

Semantic Services for Grid-Based, Large-Scale Science

William E. Johnston, Lawrence Berkeley National Laboratory
and NASA Ames Research Center

Grid technology^{1,2} has evolved over the past several years and is merging with Web Services to provide mechanisms and infrastructure for a standardized, componentized approach to building distributed, virtual systems and organizations for large-scale science. This software technology will knit hardware, data, and resources into an

infrastructure that substantially simplifies building science applications. It will simplify supporting collaborations involving large-scale computing systems, data archives, and instruments that span many different projects, institutions, and countries. This infrastructure will comprise grid-based services that integrate with the user's work environment and enable uniform, highly capable access to these widely distributed resources. These services will integrate transient-use resources (such as computing systems and scientific instruments, which are used as needed to perform a simulation or analyze data from an experiment), persistent-use resources (such as databases, data catalogs, and archives), and collaborators who'll be involved for a project's lifetime or longer.

However, as grid practitioners begin to understand the grid services environment and how the science community will use it and deploy its capabilities, we're seeing that an infrastructure for the next generation of the science process will need further capabilities.

Beyond the grid services and application-oriented services in this environment, other basic functionality is necessary. This includes, for example, virtual data services³ and application composition frameworks, such as the Common Component Architecture XCAT,⁴ that manage several styles of connectivity between componentized services. Web-based portal builders such as Xportlet⁵ that provide componentized tools for building GUIs in the Web and grid environment are also necessary.

However, even with this added functionality, the problem of making these tools usable in the science environment remains. We must provide mechanisms

to structure and manipulate various representations of the available application services, tool services, and data. One promising approach is to build on the tools being developed in the Semantic Web community, which applies the AI community's discipline-oriented descriptions of the semantic aspects of services and data to XML-based descriptions of applications components (Web Services) and data.

We envision Semantic Web-like tools that automatically check the validity of sequences of composed operations and data, and automatically construct intermediate steps in a loosely specified sequence. These tools should also automatically construct sequences of operations that are consistent with a discipline model representing permitted relationships among simulation and analysis operations and data for particular disciplines, such as climatology or high-energy physics. I call these tools *semantic services*.

Grid technology

The motivation for current large-scale, multi-institutional grid projects is to enable resource and human interactions that facilitate large-scale science and engineering (such as aerospace systems design, high-energy physics data analysis, climatology, large-scale remote instrument operation, and collaborative astrophysics based on virtual observatories). In this context, grids aim to provide significant new capabilities to scientists and engineers by facilitating routine construction of information- and collaboration-based problem-solving environments that are built on demand from large resource pools.

The process of large-scale science must evolve to facilitate the next steps of scientific discovery. Grid technology and semantic tools will be valuable in dealing with the complex multidisciplinary simulation and data environments that next-generation science will require.

Functionally, grids provide consistent tools, middleware, and services for

- Building the application frameworks that let scientists express and manage the simulation, analysis, and data-management aspects of overall problem solving
- Providing a uniform look and feel to a variety of distributed-computing and data resources
- Supporting construction, management, and use of widely distributed application systems
- Facilitating human collaboration through common security services and resource and data sharing
- Providing remote access to and operation of scientific and engineering instrumentation systems
- Managing and securing this computing and data infrastructure as a persistent service

Grids accomplish these tasks through a set of uniform software services that manage and provide access to heterogeneous, distributed resources and a widely deployed infrastructure. Figure 1 depicts grids' layered architecture.

The Global Grid Forum (www.gridforum.org), which consists of some 700 people from some 130 academic, scientific, and commercial organizations in about 30 countries, is working on defining and standardizing grid middleware. The GGF involves both scientific and commercial computing interests.

Science case studies

The US Department of Energy's Office of Science (www.er.doe.gov) recently undertook to characterize how the process of doing large-scale science must change to support scientific advances. In four workshops from 2002 to 2003, the Office of Science analyzed issues and set out networking and middleware requirements, proposed approaches to meet requirements, and examined the computing requirements and approach.⁶⁻⁹ This section presents the requirements analyses in the case studies from two of these workshops as they relate to semantic services.

