# Proceedings of the

# NSF Invitational Workshop on Distributed Information, Computation, and Process Management for Scientific and Engineering Environments (DICPM)*

**Herndon, Virginia**
**May 15-16, 1998**

**Nicholas M. Patrikalakis, Editor**
**Massachusetts Institute of Technology**

**URL: http://deslab.mit.edu/DesignLab/dicpm/**
**E-mail: DICPM@deslab.mit.edu**

January 14, 1999



*The workshop was supported by the National Science Foundation, the Information and Data Management Program and the Robotics and Human Augmentation Program of the Information and Intelligent Systems Division, the International Coordination Program of the Advanced Networking Infrastructure and Research Division, the Operating Systems and Compilers Program of the Advanced Computational Infrastructure and Research Division, the Experimental Software Systems Program of the Experimental and Integrative Activities Division, the Control, Mechanics, and Materials Program of the Civil and Mechanical Systems Division, the Integrative Systems Program of the Electrical and Communications Systems Division, the Atmospheric Sciences Climate Dynamics Program, the Office of Multidisciplinary Activities, and the Arctic Research and Policy Program, under grant IIS-9812601. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the workshop participants, and do not necessarily reflect the views of the National Science Foundation.

Additional copies of this report can be ordered at a cost of $3 per copy. Requests should be made by mail to:

DICPM Workshop Report
Massachusetts Institute of Technology, Room 5-426
77 Massachusetts Avenue
Cambridge, MA 02139-4307, USA

or by E-mail to: *DICPM@deslab.mit.edu*.

# Contents

# 1 Executive Summary

An Invitational Workshop on Distributed Information, Computation, and Process Management for Scientific and Engineering Environments (DICPM) was held at the Hyatt Dulles in Herndon, Virginia, USA, on May 15-16, 1998. The workshop brought together domain specialists from engineering and the ocean, atmospheric, and space sciences involved in the development and use of simulations of complex systems, and computer scientists working on distributed repositories, visualization, and resource management. The objective was to formulate directions for further research efforts to facilitate effective collaboration and to help increase access to information and sharing of results and tools useful in large-scale, distributed, multidisciplinary scientific and engineering environments.

Funding for the workshop was provided by the National Science Foundation (NSF). The 51 participants were drawn from academia (35), industry (4), and government (12), including program managers from NSF, National Aeronautics and Space Administration, National Institute of Standards and Technology, and National Oceanic and Atmospheric Administration. The final detailed report of the workshop will be available after September 30, 1998, at the workshop web site <*http://deslab.mit.edu/DesignLab/dicpm/*>.

## 1.1 Motivation

The simulation of complex systems encompasses many domains, including physical systems, such as the oceans and the atmosphere, with a large variety of interacting processes and dynamic geophysical, chemical, and biological phenomena at disparate spatial and temporal scales. Additionally, these simulations may include sophisticated man-made systems encountered in the design and manufacturing of land, air, space, and ocean vehicles. Research advances in the areas of complex systems generate new requirements for computational environments and infrastructure.

## 1.2 Workshop Themes

For motivational background, a series of formal presentations were given to a plenary session on topics such as distributed and collaborative systems, multidisciplinary scientific simulation, metadata for data and software, distributed workflow and process management, scientific and engineering archives and repositories, and engineering standardization efforts. Following these presentations, small focused breakout groups met to discuss specific issues and suggest avenues for future research efforts. This was followed by summary presentations by the chairs of the workgroups to a final plenary session, which was followed by an overall discussion.

Although many diverse views were aired, a consensus did emerge that the major problems inhibiting the widespread exploitation of multidisciplinary collaboration in scientific and engineering analysis and simulation were threefold:

1. Insufficient support for computational infrastructure to make accessible information for interpretation and sharing of results and tools;
2. Institutional barriers to multidisciplinary cooperation (e.g., educational focus, publication policy, promotion criteria, funding, etc.); and
3. Communication barriers stemming from narrow specialization of technical expertise and experience (e.g., domain science vs. computer science, theoretical science vs. applied science, industry vs. academia, etc.).

## 1.3 Conclusions and Recommendations

Towards alleviating these barriers to effective multidisciplinary activities, the workshop led to the following proposals:

1. The allocation of support and incentives for multidisciplinary projects by the appropriate facilitators in the research community, industry, and government, which will foster cooperation between computer and domain scientists, and encourage team-based approaches to multidisciplinary problems;
2. The establishment of a national (and possibly, international) digital library for the physical, biological, and social sciences and engineering, which will help disseminate research knowledge and resources beyond conventional domain boundaries; and
3. The establishment of a global distributed information registry and repository (a "virtual scientific marketplace" [7]) for experts, tools, and procedures, which will facilitate multidisciplinary collaboration.

## 2 Sponsorship and Acknowledgments

The DICPM Workshop was generously supported by the National Science Foundation under grant IIS-9812601. We wish to thank the following NSF Divisions, Programs, and Program Managers:

- Directorate for Computer and Information Science and Engineering (CISE)
  - Division of Information and Intelligent Systems (IIS)
    - Information and Data Management (Dr. Maria Zemankova)
    - Robotics and Human Augmentation (Dr. Howard Moraff)
  - Division of Advanced Networking Infrastructure and Research (ANIR)
    - International Coordination (Dr. Steve N. Goldstein)
  - Division of Computer-Communications Research Division (C-CR)
    - Operating Systems and Compilers (Dr. Mukesh Singhal)
  - Division of Experimental and Integrative Activities (EIA)
    - Experimental Software Systems (Dr. William Agresti)
- Directorate for Engineering (ENG)
  - Division of Civil and Mechanical Systems (CMS)
    - Control, Mechanics and Materials (Dr. Ken Chong)
  - Division of Electrical and Communications Systems (ECS)
    - Integrative Systems (Dr. Art Sanderson)
- Directorate for Geosciences (GEO)
  - Atmospheric Sciences Program (ATM)
    - Climate Dynamics (Dr. Jay Fein)
- Directorate for Mathematical and Physical Sciences (MPS)
  - Office of Multidisciplinary Activities (OMA) (Dr. Henry N. Blount)
- Office of Polar Programs (OPP)
  - Arctic Research and Policy (Dr. Charles Myers)

The DICPM Workshop would not have taken place without Dr. Maria Zemankova's leadership. She recognized the importance of the Workshop theme and worked tirelessly to help orchestrate the funding for the workshop from NSF, and provided many helpful comments on this report.

All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the workshop participants, and do not necessarily reflect the views of the National Science Foundation.

We would like to thank the Program Committee members:

- Prof. Paul J. Fortier, University of Massachusetts Dartmouth
- Prof. Christos N. Nikolaou, University of Crete and ICS/FORTH, Greece
- Prof. Allan R. Robinson, Harvard University
- Prof. Jarek Rossignac, Georgia Institute of Technology
- Prof. John R. Williams, MIT

the workgroup chairs and recorders:

- Prof. Peter Buneman, University of Pennsylvania
- Dr. Pierre Elisseeff, MIT
- Prof. Yannis Ioannidis, University of Athens, Greece
- Prof. Miron Livny, University of Wisconsin
- Prof. Christos N. Nikolaou, University of Crete and ICS/FORTH, Greece
- Prof. Jarek Rossignac, Georgia Institute of Technology
- Dr. Jakka Sairamesh, IBM Research
- Dr. Knut Streitlien, MIT
- Prof. Alvaro Vinacua, Polytechnic University of Catalonia, Spain

and the workshop staff:

- Mr. Stephen L. Abrams, MIT
- Ms. Kristin M. Gunst, MIT

for their efforts in organizing and running the workshop. We would also like to thank the invited speakers for their enlightening motivational presentations:

- Dr. Joseph Bordogna, NSF
- Prof. Peter Buneman, University of Pennsylvania
- Dr. William J. Campbell, NASA
- Dr. John C. Cherniavsky, NSF
- Prof. Allan R. Robinson, Harvard University
- Prof. Jarek Rossignac, Georgia Institute of Technology
- Prof. Amit P. Sheth, University of Georgia
- Dr. George Strawn, NSF
- Dr. Stuart Weibel, Online Computer Library Center
- Dr. Roy D. Williams, California Institute of Technology
- Dr. Peter R. Wilson, Boeing Commercial Airplane Group

Finally, we would like to thank the following scientists for their help in preparing the final workshop report: Stephen L. Abrams, Paul J. Fortier, Yannis Ioannidis, Christos N. Nikolaou, Allan R. Robinson, Jarek Rossignac, and Alvaro Vinacua, as well as all of the workshop participants for their thoughtful input.

# 3 Introduction

By focusing on topics such as:

- Application of digital library technology to disseminate the results of scientific simulations
- Metadata models and search/retrieval mechanisms for scientific software and data
- Networking issues for transparent access and sharing of data, including automatic generalized data translation, and remote software invocation
- Emerging standards for distributed simulations
- Management and visualization of complex data sets

the workshop produced this advisory document outlining open problems and proposed avenues of research in these areas. The application of further research efforts along these lines will help to increase the availability, effectiveness, and utilization of large-scale, cross-disciplinary, distributed scientific systems.

The simulation of complex systems encompasses many domains, including physical systems, such as the oceans and the atmosphere, with a large variety of interacting processes and dynamic geophysical, chemical, and biological phenomena at disparate spatial and temporal scales. Additionally, these simulations may include sophisticated man-made systems encountered in the design and manufacturing of land, air, space, and ocean vehicles. Some of the related computer science issues include:

- Cross-domain query translation
- Disciplinary and interdisciplinary ontologies and information modeling
- Distributed resource recovery
- Large-scale computational infrastructure providing the necessary services and performance

# 4 Major Themes Discussed

## 4.1 Research Environment

To illustrate the importance of the workshop theme, we present two scenaria showing how advanced distributed information systems can lead to a new generation of systems and capabilities in the fields of ocean science and collaborative engineering. Other scenaria could just as well have been presented in the areas of meteorology, space weather, seismic networks, distributed medical applications, or biology.

### 4.1.1 A Scenario for Ocean Science

Ocean science and technology is at the threshold of a new era of important progress due to the recent feasibility of fundamental interdisciplinary research and the potential for application of research results. The fundamental research areas today include bio-geo-chemical cycles, ecosystem dynamics, climate and global change, and littoral-coastal-deep sea interactions. The development and application of multiscale, multidisciplinary Oean Observing and Prediction Systems (OOPS) will substantially accelerate progress and broaden the scope of such research. The OOPS concept involves a mix of advanced state-of-the-art interdisciplinary sensor, measurement, modeling, and estimation methodologies, as well as major systems engineering and computer science developments [20]. The development of OOPS utilizing distributed oceanographic resources will significantly increase our national capability to manage and operate in coastal waters for civilian and defense purposes, including: pollution control (e.g., outfalls, spills, harmful algal blooms); public safety and transportation; resource exploitation and management (e.g., fisheries, oil, minerals); and maritime and naval operations. Moreover, the general design principles of a distributed information system architecture will be applicable to analogous systems for field estimation in other areas of earth science research and management, e.g., meteorology, air–sea interactions, climate dynamics, seismology, and to other fields in which distributed systems are essential.

Real–time operations begin with a definition of the objectives and a review of the current state of knowledge. Decisions are taken on the processes and scales to resolve, and on which multidisciplinary aspects (physical, chemical, ecosystem, etc.) to include. These decisions are driven largely by the assembled pre-operation state of knowledge. The physical characteristics (topography, water masses, etc.) and dominant processes (currents, winds, tides, and ecosystem) need to be understood. Available data is gathered and previous modeling results are examined. A relevant OOPS can be defined and assembled with the appropriate remote and *in situ* sensors and models. The OOPS is tested and refined with Observational System Simulation Experiments (OSSEs). The models are initialized with the historical synoptic and climatological data and acted upon by the present forcings. Initial field estimates and sampling strategies are made. During the real–time operation, newly acquired data is assimilated into the models. Field estimates and sampling strategies are adapted

A distributed information system for ocean processes (DISOP) can improve real–time operations by: (1) transparent transmission and ingestion of data; (2) use of distributed computational resources; (3) improved data reduction and interpretation; and (4) increased potential for autonomous operations. Consider a hypothetical scenario in which a multidisciplinary survey is carried out in a region which includes the Middle Atlantic Bight to the Gulf Stream Ring and Meander Region. Using the DISOP system, the historical physical/ecosystem data is transparently gathered from multiple sites. An OOPS is defined with appropriate data gathering platforms and models, then refined with OSSEs. Pre–operational field estimates are constructed using distributed computational resources. Data gathering begins with the large-scale context coming from remote satellite sensing and aircraft flights. Synoptic surveys are performed with ships, Autonomous Ocean Sampling Networks (AOSNs) [6], and acoustic networks. Data from these varied platforms is gathered via the DISOP system and readily assimilated into the models. The information contained in the simulations is reduced, distributed, and configured to the various specific needs of scientists and decision makers. Automated processes collect the results and error assessments and optimally adapt sampling plans. Autonomous underwater vehicles (AUVs) download their new missions automatically.

Present and future ocean observing and prediction systems require tremendous computing and network resources. To implement the system functionalities described in this example will require computing capabilities that are at least several orders of magnitude greater than current state-of-the-art, including: identification and gathering of multidisciplinary data distributed in the network; preparation and fine tuning of the OOPS and its components via extensive simulated experiments; ingression/assimilation/integration/egression of data during operations; automatic work flow management; and distribution of the results. The experience gained with the use of a prototype information system in realistic simulated and real–time set-ups will stimulate and expand the conceptual framework for knowledge networking and the scope of new computational challenges for ocean science.

### 4.1.2 A Scenario for Collaborative Engineering

In today's global economy, large-scale engineering and construction projects commonly involve geographically distributed design teams, whose members must communicate with each other in synchronous and asynchronous ways, accessing highly complex design databases and catalogs, and interacting with customers, stylists, suppliers, and manufacturing experts. Each of these groups of people use different tools to manipulate, evaluate, and annotate product models. Current generation computer-based tools for collaboration and data sharing impose considerable and undue limitations on the productivity of these teams, and possibly on the quality of their products.

For example, the design and manufacture of a new automobile involves a conceptual design phase, which usually includes external body styling, the engineering of the body and its internal structure, the engine, and different control and mechanical subsystems, lights, and interior. These design activities are usually handled by different groups in the organization, all of whom need to coordinate their design decisions and to involve external contractors or manufacturing departments.

The design cycle of an automobile relies upon a very complex and repetitive flow of information between the various design groups. For instance, usability tests may lead the "interior design" team to lower the windshield, which may conflict with styling considerations, impact the aerodynamics of the automobile, or impose new constraints on the engine layout. Resolving conflicts such as these typically requires a series of cycles through which the different design and manufacturing teams progressively refine their understanding of each other's design constraints in search of an overall optimal compromise. These discussions are full of references to the product features (e.g., shape and position of features, manufacturing and maintenance processes, etc.) and are thus difficult to perform via electronic mail, voice mail, or telephone conversations. Furthermore, these negotiations and collaborative decisions must be fully integrated within the established design process and must be properly captured and documented for future reference. Clearly, geographically distributed teams must rely upon communication and collaborative decision making tools that fully integrate access to the 3D product model with mainstream documentation and both synchronous and asynchronous communication tools. Members of the design team must be able to quickly inspect the results of the latest engineering changes, compare them to previous alternatives, discuss them with others, and capture these discussions in text, "red-line" mark-up, or voice annotations of the 3D model.

Such capabilities depend heavily on geometric compression technologies for accessing remotely located product models, whose complexity grows much faster than the communication bandwidth. They also require graphic acceleration techniques that support the real-time rendering of complex models on personal computers. To be effective, they must also offer easy-to-use graphic interfaces for manipulating the view, the model, and the annotations. Finally, they must be seamlessly integrated with personal productivity and communication tools. For example, a designer may receive an electronic mail message that opens a 3D viewer, quickly download the appropriate subassembly model, and proceed to describe an engineering problem using references and red-lining over the model. The designer will initiate a teleconference and discuss the problem with others while jointly manipulating and pointing to a shared 3D model. During their discussions, they may need to browse component catalogs or databases of prior designs, or to integrate new components directly into their new design. Finally, they will want to document the result of this interaction and append an engineering change request to the master product model database.

In summary, we must develop effective tools that will give engineers the ability to access, visualize, and annotate highly complex models, and the ability to generate, document, find, and customize reusable design components. These tools must be interactive, easy-to-use, and well integrated in the overall work flow. The development of such tools for a given domain requires a combination of:

1. A thorough understanding of domain practices and needs;
2. A firm grasp of the principles of remote collaboration and visual communication; and
3. Advances in computing technologies in a variety of computer science areas, including data compression, visualization, human-computer interaction, remote collaboration, digital libraries, and design automation.
4. Data exchange standards and efficient compression and memory utilization techniques to facilitate the detailed representation of complex objects for manufacturing processes.

## 4.2 Facilitating Multidisciplinary Collaboration

In order to develop technology and tools that will truly support the management of distributed information, computation, and processes for scientific and engineering environments, three different kinds of people should come together and play distinct, yet mutually supporting, roles.

First are the domain scientists and engineers, e.g., climatologists, biologists, oceanographers, electrical, mechanical, civil, aeronautical engineers, etc. Their needs are the driving force behind any efforts that result from the DICPM workshop. The whole purpose is to provide technology so that larger-scale and more sophisticated scientific and engineering problems may be addressed by the members of this group.

Second, the computer scientists, whose role is to develop new computational environments, motivated by the needs of the domain scientists and engineers. Of course, even if the computer science research is driven by the scientists' and engineers' needs, the goal is also to advance computer science as a discipline in its own right.

Finally come the facilitators, i.e., the government funding agencies, professional societies, standards organizations, and industry foundations that support research and development activities. They typically provide the means and organizational infrastructure for the scientists and engineers to achieve their goals.

The following technologies seem to be relatively ubiquitous and of high importance (given in arbitrary order):

- Easier Identification, Access, and Exchange of Data and Software: International repositories and banks should be created that would store important datasets and software tools, increasing the reuse of such valuable resources. Members of the community should be able to deposit the results of their efforts in such banks, a validation mechanism should check to make sure that the deposited data and software meet certain standards, and the material should then become available to the community. Identifying "general" validation mechanisms and search strategies for data and especially software are key technical challenges for this to become possible.

- Easier Coupling of Models: Modern science requires the integration of many different models that solve individual problems and are geographically distributed at different sites, so that more complex problems may be solved. Coupling "arbitrary" software is a significant technical challenge: for these models to work together mechanisms must be developed for specifying their interaction patterns.

- Easier Management and Monitoring of Processes: In conducting large-scale (or even small-scale) experiments, one must be able to specify, monitor, and generally manage the flow of control and data during the experiment, including all interactions with humans, models, and physical instruments. This raises several technical issues, some of which are shared with typical business-oriented workflow management and others that are unique to scientific experimentation environments.

- Increased Collaboration: The phenomenon of multiple people needing to work together on a problem is more and more common in scientific research. Improved technology is needed for facilitating such collaboration, which may be of one of two forms: static, which mostly requires off-line exchange of arbitrary material, i.e., software, data, metadata, publications, etc.; and dynamic, which requires the ability for multiple actors to manipulate the same "object" simultaneously. Additional tools are also needed to manage the collaborative process.

- Increased Computing Power: Despite the tremendous advances in processor speed of the last few years, scientists' hunger for more computing power never ceases. The complexity of forthcoming experimental studies will require many more cycles than are currently available to scientists. Faster processors and larger parallel systems seem necessary to address these increased demands.

- New Algorithms: The expanded and multidisciplinary nature of experimental studies requires modeling of phenomena, systems, and behaviors that are significantly more complex than in the past. For several of them, current algorithms are unsatisfactory and new, efficient ones must be developed. In addition to these problem-specific algorithmic needs, the geographic distribution of scientific teams and/or computational facilities gives rise to a general demand for the development of algorithms that are naturally distributed, i.e., they are specifically suited to use distributed computing facilities.

## 4.3 Research Themes

### 4.3.1 Integration, Heterogeneity, and Reusability

As was made apparent in the ocean science and collaborative engineering scenaria presented above, future advances in the scope, quality, and productivity of scientific and engineering activities will depend upon the close integration and reuse of cross-domain data, procedures, and expertise. Ubiquitous network connectivity provides some advantages towards achieving this goal, for example, direct access to geographically dispersed co-workers and data sources. However, this ease of communication has also encouraged the proliferation of widely distributed computational resources without any universal mechanism for resource discovery. Furthermore, the extreme heterogeneity of these dispersed resources poses an additional problem in terms of the compatibility and translation between processors, operating systems, implementation languages, data formats, etc.

To avoid unnecessary duplication while at the same time allowing the widest possible dissemination and use of scientific and engineering resources (both data *and* software), these resources should be collected in distributed public repositories. There are two possible models for designing and implementing information repositories: product-oriented and service-oriented. A product-oriented repository contains actual resources (e.g., data files, software in source code or executable format, etc.) that can be downloaded to a user's computer, where they can be used or incorporated into complicated process flows [8, 2]. Under the service-oriented model, resources do not move between the repository and the user; rather, the user sends a *request* to the repository, which finds the appropriate resource to service the request and then returns the result to the user [4]. While the product-oriented model may be more appropriate for legacy resources, it does require a higher level of computer systems expertise to use (e.g., compilation, including modification of machine-dependent characteristics of the software, the system command language, etc.). The service-oriented model helps to insulate the domain scientist or engineer from these details, accepting service requests phrased in the terminology of the problem domain. One important issue that remains to be resolved is the question of long-term maintenance of such a repository: who will manage the repository, for how long will it be maintained, and who will cover the costs?

CORBA (Common Object Resource Broker Architecture) [12, 23] is a standard object-oriented architecture for "middleware," which manages communications between distributed clients and servers transparently, and enables different applications to use standard mechanisms for locating and communicating with each other. Java [15] provides a language in which to develop portable applications that may freely interoperate on all important commercial platforms. Although these technologies are necessary enablers of the distributed collaborative environments that scientists and engineers need, they cannot satisfy all needs. Additional work is necessary in the following research areas:

- Efficient distributed indexing of available resources, through flexible, fault tolerant distributed data structures.
- Brokerage services that will enable service providers (e.g., data store administrators,

modelers, video-conferencing service providers, on-line educators, etc.) to advertise their resources and the conditions under which they may be accessed, by which users can match their needs with the services offered.

- Quality of Service (QoS) enforcement, that will allow service providers to fulfill their Service Level Agreements (SLAs) (e.g., bandwidth and processor time reservations for live video feeds, response time constraints, etc.) with information consumers such as individual researchers, professional societies, or universities.
- Mediators that can translate and possibly decompose higher-level or domain-specific queries into lower-level or legacy queries, and wrappers for data conversion.

### 4.3.2 Metadata and Tools for Reusability

Metadata means "data about data" or "information about information." The descriptive metadata fields can be different in nature. Some fields (such as titles) contain mostly terminology, some contain uninterpreted attributes (e.g., creator name), and others contain more structured information (e.g., numeric values, dates, and format or protocol specifications).

The Dublin Core and the Warwick Framework [10, 17] are a first step at establishing a metadata classification. They define a minimal set of optional fields that can describe an information object. XML [3] is becoming increasingly adopted as a common syntax for expressing structure in data. Moreover, the Resource Description Framework (RDF) [22], a layer on top of XML, provides a common basis for expressing semantics.

In addition to the Dublin Core and XML, which are efforts to provide a universal set of verbs or a universal syntax for expressing metadata, there have been efforts by various scientific communities to define their own metadata standards. For example, the geospatial community has defined the FGDC (Federal Geographic Data Committee) Content Standard for Digital Geospatial Metadata [5], the biology community has defined the NBII (National Biological Information Infrastructure) Biological Metadata Standard [14], and the acoustics community has agreed upon a standardized terminology [1]. These efforts are expected to continue and spread across all areas of scientific endeavor and should be encouraged and supported. Furthermore, metadata standards within each scientific community must also strive to address the problem of providing standardized descriptions of program functionalities. These descriptions will greatly increase legacy code reusability.

Tools need to be developed that facilitate cross-domain scientific work by addressing the problem of terms that have different meanings for different communities. At the very least, cross-domain thesauri should be built and maintained. Eventually, multilingual concerns will also need to be addressed to satisfy the needs of a worldwide scientific community.

### 4.3.3 User Interfaces and Visualization

Interactive visualization already plays an important role in manufacturing, architecture, geosciences, entertainment, training, engineering analysis and simulation, medicine, and science. It promises to revolutionize electronic commerce and many aspects of human-computer interaction. In many of these applications, the data manipulated represents three-dimensional shapes and possibly their behavior (e.g., mechanical assemblies, buildings, human organs, weather patterns, heat waves, etc.). It is often stored using 3D geometric models that may be augmented with behaviors (animation), photometric properties (e.g., colors, surface properties, textures, etc.), or engineering models (e.g., material properties, finite element analysis results, etc.). These models are increasingly being accessed by remotely located users for a variety of engineering or scientific applications. Thus, since these 3D models must support the analysis, decision making, communication, and result dissemination processes inherent to these activities, considerable progress is required on the following three fronts:

1. Access and visualization performance for complex datasets
2. Easy-to-use and effective user interfaces
3. Integration with knowledge discovery (e.g., data mining) and sharing tools

We must provide the support technologies for quickly accessing models via the Internet and for visualizing them on a wide spectrum of computing stations. The number and complexity of these 3D models is growing rapidly due to improved design and model acquisition tools, the wide-spread acceptance of this technology, and the need for higher accuracy. Consequently, we need a new level of tools for compressing the data and for inspecting it in real-time at suitable levels of resolution. Although in some applications, such as manufacturing and architecture, it may be obvious how the data should be displayed, other applications (e.g., medicine, geoscience, business, etc.) require that the raw data be interpreted and converted into a geometric form that illustrates its relevant aspects. This interpretation often requires considerable computational resources and domain expertise.

Because engineers and scientists must concentrate on the data they need to inspect, manipulate, or annotate, and because they cannot waste their time waiting for a graphics response or trying to set up a suitable viewing angle, the user interfaces that support these activities must be easy-to-use and highly effective. Although much progress has been made in virtual reality, immersive head-mounted or CAVE set-ups are often expensive and impractical. A new generation of 3D interface technologies is needed for personal computers, which are becoming increasingly mobile. The mobility of computing platforms raises important research issues in itself.

Although the availability of precise 3D models permits the automation of a variety of analysis and planning tasks, their primary role is generally to help humans gain scientific insight, make design decisions, and disseminate the results of their work in support of collaboration and education activities. Current generation design, documentation, and library tools require significant investments from users who want to make their work accessible to and reusable by others. Therefore, it is imperative to automate some of this documentation and preparation work and to integrate the access and manipulation of 3D models with a variety of communication tools including electronic mail, text editors, web publishing, product data management, shared bulletin boards, and digital libraries. A major difficulty lies in the development of practical tools for describing the data at a syntactic and semantic level and for using these descriptions to automate database access or format conversions.

### 4.3.4 Navigation and Data Exploration

There is a clear need for a new generation of user query interfaces for scientific databases. These interfaces must be designed such that the novice user ("domain scientist") can quickly and intuitively acquire expertise in the use of the query system. To support new query languages, database data models also need to support naturally scientific data formats and structures. The interface should be natural to the domain scientist and may include textual, video, voice, or gestural input.

One particular theme that seemed to arise from many discussions was that queries may be more *navigational* than *set-oriented*, which is what most database scientists envision. Navigation should allow the domain scientist to examine singular data sets as well as to combine data sets in more complex ways, possibly through functional or domain specific interfaces. For example, an oceanographer may be interested in seeing how a dataset containing temperature versus depth data relates to biological activity at various depths. The query engine must be able to "answer" these types of queries as well as conventional queries over a relational or distributed heterogeneous database.

In addition, given the exploratory nature of scientific analysis, it is important for the user interface to facilitate navigation so that the user focuses on the scientific task and not on the database interaction. Important in this regard are the issues of navigation context and data/query fusion. Regarding the former, database systems should maintain the history of a navigation so that easy-to-specify query deltas may be unambiguously interpreted within that context and based on where the user currently is. Regarding the latter, database systems should allow users to be "immersed" into data sets and query/navigate them through actions directly on them instead of through explicitly specifying a query.

Instead of navigational queries or conventional queries, some researchers suggest the need for data publishing instead of data queries. They view the problem of scope as one not easily solved ("I cannot find the information I want, since I don't know where to look"). One solution is to have domain scientists *electronically publish* their data sets, or to establish services at some number of repositories, where data sets can be viewed, retrieved, or distributed to subscribers.

The important issue underlying all of these possibilities is that of interoperability: How can we provide query languages that integrate distributed heterogeneous multimedia information and deliver this information in ways amenable to scientific discovery? It is not sufficient simply to provide data to the user; the system *must* deliver information in a manner from which the domain scientist can easily formulate and execute additional derived queries.

### 4.3.5 Data Mining of Large Datasets

Information is data with semantic associations. It is therefore imperative that stored data has not only *structural* metadata associated with it, but that *semantic* metadata is also associated with the data so as to facilitate domain-specific information acquisition and discovery. For scientific databases this semantic information typically takes the form of added metadata preceding the domain-specific (or even program specific at times) data that is to be used in data mining operations as well as in query processing.

Data mining is an important component of scientific information discovery within large domain-specific data sets. The issue is in the size of the data sets to mine and how to mine information efficiently versus coded sequences as is found in conventional data warehouses used in mining. To better investigate the issues of scale one needs to look first at the formats of the information to be mined. Typically, domain-specific scientific data has associated with it numerous metadata fields that describe pertinent information about the scientific information, for example, who generated the information, how it was generated, what is the confidence level of the information, etc. This meta-information along with conventional metadata should be used to enhance mining. Conventional mining techniques wrapped around information coding, fixed common context, and pattern analysis need to be augmented to allow the natural use of domain data sets.

If data mining is to be useful for domain scientific knowledge discovery, then new means of codifying information or naturally searching and analyzing data in raw format must be developed. For example, the intersection of pattern recognition, data compression, and data mining algorithms must be accomplished. A desirable function for many domain scientists would be the ability to discover new common features within correlated data sets. One main issue to domain scientists is conventional data mining's ability to scale-up and be useful with highly distributed heterogeneous data sets.

### 4.3.6 Distributed Algorithms for Search and Retrieval, Pattern Recognition, and Similarity Detection

Pattern recognition seeks to discover information (features) from either visual information or some other data format such as a data table or multidimensional data set. To accomplish this, it is necessary to "restate" information in unusual or different forms, so that patterns in structure or state may emerge or be discovered. The algorithms of interest to the DICPM workshop do not deal with simple pattern recognition in a local data set, but rather, with recognition within a very large heterogeneous and widely distributed multidimensional data set.

To support pattern recognition, new visualization and query navigation tools are required to allow the non-expert to use and discover knowledge from visual and multimedia data. Pattern recognition relies on the ability to express meaningful queries over the stored information. For example, patterns inhabit some region in space and time defined by its variables. This feature space can be examined to look for boundaries, spatial constraints, distances from some mean, nearest neighbors, partitions, etc. The idea is that if data can be described as being related to a simple feature or more complex feature group then patterns of similar form can be discovered from previously unknown information.

Many researchers have indicated a desire to see visualization in the form of pattern recognition applied together with data mining tools to further enhance the domain scientist's ability to learn and acquire previously unknown knowledge from raw data. In addition, it is desirable for the domain scientist to have the capability to easily use the tools without the requirement of becoming a "computer professional." For example, oceanographers wish to study how nutrients, physics, and biological entities of interest interact with higher trophic bodies over some time frame, but do not wish to be required to know how to code visualization or mining algorithms to accomplish this. They simply wish to be able to select the time frames and parameters of interest and let the machine deliver the information to them in useful forms.

The form of these and many other queries is to "find something like this pattern or object" within a particular collection of data sets. To be effective, the pattern recognition tools must be able to operate within a distributed environment, be highly operable on spatial and temporally based heterogeneous data sets, and be dynamically alterable by the user if they are to become widely useful to the domain scientist.

