Collaborative Knowledge Sharing for E-Science

Kirk D. Borne and Timothy Eastman

QSS Group Inc.
NASA Goddard Space Flight Center
Greenbelt, MD 20771 USA
kirk.borne@gsfc.nasa.gov, eastman@mail630.gsfc.nasa.gov

Abstract

Science research programs have become massive data producers. This ability to produce large data volumes must be matched by technologies that make better use of the data flood and that facilitate reuse of the data, in order to reap the maximum scientific return from our research investments. In particular, the extraction and integration of knowledge from multiple data sources must become standard practice, both for science-enabled decision support and for scientific discovery. We describe the emerging e-Science paradigm and its application to data-driven knowledge discovery and collaborative knowledge sharing.

The Science Data Flood

The flood of data in all disciplines poses a critical challenge to the advancement of science. The challenge is manifest in several areas: data discovery, data access, information retrieval, knowledge discovery, knowledge representation, and knowledge sharing. As the quantity of data increases at fantastic rates over the next several years, these challenges may become insurmountable obstacles to scientific research and discovery. The advent of semantic technologies for knowledge representation and discovery will provide some relief in this environment. But this is not enough. The quantity of knowledge gained through the scientific enterprise -- research and analysis -- will also outpace our ability to track and assimilate it all. With these challenges in mind, we are investigating an e-Science paradigm for scientific research, analysis, and discovery. This is based upon emerging computational technologies, including web services, data-centric high-performance computing, data mining, ontologies, and the semantic web.

As a base for space science data archiving and analysis, NASA's Space Physics Data Facility and National Space Science Data Center provide a rich foundation of data and data analysis tools with which to advance an e-Science research paradigm (Eastman et al. 2005). In addition, future projects, which include very large data-producing astronomy and astrophysics projects (such as the Large Synoptic Survey Telescope project), will contribute to and benefit from this evolving e-Science research framework.

Annotation Systems for Science

We are beginning to investigate the use and utility of annotation systems for collaborative scientific knowledge acquisition and knowledge-sharing (e.g., BioDAS.org). These interoperable information sources will facilitate continued growth in collaboration, knowledge-sharing, and science research that parallels the growth in data volumes and in information content within expanding science data repositories. Already in use within the bioinformatics and genomic sciences research communities, distributed annotation systems (DAS) provide a tool for knowledge feedback and sharing among a large community of researchers who are working with very large data repositories. The DAS provides an accessible on-line means for researchers to provide feedback through annotated databases.

Annotations include new scientific knowledge discovered through normal research activities. These dynamic, distributed, data-driven annotations will constitute a digital library of scientific knowledge. The annotations will be visible and reusable by the entire research community, thus further advancing science while minimizing duplication of effort in data mining, analysis, and knowledge discovery.

An annotation system will have numerous facets of particular interest and relevance to the AI and Semantic Web communities, including ontologies for knowledge representation and markup, information provenance, extracting knowledge from distributed data sources, schemas for annotated knowledge bases, ontology reconciliation, collaboration, and trust. All of this relies upon the invocation of important computational techniques in the areas of artificial intelligence, databases, data mining, machine learning, information integration, web services, security informatics, social informatics, and more.

e-Science

E-Science refers to the internet-enabled sharing of distributed data, information, computational resources, and team knowledge for the advancement of science. The technologies that enable this distributed sharing of resources are the same technologies that enable the equivalent paradigms of e-Commerce, e-Business, and e-Government. Machine-to-machine protocols and standards are used for passing messages between processes (e.g., queries and responses), for describing services and resources (e.g., data, software packages, computational environments), for discovery of these resources (e.g., through open registries), and for integrating data and responses from distributed systems. These technologies enable web portals to perform their function: one-stop shopping for airline tickets, for hotels, for books, or for data. E-Science thus builds upon and takes advantage of technologies that are built to satisfy economic and functional requirements. The scientific user requirements of e-Science are sufficiently parallel to those of e-gov and e-biz that technology transfer is relatively straightforward. Typical user requirements on modern internet-based science data systems include: find the right data right now; one-stop shopping for all data needs; integrate data from multiple sources with minimal human intervention; transparent user access to heterogeneous databases and query systems; metadata-assisted data fusion; identify the most important and/or unique events within a massive data collection; and knowledge discovery in databases (KDD).

Knowledge-Building Systems

Through the acquisition of new data and the development of data-driven models, a science research program can be seen as a knowledge-building system. The system is building new knowledge dynamically about its application domain and about its scientific results. Ideally, the acquisition of this new knowledge is not an academic exercise. Rather, this new knowledge can be shared with other systems, including other researchers, other data systems, or other scientific sensors. The use of ontologies to represent knowledge (e.g., using OWL) enables knowledge-sharing and reuse. As discoveries are communicated and knowledge is shared across distributed heterogeneous systems and research programs, the scientific knowledge-building capability is magnified.

A Research Program

Our research program is in its early stages. We are focusing initially on three research areas: DAS (Distributed Annotation System) user requirements analysis, the design of distributed annotation systems for e-Science, and the development of a new space science informatics research discipline. We are sharpening the latter into the specific development of a new discipline of astro-informatics that parallels bioinformatics and geoinformatics (geographic information systems) as stand-alone science disciplines within their main science domain. A discovery informatics infrastructure and an informatics research foundation are prerequisites to the user acceptance and usability of the DAS, as demonstrated by the use of BioDAS.org within the bioinformatics community. Our approach focuses on AI, machine learning, and semantic web technologies for collaborative knowledge acquisition and sharing.

The acquisition of scientific data in all disciplines is now accelerating and causing a nearly insurmountable data avalanche. Assimilating these data into models and using these data and models to drive scientific measurement systems are major scientific challenges for today's large scientific research projects. The application of an e-Science paradigm, including Grid computing (Borne 2005), Web Services, Semantic knowledge representation, and machine learning algorithms will enhance the scientific return and knowledge-building capabilities of future science programs. These information technologies will enable us to address data-intensive problems that would not otherwise be manageable. This will permit large research projects to make use of larger data volumes in the scientific discovery and modeling process than is currently possible.

References

Borne, K. D. 2005. “Grid-Enabled Science with the National Virtual Observatory,” in the *NASA Workshop on Grid Computing*. Downloaded on August 29, 2006 from http://romulus.gsfc.nasa.gov/msst/gridws/.

Eastman, T., Borne, K., Green, J., Grayzeck, E., McGuire, R., and Sawyer, D. 2005. “eScience and Archiving for Space Science,” *Data Science Journal*, vol. 4, pp. 67-76.