Climate-modeling requirements

To better understand climate change, we need climate models that have higher resolution and incorporate more of the real world's physical complexity. Over the next five years, climate models will become increasingly

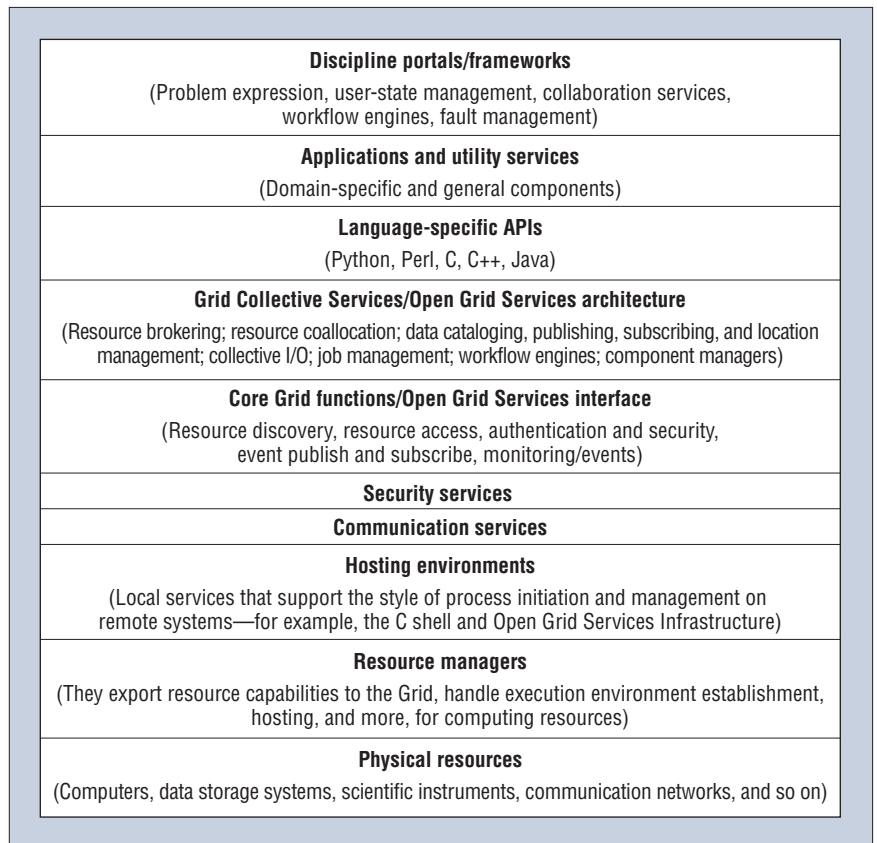


Figure 1. Grids' layered architecture.

complex through work such as the North American Carbon Project, which endeavors to fully simulate the terrestrial carbon cycle.

The need to forecast local and regional climate, as well as extreme climate changes (droughts, floods, severe storms, and other phenomena), drives these advances. Over the next five years, climate models will also incorporate the available and growing amounts of observational data, for both hindcasting and intercomparison purposes. So, instead of tens of terabytes of data per model instantiation, grid technology will enable storage of hundreds of terabytes to a few petabytes of data at multiple computing sites for climate scientists worldwide to analyze. We must fully utilize middleware systems, such as the Earth System Grid (www.earthsystemgrid.org) and its descendants, to access and manage such large, distributed, and complex pools of observational and simulation data.

In the period five to 10 years out, climate models will again improve in resolution and integrate more components. They'll be used for regional-scale modeling, which requires resolutions that range from tens to hundreds of meters instead of the hundreds-of-

kilometers resolution of the Community Climate System Model and Parallel Climate Model.

To improve climate modeling, the many institutions working on different aspects of climate must be able to easily describe, catalog, and seamlessly share their knowledge and supporting data; this facilitates the required interdisciplinary collaboration. Furthermore, all submodels must interoperate in ways that represent how the different climate elements interact.

As climate modeling becomes more multidisciplinary, scientists from oceanography, atmospheric sciences, and other fields will collaborate to develop and examine more realistic climate models. Biologists, hydrologists, economists, and others will help create additional climate model components that represent important but still poorly understood influences on climate (see Figure 2). These models will have a true carbon cycle component, with models of biological processes to—for example—simulate marine biochemistry and fully dynamic vegetation. These scenarios will include human population change, growth, and econometric mod-