### 4.3.7 Modes of Collaboration

Progress in science and engineering requires two types of communication:

1. Two-way synchronous and asynchronous communication between geographically distributed collaborators, members of a design team, or of a scientific project; and
2. One-way asynchronous dissemination of research results or design solutions for other practitioners to reuse in their own research efforts or design activities.

Currently, both types rely primarily on traditional media (e.g., text, images, blue-prints, etc.) and on access to simulation datasets or design models. In such a setting, communication is significantly less effective than when the collaborators are co-located and can discuss face-to-face, while pointing to specific features of a dataset or specific components of the design. It is therefore important to provide scientists and engineers with a new generation of easy-to-use communication systems that integrate such "natural" media (such as voice and gesture) with the geometric and visual representations of simulation results or design components. Furthermore, it is important to understand how and when such natural media should be used, and how they are related to evolving collaboration practices and with other personal productivity and electronic communication tools. In dealing with the types of highly mobile computing platforms that are becoming available, the role of process management of the collaborative enterprise becomes increasingly important.

#### 4.3.8 Performance Issues

Domain scientists and engineers will increasingly conduct their work through the Internet. All aspects of the scientific and engineering enterprise, including the writing of research papers, scheduling of on-line interactive collaborations, videoconferences, and remote experiments, will be performed over networks, making severe demands on CPU power, bandwidth, and throughput of a hitherto unknown magnitude. Significant research effort is needed towards achieving the goal of providing domain scientists with appropriate scheduling tools, mechanisms, and functionalities with Quality of Service (QoS) offerings (for example, bandwidth reservation schemes, packet loss probability enforcement schemes, Service Level Agreement mechanisms provided by networks and operating systems). Use of market mechanisms [11, 18] should be investigated in addition to other possible paradigms for on-the-fly resource allocation.

Any effective load balancing and scheduling mechanism should rely on extensive monitoring and tracking of workflow states, transitions, and events. Research should be conducted on how to trace records originating from remote parts of a heterogeneous distributed system. This information should be correlated to provide adequate information about contention and usage of network resources (both hardware and software). This information should further be correlated to yield insights into delays experienced by the users of the network. In addition, proper interfacing and visualization of flow monitoring information [13, 21] helps the users to have a direct understanding of the system's performance. Effective caching and replication schemes should be used to alleviate network traffic and minimize response time [16].

Parallel batch processing environments such as PVM and Condor [9, 19] will become ubiquitous as people increasingly make use of the power of networks of workstations and of parallel supercomputers. Difficult problems of heterogeneity and code parallelization will continue to plague the community, unless a research effort in this area provides the needed solutions.

#### 4.3.9 Scale Issues and Compression

When exchanging over a network the large amounts of data required by many scientific and engineering applications, bandwidth limitation becomes a major concern. One way around this problem is to improve the compression schemes utilized to reduce the size of the information before transmitting it, essentially trading off CPU cycles at both ends for increased bandwidth. Besides general compression methods in widespread use today, researchers have shown how specific algorithms that minimize data redundancy in specific domains may achieve even greater efficiencies.

Unfortunately, data compression has a cost in producing terse data collections that are much harder to understand and process because of the removed redundancies. Additionally, these methods may occasionally fail to provide the desired information fast enough, or they will provide the entire dataset when some simplified version might have sufficed. Multiresolution techniques complement compression by providing many levels of detail of the same dataset, so that the data can be accessed according to the needs of a problem domain. One specific kind of multiresolution also affords incremental data transfer, where a quickly delivered rough approximation is superseded by subsequent refinements. (One prominent example of this technique is the use of interlaced web page images that gradually gain increased definition while letting the user proceed with his or her work.)

## 4.4 Issues for Facilitators

In moving in the directions and towards the goals suggested here, it will be necessary to address and perhaps change the ways in which the research community, industry, professional associations, and government funding agencies have traditionally operated. Attendants at the DICPM workshop had numerous concerns and recommendations in this context that emanate from a perception that the framework in which their work is presently funded, conducted, and evaluated does not favor the pursuit of complex interdisciplinary endeavors. Issues that were explicitly raised include:

- Building inter- and multidisciplinary research teams.

  Setting up a team to work on a broad, multidisciplinary problem involves larger costs and greater effort. Members of the group have to work to find a common language or to be able to understand each other's language, concerns, constraints, etc. Since these efforts are neither valued nor encouraged, young researchers tend to shy away from these types of projects.

  Furthermore, even when such a team can be put together and is willing to go forward with a challenging project, finding funding for the project is difficult, since funding is generally allocated along pure disciplinary lines. Moreover, the fair and objective evaluation of such projects entails higher complexity, requiring teams of evaluators with similar (or even more extensive) multidisciplinary backgrounds. Also, the recommendations from diverse evaluation teams may be difficult to reconcile and merge.

- Rewards, publications, tenure, promotion

  The standards by which a researcher's work is measured today also tend to discourage the undertaking of extensive multidisciplinary projects. Because of the extra costs discussed previously, the rate at which publications can be produced in these environments may be lower (or at least may be foreseen as being lower). It may also be more difficult to get them published because of the lack of adequate forums. Since publications have a direct impact on all currently accepted mechanisms of promotion, this is an important hurdle, especially for younger researchers. If multidisciplinary projects are to prosper, it will require institutions to rethink their policies and establish premiums to leverage these (perceived) difficulties and make these projects more attractive to all researchers.

- Resources for software development and maintenance

  It is not the same to develop software in isolation to test an idea or prove a concept than to solve problems in an heterogeneous and multidisciplinary team. The latter requires much more effort in the areas of usability and robustness of the software, and furthermore, requires more substantial efforts in software maintenance and user support. These efforts are generally associated with commercial, rather than research software, and have a high cost that research teams are not currently able to cover. This is an important issue that will have to be addressed when formulating funding policies. However, in response to adequate funding for multidisciplinary software development, the result should be in the form of more reusable and sharable code from which other groups or industries can benefit.

- Resources for producing usable data

  As with software, raw data requires a large and sustained effort in terms of organization, documentation, cleaning, and archiving in order for the data to be usable by others than those directly or closely involved in its collection. This is not unrelated to the current debate on the public availability of data originating from government funded research. To achieve long-term availability, use, and reuse of the data, the extra cost of making it usable will have to be met, most likely by government funding agencies, professional societies, and possibly, the end users.

- Resources for usability and reusability of software

  It is not sufficient merely to build more robust and portable software. For these software products to be reused in diverse remote contexts, a ubiquitous distributed communications infrastructure is necessary. This capability will require substantial initial effort to establish, and should also include adequate tools to make it possible to easily discover appropriate hardware and software (data *and* programs) resources.

- Survey of data format usage (what, why)

  In order to establish a baseline for measuring future research and product deployment, it is advisable to conduct an extensive survey to determine the data and software formats being used currently (or being planned) in various scientific and engineering domains, and the reasons why they have been chosen. A broad survey of this nature would allow for more rational planning in trying to bring together the necessary infrastructure for the future open exchange and sharing of large datasets and software systems among the scientific community. This survey is suggested in the conviction that for these efforts to succeed, we will need to accommodate current practice, rather than trying to force new and different methodologies upon research groups.

It is in part by devising appropriate policies to meet these challenges that we may expect to succeed in fostering the most advanced broad-spectrum multidisciplinary research efforts in the near future.

# 5 Conclusions and Recommendations

A number of specific barriers to effective multidisciplinary activities have been identified. These barriers can be characterized as *computational*, *structural*, and *social*.

**Computational.** reliable and supported tools do not currently exist to facilitate the cooperative and collaborative sharing of data and software by domain scientists.

**Structural.** Except for some recent programs, such as the NSF Knowledge and Distributed Intelligence (KDI) Initiative, current funding mechanisms and career development practices favor narrowly focused research activities.

**Social.** Perhaps most fundamentally, there has been little emphasis placed on multidisciplinary cooperation by the members of the research community. Through education and practice we have preferred to develop and exercise expertise in narrow topic areas. Until this attitude can be changed, improvements in the first two areas will have little impact.

Towards the goal of removing these barriers, the DICPM workshop participants propose the following suggestions:

1. To address the structural barriers to effective multidisciplinary activities, the appropriate facilitators in the research community, industry, and government can foster collaborative research by providing funding explicitly for multidisciplinary projects, extending the scope of the few such initiatives currently in place.
2. Additionally, to relieve related social barriers, the facilitators must find the means to provide financial support and career incentives to foster cooperation between computer scientists and domain scientists, and to encourage team-based approaches to multidisciplinary problems. These collaborations should occur at both the national and international levels. International cooperation is expected to be enhanced by programs such as the *New Vistas in Transatlantic Science and Technology Cooperation* recently initiated by the US and European Union.
3. In an effort to provide widespread dissemination of research knowledge beyond conventional domain boundaries, reducing the social barriers to collaboration, the research community, professional societies, industry, and government should help to establish a national (and possibly, international) digital library for the physical, biological, and social sciences and engineering.
4. To further reduce the computational barriers to collaboration, the facilitators should cooperate to establish a global distributed information registry and repository (or "virtual scientific marketplace" [7]) for expert knowledge, simulation and analysis tools, and procedures. Such a repository could be organized in terms of a product-oriented model for legacy data and software, or as a service-oriented model for a more flexible and easier-to-use agent-based information system.

> The success of this repository will depend upon the efforts of teams of computer scientists, domain scientists, and software vendors to construct curated and documented repositories with advanced search capabilities. To produce the necessary tools for such a repository, research funding will be required for the theme areas identified previously in Section 4.3, including middleware and distributed resource discovery; metadata and tools for reusability; user interfaces and visualization; navigation, query languages, data exploration, and browsing; modes of collaboration; performance issues; scale, compression, and multi-resolution representation; data mining; integration, heterogeneity, and reusability; and distributed algorithms for search and retrieval, pattern recognition, and similarity detection.

The success of the STEP standard for the exchange of manufacturing product data (driven by Department of Defense requirements) provides a strong example of a collaborative development process undertaken by the research community, professional societies, industry, and government that can be applied to multidisciplinary science and engineering. These areas need their facilitators (DoD, Department of Energy, Department of Commerce, NSF, etc.) to motivate standards (e.g., metadata standardization) and multidisciplinary cooperation and collaboration.

# 6 References


[1] American National Standard Acoustical Terminology. ANSI S1.4-1994, New York, 1994.

[2] R. Boisvert, S. Browne, J. Dongarra, and E. Grosse. Digital software and data repositories for support of scientific computing. In *A Forum on Research and Technology Advances in Digital Libraries (DL '95), May 15-19, 1995, McLean, Virginia*, 1995.

[3] T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible markup language (XML) 1.0. REC-xml-19980210, February 10, 1998. <http://www.w3.org/TR/REC-xml>.

[4] H. Casanova, J. J. Dongarra, and K. Moore. Network-enabled solvers and the NetSolve project. *SIAM News*, 31(1), January 1998.

[5] Federal Geographic Metadata Committee. Content standards for digital geospatial metadata, June 4, 1994. <http://geology.usgs.gov/tools/metadata/standard/metadata.html>.

[6] T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb. Autonomous ocean sampling networks. *Oceanography*, 6(3):86–94, 1993.

[7] M. L. Dertouzos. *What Will Be: How the New World of Information Will Change Our Lives.* HarperEdge, San Francisco, 1997.

[8] J. Dongarra, T. Rowan, and R. Wade. Software distribution using XNETLIB. *ACM Transactions on Mathematical Software*, 21(1):79–88, March 1995.

[9] J. J. Dongarra, G. A. Geist, R. Manchek, and V. S. Sunderam. Integrated PVM framework supports heterogeneous network computing. *Computers in Physics*, 7(2):166–175, April 1993.

[10] Dublin core metadata element set: Resource page, November 2, 1997. <http://purl.org/metadata/dublin_core/>.

[11] D. F. Ferguson, C. N. Nikolaou, J. Sairamesh, and Y. Yemini. Economic models for allocating resources in computer systems. In S. H. Clearwater, editor, *Market-Based Control: A paradigm for distributed resource allocation*, chapter 7, pages 156–183. World Scientific, 1995.

[12] Object Management Group. OMG home page, June 12, 1998. <http://www.omg.org/>.

[13] W. Gu, J. Vetter, and K. Schwan. An annotated bibliography of interactive program steering. *ACM SIGPLAN Notices*, 29(9):140–148, July 1994.

[14] National Biological Information Infrastructure. NBII biological metadata standard, May 12, 1998. <http://www.nbii.gov/standards/metadata.html>.

[15] Java technology home page, June 16, 1998. <http://www.javasoft.com/>.

[16] S. Kapidakis, S. Terzis, J. Sairamesh, A. Anastasiadi, and C. N. Nikolaou. A management architecture for measuring and monitoring the behavior of digital libraries. In *First International Conference on Information and Computation Economies, ICE-98*, October 25–28 1998.

[17] C. Lagoze, C. A. Lynch, and R. Daniel, Jr. The Warwick framework: A container architecture for aggregating sets of metadata. Technical Report TR96-1593, Cornell University, Ithaca, NY, June 21 1996. <http:// cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR96-1593>.

[18] S. Lalis, C. N. Nikolaou, and M. Marazakis. Market-driven service allocation in a QoS-capable environment. In *First International Conference on Information and Computation Economies, ICE-98*, October 25-28 1998.

[19] M. Litzkow, M. Livny, and M. W. Mutka. Condor — a hunter of idle workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*, pages 104–111, June 1988.

[20] C. J. Lozano, A. R. Robinson, H. G. Arango, A. Gangopadhyay, N. Q. Sloan, P. H. Haley, and W. G. Leslie. An interdisciplinary ocean prediction system: Assimilation strategies and structured data models. In P. Malanotte-Rizzoli, editor, *Modern Approaches to Data Assimilation in Ocean Modeling*. Elsevier, The Netherlands, 1996.

[21] M. Marazakis, D. Papadakis, and C. N. Nikolaou. Management of work sessions in dynamic open environments. In *Proceedings of the International Workshop of Workflow Management*. IEEE Computer Society Press, 1998.

[22] R. Swick, E. Miller, B. Schloss, and D. Singer. W3C resource description framework, June 3, 1998. <http://www.w3.org/RDF/Overview.html>.

[23] S. Vinoski. CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications Magazine*, 14(2), February 1997.

# 7 Workshop Program

## 7.1 Friday, May 15

| Friday, May 15, 1998 | | |
|---|---|---|
| **Time** | **Location** | **Event** |
| 7:30– 9:00 | Concord B | Registration and continental breakfast |
| 9:00– 9:15 | Concord B | Introduction<br>Prof. Nicholas M. Patrikalakis, MIT |
| 9:15– 9:45 | Concord B | "*Collaborative Design and Visualization*"<br>Prof. Jarek Rossignac, Georgia Institute of Technology |
| 9:45–10:00 | Concord B | "*Advanced Networks for Distributed Systems*"<br>Dr. George Strawn, NSF |
| 10:00–10:15 | Concord B | "*Infrastructure for New Computational Challenges*"<br>Dr. John C. Cherniavsky, NSF |
| 10:15–10:30 | | Break |
| 10:30–11:00 | Concord B | "*Forecasting and Simulating the Physical-Acoustical-Optical-Biological-Chemical-Geological Ocean*"<br>Prof. Allan R. Robinson, Harvard University |
| 11:00–11:45 | Concord B | "*The Metadata Landscape: Conventions for Semantics, Structure, and Syntax in the Internet Commons*"<br>Dr. Stuart Weibel, Online Computer Library Center |
| 11:45–12:00 | | Break |
| 12:00– 1:30 | Concord B | Lunch<br>"*Information Science Initiatives at NASA*"<br>Dr. William J. Campbell, NASA |
| 1:30– 1:45 | Concord B | Breakout group goals, themes, and topics<br>Prof. Nicholas M. Patrikalakis, MIT |
| 1:45– 3:30 | Concord B<br>Wright<br>Loudoun<br>Clipper | Four breakout groups meet to charter future research agenda |
| 3:30– 3:45 | | Break |
| 3:45– 6:00 | Concord B<br>Wright<br>Loudoun<br>Clipper | Four breakout groups continue |
| 6:00– 6:30 | | Break |
| 6:30– 7:00 | Concord A | Cash bar/social |
| 7:00– 9:00 | Concord A | Dinner<br>Speech by Dr. Joseph Bordogna, NSF |

## 7.2 Saturday, May 16

| Saturday, May 16, 1998 | | |
|---|---|---|
| **Time** | **Location** | **Event** |
| 7:30– 8:30 | | Continental breakfast |
| 8:30– 8:45 | Earhart/Lindbergh | "*Workflow and Process Automation in Inforormation Systems: A Proposal for a Multidisciplinary Research Agenda and Follow-up*"<br>Prof. Amit P. Sheth, University of Georgia |
| 8:45– 9:00 | Earhart/Lindbergh | "*Interfaces to Scientific Data Archives*"<br>Dr. Roy D. Williams, California Institute of Technology |
| 9:00– 9:15 | Earhart/Lindbergh | "*Integrating Biological Databases*"<br>Prof. Peter Buneman, University of Pennsylvania |
| 9:15–10:15 | Earhart/Lindbergh<br>Fairfax<br>Loudoun<br>Clipper | Four breakout groups meet to summarize their discussion |
| 10:15–10:30 | | Break |
| 10:30–11:30 | Earhart/Lindbergh<br>Fairfax<br>Loudoun<br>Clipper | Four breakout groups meet to prepare plenary presentations |
| 11:30– 1:15 | Wright | Lunch, "*STEP Models and Technology*"<br>Dr. Peter R. Wilson, Boeing |
| 1:15– 3:15 | Earhart/Lindbergh | Four breakout group summary presentations to plenary session |
| 3:15– 3:30 | | Break |
| 3:30– 4:25 | Earhart/Lindbergh | Plenary discussion of group recommendations<br>Moderator: Prof. Nicholas M. Patrikalakis, MIT |
| 4:25– 4:30 | Earhart/Lindbergh | Concluding remarks to plenary session<br>Dr. Maria Zemankova, NSF |
| 4:30– | | Plenary session of workshop adjourns |
| 4:30– 6:00 | Fairfax | Executive session with breakout group chairs, workshop program committee, and NSF observers to discuss final report |
| 6:00– | | Workshop adjourns |

# 8 Workshop Participants

Prof. Nabil R. Adam, Rutgers University
Dr. Ethan Alpert, National Center for Atmospheric Research
Prof. William P. Birmingham, University of Michigan
Dr. Joseph Bordogna, NSF
Prof. Francis Bretherton, University of Wisconsin
Prof. Lawrence Buja, National Center for Atmospheric Research
Prof. Peter Buneman, University of Pennsylvania
Dr. William J. Campbell, NASA
Dr. John C. Cherniavsky, NSF
Dr. Kenneth W. Church, AT&T Labs
Dr. Geoff Crowley, Southwest Research Institute
Prof. Judy B. Cushing, The Evergreen State College
Dr. Ernie Daddio, NOAA
Dr. Donald Denbo, NOAA
Dr. Pierre Elisseeff, MIT
Prof. Zorana Ercegovac, UCLA
Prof. Gregory L. Fenves, University of California, Berkeley
Prof. Paul J. Fortier, University of Massachusetts Dartmouth
Dr. Christopher J. Freitas, Southwest Research Institute
Prof. James C. French, University of Virginia
Prof. Martin Hardwick, RPI
Dr. George Hazelrigg, NSF
Prof. Yannis E. Ioannidis, University of Athens, Greece
Prof. Alan Kaplan, Clemson University
Prof. Miron Livny, University of Wisconsin
Dr. Christopher Miller, NOAA
Dr. Terry R. Moore, University of Tennessee, Knoxville
Dr. Charles Myers, NSF
Prof. Christos Nikolaou, University of Crete, ICS/FORTH, Greece
Prof. Nicholas M. Patrikalakis, MIT
Dr. Lawrence Raymond, Oracle Corporation
Prof. Allan R. Robinson, Harvard University
Prof. Jarek Rossignac, Georgia Institute of Technology
Prof. M. Saiid Saiidi, University of Nevada, Reno
Dr. Jakka Sairamesh, IBM Research
Dr. Clifford A. Shaffer, Virginia Polytechnic Institute and State University
Prof. Amit P. Sheth, University of Georgia
Prof. Terence Smith, University of California, Santa Barbara
Dr. Ram D. Sriram, NIST
Dr. George Strawn, NSF
Dr. Knut Streitlien, MIT
Dr. Jean-Baptiste Thibaud, NSF
Prof. George Turkiyyah, University of Washington
Prof. Alvaro Vinacua, Polytechnic University of Catalonia, Spain
Dr. Stuart Weibel, Online Computer Library Center
Prof. Jack C. Wileden, University of Massachusetts
Prof. John R. Williams, MIT
Dr. Roy D. Williams, California Institute of Technology
Dr. Peter R. Wilson, Boeing Commercial Airplane Group
Prof. Clement Yu, University of Illinois at Chicago
Dr. Maria Zemankova, NSF

# 9 Invited Presentations

## 9.1 Jarek Rossignac

"*Collaborative Design and Visualization*"
Prof. Jarek Rossignac, Georgia Institute of Technology
Friday, May 15, 9:15–9:45, Concord B Ballroom

To enhance collaborative design, interactive 3D visualization must be trivial-to-use and fully integrated with the established workflow practices and other personal productivity, connectivity, and communication tools. When successfully combined with selective Internet access and multimedia annotations, interactive visualization tools will become the principal interface from which design and analysis activities are controlled and coordinated. We will report the conclusions of TeamCAD, the first GVU/NIST workshop on Collaborative CAD, and will review the state of the art in the compression, progressive download, and real-time rendering of highly complex 3D models and in interactive navigation and manipulation tools.

Jarek Rossignac is the Director of GVU, Georgia Tech's Graphics, Visualization, and Usability Center, which involves 51 faculty members and over 160 graduate students focused on technologies that make humans more effective. Prior to joining Georgia Tech, he worked at the IBM T.J. Watson Research Center as the Strategist for Visualization; the Senior Manager of the Visualization, Interaction, and Graphics department; and the Manager of several IBM's graphics products: 3D Interaction Accelerator, Data Explorer, and PanoramIX. Rossignac holds a PhD in EE from the University of Rochester, New York in the area of Solid Modeling, chaired 12 conferences, workshops, and program committees in Graphics, Solid Modeling, and Computational Geometry; guest edited 7 special issues of professional journals; and and co-authored 13 patents in these areas.

## 9.2 George Strawn

"*Advanced Networks for Distributed Systems*"
Dr. George Strawn, National Science Foundation
Friday, May 15, 9:45–10:00, Concord B Ballroom

The Science, Technology and Research Transit Access Point (STAR TAP, http://www.startap.net) project was funded by the National Science Foundation (NSF) in April, 1997 for the purpose of providing an international exchange point for high performance-networked applications. STAR TAP is one of the exchange points for Next Generation Internet (NGI). It is also the same facility as one of the emerging Internet 2 GigaPops. The Department of Energy's ESnet and NASA's NREN have also connected to STAR TAP, as have Canada's CA*net II and Singapore's SINGAREN. DoD's DREN, Taiwan's TAnet, and other high-performance networks from Asia-Pacific, Europe, and, perhaps, Latin America are expected to connect shortly. The STAR TAP is staffed with networking and applications specialists, who are charged to assist in tuning end-to-end applications that support international high-end collaborations.

Key Words: STAR TAP, high-performance applications, Internet Exchange Point, Acceptable Use Policy, vBNS, CA*net II, ESnet, NREN, Next Generation Internet, Internet 2, GigaPoP, SINGAREN, TAnet

## 9.3 John C. Cherniavsky

"*Infrastructure for New Computational Challenges*"
Dr. John C. Cherniavsky, National Science Foundation
Friday, May 15, 10:00–10:30, Concord B Ballroom

Dr. Cherniavsky is the Director of the Division of Experimental and Integrative Activities within the Computer and Information Science and Engineering Directorate of the National Science Foundation.

## 9.4 Allan R. Robinson

"*Forecasting and Simulating the Physical-Acoustical-Optical-Biological-Chemical-Geological Ocean*"
Prof. Allan R. Robinson, Harvard University
Friday, May 15, 10:30–11:00, Concord B Ballroom

The ocean is a complex multidisciplinary fluid system in which dynamical processes and phenomena occur over a range of interactive space and time scales, spanning more than nine decades. The fundamental interdisciplinary problems of ocean science, identified in general more than half a century ago, are just now feasible to research. They are being pursued via research in ecosystem dynamics, biogeochemical cycles, littoral-coastal-deep sea interactions, and climate and global change dynamics. Concomitantly, novel progress in maritime operations and coastal ocean management is underway. Realistic four-dimensional (three spatial dimensions and time) interdisciplinary field estimation is essential for efficient research and applications. State variables to be estimated include, e.g., temperature, salinity, velocity, concentrations of nutrients and plankton, ensonification, irradiance and suspended sediments. Such estimates, including real time nowcasts and forecasts as well as simulations, are now feasible because of the advent of Ocean Observing and Prediction Systems, in which: (i) a set of coupled interdisciplinary models are linked to; (ii) an observational network consisting of various sensors mounted on a variety of platforms; and (iii) via data assimilation schemes. Compatible multiscale nested grids for the models and observations are essential. The Harvard Ocean Prediction System (HOPS) is such a system: it is generic, portable and flexible, and is being and has been used in many regions of the world ocean. HOPS is a central component of both the advanced Littoral Ocean Observing and Prediction System (LOOPS) being developed collaboratively under the National Ocean Partnership Program, and of the prototype Advanced Fisheries Management Information System (AFMIS). These advanced systems exemplify the research directions and needs of the ocean science community for novel and efficient distributed information systems, which can only be realized by a powerful integrated effort in concert with the computational, distributed system and engineering communities.

Allan R. Robinson, Ph.D., is Gordon McKay Professor of Geophysical Fluid Dynamics in the Division of Engineering and Applied Sciences and the Department of Earth and Planetary Sciences at Harvard University, where he has served as the Director of the Center for Earth and Planetary Sciences and the Chairman of the Committee on Oceanography. He holds a B.A. (mcl), M.A., and Ph.D., all in physics from Harvard University, and Doctor Honoris Causa from the University of Liege. Professor Robinson is currently the Distinguished Senior Fellow of the Center for Marine Science and Technology at the University of Massachusetts (Dartmouth), and has held numerous visiting professorships, including the Slichter Professorship (UCLA), the COMNAVOCEANCOM Professorship (Naval Postgraduate School), and the Franqui Professorship (University of Liege). He has been Visiting Scientist or Professor at the NATO SACLANT Undersea Research Centre, Woods Hole Oceanographic Institution, Johns Hopkins Applied Physics Laboratory, Cambridge University, Imperial College (London), and the Indian Institute of Sciences (Bangalore). He was awarded the ONR Distinguished Educators Award in Ocean Science in 1991. Professor Robinson's research interests and contributions have encompassed dynamics of rotating and stratified fluids, boundary-layer flows, thermocline dynamics, the dynamics and modeling of ocean currents, and the influence of physical processes on biological dynamics in the ocean. He is recognized as one of the pioneer and leading experts in modern ocean prediction, and has contributed significantly to the techniques for the assimilation of data into ocean forecasting models. Professor Robinson has served on numerous national and international advisory committees, including the Ocean Sciences and Naval Studies Boards of the National Research Council. He chairs and has chaired many programs and working groups for international cooperative science, including those associated with ocean mesoscale dynamics, ocean prediction, the Mediterranean Sea, global ecosystem dynamics, and the global coastal ocean. Prof. Robinson is currently lead PI of the NOPP program LOOPS. He has authored and edited more than a hundred and fifty research articles and books, and is currently editor-in-chief of the prestigious series of treatises, *The Sea*, and the journal, *Dynamics of Atmospheres and Oceans*. Professor Robinson is a Fellow of the American Academy of Arts and Sciences, the American Association for the Advancement of Science, and the American Geophysical Union.

## 9.5 Stuart Weibel

"*The Metadata Landscape: Conventions for Semantics, Structure, and Syntax in the Internet Commons*"
DR. STUART WEIBEL, Online Computer Library Center, Inc.
Friday, May 15, 11:00–11:45, Concord B Ballroom

Stuart Weibel has worked in the Office of Research at OCLC since 1985. During this time he has managed projects in the areas of automated cataloging, document capture and structure analysis, and electronic publishing. He currently coordinates the Dublin Core Metadata

Workshop Series and related applications of World Wide Web technology and Internet protocol standardization efforts. Dr. Weibel is also a founding member of the World Wide Web Conference Coordinating Committee.

## 9.6 William J. Campbell

"*Information Science Initiatives at NASA*"
Dr. William J. Campbell, NASA
Friday, May 15, Luncheon speaker, Concord B Ballroom

A major barrier to the wider use of Earth remote sensing data is timely access to satellite data products that can be combined easily with other resource management applications already in use by the general user community. The Regional Applications Center program was initiated by Goddard Space Flight Center's Applied Information Science Branch, to extend the benefits of its information technology research and cost-effective system development to a broader user community. This talk will highlight a variety of technologies being developed to support this program.

William J. Campbell is the Head of the Applied Information Science Branch at NASA Goddard Space Flight Center. The Branch is responsible for the development of intelligent information and data management value-added systems. He also serves as an Associate Editor for the Photogrammetric Engineering and Remote Sensing Journal, and has served as a member of the Committee on Information, Robotics, and Intelligent Systems, and Human Centered Systems for the National Science Foundation and as a Committee Member of the National Resource Council of the National Academy of Sciences. He is the recipient of over 25 NASA awards including NASA's Exceptional Achievement Medal. He has been at NASA for 19 years.

## 9.7 Joseph Bordogna

Dr. Joseph Bordogna, National Science Foundation
Friday, May 15, Dinner Speaker, Concord A Ballroom

Joseph Bordogna is Acting Deputy Director and Chief Operating Officer of the National Science Foundation. Complementing his NSF duties, he has chaired Committees on Manufacturing and Environmental Technologies within the President's National Science and Technology Council, was a member of the Federal Government's Technology Reinvestment Project team (TRP), and serves on the Partnership for a New Generation of Vehicles (PNGV) Committee, and the U.S.-Japan Joint Optoelectronics Project.

He received the B.S.E.E. and Ph.D. degrees from the University of Pennsylvania and the S.M. degree from the Massachusetts Institute of Technology. As well as his assignment at NSF, his career includes experience as a line officer in the U.S. Navy, a practicing engineer in industry, and a professor.

Prior to appointment at NSF, he served at the University of Pennsylvania as Alfred Fitler

Moore Professor of Engineering, Director of The Moore School of Electrical Engineering, Dean of the School of Engineering and Applied Science, and Faculty Master of Stouffer College House, a living-learning student residence at the University.

He has made contributions to the engineering profession in a variety of areas including early laser communications systems, electro-optic recording materials, and holographic television playback systems. He was a founder of PRIME (Philadelphia Regional Introduction for Minorities to Engineering) and served on the Board of The Philadelphia Partnership for Education, community coalitions providing, respectively, supportive academic programs for K-12 students and teachers.

He is a Fellow of the American Association for the Advancement of Science (AAAS), the American Society for Engineering Education (ASEE), and the Institute of Electrical and Electronics Engineers (IEEE).

## 9.8 Amit P. Sheth

"*Workflow and Process Automation in Information Systems: A Proposal for a Multi-disciplinary Research Agenda and Follow-up*"
Prof. Amit P. Sheth, University of Georgia
Saturday, May 16, 8:30–8:45, Earhart/Lindbergh Suite

The final report of the NSF workshop on *Workflow and Process Automation in Information Systems* (http://lsdis.cs.uga.edu/activities/NSF-workflow) stated:

> "Work Activity Coordination involves such multidisciplinary research and goes beyond the current thinking in contemporary workflow management and Business Process Reengineering (BPR). In particular, instead of perceiving problems in prototypical terms such as the information factory, white-collar work and bureaucracy, we believe that this limited point of view can be explained by a lack of synergy between organizational science, methodologies, and computer science. Multidisciplinary research projects, based on mutual respect and willingness to learn from another discipline, can help to create a thriving research community that builds upon the strengths of different disciplines, such as distributed systems, database management, software process management, software engineering, organizational sciences, and others."

