DATA TECHNOLOGIES

Geospatial Data Integration

Karen Stocks, Reporter
San Diego Supercomputer Center
University of California, San Diego

Jim Quinn, Facilitator
University of California, Davis

Initial Questions

- How can information technologies be better used to facilitate integration and synthesis of geospatial data acquired via environmental networks?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

Recent advances have brought exciting changes to the landscape of geospatial information. Remote sensing techniques have created continuous, large-scale coverages of parameters previously sensed only through point-sampling. Affordable, accurate GPS systems have increased the volume of data having good spatial referencing. And upcoming wireless sensors and microchip GPS units show promise for continuing improvements in the future.

Data Issues

Research findings and management decisions will never be better than the data from which they are drawn. Improving the data available for research and management involves creating incentives for data sharing, creating quality base data sets, and focusing research and development funding on new technologies capable of sampling underrepresented data types.

Incentives for data sharing

The largest current limitation on data availability is the lack of incentives for data sharing within the research community. The reigning professional standard of publishing a journal article describing research conclusions generally provides only a text summary of the data. What is needed to facilitate data integration for large-scale, long-term, or multi-factor research is access to fully documented, electronic data sets. It is unrealistic to expect individual researchers to take on the challenge of providing

such data sets when it is both unfunded and unrewarded as a professional accomplishment.

Data management and access needs to become a defined and funded part of any proposal that creates new data sets, and funding agencies must make their expectations (including timelines) for this explicit. Information management typically constitutes 10% or more of the cost of commercial R&D, and funding agencies should expect similar resources to be devoted to making geospatial data interoperable and readily shared.

Professional recognition for publishing data is equally important. Methods for crediting data resources that are parallel to literature citation need to be developed. Scientific societies and publishers should be encouraged to follow the GenBank model, requiring that the raw data for any published article be placed in a database and made publicly available after a set period of time. More generally, the development of robust data resources is often a creative exercise fully equivalent to producing journal articles. With the advent of all-electronic professional outlets for publication, it is feasible and would be highly desirable in terms of both professional recognition and traditional quality-assurance, to have peer review of data sets and accompanying metadata equivalent to the traditional review of articles and books.

Finally, there is a need for a formal "code of ethics" for data use covering the issue of how long an investigator can keep a data set proprietary, how intermediate data products (such as Web resources compiled from published data) are credited and cited, etc. Once the expectations are clear, then institutions and funding agencies (and reviewers) can begin to evaluate researchers based on these expectations.

Base Data Sets

Good research and good policy require the creation of high-quality, standard data coverages that are applicable to a broad spectrum of users. Outside of remote sensing, many geospatial data sets are composed of point data measurements (e.g. soil samples, rare species locations). To create useful products, these points must be integrated and interpolated to create continuous views, using models whose assumptions, limitations, and uncertainties are communicated. The assessment and visualization of uncertainty is a particular research and technological challenge for mapped data, particularly when there are repeat measurements. Often the variables measured (e.g. remote sensing "color") are not the variables of true interest ("land use"). Those coverages that do exist, such as watershed delineations and vegetation indices, have proven to be valuable resources. The aggregation of relevant point data is time-consuming and the process of creating a coverage from point data is best done by scientists familiar with the characteristics of the base data sets in collaboration with statistical/analytical experts. Creating standard products properly and making them available in a variety of formats will reduce redundancy and improve decision-making.

A related problem is the availability and appropriateness of "framework data" - the "base layers" used to spatially reference geospatial data from research projects and monitoring. The Federal Geographic Data Committee has recognized a set of framework data sets (elevation, hydrography, roads, etc.) that are essential for landuse planning and related disciplines, and most have complete national coverages or national initiatives to complete coverages. There is less consensus on the "framework" data essential for environmental research (soils? vegetation? land management practices?), and efforts to address these data needs remain fragmented and underfunded.

New Sensor Technologies

Remote sensing technologies can now create large-scale, high-quality maps of a variety of parameters. However, data types that cannot be remotely sensed are still only represented by limited data points. Priorities for the next generation of sampling technologies must include new methods for measurements traditionally taken through *in-situ,* human-mediated, time-intensive point sampling. Promising avenues include computer-aided video identification of spe-

cies, automated processing of genetic samples, and new acoustic techniques.