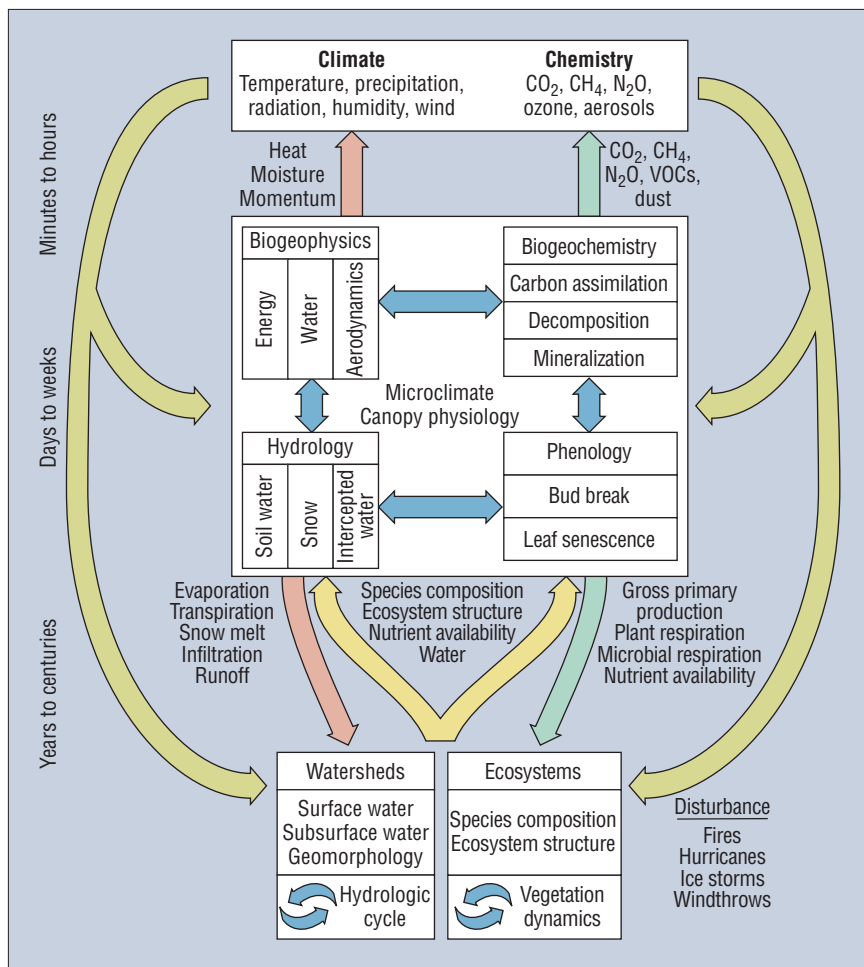


Figure 2. A “complete” approach to climate modeling involves the many interacting processes and data of terrestrial biogeoscience. (This figure is based on an illustration from *Ecological Climatology: Concepts and Applications* by Gordon Bonan [Cambridge Univ. Press, 2002].)

els that simulate potential changes in natural-resource use and efficiency. Additionally, climate models will integrate models representing solar processes to better simulate incoming solar radiation.

Specialized scientific groups working on a comprehensive, multidisciplinary model’s components build many specialized software and data environments that almost certainly can’t be combined on a single computing system. Almost all such multidisciplinary simulation is inherently distributed, with the overall simulation combining software and data from many systems into one virtual system using tools and facilities for building distributed systems.

The following capabilities, which result from combining computing, communication, and data-storage systems with grid services, will enable this sort of science process in the future:

- The computing capacity required for a task is available when the science needs that task. In particular, you must be able to incorporate supercomputers into virtual systems so that simulations whose components run on supercomputers can integrate with the science community’s many computing systems.
- The data capacity sufficient for the science task is available independent of location and is managed by the information systems that build, maintain, and allocate sharing of knowledge bases.
- The fundamentally distributed science community has remote access to computing, data, and distributed collaboration tools.
- Communication capacity and capability are sufficient to support the aforementioned in a way that’s transparent to both systems and users.

- Virtual data catalogs and work planners automatically reconstitute derived data on demand.
- Software services provide a rich environment that lets scientists build multidisciplinary simulations in ways that are natural to the scientific process; they don’t need to focus on the details of managing the underlying computing, data, and communication resources.

The climate community’s future science process also requires the informed interoperation of diverse submodels and integration of knowledge from many disciplines so that a realistic overall climate model can make valuable predictions for society. Constructing and managing multidisciplinary models will require tools that can use subdiscipline knowledge to help structure the multicomponent processing needed for comprehensive simulations. That is, these tools must not only build and manipulate complex domain models but also guide the interactions of different domain models. I’ll discuss this sort of semantic service, which addresses building and managing models whose components are also complex models, in more detail later in the article.