After a review of workshop results, we will discuss our research agenda in extending this emphasis on coordination to also include collaboration, information management and other scientific domains. We will discuss some new challenges in building a work coordination and collaboration environment to support virtual teams, which can support more effortless and productive human participation, and more flexible or dynamic interactions that are distributed in space and time.

We will also briefly discuss the follow-up to the workshop in two community building efforts. First is the launching of the NSF co-sponsored web-based open resource and repository, WORP, for research in WORkflow and Process management in Information Systems (http://lsdis.cs.uga.edu/ worp). The second is a group effort in initiating the International Joint Conference on Work Activities Coordination and Collaboration with the sponsorship of four ACM SIGs.

Dr. Amit Sheth directs the Large Scale Distributed Information Systems Lab (LSDIS, http:// lsdis.cs.uga.edu), is an Associate Professor of Computer Science at the University of Georgia, and the President of Infocosm, Inc. (http://www.infocosm.com). Earlier he worked for nine years in the R&D labs at Bellcore, Unisys, and Honeywell. His research interests are in developing work coordination and collaboration systems through intelligent integration of collaboration, collaboration and information management technologies, and in enabling Infocosm through semantic interoperability and information brokering. His research has led to two commercial products. A selection of his professional activities include four recent conference/workshop keynotes in the areas of workflow management and semantic interoperability, organization of the NSF workshop on *Workflow and Process Automation in Information System*, serving as a co-director of NATO ASI on Workflow Management Systems and Interoperability, and serving as the steering committee chair of the ACM International Joint Conference on Work Activities Coordination and Collaboration.

## 9.9 Roy D. Williams

"*Interfaces to Scientific Data Archives*"
Dr. Roy D. Williams, California Institute of Technology
Saturday, May 16, 8:45–9:00, Earhart/Lindbergh Suite

Dr. Williams will report some of the findings, open issues, and conclusions from the workshop *Interfaces to Scientific Data Archives*, held in Pasadena, California, March 25-27, 1998.

Dr. Williams received a Bachelors Degree in Mathematics at Cambridge and a Ph.D. in Physics from Caltech (1983). After postdocs in nuclear and solid-state physics at Caltech and Oxford, he worked with Fox and Messina at Caltech making parallel computers work for scientific simulation. His current work concerns building active digital libraries for scientific data, with collaborative archive projects in gravitational waves, remote sensing, geology, astronomy, and high-energy physics.

## 9.10 Peter Buneman

"*Integrating Biological Databases*"
Prof. Peter Buneman, University of Pennsylvania
Saturday, May 16, 9:00–9:15, Earhart/Lindbergh

Data associated with the Human Genome Project is widely distributed and stored in an extraordinary variety of data formats and data management systems. The structure of these databases is complex and rapidly evolving. I shall describe some of the challenges and successes in integrating these data sources, which is required for almost any form of hypothesis testing or data mining.

Peter Buneman is a Professor of Computer Science at the University of Pennsylvania. He is interested in databases, programming languages, and in the common principles of these subjects. He has also worked extensively in problems of data integration, and, with the database group at Penn, has developed integration software that is being applied in biological and other domains.

## 9.11 Peter R. Wilson

"*STEP Models and Technology*"
DR. PETER R. WILSON, Boeing Commercial Airplane Group
Saturday, May 16, Luncheon speaker, Wright Suite

Dr. Peter Wilson is a Principal Engineer at the Boeing Company working in the area of information modeling and STEP. Early in his career he was a semiconductor physicist in the UK before moving into solid modeling. He has worked in industry, academia and government research laboratories in the US and has been active in data exchange standards for almost twenty years. He is a past Editor-in-Chief of *IEEE Computer Graphics & Applications* and co-author of the book *Information Modeling the EXPRESS Way*.

# 10 Workgroups

**Group 1** [*Friday*: Concord B Ballroom / *Saturday*: Earhart/Lindbergh Suite]:

Jarek Rossignac (*Chair*)
Pierre Elisseeff (*Recorder*)

Ethan Alpert
Francis Bretherton
William P. Birmingham
Judy B. Cushing
Zorana Ercegovac
George Hazelrigg
James Kinter
Lawrence Raymond
Terence Smith
Jack C. Wileden
John R. Williams
Clement Yu

**Group 2** [*Friday*: Wright Suite / *Saturday*: Fairfax Boardroom]:

Peter Buneman (*Chair*)
Yannis Ioannidis (*Co-chair*)
Knut Streitlien (*Recorder*)

Lawrence Buja
William J. Campbell
Kenneth W. Church
Geoff Crowley
Alan Kaplan
Christopher Miller
Terence Moore
Charles E. Myers
Allan R. Robinson
George Turkiyyah
Maria Zemankova
Stuart Weibel

**Group 3** [Loudoun Boardroom]:

Christos N. Nikolaou (*Chair*)
Jakka Sairamesh (*Recorder*)

Donald Denbo
James C. French
Nicholas M. Patrikalakis
Ram D. Sriram
Jean-Baptiste Thibaud
Roy D. Williams
Peter R. Wilson

**Group 4** [Clipper Boardroom]:

Miron Livny (*Chair*)
Alvaro Vinacua (*Recorder*)

Ernest Daddio
Gregory L. Fenves
Paul J. Fortier
Christopher J. Freitas
Martin Hardwick
M. Saiid Saiidi
Clifford A. Shaffer
Amit P. Sheth

# 11 Position Papers

## 11.1 Nabil R. Adam

"*Digital Libraries in Environmental and Earth Sciences*"[1]

Prof. Nabil R. Adam
Rutgers University, Center for Information Management and Connectivity
180 University Avenue
Newark, NJ 07102
E-mail: adam@adam.rutgers.edu
URL: http://cimic.rutgers.edu

In this paper, we focus on digital libraries in the domain of Environmental and Earth sciences. We view digital libraries (DL) as an infrastructure for supporting the creation of information sources, the movement of information across global networks, the effective and efficient interaction among knowledge producers, librarians, and information and knowledge seekers.

Recently, Rutgers University Center for Information Management, Integration, and Connectivity (CIMIC) has been designated as a NASA Regional Applications Center (RAC). Rutgers CIMIC and the Hackensack Meadowlands Development Commission (HMDC) of the state of New Jersey have agreed to collaborate on establishing a space and land-based environmental monitoring system to address HMDC's primary areas of interest including land monitoring, fire detection, water and air quality, urban planning as well as development of outreach educational materials and programs for environmental education. Specifically, the establishment of the NASA RAC at Rutgers CIMIC will help contribute to the region in three areas:

1. Environmental and economical development of the region;
2. multi- disciplinary instructional and research opportunities for faculty and students at collaborating institutions; and
3. curriculum enrichment opportunities for elementary/high schools in the surrounding urban areas.

The NASA RAC at Rutgers CIMIC has just been augmented by a new major initiative supported by the HMDC: establishing the Meadowlands Environmental Research Institute (MERI) as a world-class center for the scientific investigation of urban wetlands, their functioning, restoration, and sustainable management. The research team (N. Adam, F. Artigas, V. Atluri, A. Gates, J. M. Hartman, G. Henebry, V. Hover, and J. Weis), which is made up of faculty drawn from various academic disciplines including Biological, ecological, Environmental, Geological, and Information Sciences has identified the following as MERI's research themes: vegetation patterns, mudflats, contaminant hotspots, scientific data management, and water and air quality.

[1]N. Adam would like to acknowledge support from the NASA Center of Excellence in Space Data and Information Sciences.

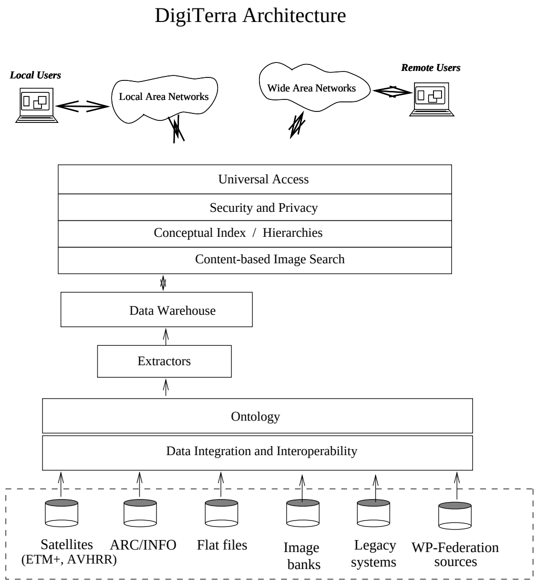



Figure 1: DigiTerra architecture.

We are in the process of developing an Environmental Digital Library called *DigiTerra* to meet the goals outlined above and to serve our diverse scientific user community. DigiTerra will facilitate the collection, assimilation, cataloging and dissemination/retrieval of a vast array of environmental data that includes images from a variety of space borne satellites, ground data from continuous monitoring weather stations, water and air quality, maps, reports and data sets from federal, state and local government agencies.

We are developing DigiTerra as a multilayered architecture consisting of the following layers:

**The integration and interoperability layer**. Concerned with collection, assimilation, cataloging and dissemination/retrieval of a vast array of environmental data, e.g., Images from a variety of space borne satellites, ground data from continuous monitoring weather stations, and maps, reports data sets from federal, state and local government agencies. There is a need efficient mechanism to integrate new data resources and provide seamless access to heterogeneous and autonomous sources.

**The data warehousing/data mining layer**. Provides fast and efficient access to the integrated data and to provide efficient data analysis and complex queries without interrupting the operational processing at the underlying information sources. Specific issues to address include:

- Appropriate data models given that we are no longer just concerned with numeric/character data
- Levels of aggregation that are meaningful for multimedia data, e.g., a collection of videos, maps, and images
- Methods to employ to allow a data warehouse to effectively evolve as new sources become available.

- Schema evolution that naturally follows as scientists develop better understanding of information at sources
- Enable scientists to push their findings back to the data warehouse

**The concept indexing and content-based retrieval layer**. Provides efficient retrieval by suitably organizing the multimedia data based the concepts associated with the objects. Issues to be addressed include:

- Developing a more expressive (versus keywords) and robust representation of multimedia objects that uses syntax and semantics
- Automating/semi-automating concept and feature discovery and extraction
- Classifying objects and indexing them based on extracted concepts and features
- Having a classification scheme that adopts to user's expertise, extensible, domain portable, and provides incremental modeling?

**The universal access layer**. Scientists wishing to access digital library objects possess varying capabilities and characteristics. User's capabilities include physical, technical, linguistics, and domain expertise. User's characteristics, on the other hand, include mobility, interests, preference and profile, and information appliances, e.g., PC, PDA, and TV. Universal access is concerned with having objects have built-in intelligence so that they can automatically manifest themselves to cater to different users' capabilities and characteristics. Specific issue to address include the ability to author objects suitable for automatic manifestation, to detect/identify and manage user capabilities and preferences/profiles, to accommodate the user mobility, and to detect the information appliance capabilities.

**The ontology layer**. Enable users with diverse backgrounds to query across multiple domains. Such a layer must be able to cater to users with diverse backgrounds by offering broader, more general, ontologies that are interlinked to cover many domains.

**The security and privacy layer**. Provides support for security of objects through suitable access control, watermarking and secure networking technologies and protect privacy of users. There is a need to be able to have authorizations be specified based on users' qualifications and characteristics rather than the users themselves; to authorize access to an object be specified based on the contents within an object rather than the object identifier; grant access to a specific portion(s) of the object rather than the entire object; be bale to have security policies of different heterogeneous and autonomous sources integrated and enforced; and be able to support copyright protection and copy prevention.

Each of these layers engenders a unique set of research issues that will be addressed in the course of project and tested in the DigiTerra prototype.

## 11.2 William P. Birmingham

"*Essential Issues in Distributed Computational Systems*"
Prof. William P. Birmingham
The University of Michigan, Electrical Engineering and Computer Science Department
Artificial Intelligence Laboratory
Ann Arbor, MI 48109
E-mail: wpb@eecs.umich.edu
URL: http://ai.eecs.umich.edu/people/wpb/home.html

Complex computational systems are manifested in many applications, such as engineering design (also known as concurrent engineering), digital libraries, and the simulation of various natural and man-made phenomena. Two common characteristics of these systems are the following: they are cast as distributed systems, composed of many elements that coordinate their computation to achieve shared (and perhaps individual) goals; and, they are open-ended in their requirements, implying that they evolve in terms of computation and tasks they perform, and the support that they give to users.

We have developed two complex computational systems for large-scale engineering-design and digital-library applications. These systems are distributed, as they are composed of hundreds of computational elements; for these applications the elements are software agents. These agents make various types of decisions using private (non-shared) knowledge and preferences. They coordinate their actions using a variety of mechanisms, some based on microeconomic methods, and some based on constraint-satisfaction and decision-theoretic methods.

An important property of the systems we have built is that they have decentralized control. We define informally such a system as one where there is no centralized decision maker, database, or any other computational element upon which the system critically depends. A common example of a decentralized system is the World Wide Web, where there is essentially no single controlling mechanism. We believe that decentralized control is essential to developing complex computational systems.

In view of the characterization of complex computational systems given earlier, we see some of the advantages of constructing decentralized systems as the following: decentralization implies a very modular software architecture that is easy to scale; incentive mechanisms can be designed to entice third parties to add, or remove, capabilities to the system, making the system more adaptive to changing requirements; control over proprietary resources can be maintained by individual organizations; and, it can be easier to model complex scientific problems, and then design and engineer complex computational systems as collections of smaller elements. Thus, we believe there are many advantages to a strongly decentralized view of computation.

From the perspective of the systems we have developed, some of the requirements for a decentralized system are the following:

- **Coordination**: An agent of the type we have developed is usually incapable of solving by itself the types of problems a user puts forth to the system. Thus, an agent must enlist the services of other agents. For example, in our digital library, many agents are involved in responding to a document query: one agent interfaces the user to the library, another plans a search strategy, another searches for documents in various document collections, yet another may reformulate the query, and so forth. This type of interaction requires coordination among the agents, both in terms of the types of computation they jointly perform, but also in terms of the incentive structure that motivates the agents to work together.
- **Common Communication Structures**: Coordination requires that agents in our systems send messages with intent. This, in turn, requires that the agents have a common message structure that includes syntax and semantics. In addition, various "rules of the game" (e.g., all agents are truthful) must be clear to all agents.
- **Ontologies**: A shared ontological commitment among agents is essential. Not only is this necessary for communication and coordination, but the ontology can be used for other tasks. For example, in the digital library, we use formal ontologies for both "advertising" services for agents, but also for search and query reformulation. Thus, formal ontologies with concomitant inference abilities have many benefits for decentralized systems.

## 11.3 Francis Bretherton

"*Metadata from the Perspective of an Environmental Scientist*"

PROF. FRANCIS BRETHERTON
University of Wisconsin–Madison
Space Science and Engineering Center
Madison, WI 53706
E-mail: fbretherton@ssec.wisc.edu

**Background**

Environmental sciences centered on climate and the atmosphere, but with concern for broad interdisciplinary issues such as Global Climate Change. Sound management and integration of environmental information from very diverse sources is central to meeting policy makers' concerns.

**Problem Areas**

1. Locating information–e.g., a web search for distributed technical collections. There is no existing superstructure for Keyword Schemas, though the FGCDC content standards provide a start. We need standards for registering and accessing specialized subsets and extensions emerging from diverse disciplines, and a framework for negotiating commonalities, synonyms, and homonyms among such subsets. The Dublin Core is a *must*, but tools for implementing it unambiguously are not quite there yet.

2. A critical requirement for understanding environmental change is documenting existing data streams for long-term analysis. A good test is to imagine what our successor scientists will think 20 years from now, when they are trying to determine whether an environmental variable has really changed, or whether apparent changes are due to the way we encoded or processed the measurements. The only evidence is the metadata record, and there are no insiders around to be asked.

   This issue involves an open-ended class of users, and the biggest problem is failure to record things that assumed to be either unimportant or generally known. The only strategy is:

   $$\text{Metadata content check lists} + \text{Scientific judgement}$$

   but there is little technical support for the activity. The inadequacies of the past require very expensive and difficult data archeology. With growing automation the problem is rapidly getting worse, not better.

3. Documenting complex computer models. Environmental science involves many high-volume observing systems such as satellites. Transforming these observations into usable data products frequently involves complex computer models (e.g., atmospheric general circulation models). The documentation from such models is almost never adequate to reconstruct the model from scratch, and even then rarely produces the same results. We need a framework for explicitly modeling the science concepts in a

language that links the conceptual level to declarative data structures and standard algorithms. A fundamental obstacle is the need to treat "approximate equivalence" of distinct representations of real-world objects, as opposed to "mathematical equivalence" expressed by axioms of equivalence classes. Such science concept modeling will inevitably be local and incomplete, raising downstream issues of merging independent, and possibly inconsistent, sub-models.

## 11.4 Lawrence Buja and Ethan Alpert

"*Computational and Large Data Issues in NCAR's Climate System Model*"

Dr. Lawrence Buja and Dr. Ethan Alpert
National Center for Atmospheric Research
Boulder, Colorado 80307-3000
E-mail: southern@ncar.ucar.edu, ethan@ncar.ucar.edu
URL: http://www.cgd.ucar.edu/csm/

The Climate System Model (CSM) project at the National Center for Atmospheric Research (NCAR) addresses one of the most difficult and urgent challenges in science today: understanding and predicting the Earth's climate, particularly climate variation and possible human induced climate change. Its user community spans a wide range of domains, ranging from the science arena, to the classroom and extending into the political circles of national policy making and societal impacts assessment.

As this model and its large data holdings are made available to its diverse community, a number of hurdles must be cleared. First, much of the existing CSM model output remains virtually inaccessible to all but those who are closely associated with the model development. Second, even when available, the data products themselves are in a discipline-specific form which inhibits easy cross-domain information exchange. Finally, although the CSM code itself is freely available, many researchers lack the computation resources necessary to carry out further scientific investigations with such a computationally intensive model.

At the National Center for Atmospheric Research, we are engaged in a number of efforts to address these issues.

**Climate Model Data Access**

The scientists who comprise the CSM research community span many disciplines and are located at a variety of sites around the globe. This presents a huge challenge to provide these researchers fast, intuitive and secure remote access to the data for research and decision making.

For typical climate modeling scenarios, the CSM typically produce tens of gigabytes of data. Further, the real-world observational datasets needed for model evaluation also consist of terabytes of data. Current data access methods (e.g., FTP) are inadequate to supply these data products to the CSM user community. Even the Next Generation Internet will not come close to addressing the massive data delivery needs of the CSM.

There is a clear need to provide easy and open access to the CSM data archives at NCAR, allowing outside researchers to remotely analyze and reduce the data to manageable volumes. Current projects aimed at this include extending the development of University of Rhode Island's Distributed Ocean Data System (DODS), evaluating the applicability of CORBA methodologies and broadening the NCAR Command Language (NCL) to utilize the unique computational and storage resources of NCAR's supercomputing center.

**Toward a Common Currency for Scientific Information Exchange**

Even greater than distance, as a barrier to understanding interactions among complex systems, are the discipline-specific forms in which the products of advancing knowledge typically are stored, transmitted and processed. Data containing complex structures or large values of numerical values can only be represented and used in a computer, which has resulted in the development of many specialized software tools and applications for data analysis and processing. These tools are often highly focused in a specific discipline and often don't address functionality beyond the scope of the scientist's latest endeavor. In many ways, scientific discourse is limited because often only a select group of scientists have both access the necessary computational resources and the necessary software skills to work with digital data. This is compounded further when scientists from other disciplines wish to investigate cross disciplinary data and aren't experts in the specific domain. What is needed is a "common currency" for scientific data exchange and analysis. A common currency would handle the details of translating data from its original discipline specific form and representation into a variety of equivalent representations. A common currency would also facilitate common ways of specify data reduction and processing tasks.

Defining a common currency for scientific data exchange represents many challenges. Systems that would handle this common currency exchange would be widely distributed, and require discipline specific intelligent "objects" to handle and sequence data conversion and processing. Additionally, user interfaces need to be flexible to allow a variety of users from different disciplines and skill levels to locate, understand and use effectively.

Much research is need to begin to understand how to implement a common currency for scientific data. The following are areas of key research needed:

- Can sufficient abstractions of scientific data be developed to support the object-oriented classification of scientific data representations?
- What types of meta-data are needed to support enabling the determination of computational semantics in an automated fashion?
- Can extant data belong to multiple classes of data and still have a single format and original representation?
- What types of interfaces (natural language, GUI) will best facilitate cross disciplinary distributed data processing application development? Can these interfaces be sufficiently intuitive for scientific researchers?
- How can classification of scientific data assist in locating data?

### Distributed Climate Modeling

The CSM is a loose coupling of distinct, yet simultaneously executing, atmospheric, ocean, land surface and sea-ice climate models linked together by small, well-defined, inter-component interface exchanges. The complex nature and enormous resources consumed by climate system models of this class have generally limited their application to studies at coarse geographical resolutions carried out at national or international supercomputer centers.

The advent of high-speed research networks, powerful mid-level computers and high-capacity storage media provides the necessary infrastructure to conduct large climate simulations at widely-distributed research centers. Spreading the large climate modeling computational load among a number of smaller research institutions would make it possible to extend the traditionally coarse climate modeling framework to interactively incorporate any number of fine-scale regional models at a scale impossible to achieve at a single supercomputer center. While proof of concept integrations show this is technically possible, arranging the necessary logistics and political cooperation to carry it through may be quite another matter.

**Summary**

Distributing a complex climate model and its large data archives to a diverse user community presents a number of challenges and we realize that these issues are by no means unique to the climate modeling community.

In the data arena, our basic premise is that distributed, object-oriented design and development methods, applied to suitable abstractions, can yield common currency for open cross-disciplinary scientific and technology information analysis and exchange.

Computationally, a distributed modeling framework holds the promise for allowing widely separated research centers to coordinate their resources to carry out major climate simulations, releasing large climate modeling out of the tight confines of the major supercomputer centers and into a much wider and diverse research community.

## 11.5 Peter Buneman

"*What is Special About Scientific Data?*"

Prof. Peter Buneman
University of Pennsylvania
Philadelphia, PA 19119
E-mail: peter@cis.upenn.edu
URL: http://www.cis.upenn.edu/∼peter

Scientific databases have a history that is as long as the history of database technology. It is therefore surprising to find relatively little penetration of database technology into the field of scientific databases. Moreover there appears to be very little technology that is common to scientific databases in general. Why is this?

To answer this question one has to look at how scientific data differs from "business" or "administrative" data upon which database technology has had a considerable impact. Most scientific data sets follow a data/metadata paradigm. The data is typically an image, a time series, a nucleic acid sequence, etc. – the result of some experiment or survey. The metadata typically consists of "administrative" data, such as where and when the image was recorded and perhaps some data on the experimental conditions. While administrative data can usually be represented cleanly in a relational database, the data itself cannot. Traditional relational systems cannot deal with the storage requirements for these special data types, nor can they accommodate the associated operations in the database query language.

The picture has changed somewhat with the development of object-oriented and object-relational systems where there has been some application of database technology to scientific data [3,4,5]. However, given the ten year history of these database management systems, the impact has not been great for systems that have effectively overcome most of the limitations of relational databases.

There are, I believe, other reasons why database technology has not penetrated scientific computing. If the metadata is relatively simple and there is no need for concurrency control (typically scientific data sets are archival) then the added functionality of a database management system, when compared with a simple indexing system, may not be worth the effort. But there are also two more fundamental reasons that I see in at least molecular biological databases.

First, the metadata in these databases is more than "administrative" data. It contains annotations that individuals have added in the process of analyzing and correcting the data. These annotations may have a complex structure and that structure may change. Even in object-oriented database management systems, that structure may be difficult to capture, and managing change in structure may be very difficult in an object-oriented database management system.

Second, what, if any, database management is used may not be the problem. The problem is integration. A "database of databases" [6] lists some 400 databases of general interest to researchers. This figure is growing at some 30 databases a year. Each database is heavily annotated or "curated" and typically deals with some aspect of research. One finds databases centered around diseases, organisms, or some genetic control mechanism. Because of the volatility of structure mentioned earlier, there is very little hope of producing an all-embracing database. But any given line of research will typically require access to several of these databases. Database technology is only now being developed to solve the integration problem.

At Penn we have been developing database integration technology that is robust with respect to structural changes and does not require the data to be in conventional databases. It can deal with the formats in which the data is currently held, typically some kind of structured text. It is based upon a new approach to database transformation that can handle the complex data types found in scientific databases [1] and we have had some success in applying this in bio-informatics [2]. The software is generic in the sense that once we have built an interface for the formatting convention or database interface used by the source there is no need to do further work when we are given a data source that conforms to that interface.

The issue of genericity brings me to the observation that each scientific discipline appears to be developing its own standards and its own formats. While it is true that the data (as described above) may require some special representation, there is no obvious need to develop special-purpose formats for metadata. There are several generic formats and data exchange protocols such as ASN.1, XML, CORBA that can readily be adapted to handle data as well as metadata. Coming to some consensus on interchange protocols and formats will simplify integration problems and will, perhaps, enable more interdisciplinary scientific research.

1. Peter Buneman, Shamim Naqvi, Val Tannen, and Limsoon Wong, " Principles of programming with complex objects and collection types," *Theoretical Computer Science*, 149(1):3–48, September 1995.
2. Susan Davidson, Christian Overton, Val Tannen, and Limsoon Wong "Biokleisli: A digital library for biomedical researchers," *Journal of Digital Libraries*, 1(1), November 1996.
3. N. Goodman, S. Rosen and L. Stein. "Building a laboratory information system around a C++-based object-oriented DBMS," *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.
4. R. L. Grossman, D. Hanley and X. Qin, "PTool: A Light Weight Persistent Object Manager," *Proceedings of SIGMOD 95*, 1995, 488–458.
5. D. Maier and D. M. Hansen. *Object Databases for Scientific Computing*, SSDBM, 1994, 196–206.
6. http://www.infobiogen.fr/index-english.html.

## 11.6 Kenneth W. Church

"*Data Publishing: An Alternative to Data Warehousing*"

Dr. Kenneth W. Church
AT&T Laboratories
Florham Park, NJ 07932
E-mail: kwc@research.att.com

Data publishing is like data warehousing, but with an emphasis on distribution. We don't want "Roach Motels," where data can check in, but it can't check out. What is important is not the size of the inventory, but how much is shipped. The more people who look at the data, the more likely someone will figure out how to put it together in just the right way. But unfortunately, as a practical matter, computer centers are very expensive, so expensive that only big ticket applications can justify the expense. Others cannot afford the price of admission. So much more could be accomplished if only we could find ways to make it cheap and convenient for lots of people to look at lots of data.

| Data Warehousing | Data Publishing |
|---|---|
| Centralization | Distribution |
| Large Investment | Affordability |
| – Big Databases and Indexes | – CD-ROM/Web |
| – Support (computer center) | – Shrink-wrapped |
| – Big Iron | – No Iron |
| – Provision for peak load | – "No load" |
| – On-line | – Off-line |
| High payoff applications | Mass Market |

Datamining need not be expensive. Large datasets are not all that large. Customer datasets are particularly limited. AT&T has a relatively large customer base, approximately 100 million residential customers. If we kept a kilobyte of data on each customer to store name, address, preferences, etc., the customer dataset would be only 100 gigabytes. The hardware to support basic datamining operations on a dataset of this size costs less than $15,000 today ($100/gigabyte for the diskspace + $5000 for the CPU). In a few years, I would hope that anyone could process a mere 100 gigabytes with whatever word processor is already on their desk.

A more interesting challenge is to mine purchasing records, but even these datasets are not all that large. The 100 million customers purchase about 1 billion calls per month. This kind of data feed is usually developed for billing purposes, but many companies are discovering a broad range of other applications such as fraud detection, inventory control, provisioning scarce resources, forecasting demand, etc. If we were to store 100 bytes for each call, a generous estimate, then we would need a 100 gigabytes of diskspace per month (= $10,000/month at $100/gigabyte). Thus, a month of purchasing records is about the same size as the customer dataset. A year of purchasing records is about a terabyte ($100,000 of disk space).

A hundred gigabytes or a terabyte might seem like a lot, but it really is not. The costs to replicate a dataset of this size are very reasonable, approximately \$10-\$100 thousand, or about what we used to spend for a workstation in 1980. Assuming that computer prices continue to fall, datamining stations will soon become about as commonplace as workstations are today.

Of course, there are many other challenges that need to be overcome before datamining stations become a commodity:

1. **Machine models**: With terabytes of external memory and only gigabytes of physical memory (and kilobytes of cache), these datamining stations look more like the tape-to-tape machines of the 1960s than random access memory. We need to rethink many of the standard algorithms such as sorting. Datamining stations are quite different from what we are currently used to working with, e.g., workstations and transaction databases. There are huge opportunities for parallelism because these problems tend to be embarrassingly parallel, and ideal for shared-nothing architectures.
2. **Bandwidth**: Although each datamining station has plenty of space, and there are only very minor communication requirements, communications are still a major problem. How do we conveniently ship terabytes of data? CD-ROMs are too small. Networks are too slow. Tape standards are still emerging. All of this could change, but what should we do in the meantime? Compression?
3. **User Interfaces**: Data workstations need to appeal to a mass audience (mostly non-programmers). The query language should be as easy to use as a spread sheet. Datamining projects have failed because they were too demanding on the users. Although experts can write queries that do useful things and do well on benchmarks, real users can't. Bugs are a reality, especially for naive users. Queries don't work. They don't do anything useful, and they consume infinite resources in the process. Success depends on typical performance, not expert performance. Desiderata: Make it easy/possible for analysts (non-programmers) to summarize months/years of purchasing records in a couple of hours. If they think of something they want to look at in the shower, they should have some preliminary results before lunch (if not sooner).

## 11.7 Geoff Crowley and Christopher J. Freitas

"*Multidisciplinary Space Science Simulation in a Distributed Heterogeneous Environment*"

Dr. Geoff Crowley, Dr. Christopher J. Freitas, Aaron Ridley, David Winningham, Richard Murphy, and Richard Link
Southwest Research Institute, San Antonio, TX 78238-5166
E-mail: crowley@picard.space.swri.edu
URL: http://espsun.space.swri.edu/spacephysics/home.html

The Earth's space environment includes the middle atmosphere at altitudes of about 50 km, through the upper atmosphere and magnetosphere, to the solar wind and ultimately to the Sun itself. The solar wind streams out from the sun at speeds of several hundred kilometers per second, carrying energetic particles and its own magnetic field (the Interplanetary Magnetic Field, IMF). The magnetosphere (the magnetic cavity surrounding the Earth) protects the atmosphere from direct bombardment by most of these particles. The magnetosphere accumulates energy through its interaction with the solar wind. Periodically, the energy is released in the form of particle precipitation, strong electric fields and large currents flowing in the ionosphere at high latitudes. The particle precipitation causes bright aurorae in both the Northern and Southern hemispheres. The large currents produce perturbations in the magnetic field measured at the Earth's surface, resulting in the phenomenon known as a magnetic storm. The large amount of energy and momentum deposited in the thermosphere and ionosphere during magnetic storms changes the global circulation patterns, density, temperature and composition above about 100 km altitude. Some of the most energetic particles penetrate to lower altitudes, where they alter the mesospheric and stratospheric chemistry.

The term "space-weather" refers to conditions on the sun, in the solar wind, magnetosphere, ionosphere, thermosphere, and mesosphere, that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life or health. The Nation's reliance on advanced technological systems is growing exponentially, and many of these systems are susceptible to failure or unreliable performance because of extreme space-weather. Adverse conditions in the space environment can cause disruption of communications, navigation, electric power distribution grids, and satellite operations, leading to a broad range of socio-economic losses. The National Space Weather Program (NSWP) is a new initiative designed to address many of the unresolved aspects of space weather (including theory, modeling and measurements) in a unified manner. The NSWP initiative is jointly funded by the Air Force, Navy, NSF and NASA. One of the goals of the NSWP is to produce weather forecasts for the various regions of space ranging from the sun to the Earth's middle atmosphere. The NSWP has the potential to make important contributions to the safety and reliable operation of many technological assets ranging from communication satellites to transformers on industrial power distribution grids. The NSWP needs new capabilities in modeling if it is to achieve its forecasting goals.