Interoperability and Standards

Addressing complex environmental questions requires the integration of data from many resources and the application of multiple informatics tools: GIS, databases, visualization tools, knowledge representations, statistical packages, etc. Current barriers to bringing together heterogeneous data sets and to moving between multiple software platforms form logistical barriers to research progress. While these barriers can be overcome, they require large investments of human effort.

Data format incompatibilities may be partially addressed through standards. Once metadata standards are adopted by the community, this will allow the development of tools that can interact automatically with the metadata. While standards for geospatial data do exist (e.g. Federal Geospatial Data Content standard and Geographical Markup Language), they are not widely used and are not implemented by commercial software packages, in part because they are highly complex. Moreover, the standards address the format of expression but not the actual vocabularies (semantics) used. Much of the power of metadata for information discovery rides on consistent or crosswalked uses of language, which are necessarily tied to particular user communities.

It is recognized that comprehensive documentation of data sets is a worthy goal and that it is unlikely that any single standard will ever suit the plethora of ways in which geospatial data is used. However, the reality is that unimplemented standards are not effective - a data provider creating a small data set that contains location information but is not aimed at geospatial description per se simply will not invest much time in standards compliance without adequate incentives and support.

Software Research and Development

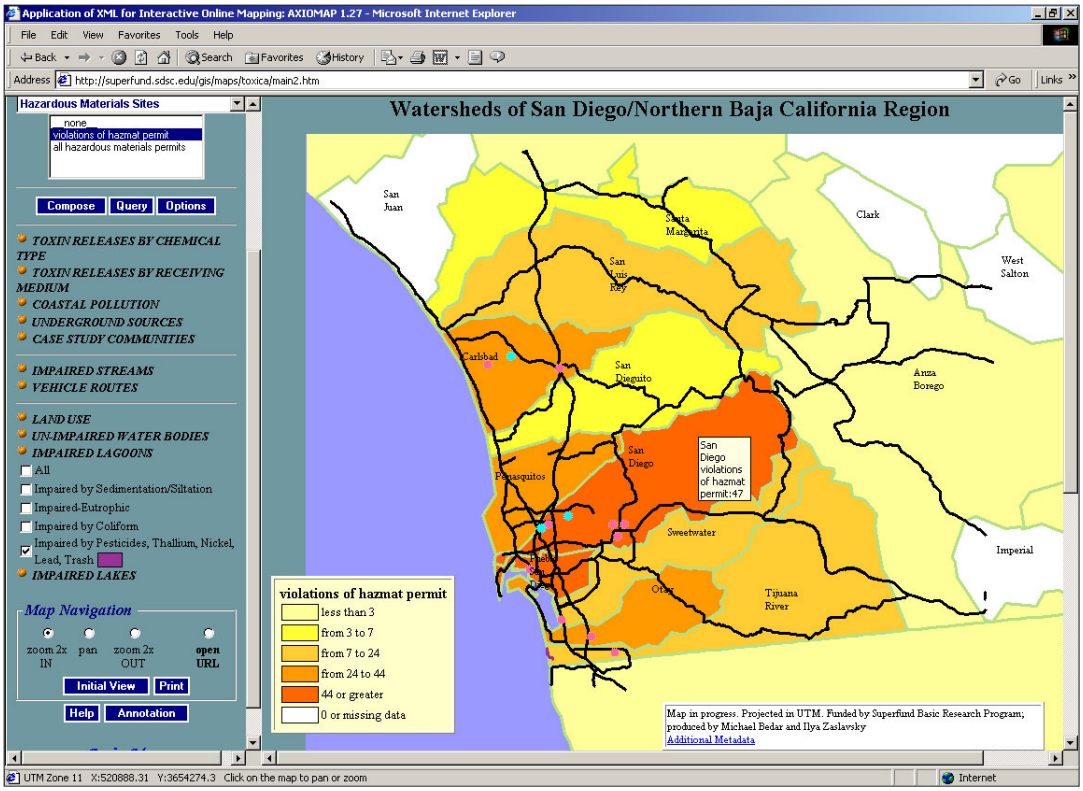
In addition to streamlining the software currently available, new tools and new approaches for working with data are also needed. By streamlining and increasing the capabilities of informatics software, analyzing and processing data can become more efficient and powerful. GIS systems were developed from a cartographic paradigm that does not scale well to today's 4-D data needs: height/depth and time are often poorly represented, and connections to quasi-spatial information are poor. In part this is because the small number of commercial vendors producing mass-market tools cannot respond to the specific needs of small user groups. And when small, specific, individually-built tools are developed to overcome holes in the commercial products, these tools are difficult to integrate with other software and are often not widely available.

