6.1 CHALLENGES AND OPPORTUNITIES FOR A NEW GENERATION OF DATA SERVICES FOR GEOSCIENCE EDUCATION AND RESEARCH

Mohan K. Ramamurthy*
Unidata, University Corporation for Atmospheric Research
Boulder, Colorado

1. INTRODUCTION

A revolution is underway in the role played by cyberinfrastructure and data services in the conduct of research and education. We live in an era of an unprecedented data volume from diverse sources, multidisciplinary analysis and synthesis, and active, learner-centered education emphasis. For example, modern remote-sensing systems like hyperspectral satellite instruments and rapid scan, phased-array radars are capable of generating terabytes of data each day. Complex environmental problems such as global change and water cycle transcend disciplinary and geographic boundaries, and their solution requires integrated earth system science approaches. Contemporary education strategies recommend adopting an Earth system science approach for teaching the geosciences, employing new pedagogical techniques such as enquiry-based learning and hands-on activities. In essence, today's education and research enterprise depends heavily on robust, flexible and scalable cyberinfrastructure, especially on the ready availability of quality data and appropriate tools to process, manage, analyze, integrate, and visualize those data.

Fortuitously, rapid advances in computing, communication and information technologies have also revolutionized the use of data, tools and services in education and research. The explosive growth in the use of the Internet in education and research, largely due to the advent of the World Wide Web, is by now well documented. On the other hand, how other technological, social and cultural trends have shaped the use of data services is less well understood. For example, the computing industry is converging on an approach called Web services that enables a standard and yet revolutionary way of building applications and methods to connect and exchange information over the Web. This new approach, based on XML – a widely accepted format for exchanging data and corresponding semantics over the Internet - enables applications, computer systems, and information processes to work together in a fundamentally different way. Likewise, the advent of digital libraries, grid computing platforms, interoperable frameworks, standards and protocols, open-source software, and community atmospheric models have been important drivers in shaping the use of a new generation of end-to-end cyberinfrastructure for solving some of the most challenging scientific and educational problems.

2. KEY DRIVERS

Contemporary data services need to be firmly grounded in scientific and education drivers, community needs, technological and socialogical trends.

2a. Science Driver:

Numerous national and international reports underscore the importance of interdisciplinary environmental research and education. Among them are Grand Challenges in Environmental Science (NRC 2001) and Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century (NSF 2003). The NRC report points to a growing recognition that "natural systems, ecosystems, oceans, drainage basins, including agricultural systems, the atmosphere, and so on are not divided along disciplinary lines". Two of the grand challenges identified by the NRC, biogeochemical cycles and climate variability, depend heavily on integration of data from several disciplines. Another excellent example is hydrologic forecasting, one of the four challenges prioritized as deserving immediate investment.

The NSF decadal plan for Environmental Research and Education (ERE) also echoes the need for improving our understanding of the natural and human processes that govern quantity, quality and availability of freshwater resources in the world. While recent advances in remote sensing, combined with a new generation of coupled atmospheric-hydrological models, are driving a new revolution in hydrometeorological predictions, future research and education in this area will require finding and integrating observational and model data from the oceans, the atmosphere, the cryosphere, and the lithosphere, crossing the traditional disciplinary boundaries.

According to NSF director Rita Colwell (Colwell 1998), "Interdisciplinary connections are absolutely fundamental. They are synapses in this new capability

* *Corresponding author address*: Dr. Mohan Ramamurthy P. O. Box 3000, Unidata/UCAR, Boulder, CO 80307-3000. Email: mohan@ucar.edu

to look over and beyond the horizon. Interfaces of the sciences are where the excitement will be the most intense... ." For example, studies on societal impact of and emergency management during hurricane-related flooding involve integrating data from atmospheric sciences, oceanography, hydrology, geology, geography, and social sciences with data bases in the social sciences.

Similar multidisciplinary needs are emerging to solve certain disaster/crisis management problems. Two highly topical examples are fire-weather forecasting and environmental modeling for homeland security. In homeland security, for instance, there is a need to forecast the dispersal of hazardous radioactive, biological, and chemical materials that may be released (accidentally or deliberately by terrorism) into the atmosphere. For the latter scenario, detailed, four-dimensional information on transport and dispersion of hazardous materials through the atmosphere, and their deposition to the ground are needed at a resolution of individual community scales. Moreover, this information needs to be linked in real-time to databases of population, evacuation routes, medical facilities, etc. to predict the consequences of various release scenarios (e.g., how many people will be exposed to or will be injured by potentially dangerous concentrations of those materials).

