COVER FEATURE

Standards-Based Computing Capabilities for Distributed Geospatial Applications

Craig Lee, Open Grid Forum and The Aerospace Corporation
George Percivall, Open Geospatial Consortium

Researchers face increasingly large repositories of geospatial data stored in different locations and in various formats. To address this problem, the Open Geospatial Consortium and the Open Grid Forum are collaborating to develop standards for distributed geospatial computing.

It is said that 80 to 90 percent of all information is geospatially related. Whatever the exact percentage, geospatial information clearly has immense applicability across the spectrum of human endeavor. Examples include oil and gas exploration, weather forecasting and tracking, aviation, satellite ground systems, environmental planning, disaster management, public administration (e-government), civic planning and engineering, and all manner of e-sciences. All such activities entail gathering significant amounts of data and other critical information that must be stored, accessed, and managed.

With our planet's growing "infomass," researchers face increasingly large repositories, archives, and libraries of geospatial data that are inherently distributed across many institutions and countries. Even within organizations, such data collections can be stored in different locations and in various formats and schemas. In addition, a growing number of in situ sensors gather data continuously.

Many geospatial applications require not only data from multiple heterogeneous sources but also special processing resources available at remote sites, including high-performance computing. Grid computing addresses this need for data diversity, processing resource chains, and HPC, along with other requirements such as security and digital rights management, in a distributed environment. The "Distributed Geospatial Computing Initiatives" sidebar lists a few important projects that use grid-based geospatial applications.

NEED FOR STANDARDS

The ability to access, integrate, analyze, and present geospatial data across a distributed computing environment using common tools has tremendous value. Indeed, with the growing connectedness of our world—through data-collecting instruments, data centers, supercomputers, departmental machines, and personal devices such as cell phones, PDAs, and smart phones—as a society we expect a wide range of information to be instantly accessible from anywhere. This expectation is motivated not only by personal convenience; there are also solid business and policy reasons for enabling such distributed geospatial applications.

Achieving such ease of access and interoperation will require best practices that are codified into widely adopted standards. In the context of e-Science, the National Science Foundation's Cyberinfrastructure Council argues that "the use of standards creates economies of scale and scope for developing and deploying common resources, tools, software, and services that enhance the use of cyberinfrastructure in multiple science and engineering communities. This approach allows maximum interoperability and sharing of best practices. A standards-based approach will ensure that access to cyberinfrastructure will be inde-

Published by the IEEE Computer Society

Distributed Geospatial Computing Initiatives

The following grid-based initiatives illustrate the broad applicability of and growing demand for distributed geospatial computing resources.

CYCLOPS

The Cyberinfrastructure for Civil Protection Operative Procedures project (www.cyclops-project.eu) brings together Global Monitoring for Environment and Security (GMES), a major European space initiative, and grid computing with a focus on the needs of European civil protection. CYCLOPS will use the EU grid infrastructure built by the Enabling Grids for E-SciencE project.

GEO GRID

The Global Earth Observation Grid project (www.geogrid.org) at Japan's National Institute of Advanced Industrial Science and Technology also applies grid technology to global Earth observation. GEO Grid integrates satellite imagery, geological data, and terrestrial sensor data to monitor and respond to disasters such as earthquakes, landslides, and flooding.

GEOSS

The Global Earth Observation System of Systems (www.earthobservations.org/) is an activity of Group on Earth Observations, a partnership of 124 governments and international organizations. GEOSS aims to enable continuous monitoring of the Earth and provide access by researchers to a vast shared set of information resources. The Open Geospatial Consortium has led a series of GEOSS interoperability demonstrations at IEEE and International Society for Photogrammetry and Remote Sensing meetings. OGC also leads a GEOSS Architecture Implementation Pilot involving more than 120 organizations (www.ogcnetwork.net/aipilot).

SANY

The Sensors Anywhere integrated project (http://sany-ip.eu) focuses on interoperability of in situ sensors and sensor networks. The SANY architecture provides a quick and cost-efficient way to reuse data and services from currently incompatible sensor and data sources. The project contributes to GMES and GEOSS in the area of in situ sensor integration.