The Open Geographical Information System initiative [OpenGIS] partially addresses these issues. However there are still considerable conceptual gaps among the approaches and paradigms of the GIS/cartography/remote sensing community, the visualization community, and the relational database community that need to be bridged to produce an integrated data environment.

It was also noted that there is a spectrum of geospatial data users. While there is a need for powerful and advanced capabilities for leading-edge IT research, there is also a need for user-friendly, easy-to-learn tools for those basic operations common to a broad spectrum of environmental researchers.

Infrastructure

The informatics infrastructure needs to continue to grow and mature in order to support the new data sources and new tools. Overall, advanced geospatial data processing is pushing the computing-power and storage capabilities of the country's infrastructure. Initiatives such as funding for grid-computing projects are welcome additions, and continued support for growing computing systems of a range of sizes is required. Beyond general computing power, there are three specific areas that form, or will soon form, substantial barriers to progress.



Web-Based Interactive Mapping - For Regional Environmental Health Information
(Web EI for San Diego County and Northwestern Baja California)

Web-Based Environmental Informatics (WebEI) is an interactive mapping service that concentrates on integrating and visualizing distributed environmental health data in the San Diego-Baja California border region. Two foci are on issues related to the Total Maximum Daily Load (TMDL) process--an approach to conserving aqueous resources by attending to the total amount of each pollutant that a water body receives--and on community development in the Colonia area of Tijuana.

Available data include impaired waterbodies, watershed boundaries, toxic releases, land use, and soon, health demographics, urban infrastructure (e.g., sewage and power), biodiversity and habitat, and responsible authority. These layers can be overlaid and grouped in various combinations for spatial insight. Users can look in more depth at the issues at work in a particular location by clicking on the point features. As additional data becomes available and integrated into Web EI, it is hoped that this information system will aid in decisions that lead to the sustainable development of the border region.

The first limitation is the lack of organizing elements within online data resources. Data is being served by many groups and at many levels, from individual researchers with desktop servers through field stations, research institutions, libraries, journals, and government agencies from local to federal. While this new data accessibility is an exciting step forward, the hodgepodge of data resources makes it difficult to find a particular data type of interest or to evaluate its quality. Facilities for data clearinghouses/catalogs, tailored search engines, and a method for peer-review and/or user ratings of a resource to indicate its quality are all needed. A key to facilitating the creation of catalogs and searches will be the adoption of metadata standards, including controlled vocabularies, for describing data contents.

Past experience with attempts to establish central repositories for data gathered by individual investigators and programs have not been encouraging. It is likely that authoritative source copies of much of the important geospatial data will remain distributed among thousands of sources and will be somewhat idiosyncratic in content and format, meaning that archival and bandwidth challenges will if anything increase. Mirror sites, portals, and clearinghouses will need robust methods for extracting and validating integratable shared elements from heterogeneous sources, and successively abstracting them, as geographical domains of application increase. Such scalability poses fundamental problems in knowledge representation. It also poses substantial challenges in the sociology of science, since the ability to integrate data requires some consensus in the provider community on the expression of information (semantics and ontologies) in their fields of research.