Climate’s complexity is typical of most macroscale phenomena—from cosmology to cellular function. So the issues that arise when looking at how climate modeling’s science process must evolve are characteristic of much of science.

High-energy physics requirements

The major high-energy and nuclear physics (HENP) experiments of the next 20 years will break new ground in understanding the fundamental interactions, structures, and symmetries that govern the nature of matter and space-time. The largest collaborations today—such as the CMS (Compact Muon Solenoidal detector) and ATLAS (A Toroidal Large Hadron Collider Apparatus) collaborations—are building experiments for CERN’s (European Organisation for Nuclear Research) Large Hadron Collider and encompass 2,000 physicists from 150 institutions in more than 30 countries.

HENP problems are among the most data-intensive known. The current generation of operational high-energy physics experiments at Stanford Linear Accelerator Center and Fermi National Accelerator Laboratory, as well as the nuclear physics experiments at the Relativistic Heavy Ion Collider program at Brookhaven National Laboratory, face many

data and collaboration challenges. SLAC's experiment, for example, has already accumulated data sets approaching a petabyte. These data sets will increase in size by a factor of a thousand within the next decade. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals from the detector data, better calibrate the detector, and better reconstruct the quantities of interest. The globally distributed ensemble of computing and data facilities available to HENP, although large by any standard, is less than physicists need to do work in a fully creative way. So a need exists to solve the problem of optimally managing global resources to maximize major experiments' potential for breakthrough discoveries.

Physicists wouldn't have attempted these global-scale collaborations if they couldn't count on highly capable networks to interconnect physics groups throughout the experiments' life cycles. They need networks to enable the construction of grid middleware with data-intensive services that help access, process, and analyze massive data sets. They must also be able to count on highly capable middleware to help manage worldwide computing and data resources to bear on the data-analysis problem of high-energy physics.

To meet the technical goals, the experiment management team must set priorities for using the available computing and network and manage and monitor the system globally end to end. A new mode of human-grid interactions must be developed and deployed so that the physicists and a grid system can learn to maximize the workflow through the system. Developing an effective set of trade-offs between high resource utilization levels and rapid turnaround time and matching resource-use profiles to each scientific collaboration's policy over the long term present new challenges (in scale and complexity) for distributed systems.

This will involve

- Managing authorization to access secured, worldwide resources
- Data migration in response to use patterns and network performance
- Naming and location transparency for data and computing resources
- Direct network access to data management systems
- Publish and subscribe and global discovery
- Monitoring to optimize use of network, computing, and storage resources

- Policy-based scheduling and brokering to reserve the resources needed for a task
- Automated planning and prediction to minimize the time to complete tasks and maximize utilization, which includes tracking worldwide resource-use patterns

In the context of semantic services, AI-based planning techniques increasingly (and necessarily) are being used to optimize resource use.^{10,11}

However, at problem-solving abstraction's highest level, where physicists interact with data that's as refined as possible through automated techniques, a need for knowledge management exists. Stewart Loken, from the Lawrence Berkeley National Lab's physics division, provides this example:

HEP experiments collect specific types of data for the particles that result from high-energy collisions of the protons, electrons, ions, etc. that are produced by the accelerators. The types of data are a function of the detector and include things like particle charge, mass, energy, 3D trajectory, etc.

However, much of the science comes from inferring other aspects of the particle interactions by analyzing what can be observed. Many quantities that are derived from what is observed are used in obtaining the scientific results of the experiment. In doing this more abstract analysis, the physicist typically goes through a process like the following.

Events of interest are usually characterized by a combination of jets of particles (coming from quark decays) and single particles like electrons and muons. In addition, we look for missing transverse energy (an apparent failure of momentum conservation) that would signal the presence of neutrinos that we cannot detect.

The topologies of individual events follow some statistical distribution, so it is really the averages over many events that are of interest. In doing the analysis, we specify what cone angle would characterize a jet, how far one jet needs to be from another (in three dimensions), how far from the single particles, how much missing transverse energy, the angles between the missing energy vector and the other particles, etc.