SEADATANET

SeaDataNet (www.seadatanet.org) is a pan-European project that provides a distributed infrastructure for managing ocean and marine data collected from in situ and remote sensing sources in 35 countries.

pendent of operating systems; ubiquitous, and open to large and small institutions."[1] In fact, these issues are not specific to e-Science but are applicable across many enterprises.

OGC-OGF COLLABORATION

The Open Geospatial Consortium (www.opengeospatial.org) and the Open Grid Forum (www.ogf.org) are collaborating to develop open standards that address the distributed computing needs of geospatial applications while accommodating the inevitability of diverse formats, schemas, and processing algorithms.

OGC is an international consortium that uses a consensus process to develop publicly available interface standards for geospatial information and services. OGF is a similar organization dedicated to the development of standards for the management of distributed computing resources such as servers, networks, and storage.

The integration of OGC and OGF standards will provide the necessary infrastructure for developing tools, software, and services that work together and can be used by multiple communities. They will support interoperability and the development and wide adoption of shared best practices, as well as help both small and large technology providers and users innovate and leverage one another's strengths and accomplishments. Layered on basic Internet and Web standards like HTTP, HTML, and XML, these efforts enable powerful new capabilities available through the huge installed base of Internet and Web clients and servers.

NOTIONAL GEOCOMPUTING ARCHITECTURE

Figure 1 shows a notional architecture for distributed geospatial applications. Such a cyberinfrastructure must process sensor data from its raw observational state to georeference its location and remove artifacts associated with the sensor network's acquisition specifics. It must also transform the data from sensor measurements to calibrated physical parameters of interest—that is, engineering units.

Researchers can use this data directly or store it in a repository. In either case, to enable proper tagging (identification), processing, and decision making, the cyberinfrastructure must provide user access to direct sensor observations or to geoscience repositories in an interoperable fashion that includes metadata about the observational data and the sensor itself. In many cases, the data's provenance is important.

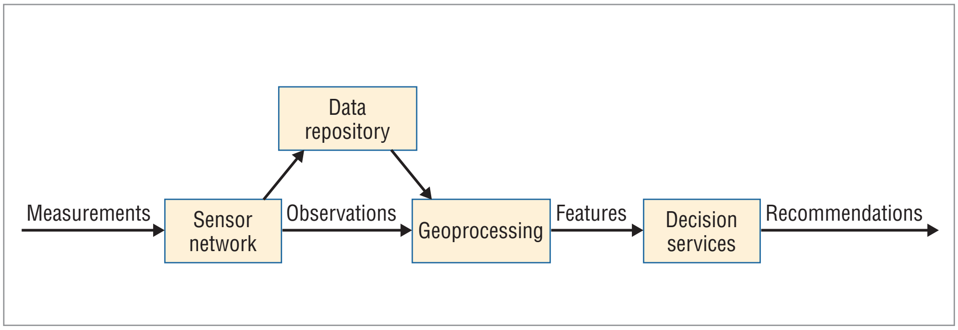

Figure 1. To support critical decision making, a geoscience cyberinfrastructure must provide end-to-end support including sensing, data storage, and data processing. Because sensor networks and data repositories are usually distributed, the infrastructure should support ease of access, data interoperability, and user authentication and authorization.

Because sensor networks and data repositories are usually distributed, the infrastructure should support ease of access, data interoperability, and user authentication and authorization as constrained by policy. Providing interoperability with other domains, including the mass market for geospatial data on the Internet, is another consideration.

Data processing requirements can be straightforward or complex and intensive. In either case, it might be necessary to discover and schedule computing resources that are suitable to the requirements on a grid or in a computing cloud.

Decision-support services typically consume the final data products. These services might simply present the data to human decision-makers in a format appropriate for human comprehension. They could also be rule-based or other automated systems that infer appropriate decisions. In either case, such decisions could drive subsequent actuator systems that "close the loop" by affecting the system under observation. Such systems are sometimes called *dynamic data-driven application systems* (www.dddas.org).

Certain basic capabilities are necessary for an open-standards-based infrastructure for geospatial applications. First, information models, encodings, and metadata are needed to enable integration of data collected from different sources into representations of scientific knowledge. Second, standards for geospatial catalog services and standards for describing sensors, geospatial Web services, and geospatial data are necessary to enable discovery of and access to extended networks of such resources. Finally, processing and workflow architectures and standards must accommodate large, distributed datasets and grid computing resources.