The second limitation is the lack of long-term data archiving provided in the traditional 3-5 year grant tenure. While national data centers can play a role in this process, enforced data "drop-offs" at the conclusion of a grant tenure will not provide the highest quality data resources. In reality, no data set is ever fully finished, and allowing data authors to have continued access to update and expand their data will improve data quality. It will be crucial to have facilities (and long-term funding) for distributed data centers that allow data management to be kept in the hands of either the authors or the user groups (such as a scientific society or a field station) while still providing a robust framework for data maintenance and access.

The third limitation is bandwidth. Wireless communications paired with micro-GPS and other sensors have ushered in a new era in spatially-referenced environmental sensing. However, the current FCC restrictions on bandwidth are crippling potential applications of sensor networks. Old regulations must be reevaluated in light of current technologies to allow scientific access to bandwidth.

Education

In addition to facilitating interdisciplinary research in geospatial tool development and application, progress in environmental science would be advanced by raising informatics literacy among domain scientists. Just as statistical packages, spreadsheets, and word processors are considered required tools in any scientific domain, environmental scientists today need to have basic familiarity with data management practices and the uses of GIS, database, and visualization software. Efficiently finding, accessing, and using data is intrinsic to the modern process of research and resource management in all fields. Unfortunately, the teaching of geospatial concepts and technologies is particularly fragmented, as important applications cross traditional disciplinary departments. The cartographic conventions underlying modern GIS software have traditionally been taught in geography departments, many of which are struggling and disappearing, and offerings in other departments (optics and remote sensing in Physics, data models in CS, CAD in Engineering, vegetation maps in Biology, geomorphology in Geology) are typically uncoordinated, usually due to institutional barriers to teaching outside one's department or college.

Support for model undergraduate curricula to bring together computer science, geography, and other domain sciences would help provide courses with the appropriate balance of theoretical and applied aspects. Both full courses and IT components integrated into existing domain-science courses are appropriate. The working group also recognized that having people cross-trained in both environmental sciences and informatics (programming, database design, GIS technologies, data server design, etc.) will be critical to future progress and that there is a role for full undergraduate majors or concentrations in interdisciplinary Environmental Informatics. Targeted funding will help institutions develop model curricula that cross traditional departmental lines. At the graduate level, models are needed for facilitating interdisciplinary research through graduate students shared between Computer Science and domain departments.

There is also a need for continuing education for current researchers in environmental sciences. Support for a variety of workshops, distance-learning programs, and related resources to reach current academic and governmental researchers and managers can address this need. In particular, we note that most departments and agencies will be unable to support full-time experts versed in the full breadth of informatics techniques. Thus, there is a need for expert centers offering "consulting-style" advice to projects in managing data, setting up data-access Web pages, integrating data from multiple sources, etc.

Fostering Interdisciplinary Informatics Research

The sections above list many research and development steps that are central to continuing progress in environmental informatics. Critical to all these efforts is a concerted cooperation between

computing/information sciences and the domain sciences that use geospatial data. Targeted funding from NSF to support these projects has gone far to foster these partnerships (e.g. BDI and ITR). Further funding support for data integration activities is required for them to continue, but there are also social/institutional barriers that need to be addressed. Interdisciplinary IT work must be professionally rewarded. Tenure decisions, job descriptions, etc. need to recognize the value of this work. Joint faculty and interdisciplinary education programs can help cross traditional departmental boundaries.

Most critical, however, is an equivalent to journal publication for IT work. Peer-reviewed articles are the coin of the realm in academia. They are the measuring stick through which applicants are hired, tenure decisions made, and salaries negotiated. But the development of information systems, data resources, and software tools does not lead to journal publications. There needs to be a mechanism for attaching peer-review status to the actual data product or tools that are produced; scientific societies can take a leadership role in creating a new process for community evaluation of data resources and tools.

Recommendations

Data Recommendations

- Create a data "code of ethics" to cover expectations and timelines for data sharing, methods for crediting intermediate data resources, etc.
- Promote the identification and creation of base data sets for widely-used variables. This includes targeted sampling to fill gaps in data as well as analytical efforts to gather and integrate point data.
- Target funding to develop technologies beyond *in-situ,* human-mediated point sampling, particularly species- and gene-level biological sampling.

