What can the Semantic Grid do for Science and Engineering?

Jim Myers
Associate Director, Collaborative Cyberservices
NCSA

Extended Abstract:

Scientists and Engineers have been happily performing research and Analyses for hundreds of years without the Semantic Grid. What's changing in their world now that would motivate them to look to the Semantic Grid? Which of their problems can it solve? And how can we recognize the low-hanging fruit – the combinations of communities and issues where introducing the Semantic grid now will create the most scientific value?

With advancing technologies, scientific productivity, roughly the number of experiments per researcher, has been climbing rapidly. Instruments have become faster, more accurate, more manageable, more automated, and have greater signal-to-noise. One result is that traditional science can be done faster and more broadly (experiments on a broader range of samples), leading to information overload. Combined with the increased ability to analyze and share data, another result is researchers can perform new types of analyses – looking for higher dimensional relationships through data mining, feature detection, correlation methods, clustering, etc. Thus, researchers are building community data stores and analyzing colleagues' data. It has also become possible to work across communities – to do systems-science and ab-initio engineering. The issues involved can be understood by 'Considering a spherical cow'. A spherical cow model is adequate for example, for studying cow metabolism. Expanding the model for use in studying cow locomotion would add unnecessary complexity for studying cow metabolism and could lead to requirements to take measurements (i.e. cow height) to be able to share data that would not otherwise be required for the experiment being performed. Should we reach consensus here? Or should researchers keep their current models and be able to integrate data via assumptions (.e.g. assuming a standard cow shape and inferring height from the mass in the spherical model)? What if a researcher in, for example, quantum mechanics, is involved in communities modeling combustion, materials science, atmospheric science, biology, etc. – reaching consensus suddenly involves all of science. Clearly a way to describe models and the mappings between them is needed.

If one then looks at semantic technologies widely – i.e. at not only the stack of RDF, OWL, etc., but at topic-based messaging, aspect-oriented computing, translation as a first class operation, agents, etc. – the power of semantic technologies in science becomes clear. It's useful to first ask 'what can semantic technologies do in general' – make information computer readable? Enable collaboration through standard models? Provide decoupling between models and between models and software? Or standardization of model representation allowing model manipulation? In fact, on the last two are unique to semantic technologies – binary files are computer readable and can function as a shared

Dagstuhl Seminar Proceedings 05271
Semantic Grid: The Convergence of Technologies
http://drops.dagstuhl.de/opus/volltexte/2005/395

standard. Semantic technologies make it easier to discuss models separately from implementation and allow the development of software that can visualize and analyze multiple models.

A common syntax for describing semantics is clearly useful – analogous to the use of common ways to describe mathematical equations. And it can be seen as decoupling – decoupling human discussion of semantics from the syntax used to describe them. Beyond this, decoupling can be important in science and engineering in a number of (arguably related) areas. Data described by models can be transformed more easily using generic translator/parser software, providing 'data virtualization' (beyond, for example, data location virtualization). Similarly, one can use ontologies to simplify mapping event messages, enabling easier assembly of modules into workflows or applications into problem solving environments. At the highest level, one can describe the scientific models being instantiated in the data and applications and consider mapping between models of different phenomena to enable systems science/ab-initio engineering.

To put this in some context, consider the functionality related to 'scientific logistics' – project organization, experiment planning, data organization, note-taking, process logging, etc. These applications have shares concepts such as resource names and hierarchical structures, but most also have more specialized or unique concepts – from owners and access policies to notebook pages, quality/trust (e.g. in curation), etc. Since it is important for coordination of this functionality – authoring notes about logged processes used to make quality judgments for curation, it is tempting to consider an object model – develop a 'resource' object that has all the methods and attributes required. However, an aspect/semantic model makes more sense, i.e. giving control of the definition of attributes and methods to the developers on individual tools and using a metadata (triple) store to capture the tools outputs in a common information space (and then allowing, for example, a user to instruct a notebook to interpret the project/experiment hierarchy created by a project planning tool as a notebook/chapter hierarchy to synchronize these tools). The Architecture breakout group at the 2nd International Data Provenance and Annotation Workshop (see the Background link from the Semantic Data Grid project site given below) described a number of high-level aspects of data and attempted to start mapping them to lower level shared concepts/services. Although the details are outside the scope of this discussion, these are some of the ideas we've been exploring in our Scientific Annotation Middleware and Collaboratory for Multiscale Chemical Science, and Semantic Data Grid projects (http://www.scidac.org/SAM/, http://cmcs.org/, http://collaboratory.pnl.gov/sdg/).

With this background, one can start to map the strengths of semantic grid technologies to the issues facing science and engineering researchers through use cases. A common syntax for describing semantics is useful in standardizing science and engineering reference data – such as heat capacities for different substances – as well as the outputs of standard analytical techniques such as mass spectroscopy. Data virtualization helps in assembling/integrating reference data in 'legacy' systems. Similarly, research communities seeking to standardize experiment protocols (e.g. data analysis pipelines) can benefit from communication virtualization, allowing different programs to be

assembled into workflows. In some cases, communities may be seeking to standardize specific aspects of their work, i.e. ways of annotating data, and application virtualization can help with the development of tools that can easily be integrated into environments. Finally, researchers working across disciplines or across scales – systems science, applications-driven research, consequence-based or ab-initio engineering, etc – can use model-level virtualization to map between disciplinary world views. Note that for all of these cases, although the semantic web aspect is highlighted in the description, the grid concepts of distributed, independent resources and virtual-organization-level agreements about appropriate policies and shared understandings are also critical.

A number of recent reports highlight these trends in science and information technology requirements that match the semantic grid. In the US, the NSF CyberInfrastructure[1] and Long-Lived Data Collections[2] reports, the DOE Data Management[3] and National Collaboratories[4] reports (and presumably reports from upcoming workshops to define the successor of the Scientific Discovery through Advanced Computing (SciDAC) program) identify issues such as data heterogeneity, the growing need to share and integrate data, the management of large, complex, dynamic analysis processes and coping with the output of high-throughput techniques. In all of these areas, standardized syntax for semantics, the decoupling inherent in the aspect/metadata approach, the scalability of the web/web service approach, and the ability to scope resources, properties, and agreements in virtual organizations can all simplify the development – and evolution – of solutions.

While the semantic grid may not be the only way to make progress on these issues, it provides a very powerful model that addresses issues that are often under-appreciated. Anyone who has served on a standards committee(s) knows why the joke about 'the good thing about standards is that there are so many of them' will continue to evoke wry smiles. A quote from Alice In Wonderland gets to the heart of the matter:

> "When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean--nether more nor less."
> "The question is," said Alice, "whether you can make words mean so many different things."
> "The question is," said Humpty Dumpty, "which is to be the master--that's all."

Heterogeneity of understanding is central to creativity, and semantic grid technologies promise to allow researchers (i.e.. science and engineering researchers, *and* computer scientists) to remain masters of their domains while collaborating and coordinating at an unprecedented scale.

[1] http://www.nsf.gov/cise/sci/reports/toc.jsp
[2] http://www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf
[3] http://www.sc.doe.gov/ascr/Final-report-v26.pdf
[4] http://dsd.lbl.gov/Collaboratories/NCWorkshop/agenda.htm