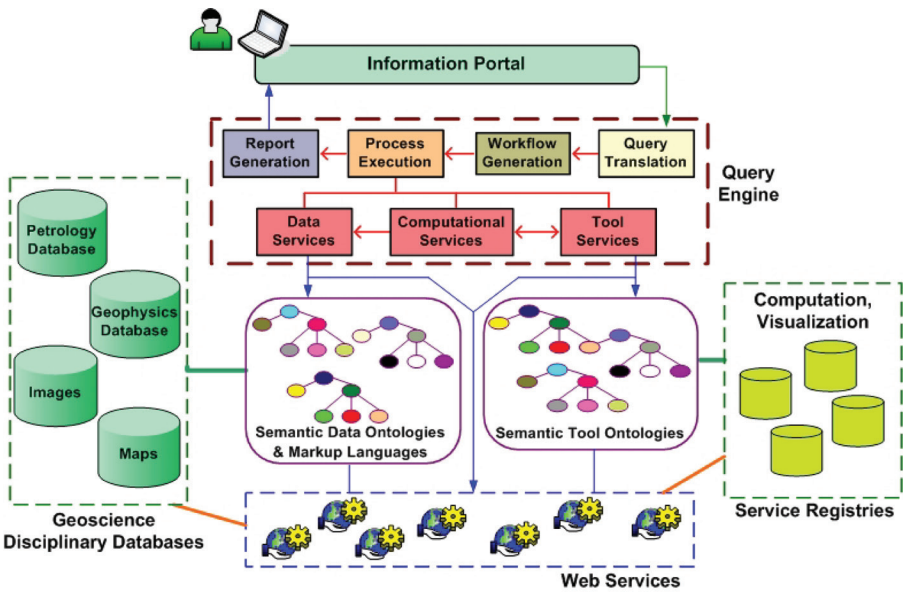

Figure 4. Schematic representation of geoinformatics components linked for efficient access to data, tools, and services for solving complex problems. A user query presented through a portal would automatically retrieve the appropriate tool defined as a Web service (registered to semantic tool ontology), which in turn would identify the required data set (registered to a data ontology), facilitating integration of heterogeneous and distributed data resources.

enable them to be automatically selected to answer geoscience queries. For example, the problem of integrating heterogeneous volcanic and atmospheric chemical-compound data used to assess the atmospheric effects of volcanic eruptions can be accomplished through semantically enabled registration and integration engines (Malik et al., 2007b; Rezgui et al., 2007; Fox et al., 2008). A simple query, such as "Find A-type plutons in Virginia and identify the correlation between these plutons and their gravity properties," requires Web-based access to distributed data resources (geochemical, gravity, and map databases, as well as computational and visualization tools) (Rezgui et al., 2007). Clearly, continued participation by geoscientists in ontology development and engineering and registration of data and tools will enable the community to move ahead into the emerging world of the Semantic Web.

THE FUTURE: THE SEMANTIC WEB AND DATA WITH NO BORDERS

The emerging Semantic Web is an extension of the existing Web, in which all information is given a well-defined meaning (Berners-Lee et al., 2001). The ultimate goal of the Semantic Web is to transform the present-day Web into a medium through which data and applications can be automatically understood and processed without geographical or organizational boundaries. The Semantic Web allows understanding, sharing, and invocation of data and services by automated tools associated with ontologies (Alonso-Jiménez et al., 2006), and it is already in use within the corporate world (Oracle, 2010; W3C, 2009a, 2009b). Other advantages of Semantic Web technologies for the geosciences include (1) facilitated knowledge management (capturing, extracting, processing, and storing knowledge) (Alonso et al., 2003); (2) integration across heterogeneous domains through ontologies (Fox et al., 2008); (3) the ability to handle non-text items, such as images and multimedia (Schreiber et al., 2001); (4) efficient information filtering (sending selective data to the right clients); (5) machine understanding (the ability to take humans out of the "integration loop"); (6) the formation of virtual communities (Reitsma and Albrecht, 2005); (7) legacy capture for long-term archiving; (8) serendipity (finding unexpected collaborators); and (9) Web-based education (Ramamurthy, 2006).