Standards Recommendations

- Standardized expression of point data. Space and time are unifying factors that can serve to integrate a large and heterogeneous universe of data that is evolving, if properly applied. There needs to be a simple, standard way to represent x, y, z, and t location *with accuracy and precision estimates* that can be easily implemented in any data set with spatial-temporal components, along with libraries of names and attributes of the entities being temporally and geo- referenced. An example for species data is the [Species Analyst].
- Endorsement of self-describing data formats. There is currently no "common denominator" data format or generally accepted standard. Until that time, the use of self-describing data formats such as [NetCDF] is strongly encouraged to ensure that the information necessary for extracting and understanding the data is always preserved.
- Creation of tiered metadata standards.
- Development of tools and clearinghouses based on metadata standards.

R&D Recommendations

Prioritize development of key software and tools:

- Automated feature extraction and change detection. For very large data sets such as satellite remote sensing, the entire data set cannot be evaluated by a person. Tools are needed to identify and flag "interesting" features to be examined by a researcher.
- Data mining and time-series data analysis tools. Ideally, geospatial and temporal tools need to be integrated for 4-D analysis of data.
- Estimating, visualizing, and appropriately handling uncertainties in values.
- Automated or semi-automated raster/vector data conversion.
- Creating coverages effectively from point data.
- Visualization of high-dimensionality data. ODBC is not sufficient - there is a need for virtual database tools.
- Interoperability functions, particularly for moving between off-the-shelf products; for linking geospatial data with model/simulation output effectively; and for integrating the idiosyncratic, individually-built tools that exist.

Support OpenGIS development:

- Create online workbenches and software for common geospatial operations that are designed for quick learning and ease of use for unsophisticated users.

Infrastructure Recommendations

- Continue growth of computing infrastructure.
- Reevaluate FCC regulations to facilitate scientific use of bandwidth.
- Create metadata catalogs and clearinghouses for data access.
- Define a framework for distributing portions of the national data centers to allow groups interested in a particular type of data to be its caretakers, with long-term, low-level funding provided as long as performance standards are met.
- Develop an initiative on knowledge representation in geospatial environmental data.

Education Recommendations

- Expand interdisciplinary courses, majors, curricula, and workshops for teaching Information Technology applications within the environmental sciences at the undergraduate, graduate, and continuing-education level.
- Create expert-centers to provide data management and analysis advice to the environmental research community.

Interdisciplinary Recommendations

- Create an equivalent to the peer-reviewed publication to foster recognition for data resources and tools.

DATA TECHNOLOGIES

Distributed Data Access and Retrieval

Jim Beach, Facilitator
University of Kansas

Bertram Ludaescher, Reporter
San Diego Supercomputer Center
University of California, San Diego

Initial Questions

- How can information technologies be better used to facilitate distributed access and retrieval of data acquired via environmental networks?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

Enabling Internet Knowledge Discovery: Beyond Keyword Search and Retrieval

Anyone who has used the Internet for knowledge discovery in environmental biology knows what a bountiful information morass it is. Internet search engines in tenths of a second retrieve daunting numbers of hits from keyword-based queries. In January 2002, a search using Google (www.google.com) for "pacific salmon" generated 325,000 hits, "water chemistry" produced 1,030,000; "fish reproduction" yielded 387,000 linked pages. The overwhelming size of these result sets is only matched by the heterogeneity of linked documents they point to. They include all classes of documents known to man - narratives, research studies, historical essays, data, maps, pictures, sounds, resumes, textbooks, commercial products, as well as regulatory, policy, and educational documents. At the other extreme, a Google search on "pacific salmon AND water chemistry AND fish reproduction" returns exactly four hits: 1) an EPA report to the US Congress, 2) a fisheries management plan for Lake Superior, 3) an encyclopedia of ocean sciences, and 4) a perspectives article on freshwater ecosystems from *Ecological Monographs*.

Now imagine a fisheries biologist with a need to identify the relationship between water quality and spawning rates for pacific salmon. To determine this, she would need to locate formal research investigations that have looked at the impact of physical, chemical, and temperature characteristics of rivers on the physiological and reproductive behavior on any salmonid or other related fish species. She

would need to know what variables were under study, be able to review the results, and probably inspect the field data, in order to assess their relevance and suitability to address the research issue at hand.