What I would like to see is a set of tools to describe these topologies without typing in lots of code—e.g. a graphical interface that lets you draw the average event and trace out how statistical variations would affect that. We do simulation of interesting processes and they guide the selection of events, so we would want to learn from that as well. In order to transform these sorts of queries into combinations of existing tools and appropriate data queries, some sort of knowledge-based framework is needed.

In the next section, I describe this sort of semantic service, which organizes operations within a single domain model.

Semantic services

To realize the benefit of a componentized science simulation environment that's rich with discipline data, three types of capabilities related to automatic query structuring are necessary. That is (at least initially), the semantic services noted earlier are primarily related to automatically verifying and structuring various forms of queries within a scientific discipline's fairly well-defined and stylized environment.

Category 1

The first necessary capability is being able to check the validity of complex sequences that the user manually constructs and provide guidance if they're incorrectly structured.

A scientist might well know how to formulate an abstract sequence of operations on data that will answer a question or get a desired result in terms of the science analysis steps. However, the exact forms of analysis and simulation components and available data might not be directly suitable for the desired sequence at the science level, or the available components might produce the desired transformation only if invoked in certain ways. The specifics on permitted connections between components or data formats must be encoded in semantic models. Then the models can provide higher-level constraints on interrelationships and inform users of constraint violations.

Some precursors to this capability exist, such as graphical model builders that enforce semantic data compatibility for a few data types when building the workflow network that represents the discipline model. The user gets help through a graphical programming language that enforces certain constraint types among the building blocks. However, this approach is rigid and can represent only a limited range of interconnection relationships.

A generalization of this capability is necessary. It should provide detailed descriptions of data through metadata and XML schema, and corresponding descriptions of the kind of data needed as a simulation component's input and produced as output. So, we need tools that can check constructed workflows' validity through a complete compatibility analysis of input, output, and data types and formats as well as the semantic relationships among components. Tools should report incompatibilities in a meaningful way that indicates what components might correctly interact, what data characteristics an operation needs, what data formats are available, and so on.

Category 2

The second capability is to automatically build simple composite operations from libraries of simpler ones on the basis of components' semantic relationships. That is, given the semantic relationships among a fairly limited and well-defined set of primitive operations and data in a well-defined discipline model, semantic tools should automatically construct compound operations that transform the data by invoking primitive operations in the correct order. For example, if a user wants a particle's linear velocity components and the available data provides angular momentum and mass, the tools should automatically assemble the sequences of transformations that derive linear velocity. This capability would address the requirements of the example query from the high-energy physics case study I discussed earlier.

Category 3

The third capability is to describe not only complex discipline models but also these models' interactions. These semantic services should provide higher-level constraints on interrelationships to automatically order the various models' simulation components and data transformations in response to certain queries. For this, it's necessary to represent the multidisciplinary relationships that make up models such as the terrestrial biogeoscience environment in Figure 2, and the types of questions that might be asked related to these models.

This is a critical capability. As we tackle broader, more realistic problems, problem solving will always result from multidisciplinary simulation and data analysis. But to realize this process's full benefit, it must be available to a wide range of practitioners. If we must assemble a team of experts representing each discipline of the multidisciplinary model every time we want to make changes, multidisciplinary modeling's utility will be limited. We need to encode enough discipline knowledge in general semantic models so that they can answer what-if questions in specific areas. In other words, if subdiscipline specialists need to change their model components or configuration to experiment or solve individual problems, scientists should still be able to use higher-level discipline models to ensure the configuration's overall correctness without needing to consult other experts.

Furthermore, practitioners who aren't specialists in any of a model's subdisciplines should be able to reliably use the model. A

complex, multidisciplinary simulation characterizes a jet engine's operation. However, an aircraft designer only wants to know how to configure that simulation to provide the appropriate responses when coupled with a particular aircraft design, particular atmospheric conditions, and so on.

This sort of scenario is characterized by the following two examples. In these examples, answering the what-if question requires assembling different components in different ways within a general discipline-model framework that imposes constraints to ensure the combined components operate correctly. Different combinations of submodels are put together automatically.

Consider this example. What will my itinerary look like if I wish to go from San Francisco to Paris to Bucharest? In Bucharest, I want a three- or four-star hotel within three kilometers of the Palace of the Parliament. The hotel cost can't exceed the US Department of State foreign per diem rates.

To answer such a question—relatively easy but tedious for a human—the system must understand the relationships between maps and locations and per diem charts and published hotel rates, and it must apply constraints (for example, < 3 km, 3- or 4-star, cost < \$ per diem rates).