OGC WEB SERVICES

The notional geocomputing architecture is largely supported by the OpenGIS standards that OGC already has in place (www.opengeospatial.org/standards).

OWS standards

OGC Web Service standards are layered on top of open Internet standards—in particular, the HTTP, uniform resource locators (URLs), multipurpose Internet mail extensions (MIME), and XML World Wide Web standards. The main OWS standards include the following:

Web Map Service. WMS standardizes the display of registered and superimposed maplike views of information that come simultaneously from multiple remote and heterogeneous sources.

Web Feature Service. WFS standardizes the retrieval and update of digital representations of real-world entities referenced to the Earth's surface.

Web Coverage Service. WCS standardizes access to spatially extended coverages, usually encoded in a binary format and offered by a server.

Catalogue Service for the Web. CSW standardizes interfaces to publish, discover, browse, and query metadata about data, services, and other resources. Metadata standards from ISO/TC 211 (*ISO 19115:2003, Geographic information—Metadata* and *ISO/TS 19139:2007, Geographic information—Metadata—XML schema implementation*) also play a critical role in the Web-based publishing and discovery of geospatial resources.

SWE standards

OGC Sensor Web Enablement (SWE) standards make all types of sensors, instruments, and imaging devices accessible and, where applicable, controllable via the Web:

Sensor Observation Service. SOS standardizes the retrieval of observations and sensor system information. Observations can be from in situ sensors (for example, water monitoring) or dynamic sensors (for example, satellite imaging).

Sensor Planning Service. SPS standardizes an interface to task sensors or models. It can be used to reprogram or calibrate sensors, start or change sensor missions, and execute and control simulation models.

Sensor Alert Service. SAS defines an interface for publishing and subscribing to alerts from sensors. If an event occurs, the SAS will notify all clients that subscribe to this event type.

Observations and Measurements. O&M defines an abstract model and an XML schema (www.w3.org/XML/Schema) encoding for observations and provides support for common sampling strategies. O&M also