Capabilities based on semantic integration of data, services, and processes will become the new paradigm in scientific endeavors and will provide a significant boost to the visibility of geoscience research and education in a competitive world. Significant industry and government funding will be necessary for geoinformatics to grow to the level enjoyed by its sister program in bioinformatics (e.g., Mohan-Ram, 2000; Tracor Systems Technologies, 1998). We also support the establishment of a consortium to provide an organizational platform for promoting long-term management of data and resources. Researchers in bioinformatics have already recognized the need to establish economically viable models for the long-term survival of public data on the Web (Ellis and Kalumbi, 1998); geoscientists can utilize the voice of the consortium to provide stability for existing data, because those data represent the fundamental infrastructure for future geoscience research and its applications.

SUMMARY

Earth has a complex record of the dynamic interaction among plates, materials, and life that provides clues to the physical and chemical evolution of continents, oceans, atmosphere, and life forms. Extremely heterogeneous data from rocks that preserve ~4.5 billion years of history have been meticulously gathered through observations over the centuries, and this highlights the integration problems associated with studies of biodiversity, climate change, planetary processes, and natural hazards and resources. The vision of geoinformatics is to create a fully integrated geosciences information network with free access to earth-science data, tools, and services. Research in all categories of geoinformatics will support the emerging challenges posed by the building of knowledge societies:

> First, to narrow the digital divide that accentuates disparities in development, excluding entire groups and countries from the benefits of information and knowledge; second, to guarantee the free flow of, and equitable access to, data, information, best practices and knowledge in the information society; and third to build international consensus on newly required norms and principles.
>
> (UNESCO, 2003, preface)

ACKNOWLEDGMENTS

This paper was written on behalf of the GSA Geoinformatics Division. We acknowledge the support of the National Science Foundation's Division of Earth Sciences, the U.S. Geological Survey, the U.S. National Aeronautical and Space Administration, the British Geological Survey, the Geological Society of America, and the American Geophysical Union.

REFERENCES CITED

Alonso, G., Casati, F., Kuno, H., and Machiraju, V., 2003, Web services: Concepts, architecture, and applications: Berlin, Springer Verlag, 354 p.

Alonso-Jiménez, J.A., Borrego-Díaz, J., Chávez-González, A.M., and Martín-Mateos, F.J., 2006, Foundational challenges in automated semantic Web data and ontology cleaning: IEEE Intelligent Systems, v. 21, no. 1, p. 42–52.

Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerchmitt, D., Messina, P., Ostriker, J., and Wright, M., 2003, Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Advisory Panel on Cyberinfrastructure, 84 p.: http://www.nsf.gov/od/oci/reports/atkins.pdf (7 Aug. 2010).

Baader, F., Horrocks, I., and Sattler, U., 2004, Description logics, *in* Staab, S., and Studer, R., eds., Handbook on Ontologies: New York, Springer Verlag, p. 3–28.

Berners-Lee, T., Hendler, J., and Lassila, O., 2001, The semantic Web: Scientific American, v. 284, p. 34–43.

BioBasics, 2007, BioPortal, Glossary, Standards: Government of Canada: http://www.biobasics.gc.ca/english/View.asp?mid=427&x=696 (27 Aug. 2010).

Boisvert, E., Johnson, B., Schweitzer, P., and Anctil, M., 2003, XML Encoding of the North American Data Model: U.S. Geological Survey Open-File Report 03-471: http://pubs.usgs.gov/of/2003/of03-471/boisvert/index.html (7 Aug. 2010).

Cardoso, J., and Sheth, A., 2006, The semantic Web and its applications, *in* Cardoso, J., and Sheth, A., eds., Semantic Web Services, Processes and Applications: New York, Springer, v. 3, p. 3–33.