Clearly, keyword or full-text searching on the Internet is of limited value for enabling such environmental research. On the one hand our biologist has the untenable option of browsing through millions of linked documents; on the other, she could take the traditional approach of using the four perspective/summary documents as starting points into the research literature. The research objectives are manageable with skilled library research and a three-month review of the literature, but manually assembling research knowledge in this way is a slow and costly process. Although searchable publication abstracts and indexes provide shortcuts to the relevant literature, discovering other species models, from *in situ* or artificial *in vitro* studies of water quality and reproduction, is a hit-or-miss proposition, heavily dependent on keyword indexing. Finding and managing her findings with photocopies of research publications of course provides no support for utilizing pre-existing data sets for re-analysis, extrapolation, or the development of predictive models.

Web interface for LIFEMAPPER *(beta.lifemapper.org) a NSF KDI-funded project which uses the Species Analyst distributed search and retrieval network to obtain biological museum specimen data records that it then utilizes in a distributed SETI@ Home-like screensaver architecture to parallelize the computation of species distribution models based on the museum specimen data. Those models are then archived and visualized on the Lifemapper server.*

The emerging "Semantic Web" aims at changing all of that and the way we do science by transforming the process of networked knowledge discovery and retrieval [Berners-Lee, 2001]. Knowledge representation and linking technologies now being developed and deployed in the sciences aim to tame the Internet from an uncontrolled firehose spewing links to hundreds of thousands of documents in milliseconds in response to a simple query, into a rich distributed corpus of contextualized research information, linked by a deep semantic framework with analysis engines. This matrix of semantic relationships will enhance integration and analysis capabilities well beyond today's keyword and full-text search and retrieval facilities to make the Internet a dynamic workbench for ad hoc knowledge discovery and generation. The conceptual mapping of environmental data, information, and knowledge will enable us to expose the deeper foundation of structure and process in natural systems.

Although the infrastructure for the Semantic Web will be standards and protocols that have just recently become the objects of attention (see below), the content and knowledge linking of the Semantic Web will evolve slowly and likely in response to conceptually localized efforts delimited by funding or disciplinary scope.

In addition to the intellectual contributions of the designers and builders, how do we build something that we know has a very high probability of being used? How do we identify and focus on long-term priorities, with our feet in the shifting sands of technology, and continually implement more efficient systems with next year's technology? What is the minimal payoff that should be expected and measured with NSF funding of infrastructure projects?

Challenges in Distributed Access and Retrieval

Data from environmental networks is being collected, transported, stored, analyzed, and disseminated in a highly distributed fashion. Environmental networks can provide the Data Grid under development across the nation [Foster, 1998] with different kinds of environmental sensing capabilities, often combined with real time or near real time accessibility [HPWREN, ROADNet]. Distributed data networks may reflect cached data as well as sensor data and museum data.

One challenge is that from the field, where environmental sensors gather data, to the intermediaries and end users of information, there is an enormous variety of data transport and access demands, data uses, and data users. This absence of a common data and user profile in the environmental sciences community prevents a "one size fits all" approach to distributed data access and retrieval for environmental networks. Seamless distributed data access and interoperability is a challenging goal in the presence of significant heterogeneity of data, infrastructure, and user requirements.

The profile of data usage varies along different dimensions: Technically, data traveling from field sensors through intermediate nodes and different "aggregate states" (e.g., raw data can be transformed, analyzed, annotated with metadata, cleaned, aggregated, and finally stored in a curated digital library or archive) may encounter different bandwidth bottlenecks along the way before it reaches its destination, say a client application on a scientist's laptop. Ideally, dealing with different bandwidths should not be the burden of the end user or even the data provider but should be handled by adaptive software that balances users' needs and available network bandwidth. Parameters and models need to be developed that can describe user demands and usage scenarios. These would address questions and issues such as following:

- How "fresh" and recent should data be?
- How much precision and accuracy makes sense? What sampling rates are adequate?
- How much persistency is needed? For example, does a ring buffer holding one week of data provide enough persistence to guarantee that all relevant analyses and archival requirements are met before data is overwritten?
- How is data quality described, measured, and guaranteed? In particular, if data is automatically published from the field to the Web, how is quality assurance and quality control maintained?
- What access methods will best support users' requirements? Are http and ftp sufficient, or are database languages and APIs (e.g., SQL, JDBC) needed? How about digital library protocols and methods for data access in archived collections?
- How can data from different sources be combined and integrated? When such value-added mediation services are provided, how can the origin and provenance of data be tracked in order to give credit to the data providers?