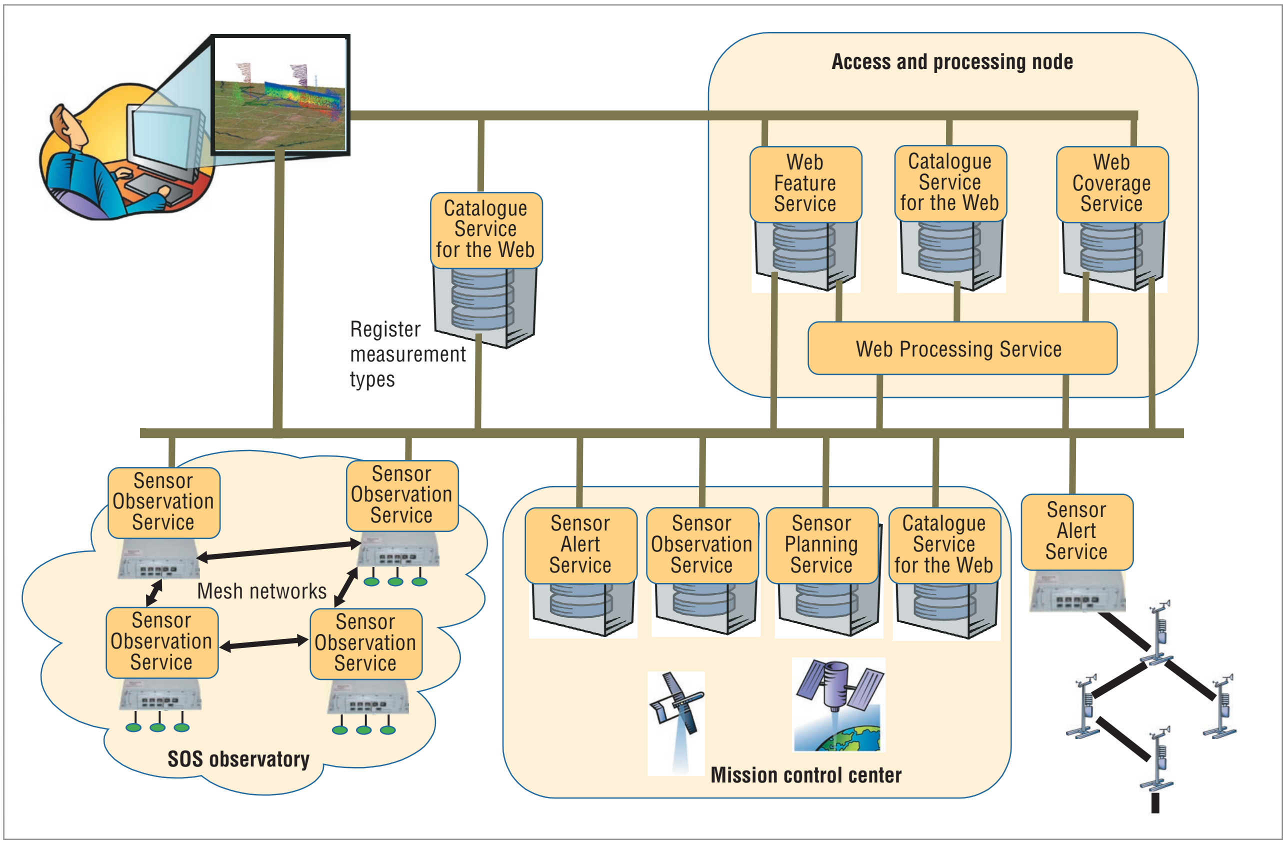

Figure 2. OGC Web services in support of geospatial applications. Multilevel integration enables access and sharing of resources over networks in e-Science communities.

provides a general framework for scientific and engineering systems that deal in technical measurements.

Sensor Model Language. SensorML specifies models and XML encoding that provide a framework within which the geometric, dynamic, and observational characteristics of sensors and sensor systems can be defined and encoded in XML.

e-Science communities

Figure 2 depicts OWS and SWE services in support of geospatial applications that process measurements from sensor webs and make them available to users. Multilevel integration enables access and sharing of resources over networks in *e-Science communities*. Different organizations can own and operate sensor systems and processing nodes for different purposes.

Sensor-observation value chains define the life cycle of sensor-produced observations, from raw unprocessed data to information products and services delivered to applications and consumers. Depending on requirements, geospatial applications can provide sensor data in its most elemental unprocessed form, as an observation object complete with the metadata and processes used to estimate a value describing a phenomenon, or, for example, into single-valued geographic feature or multivalued coverage representations.

Web Processing Service

While the OWS and SWE Web services provide access to remote geospatial data and sensor networks, they do not address general distributed computing, resource management, or the geoprocessing part of the notional architecture.

As a first step in this direction, OGC has developed the Web Processing Service (WPS) standard to define basic request-response interaction for remote execution of a service. This service can include any algorithm, calculation, or model that operates on spatially referenced data. WPS clearly could serve as an interface to a wide range of computing resources, from local servers to high-end computational grid resources. Providing and managing such access to distributed computing resources is the key goal of OGC's collaboration with OGF.

OGF DISTRIBUTED COMPUTING TOOLS

Since its emergence in the late 1990s, the grid community has focused on distributed resource management. The term "resource" refers to any type of hardware, software, service, or capability. Hence, in addition to typical hardware resources such as processors, storage, and networks, resource management also can involve bandwidth; data catalogs, archives, and repositories; data-producing instruments and sensor networks;

and computational services, including peer-to-peer networks.

Grid capabilities

While the grid concept grew out of the need to collaborate and share data and computing resources on large scientific projects, grid capabilities are fundamental to general distributed computing and directly support the notional geospatial architecture.

Information models and architectures. Basic resources in the distributed computing environment must be describable in a well-known information model. OGF is working with the Distributed Management Task Force (www.dmtf.org/home) to integrate elements of the Grid Laboratory Unified Environment into the Common Information Model, a widely used information model standard. Even with such models, every application domain user group must define its own information architecture that captures domain-specific knowledge.