Commission for the Management and Application of Geoscience Information, 2008, Why do we need GeoSciML?": CGI-IUGS, http://

www.cgi-iugs.org/tech_collaboration/docs/Why_do_we_need_GeoSciML_v1.doc (7 Aug. 2010).
Doan, A., and Halevy, A., 2005, Semantic integration research in the database community: A brief survey: American Association for Artificial Intelligence Magazine, v. 26, p. 83–94.
Ellis, L.B.M., and Kalumbi, D., 1998, The demise of public data on the Web?: Nature Biotechnology, v. 16, p. 1323–1324.
Fensel, D., 2004, Ontologies: A silver bullet for knowledge management and electronic commerce: New York, Springer Verlag, 162 p.
Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J., 1992, Knowledge discovery in databases: An Overview: AI Magazine, v. 13, p. 57–70.
Fox, P., Sinha, A.K., McGuinness, D., Raskin, R.G., and Rezgui, A., 2008, A volcano erupts: Semantic data registration and integration: U.S. Geological Survey Scientific Investigations Report 2008-5172, p. 72–75.
Glasgow Caledonian University, 2008, Learning Services Support, Useful Definitions: http://www.learningservices.gcal.ac.uk/it/staff/definitions.html (27 Aug. 2010).
Gruber, T.R., 1993, A translation approach to portable ontologies: Knowledge Acquisition, v. 5, p. 199–220, http://tomgruber.org/writing/ontolingua-kaj-1993.pdf (27 Aug. 2010).
Lin, K., and Ludäscher, B., 2003, A system for semantic integration of geologic maps via ontologies, *in* Proceedings, Semantic Web Technologies for Searching and Retrieving Scientific Data (SCISW), Sanibel Island, Florida.
Malik, Z., Rezgui, A., and Sinha, A.K., 2007a, Ontologic Integration of Geoscience Data on the Semantic Web: U.S. Geological Survey Scientific Investigations Report 2007-5199, p. 41–43.
Malik, Z., Rezgui, A., Sinha, A.K., Lin, K., and Bouguettaya, A., 2007b, DIA: A Web services-based infrastructure for semantic integration in geoinformatics, *in* Proceedings, IEEE International Conference on Web Services, Salt Lake City, Utah, p. 1016–1023.
Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L., 2003, WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies, Preliminary Report: LADSEB-CNR, Padova, Italy: http://wonderweb.semanticweb.org/deliverables/documents/D17.pdf (7 Aug. 2010).
McGuinness, D.L., 2003, Ontologies come of age, *in* Fensel, D., Hendler, J., Lieberman, H., and Wahlster, W., eds., Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential: Cambridge, Mass., MIT Press, p. 171–196.
McGuinness, D.L., and van Harmelen, F., 2004, OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004: http://www.w3.org/TR/owl-features/ (7 Aug. 2010).
Mohan-Ram, V., 2000, Federal Funds and Bioinformatics Grants: A Match Made in Heaven?: Science Career Magazine, http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2000_09_01/noDOI.4689950319642425983 (7 Aug. 2010).
Niles, I., and Pease, A., 2001, Towards a standard upper ontology, *in* Welty, C., and Smith, B., eds., Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, October 17–19, p. 2–9.
Noy, N.F., 2004, Semantic integration: A survey of ontology-based approaches: SIGMOD Record, v. 33, p. 65–70.
Noy, N.F., and McGuinness, D.L., 2001, Ontology development 101: A guide to creating your first ontology: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html (7 Aug. 2010).
Object Management Group, 2010, Unified Modeling Language, UML® Resource Page: Object Management Group, Inc., http://www.omg.org/uml (27 Aug. 2010).
Obrst, L., 2003, Ontologies for semantically interoperable systems, *in* Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, November 03–08, p. 366–369.
OneGeology, 2010, Making Geological Map Data for the Earth Accessible: http://www.onegeology.org/home.html (27 Aug. 2010).
Oracle, 2010, Oracle Database Semantic Technologies: http://www.oracle.com/technetwork/database/options/semantic-tech/index.html (27 Aug. 2010).
Ramachandran, R., Graves, S.J., and Raskin, R., 2006, Ontology re-engineering use case: Extending SWEET to map climate and forecasting vocabulary terms, *in* Proceedings, American Geophysical Union Spring Meeting, v. 87, 7 p.
Ramamurthy, M.K., 2006, A new generation of cyberinfrastructure and data services for earth system science education and research: Advances in Geosciences, v. 8, p. 69–78.
Raskin, R.G., and Pan, M.J., 2005, Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET): Computers and Geosciences, v. 31, p. 1119–1125.
Raskin, R.G., 2006, Development of ontologies for earth system science, *in* Sinha, A.K., ed., Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, p. 195–200.
Reitsma, F., and Albrecht, J., 2005, Modeling with the Semantic Web in the geosciences: IEEE Intelligent Systems, v. 20, p. 86–88.
Rezgui, A., Malik, Z., and Sinha, A.K., 2007, DIA engine: Semantic discovery, integration, and analysis of earth science data: U.S. Geological Survey Scientific Investigations Report 2007-5199, p. 15–18.
Schreiber, A.T., Dubbeldam, B., Wielemaker, J., and Wielinga, B., 2001, Ontology-based photo annotation: IEEE Intelligent Systems, v. 16, no. 3, p. 66–74.
Sedris Technologies, 2007, Glossary, Interoperability: http://www.sedris.org/glossary.htm#I-grp (27 Aug. 2010).
Semy, S., Pulvermacher, M., and Obrst, L., 2004, Toward the use of an upper ontology for U.S. government and U.S. military domains: An evaluation: The MITRE Corporation (04-0603), http://handle.dtic.mil/100.2/ADA459575 (7 Aug. 2010).
Sheth, A., 1998, Changing focus on interoperability in information systems: From system, syntax, structure to semantics, *in* Goodchild, M., Egenhofer, M., Fegeas, R., and Kottman, C., eds., Interoperating Geographic Information Systems: Netherlands, Kluwer, p. 5–30.
Simmhan, Y.L., Plale, B., and Gannon, D., 2005, Survey of data provenance in e-science: ACM Sigmod Record, v. 34, no. 3, p. 31–36.
Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T., Johnson, B., Laxton, J., and Richard, S., 2006, GeoSciML: Enabling the Exchange of Geological Map Data, *in* Proceedings, Australian Earth Sciences Convention (AESC) Melbourne, 4 p.
Sinha, A.K., Zendel, A., Brodaric, B., Barnes, C., and Najdi, J., 2006, Schema to ontology for igneous rocks, *in* Sinha, A.K., ed., Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, p. 169–182.
Sinha, A.K., McGuinness, D., Fox, P., Raskin, R., Condie, K., Stern, R., Hanan, B., and Seber, D., 2007, Towards a Reference Plate Tectonics and Volcano Ontology for Semantic Scientific Data Integration: U.S. Geological Survey Scientific Investigations Report 2007-5199, p. 43–46.
Steven, R., Wroe, C., Lord, P., and Goble, C., 2004, Ontologies and bioinformatics, *in* Staab, S., and Studer, R., eds., Handbook on Ontologies: New York, Springer Verlag, 657 p.
Takarada, S., Kawabata, D., Kouda, R., Miyazaki, J-C., Fusejima, Y., and Asaue, H., 2007, Integrated geological map database (GeomapDB) in Geological Survey of Japan, AIST: U.S. Geological Survey Scientific Investigations Report 2007-5199, p. 5–7.
Tracor Systems Technologies, Inc., 1999, Bioinformatics in the 21st century: http://clinton4.nara.gov/WH/EOP/OSTP/NSTC/html/bioinformaticsreport.html (7 Aug. 2010).
TopicMaps.Org Authoring Group, 2001, XML Topic Maps (XTM) 1.0: TopicMaps.Org, http://www.topicmaps.org/xtm/index.html (27 Aug. 2010).
Tripathi, A., and Babaie, H.A., 2008, Developing modular hydrogeology ontology by extending the SWEET upper-level ontologies: Computers and Geosciences, v. 34, no. 9, p. 1022–1033.