Below we outline some promising directions toward facilitating distributed access, seamless retrieval, and interoperability of information from environmental networks. Detailed usage models and scenarios describing different types of users (scientists, policy makers, students, etc.) and their requirements will be helpful to determine specific instantiations of the frameworks described.

Information Technology for Data Exchange and Information Integration

Notwithstanding the specific needs of individual communities, the broad goals of seamless distributed access and retrieval from environmental networks are in fact common to many disciplines: Information systems have to be made interoperable such that heterogeneities in platforms, physical location and naming of resources, data formats and data models, supported programming interfaces and query languages, etc., all become transparent to the user. The need for such an interoperable *Grid infrastructure* [Foster, 1998] that can enable new science based on distributed computing, data sharing, and information integration is driving many national-scale projects in several disciplines, e.g. [NEON, NCEAS,

LTER, KNB, NPACI, GBIF], as well as international efforts e.g. [GGF, 2001].

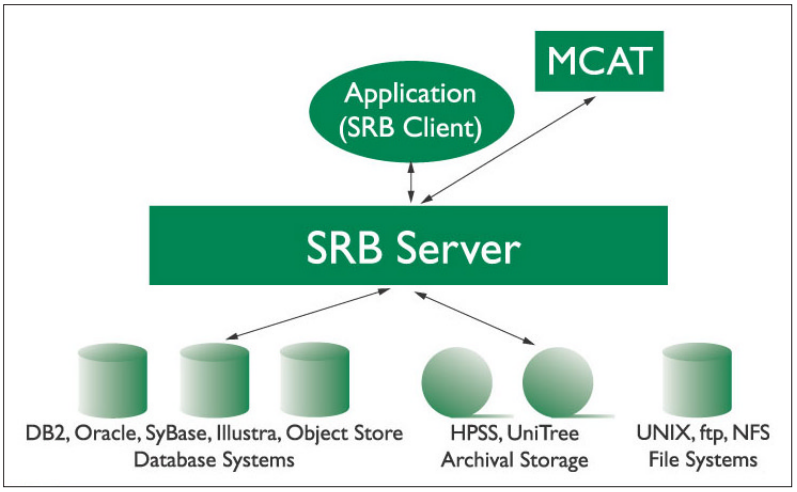
The services provided by such an infrastructure can be roughly classified as: (1) system and data interoperability issues, addressed by *Data Grid Services,* and (2) semantic interoperability and information integration issues, addressed by the *Semantic Mediation Services* of a future *"Knowledge Grid."*

Data Grid Services

System aspects of interoperability include distributed storage across heterogeneous devices, data transport, access protocols, and distributed computing services. A prominent grid tool that addresses many system aspects is SDSC's Storage Resource Broker [SRB]. Instead of installing your own ftp site or Web server, and worrying about different device drivers, access control, location of distributed storage systems, disks, etc. a user simply becomes a member of a data grid by registering as an SRB user and installing a lightweight client or any standard Web browser. To the end-user, the SRB appears to be a *virtual drive* (the so-called "SRB space") into which environmental data can be put and from which other users (limited to authorized ones, if appropriate) can obtain data. The SRB makes transparent to the user such system aspects as:

- *How to access a specific storage device* (disk, tape, database, etc.). The SRB has an extensive and extensible set of "drivers" (aka "cartridges," "blades," "plug-ins") for storage devices.
- *Where the data set is located*. A user does not have to know or be concerned with the physical location of data sets. The SRB relieves her of having to deal with these details by managing all this information through a metadata catalog (MCAT).