The second example is a similar but more realistic query that relates to the climate model described earlier. Consider this prototype query: "Within 20 percent, what will be the water runoff in the Comanche National Grassland creeks if we seed the clouds over Southern Colorado in July and August next year?"

To answer such a question, you'd have to

- Understand the details of models and data for precipitation, evapotranspiration, and evaporation
- Figure out what runoff basins are in the Comanche National Grassland
- Locate stream network models
- Obtain historical cloud cover data for July and August
- Determine inputs and outputs for an appropriate precipitation process chain model to characterize seeding results
- Incorporate historical (or current) stream runoff rates
- And more

Each of these models will establish resolution, accuracy, regions of validity, and other characteristics. The data will have to be transformed into specific input types with specific

units, girding, and so on, so that it can be used with the available numerical simulations. This information will be in documentation for the models, online data set descriptions, reference documents describing the data's accuracy, and other forms.

A human would have to extract this information and identify appropriate data conversion programs, figure out how the models relate to each other, set up scripts to run the models and data conversions, organize the intermediate files so that downstream processing steps may refer back to them, and more. It would likely take weeks or months to assemble the required information and gain enough understanding of the models to correctly structure the required operations.

On the other hand, the necessary information can be encoded in metadata about the related data and the services (component input and output data structures). Ontologies can represent the relationships among related components and data, and factors such as accuracy and resolution dependencies.

Intersecting the Comanche National Grassland's geographic location with the runoff basin and stream locations will yield the hydrology basin and the associated stream networks. Ontologies describing hydrologic simulations should give relationships among precipitation, evapotranspiration, and evaporation models; their required input; and so on. Ontologies describing atmospheric moisture data from Earth Observing System satellites should indicate how to appropriately transform this data into the form the models require.

Higher-level ontologies describing relationships among the relevant geophysical systems should describe how to establish the relationships among the subsystem-level ontologies describing the models I noted previously. From these system-level ontologies, tools should construct generalized workflows to get from cloud seeding to runoff. Similar tools applied to the subsystem models will fill out the workflow, generating abstract grid workflows that specify data to obtain, simulations to run, intermediate files to store, and more.

Tools then pass this abstract or general grid workflow to an AI-based planner that constructs and adaptively manages the code execution and data movement. That is, the abstract or high-level workflow describes the relationships among the simulation components and data for solving a particular problem or query. The AI planner optimizes use of available computing and storage systems. These systems execute simulations and manage the resulting

data in a dynamic environment where computers and storage systems can come and go or even fail. Moving simulation codes and data among the available resources to keep the general model workflow progressing toward a solution is a problem distinct from constructing the original, high-level workflow.^{10,11}

With tools that can apply constraint queries against the ontologies, the user should be able to decompose the question into a few constituent parts and quickly find out how the parts must interoperate, what data is needed, and so forth. Ontologies associated with the data should describe how to accomplish transformations of coordinates and units, change resolution, and more; what the required data transformations are; and how to configure them.

Even if you had to manually locate relevant data (we aren't assuming a broad data discovery capability in these examples—that's a separate topic), you would avoid tedious human interventions in the end. This makes possible a broader use of the complex, underlying knowledge and information base. It also lets nonspecialist practitioners more easily get answers to their top-level planning, prediction, and strategy questions.

Anonspecialist should be able to formulate quantitative or qualitative, and declarative or constraint-based queries in problem-solving environments with multiple, related data and simulations operating in several discipline models. Semantic models and tools should generate correctly structured sets of operations—sequencing and parameterizations—and manage acquiring or generating the data to input to the analysis and simulations that will resolve the query. This should be possible across multiple domain models, such as for topography, hydrology, and climate, as the terrestrial biogeoscience example I discussed earlier illustrates. The general data and simulation workflow must be automatically mapped onto appropriate computing and data resources using grid resource brokering and planning services. This involves integrating AI techniques and tools with grid services technology to produce a Semantic Grid.