UKOLN, 2006, Interoperability Focus: About: University of Bath, http://www.ukoln.ac.uk/interop-focus/about/ (27 Aug. 2010).
UNESCO, 2003, Science in the Information Society: United Nations Educational, Scientific and Cultural Organization Report CI2003/WS/6, 55 p.
W3C, 2003, Extensible Markup Language (XML): http://www.w3.org/xml (27 Aug. 2010).
W3C, 2004a, Resource Description Framework (RDF): http://www.w3.org/rdf (27 Aug. 2010).
W3C, 2004b, Web Ontology Language (OWL): http://www.w3.org/2004/OWL (27 Aug. 2010).
W3C, 2006, Web Services Architecture: http://www.w3.org/2002/ws/ (7 Aug. 2010).
W3C, 2009a, Semantic Web, W3C Celebrates Semantic Web Progress at SemTech 2009: http://www.w3.org/2009/06/SemTech-pressrelease.html.en (27 Aug. 2010).
W3C, 2009b, Semantic Web Case Studies and Use Cases: http://www.w3.org/2001/sw/sweo/public/UseCases (27 Aug. 2010).
W3C, 2009c, W3C Semantic Web Frequently Asked Questions: http://www.w3.org/2001/sw/SW-FAQ#What1 (27 Aug. 2010).