In addition to identifying national priorities and computational grand challenges in the sciences, many of the NSF and NAS reports cited above have also documented infrastructure needs, including comprehensive data collection, management and archival systems and new methods of datamining and knowledge extraction. For example, the NSF ERE Advisory Committee calls for building infrastructure and technical capacity with a new generation of cyberinfrastructure "to support local and global research and to disseminate information to a diverse set of users including environmental professionals, the public, and decision makers at all levels." Toward building the cyberinfrastructure, the ERE agenda foresees the need for a comprehensive suite of data services that will facilitate synthesis of datasets from diverse fields and sources, information in digital libraries, data networks, and web-based materials so that they can serve as essential tools for educators, students, scientists, policy-makers and the general public. Similar needs for web-based real-time and archived data services, including digital library integration and fusion of scientific information systems (SIS) with geographic information systems (GIS), were expressed at the NSF-sponsored Workshop on Cyberinfrastructure for Environmental Research and Education.

A growing number of universities are engaged in real-time modeling activities, and this number is expected to increase as advances in computing and communication technologies facilitate local atmospheric modeling. A new generation of models (e.g., the Weather Research and Forecasting [WRF] model) can predict weather on the sub 1-km scale, with the potential to address community-scale concerns. Providing initial and boundary condition data along with analysis and visualization tools for these efforts requires an extensive cyberinfrastructure.

2b. Education Driver

Challenges facing science education have been well articulated in a number of documents (e.g., Shaping the Future (AGU 1997) and Geoscience Education: A Recommended Strategy (NSF 1997)). They recommend adopting an Earth system science (ESS) approach for teaching the geosciences, integrating research experiences into curricula, employing contemporary pedagogies, and making appropriate use of educational technologies. Science education should also be about teaching students the language of science and providing students with opportunities to engage in scientific inquiry and investigation (Lemke, 1990).

Shaping the Future also calls for an inquiry-based approach to science education. For example, hands-on, learner-centered education in meteorology depends on the availability of meteorological data and analysis and display tools of high quality. By supplying these data and tools, programs like Unidata have been instrumental in transforming learning in the atmospheric sciences. Digital libraries (exemplified by efforts like the National Science Digital Library (NSDL) and the Digital Library for Earth System Education (DLESE)) augment web-based learning resources with high-quality data resources that can be embedded in interactive educational materials. Internet-based tools also open data access for faculty and students at small colleges where little system administration support is available for the installation of advanced data systems and applications. Engaging students with real-world data is a powerful tool for not only motivating students but also helping them learn both scientific content and principles and the processes of inquiry that are at the heart of science (Manduca, 2002). For example, working with real-world examples and actual data can place learning in a context that is both exciting and relevant by providing connections between classroom instruction and a student's experience with their local environment (e.g., diurnal temperature changes and seasons), major weather events (e.g. tornadoes, hurricanes, blizzards) and climate events (e.g., global warming).

In essence, making science education more authentic and process-oriented requires access to the same kinds of data, tools, and models as those used by scientists, needing new types of data services that are far easier to use by the science education community.

2c. InformationTechnology trends

Computers and information technologies are now playing a central role in this complex and ever-changing world in which we live and work, with the World Wide Web reshaping almost every aspect of our work, including education, research, and commerce. Computing, communications and information technology trends of recent years have not only had a democratizing effect on daily life, but they have also changed the very nature of data services for education and research. For example, below is a partial list of key technologies and trends that have enabled a new generation of end-to-end data services in the scientific community:

- Internet
- Commodity microprocessors
- World Wide Web
- Open source model
- Object-oriented programming
- Open standards, frameworks, and conventions
- Extensible Markup Language (XML)
- Standards-based web services
- Digital libraries
- Collaboratories
- Grid computing
- Data portals and federated, distributed servers
- Geographic Information Systems
- Ontologies and Semantic web
- Data mining and knowledge discovery

This section highlights a few of the above key information technology (IT) advances and trends that have revolutionized the provision and use of data in the geosciences. Taken together, the above technologies have enhanced the ability of data providers to better serve their communities, lower the costs for the users, and allowed a greater participation in the data activities in new networked world.