Southwest Research Institute (SwRI) is developing a space weather model spanning the mesosphere, ionosphere and thermosphere. The new model is based on an existing computer code which runs on CRAY Supercomputers at the National Center for Atmospheric Research (NCAR). This code is called the Thermosphere-Ionosphere-Mesosphere-Electrodynamics General Circulation Model (TIME-GCM), and is widely acknowledged as the premier space weather code in existence. The TIME-GCM predicts winds, temperatures, major and minor composition, and electrodynamic quantities globally from 30 km to about 500 km. It does this by solving the momentum, hydrostatic, energy and continuity equations with the appropriate physics and chemistry for each altitude. The current model uses a fixed geographic grid with a 5 x 5 degrees horizontal resolution, and a vertical resolution of half a pressure scale height. This fixed grid size restricts the number of geophysical problems which can be addressed with the NCAR-TIMEGCM. The code is being modified at SwRI to run in a distributed parallel computing environment, and will be enhanced by providing a variable grid size capability to permit mesoscale science problems to be addressed that are not within the current model's capability.

The Research Initiative Program in Advanced Modeling and Simulation (RIP-AMS) was an inter-divisional collaboration at SwRI which resulted in the enhancement and expansion of SwRI capabilities in high performance parallel computing. The RIP-AMS program resulted in parallel computing techniques which permit significant improvements in the runtime of computer codes. Specifically, algorithms based on domain decomposition strategies have been developed, providing a framework which will be applied to the TIME-GCM code, allowing a natural method for parallelization and incorporation of variable grid size regions. The message passing control system used in the RIP-AMS is based on the Parallel Virtual Machine (PVM) library. Here, PVM provides a control environment which seamlessly couples together different workstations, providing a common data communication format and system control over the message passing process. The SwRI heterogeneous parallel virtual machine consists of 40 high end workstations connected by a 155 Mb/s fibre-optic ATM network.

We are also planning to couple the SwRI parallelized TIMEGCM to models of the inner magnetosphere and the solar wind located at different locations, extending the virtual machine across the continental US. The challenges of these ambitious projects will be outlined.

## 11.8 Judith B. Cushing

"*Mediated Model Coupling*"

PROF. JUDITH B. CUSHING
The Evergreen State College
Olympia, WA 98505
E-mail: judyc@evergreen.edu
URL: http://192.211.16.13/individuals/judyc/home.htm

Advances in computer science have revolutionized research in the natural sciences. The scientific community now routinely relies on simulations involving complex computations on huge, multi-dimensional data. Many of these simulations have become valuable research tools in understanding isolated phenomena and, thus, constitute the important but small, first steps towards an ultimate goal of understanding complex interactions among physical processes.

To achieve this ultimate goal, existing models must be coupled into larger simulations. The support needed for this coupling goes significantly beyond traditional mechanisms for interface specification, input/output format conversions, and interprocess communication. The scientists, concerned about the scientific viability of the coupled models, must be given the tools to dynamically explore model correlations and relationships at a very high, domain-specific level, shielding them from the details of high-performance computing while supporting the translation of their hypotheses into real, working systems for evaluation.

Computational infrastructure to support this class of activity will involve considerable new research. Jan Cuny, Al Malony and Doug Toomey at the University of Oregon, Lawrence Snyder at the University of Washington and myself are working towards developing such support, which we call "mediated" model coupling.

My own role in this work will be for database support for domain specific computational environments. I think of this as being a persistent store, not only of empirical data that can be used as inputs to the programs, and the program inputs and outputs, but also the application profiles themselves. An application profile includes:

- Application semantics (signature) - inputs/outputs/parameters,
- Application function - such that one familiar with the domain can understand what the program does,
- Methods (transformations) between domain specific data types (defined below) and the application inputs (and parameters, if any of those require calculation from other data).
- Installation-specific information (optional) - which provides a place for the environment to look to determine where the program resides.
- Process information - which could be used to store inputs, parameters, and outputs of specific invocations of an application (i.e., a computation), as well as current status

and restart information.

There are many interesting CS (DBMS) questions that are more widely applicable than to this domain, or even just to computational science:

- How to characterize programs (and maybe processes) as data (objects), so that the DB can hold information about them.
- Semi-structured queries (parsing) of somewhat formatted text, e.g., program output for input to the DB (so that one can do queries and other operations on this data). The data is not necessarily well behaved, and the query or parsing may have to be mediated by a real person. One would also like a domain specific query language for the database, so that it can generate inputs to programs in the right format.
- Initially, we assume that the applications are black boxes, but eventually might want to define the domain specific database interface as a pattern to which application programmers (domain scientists) can write, such that newly interfaced computational applications could take inputs from the database and write outputs to the database, without the programmer of the application being a database programmer.
- More easily extensible schema languages would be critical for such an infrastructure, so that specializations to existing data types can be added to the schema, or one data type can be replicated across different applications.

## 11.9 Jack Dongarra and Terry R. Moore

"*Computational Grids — The New Paradigm for Computational Science and Engineering*"

Prof. Jack Dongarra and Dr. Terry R. Moore
University Of Tennessee, Department of Computer Science
Innovative Computing Laboratory
Knoxville, TN 37996-4030
E-mail: terry-moore@utk.edu
URL: http://www.cs.utk.edu/∼tmoore/

There is the evidence on a number of different fronts that we are entering a new stage in the evolution computational science and engineering. Fueled by constant advances in computing and networking technologies, the research community has begun to develop and implement a new vision for science and engineering that aims to accelerate progress by focusing on computer generated models and simulations of unprecedented scope and power. Well funded prototypes for this new, simulation-intensive approach are already under development in NASA, NSF's Partnerships for Advanced Computational Infrastructure (PACI) program and DOE's ASCI program. What these initial efforts reveal is that this new way of structuring the work of the science and engineering community requires a corresponding change of paradigm in the use of computational resources. The terms "information power grid," "computational power grid," or just "computational grid" have gained currency as interchangeable names for this new paradigm.

The concept of a computational grid is intended to capture the idea of using high-performance network technology to connect hardware, software, instruments, databases, and people into a seamless web that supports a new generation of computationally rich problem-solving environments for scientists and engineers. The underlying metaphor of a "grid" refers to distribution networks that developed at the beginning of the twentieth century to deliver electric power to a widely dispersed population. Just as electric power grids were built to dependably provide ubiquitous access to electric power at consistent levels, so computational grids are being built to provide access to computational resources that is equally reliable, ubiquitous, and consistent.

The most important aspect of computational grids, however, is their ability to provide on-demand concentrations of the massive computational and information resources required for simulation-intensive research. The general problem is that this escalated use of computer generated modeling and simulation requires a tremendous amount of resources of all kinds. The scientific problems under investigation are computationally intensive, data intensive, and can require real-time interactions of various kinds as well. To solve such problems, tremendous computational power has to be combined in a timely way with highly skilled research teams, instruments, simulators, and large databases. But for a variety of powerful economic and practical reasons, these elements are almost always dispersed geographically, so that the only feasible way to bring them together in the required way is through innovative applications that build on high-performance networking. In other words, this approach to science requires the formation of computational grids.

Our group has several on-going projects that can help to address the needs of this new approach to computing. Two projects that are especially pertinent:

- NetSolve (http://www.cs.utk.edu/netsolve/): NetSolve is highly modular and composable middleware for building computational Grids. It is designed to transform disparate computers and software components into a unified, easy-to-access computational service. It aggregates the hardware and software resources of any number of computers that are loosely connected across a network and offers up their combined power through client interfaces that are familiar from the world of uniprocessor computing. NetSolve's modular client-agent-server architecture has at least three key advantages: (1) NetSolve already supports a variety of widely used and familiar client interfaces; (2) The NetSolve server pool can be arbitrarily large, heterogeneous and distributed; and (3) NetSolve's intelligent agents create a modular environment where both new resources and new features are easy to add.
- Repository in a Box (http://www.nhse.org/RIB/): Building on our work with NetLib (http: //www.netlib.org) we are participating in the development and deployment of Repository in a Box (RIB). RIB is a software package for setting up and maintaining software repositories has been developed by the National High Performance Software Exchange Technical Team. For examples of repositories that use RIB, see PTLIB (http://www.nhse.org/ptlib/) and HPC-Netlib (http://www.nhse.org/hpc-netlib/). Using RIB will allow your repository to interoperate with other NHSE repositories, as well as with a growing number of other software repositories that use the Basic Interoperability Data Model (BIDM, http://www.nhse.org/ RIB/help/help.html#data_model), which is the IEEE standard for interoperable software cataloging on the Internet. If you want to set up a repository at your site using RIB, you can get the a copy of the software (http://www.nhse.org/RIB/software/) and register (http://www.nhse.org/RIB/ interoperate/) your repository so that other repositories can interoperate with you.

We believe that the ongoing development computational grids will play a central role in the work of scientific and engineering community over the next decade, and therefore, that the emergence of computational grids will have a dramatic impact on key issues for the DICPM workshop. The impact of the grid and of grid-based, simulation-intensive research on the world of scientific database management relates primarily to two key factors: (1) the much more fluid and dynamic characteristics of information locality allowed by grid-based computation; and (2) the higher level of compositionality required in software systems that can take advantage of the power and flexibility that computational grids will permit. Among more specific issues are the following:

- Ability to cache and replicate data in a convenient and transparent way without compromising consistency.
- Need for language independent characterization of computation and data types.

- Need to be able to generate good model of the dynamic topology of a given grid, including the loads on its network and compute nodes.
- Need to establish standards for the integrity and quality of grid-based computations, i.e., for computations that use computational components made available on the grid.
- Ability to provide real fault-tolerance in grid-based computing environments.

## 11.10 Pierre Elisseeff, Nicholas M. Patrikalakis, and Henrik Schmidt

"*Computational Challenges in Ocean Acoustics*"

Dr. Pierre Elisseeff, Prof. Nicholas M. Patrikalakis, and Prof. Henrik Schmidt
Massachusetts Institute of Technology, Department of Ocean Engineering
Cambridge, MA 02139-4307
E-mail: elisseef@mit.edu, nmp@mit.edu, henrik@keel.mit.edu
URL: http://dipole.mit.edu:8001, http://deslab.mit.edu/DesignLab/

The past two decades have provided an extremely fertile ground for computational ocean acoustics. Thanks to unprecedented developments in the availability of affordable computer power, the field has evolved from being a mere analytical branch of wave propagation physics to being a mature numerical discipline with original contributions of its own. The current computational techniques can be sorted, for the sake of simplicity, into three main branches: propagation models, scattering models, and inversion methods. A brief review of the more numerically mature techniques, namely propagation models and to some extent inversion methods, follows. Pending and upcoming computational challenges are then discussed.

Acoustic propagation numerical models rely on the same underlying multi-dimensional Fourier transform approach to the wave equation. Once the implicit analytical solution to the wave equation is put in integral form, the numerical scheme chosen to perform the integration will determine the type of propagation code: the stationary phase method yields the ray approach, the method of residues yields the normal mode approach, the direct trapezoidal integration rule yields the wavenumber integration approach. The parabolic equation approach stands aside, being based on a factorization of the wave equation differential operator. The reader is referred to [2] for a full discussion. All four classes of codes correspond to specific, partially overlapping regimes determined by parameters such as frequency, range-dependence, elasticity. Provided accurate environmental information is supplied, the appropriate numerical model will predict the acoustic field within a fraction of a decibel. Inversion techniques aim at measuring environmental or source parameters given some acoustic measurements. They can be viewed as a special class of data assimilation algorithms. The first inversion technique, Ocean Acoustic Tomography, is an adaptation of medical and seismic imaging schemes; it relies on a perturbational approach and assumes a good *a priori* knowledge of the environment [3]. It is the acoustic equivalent of objective analysis techniques used by oceanographers. The other inversion technique, Matched -Field Processing (MFP), leverages the accuracy of modern propagation models by computing a large number of simulated signals assuming different environmental parameters [4]. It then identifies the synthetic signal most closely correlated with the measured one, therefore inferring what the true environment is. MFP has proven to be a very fertile research ground, providing a truly original approach to non-linear inversion problems such as source localization and environmental estimation.

Computational ocean acoustics now stands at a threshold in the history of the discipline. Having to a large extent solved the forward (prediction) problem for a given deterministic environment, it is now severely limited in its ability to accurately represent, store and use complex environmental information for increasingly relevant scenarios such as coastal and basin-scale operations. Further computational bottlenecks include MFP and scattering studies. Environmental uncertainty is dealt with through the use of stochastic formulations, which require significant computational resources. On the other hand, acoustic waves are the only means of remotely sensing the depths of the ocean, and can provide oceanographers and students of the ocean in general with significantly more field data than what is currently available. Much stands to be gained by coupling acoustic and oceanographic models across disciplinary boundaries. The latter will gain access to vast amounts of untapped data and improve current ocean forecasts. The former will gain the much need environmental information necessary to understand and quantify propagation mechanisms and ultimately improve signal processing performance. Some of the computational challenges ocean acoustics will face in order to achieve this multidisciplinary convergence are:

- **Acoustic data assimilation**. As argued above much stands to be gained by coupling acoustic and oceanographic models. It will enable concurrent progress in both disciplines. A number of issues must be resolved before even tackling the assimilation problem. Data formats and databases must be reconciled, possibly in a distributed manner in order to preserve issues of data ownsership and maintenance. The mutual resolution and sensitivities of acoustic and oceanographic models must be understood. Acoustic models, to take the simplest example, operate on a smaller scale and can be very sensitive to the sub-grid numerical noise of oceanographic model outputs. In order to achieve proper model coupling, the transfer of signal and noise between models and scales must be investigated using an adequate stochastic framework. Once models are coupled, efficient algorithms will be required in order to handle the increased computational complexity as well as the significantly increased amount of data being assimilated.
- **Object-oriented computing**. All numerical modeling is currently done on a procedural basis, most often in FORTRAN. This separation of data and processing functions, while very convenient for the direct implementation of relatively simple numerical algorithms, has major and known defaults [1]. An object-oriented methodology must be adopted when combining cross-disciplinary computational modules in order to contain unavoidable bugs and build reliable and maintainable multi-disciplinary computational systems integrating the different physical mechanisms at play, as is the case in ocean data assimilation. The use of object-oriented technique will further benefit the development of hybrid three-dimensional propagation modeling techniques by facilitating the combination of existing separate computation modules such as wavenumber integration and boundary elements. Finally, the use of features such as operator overloading enables the use of a meta-modeling language, whereby the scientist can optimally merge two data sets by simply "adding" them and is then free to focus on scientific questions.
- **Distributed computing**. Given the increasingly multidisciplinary nature of ocean sciences including acoustics, future ocean observing systems will combine a multitude of heterogeneous assets which are distributed in nature. While acoustic data might be acquired on a given node of a generalized observation network, a mobile au-

tonomous underwater vehicle (AUV) might acquire *in situ* data at a different location, and oceanographic data might be gathered at a third location. AUV navigation as well as real-time data assimilation will entail a large amount of acoustic computations which some nodes, such as the AUV, might not be able to handle. One can then envision a distributed system in which a number of modeling computational tasks are performed on a service basis by the relevant nodes. This approach is also beneficial to MFP, which is an inherently parallel approach. Distributed computing will require some measure of encapsulation readily provided by object-oriented techniques.

- **Software metadata**. In order to facilitate large scale distributed computing as well as cross-disciplinary interaction among the various ocean sciences, knowledge about existing numerical prediction codes is required. It must list and quantify the actual modeling assets available to the community as well as guide the choice of a potentially inexperienced user. Software metadata would then assist in the localization of all existing modeling assets relevant to a given operational problem, for instance in the course of a real-time environmental assessment exercise. The appropriate model would then be used for the assimilation of data into a complete multi-disciplinary ocean model.

1. Buzzi-Ferraris, G., *Scientific C++ – Building Numerical Libraries the Object-Oriented Way*, Addison-Wesley, 1993.
2. Jensen, F. B., Kuperman, W. A., Porter, M. B., and Schmidt, H., *Computational Ocean Acoustics*, AIP, 1994.
3. Munk, W., Worcester, P., and Wunsch, C., *Ocean Acoustic Tomography*,Cambridge University Press, 1995.
4. Tolstoy, A., *Matched Field Processing for Underwater Acoustics*, World Scientific, Singapore, 1993.

## 11.11 Zorana Ercegovac

"*Resources Metadata+User Metadata: Toward Knowledge Metadata*"

Prof. Zorana Ercegovac
University of California, Los Angeles
Department of Library and Information Science
Los Angeles, California 90095-1521
E-mail: zercegov@ucla.edu
URL: http://www.gslis.ucla.edu/LIS/faculty/zercegov/ercegovac.html
http://www.lainet.com/infoen/

### Context

In order to enhance the organization of digital resources on the Internet, it might be useful to look at some of the traditional features that have been successful in organizing and describing printed resources in bibliographies and periodical indexes. Bibliographies are typically compiled and annotated by experts in a given field (e.g., *Bibliography of American Imprints to 1901*; *The Law-of-the-Sea: a Bibliography*). Periodical indexes (e.g., Engineering Index, Inspec) are designed to provide timely access to the periodical literature, papers presented at conferences, and parts of books. These features include Preface, Intended Users, and Instructional model of how to use a certain tool; together, they provide a rich layer of information about the collections and its uses. We believe that these techniques have not been sufficiently exploited in the context of designing metadata standards for describing large heterogeneous distributed datasets on the Web; their conceptual simplicity and complementarity to the model of library catalogs might help the user see the inner structure of the collections better than it has been possible so far, and increase the probability that only relevant objects from these collections will be retrieved.

### Content metadata is not sufficient: we need user metadata

Library catalogs have served well as a primary model for many metadata standards. However, while they offer guidance in representing the contents of a given collection, they fall short of representing the user or different classes of users of virtual collections. Library catalogs, however, make several assumptions; the collection is well defined and available locally; the user typically comes to the library; and reference librarians are there to help and instruct. So, library catalogs do not have preface to their collection (note: online catalogs that are accessible remotely do), statement of intended beneficiaries, or instructional guides to teach people how to use catalogs; users' guides are relatively novel phenomenon and limited to online catalogs. Unlike library catalogs, bibliographies and periodical indexes are not limited to any particular physical collection of objects; their readers are more heterogeneous than those who use library catalogs; and typically, there are no intermediate experts to teach the user how to use a particular tool. As such, bibliographies are closer to the nature of metadata that represent digital virtual libraries than they are to library catalogs that represent local holdings. Therefore, bibliographies define the extent of their scope and domain (explicated in "preface"); specify user classes (via "intended user"); and describe effective ways for discovering resources in virtual collections. Drawing on the model of primarily library catalogs, current metadata standards have attempted to provide the descriptive, locational, and to a certain degree collocational power. The inclusion of preface, use, and usage could extend the functionality of metadata standards beyond the objectives of traditional library catalogs; the value-added functions are contextualization, appraisal, quality control and preservation, legal control, multiple access control, and, in particular, the user complex layer.

### Case study: Putting the user in center

The user complex layer is being incorporated into Learning Portfolio for Accessing Global Engineering Information in User-Centered Environments, a research project that is currently being investigated by Ercegovac and sponsored by the Engineering Information Foundation. The following three features are explicated in the project.

- **PREFACE metaphor**. Physical library is self-describing in many important ways. It is confined with physical boundaries, and typically contains charts, signage, reference librarians, a catalog, an information (smart) kiosk, and a myriad of discovery tools each designed for different purposes and user needs. Today's digital libraries lack this layer of information that shows content, context, structure, capabilities, and limits of collections. What is needed is a "preface" describing the scope, domain, and how collections and objects relate to other collections and objects.
- **USER metaphor**. So far metadata standards have directed their attention to describing content analysis rather than user characteristics. We need attributes containing time-generated profile of users along with their capabilities, preferences, and path models which focus on usage histories rather than content analysis alone.
- **INSTRUCTIONAL metaphor**. People continue to have numerous difficulties even with graphical user interfaces and other advanced capabilities; many powerful tools go little or rarely used. In other words, it is not only important that tools exist; as important is the deployment of effective literacy guides.

### Seamless metadata

People, by definition, are processors of multimedia inputs. We are engaged in processing audio, graphical, animation, textual and numerical inputs. However, conceptually different metadata standards have been created independently to describe relatively homogeneous data sets within different disciplinary traditions. For example, the library community designed the Anglo-American Cataloging Rules / Machine Readable Cataloging (AACR2/USMARC) as a metadata standard that was initially and predominantly used to describe books; while the standard was later extended to include non-book formats such as rare books and manuscripts, cartographic materials, graphical objects, and sound recordings, many non-book formats were represented with book- related attributes. Similarly, other traditions have developed their own metadata (e.g., Encoded Archival Description (EAD), Federal Geographic Data Committee (FGDC), Text Encoding Initiative (TEI), and Visual Resources Association (VRA). Only recently have we seen the attempt to, for example, mesh FGDC Content Standard for Spatial Metadata with the AACR2/MARC Format.

The Dublin Core with its guiding principles represents a new direction that accommodates multiple data types into a single underlying super-structure model. What we need next is to incorporate user profiles into user- centered metadata standards. One of the early attempts in this direction is the Learning Portfolio project which contains multiple layers of learning tools that reflect learning stages of different user groups in different engineering disciplines. Some of the research questions pertain to vocabulary control (e.g., semantic clusters of, for example, the author concept, which may be known as a creator, responsible party, provenance, originator, etc., in different metadata); ability to locate resources throughout their lives; deciding what should be described, who would describe, and to what extent, to mention just a few issues before us.

## 11.12 Gregory L. Fenves

*"A Computational and Communication Infrastructure for a High Performance Seismic Simulation Network"*

Prof. Gregory L. Fenves
University of California, Berkeley
Pacific Earthquake Engineering Research Center
Berkeley, CA 94720-1710
E-mail: fenves@ce.berkeley.edu
URL: http://www.ce.berkeley.edu/~fenves/

**Background**

Earthquake engineering is a multi-disciplinary field that creates technology to improve the seismic performance of civil systems (buildings, transportation and lifeline infrastructure, and industrial facilities). The range of scientific and engineering issues addressed in earthquake engineering is broad, ranging from fundamental behavior of advanced materials, earthquake response prediction and modification of soils, structural behavior and control, design methodologies, and public policy and economic decision making for seismic risk reduction.

Research in earthquake engineering is now oriented towards performance based engineering design. The objective of PBED is to provide a rational framework for owners and decision-makers to understand the tradeoffs between lifecycle cost and seismic performance. In this context, seismic performance is a complex and not yet fully defined metric of the probability that a system will meet the expected behavior in future earthquakes. In contrast with current design methodologies, which primarily protect life but not necessarily the system itself, PBED provides options for performance up to continued operation during an earthquake. Whereas understanding of system behavior associated with collapse of structural systems, important for protecting life safety, is fairly well established, the earthquake engineering research enterprise must be re-oriented to address higher levels of system performance such as limited damage. For this workshop, the tools for new research can be categorized as two broad initiatives:

1. High fidelity computational simulation of seismic performance
2. Advanced experimental simulation of seismic performance

The two initiatives are closely linked; one cannot be effective without the other. The computational and information needs for the simulation of realistic performance of soil and structural systems are addressed in a companion paper by G. Turkiyyah. The second initiative is addressed in this paper. These two initiatives form the basis for a proposed networking of earthquake engineering experimental facilities in the US. The earthquake engineering research community is working with the National Science Foundation to create a National Network for High Performance Seismic Simulation.

**Communication Infrastructure for the Seismic Simulation Network**

The seismic simulation network will consist of a distributed physical testing facilities and computational resources. Testing facilities will include loading and control systems for simulating the deformations of large structural and geotechnical components and subsystems; earthquake simulators ("shaking table") to dynamically test subsystems and system models; centrifuges for testing soil systems with scaled gravity; and other testing methods under development (some described below). The utility of this "network" of facilities will be enhanced by a communication infrastructure. The objectives of the communication infrastructure are to: (1) allow collaborative design and conduct of experiments among participants in different disciplines and geographically dispersed; (2) disseminate and "mine" information from data that are expensive to collect; (3) provide opportunities for advanced testing methods; and (4) allow integration of physical testing with simulation by computational methods. Some of the scientific issues to be dealt with in the communication infrastructure are:

- **Heterogeneous data**. Testing data is very diverse and unique for each experiment. It can range from sampled analog data from instrumentation for displacement, acceleration, temperature, and other physical phenomena, to optical, acoustic, electrical field, and other emission data. High definition video images may be used in the future, with possibilities for direct measurement of some data from the images (e.g., crack formation, deformation fields, etc.), that bring in issues of image recognition.
- **Advanced testing methodology**. It is not possible to conduct an experiment on a large scale structure and its supporting foundation and soil. New developments in on-line control of dynamic simulation will be an important feature of the network. Advanced communication will enable testing different components in different laboratories with the linking of experimental facilities. Components representing parts of a large system can be tested in different laboratories. One example is simulating the dynamic response of structure on a shaking table and the foundation in a centrifuge. The interface forces and deformations, with some processing, is transmitted between the controllers for the two testing facilities. Other possibilities may include using advancing sensing instruments, such as electron scanning microscopes, to observe phase changes, flaw propagation, and other changes in materials during severe cycles of deformation during a simulated earthquake. New testing methodologies will depend on a low latency, high reliability communication networks and protocols, in addition to development of the testing technology and methods themselves.
- **Computational technology**. The Turkiyyah position paper outlines the scientific needs for improving the computational simulation of earthquake behavior of civil systems. In regards to testing, a promising area is hybrid methods in which a portion of the system is modeled computationally and another portion is modeled and simulated physically in the laboratory. This is done now to simulate dynamic behavior in a "slow" loading environment. With improved computational speeds, communication speed, and new control systems, this can be done dynamically (which will avoid some of the scaling and velocity dependent problems with slow testing).
- **Visualization**. Another area which needs more attention in earthquake engineering simulation is visualization. The massive amounts of data generated during a test or

computational simulation cannot be appreciated or fully understood with the current methods for plotting data. Much of what is interesting during a test is not visually observed. However, collected data (e.g., strain fields) can be visualized in real and simulated time. Visualization tools are needed that closely replicate the physical processes taking place during a test.

- **Collaborative environments**. As mentioned previously, a distributed experimental network will require a collaborative communication infrastructure to be effective. The environment should facilitate test specification, design, and conduct. It should allow participants to customize their view of the test and provide interfaces to specialized software.
- **Data protocols**. Issues of quality assurance and control, ownership of data (intellectual property), liability, are not only technical issues but relate to how data are ultimately used in the design of buildings, transportation systems, and other civil systems. Society has high expectations that the data used in the design is reliable and interpreted correctly. A communication infrastructure for seismic simulation must go beyond exchanging uncorrupted data to providing mechanisms to assure the quality of the data and its interpretation.

## 11.13 Paul J. Fortier

"*Advanced Fisheries Management Information System (AFMIS)*"

Prof. Paul J. Fortier
University of Massachusetts Dartmouth, Electrical and Computer Engineering
North Dartmouth, MA 02747
E-mail: pfortier@umassd.edu
URL: http://www.ece.umassd.edu/ece/faculty/pfortier/pfortier.htm

Ocean systems science seeks a regional and ultimately a global scale understanding of the components, interactions, and evolution of the complete ocean system, including the physical, biological and fish components. The raw data supporting ocean system sciences is rapidly growing per year and shows no signs of abating. The intrinsic problem of archiving, searching, and distributing such hugh data sets are compounded by the heterogeneity of the data and the generators of the information. In addition to these issues, users of information come from a variety of scientific disciplines with differing views and requirements for this information. A successful ocean data management system must provide seamless access to arbitrary local and remote data sets, and be compatible with existing and evolving data analysis, data location and data transfer environments such as HOPS, CORBA or the web, if it is to be available and useful to a diverse group of scientists

Technical and scientific activities such as fishery management within the seas, require knowledge of the multi-dimensional spatial distributions of the oceans and embedded biological properties along with related time dependencies. An understanding of these multi-dimensional fields and their relationships is essential to further knowledge of this domain's dynamic nature. Field estimation in the ocean domain is very complex and made even more difficult due to the sparseness of information available. In order to make continuous, accurate, and complete oceanic field estimates requires more complete data sets describing the measured ocean and its inhabitants. To develop these data sets requires the collaboration of experts from a variety of disciplines; oceanography, physics, electrical engineering, computer engineering, computer science, ocean biology, chemistry, and geology, each providing insights on their unique understanding of the ocean environment and technologies applied to ocean research. The primary outcome of such interaction is an integrated information model of the multi-dimensional ocean domain.

The ocean sciences area of interest to our research is the inhabitants of the commercial fisheries. The present situation in fishery management is deemed prehistoric at best. The criticism may not be fully warranted, but the present situation is not adequate, nor time effective for the participants, as well as fishery policy and management agencies. Present fishery data, for end participants, is sparse to non existent, and for fishery management units it is not much better. Typical fishery management units use data relating to single populations of fish, which consist mainly of aggregate, non-validated data, collected over some long temporal period, typically an entire year. Due to this, fishery management decisions are generally referable to the regulation of size or age specific fishing mortality covering long temporal periods. Improvements in data capture, timeliness, assimilation and presentation can greatly improve the performance of present fishery management and the lives of those germainly dependent on sound and timely policy and management directives.

Fishery management and fisherman's commercial operations in the seas depend on the prey they both seek. The populations of fish are interdependent, so it makes sense to manage them in the context of all of the species involved and the environment they exist within. In addition it would be desirable to manage a single population or multiple populations within time scales that are much shorter than the present yearly basis. In fact the most desirable outcome would be to have the ability to model and therefore manage fisheries much like we monitor atmospheric weather today, in near real-time fashion. Beyond the fisheries, another important component within a fisheries management and prediction system is the environment. The environment changes over a spectrum of time scales, and these multiscale environmental interactions are important sources of information that contribute to fish mortality and fish population dynamics. The atmospheric, physical ocean, chemical ocean, geological, biological and fish information must be extracted, assimilated into the database, modeled and mined for knowledge when forecasting fisheries.

There are several problems with present fishery data management as introduced above. The first deals with the inadequacy of an integrated data model describing the multimedia environmental, biological and fisheries information which in turn dictate the usefulness of this data to diverse user groups. A second problem, deals with present environmental and in particular fishery data and data collection processes. The fishery data is based on sparse, temporally sampled information which is not at present highly correlated, nor validated to physical systems data, for example a complete spatial and temporal ocean model. The third problem deals with the lack of a comprehensive data model for integrating numerous, diverse non homogeneous, multimedia data sets of the ocean and fishery environment into one meta data model, useful to a variety of research, development and operational organizations. Data collection for many ocean environmental components is periodic (e.g., satellite scans, in situ sensors), while others are aperiodic, sporadic or ad hoc. Many assessments are even only performed on a yearly or bi-annual basis (e.g., fish catches, surveys etc.). The true environmental and biological population effects are continuous and highly dependent on a variety of conditions, implying the inadequacy of the present scheme. Presently data is collected from a variety of sensors and merged into some existing and proposed experimental systems, for example LOOPS, HOPS, DODS, and ultimately AFMIS. Existing systems provide partial data on the ocean environment and hopefully in the future, more complete biological and pollution information, however they do this in forms useful to a subset of interested parties both commercial and governmental. Many of these efforts are only in initial conceptual phases and are also struggling with concepts of how to view this multidimensional data within some all encompassing meta data framework. The SQL3 database standard to be released this spring-summer, provides a model (object/relational) amenable to capturing the variety of meta and real information described, as well as means to uniformly manage and operate upon diverse data sets. In addition numerous vendors, are delivering commercial systems implementing some subset of this language, which can form a base to build the prototype research system and ultimately a commercial system upon making this system more widely available.