The first two semantic services categories I described earlier are probably within the scope of current technology, and the third is more visionary. However, such services are necessary to move grids to a central position in the next-generation science process. ■

The Author



William E. Johnston is a senior scientist and manager of the US Department of Energy, Energy Sciences Network in the Information Technologies and Services Division of the Computing Sciences Directorate of Lawrence Berkeley National Laboratory. He is also task manager for the Prototype Data Services in the Grid Common Services project at the NASA Ames Research Center. His research interests include high-speed, wide-area network-based distributed systems, widely distributed computational and data Grids, Public Key Infrastructure-based security and authorization systems, and use of the Internet to enable remote access to scientific, analytical, and medical instrumentation. He received his MA in mathematics and physics from San Francisco State University. Contact him at Lawrence Berkeley Nat'l Laboratory, MS 50B-2239, Berkeley, CA, 94720; wejohnston@lbl.gov, <http://dspd.lbl.gov/~wej>.

Acknowledgments

I thank Dieter Fensel and Mark Musen, the guest editors of *IEEE Intelligent Systems'* March/April 2001 issue on Web ontology languages, along with the authors who contributed to it, for a solid introduction to this topic that will be at the heart of the semantic services envisioned in this article. Thanks also to Horst Simon of the Lawrence Berkeley National Laboratory Computational Research Division for comments on this article's first draft and to an anonymous reviewer whose comments resulted in many clarifications.

The "Climate-modeling requirements" subsection is based on material from Gary Strand (strandwg@ucar.edu, National Center for Atmospheric Research) that was adapted from the High Performance Network Planning Workshop,⁶ and material that Tim Killeen (NCAR) presented at the Blueprint for Future Middleware and Grid Research and Infrastructure⁸ middleware workshops. The "High-energy physics requirements" subsection is based on material from Julian J. Bunn (julian@cacr.caltech.edu) and Harvey B. Newman (newman@hep.caltech.edu) of California Institute of Technology that was also adapted from the High Performance Network Planning Workshop. Barney Pell, Keith Golden, and Piyush Mehrotra of NASA Ames Research Center contributed to the example about water runoff in the Comanche National Grassland creeks in the "Category 3" subsection.

This work was funded by the US Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information, and Computational Sciences Division, Collaboratory Environments Program, under contract DE-AC03-76SF00098 with the University of California; and the NASA Aerospace Technology Enterprise CICT (Computing, Information, and Communications Technology) Program's Computing, Networking, and Information Systems Project. This document is LBNL report no. 54249.

References

1. I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, 2nd ed., Morgan Kaufmann, 2003.
2. F. Berman, G. Fox, and T. Hey, eds., *Grid Computing: Making the Global Infrastructure a Reality*, John Wiley & Sons, 2003.

3. I. Foster et al., "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," *Proc. 14th Int'l Conf. Scientific and Statistical Database Management (SSDBM 02)*, IEEE CS Press, 2002, pp. 37–46.
4. M. Govindaraju et al., *XCAT 2.0: A Component-Based Programming Model for Grid Web Services*, tech. report TR562, Dept. Computer Science, Indiana Univ., 2002; www.extreme.indiana.edu/xcat/publications/tr-xcat.pdf.
5. D. Gannon and R. Bramley, "Middleware Technology to Support Science Portals: A Gateway to the Grid," US Dept. of Energy National Collaboratories Program, 2003; <http://doecollaboratory.pnl.gov/research2>.
6. *High Performance Network Planning Workshop*, US Dept. of Energy Office of Science, 2002; <http://doecollaboratory.pnl.gov/meetings/hpnpw>.
7. *DOE Science Networking Challenge: Roadmap to 2008*, US Dept. of Energy Office of Science, 2003; <http://gate.hep.anl.gov/lprice/Roadmap/index.html>.
8. *Blueprint for Future Science Middleware and Grid Research and Infrastructure*, Large Scale Networking Coordinating Group's Middleware and Grid Infrastructure Coordination Committee, 2002; www.nsf-middleware.org/MAGIC.
9. *A Science-Based Case for Large-Scale Simulation*, US Dept. of Energy Office of Science, 2003; www.pnl.gov/scales.
10. E. Deelman et al., "Mapping Abstract Complex Workflows onto Grid Environments," *J. Grid Computing*, vol. 1, no. 1, 2003, pp. 9–23; www.isi.edu/~blythe/papers/grid-journal03.html.
11. M.D. Rodriguez-Moreno, P. Kearney, and D. Meziat, "A Case Study: Using Workflow and AI Planners," *Proc. 19th Workshop UK Planning and Scheduling Special Interest Group (PLANSIG 2000)*, Open Univ., 2000; <http://mcs.open.ac.uk/plansig2000/Papers/MDR-Moreno.pdf>.