The introduction of microprocessor based computer systems in the 1980s, combined with the increased connectivity of college campuses to the Internet, led to a transition from large scale, mainframe based technologies to low-cost distributed systems, making it possible for widespread access to and use of scientific data. The wiring of universities for Internet connectivity was a prerequisite for receiving data via, for example, the Unidata Internet Data Distribution system and the Local Data Manager, which use TCP/IP communication standards for data transport.

The advent of the World Wide Web (or simply the Web) in the 1990s brought about a revolution in information services. It was directly responsible for not only the explosive growth of the Internet and its users, but it also provided the ability to provide interactive, remote services. In the process, the Web radically transformed the sharing of data and information and resulted in greater use of communication infrastructures to create and store information and then to deliver it from providers to end users. The Web also brought with it a massive proliferation of online educational materials, many of them based around extensive use of interactive services. Services and tools were created to help one communicate, search for information and data, and make information and data available on the Internet. In the process, library services evolved from local traditional collections to global resources provided on demand via the Web, ushering in the era of digital libraries.

The 1995 NSF-sponsored Digital Libraries Workshop entitled "Interoperability, Scaling, and the Digital Library Research Agenda defined digital libraries as: "An organized collection of multimedia data with information management methods that represent the data as information and knowledge."

Even though retrieval of information efficiently is arguably the most important role of digital libraries, a potentially even more valuable contribution of digital libraries is their ability to preserve, catalog, and curate information, extend discourse, build communities that provide richer contexts for people to interact with information and each other, all toward the creation of new knowledge. According to Griffin (1998), the real value of digital libraries may ultimately prove to be in their ability to “alter the way individuals, groups, organizations etc, behave, communicate, collaborate, and conduct business.” In essence, digital libraries, much like other aspects of the Web, are becoming powerful instruments of change in education and research.

Specifically, modern environmental studies rely on diverse datasets, requiring tools to find and use the data. The data discovery process has become an important dimension of the scientific method, complementing theory, experimentation, and simulation as the tools of the trade. Future success will depend on how well researchers are served by tools and services pertaining to data discovery and use.

Another notable trend information technology trend is the desire to integrate all information, including data and a variety of services behind a single entry point or a portal. Portals often include personalization features, allowing users a tailored view into the information. The customization permits: a) a single point of authentication to validate access permissions and enable links to available resources. b) the ability to design a customized view of available information.

The open access, open source and open standards are inter-related concepts that are gaining momentum and developers of data service are aggressively rethinking how they might both contribute to and benefit from these trends toward “openness”. The benefits of open access, open source, and open standards for are numerous.

Open source software is software that includes source code and is usually available at no charge. The open source model for software has many benefits. For instance, it has the advantage of harnessing the collective wisdom, experiences, expertise and requirements of large communities. Additional features and benefits include scalability, extensibility, and customizability. For example, people using a wide variety of hardware platforms, operating systems, and software environments can test, modify, and run software on their system to test for portability.

Open source software also increases opportunities for software reuse, adaptation to different hardware and software environments, and customization to user needs. The best example, perhaps, is the use of Linux in a wide range of electronic and computer systems. Linux has been ported to everything from embedded microcontrollers, videogame consoles, mobile phones, PDAs, network routers and wireless access points, personal and mainframe computers, to massively parallel high-performance computing systems. The pervasive use of Linux, coupled with the availability of inexpensive, commodity microprocessors and storage devices, has dramatically reshaped the scientific and data services landscape in the geosciences.

The use of open source model for middleware, a special kind of software between client and server processes to ensure consistency and interoperability, is particularly important for developing new data services. For example, it enables the provision of a standard, stable, consistent interface to a wide variety of applications, on a broad set of platforms and enable their inter-operability. In the process, it decouples data service providers from users, allowing end users with multiple clients to access the same services. This can accelerate the migration of data services to new and diverse platforms. Furthermore, it facilitates the “wrapping” of legacy systems in standard interfaces, giving them the ability to integrate with other distributed components and systems. Given the demand for standards-based, open systems that easily integrate, the open source development process provides a significant advantage over proprietary approaches to software development and use.