An integrated "synthetic ocean" database system which possesses the capability to search out (data mine), collect, integrate and deliver information describing the continuous (or near continuous) nature of the ocean environment and its inhabitants as described above, will provide a platform to develop computer based simulation and modeling tools to aid participants in numerous commercial and policy organizations involved in the oceans. For example fisherman, conservationalist, politicians, fish processors, fish traders, etc,. all have an interest in a more complete understanding of the dynamics which impact the environment they participate in or manage. A database system which provides integrated record, abstract data sets, objects, temporal and spatial data sets can be used to develop a variety of ocean and biological forecasting tools much like those available at present for weather forecasting. These tools and data sets, in turn will spur on further research and development of extensions useful to the full spectrum of interested parties. Such tools would aid DoD, NOAA, NMFS, NASA, EPA, state organizations, educators, students, fishing cooperatives, fish processors, regional fishery councils and ultimately fisherman in better understanding and managing this important earth resource. The developed systems could provide a near real-time information service to parties much as the weather services do today to aid in day to day activities of these organizations.

With the basic concept of an ocean information system, we can then move towards development of additional products that use this information bank, much like what occurred based on the advent of the GPS systems numerous related product spinoffs. One initial possibility is an informational source available for access in real-time by anyone with a computer and communications capabilities. Users could register for specific information and have it automatically down loaded to their site. For example, maps forecasting the best locations for finding flounder over the next few days, or fishing grounds to be shut down to foster spawning and population recovery.

Present research at UMD and CMAST focuses on the development of a state of the art management information system for fisheries management and to support the implementation of a long-term full-spectrum environmental management system for the southern New England coastal waters. The Advanced Fisheries Management Information System (AFMIS), will be used for the real-time management of marine fisheries and to develop a fishery information service. The basic conceptual structure of the system consist of an integrated simulation and sensor system centered on a sensor, assimilation and simulation structure, all supported by an underlying information management component. Ocean, biological, and fish dynamic models form the foundation of the system. The models are driven by satellite, aircraft, extracted legacy data and in-situ sensor data. This ocean database will be supported and augmented by fishing boat and port sampling data acquired from the National Marine Fisheries Service and other commercial sources. The sensor, assimilation and simulation system structure generates an environmental assimilation, which can be thought of as the "computed ocean environment." The computed ocean environment includes; the status of physical ocean properties, chemical ocean properties, geologic properties of the ocean floor, ocean lower trophic biological properties, the status of spatially integrated size-class fish populations and other high trophic levels in the biological food chain.

## 11.14 Christopher J. Freitas

"*Multi-physics Simulations on Distributed Computer Resources*"

Dr. Christopher J. Freitas
Southwest Research Institute, Computational Mechanics Section
San Antonio, TX 78238-5166
E-mail: cfreitas@swri.edu
URL: http://www.swri.org

### A Perspective

Southwest Research Institute (SwRI) has worked in the area of high performance parallel computing for the past several years, starting in the late 1980's. Our work has encompassed MPP class computing, in conjunction with our relationships with the Department of Energy Laboratories (i.e., Sandia National Laboratories and Lawrence Livermore National Laboratory), and Distributed Parallel computing based on Parallel Virtual Machine (PVM), Message Passing Interface (MPI), and high-speed networks. Although, SwRI has extensive experience with traditional MPP class computing, we have focused, from the very beginning on fully distributed parallel computing.

Early on in our research we had extensive interactions and cooperation with nearly every MPP hardware vendor (e.g., Intel, IBM, Convex, Cray, KSR, etc.). All these vendors were offering collaboration and their latest generation of parallel computer for our free use. But, with all these interactions and collaborations we were left feeling uneasy about the direction and focus of the HPC vendor community with regard to parallel computing. It seemed they were still operating in the vector-computing mode in which "big-iron" computers were the norm, and they were focused on getting the most number of CPUs in a single- or multiple-cabinet computer, rather then addressing the true nature of the paradigm of parallel computing. It was and still is the view of SwRI, that the strength of parallel computing is not necessarily in having the next best and biggest machine, but rather to take advantage of the aggregate computing power of existing facilities, supplementing that hardware where required. America and American industry are becoming ever more cost conscious. What they want is the ability to use existing compute cycles, supplemented by moderate cost improvements to achieve improved computational speed. The issue is no longer the amount of speed-up versus number of compute nodes, but rather the amount of reduced computational time it takes to get the compute job done. If one can achieve even a factor of two reduction in wall-time in a real application, this can translate into enormous savings in man-hours and provide potentially more accurate solutions at lower cost.

The validity of our belief has been borne out in a large number of externally funded research programs at SwRI in which distributed parallel computing methods are being used in the service of projects. In particular, SwRI's vision of parallel computing is being refined for NASA applications through our current project on Interactive Meta-Computing Technologies for Industrial Deployment, NASA Contract No. 961000.

**Research Activities**

SwRI has focused on the technologies related to the efficient use of workstation clusters connected by high-speed networks to solve real world engineering and science problems. This research has been based on the message passing paradigm using primarily Parallel Virtual Machine (PVM) and Message Passing Interface (MPI) environments. These tools have been coupled to numerical techniques such as domain decomposition, in which complex geometries are broken into simpler subregions and in which independent grid systems may be created. These independent grid systems or operators are then used to solve, in an ordered fashion, the governing partial differential equations, directly accounting for the internal boundary interfaces that now exist between grid blocks or data structures required by the multiple operators. Coupled to these techniques for the solution of the governing equations has been the development of new methods for scientific visualization in which computational data may be retrieved during the course of the computation for real-time evaluation. The kernel to this visualization system is a data cache for which tools have been developed at SwRI for cache manipulation. In addition to the software methodologies developed for equation solving and scientific visualization, SwRI has also created a second generation ATM Testbed. ATM, Asynchronous Transfer Mode, is a high-speed switch-based network that promises performance levels of more than an order-of-magnitude over standard or switched Ethernet or other competing techniques. The SwRI ATM system connects several clusters of workstations, two MPP computers, one second-generation Distributed Array Processor (DAP) computer, and two multiprocessor graphics systems via multiple FORE ASX-1000 and ASX-200 ATM switches, and dual OC-3c optical links.

A broad spectrum of scientific applications using parallel computing strategies have been developed at SwRI. These application areas include: turbulent, compressible/incompressible, reactive flow in complex geometries, subsurface reactive flow transport, space weather simulation, probabilistic structural dynamics, and large deformation material response modeling. The algorithms that solve these problems use a range of sub-algorithms with different computational characteristics. These sub- algorithms may then be mapped to a distributed virtual computer in which different computer resources are matched to the requirements of the different sub-algorithms. The efficient use of these distributed computer resources requires an understanding of the scalability of the solution process, and tools for dynamic load balancing and fault tolerance and recovery. Tools that perform dynamic resource allocation, in which both the computer and the pipeline connecting it to other resources are monitored for performance in the dynamic balancing of computations, are being developed at SwRI.

## 11.15 James C. French

"*Quality of Information (QoI): Mitigating the Lack of Authority Control*" [2]

James C. French
University of Virginia, School of Engineering and Applied Science
Department of Computer Science
Charlottesville, VA 22903
E-mail: french@cs.virginia.edu
URL: http://www.cs.virginia.edu/brochure/profs/french.html

Linking disparate information resources into a cohesive knowledge base involves solving a number of vexing interoperability issues. Even if we adopt unambiguous standards for required metadata and agree completely on the semantics, we still have a serious problem in resolving individual field values. The quality of information (QoI) is the issue.

Integrated research environments for scientists and engineers will have many problems related to the integration of fielded information. Consider bibliographic records as an example. Here we have fields (attributes) such as title, author names, author affiliations, journal name, date of publication, etc. If we were to attempt to merge multiple bibliographic databases from several sources, we would have to grapple directly with the issue of object identity: when are two records denoting the same object? It might be quite difficult, indeed undecidable in general, to make these associations. However, when domain knowledge is introduced the problem becomes more tractable.

In earlier work [3,2,1,4], we have used lexical methods and clustering analysis for this problem and have had excellent results. However, it has become clear that we are at the limit of what can be achieved by syntactic methods alone. To make further progress we must introduce additional knowledge into the process, first general knowledge and finally very specific domain knowledge. We are now considering approaches involving combining evidence to improve belief and that appears to be promising.

Librarians attempt to solve this problem by use of a controlled vocabulary; they refer to this as authority control. An authority file is used to keep track of all permissible strings allowable in a field. We believe that strict authority control will prove to be impossible in a dynamic information environment and view our approach toward value resolution as mitigating the lack of authority control, a more realistic goal.

1. J. C. French, A. L. Powell, and W. R. C. III, "Efficient Searching in Distributed Digital Libraries," *Proceedings of the 3rd ACM International Conference on Digital Libraries (DL '98)*, Pittsburgh, PA, 23-26 June 1998.
2. J. C. French, A. L. Powell, and E. Schulman, "Applications of Approximate Word Matching in Information Retrieval," *Proceedings of the 6th International Conference on*

[2] This work supported in part by DARPA contract N66001-97-C-8542, NSF grant CDA-9529253, and NASA GSRP NGT5-50062.

*Information and Knowledge Management (CIKM '97)*, pages 9-15, Las Vegas, Nevada, 10-14 November 1997.

3. J. C. French, A. L. Powell, E. Schulman, and J. L. Pfaltz, "Automating the Construction of Authority Files in Digital Libraries: A Case Study," in C. Peters and C. Thanos, eds., *First European Conference on Research and Advanced Technology for Digital Libraries*, volume 1324 of *Lecture Notes in Computer Science*, pages 55-71, Pisa, 1-3 September 1997. Springer-Verlag.
4. E. Schulman, J. C. French, A. L. Powell, G. Eichhorn, M. J. Kurtz, and S. S. Murray, "Trends in Astronomical Publication Between 1975 and 1996," *Publication of the Astronomical Society of the Pacific*, 109:1278-1284, 1997.

## 11.16 Martin Hardwick

"*Information Infrastructure for the Industrial Virtual Enterprise*"

Prof. Martin Hardwick
Rennselaer Polytechnic Institute
President, STEP Tools, Inc.
Troy, NY 12180-3590
E-mail: hardwick@rdrc.rpi.edu, hardwick@steptools.com
URL: http://www.rdrc.rpi.edu/staff/hardwick.html

The National Industrial Information Infrastructure Protocols (NIIIP) Consortium is a team of organizations that entered into a cooperative agreement with the U.S. Government, in 1994, to develop inter-operation protocols for manufacturers and their suppliers.

The NIII protocols make it easier for engineering organizations to share technical product data over the Internet. They do this by building on the STEP standard for product data exchange. STEP provides common definitions for product data that can be read and written by many CAD/CAM/CAE and PDM systems.

If engineering organizations can share data using STEP, then suppliers are no longer constrained to own and operate the same systems as their customers. The potential benefit is large. Suppliers are spending a lot of money buying the systems of their customers, more money training people to use those systems, and even more money re-entering data from their preferred systems into the system required by each customer.

The protocols selected and developed by the NIIIP Consortium have been validated in three end-of-cycle demonstrations. In each cycle, a team with expertise in technical product data, object modeling, workflow management, security, and knowledge representation came together and demonstrated how technical barriers to the dynamic creation, operation and dissolution of "virtual enterprises" are overcome by the NIII protocols.

The challenge problem for the NIIIP project was to make it possible for a client program to perform an operation on technical product data over the Internet. In Cycle 1 the operation was "CAD visualization." In Cycle 2 the operation was "engineering change." In Cycle 3 the operation was "create assembly." To enable these operations, the NIIIP protocols had to make it reasonable for the client application to access product data across the Internet, modify product data across the Internet and to integrate product data across the Internet.

STEP defines an integrated suite of information models for the life cycle of a product. The information models are defined in a language named EXPRESS that is similar to other data definition languages, but unique because it can describe the complex data structures and constraints needed for properties such as geometry and topology. Today, STEP information is usually exchanged using files. However, STEP information can also be shared using an EXPRESS-driven programming interface called the STEP Data Access Interface (SDAI). The SDAI lets an application program access STEP data in a repository. When the NIIIP program began, the repository and application program were assumed to be operating in the same process.

Four NIII PD protocols have been layered on the STEP standards. The first two are being made into new ISO STEP standards, and the fourth is being made into an Object Management Group (OMG) standard. An appropriate standardization body has yet to be selected for the third protocol.

1. SDAI Java and SDAI IDL are programming environments that let applications access and manipulate product data stored in remote repositories over the Internet [1, 2].
2. EXPRESS-X is a mapping language that creates application-specific objects out of the neutral data defined by STEP [3].
3. STEP Services is a set of operations for cutting and pasting objects defined by EXPRESS-X between data sets [4].
4. The OMG Product Data Management (PDM) Enabler defines a programming interface for servers that integrate product assembly information [5].

The SDAI Java binding lets applications manipulate product data using World Wide Web browsers. The SDAI IDL binding lets applications that are written in different programming languages manipulate product data over the Internet. With these two bindings, an application can open a repository, find data, and transfer it to a remote client. The repository can be defined by one of the EXPRESS models defined by STEP or another EXPRESS model.

EXPRESS-X is a language for mapping data between EXPRESS models. It can be used to map data between the application-neutral EXPRESS model that defines a STEP standard, and an application-specific model. The STEP models are designed to support operations that a large range of enterprises might want to perform against a type of product. An application-specific model is one that can handle the more limited range of operations needed by a single application.

The STEP Services protocol defines data transfer and integration operations. With these operations, an application can integrate objects from multiple sources to create a database. The services define how an application can traverse a set of objects, find the ones it needs, copy them to another data set, and integrate those copies with the existing objects in the database.

The PDM Enabler is a protocol that defines programming interfaces for product data management. Part of the PDM Enabler defines a standard programming interface for servers that manage product assemblies. OMG plans to define similar standards in many other areas, including Manufacturing Execution Systems (MES) and Enterprise Resource Planning (ERP). These standards will be useful for servers that need to manage manufacturing execution and planning information.

When the protocols are complete they will make it easier for enterprises to share technical product data over networks. Programs such as the Boeing 777 design project have shown the value of product data sharing using a homogeneous database. The NIIIP protocols should allow teams of enterprises to share product data from distributed, heterogeneous product databases.

1. *Industrial Automation Systems. Product Data Representation and Exchange. Part 27, Java Binding for the Standard Data Access Interface*, ISO 10303-27, (http://www.nist.gov/sc4/ stepdocs), International Organization for Standardization, Geneva, Switzerland, 1998.
2. *Industrial Automation Systems. Product Data Representation and Exchange. Part 26, IDL Binding for the Standard Data Access Interface*, ISO 10303-26, (http://www.nist.gov/sc4/ stepdocs), International Organization for Standardization, Geneva, Switzerland, 1996.
3. *Industrial Automation Systems. Product Data Representation and Exchange. Part 14, EXPRESS-X Mapping Language*, ISO 10303-14 (http://www.nist.gov/sc4/stepdocs), International Organization for Standardization, Geneva, Switzerland, 1998.
4. M. Hardwick, *The STEP Services Reference Manual*, Technical Report 96004, Laboratory for Industrial Information Infrastructure, Rensselaer Polytechnic Institute, Troy, New York, 1995 (http://www.rdrc.rpi.edu/).
5. *White Paper to the OMG in Response to OMG Manufacturing Domain*, Task Force OMG TC Document 97-09-1, Object Management Group, Framingham, MA, December 1993. (ftp://ftp.omg.org/pub/docs/mfg/97-09-01.pdf)

## 11.17 Catherine Houstis, and Christos N. Nikolaou

"*Open Scientific Data Repositories Architecture*" [3]

Prof. Catherine Houstis, Prof. Christos N. Nikolaou, and Spyros Lalis
University of Crete and Foundation for Research and Technology - Hellas

Sarantos Kapidakis and Vassilis Christophides
Foundation for Research and Technology - Hellas
E-mail: {houstis, nikolau, lalis, sarantos, christop}@ics.forth.gr

The improvement of communication technologies and the emergence of the Internet and the WWW have made feasible the worldwide electronic availability of independent scientific repositories. Still, merging data from separate sources proves to be a difficult task. This is due to the fact that the information required for interpreting, and thus combining, scientific data is often imprecise, documented in some form of free text or hidden in the various processing programs and analytical tools. Moreover, legacy sources and programs understand their own syntax, semantics and set of operations, and have different runtime requirements. These problems are multiplied when considering large environmental information systems, touching upon several knowledge domains that are handled by autonomous subsystems with radically different semantic data models. In addition, there is a widely heterogeneous user community, ranging from authorities and scientists with varying technical and environmental expertise as well as the public.

To address these problems we envision an open environment with a middleware architecture that combines digital library technology, information integration mechanisms and workflow-based systems to address scalability both at the data repositories level and the user level. A hybrid approach is adopted where the middleware infrastructure contains a dynamic scientific collaborative work environment and mediation to provide both the scientists and simple users the means to access globally distributed heterogeneous scientific information. The proposed architecture consists of three main object types: data collections, mediators, and programs.

Data collections represent data that are physically available, housed in the system's repositories. Scientific data collections typically consist of ordered sets of values, representing measurements or parts of pictures, and unstructured or loosely structured texts containing various kinds of information, mostly used as descriptions directed to the user. Each data collection comes with a set of operations, which provide the means for accessing its contents; the access operations essentially combine the model according to which items are organized within the collection and the mechanisms used to retrieve data.

Mediators are intermediary software components that stand between applications and data collections. They encapsulate one or more data collections, possibly residing in different repositories, and provide applications with higher-level abstract data models and querying facilities. Mediators are introduced to encode complex tasks of consolidation, aggregation, analysis, and interpretation, involving several data collections. Thus mediators incorporate

[3] This work is funded in part by the THETIS project, for developing a WWW-based system for Coastal Zone Management of the Mediterranean Sea, EU Research on Telematics Programme, Project Nr. F0069.

considerable expertise regarding not only the kind of data that is to be combined to obtain value-added information, but also the exact transformation procedure that has to be followed to achieve semantic correctness. In the simplest case, a mediator degenerates into a translator that merely converts queries and results between two different data models. Notably, mediators themselves may be viewed as abstract data collections, so that novel data abstractions can be built incrementally, by implementing new mediators on top of other mediators and data collections. This hierarchical approach is well suited for large systems, because new functionality can be introduced without modifying or affecting existing components and applications.

Programs are procedures, described in a computer language that implement pure data processing functions. Rather than offering a query model through which data can be accessed in an organized way, programs generate data in primitive forms (e.g. streams or files). They also perform computationally intensive calculations, so that appropriate recovery and caching techniques must be employed to avoid repetition of costly computations. Last but not least, program components can be very difficult to transport and execute on platforms other than the one where they are already installed, so that they must be invoked remotely.

Every component has metadata associated with it. Thus, registered objects can be located via corresponding meta-searches about their properties. Data collections and mediators bear metadata describing their conceptual schemata and querying capabilities. Programs are described via metadata about the scientific models implemented, the run time requirements of the executable, and the type of the input and output. Metadata are the key issue in an cross-disciplinary system with numerous underlying components: only when we know all significant details about data and programs we will be able to combine them and use them with minimal extra human effort.

Access to the system, is given through a single web-based interface, which allows users to browse through the object collections in order to discover data sets, mediators, and programs of interest. This is achieved by invoking a metadata search engine that locates objects based on their metadata descriptions and returns a list of links to the corresponding objects. In addition, data collections, mediators and programs are put together to offer the user two advanced modes of use: applications and workflows. Applications are rigid software components tailored to a specific information-processing task. Each application uses a well-defined subset of the system's data collections, combines their content in a certain way, and provides value-added services to a known user group. Hence, data integration and processing functions of applications are introduced via carefully designed mediators. Workflows are program pipelines that can be set up at run time, by locating the appropriate components and interactively linking them with each other. Based on the type of the components selected, the system checks for incompatibilities and runtime restrictions, creates an execution context, set up the network connections among the various components, and invokes them in the appropriate order.

This hybrid approach greatly enhances flexibility. On the one hand, mediation is used to offer robust data abstraction with advanced querying capabilities, and to promote incremental development, which is required to build heavy-weight applications such as decision making tools. On the other hand, the collaborative workflow environment is addressing the needs of the scientific community for flexible experimentation with the available system components, at run time. In fact, a workflow can be used to test processing scenaria for a particular user community, in a "plug-and-play" fashion, before embodying them into the system as mediators and applications.

## 11.18 Yannis Ioannidis and Miron Livny

"*Desktop Scientific Experiment Management*"

Prof. Yannis Ioannidis
University of Athens, Greece, Department of Informatics
Prof. Miron Livny
University of Wisconsin, Department of Computer Sciences
Madison, WI 53706
E-mail: {yannis,miron}@cs.wisc.edu
URL: http://www.cs.wisc.edu/~{yannis,miron}/

### Motivation

In the past few years, several scientific communities have initiated very ambitious and broad-ranged projects in their disciplines. The NASA Eos effort and the NIH Human Genome project are two examples of such national and international scientific endeavors. A major part of these projects is the collection of huge amounts of data (sometimes measured in petabytes) on complex phenomena. Managing this surge of scientific data poses significant challenges, many of which cannot be effectively addressed by existing database technology. This has resulted in much research activity in the area of *Scientific Database Systems*.

Nevertheless, still little attention has been devoted to the needs of small teams of scientists who perform individual experimental studies in their laboratories. In particular, a major problem that many experimental scientists are facing is that there are no adequate experiment management tools that are powerful enough to capture the complexity of the experiments and at the same time are natural and intuitive to the non-expert. A small laboratory that can easily generate and store several megabytes of data per day is still dependent on the good old paper notebook when it comes to keeping track of the data.

In addition to generating a significant amount of data during its course, an experimental study often involves accessing many *data-providing* "systems", including simulation tools, laboratory equipment, statistical analysis packages, visualization packages, administrative tools, and others. They typically form a heterogeneous collection of legacy systems whose interactions capture much of the *experiment flow* of the study.

To address the above challenges, what is needed is a management system offering high-level modeling of and access to the data generated by an experimental study, integrated access to all external data-providing "systems", the ability to plug in any such system without touching it, and an advanced conceptual interface that hides the input/output details of these systems.

### The ZOO Project

Over the past four years, based on NSF funding (partly within the "Scientific Databases Initiative"), we have collaborated with several domain scientists, studied the needs of a wide range of experimental disciplines, developed solutions to some of the basic problems in experiment management, and made significant progress towards implementing a simple

*Desktop Experiment Management Environment (DEME)* called *Zoo*. Our work has proceeded in a tight loop between developing *generic* experiment management technology that is implemented in a *generic* tool, Zoo, and installing *customized* enhancements of the tool that constitute full systems (complete *Customized Desktop Experiment Management Systems (CDEMSs)*) in laboratories[4] of interest. The following are the main features of the system:

1. A generic architecture that captures the essence of experimental studies in arbitrary disciplines.
2. An open architecture that permits its customization to specific studies or environments with little effort, offering tools for connecting the system to many external applications.
3. An object-based database server that supports a high-level data model and query language, captures and leads the entire experimental process, can deal with some "status" queries for in-flight experiments, and offers many desirable features of complex scientific workflows.
4. A modular translator of data from their conceptualization and representation in the database to those required by the experimentation environment (typically flat files) and vice-versa.
5. A uniform visual interface that can be used throughout the life-cycle of experimental studies.

A key characteristic that brings many of the above together is that the conceptual/logical schema of the experiment data has been placed at the center of Zoo. Whether the activity is designing the study, invoking an experiment, querying the data, or analyzing query results, a (visual) representation of the schema is used to perform the activity. The schema essentially captures the study itself, including its data and all forms of metadata.

Zoo has been successfully tested by a collaborating group in Soil Sciences for plant-growth simulation experiments. ZOO has been customized to communicate with several simulators used for such experiments, and the resulting CDEMS has been used to drive test runs on them. Likewise, Zoo has also been tested by a collaborating group in Biochemistry for NMR experiments run on spectrometers, with very positive results again.

**Research Challenges**

Our experience from installed software as tested and evaluated in real-life settings has shown that one of the biggest challenges of experiment management systems like Zoo lies with their user interface. These include problems related to schema visualization, visual query languages, dynamic queries, approximate query answers, incremental queries, visual translation specification between internal objects and external files, and others. Another major challenge lies in dealing with arbitrary external systems. Especially critical is the semantic integration of "schemas" of external software, dealing with interactive (as opposed to batch) external

[4] We use the term "laboratory" to indicate any scientific environment where experiments are conducted, be it a physical laboratory in the traditional sense, a virtual laboratory involving scientists collaborating across the network, simulation-based modeling, etc.

systems, and querying external systems while they are running and obtaining useful status information. Finally, modeling advanced experiment flows still remains an elusive goal, which is nevertheless very important for scientific experiments.

## 11.19 Alan Kaplan and Jack C. Wileden

"*Practical Foundations for Transparent Interoperation*"

Prof. Alan Kaplan
Clemson University, Department of Computer Science
Clemson, SC 29634-1906
E-mail: kaplan@cs.clemson.edu
URL: http://www.cs.clemson.edu/~kaplan/

Prof. Jack C. Wileden
University of Massachusetts, Department of Computer Science
Amherst, MA 01003-4610
E-mail: wileden@cs.umass.edu
URL: http://www-ccsl.cs.umass.edu/~jack/

Scientific software systems, such as computer-aided design/computer-aided modeling applications, geographic information systems, earth observation systems, and bioinformatic applications, are increasingly focused on the exchange and integration of distributed information. Bioengineers, geneticists, biologists and other scientists are creating, using and managing ever larger amounts of ever more complex data. Moreover, rather than relying on the data produced by an individual application, scientists are becoming increasingly dependent on data originating from multiple sources. Such data may be produced by applications developed by a single scientist or various scientists in a particular laboratory, or more likely with the proliferation of the World Wide Web, acquired from other scientists. In a related manner, independently developed tools may produce individual data sets that need to be integrated in order to perform a required analysis. Legacy data, i.e., data produced by applications that are no longer accessible or maintainable, is also another source of information often required by scientists.

Developers and users of such data, however, are faced with a difficult tradeoff in the design and construction of software that creates and uses this data. They need to be able to model data so that it can be easily understood and efficiently used by an individual scientific application; at the same time they need the data to be in a form such that it can be used, integrated, and shared by other scientific applications engaged in integration of distributed data, even though the data might be described and defined using various formats, notations, models and/or languages.

Despite these problems, computer science has provided little foundational support to developers of scientific software systems. Instead, scientific applications traditionally have overcome such problems by resorting to relatively low-level techniques, such as using flat files, standard data interchange formats, IDL-mediated data exchange mechanisms or ad hoc wrapping of legacy data repositories. Such approaches tend to have various shortcomings. For example, standardized interchange formats typically require explicit translations, which are often inefficient and prone to error. IDL-mediated mechanisms, such as CORBA and DCOM impose foreign data models that represent a least common denominator type system [3]. Furthermore, once a developer has committed to a particular IDL-mediated mechanism (e.g., CORBA, DCOM), changing to a different mechanism is extremely difficult and very costly. Another serious drawback associated with such approaches is that their use generally requires that developers and users be aware of the boundaries between various data repositories and the applications that need to access them. As a result, software based on these approaches that needs to access and manipulate scientific data is difficult to develop and maintain.

Our research is directed toward developing computer science foundations for transparent interoperation, in particular, exploring both theoretical and practical aspects of this problem domain. The primary objective of our work is to hide the boundaries or seams between heterogeneous data repositories or between data repositories and applications that need access these repositories. Developing appropriate theoretical and practical foundations for transparent interoperability results in software and data that it easier to develop, reuse, share and maintain.

A companion paper [5] appearing in this workshop outlines some formal models of type compatibility, type safety, and name management. In this paper, we give an overview of the practical aspects of our work, specifically the development of a new, highly transparent approach to interoperation, called PolySPIN. A collection of prototype, automated tools supporting the use of the PolySPIN approach, as well as our experience with their application, is also described.

Based on our formal foundations, we have been designing, developing, and experimenting with tools that facilitate interoperability. The PolySPIN approach provides a transparent interoperability mechanism for programming languages. More specifically, it provides support for polylingual interoperability [4] where applications can access compatible types defined in distinct languages as if they were defined in the language of the application. The fact that the types are defined and implemented in a different programming language is hidden from the application. Related to this mechanism is PolySPINner, a collection of tools that automates PolySPIN and supports type-safe polylingual interoperability [1,2].

Although approaches such as standard file formats, relational databases, IDLs (e.g., CORBA, OLD/DCOM) support certain aspects of polylingual interoperability, our approach offers several advantages over such mechanisms:

- The same data type model is used to define, manipulate and store the data.
- It is megaprogramming friendly. Applications can share and integrate data after it has been defined and created.
- It has minimal impact on software developers. Existing applications do not have to be manually modified in order to access complex data defined in different programming languages.
- It supports realistic interlanguage type compatibility. For example, applications can access data where only a subset of features is needed.

To help illustrate these advantages, we describe how the PolySPINner toolset can be used to create a Java genome sequence application that accesses and manipulates genome sequences defined using both Java and C++ type systems. The toolset takes as input the class definitions (i.e., interface and implementation) for the C++ and Java types. With respect to our scenario, there are Java and C++ type definitions for genome sequences. The types may match exactly, or more likely, there is an intersection of features that is relevant to the new Java genome application. In any event, PolySPINner first parses the type definitions for each language and then determines whether the types are compatible (where compatibility can be specified as discussed in [5]). If the types are deemed compatible, PolySPINner automatically re-engineers the implementations of all relevant operations according to the PolySPIN framework. The re-engineered operations include code that checks for the actual language implementation for objects and then invokes the appropriate operation implementation in the appropriate language.

At first blush, PolySPIN and the PolySPINner toolset provide similar functionality compared to popular distributed object technologies such as CORBA and OLE/DCOM. The primary advantages of our approach over these contemporary approaches are enumerated above. Although the details are beyond the scope of this position paper, our approach also hides the underlying interoperability technology. This means that IDL-mechanisms (such as CORBA and OLE/DCOM), as well as non-IDL mechanisms such as Remote Method Invocation and Java Native Interface can be potentially made transparent. Thus, a developer can change the underlying interoperability mechanism with minimal impact on existing applications.

In summary, our position is that interoperability should be transparent in distributed, complex software systems. Defining practical foundations for transparent interoperability (as opposed to using ad hoc and/or cumbersome approaches) permits developers of scientific applications to focus on the problem domain rather than on the underlying interoperability mechanism. We claim that this results in scientific applications and data that are easier and less costly to design, build, maintain and share.

1. Barrett, D. J., Kaplan, A. and Wileden, J. C., "Automated Support for Seamless Interoperability in Polylingual Software Systems," *Fourth Symposium on the Foundations of Software Engineering*, San Francisco, CA, October 1996.
2. Kaplan, A., *Name Management in Convergent Computing Systems: Models, Mechanisms and Applications*, PhD Thesis, Technical Report TR–96–60, Department of Computer Science, University of Massachusetts, Amherst, MA, May 1996.
3. Kaplan, A., Ridgway, J. V. E. and Wileden, J. C., "Why IDLs Are Not Ideal," *Ninth IEEE International Workshop on Software Specification and Design*, Ise-Shima, Japan, April, 1998.
4. Kaplan, A. and Wileden, J. C., Toward Painless Polylingual Persistence, *Proceedings Seventh International Workshop on Persistent Object Systems*, Cape May, NJ, May 1996.
5. Wileden, J. C. and Kaplan, A., "Formal Foundations for Transparent Interoperation," *NSF Workshop on Distributed Information, Computation, and Process Management for Scientific and Engineering Environments*, Herndon, VA, May 1998.

## 11.20 Miron Livny

"*High Throughput Computing on Distributively Owned Grids*"

Prof. Miron Livny
University of Wisconsin
Madison, WI 53706
E-mail: miron@cs.wisc.edu
URL: http://www.cs.wisc.edu/~miron/

### Introduction

As more scientists, engineers, and decision makers use computers to generate behavioral data on complex phenomena, it is not uncommon to find customers of computing facilities who ask the question: How much behavioral data can I generate by the time my report is due? This question represents a paradigm shift. In contrast to other users, these *throughput oriented users* measure the power of the system by the amount of work it can perform in a fixed amount of time. The fixed time period is measured in the relatively coarse units of days or weeks, whereas the amount of work is seemingly unbounded — one can never have too much data when studying a biological system, testing the design of a new hardware component, or evaluating the risk of an investment.