> **Grid capabilities are fundamental to general distributed computing and directly support the notional geospatial architecture.**

Catalogs and discovery services. Once such information models and architectures are in place, the grid community can use concrete instances to populate catalogs of information for discovery by users and systems. Such catalogs must support discovery or "look-up" queries to find useful and relevant data, services, and so on. Depending on the size of the distributed environment to be supported, multiple catalogs might have to federate together to serve a large, distributed user community. An example of this is the Globus Monitoring and Discovery Service (www.globus.org/toolkit/mds).

Data, execution, and workflow services. There must be a way to access data, and for services there must be a way to schedule execution. Data access could mean simply transferring remote files using the GridFTP protocol (http://dev.globus.org/wiki/GridFTP), posing queries to remote databases, or even using a grid file system that provides a uniform file namespace across multiple institutions.

Similarly, service execution means being able to schedule the execution, monitor its progress until completion, and then retrieve any results. This can be accomplished using the HPC Basic Profile (www.ogf.org/documents/GFD.114.pdf) or the Distributed Resource Management Application API (www.drmaa.org). Workflow is the management of a sequence of operations or services that typically have dependencies and must be executed in a given order. The grid community has developed many workflow management systems.

Event, monitoring, and management services. Event notification is a key part of any distributed system and can be accomplished using OGF standards such as Information Dissemination (INFOD) that support the publish-subscribe paradigm. Monitoring an execution's progress is important to an individual user, but depending on the size of the distributed infrastructure to be maintained, it might be necessary to monitor which machines are up, which networks are up, and so on. Administrators must also be able to manage which machines and services are available to authorized users and their physical configuration.

Security. Users must be issued a grid identity that enables authentication and authorization when they request use of remote resources. This is typically done with X.509 certificates using the Grid Security Infrastructure (GSI) or Shibboleth.

A key development in grid security is the notion of virtual organizations—different users and institutions can participate in different VOs and have different roles and capabilities depending on their requirements. A good example of such a system is the Virtual Organization Membership Service (www.globus.org/grid_software/security/voms.php).

OGF standards

OGF has ongoing work across all of these areas, though in the context of the OGS WPS standard, which is a key point of collaboration between the organizations, the following OGF standards (www.ogf.org/gf/docs) are the most directly relevant:

HPC Basic Profile. This is a cluster-scheduling interoperability layer that coordinates the use of the Job Submission Description Language and the OGSA Basic Execution Service, along with interoperable security mechanisms, to address the primary high-performance use case of batch-job scheduling for scientific and technical applications. Demonstrated at Supercomputing 2006 and 2007, HPC Basic Profile is being developed commercially by Microsoft, Platform Computing, and Altair Computing.

Simple API for Grid Applications. SAGA is an object-oriented API for programming grid applications with a common "look and feel" across all operations when programming in a distributed environment. This includes support for job and task management, file management, replica management, name spaces, streams, monitoring, sessions, security contexts, and error handling. SAGA currently has C++ and Java bindings.

GridRPC. This standard generalizes the well-known remote procedure call model for grid environments. It provides a simple model and API whereby function handles can be used to reference and invoke remote services as though a procedure call was being made. Such calls can, however, be synchronous or asynchronous depending on application requirements. The GridRPC API is currently used by the NetSolve/GridSolve (http://

icl.cs.utk.edu/netsolve), Ninf-G (http://ninf.apgrid.org), and Distributed Interactive Engineering Toolbox (http://graal.ens-lyon.fr/~diet) systems, which provide a registry to catalog function handles for later discovery and use.

Data Access and Integration specifications. WS-DAI defines a general data service model and associated message patterns that can be applied to all manner of data resources in a grid environment. In WS-DAI, data resources are addressed using WS-Addressing end-point references that contain the resource's abstract name and uniquely identify the resource. XML-based property documents can be associated with a data resource and read/set through the data access service's interface.

Information Dissemination. INFOD provides a general means to determine which messages are to be sent from publishers to subscribers based on information kept in a registry. More specifically, it specifies interfaces that allow the characterization of publishers, subscribers, and various other components using vocabularies that are meaningful to members of the communities to which they belong.

> **OGC and OGF have identified numerous possible areas for initial integration efforts.**

INITIAL INTEGRATION EFFORTS

With these models and standards in mind, OGC and OGF have identified numerous possible areas for initial integration efforts. As indicated earlier, a key goal is implementing WPS with standard grid tools that enable access to a range of back-end processing environments. Beyond WPS, there is a need to develop joint interoperability specifications for workflow management, event notification, and security.