Interfaces, based on open standards are by definition publicly documented and based on an explicit or de facto standard. There is evidence that well developed open standards for data formats are less likely to become quickly obsolete and are more reliable and stable then proprietary formats. Having access to the file format also allows users and developers to create data conversion utilities into other formats. File formats that use open standards can assist in long-term archiving because they allow for software and hardware independence. Open standards also allow for greater flexibility and easy migration to different systems and interoperability of diverse systems. Open access, open source software models, and open standards each offer a number of significant benefits in the provision of data services. However, when they are combined the benefits can be even greater.

In the data services area, there exist many excellent examples of open source software that are highly reliable and supported by a large community, including Linux, Apache, MySQL, and similar projects. Network Common Data Form (netCDF) and Open-source Project for a Network Data Access Protocol (OPeNDAP) are leading examples of open source software in the geosciences data infrastructure area. Because of its free and open source nature, netCDF software has been incorporated into over 50 other open source software packages and 15 commercial packages, resulting in its widespread use and status as a de facto standard for data format in atmospheric and related sciences. Likewise, the OPeNDAP software has found wide use outside the core oceanographic community where it originated, and is now used for several different kinds of science data.

Extensible Markup Language, XML, is a simple, highly flexible, text-based framework for defining mark up languages. This standard for classifying, structuring, and encoding data allows organizations and services to exchange information more easily and efficiently. Although originally developed to facilitate Web-based publishing in a large scale, XML has since rapidly gained acceptance and usage in the exchange of a wide variety of data on the Web. An important emerging standard for interoperability of data systems is in the metadata area, which can use XML to share descriptions of underlying datasets.

The ability of XML to organize data into a computer-interpretable format that is also easy to code and read by humans is quickly making XML the lingua franca for business services and electronic commerce and also rapidly becoming a widely used standard in the data services world. Because of its simplicity and elegance, XML has radically transformed the provision of data services in the scientific community. Some of its principal benefits include: a) ability to delineate syntactic information from semantic information; b) allows the creation of customizable markup languages for different use cases and application domains; c) platform

independence. For example, XML makes it possible for providers of data services to send information about data sets, metadata, in a form completely separate from the presentation of the underlying data. Furthermore, service providers can present the same information in multiple forms or views using XML style sheets, customized to the needs of particular users. For example, Really Simple Syndication (RSS) is a lightweight XML format for sharing news and bulletins, and it has been successfully used by the U. S. National Weather Service to disseminate weather information such as local forecasts, watches, and warnings to Internet users. The same technology can also be used in the data services context to notify users when new data becomes available in a data system.

Web services, based on XML and HTTP, the two open standards that have become ubiquitous underpinnings of the Web, are emerging as tools for creating next generation distributed systems. Besides recognizing the heterogeneity as a fundamental ingredient, web services, independent of platform and development environment, can be bundled, published, shared, discovered, and invoked as needed to accomplish specific tasks. Because of their building-block nature, web services can be deployed to either perform simple, individual tasks or they can be chained to perform complicated business or scientific processes. As a result, web services implemented in a Service Oriented Architecture (SOA) or framework, are quickly becoming a technology of choice for deploying cyberinfrastructure for data services. By wrapping existing applications as web services in a SOA, the traditional obstacles to interfacing legacy and packaged applications with data systems are being overcome through loosely coupled integration. Such an approach to integration affords an easier pathway to interoperability amongst disparate systems. The new software architectures based largely on Web Services standards are enabling whole new service-oriented and event-driven architectures that is challenging traditional approaches to data services.