Given their unbounded need for computing resources, the throughput oriented community is closely watching activity in the computational grids area, and anxiously awaiting the formation of productive High Throughput Computing (HTC) grids. To meet the seemingly infinite appetite for computing power of this community, a HTC grid has to be continuously on the lookout for additional resources and therefore must have the mentality of a scavenger. The services of a provider of computing power are always accepted regardless of the resource's characteristics, degree of availability, or duration of service. As a result, the pools of resources HTC grids draw upon to serve their customers are large, dynamic and heterogeneous collections of hardware, middleware, and software.

As a result of the recent decrease in the cost/performance ratio of commodity hardware and the proliferation of software vendors, resources that meet the needs of HTC applications are plentiful and have different characteristics, configurations and flavors. A large majority of these resources reside today on desk-tops, owned by interactive users, and are frequently upgraded, replaced or relocated. The decrease in the cost/performance ratio of hardware also rendered the concept of multi-user time-sharing obsolete. In most organizations, each machine is usually allocated to one individual in the organization to support his/her daily duties. A small fraction of these machines will be grouped into farms and allocated to groups who are considered to be heavy users of computing by management. As a result of this trend, while the absolute computing power of organizations has improved dramatically, only a small fraction of this computing power is accessible to HTC users due to the ever increasing fragmentation of computing resources. In order for a HTC grid to productively scavenge these *distributively owned* resources, the boundaries marked by owners around their computing resources must be crossed.

However, crossing ownership boundaries for HTC requires that the rights and needs of individual resource owners be honored. The restrictions placed by owners on resource usage for HTC can be complex and dynamic. The constraints attached by owners to their resources prevent the HTC grid from planning future allocations. All the resource manager of the grid knows is the current state of the resources. It therefore has to treat them as sources of computing power that should be exploited *opportunistically*. Available resources can be reclaimed at any time and resources occupied by their owners can become available without any advance notice. The resource pool is also continuously evolving as the mean time between hardware and software upgrades of desk-top machines is steadily decreasing.

In addition to ownership boundaries, HTC grids must cross administrative domains as well. The main obstacle to inter-domain execution is access to the grid from which the application was submitted, such as input and output files. The HTC grid has to provide means by which an application executing in a foreign domain can access its input and output files that are stored at its home domain. The ability to cross administrative domains not only contributes to the processing capabilities of the grid, but also broadens the "customer circle" of the grid. It makes it very easy to connect the computer of a potential customer to the grid. In a way, the HTC grid appears to the user as a huge increase in the processing power of her personal computer since almost everything looks the same except for the throughput.

HTC applications usually follow the master-worker paradigm, where a list of tasks is executed by a group of workers under the supervision of a master. The realization of the master and the workers and their interaction may take different forms. The workers may be independent jobs submitted by a shell script that acts as the master and may collect their outputs and summarize them, or they can be implemented as a collection of processes which receive their work orders in the form of messages from a master process that expects to receive the results back in messages. Regardless of the granularity of the work-units and the extent to which the master regulates the workers, the overall picture is the same — a heap of work is handed over to a master who is responsible for distributing its elements among a group of workers.

Making high throughput computing grids a reality will require intensive and innovative research in a wide range of areas. We list here some of challenges we face on the way to the deployment of large scale productive HTC computational grids.

1. **Understanding the sociology of a very large and diverse community of providers and consumers of computing resources.** In a HTC grid, every consumer can also be a provider and every provider can also be a consumer of services. We need to understand the needs of these communities and develop tools that will enable them to interact with the grid, monitor its behavior and express their policies.
2. **Throughput oriented parallel applications.** The effectiveness of a HTC grid depends on the ability of the applications to adapt to changes in the availability of processors and memory. Applications that require a fixed number of processors and memory are unlikely to benefit from the services of a HTC grid.
3. **Understanding the economics (relationship between supply and demand) of a HTC grid.** While basic priority schemes for guaranteeing fairness in the allocation

of resources are well understood, mechanisms and policies for equitable dispersion of services across large computational grids are not. A major aspect in the development of such policies involves understanding supply and demand for services in the grid.

4. **Data staging — moving data to and from the computation site.** When the problem of providing vast amounts of CPU power to applications becomes better understood, the next hurdle that must be crossed will be that of providing sufficiently sophisticated Resource Management services to individual throughput oriented applications. A major aspect of this problem is that of integrated staging, transport and storage media management mechanisms and policies for high throughput.

5. **Tools to develop and maintain robust and portable resource management mechanisms for large, dynamic and distributed environments.** Current approaches to developing such complex resource management frameworks usually involve a new implementation of large fractions of the framework for each instance of a marginally different Resource Management System. We also lack tools to monitor the behavior of large distributed environment and keep track of large collections of jobs submitted to such environments. Given the rapid changes in hardware and operating systems, the biggest potential source of throughput inefficiency is exclusion of resources due to the inability of the mechanisms of a HTC grid to operate on new computers. There is nothing more frustrating for a HTC customer than new resources that are likely to be the biggest and fastest being excluded from the HTC grid due to porting difficulties.

## 11.21 Nicholas M. Patrikalakis and Stephen L. Abrams

"*Towards a Distributed Information System for Ocean Processes*"

Prof. Nicholas M. Patrikalakis and Mr. Stephen L. Abrams
Massachusetts Institute of Technology, Department of Ocean Engineering
Cambridge, MA 02139-4307
E-mail: nmp@mit.edu, sabrams@deslab.mit.edu
URL: http://deslab.mit.edu/DesignLab/

With the advent of new sensors, storage technologies, and networking technologies, the potential exists for a new era of ocean science investigation where students, scientists and researchers from different disciplines, and government policy makers and officials have frictionless access to oceanographic data, simulation results, and software. Our goal is to create standards and an information processing environment that allows these users to extract and employ marine data from ocean sampling sensors and the simulated ocean as timely input to visualization, analysis, and ocean modeling software. The intent of the system is to make scientific inquiry easier, quicker, and more collaborative in nature by creating an oceanographic metadata standard and a software system that uses that standard to create a frictionless simulation and analysis environment.

The premises behind this project are:

1. The trend in oceanographic research is towards interdisciplinary experiments where scientists exchange data and results in real or near real-time
2. Lack of a complete oceanographic metadata standard hinders scientific inquiry and the sharing of data and software
3. Data from ocean observations, the tools to analyze them, and numerical prediction models reside at many different locations and have heterogeneous formats
4. Scientists, planners, and managers need to apply different analyses to data, both for reducing the uncertainty of results and for assessing sensitivity to crucial assumptions
5. Oceanographic simulation software is extremely complex with steep learning curves
6. The existence of the Internet makes it possible to rapidly access the different resources without investments in hardware for the user

Our system will provide the oceanographic scientist, researcher, or planner with the computational infrastructure necessary for the flexible creation and execution of complex simulation and analysis workflows, whose individual components are distributed oceanographic data and software resources.

The three main research thrusts of this project are:

1. Distributed resource search and retrieval
2. Interoperability in a heterogeneous computational and communications environment

3. Flexible and efficient process management

To facilitate the identification of and access to the appropriate resources, we will utilize a uniform marine ontology, or controlled formal vocabulary, to attach descriptive domain-specific metadata to all resources registered with the system, enabling efficient metadata-based searching for distributed data and processing systems. Our ontology and metadata efforts will draw on existing standards in the geoscientific and oceanographic fields whenever possible.

To gain widespread acceptance and effective use, and to ensure future extensibility, a distributed system must allow for the widest possible heterogeneity in its operating environment. Since each resource component of a workflow potentially exists on a unique host, it is imperative that we use a common information transport mechanism that supports transparent interoperability.

Process management encompasses the various administrative capabilities of the executive system for registering resources, attaching appropriate metadata to those resources, and generating, validating, and executing workflows. One important aspect of workflow management is automatic generalized data translation, which is necessary to ensure the smooth flow of data between the computational nodes of the workflow, regardless of syntactic and semantic mismatches between various input and output formats. Another important issue is the mechanism for smoothly integrating established legacy systems — despite possible obsolescence of implementation environment, non-interactive execution mode, or non-network connectivity — into the standard operating framework.

Oceanographic processes, whether physical, chemical, biological, acoustical, geological, or optical, are all characterized by a wide range of spatial and temporal scales and variability. Furthermore, these processes exhibit complex couplings and interactions. Our system will explicitly address these issues by providing an extensible computational environment for the manipulation of complex oceanographic workflows. No longer burdened with the technical minutiae of networking, distributed search and retrieval, query languages, and data management, the user of the system is freed to concentrate on domain-specific scientific research issues. By providing a mechanism for rapidly performing numerous simulation scenarios, the system will enable a more efficient and exhaustive exploration of the potential solution spaces of complex, multi-disciplinary oceanographic problems or research areas. One component of the research program concerns the investigation of multiscale, multidisciplinary ocean phenomena. Our system will, in general, facilitate the exploration of these type of complex geoscientific systems.

A detailed investigation of many interesting oceanic processes requires in situ measurements of the ocean environment. One highly effective and efficient means of utilizing sensor resources is through adaptive sampling with autonomous unmanned vehicles. Through a dynamic process of cooperative autonomous decision-making in response to changing environmental or experimental conditions, these vehicles can optimize their mission for the exploration of regions of high interest, lowering the time and expense of acquiring relevant high-quality data.

Our system will be a prototype of a domain-specific, multi-disciplinary knowledge system that addresses issues of scale, heterogeneity, and distributed interoperability of the computing environment.

## 11.22 Lawrence P. Raymond

"*Pattern Analyzers*"

Dr. Lawrence P. Raymond
Oracle Corporation, Special Projects
Kingwood, TX 77345
E-mail: lraymond@us.oracle.com

Forget death and taxes. Change is perhaps the only process one can depend on. This includes change in natural properties such as alterations of state, condition, location, composition, dynamics, and impact. And it also includes to changes in markets, resources, costs, prices, rules, and regulations, geopolitical status, and other similar risk and opportunity measures. Those able to measure and monitor change meaningful to their purposes are the most likely to find their missions realized. Those able to relate the dynamics in business and natural properties to root causes, potentially can manage risk and foresee opportunities to great advantage.

This explains the interest in data management, computerized simulations and analytical advances. Many growth interests in science and business are strategic. Business future is dependent upon sustainability not of things, but of dynamic balance. For man to advance quality of life, he must learn to balance resource and business consumption with productivity. And we have much to learn. We need a mechanism that allows timely comparison of decisions and policy with their consequences. Suggested here is that we need real-time measures of the change resulting from actions we can control, in context with those events we cannot. Given this, it should be possible to start with what we think we know, compare results against expectations, and make rapid adjustments as results dictate. This should lead to continuous learning with minimal redundancy and maximum justification.

How close are we to achieving this? The answer is indicated by the current state of technologies for data collection, management, and transformation, as well as reduction, modeling, and interpretation techniques. This perhaps is why we are here.

**Current State-of-the-Art**

Technology in database management no longer is limiting creativity. It is now possible to create databases directly from data models using computerized tools that write much of the script directly. Tools also exist for loading legacy data ranging from character, audio, video, temporal, image, text and spatial; these may be in flat files or multi-dimensional. Data can be housed in universal servers, and tapped by a variety of application servers directly by network computers or client-server environments. Network servers can push files and programs through the system, or can link directly to a single client sharing software throughout the enterprise. Data storage has become almost limitless, whether in co-located or distributed systems, such that the risk of data loss is approaching zero. With the advent of Java and other open-architectures, restrictions placed by operating systems potentially are removed. With the advent of the Internet, any data can now be accessed anywhere at any time. And, through data warehousing and metadata, data can be mined and manipulated in a variety of ways to build upon our knowledge base.

Speed, capacity, and system flexibility will continue to evolve, but none of these are limiting knowledge evolution at this time. Limitations appear in the amount of knowledge that can be gained from these huge data stores, as well as in the efficiency and utility of data acquisition. These, in turn, depend upon the number of persons educated in science, systems, and data analysis, as well as upon advancements in ways to capture and analyze data.

**Changing Paradigms**

Several fundamental changes are occurring in the ways we think about information. One is that patterns reflected in acquired data sets are now being investigated instead of parameter trend analyses. This derives from concepts that organization exits in chaos, and strings of data exist that depict new relations when sufficient numbers are brought together and viewed in new ways.

It is now recognized that all natural phenomena oscillate within ranges measured by levels, frequency and rate, and that within a given time frame, are characteristic of normal conditions. Thus, all parameters can be viewed as zero deviations from normal, with corresponding values zeroed as floating points. Any changes from normal conditions can be arbitrarily defined as significant, and described by a multi-dimensional control surface. Data patterns shown by these control surface images likely reflect environmental perturbations that can be defined through research.

Examples are patterns in pH, temperature, conductivity, and redox potential measured in biological fluids during normal and diseased states. Each of these parameters normally cycles over a given range, dependent primarily on time, food source, and fluid intake. Natural system regulators maintain this condition in homeostatic state. When abnormal factors are introduced, such as particles, antigens, and antagonists, these natural regulators are stressed. The first indication of stress occurs as spikes in parameter values that manifest as change in standard deviation. Standard deviation increases until the capacity of the regulatory system is surpassed. At this point, the rate of parameter change increases and continues as normal level thresholds are exceeded, without return to normal. This occurs most notably for temperature, with smaller, related changes occurring in other parameters.

The time interval varies between increased standard deviation and increased rate, dependent upon causal agents. Similarly, characteristic patterns develop for the other parameters, creating different surface contours based upon both magnitude and direction of response. Viewed as a multi-dimensional surface, each pattern was associated with specific causes, some with remarkable resolution. The outcome was an ability to quickly predict many root causes, based upon a comparative analysis of multi-dimensional response curves. This analytical approach could be improved by coupling with mathematical models; this may extend and improve predictability, as well as help address elements of uncertainty.

**Technology Transfer**

Migrating technology to the commercial is facilitated by building upon existing industry standards in ways that help answer difficult problems in a timely fashion. The above suggest that principal opportunities exist in sensor system research and in coupling pattern-based analytical methods with current modeling efforts.

This research was conducted in relatively small, controlled environments, up to warehouse size and neonatal ward complexity. It related only cause and effect. What needs to be done now is to expand this to real world regions, greater amounts of data, sensor systems able to measure and resolve regional data, all within an analytical framework that helps identify and justify further research needs.

## 11.23 Allan R. Robinson

"*Ocean Observation and Prediction Systems and the Potential for Progress via Distributed Computing*"
PROF. ALLAN R. ROBINSON
Harvard University, Division of Engineering and Applied Sciences
Department of Earth and Planetary Sciences
Cambridge, MA 02138
E-mail: robinson@pacific.harvard.edu

The ocean is a complex multidisciplinary fluid system in which dynamical process and phenomena occur over a range of interactive space and time scales, spanning more than nine decades. The fundamental interdisciplinary problems of ocean science, identified in general more than a half century ago, are just now feasible to research. They are being pursued via research in ecosystem dynamics, bio-geo-chemical cycles, littoral-coastal-deep sea interactions, and climate and global change dynamics. Concomitantly, novel process in maritime operations and coastal ocean management is underway. Realistic four-dimensional (three spatial dimensions and time) interdisciplinary field estimation is essential for efficient research and applications. Ocean field estimation must take into account multiple scale synoptic circulation (mesoscale, jet-scale, regional, sub-basin, and basin) and episodic and intermittent variability. The multiplicity of scale, in both space and time, require nested high-resolution observational and modeling domains with accurate sub-gridscale parameteriza-

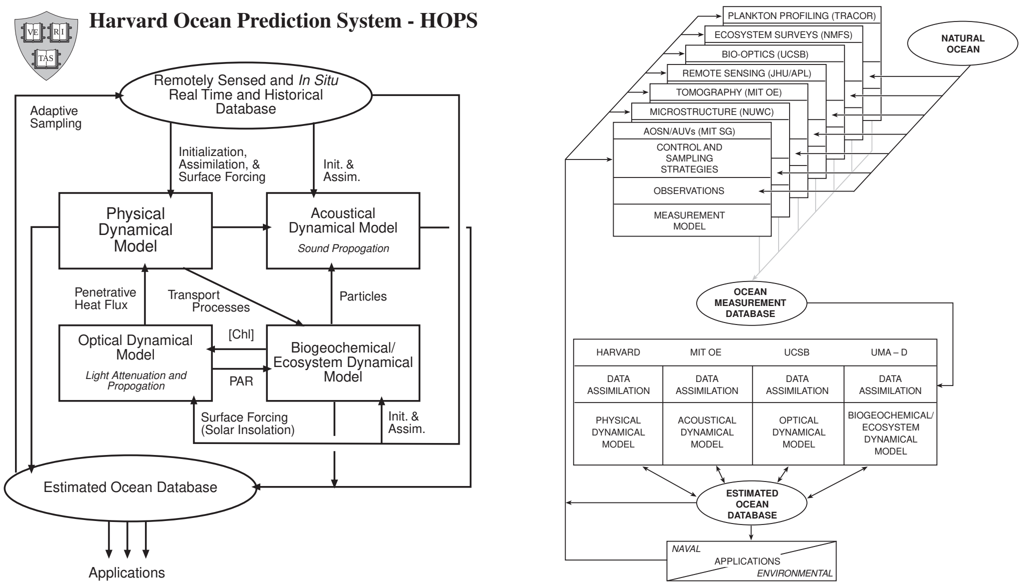



Figure 2: (a) Harvard Ocean Prediction System (HOPS); (b) Littoral Ocean Observing and Prediction System (LOOPS) architecture.

tions. State variables to be estimated include, e.g., temperature, salinity, velocity, concentrations of nutrients and plankton, ensonification, irradiance and suspended sediments. Physical process which must be represented, in the case of the coastal ocean, include: tides, fronts, waves, currents and eddies, boundary layers, surf zone effects, estuarine processes, etc. Other oceanic regimes will include other processes as well. Remote sensing of sea surface temperature, sea surface color (chlorophyll), sea surface height from satellites which are now all flying together allow for continuous ocean monitoring with high resolution (alongtrack for SSH) all over the globe. Realistic field estimates, including real-time nowcasts and forecasts as well as simulations, are now feasible because of the advent of Ocean Observing and Prediction Systems (OOPS), in which:

1. a set of coupled interdisciplinary models are linked to,
2. an observational network consisting of various sensors mounted on a variety of platforms,
3. via data assimilation schemes.

Compatible multiscale nested grids for the models and observations are essential. Data assimilation methods meld various types of data with dynamical estimates to improve field estimates and provide feedback for adaptive sampling and model improvement. The effective assimilation of data helps to: control model phase error, correct model dynamical deficiencies, provide parameter estimates, etc.

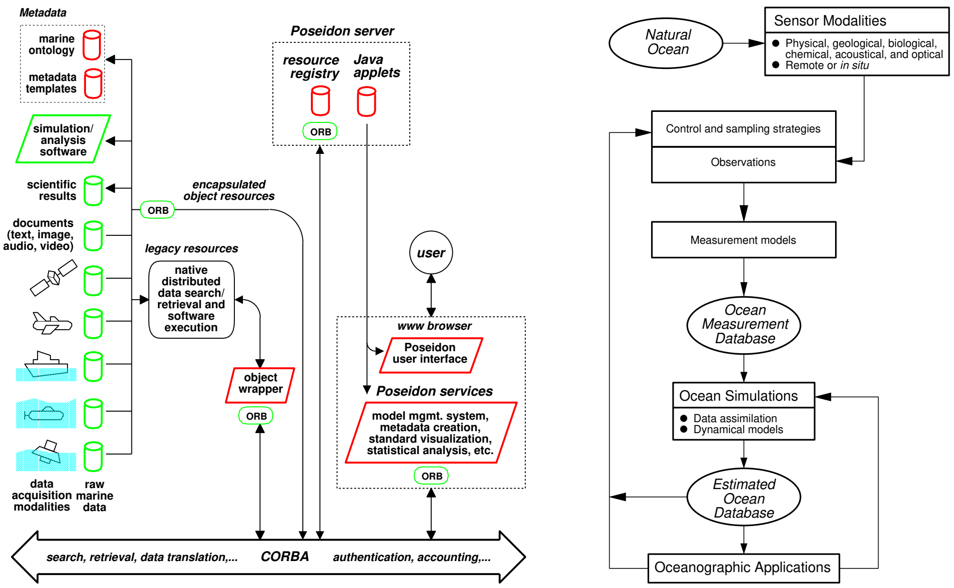



Figure 3: (a) CORBA-based OOPS; (b) Generic OOPS.

The Harvard Ocean Prediction System (HOPS) (Fig. 2(a)) is an example of an OOPS: it is generic, portable, and flexible, and is being and has been used in many regions of the world ocean. HOPS is a central component of both the advanced Littoral Ocean Observing and

Prediction System (LOOPS) (Fig. 2(b)) being developed collaboratively under the National Ocean Partnership Program, and of the prototype Advanced Fisheries Management Information System (AFMIS). Ocean science is at the threshold of a new era of progress as a result of the development of multiscale, multidisciplinary OOPS. These OOPS involve novel interdisciplinary sensor, measurement, modeling and estimation methodologies and systems engineering techniques. Efficient OOPS require distributed network systems. Distributed network-based OOPS (Fig. reffig:oops2(a)) will increase the capability to manage and operate in coastal waters for: pollution control, public safety and transportation, resource exploitation and coastal zone management, and maritime and naval operations.

Experience gained with prototype OOPS will stimulate and expand the conceptual framework for knowledge networking and the scope of computational challenges for ocean science. The general principles of OOPS system architecture (Fig. 3(b)) will be applicable to analogous systems for field estimation in other science areas: meteorology, air-sea interactions, climate dynamics, seismology, etc. These advanced systems exemplify the research directions and needs of the ocean science community for novel and efficient distributed information systems, which can only be realized by a powerful integrated effort in concert with the computational, distributed system and engineering communities.

## 11.24 Jarek Rossignac

"*Collaborative Design and Visualization*"
Prof. Jarek Rossignac
Georgia Institute of Technology, Graphics, Visualization, and Usability Center
Atlanta, GA 30332-0280
E-mail: jarek@cc.gatech.edu
URL: http://www.gvu.gatech.edu/people/faculty/jarek.rossignac/

In the past, 3D was restricted to specialists having the appropriate skills, professional software, expensive graphics hardware, and access to corporate 3D databases. Real-time graphics on personal computers, quick download of 3D models over the internet, and the standardization of data formats and libraries will open the door to a "democratization" of 3D access and will permit to link 3D databases with personal productivity and communication tools. Employees, customers, teachers, and suppliers, although not having any 3D expertise, will be able to use 3D databases for collaborative design reviews, problem solving, and tracking; for 3D-based multi-media problem reports; for online education, training, and documentation; for internet-based part purchasing and subcontracting; for demonstration to customers; or for advertising.

A successful deployment of computer mediated tools for the collaborative design and exploitation of 3D datasets requires major advances on many fronts.

- The complexity of detailed 3D models increases faster than the communication bandwidth. Dedicated 3D compression schemes and progressive transmission techniques will provide users with instant access to remote models represented by millions of geometric primitives.
- Interactive manipulation of models of industrial complexity requires graphics performances that are 2 to 3 orders of magnitude greater than what is currently available on personal computers. Multi-resolution models and adaptive rendering algorithms that automatically manage the optimal trade-off between performance and image quality will be combined with image manipulation techniques and with progressive transmission of compressed models.
- There is a clear need for a new generation of user-interfaces that enable everyone to very quickly become an expert in navigating possibly animated 3D worlds and in using them effectively for design, marketing, education, or entertainment. Although immersive VR may not be appropriate in many applications, 3D input devices and stereo displays can make a considerable difference, even when used by team members in design review sessions.
- To be truly effective, interactive 3D viewing and design environments must relieve users from the most tedious activities, such as interference detection or the verification of compliance with design guidelines, and be fully integrated with personal productivity and product data management tools. Advances in graphics hardware and progress in

computational geometry may soon eliminate the computational bottleneck of interference tracking and inspection.

- Computer assisted collaboration environments must support natural communication modalities, which include video, voice, and gesture annotations. New techniques must be developed to support the intuitive creation and the interactive exploitation of cross-references between these modalities.

### "*TeamCAD: The First GVU/NIST Workshop on Collaborative Design*"

TeamCAD, the first GVU Workshop on Collaborative Design, was sponsored by NIST, chaired by Jarek Rossignac, and hosted on May 12-13, 1997 by Georgia Tech's GVU Center in Atlanta. The workshop brought together an international team of experts in collaborative design and in mechanical and architectural CAD, including 50 faculty members from 23 universities and 18 industrial participants representing software vendors and users. The objectives of the workshop were to assess the state of the art in collaborative CAD (i.e., in the technologies and practices for computer enhanced collaborative design in mechanical, architectural, and construction applications), to identify key social and technological issues, and to provide opportunities for research collaborations in this area. Workshop presentations included 10 project overviews, 20 research papers describing new paradigms for designing and inspecting architectural and mechanical CAD models in collaborative environments and new software and database technologies that support shared models.

During the break-out sessions the participants were divided into 3 teams and were asked to investigate specific aspects of collaborative design:

1. Carlo Sequin chaired a broad discussion focussed on the teamwork practices and on the need for new collaboration tools.
2. Chandrejit Bajaj chaired the charge towards the identification of the research challenges in collaborative design and inspection.
3. Bill Regli chaired discussions on how does the Internet and standardization affect data sharing issues.

The workshop was focused on the following questions: What are the various forms of collaborative design? How do people perform collaborative design in traditional engineering environments? Why is Computer Assisted Collaborative Design important? How will it affect the work practices? What major research issues must be addressed before significant progress is made in this area?

The breadth and diversity of the attendees enabled the workshop to assess current technology and its limitations:

- Tools for collaborative design, and how good they are
- Use of the internet and what are the technology issues

- Data exchange standards, their reliability and cost

While most attendees were technologists, we wanted to determine the extent to which human issues and corporate practices impede or enhance collaborative CAD.

- Why aren't current technologies inadequate for the users?
- What must be changed in current industrial practices to enhance the positive impact of digital collaboration?
- What other issues are standing in the way of a successful deployment of this technology?

Lastly, and with an important eye toward the future, the workshop attendees have formulated additional requirements for a successful future research strategy in TeamCAD:

- Involve representation of users in defining research directions and validating the results
- Establish procedures for measuring the benefits of TeamCAD technologies
- Address the usability issues (barrier for non-experts)
- Define a comfortable paradigm for interdisciplinary research in TeamCAD

From the numerous discussions, a few points of agreement have emerged.

- TeamCAD environments span a wide spectrum of possible embodiments: from purely asynchronous text-based product data management tools, all the way to fully immersive 3D collaborative design environments, where virtual models of the product, the collaborators, and their annotations are available in a shared space. Progress in these areas is substantial. However, many of the fundamental problems in TeamCAD are not purely technological, but also cultural and influenced by current business practices. There is a great deal of speculation about "what users want," but too little scientific data about real practices and expectations. Therefore, research in these technologies must be accompanied with empirical studies and formal evaluations. Yet, there is a clear lack of evaluation metrics for collaborative CAD. We must develop a common set of benchmark tasks and data and use them early and systematically to validate research results.
- Internet technologies are the essence of TeamCAD. Yet these technologies are progressing at a much faster rate than their exploitation by academics and industrial developers of CAD products. We must understand how to leverage object technologies and how to provide flexible security and higher-level data consistency checking on top of general purpose network and telecommunications technologies.
- Finally, little is available to TeamCAD developers to support true knowledge exchange techniques, such as design rationale, which greatly facilitate design reuse and maintenance.

The Workshop proceedings may be obtained by contacting the GVU center at (404) 894-4488.

## 11.25 M. Saiid Saiidi and E. Manos Maragakis

"*Bridge Research and Information Center (BRIC) at the University of Nevada, Reno*"

Prof. M. Saiid Saiidi and E. Manos Maragakis
University of Nevada, Reno, Civil Engineering Department
Reno, NV 89557
E-mail: saiidi@unr.edu

### Introduction

The deterioration and obsolescence of highway bridges is a very important component of the major infrastructure problem that the nation is facing today. At the present time nearly 40 percent of the bridges in the United States are either structurally deficient or obsolete. They need to be repaired, retrofitted, or replaced in many cases. At the same time, many recent earthquakes have, once more, reiterated the vulnerability of highway bridges to earthquake loadings. They have also reemphasized the necessity of continuing research and implementing the results related to effective seismic strengthening and earthquake resistant design of new bridges.

In recent years, extensive research has been performed on many aspects of bridge engineering both in the US and abroad. Some of these aspects include seismic design and retrofitting, a variety of analytical studies, development of non-destructive evaluation techniques, use of new materials and construction types, etc. It is becoming increasingly evident that a multi-disciplinary approach is necessary to address many of the problems. The wide variety of the research, along with its global scatter and the fact that it crosses several disciplines, make it necessary to have a mechanism for an effective dissemination of information on both past and current research, as well as on documented needs for new research. At the same time, significant technological advancements are available today for the dissemination, storage and presentation of information (Internet, CD ROM, magnetic tape, etc). This technology should be widely used for the dissemination of information on advancements in bridge engineering. It is also very important that a major national and international cooperative effort be established among bridge researchers in order to exchange information on research and design practices and to try to establish common research strategies. The proper dissemination of information along with the establishment of cooperative research efforts, will facilitate the identification of the most important research problems, will enhance the quality of research and will increase its efficiency by avoiding unnecessary duplication.

### Objectives

A five-year project was awarded by the National Science Foundation to the University of Nevada, Reno (UNR), in 1997. The main objective of the project is the establishment of a Bridge Research and Information Center (BRIC) in the Civil Engineering Department of the University of Nevada, Reno, with the following goals:

1. To establish complete collection of information on past, current, and future research

on a wide variety of topics on bridge engineering.

2. To use the latest computer technology for the processing and storage of the information and its dissemination to bridge researchers, engineers, consultants, companies as well as institutions, libraries and other similar organizations all over the world.
3. To provide on line video access of highlights of major tests and experimental data.
4. To continuously expand the center activities and to respond to users' requests by making adjustments and modifications to serve their needs.

As the first step, the job description for a full-time manager for the center was developed. The position was advertized and appropriate candidates were interviewed. A candidate has been selected and is due to assume the position in mid May, 1998. Among the initial tasks to be performed by the Center Manager will be the development of scope of the networking activities and the establishment a core user profile. Another key task will be the development of a plan and specifications for hardware and software acquisition, as well as milestones to monitor and assess the center's progress towards its goal.

**Link with National Center for Earthquake Engineering Research**

The University of Nevada, Reno (UNR), is one of the research institutions involved in the recently funded earthquake engineering center at the State University of New York in Buffalo (NCEER). One of the components of UNR activities related to NCEER is facilities and information networking. Because the focus of the earthquake engineering group at UNR has been on bridge structures, it is foreseen that BRIC will collect and disseminate bridge- and life line-related research and information elements including the educational component of earthquake engineering.

**Contribution to National Earthquake Engineering Facilities Network**

It has been a strong desire of the National Science Foundation and other research funding agencies to see researchers share testing facilities and research information in on efficient and streamlined basis. One of the major components of BRIC activities will be to facilitate networking of research data and facilities. BRIC will lead an effort to develop a standardized data collection protocol to simplify the exchange of information. Guidelines will be developed regarding instrumentation, format for data storage, and data types to be collected and made available to researchers and engineers. Initial demonstration projects will include video tapes that will be available on the Web showing reinforced concrete bridge columns and pipelines tested on large-capacity shake tables recently installed at UNR. A bridge column shake-table testing video recorded on May 7, 1998, is currently being processed for viewing on the Web.

## 11.26 Jakka Sairamesh and Christos M. Nikolaou

*"Frameworks for Distributed Query and Search in Scientific Digital Libraries"*

Dr. Jakka Sairamesh
IBM T. J. Watson Research
Hawthorne, NY, 10532
E-mail: jramesh@watson.ibm.com
URL: http://www.cs.columbia.edu/~jakka/home.html

Prof. Christos N. Nikolaou
University of Crete and ICS-FORTH
Leoforos Knossou, Heraklion, Crete 71409, Greece
E-mail: nikolau@ics.forth.gr
URL: http://www.ics.forth.gr/~nikolau/nikolau.html

We present an architecture for storing, managing and presenting geographic and coastal information of various kinds to a variety of users. The primary goal is to provide *transparent* access to information stored in various databases spatially distributed. We first present a scenario and then the architecture. Our architecture is based on OMG CORBA services and Object frameworks for access control and repository services. This work grew out of work mentioned in [1,2].