Because WPS employs a relatively simple model whereby a user can send data for processing on a (possibly remote) server and retrieve the results, implementations using HPC Basic Profile, SAGA, or GridRPC would all be possible. Depending on the security requirements and discovery support requirements, however, implementation strategies could also include generic Web services or even Web 2.0 mashups based on asynchronous JavaScript technology and XML.

Back-end hardware provisioning

In terms of back-end environments, obvious candidates include traditional grid infrastructures with large national and international backing such as the EU Enabling Grids for E-SciencE (EGEE; www.eu-egee.org), the Japanese National Research Grid Initiative (NAREGI; www.naregi.org), the UK National Grid Service (NGS; www.grid-support.ac.uk), and the US TeraGrid (www.teragrid.org).

Other possibilities include virtualization and cloud computing technologies, which established grids are beginning to use already. This development supports the viewpoint that the traditional grid concept is essentially a distributed framework that encompasses the notion of dynamic provisioning by cloud resources.[2]

Workflow management

Workflow management tools are typically used to manage sequences of calls to multiple remote services. Rather than having the output from service A "bounce" through the user client to be input to service B, workflow tools can coordinate the direct transfer of data from A to B with the scheduling and completion of these services.

Grid programming tools such as SAGA and GridRPC could manage basic workflow requirements. While commercially oriented workflow tools such as the Business Process Execution Language are available, the grid community has also developed numerous open source workflow tools for scientific and engineering applications, including Directed Acyclic Graph Manager (www.cs.wisc.edu/condor/dagman), Pegasus (http://pegasus.isi.edu), Kepler (http://kepler-project.org), Taverna (http://taverna.sourceforge.net), and Triana (www.trianacode.org).

Event notification

Many geospatial tools require event notification and information dissemination. Besides geoprocessing workflows, event notification is needed for decision support services and sensor web enablement. Hence, standards like INFOD that provide content-based publish-subscribe capability could find wide applicability.

Initiatives

To coordinate integration of standardized tools, OGC and OGF have organized a series of successful workshops at OGC and OGF meetings. OGF has also participated in the planning efforts for OGC's sixth OGC Web Services Testbed (OWS-6), which began 30 June 2008. OGC testbeds and other "interoperability initiatives" are fast-paced, practical, hands-on, scenario-focused projects designed to test different approaches and to deliver draft standards into the OGC's Technical Committee. Grid computing plays a central role in the Geo-Processing Workflow thread of OWS-6.

In addition, the broad applicability and demand for distributed geospatial computing capabilities that are driving this effort have spawned numerous practical integration initiatives that are well under way and represent potential participants. The "OGC-OGF Integration Initiatives" sidebar describes several of these efforts.

OGC-OGF Integration Initiatives

The following initiatives are examples of OGC-OGF integration efforts driven by the demand for distributed geospatial computing capabilities.

GEOSPATIAL KNOWLEDGE GRID

The George Mason University Center for Spatial Information Science and System (CSISS) Geospatial Knowledge Grid (http://grid.laits.gmu.edu) is a Web portal that integrates OGC and grid technologies for Earth science modeling and applications. The project is sponsored by NASA's Advanced Information Systems Technology program and conducted by CSISS in conjunction with the Ames Research Center and the Department of Energy's Lawrence Livermore National Laboratory. The Geospatial Knowledge Grid is part of the Committee on Earth Observation Satellites (www.ceos.org) testbed for evaluating the usability of grid technology in geospatial disciplines.

52° NORTH WPS-G

WPS-G is a powerful open source enhancement of 52° North's Web processing service (http://52north.org) that is publicly available. Existing implementations of 52° North WPS processes can be reused and distributed over the grid. The service supports Unicore 6 middleware (www.unicore.eu/download/unicore6), and could easily support other grid middleware packages such as Globus. WPS-G requires no administrative access, only a valid certificate. 52° North plans to "gridify" other mainstream geographic information system tools such as ESRI's ArcGIS Server (www.esri.com/software/arcgis/arcgisserver/index.html) and the Geographic Resources Analysis Support System (http://grass.osgeo.org) via WPS.