Manuscript received 3 Nov. 2009; accepted 13 Apr. 2010. ❖

APPENDIX A

Glossary of Selected Terms

Conceptual model—uses a comprehensive idea that brings diverse elements into a basic relationship.

Data—values derived from scientific experiments and factual information, especially information organized for analysis.

Database—a structured collection of data managed to meet the needs of a community of users. The structure is achieved by organizing the data according to a database model.

Data model—an abstract model that describes how data are represented and used.

Description logics—a family of knowledge-representation languages that can be used to represent the terminological knowledge of an application domain in a structured and formally well-understood way.

Foundation technologies—technological resources for creation, communication, storage, and interpretation of data (e.g., spreadsheets, databases, word processors, bandwidth, HPC, Internet, etc.).

Interoperability—"enables distributed heterogeneous systems to be interactive so that a meaningful exercise may be conducted" (Sedris Technologies, 2007); the ability to exchange and use information across heterogeneous data.

Integration—the process of combining data residing at different sources and providing the user with a unified view of such resources.

Integration through layering—overlay of data products as is commonly utilized in GIS methods.

Integration through semantics—a set of technologies, drawn from artificial intelligence, linguistics, and knowledge management, designed to help make sense of complex information and allow improved integration between systems.

Markup language—"a notation for identifying the components of a document to enable each component to be appropriately formatted, displayed, or used" (Glasgow Caledonian University, 2008). A markup language (e.g., XML) provides a way to combine text and extra information about that text.

Ontology—a set of knowledge terms, including the vocabulary, the semantic interconnections, and explicit rules of inference and logic for some particular topic (Gruber, 1993).

OWL—Web Ontology Language is a family of knowledge representation languages for authoring ontologies endorsed by the W3C (2004b).

Portal—Web site considered to be an entry point for discovery and access of multiple resources and other Web sites.

Provenance—tracking the source of data and services.

Registration—adding new descriptions to a repository.

Relational model for database—based on first-order predicate logic.

Schema—structure and organization of databases, including information on the type of content and relationship within the structure (also XML and RDF schemas).

Service registry—a network-accessible directory that contains information about the available services.

Standards—defined by the International Organization of Standardization (ISO) as "documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose" (BioBasics, 2007).

Semantic—the implied meaning of data. Used to define what entities mean with respect to their roles in a system (Sedris Technologies, 2007).

Semantic interoperability—refers specifically to the meanings that are embedded in this exchanged information and to the effective and consistent interpretation of these meanings.

Semantic Web—an evolving extension of the World Wide Web in which Web content can be expressed not only in natural language but also in a form that can be understood, interpreted, and used by software agents, thus permitting them to find, share, and integrate information more easily (W3C, 2009c).

Structural interoperability—incompatibilities between hardware, operating systems, etc.

Syntactic interoperability—form of interoperability concerned with the technical issues and standards involved in the effective communication, transport, storage, and representation of metadata and other types of information (UKOLN, 2006).

Taxonomy—classification scheme for terms, structured collection of terms, generally hierarchical, that is used for both classification and navigation.

UML—Unified Modeling Language is the industry-standard language for the specification, visualization, construction, and documentation of the components of software systems. UML helps to simplify the process of software design, making a model for construction with a number of different views (Object Management Group, 2010).

Web service—defined by a set of technologies that provide platform-independent protocols and standards used for exchanging data between applications. Web services are frequently just Web application programming interfaces (APIs) that can be accessed over a network, such as the Internet, and executed on a remote system hosting the requested services.

XTM—provides a model and grammar for representing the structure of information resources used to define topics, and the associations (relationships) between topics (TopicMaps.Org Authoring Group, 2001).