**Introduction**

Suppose we have an extensive database of the physical, chemical and biological properties of the coastal region under consideration. This database includes the bathymetry of the region and various physical, chemical and biological properties of the water column. Such properties include such phenomena as currents, wave spectra, wind spectra, salinity, temperature, chemical and biological concentrations [1,2].

A typical query of the database might be phrased as follows: "Find the region of 3d space within the given coastal region and the time interval, within which the concentration of a certain chemical or microorganism may exceed a certain value." Scientists may be interested in questions of this form in order to be better able to understand the physical and chemical processes in a coastal region. Local civil authorities may be interested in issuing permits for fishing or declaring certain coastal zones as health risks, inappropriate for tourism, swimming, etc.

More generally interrogation of the database can be thought of as a query of the type "find a subset of a given set containing points with a specified property." From a logical set theoretic point of view, this interrogation operation is the computation of an intersection set created from two sets characterized by respective logical properties which are supposed to hold simultaneously for all elements of the set. This logical Boolean intersection operation is one of the most common DB interrogations and may require time consuming searches in very large databases. Here the sets involved are not amorphous clouds of discrete records but rather connected, smoothly shaped high-dimensional objects that represent certain multivariate functions.

Scientists interested in using existing programs (or algorithms, numerical techniques) to study the properties of the fresh data points (collected by the sensors) in the databases, and also look at previous research papers, and possibly some annotations, could ask for very complex queries. From the users view, the information system must provide a transparent view to the existing programs or numerical techniques, databases, and documents in an integrated fashion. This could mean searching for existing programs (which are indexed by keywords) and applying them to new data which could be located elsewhere.

**Architecture**

The web provides a simple way to represent and access documents, but for a large distributed system that needs to work together to solve the kinds of queries mentioned above, several issues of naming, access control, metadata, repositories naturally emerge. Though these technologies currently exist independently, an integrated solution to indexing, querying and presenting information objects customized to the various users of the system is still in its early stages [2].

Several efforts [1,2,3] are underway to solve these issues for a large spatially distributed database system. With the rapidly emerging Java enabled technologies, access to various legacy databases and legacy systems is becoming feasible via the Web. In addition, the Web is steadily embracing object-oriented frameworks such as OMG CORBA for better services and flexibility. Leveraging on these new technologies, the *basic* components of our architecture are the following:

- **Metadata and indexing/searching** services for GIS information such as areas/regions of maps selected by the users [2,3]. There are a few standards for Metadata definitions such as FGDC, developed by the US federal agency for digital geospatial data. The FGDC standard provides definitions for a few fields, along with their relations within a hierarchical structure.
- **Naming and Repository** services (servers). We use Naming services from OMG for locating and interacting with the information objects. The name services are very essential for efficient retrieval of objects. We envision, agents to help users search for objects by using the meta, name and index services. We plan to use OMG CORBA repository interface frameworks to access data objects located across the network.
- **Interfaces** for information sources such as underwater sensors, satellites (images) and aerial photography that provide data and image collection services. Also included are simulation models that populate the databases.
- **QoS in search and retrieval**: We address issues on QoS (quality of service) for search and retrieval of information objects. Our goal is to provide mechanisms for efficient usage of resources; such as processing, I/O and memory in the servers, and bandwidth and buffers in the networks. Novel performance based architectures are being investigated for optimized retrieval and presentation in large scale information systems.

**Description of Architecture**

Clients submit simple or complex queries via the world-wide web interface to the Digital Library system. The queries are submitted via Java enabled browsers. The system is accessible by scientists, engineers, local authorities and system administrators, but they all have different access restrictions. Users requests invoke various services such as meta, index and search to locate the objects which match the user-query. The agents (agents representing users) will fork various sub-agents to search in parallel across various databases (legacy and new) and collect/present the various information objects to the user. We assume that documents are stored in Digital Library or document databases, and metadata services are provided to access the documents.

From the scenario above, when a user selects a region of a map through the WWW browser, the coordinates of the region will be used to index the appropriate information about the region. This implies a metadata service that maps the regions of the map to the appropriate region-information. Therefore, users can zoom into a region of the map (or image) and query for various properties about the region or perform some operations on-line. It is likely that the information about a region could be dispersed across several database sites (for example the detailed image of the region will be stored separately from the data objects). For this service, distributed search queries will be sent to the various databases to obtain the objects. Metadata services describing the GIS objects will be used to index/search for the appropriate GIS objects (multi-dimensional data and images).

**Conclusions**

In this paper, we outline the issues in providing a distributed architecture for accessing the multimedia information from spatially separated scientific databases. We provide a OMG based object framework and architecture for scientists, engineers and administrators to access the scientific databases.

1. C. Houstis, C. Nikolaou, M. Marazakis, N. M. Patrikalakis, R. Pito, J. Sairamesh, and A. Thomasic, *Design of a Data Repositories Collection and Data Visualization System for Coastal Zone Management of the Mediterranean Sea*, ECRIM Technical Report TR97-0200, June 1997.
2. *The Alexandria Project: Towards a Distributed Digital Library with Comprehensive Services for Images and Spatially Referenced Information*, University of California, Santa Barbara. http://alexandria.sdc.ucsb.edu/.
3. *An Electronic Environmental Library Project*, University of California, Berkeley. http://elib.cs.berkeley.edu/.

## 11.27 Clifford A. Shaffer

"*Collaborative Problem Solving Environments*"

Prof. Clifford A. Shaffer
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061
E-mail: shaffer@cs.vt.edu
URL: http://www.cs.vt.edu/∼shaffer/

Long before digital libraries, code reuse, or the Internet became popular terms — or even distinct topics of research — the numerical analysis community created a digital library of reusable code available via the Internet. Netlib is still widely used today by scientists and engineers in a wide range of disciplines. Today, successor efforts such as NetSolve are improving the access of scientists and engineers to high-quality software. It is important to note that, while crucial to the scientific endeavor, these efforts have little to do with numerical analysis or high performance computing. Instead, they have everything to do with marshaling and distributing resources so that a wide community can make the most of its investment in high-quality software.

The scientific computing research community would achieve greater productivity if it can address a complex of process management issues that are endemic to scientific and engineering research. Our next goal should be to provide appropriate software support and process management standards that will allow scientists and engineers to make the most of their existing codes and computing resources.

The experience at Virginia Tech and elsewhere in dealing with large-scale engineering design and analysis problems has demonstrated the following critical problems: (1) It is difficult to integrate codes from multiple disciplines developed by a diverse group of people on multiple platforms located at widespread locations; (2) It is difficult to share software between potential collaborators in a multidisciplinary effort — difficult even for a team to continue using a research code once the author has left the group; (3) Current tools for synchronous collaboration are inadequate.

Integrating codes from different disciplines raises both pragmatic and conceptual issues. Pragmatically, the issue is how best to support the interoperability of independently-conceived programs residing on diverse, geographically distributed computing platforms. Another pragmatic concern is that large, complicated codes now exist that cannot simply be discarded and rewritten for a new environment. However, interoperability is best achieved by adherence to common protocols of data interchange and the use of clearly identified interfaces. The notions of interfaces and protocols lead directly to the domains of object-oriented software and distributed computing. Thus, a key issue is how to unify legacy codes, tied to specific machine architectures, into an effective whole. Conceptually, the key issue is how to foster coordinated problem solving activities among multiple experts in different technical domains, and leverage existing codes and computer hardware resources connected by the Internet.

Many of these problems can be resolved through the use of collaborative problem solving environments (CPSE). A CPSE is a system that provides an integrated set of high level facilities to support groups engaged in solving problems from a proscribed domain. A CPSE allows users to define and modify problems, choose solution strategies, interact with and manage the appropriate hardware and software resources, visualize and analyze results, and record and coordinate problem solving tasks. Perhaps most significantly, users communicate with a CPSE at a higher level than the language of a particular operating system, programming language, or network protocol. Expert knowledge of the underlying hardware or software is not required.

Key elements to CPSEs include the following:

- CPSEs should be based on aggressive use of new software engineering technology for code reuse and interoperability such as component frameworks (one example is JavaBeans). A properly designed component framework architecture for scientific computing would allow users to link together data generators such as simulations, data filters to convert data formats, and data visualizations.
- An integral part of the CPSE must be support for access to distributed resources. For example, our own work uses JavaBeans acting as proxies for simulation codes and files on high-performance computers. These proxy beans are linked together on a local workstation in ways that allow scientists and engineers to combine simulation codes to produce new results, and to link these results in turn to visualization programs.
- The CPSE must also support real-time (synchronous) collaboration among teams of scientists and engineers. Team members should be able to see and manipulate the workspace together despite being in separate places.
- The CPSE must include process support in ways that encourage effective documentation of simulation and visualization codes, along with true interoperability. As mentioned above, a primary problem encountered by multidisciplinary groups is that existing codes are hard to use by newcomers. Within a CPSE, each component should carry with it effective support for understanding the conditions under which the component should be used, and how it should link to other components. It is equally important that the builders of new components be motivated to provide the necessary documentation.
- The CPSE should support changing models and conducting multiple experiments. Basic tools such as effective version control for experiments, and automatically documenting the models and parameters used for a given run, would be helpful. So would automatic recompilation of the underlying simulations when the model is changed by selecting a new subroutine within a larger simulation code.
- Finally, the CPSEs should be created with a proper understanding of how the application domain works. We still do not understand enough about how multidisciplinary teams work to be sure what the best software tools will be, and so we need to study such teams in operation. Participatory design techniques will be crucial as Computer

Scientists work with the scientists and engineers who are the domain experts to create effective CPSEs.

### URLs to ongoing work

JAMM: Java Applets made Multi-user. `http://simon.cs.vt.edu/rJAMM/`

Sieve: A Collaborative Interactive Modular Visualization Environment.

`http://simon.cs.vt.edu/sieve/`

WBCsim: a prototype Problem Solving Environment for wood-based composites.

`http://wbc.forprod.vt.edu/pse/`

## 11.28 Amit P. Sheth

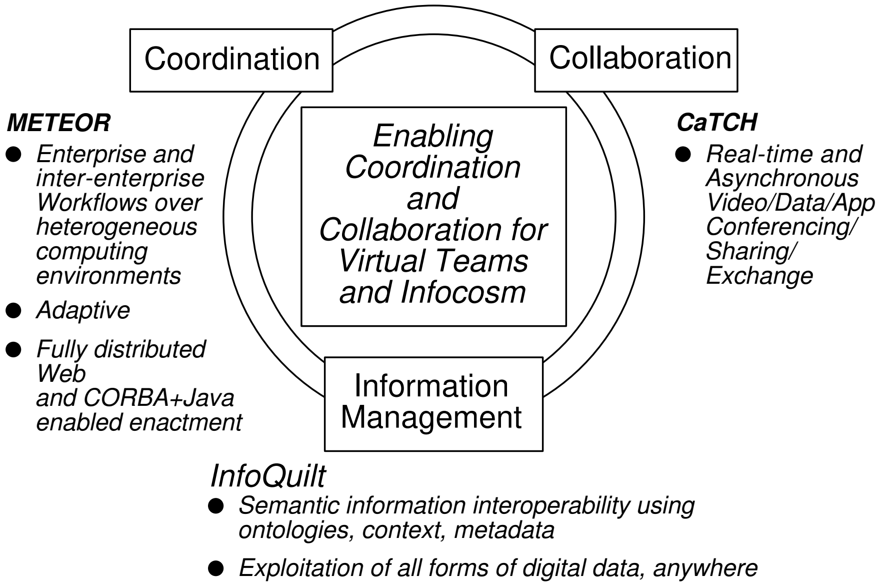
*"Enabling Virtual Teams and Infocosm: A Research Agenda Involving Coordination, Collaboration and Information Management"*

Prof. Amit P. Sheth
University of Georgia
Large Scale Distributed Information Systems Laboratory
Athens, GA 30602-7404
E-mail: amit@cs.uga.edu
URL: http://lsdis.cs.uga.edu

This position paper outlines a research agenda pursued at the LSDIS Lab. Our current research is in workflow management, collaboration and integration of heterogeneous digital media. It is increasingly moving towards seamless coordination and collaboration for virtual teams, and Infocosm enabled by semantic interoperability and information brokering.

**Current Research**

Figure below provides an overview of three of our main projects: METEOR, CaTCH, and InfoQuilt.




**METEOR for Coordination**

The METEOR (Managing End-To-End OpeRations) research, technology and system focus on *enterprise wide and inter-enterprise mission critical workflow applications in heterogeneous computing environments.* The subset of key technical features or capabilities of METEOR include:

- a workflow model with its unique approach to specify human and automated tasks and inter-task dependencies, and a comprehensive graphical designer

- fully distributed architecture and scheduling
- automatic code generation from a graphical design, repository for reuse
- two fully distributed enactment services (engines) for heterogeneous computing environments using Web (WebWork) and CORBA+Java (ORBWork) infrastructures/technologies
- techniques to support error handling, recovery, and scalability
- standards support, usage, and influence

Components of the METEOR system have been trialed and evaluated by LSDIS's industry partners, used for education, and the technology is now commercialized and licensed for real-world applications. The current research focuses on adaptive workflows with support for dynamic changes, exception handling, and (multi-level) security.

**CaTCH for Teleconsulting and Collaboration**

The CaTCH (Collaborative TeleConsulting for Healthcare) project investigates integration of several fast-evolving technologies to support *real-time* or *asynchronous* (store-and-forward), "high bandwidth collaboration." Component technologies include:

- Desktop and mobile conferencing including videoconferencing, white-boarding, and application sharing
- Multimedia capture and repositories
- Web, network computing and streaming video
- Scheduling/workflow
- Agents, web-objects, security, etc.

A prototype is undergoing trial with the participation of physicians at the Medical College of Georgia, to determine if it can support Web-based teleconsulting between a remote healthcare organization and a hospital based sub-specialist (pediatric echocardiologist). The scenario for the current trial includes referral and evaluation of pediatric echocardiograms.

**InfoQuilt for Semantic Interoperability, Information Brokering and Heterogeneous Digital Media**

Exploiting increasing amount of diverse Web-accessible information assets is a well-recognized opportunity and challenge. The InfoHarness system (jointly with Bellcore) developed a metadata-based access to heterogeneous textual data and structured databases, without relocating, reformatting and restructuring the data, and also resulted in a commercial system AdaptX/Harness marketed by Bellcore. Its extension, the VisualHarness system added access to distributed image repositories, and supported customizable access with a combination of keyword-based, attribute-based and (image) content-based search.

## 11.29 Knut Streitlien and James G. Bellingham

"*Distributed Computation Issues in Marine Robotics*"

Dr. Knut Streitlien and Dr. James G. Bellingham
Massachusetts Institute of Technology
Sea Grant College Program, Autonomous Underwater Vehicles Laboratory
Cambridge, MA 02139-4307
E-mail: knut@mit.edu, belling@mit.edu
URL: http://web.mit.edu/seagrant/www/auv.htm

The Autonomous Underwater Vehicles (AUV) Laboratory at MIT has for the last 10 years designed, built, and operated robotic vehicles for underwater exploration. The current generation, Odyssey IIb, is 3 years old and have completed 400 dives in the course of 17 field expeditions. There are 5 such vehicles in existence. Unencumbered by umbilical cables and free from the requirements of supporting human pilots, they provide an economical and agile platform for a variety of ocean measurements. Each vehicle weighs about 150kg and has a length of 2m, which is a kind of middle ground in AUV size; large enough to allow a variety of off the shelf payloads, small enough for safe deployment off almost any vessel in sea states up to 5. The batteries, computers and attitude sensors are housed in two 17" glass spheres set in a flooded polyethylene hull. The spheres permit deployment to 6000m depth, a capability unique to the Odyssey IIb. The current operational envelope is a speed of 1.5m/s and a duration of 3h between recharges. Typical deployments take place from our mission command central, inside a standard 20' cargo container where the operator configures missions via a TCP/IP tether connection. Once he or she starts a mission, the deck crew disconnects the AUV and off it goes after a preset delay. The vehicle's intelligence is based on state configured layered control, which provides good real time performance and a flexible development environment for complex behaviors, with low computational effort. At the end of a mission, the vehicle floats on the surface, and upon retrieval it is connected again for data transfer.

Recent developments in the laboratory have focused on enhanced autonomy, extended prescence in remote areas, and higher endurance, through a program called Autonomous Ocean Sampling Networks. We have designed and built, in cooperation with the Woods Hole Oceanographic Institution, a dock station for the AUV, placed on a mooring 500m below the surface. The vehicle is now able to home in on the dock, attach itself, and transfer data and power autonomously. Through a satellite connection at the mooring surface expression, the operator can access the AUV from anywhere in the world, and have remote control of sampling programs that last several months. Vehicle missions, up to 12h, can run on an automated schedule or be initiated via satellite. Low bandwidth acoustic communication to the dock from the vehicle during missions is now also available. We expect to exercise this system in full by the end of the year.

One objective of the AUV Laboratory is to further develop the operational and physical capabilities of its vehicles through integration with powerful field modeling software. The finite velocity of AUVs results in a coupling of space and time through the survey trajectory, and most energy storage devices places constraints on the extent of their surveys. To get efficient estimates for synoptic oceanic fields, AUVs must therefore adapt their sampling strategies by concentrating measurements in scientifically interesting regions of the survey domain. This is best achieved by modeling software such as the Harvard Ocean Prediction System, which assimilates observations from all available sources by melding these optimally with dynamical models to produce nowcasts and forecasts for all the modeled quantities and their error variances. The AUV will be a node in a heterogeneous network of modeling and observation resources, seeking to minimize the uncertainty in the field estimates through model based adaptive sampling.

The laboratory also has a computer model that can represent the AUV in simulations. This model uses the same algorithms as the vehicle, only the physical properties of the AUV is represented in an idealized way. This simulator will play an important role in Observing System Simulation Experiments, where the entire network described above is exercised in simulation before actual experiments are fielded.

Another trend that is likely to continue is the miniaturization of ocean sensors, which then become potential payloads for the AUV. These can take novel *in situ* measurements such as optical or biological properties, or in the case of acoustics; integral as well as local samples. Our data holdings are therefore bound to become more complex and multidisciplinary in nature. It is important to the laboratory to make this data readily available to other investigators, both in real time and the long term.

The AUV Laboratory's interest in this workshop follows from these objectives. We want to be able to register our vehicles, simulation tools, and data through metadata or other means, so that they can be made part of larger scale workflows. Our ambition is to make AUVs full nodes in a communication/knowledge/observing network, in a way that deals naturally with the limitations of low bandwidth and long time lags inherent in the communication channels. The AUVs will rely on the distributed computing services of the network for their model based adaptive sampling, which is computationally intensive. We wish to participate in an early stage in the development of standards in this area, both to contribute to them and to guide our future developments.

1. J. G. Bellingham, C. A. Goudey, T. R. Consi, J. W. Bales, D. K. Atwood, J. J. Leonard, and C. Chryssostomidis, "A second generation survey AUV," *IEEE Conference on Autonomous Underwater Vehicles*, MIT, 1994.

2. J. G. Bellingham and J. S. Willcox, "Optimizing AUV Oceanographic Surveys," *IEEE Symposium on Autonomous Underwater Vehicle Technology*, Monterey, CA, June 1996.

3. T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb, "Autonomous Oceanographic Sampling Networks," *Oceanography*, 6(3):86–94, 1993.

4. P. Malanotte-Rizzoli, ed., *Modern Approaches to Data Assimilation in Ocean Modeling*, Elsevier Oceanography Series, 1996.

5. A. R. Robinson, "Physical Processes, Field Estimation and an Approach to Interdisciplinary Ocean Modeling," *Earth-Science Reviews*, 40:3–54, 1996.

## 11.30 George Turkiyyah and Gregory L. Fenves

*"Simulation Tools to Support Performance-Based Earthquake Engineering Design"*

Prof. George Turkiyyah
University of Washington, Department of Civil Engineering
Seattle, WA 98195
E-mail: george@ce.washington.edu
URL: http://www.ce.washington.edu/~george/

Prof. Gregory L. Fenves
University of California, Berkeley, Pacific Earthquake Engineering Research Center
Berkeley, CA 94720-1710
E-mail: fenves@ce.berkeley.edu
URL: http://www.ce.berkeley.edu/~fenves/

The push towards more rational and mechanistic procedures for the assessment of the behavior of structures subjected to seismic loads has fueled the need for software tools that allow engineers to develop simulation models that can be used in such assessments. We briefly describe some characteristics of these simulations as they relate to distributed intelligent repositories and scientific problem solving environments.

The simulations are multi-domain. Earthquake engineering simulations use information from basin-scale simulations as generally performed by geophysicists, local detailed nonlinear soil simulations as generally performed by geotechnical engineers, and nonlinear simulation of the elements of the structural system as generally performed by structural engineers. These simulations may be coupled, or the results from one serve as a partial set of inputs and boundary conditions to another. Therefore, there is a great need for tools to readily translate the output of one model to a form that can be used by the next model. Such intelligent translators would go a long way towards more realistic, collaborative models.

The data needed to generate the necessary models often resides in different locations and may not always be readily available. For example, soil information, hazards information, etc. are stored in GIS databases maintained by various organizations (USGS, utility companies, departments of transportations, local agencies, etc.). However the data may be at different resolutions, may be very sparse in some spatial regions while a lot denser in others, and may have different levels of reliability associated with it. Common complaints from engineers trying to access the necessary data are that it is difficult to find information at the scale and quality needed, not to mention the heterogeneous formats in which the data is available. Technologies being developed to search large digital libraries may provide the necessary query mechanisms to search and merge/extract the data needed from distributed national repositories. Associated with this access problem, are also some public policy questions: how much information should be available? Access restrictions may have to be imposed depending on the quality, and level of detail of the data, its intended uses, etc.

Rarely are the simulations performed one-shot processes. It is often the case that an engineer would build a model and perform a large number of parametric studies to evaluate the sensitivity and relative importance of various factors on the response of the system. Using today's modeling and analysis tools, the generation of these parametric models and the exploration of model response is done at a fairly low level, with most engineers developing their models "from scratch." This is obviously inefficient. We envision a software library of reusable model fragments—"component models"—that represent the behavior of the various entities of earthquake simulations at various levels of granularity. This would allow a user to quickly build a model by plugging together the components needed. A high level scripting language would provide an interpreted environment that allows the "gluing" of these reusable components and the efficient generation of the models needed.

The simulations are generally large scale requiring significant computing resources. Many engineers who may wish to perform these simulations may not have the necessary resources locally and will need to have access to remote servers. Such "analysis servers" would accept a model and a specification of the resources needed, configure the hardware needed to run the simulations and return the results to the user. The visualization and interpretation of model results is another challenging area where interaction with intelligent repositories may be needed. For example, model interpretation requires checking analysis results against standard specifications in building codes. The presentation of the results at a high level and the identification/checking of the applicable specifications from "specification servers" are interesting tasks that have not been explored. They may be difficult because they require access not only to the model results but to the CAD models and GIS maps from which the model was generated.

## 11.31 Alvaro Vinacua

"*Issues in Distributed CAD*"

PROF. ALVARO VINACUA
Polytechnic University of Catalonia, I.R.I.
Edifici Nexus 202, C. Gran Capità 2-4
E08034 Barcelona, Spain
E-mail: alvar@turing.upc.es
URL: http://www.lsi.upc.es/~alvar/

The views expressed herein are cast from the point of view of CAD and CAGD technology, and are undoubtly colored.

As our resources grow, we are attempting and executing the design of ever more complex, intricate systems. In doing so, some aspects gain more and more relevance:

- Design over a wide geographic area: Some of the actors in the process may be located at distant spots.
- Concurrency: To shorten the time from conception to manufacturing, concurrency in the design becomes essential when dealing with ever larger projects.
- Multidisciplinary component: often very large, complex systems will involve many different areas of expertise in their conception and design. It brings about the need to address different views and intuitive user interfaces, but also to handle transparently mixes of data from diverse domains.

The problems we face may thus have partial solutions in different realms. Many current techniques become relevant, but possibly in a mix not addressed before. These include:

- Distributed Databases
- Metadata
- Standard Data Interchange Formats
- Distributed Applications
- Networking
- Data Compression Techniques (and *ad hoc* data compression, tuned for certain kinds of data).
- Computer Supported Collaborative Work

Concerning the global sharing of geometrical and product data, many other specific aspects deserve attention:

- Data sharing. Over many years of development of CAD applications and after a widespread consolidation of CAD within the industry, this issue has shown large difficulty. A case may be made, even today, that it is still wide open. In any case present data interchange standards are not designed to share those models over large, probably slow, connections, nor are they geared to be easily accepted by applications from other domains. Standards supporting incremental transmission, partial model extraction and multiple views seem desirable.

- If effective collaboration and concurrent design are going to happen, we need to explore further mechanisms that allow the use of partial information without acquiring the whole model. Simple data interchange will just not suffice. It is unclear how far we may expect to go in this direction, and how feasible it may be to use this method in meshing with applications and data from other realms.
- Methods of partial publication of design data may allow companies to publish data to facilitate their customers the incorporation of their technology without compromising their own privileged information.

While the communications resources continue their non-stop growth and improvement, we must strive to find better ways to exploit them within the realms of research and the productive economy.

## 11.32 Jack C. Wileden and Alan Kaplan

"*Formal Foundations for Transparent Interoperation*"

PROF. JACK C. WILEDEN
University of Massachusetts, Department of Computer Science
Amherst, MA 01003-4610
E-mail: wileden@cs.umass.edu
URL: http://www-ccsl.cs.umass.edu/~jack/

PROF. ALAN KAPLAN
Clemson University, Department of Computer Science
Clemson, SC 29634-1906
E-mail: kaplan@cs.clemson.edu
URL: http://www.cs.clemson.edu/~kaplan/

### Introduction

Our research is directed toward developing computer science foundations, both theoretical and practical, for transparent interoperation. The primary objective of our work is to hide the boundaries between heterogeneous dat repositories or between data repositories and applications that need access to these repositories. Scientific and engineering applications increasingly depend upon software systems comprising both code and dat components that come from a wide variety of diverse sources. Hence, appropriate theoretical and practical foundations for transparent interoperation will make it easier to develop, reuse, share, and maintain scientific and engineering applications.

In the remainder of this position paper, we suggest some features that we believe are critically important to formal foundations for transparent interoperation. We also briefly discuss some initial steps that we have taken toward developing such foundations. In a companion position paper [9] we discuss the practical aspects of our work, specifically the development of a new, highly transparent approach to interoperation called PolySPIN. A collection of prototype, automated tools supporting the use of the PolySPIN approach and some experience with its application are also described here.

### Toward Formal Foundations

Interoperation, already an important issue for scientific and engineering software systems, will become even more important as data generation rates continue to escalate and connectivity makes both data and code increasingly available. Unless interoperation is essentially transparent, it could become a major impediment to the development, reuse, sharing, and maintenance of scientific and engineering applications. Unfortunately, current approaches to interoperation are based on such things as standardized data interchange formats, intermediate notations for interface description, or ad hoc wrapping of data repositories and code components. None of these provide the required transparency and hence we advocate development of suitable alternatives.

- *No intermediaries*: Many current approaches define interoperation in terms of an intermediate representation, model, of notation, which results in loss of potential interoperation and weakened compositions [5]. Foundations for a direct, intermediary-free approach to interoperation will support stronger compositions, both of software components and data, and of representations, analyses, and tools used in scientific and engineering applications. Such an approach will also lead to reduced inertia when components, notations, representations, analyses, and tools evolve.

- *Contextualized analyses*: Current approaches implicitly assume a fixed, common context for interoperating software components and data, which is unrealistic for the complex interoperation typical of scientific and engineering applications. More suitable foundations will explicitly address context, provide analysis of consistency, and other properties of context compositions of software components and data.
- *Applicability to dynamic composition*: Current approaches implicitly assume that application software systems are statically structured. This is increasingly unrealistic as evolving software systems proliferate and as modern programming languages and other infrastructure provide growing support for dynamic structure. Appropriate foundations for transparent interoperation will provide modeling, analysis, and implementation support for dynamically-structured systems.

Our work to date on foundations for transparent interoperation has focused on modern programming languages, complex object-oriented user-defined abstract types, and persistent objects (e.g., [4, 10]). More specifically, we have been concerned with interoperability problems in programming languages, e.g., how an application written in a particular programming language accesses data that was defined and created by an application written in a different distinct programming language. To improve our understanding and assessment of suitability of different interoperability mechanisms, we have begun to develop a taxonomy of important facets of interoperability approaches. (See [7, 2] for details.) We have also developed a formal model of name management, called Piccolo, with an operational semantics based on evolving algebras [6, 3]. This model allows for the precise specification of some formal properties and analyses of name management approaches, and hence provides a starting point for contextualized interoperation analyses. In addition, we are developing an intial formal model, based on concepts from signature matching and object-oriented type theory, on which to base reasoning about, and automated support for, cross-language type compatibility [1]. This initial formal model allows developers to precisely express compatibility relationships among object-oriented types that are defined in different programming languages.

**Research Directions**

Although much of our work has set in the context of programming languages, preliminary work [8] suggests that our results can contribute to development of suitable formal foundations for transparent interoperation in scientific and engineering software systems. Towards that end, we plan to extend and improve our initial models so that they can account for other crucial aspects of transparent interoperation, such as versions, evolution, and dynamic systems modification.

1. Barrett, D. J., *Polylingual Systems: An Approach to Seamless Interoperability*, Ph.D. thesis, University of Massachusetts, Amherst, MA, February 1998.
2. Barrett, D. J., Kaplan, A. and Wileden, J. C., "Automated Support for Seamless Interoperability in Polylingual Software Systems," *Fourth Symposium on the Foundations of Software Engineering*, San Francisco, CA, October 1996.
3. Kaplan, A., *Name Management in Convergent Computing Systems: Models, Mechanisms and Applications*, Ph.D. Thesis, Technical Report TR–96–60, Department of Computer Science, University of Massachusetts, Amherst, MA, May 1996.
4. Kaplan, A., Myrestrand, G. A., Ridgway, J. V. E., and Wileden, J. C., "Our SPIN on Persistent Java: The JavaSPIN Approach," *Proceedings of the First International Workshop on Persistence and Java*, Drymen, Scotland, September 1996.
5. Kaplan, A., Ridgway, J. V. E., and Wileden, J. C., "Why IDLs Are Not Ideal," *Ninth IEEE International Workshop on Software Specification and Design*, Ise-Shima, Japan, April, 1998.
6. Kaplan, A. and Wileden, J. C., "Conch: Experimenting with Enhanced Name Management for Persistent Object Systems," *Proceedings of the Sixth International Workshop on Persistent Object Systems*, Washington, D.C., October 1995.
7. Kaplan, A. and Wileden, J. C., Toward Painless Polylingual Persistence, *Proceedings of the Seventh International Workshop on Persistent Object Systems*, Cape May, NJ, May 1996.
8. Kaplan, A. and Wileden, J. C., "Foundations for Transparent Data Exchange and Integration," *CODATA Conference on Scientific Technical Data Exchange and Integration*, Washington, D.C., December 1997.
9. Kaplan, A. and Wileden, J. C., "Practical Foundations for Transparent Interoperation," *NSF Workshop on Distributed Information, Computation, and Process Management for Scientific and Engineering Environments*, Herndon, VA, May 1998.
10. Ridgway, J. V. E., Trall, C., and Wileden, J. C., "Toward Assessing Approaches to Persistence for Java," *Proceedings of the Second International Workshop of Persistence and Java*, Half Moon Bay, CA, August 1997.