SAW-GEO AND SEE-GEO

The UK Joint Information Systems Committee is responsible for funding research into the provision of cutting-edge information services for the UK academic community. In 2006, JISC commissioned the Grid OGC Collision Program to investigate integration of grid computing and Web services based on OGC standards (http://edina.ac.uk/projects/seesaw). The program involved two themes, workflows and security, which led respectively to the Semantically Aware Workflow Engines for Geospatial Web Service Orchestration and Secure Access to Geospatial Services projects. SAW-GEO involves chaining multiple Web services together in a semantically informed workflow management system and workflow engine, while SEE-GEO involves demonstrating how to securely access OGC Web services using GSI, Shibboleth, and WS-Security.

SENSORNET

OGC is involved with Oak Ridge National Laboratory's SensorNet project (www.sensornet.gov), a comprehensive incident management system for all manner of geospatially related threats—chemical, biological, radiological, nuclear, and explosive.

CHALLENGES AND FUTURE WORK

The OGC-OGF collaboration is off to a good start, but many unresolved issues are critically important in the context of the rapidly emerging global need to migrate data from technically or physically isolated repositories with poor or nonstandard metadata to more open environments.

One key challenge is hiding grid complexity and simplifying tools for nonspecialists in geospatial application domains. Progress in ease of use might depend on a large community adopting single approaches, such as the WPS standard, that provide a standard way to exploit grid technology. However, this requires gaining more experience in using WPS with a range of back-end processing environments and various workflow management tools.

Handling massive data is another major issue, and this requires involving organizations such as EGEE, NAREGI, NGS, and TeraGrid that are capable of supporting vast computing resources as well as interoperability requirements. Similarly, the need for real-time data processing will accelerate as more experimenters get access to large-scale distributed computing environments.

Workshops on digital repositories, with participation by geospatial projects such as Dissemination and Exploitation of Grids in Earth Science (www.eu-degree.eu) and Ground European Network for Earth Science Interoperations—Digital Repositories (www.genesi-dr.eu) are planned for future OGF and OGC meetings, with a focus on topics such as security, workflow management, and programming models. Other organizations are also interested in these issues. For example, some of the European Geophysical Union's recent Earth and Space Science Informatics sessions explicitly covered grids, and more such sessions are in the works.

OGC and OGF groups have a shared interest in discussing overlapping or complementary work in digital rights management, information communications, semantics, and workflow management. This work will contribute to the development of best practices for integrating registries with grid data movement tools and security models. Data consumed and produced by WPS calls could come from and be returned to OGC federated catalogues in a simple client-server fashion, but secure, third-party transfers might also be desirable.

Metcalfe's law states that a telecommunication network's value increases in proportion to the square of the number of nodes on the network (http://en.wikipedia.org/wiki/Metcalfe's_law). A corollary to this law is that the value of online geospatial data increases according to its accessibility and interoperability.

The world's major funding agencies, foundations, and corporations are coming to realize the tremendous value of maintaining geospatial data online for verification, ongoing longitudinal studies, and cross-disciplinary studies of all sorts. This will certainly require the widespread adoption of robust metadata, distributed computing, and system management standards. Ultimately, grid-based geoprocessing will become an integral and routine part of a resource-rich Web environment. ■

References

1. National Science Foundation Cyberinfrastructure Council, "Cyberinfrastructure Vision for 21st Century Discovery," NSF, Mar. 2007; www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.
2. D. Gannon, "The Computational Data Center—A Science Cloud," Indiana Univ.; www.extreme.indiana.edu/~gannon/science-data-center.pdf.

Craig Lee *is president of the Open Grid Forum as well as a senior scientist at The Aerospace Corporation. His research interests include clusters, grids, clouds, and all aspects of parallel and distributed computing. Lee received a PhD in computer science from the University of California, Irvine. He is a member of the IEEE Computer Society and the ACM. Contact him at lee@aero.org.*

George Percivall *is the chief architect and an executive director of the Open Geospatial Consortium. His research interests include systems architecture, information science, geoinformatics, and semiotics. Percivall received an MS in electrical engineering from the University of Illinois at Urbana-Champaign. He is a senior member of the IEEE. Contact him at gpercivall@opengeospatial.org.*