According to Foster and Kesselman (1997), the Grid refers to “an infrastructure that enables the integrated, collaborative use of high-end computers, networks, databases, and scientific instruments owned and managed by multiple organizations.” Grid applications often involve large amounts of data and/or computing and often require secure resource sharing across organizational boundaries. Grid computing and the science enabled by it, eScience, are two major trends in distributed computing. A key advantage of grid computing over historical distributed computing systems is that the Grid concept permits the virtualization of computing resource such that end-users have the illusion of using a single source of ``computing power'' without knowing the actual location where their computations are performed. The use of digital certificates to access systems on behalf of a user and third party file transfer between grid nodes authenticated via certificates are specific examples of how Grid technology enables virtualization. Grid Services, which implement web services in a Grid architecture, are in still in their infancy, although several proof-of-concept testbeds have been deployed and in a number of disciplines, including earth and atmospheric sciences, high energy physics, and biomedical informatics.

3. DATA SERVICE ATTRIBUTES:

As articulated by Cornillon (2003), the ultimate objective of a data system or service is to provide requested data to the user or user’s application (e.g., analysis or visualization tool) in a transparent, consistent, readily useable form. The users do not care as much about the technology behind those systems or services, but do about transparency and usability. The key to achieving those two objectives is through interoperability of components, systems and services, via the use of standards.

An ideal data service should have the following attributes:

- User-friendly interface (e.g., portal)
- Transparency (format, protocol, etc.)
- Customization of services
- Server-side operations (e.g., subsetting, subsampling, etc)
- Aggregation of data and products
- Provision of rich metadata
- Integration across data types, formats, and protocols
- Intelligent client-server approaches
- Interoperability across components and services
- Flexibility, extensibility, and scalability
- Ability to chain services
- Support an array of tools for access, processing, management, and visualization

Given the aforementioned trends, the last decade has seen a rapid evolution from proprietary data systems (e.g., EOSDIS) towards more open data services (e.g., Community Data Portal at the National Center for Atmospheric Research, USA; British Atmospheric Data Centre, UK) However, the transition has not been without challenges for a number of reasons, including:

- Heterogeneity and complexity of distributed observing, modeling, data, and communication systems
- Nature of data coverage: diversity and multiple spatial and temporal scales

- Many data systems have been using legacy and contemporary technologies
- In some areas, there is still a lack of standards and interoperability
- User community is not monolithic
- Political, technological, and cultural and regulatory barriers, especially in global access to data

Broad data categories:

While there exist far too many data categories to list in detail, typical data systems in atmospheric sciences must provide a seamless, end-to-end services for accessing, utilizing and integrating data across the following data types:

- Real-time data
- Archived data
- Field and demonstration project data
- Episodic or case Study)
- Data from related disciplines (hydrology, oceanography, cryosphere, chemical and biosphere - soil, vegetation, canopy, evapotranspiration)
- GIS databases

The first four categories, include data from in-situ and remote sensing observations, and output from models.

In addition to data, a broad set of tools and support services should be provided for the most effective use of data.

Given the very high data rates from current and future generation observing systems such as GOES-R and NPOESS satellites, the user community will need a hybrid solution that couples a satellite-based data reception system with a terrestrial, Internet-based data access system. Both local and remote data access mechanisms will be required to deal with the large volumes of data. Both push systems for distributing data (e.g., Unidata Local Data Manager) or just notifications (using RSS feeds) and pull systems for remote access (e.g., OPeNDAP) will be required.

Data Glut, Data Mining, and Knowledge Discovery:

Advances in computing, modeling, and observational systems have resulted in a veritable increase in the volume of data. These data volumes will continue to see an exponential growth in the coming years. For example, data from current and future observing systems will result in a 100 fold increase in volume in the next decade. The GOES-R satellite, scheduled for launch in 2012, will have a hyperspectral sounder with approximately 1600 channels. In contrast, the current generation GOES satellite sounders have 18 thermal infrared channels. Similarly, each NPOESS satellite when fully deployed will have raw data rates of nearly 1 Terabyte each day.

The resulting data glut clearly requires extraction of higher level information useful to users. The process of extracting higher level information is referred to as *data mining.* Data mining is a key step toward data reduction and knowledge discovery. An ideal data system or service should include algorithms and facilities for data mining that can be applied to data sets as needed by users.

4. CONCLUDING REMARKS

This presentation will provide an overview of the above issues and discuss the how these developments are enabling new approaches to applying data services for solving geoscientific problems. Particular focus will be given to the discussion of opportunities and challenges for the geosciences community in light of the trends in observational and information technology, science, and education.

5. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under grants ATM-0317610.