In addition to transparent file access across heterogeneous devices and physical distribution, the SRB also provides solutions to other interoperability problems. For example, in addition to using the SRB as a sophisticated virtual drive (with access control, replica management, support for very large data sets, and other grid capabilities), it can also be used as a *relational data mediator.* By putting data into an SRB-accessible relational database, an *attribute-based* query and mediation mechanism becomes available to the user. This means that a user does not have to be concerned with the detailed structure of relational tables. Instead, the user can pick a set of attributes and search conditions on those attributes, after which the SRB will generate plans that span multiple tables (that may even reside in different parts of the world) and retrieve the desired data.



Client Service Middleware

The Storage Resource Broker (SRB) was developed at the San Diego Supercomputer Center (SDSC) and the National Partnership for Advanced Computational Infrastructure (NPACI) as client-server middleware to provide a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. SRB, in conjunction with the Metadata Catalog (MCAT), provides a way to access data sets and resources based on their attributes rather than their names or physical locations.

Semantic Mediation Services

There is a recent trend toward "deeper" interoperability and integration of information beyond simple distributed access of data files. First, data sets need to be "wrapped" into a suitable *metadata envelope,* in order to facilitate deeper information integration beyond the data level. The metadata may provide all kinds of descriptive information about the data, including origin and provenance, data quality, accuracy, and last but not least, context information, e.g., the terminology or taxonomy used and a spec-

ification of the semantic context within a given domain ontology. Syntactically, metadata should be encoded in XML, the de facto standard for information exchange over the Web. XML is a flexible data format that can encode both regular data (from relational or object-oriented databases) as well as semi-structured data (e.g., from system-generated Web pages). By using XML, a large number of tools for storing, querying, and manipulating XML-encoded information then become readily available. W3C standards related to applications of XML such as SOAP (for distributed object access), XML Schema (for modeling XML data), XQuery (for querying XML databases), and XSLT (for transforming XML output into a presentable form) provide a generic interoperability infrastructure based on open standards and tools, and are also employed in the development of grid services. Persistency and archival requirements can also benefit from an XML-based approach, as XML provides largely infrastructure-independent, self-describing means to represent information.

Agreed-upon metadata standards for environmental data are key to the reuse, interoperability, and integration of information. Meaningful links between disparate data are established and become "visible" and manageable to mediation services by using a set of *predefined attributes.* For complex scientific domains that require "semantically deep" dynamic querying of sources from different domains, new approaches such as *Model-Based Mediation* seem promising: In such a knowledge-based approach, the sets of attributes of different metadata standards do not stand in isolation but are mutually related to one another. Relationships between attributes and concepts across standards can be captured by a formalization of those relationships, for example, using logic rules directly [Ludaescher, 2001], or indirectly via the emerging standards developed in the context of the *Semantic Web* [Berners-Lee, 2001] effort, which aims at providing a generic infrastructure for semantic interoperability. The use of widespread, open standards and tools is also likely to positively influence the buy-in of the community, which is essential in order to create the desired high quality data and information content.

Recommendations

- Quality Control and Quality Assurance should be integrated into all aspects of data management, capture, transformation, integration and analysis.
- Maintain emphasis on funding biological informatics, especially collaborations between information technology researchers and biology research laboratories.
- Pay attention to usability and user needs. To enable new research with new kinds of people, the services and applications must be usable, and the NSF should pay close attention to mechanisms that set up feedback loops with the community the architectures serve. Establish a framework for evaluating the usability, use, and impact of evolving architectures and tools.
- Sustainability. Encourage enough labs to this kind of work to reach critical mass. Ensure that this scales socially and professionally. Establish peer review and formal mechanisms for collaboration. Support outreach activities.
- Encourage collaboration with broader, larger activities such as the NSDL.
- To be broadly interdisciplinary with other ESS disciplines, Biology needs to exert its research strengths in addition to a geoinformatics view of the world.
- Semantic Web is an interesting development; research proposals need to track it and build upon it.
- Specifications of user requirements need to be developed in detail and to inform and guide the process every step of the way.