## 11.33 Roy D. Williams

"*Questions and Recommendations for Scientific Digital Archives*"

Dr. Roy D. Williams
California Institute of Technology, 158-79
Pasadena, CA 91125
E-mail: roy@caltech.edu
URL: http://www.cacr.caltech.edu/∼roy/

**Who are the Librarians?**

There are many creators or users of a scientific data archive, but few librarians or administrators. So the question we are left with is who are the librarians for this increasing number of ever more complex libraries? Who will be ingesting and cataloguing new data; maintaining the data and software; archiving, compressing and deleting the old. Somebody should be analyzing and summarizing the data content, assuring provenance and attaching peer-review; encouraging interaction from registered users and project collaborators. Perhaps the most important function of the librarian is to answer questions and teach new users.

**How Long Will the Archive Last?**

Scientific data archives contain valuable information that may be useful for a very long time, for example climate and remote-sensing data, long-term observations of astronomical phenomena, or protein sequences from extinct species. On the other hand, it may be that the archive is only interesting until the knowledge has been thoroughly extracted, or it may be that the archive contains simulations that turn out to be flawed. Thus a primary question about such archives is how long the data is intended to be available, followed by the secondary questions of who will manage it during its lifetime, and how that is to be achieved.

Data is useless unless it is accessible, unless it is catalogued and retrievable, unless the software that reads the binary files is available and there is a machine that can run the reader programs. While we recognize the finite lifetime of hardware such as tapes and tape readers, we must also recognize that files written with specific software have a finite lifetime before they become incomprehensible binary streams. Simply copying the whole archive to newer, denser media may solve the first problem, but to solve the second problem the archive must be kept "alive" with upgrades and other intelligent maintenance.

A third limit on the lifetime of the data archive may be set by the lifetime of the collaboration that maintains it. At the end of the funding cycle that created the archive, it must be transformed in several ways if it is to survive. Unless those that created the data are ready to take up the rather different task of long-term maintenance, the archive may need to be taken over by a different group of people with different interests; indeed it may pass from Federal funding to commercial product. The archive should be compressed and cleaned out before this transformation.

**Text-based Interfaces**

While the point-and-click interface is excellent for beginners, mature users prefer a text-based command stream. Such a stream provides a tangible record of how we got where we are in the library; it can be stored or mailed to colleagues; it can be edited, to change parameters and run again, to convert an interactive session to a batch job; the command stream can be merged with other command streams to make a more sophisticated result; a command stream can be used as a start-up script to personalize the library; a collection of documented scripts can be used as examples of how to use the library.

We should thus focus effort on the transition between beginner and mature user. The graphical interface should create text commands, which are displayed to the user before execution, so that the beginner can learn the text interface.

In a similar fashion, the library should produce results in a text stream in all but the most trivial cases. The stream would be a structured document containing information about how the results where achieved, with hyperlinks to the images and other large objects. Such an output would then be a self-contained document, not just an unlabeled graph or image. Because it is made from structured text, it can be searched, archived, summarized and cited.

### XML as a Document Standard

Extensible Markup Language (XML), which has been developed in a largely virtual W3 project is the new "extremely simple" dialect of SGML for use on the web. XML combines the robustness, richness and precision of SGML with the ease and ubiquity of HTML. Microsoft and other major vendors have already committed to XML in early releases of their products. Through style sheets, the structure of an XML document can be used for formatting, like HTML, but the structure can also be used for other purposes, such as automatic metadata extraction for the purposes of classification, cataloguing, discovery and archiving.

A less flexible choice for the documents produced by the archive is compliant HTML. The compliance means that certain syntax rules of HTML are followed: rules include closing all markup tags correctly, for example closing the paragraph tag <p> with a </p> and enclosing the body of the text with <body> ... </body> tags. More subtle rules should also be followed so that the HTML provides structure, not just formatting.

### Authentication and Security

There is a need for an integration and consensus on authentication and security for access to scientific digital archives. Many in the scientific communities are experts in secure access to Unix hosts through X-windows and text interfaces. In the future we expect to be able to use any thin client (Java-enabled browser) to securely access data and computing facilities. There should be (at least) access-control levels as follows:

- Public access: anybody on the Internet can get data in the public area, and they can also run certain query engines that use limited resources. This area also contains the "home-page" for the project with the usual items.
- Low-security access: this area can be accessed with a clear-text password, control by domain name, HTTP authentication, a password known to several people, or even "security through obscurity". This kind of security emphasizes ease of access for authenticated users, and is not intended to keep out a serious break-in attempt. Appropriate

types of data in this category might be prototype archives with data that is not yet scientifically sound; or data where the principal investigator still has first-discovery rights.

- High-security access: access to these data and computing resources should be available only to authorized users, to those with root permission on a server machine, or to those who can watch the keystrokes of an authorized user. The data may be valuable intellectual property; and access at this level allows copying and deletion of files and long runs on powerful computing facilities. Appropriate protocols include Secure Socket Layer (SSL), Pretty Good Privacy (PGP), secure shell (ssh), and One-Time Passwords (OTP).

Once a user is authenticated to one machine, we may wish to do distributed computing, so there should be a mechanism for passing authentication to other machines. One way to do this is to have trust between a group of machines, perhaps using ssh; another way would be to utilize a metacomputing framework such as Globus, that provides its own security. Once we can provide effective access control to one Globus server, it can do a secure handover of authentication to other Globus hosts. Just as Globus is intended for heterogeneous computing, the Storage Resource Broker provides authentication handover for heterogeneous data storage.

**Exceptions and Diagnostics**

One of the most difficult aspects of distributed systems in general is exception handling. For any distributed system, each component must have access to a "hotline" to the human user or log file, as well as diagnostics and error reporting at lower levels of urgency, which are not flushed as frequently. Only with a high quality of diagnostic can we expect to find and remove not only bugs in the individual modules, but the particularly difficult problems that depend on the distributed nature of the application.

**Data Parentage, Peer-review, and Publisher Imprint**

Information without provenance is not worth much. Information is valuable when we know who created it, who reduced it, who drew the conclusions, and the review process it has undergone. To make digital archives reach their full flower, there must be ways to attach these values in an obvious and unforgeable way, so that the signature, the imprint of the author or publisher stays with the information as it is reinterpreted in different ways. When the information is copied and abstracted, there should also be a mechanism to prevent illegal copying and reproduction of intellectual property while allowing easy access to the data for those who are authorized.

## 11.34 Peter R. Wilson

"*Product Knowledge*"
Dr. Peter R. Wilson
Boeing Commercial Aircraft Company Group
Seattle, WA 98124-2207
E-mail: peter.r.wilson@boeing.com

There are at least three major trends affecting industry, each of which has an impact on the broad area of computer science and computer applications.

**Increased globalization and consolidation:**

The trend towards increased globalization is forcing companies to do business with new partners and customers. On the other hand, consolidation is forcing companies to deal with the problems of trying to integrate computer applications from disparate vendors when two formerly independent businesses merge.

**Employee empowerment and 'do more with less':**

Employee empowerment, to be successful, means moving decision making lower down the management chain, thus leading to a need to provide more information across a broader spectrum of recipients and also for the generation of information to occur lower in the organization.

**Technologies like the Web and CORBA:**

Technologies like the Web and CORBA open up new ways of delivering and receiving information. They provide infrastructures for exchanging information between both humans and computer applications.

Traditionally the exchange of information and knowledge in an industrial environment has been paper-based. That is not to say that computer systems and applications are not utilized but rather that the systems and applications do not talk to each other, requiring human intervention to translate between the output from one application to the input for a following application. As one example, 90% of the CPU time used by industries involved in the oil exploration field is spent on data translation and transformation activities compared with the 10% used to perform the end goal of actually analysing the data. Another example of the quantity of information required is that the weight of the paper documentation for a weapons system is often at least as much as that of the weapon itself.

In the early eighties standards started to appear to enable the representation of the geometric shape of a product in an application-neutral form. In the nineties the standard informally known as STEP (**ST**andard for the **E**xchange of **P**roduct Model Data) was approved as ISO International Standard ISO 10303:1994. STEP extended prior CAD-related exchange standards to handle non-geometric aspects of products and also extended the view of the world from being limited to the design aspects of a product to a birth-to-death viewpoint. That is, STEP (which incidentally is still being extended and enhanced) is intended to provide a computer-sensible, application-neutral view of data related to a product from the time of the product's conception to the time of its disposal, where a product can be anything from a microchip to a battleship.

The STEP standards can be thought of in two parts. One, the technology, provides an object-flavored conceptual information modeling language called EXPRESS (ISO 10303-11:1994) [1,2], an exchange file format for data corresponding to an EXPRESS model (ISO 10303-21:1994), and an API in various language bindings to manipulate EXPRESS data instances (ISO 10303 parts 22+). The other aspect is a set of product-related information models formally defined using EXPRESS (ISO 10303 parts 40+, 100+, 200+). STEP is principally focussed on information exchange within and between manufacturers of products.

There are other aspects of products that are of more interest to the owners and users, for example repair and maintenance manuals which are typically provided in paper form. These often contain product illustrations that conceptually can be derived from the engineering drawings used by the manufacturers.

The acceptance of the Web by industry is profoundly affecting the way in which companies want to do business. Here is a way in which information can be made readily available electronically in a pull rather than the traditional push mode. Industry is using Web technology not only to exchange information internally but also to make it available to their outside suppliers and customers. Essentially Web technology can be viewed as an infrastructure enabling easy information exchange. Technologies like CORBA provide an infrastructure for piping information between applications. The major problem is in defining the information that can be exchanged using these kinds of infrastructures so that it is consistently understandable at both ends of the pipe.

Another requirement on industrial data is that some of it, for business, safety or regulatory purposes, must be archived for long periods of time — in some instances for decades. If archived electronically, the data must be both retrievable and understandable when retrieved, which may be long after the original application software and hardware has died and is no longer available to perform any interpretation. Thus there must be an implementation independent method of describing the meaning of data.

There are several languages related to information modeling and representation that are either standards, or are on the way to being standardized. For example, EXPRESS and KIF for information modeling, SGML and XML for representing logical views of a document, CORBA and IDL for communication between distributed data stores and applications, UML for OO applications design, and so on.

Recognizing that a language is designed for a particular purpose, it is still an industry requirement that only a limited number of languages be actually used, and perhaps more importantly, that those that are used can interoperate with each other. This is a major challenge in both the technical and human sense. Further, because of the trend towards employee empowerment, increasing numbers of non-professional computer users will find themselves having to use these languages in their daily work in order to document their information generating activities and in order to understand the information that they might receive from others. Somehow the languages must be made natural to them. There are many good theoretically based (e.g., Set theory, first order logic) languages whose underpinnings show through into the syntax and thereby are incomprehensible to the vast majority of potential end users.

There are moves afoot to couple some of the languages. As examples:

- I have seen a demonstration of a pre-production version of a bidirectional EXPRESS-UML (class/package diagrams) translator;
- I have seen several drafts of a potential British Standard (and then an ISO Standard) for mappings of EXPRESS and CDIF (ISO/IEC FCD 15475);
- I have heard a rumour, yet to be confirmed, that someone at NASA Ames has developed an EXPRESS/KIF translator (separately see [3]);
- At the February 1998 meeting of ISO TC184/SC4 (STEP) the liaison from ISO/IEC JTC1/ WG4 (SGML), Eliot Kimber, stated

  > "SGML needs to use EXPRESS to formally define the abstract data models for SGML, HyTime, and related standards. We feel that in doing so we will greatly increase the ability to integrate STEP-based data with SGML-based data. We are asking for the help of SC4 in doing this work so that the result is of the greatest benefit to all."

- EXPRESS-G, the graphical subset of EXPRESS, is being used as the basis for a CASE tool for Java (see *http://www.vtt.fi/cic/java/java-g*).

As EXPRESS has been an ISO standard for 4 years, it is undergoing a revision and extension. The extension includes facilities for the description of processes, and the result should be an integrated language for the modeling of both information and process in an implementation neutral manner. The current state of the work is available at *http://www.eurpc2.demon.co.uk/part11.htm*.

Standardization activities do not come high on any academic priority list (neither, to be honest, on many of those who work in industry). However, I see the standards bodies as the places where the coupling of modeling languages might best be carried out. In the past, most standards activities have focussed on codifying existing practice. Increasingly I see the activities moving towards pushing the state of the art. Don't be fooled; this kind of activity is both cross-disciplinary and intellectually challenging.

1. ISO 10303-11:1994, *Industrial automation systems and integration — Product data representation and exchange — Part 11: Description methods: The EXPRESS language reference manual*, 1994.
2. Douglas A. Schenck and Peter R. Wilson, *Information Modeling the EXPRESS Way*, Oxford University Press, 1994.
3. Peter R. Wilson, *On the Translation of KIF/Frame Ontologies to EXPRESS*. NIST Report NISTIR 5957, January 1997.

## 11.35 Clement Yu

*"Searching Information from Multiple Sources"*

Prof. Clement Yu
University of Illinois at Chicago
Department of Electrical Engineering and Computer Science
Chicago, IL 60607-7053
E-mail: yu@eecs.uic.edu
URL: http://www.eecs.uic.edu/eecspeople/yu.htm

Currently, enormous amount of information is generated by numerous people. As a consequence, information is distributed and stored in various sites. The Internet is one such example. In order to retrieve useful information in response to a query, a naive method is to broadcast the query to all sites. Then the search engine at each site processes the query and retrieves a set of documents. These documents are merged and then an appropriate ranked list of documents is presented to the user. The method described above does not make use of system resources efficiently, because the query may be sent to many sites which do not contain useful information. This is a waste of communication resources. In addition, the search engine at those sites will need to process the query. They may even return some useless documents. Finally, the transmission of useless documents and their subsequent merging waste further system resources. In order to reduce the waste, the contents of each database is represented by a representative. When a query is submitted by a user, it is compared to the representatives of the databases. Then the system will estimate the number of documents in each database which are most similar to the query. Based on these estimates, only those databases which contain sufficiently large number of most similar documents will be searched. The research issues we have been studying consist of:

- What information should be contained in a database representative?
- How an accurate estimate of the number of most similar documents in a database to a query can be obtained?
- How can such estimates be obtained efficiently?
- What is a minimal information that should be used in a database representative? What is the trade-off between the space required to store a database representative and the accuracy of the estimate described above?
- Suppose a standard for each site to supply the information needed to construct the representatives cannot be reached, how should the system respond to the user queries?
- Many sites are autonomous and employ different similarity functions to retrieve documents. Which documents should be retrieved by a site and be consistent to "the global similar function" desired by the user?
- While some sites may be willing to disclose their similarity functions to the "global system" in order to help the system to utilize system resources efficiently, other sites may

be unwilling to collaborate. How can the global system "discover" such information without explicit help from these local systems?

- The issues described above have been studied to some extent for text databases. Do the partial solutions for text databases apply to image databases as well? How about structured databases such as relational or object-oriented databases?

# 12 Participants' Contact Information


- Prof. Nabil R. Adam
  Rutgers University, Center for Information Management and Connectivity
  180 University Avenue
  Newark, NJ 07102
  Tel: (973) 353-5239 · Fax: (973) 353-5003
  E-mail: adam@adam.rutgers.edu
  URL: http://cimic.rutgers.edu
- Dr. Ethan Alpert
  National Center for Atmospheric Research, Scientific Computing Division
  1850 Table Mesa Drive
  Boulder, CO 80307-3000
  Tel: (303) 497-1832 · Fax: (303) 497-1804
  E-mail: ethan@ncar.ucar.edu
- Prof. William P. Birmingham
  The University of Michigan, Artificial Intelligence Laboratory
  Electrical Engineering and Computer Science Department
  1101 Beal Ave
  Ann Arbor, MI 48109-2110
  Tel: (734) 936-1590 · Fax: (734) 763-1260
  E-mail: wpb@eecs.umich.edu
  URL: http://ai.eecs.umich.edu/people/wpb/home.html
- Dr. Joseph Bordogna (*Invited Speaker*)
  Acting Deputy Director, National Science Foundation
  4201 Wilson Boulevard, Room 1205 N
  Arlington, VA 22230
  Tel: (703) 306-1001 · Fax: (703) 306-0109
  E-mail: jbordogn@nsf.gov
  URL: http://www.nsf.gov/od/lpa/forum/bordogna/jborbio.htm
- Prof. Francis Bretherton
  University of Wisconsin, Space Science and Engineering Center
  1225 West Dayton Street
  Madison, WI 53706
  Tel: (608) 262-7497 · Fax: (608) 262-5974
  E-mail: fbretherton@ssec.wisc.edu

- Dr. Lawrence Buja
  National Center for Atmospheric Research, Climate System Modeling Group/CGD
  PO Box 3000
  Boulder, CO 80307-3000
  Tel: (303) 497-1330 · Fax: (303) 497-1324
  E-mail: southern@ucar.edu
  URL: http://www.cgd.ucar.edu/csm/
- Prof. Peter Buneman (*Invited Speaker/Workgroup Chair*)
  University of Pennsylvania, Computer and Information Science
  200 South 33rd Street
  Philadelphia, PA 19119
  Tel: (215) 898 7703 · Fax: (215) 898-0587
  E-mail: peter@cis.upenn.edu
  URL: http://www.cis.upenn.edu/~peter/
- Dr. William J. Campbell (*Invited Speaker*)
  NASA, Applied Information Sciences Branch, Code 935
  Greenbelt, Maryland 20771
  Tel: (301) 286-8785 · Fax: (301) 286-1776
  E-mail: campbell@sauquoit.gsfc.nasa.gov
  URL: http://code935.gsfc.nasa.gov/
- Dr. John C. Cherniavsky (*Invited Speaker*)
  National Science Foundation, CISE/EIA
  4201 Wilson Blvd
  Arlington, VA 22202
  Tel: (703) 306-1980
  E-mail: jchernia@nsf.gov
  URL: http://www.cise.nsf.gov/eia/staff/eia_ddhome.html
- Dr. Kenneth W. Church
  AT&T Labs–Research
  PO Box 971
  180 Park Ave
  Florham Park, NJ 07932
  Tel: (973) 360-8621 · Fax: (973) 360-8187
  E-mail: kwc@research.att.com
- Dr. Geoff Crowley
  Southwest Research Institute, Instrumentation and Space Research Division
  6220 Culebra Road
  San Antonio, TX 78238-5166
  Tel: (210) 522-3475 · Fax: (210) 647-4325
  E-mail: crowley@picard.space.swri.edu
  URL: http://espsun.space.swri.edu/spacephysics/home.html

- Prof. Judith B. Cushing
  The Evergreen State College, Lab I
  Olympia, WA 98505
  Tel: 360-866-6000 x6652 · Fax: 360-866-6794
  E-mail: judyc@evergreen.edu
  URL: http://192.211.16.13/individuals/judyc/home.htm
- Dr. Ernest Daddio
  National Oceanic and Atmospheric Administration, NESDIS/EIS
  1315 East-West Highway, Suite 15548,
  Silver Spring, MD 20910
  Tel: (301) 713-1262 · Fax: (301) 713-1249
  E-mail: edaddio@esdim.noaa.gov
  URL: http://www.esdim.noaa.gov/NOAAServer/
- Dr. Donald Denbo
  University of Washington, NOAA/PMEL/OCRD
  7600 Sand Point Way NE
  Seattle, WA 98115
  Tel: (206) 526-4487 · Fax: (206) 526-6744
  E-mail: dwd@pmel.noaa.gov
  URL: http://www.pmel.noaa.gov/
- Dr. Pierre Elisseeff (*Workgroup Recorder*)
  Massachusetts Institute of Technology, Department of Ocean Engineering
  77 Massachusetts Avenue, Room 5-435
  Cambridge, MA 02139-4307
  Tel: (617) 253-1032 · Fax: (617) 253-2350
  E-mail: elisseef@mit.edu
  URL: http://dipole.mit.edu:8001/
- Prof. Zorana Ercegovac
  UCLA, InfoEN, Graduate School of Education and Information Studies
  300 Circle Dr North, UCLA, MC 951521
  Los Angeles, CA 90095-1521
  Tel: (310) 206-9361 · Fax: (310) 391-3923
  E-mail: zercegov@ucla.edu
  URL: http://www.gslis.ucla.edu/LIS/faculty/zercegov/ercegovac.html
- Prof. Gregory L. Fenves
  University of California, Berkeley, Department of Civil Engineering, #1710
  729 Davis Hall
  Berkeley, CA 94720-1710
  b Tel: (510) 643-8543 · Fax: (510) 643-8928
  E-mail: fenves@ce.berkeley.edu
  URL: http://www.ce.berkeley.edu/∼fenves/

- PROF. PAUL J. FORTIER (*Program Committee*)
University of Massachusetts Dartmouth, Electrical and Computer Engineering
285 Old Westport Road
North Dartmouth, Massachusetts 02747-2300
Tel: (508) 999-8544 · Fax: (508) 999-8489
E-mail: pfortier@umassd.edu
URL: http://www.ece.umassd.edu/ece/faculty/pfortier/pfortier.htm
- DR. CHRISTOPHER J. FREITAS
Southwest Research Institute, Computational Mechanics
6220 Culebra Road
San Antonio, Texas 78238-5166
Tel: (210) 522-2137 · Fax: (210) 522-3042
E-mail: cfreitas@swri.edu
URL: http://www.swri.org/
- PROF. JAMES C. FRENCH
University of Virginia, School of Engineering and Applied Science
Department of Computer Science
Thornton Hall
Charlottesville, VA 22903
Tel: (804) 982-2213 · Fax: (804) 982-2214
E-mail: french@cs.virginia.edu
URL: http://www.cs.virginia.edu/brochure/profs/french.html
- PROF. MARTIN HARDWICK
Rensselaer Polytechnic Institute
Laboratory for Industrial Information Infrastructure
110 8th Street - CII 7015
Troy, New York 12180-3590
Tel: (518) 276-6751 · Fax: (518) 276-2702
E-mail: hardwick@rdrc.rpi.edu
URL: http://www.rdrc.rpi.edu/staff/hardwick.html
- DR. GEORGE HAZELRIGG
National Science Foundation, DMII Room 550
Arlington, VA 22230
Tel: (703) 306-1330 or 5299 · Fax: (703) 306-0298
E-mail: ghazelri@nsf.gov
URL: http://www.eng.nsf.gov/staff/ghazelri.htm

- Prof. Yannis Ioannidis (*Workgroup Co-chair*)
  University of Athens and University of Wisconsin, Department of Informatics
  Panepistimioupolis, TYPA Buildings
  Athens 157-71, Greece
  Tel: +30 (1) 727-5224 · Fax: +30 (1) 727-5214
  E-mail: yannis@cs.wisc.edu, yannis@di.uoa.gr
  URL: http://www.cs.wisc.edu/~yannis/yannis.html
- Prof. Alan Kaplan
  Clemson University, Department of Computer Science
  Box 341906
  Clemson, SC 29534-1906
  Tel: (864) 656-7678 · Fax: (864) 656-1045
  E-mail: kaplan@cs.clemson.edu
  URL: http://www.cs.clemson.edu/~kaplan/
- Dr. George K. Lea
  National Science Foundation
  Division of Electrical and Communications Systems
  4201 Wilson Boulevard, Room 675 S
  Arlington, VA 22230
  Tel: (703) 603-1345 x5049 · Fax: (703) 306-0305
  E-mail: glea@nsf.gov
- Prof. Miron Livny (*Workgroup Chair*)
  University of Wisconsin-Madison, Computer Sciences
  1210 West Dayton Street
  Madison, WI 53706
  Tel: (608) 262-0856 · Fax: (608) 262-9777
  E-mail: miron@cs.wisc.edu
  URL: http://www.cs.wisc.edu/~miron/
- Dr. Christopher Miller
  National Oceanic and Atmospheric Administration, NESDIS/EIS, E/EI
  1315 East-West Highway
  Silver Spring, MD 20910
  Tel: (301) 713-1264 · Fax: (301) 713-1249
  E-mail: miller@esdim.noaa.gov
- Dr. Terry R. Moore
  University of Tennessee, Department of Computer Science
  104 Ayres
  Knoxville, TN 37996-1301
  Tel: (423) 974-5886 · Fax: (423) 974-8296
  E-mail: tmoore@cs.utk.edu
  URL: http://www.cs.utk.edu/~tmoore/

- Dr. Charles Myers
  National Science Foundation, OPP 755.81
  4201 Wilson Boulevard
  Arlington, VA 22230
  Tel: (703) 306-1029 · Fax: (703) 306-0648
  E-mail: cmyers@nsf.gov
  URL: http://www.nsf.gov/od/opp/
- Prof. Christos Nikolaou (*Program Committee/Workgroup Chair*)
  University of Crete and ICS-FORTH, Office of the Rector
  Leoforos Knossou
  Heraklion, Crete 71409, Greece
  Tel: +30-81-393199 · Fax: +30-81-210106
  E-mail: nikolau@ics.forth.gr
  URL: http://www.ics.forth.gr/∼nikolau/nikolau.html
- Prof. Nicholas M. Patrikalakis (*Workshop Chair*)
  Massachusetts Institute of Technology, Department of Ocean Engineering
  77 Massachusetts Avenue
  Cambridge, MA 02139-4307
  Tel: (617) 253-4555 · Fax: (617) 253-8125
  E-mail: nmp@mit.edu
  URL: http://oe.mit.edu/people/patrikal.htm
- Dr. Lawrence Raymond
  Oracle Corporation, Spatial Technologies Practice
  5102 Windy Lake Drive
  Kingwood, TX 77345
  Tel: (281) 361.4334 · Fax: (281) 361.4428
  E-mail: lraymond@us.oracle.com
- Prof. Allan R. Robinson (*Program Committee/Invited Speaker*)
  Harvard University, Division of Engineering and Applied Sciences
  29 Oxford Street
  Cambridge, MA 02138
  Tel: (617) 495-2819 · Fax: (617) 495-5192
  E-mail: robinson@pacific.harvard.edu
  URL: http://www.deas.harvard.edu/people/Faculty/ARobinson.html
- Prof. Jarek Rossignac (*Program Committee/Invited Speaker/Workgroup Chair*)
  Georgia Institute of Technology, GVU Center/College of Computing, 241
  801 Atlantic Drive NW
  Atlanta, Georgia 30332-0280
  Tel: (404) 894-0671 · Fax: (404) 894-0673
  E-mail: jarek@cc.gatech.edu
  URL: http://www.gvu.gatech.edu/people/official/jarek.rossignac/

- Prof. M. Saiid Saiidi
  University of Nevada, Reno, Department of Civil Engineering (258)
  Reno, NV 89557
  Tel: (702) 784-4839 · Fax: (702) 784-1390
  E-mail: saiidi@unr.edu
- Dr. Jakka Sairamesh (*Workgroup Recorder*)
  IBM Watson Research
  30 Sawmill River Road
  Hawthorne, NY 10532
  Tel: (914) 784-7372
  E-mail: jramesh@watson.ibm.com
  URL: http://www.cs.columbia.edu/~jakka/home.html
- Prof. Clifford A. Shaffer
  Virginia Polytechnic Institute and State University, Department of Computer Science
  Blacksburg, VA 24061
  Tel: (540) 231-4354 · Fax: (540) 231-6075
  E-mail: shaffer@cs.vt.edu
  URL: http://www.cs.vt.edu/vitae/Shaffer.html
- Prof. Amit P. Sheth (*Invited Speaker*)
  University of Georgia, LSDIS Lab, Department of Computer Science
  415 Boyd GSRC
  Athens, GA 30602-7404
  Tel: (706) 542-2310 · Fax: (706) 542-2966
  E-mail: amit@cs.uga.edu
  URL: http://lsdis.cs.uga.edu/~amit/
- Prof. Terence Smith
  University of California at Santa Barbara, Department of Computer Science
  Santa Barbara, CA 93106
  Tel: (805) 893-2966 · Fax: (805) 893-8553
  E-mail: smithtr@cs.ucsb.edu
  URL: http://www.cs.ucsb.edu/
- Dr. Ram D. Sriram
  National Institute of Standards and Technology, Building 220/A127
  Gaithersburg, MD 20899
  Tel: (301) 975-3507 · Fax: (301) 258-2749
  E-mail: sriram@cme.nist.gov
  URL: http://www.mel.nist.gov/msidstaff/sriram.ram.htm

- Dr. George Strawn (*Invited Speaker*)
  National Science Foundation, Director of the Networking Division/CISE
  4201 Wilson Boulevard, Room 1175
  Arlington, VA 2230
  Tel: (703) 306-1950 · Fax: (703) 306-0621
  E-mail: gstrawn@nsf.gov
  URL: http://www.cise.nsf.gov/anir/Georgehome.html
- Dr. Knut Streitlien (*Workgroup Recorder*)
  Massachusetts Institute of Technology
  Sea Grant Autonomous Underwater Vehicles Laboratory
  Building E38-372, 292 Main Street
  Cambridge, MA 02139
  Tel: (617) 258-7192 · Fax: (617) 258-5730
  E-mail: knut@mit.edu
  URL: http://seagrant.mit.edu/~knut/
- Dr. Jean-Baptiste Thibaud
  National Science Foundation, Office of Polar Programs
  4201 Wilson Boulevard
  Arlington, VA 22230
  Tel: (703) 306-1045 x7342 · Fax: (703) 306-0648
  E-mail: jthibaud@nsf.gov
  URL: http://www.nsf.gov/od/opp/
- Prof. George Turkiyyah
  University of Washington, Department of Civil Engineering
  Box 352700
  Seattle, WA 98195
  Tel: (206) 543-8741 · Fax: (206) 543-1543
  E-mail: george@ce.washington.edu
  URL: http://www.ce.washington.edu/~george/
- Prof. Alvaro Vinacua (*Workgroup Recorder*)
  Polytechnic University of Catalonia, I.R.I., Edifici Nexus 202
  C. Gran Capità 2-4
  E08034 Barcelona, Spain
  Tel: +34-3-401-5786 · Fax: +34-3-401-5750
  E-mail: alvar@turing.upc.es
  URL: http://www.lsi.upc.es/~alvar/

- Dr. Stuart Weibel (*Invited Speaker*)
  Online Computer Library Center, Office of Research, MC 710
  6565 Frantz Road
  Dublin, OH 43017
  Tel: (614) 764-6081 · Fax: (614) 764-2344
  E-mail: weibel@oclc.org
  URL: http://www.oclc.org/oclc/menu/home1.htm
- Prof. Jack C. Wileden
  University of Massachusetts, Computer Science Department, LGRC
  North Pleasant Street
  Amherst, MA 01003
  Tel: (413) 545-0289 · Fax: (413) 545-1249
  E-mail: wileden@cs.umass.edu
  URL: http://www-ccsl.cs.umass.edu/~jack/
- Prof. John R. Williams (*Program Committee*)
  Massachusetts Institute of Technology
  Department of Civil and Environmental Engineering
  77 Massachusetts Avenue, Rm. 1-250
  Cambridge, MA 02139
  Tel: (617) 253-7201 · Fax: (617) 253-6324
  E-mail: jrw@mit.edu
  URL: http://rx7.mit.edu/john/
- Dr. Roy D. Williams (*Invited Speaker*)
  California Institute of Technology, 158-79
  Pasadena, CA 91125
  Tel: (626) 395-3670 · Fax: (626) 584-5917
  E-mail: roy@cacr.caltech.edu
  URL: http://www.cacr.caltech.edu/~roy/
- Dr. Peter R. Wilson (*Invited Speaker*)
  Boeing Commercial Airplane Group
  P.O. Box 3707, MS 6H-AF
  Seattle, WA 98124-2207
  Tel: (425) 237-3506 · Fax: (425) 237-3428
  E-mail: peter.r.wilson@boeing.com
- Prof. Clement Yu
  University of Illinois at Chicago
  Department of Electrical Engineering and Computer Science
  851 South Morgan
  Chicago, Illinois 60607-7053
  Tel: (312) 996-2318 · Fax: (312) 413-0024
  E-mail: yu@eecs.uic.edu
  URL: http://www.eecs.uic.edu/eecspeople/yu.htm

- Dr. Maria Zemankova
  National Science Foundation, Information and Data Management/IIS/CISE
  4201 Wilson Boulevard, #1115
  Arlington, VA 22230
  Tel: (703) 306-1926 · Fax: (703) 306-0599
  E-mail: mzemanko@nsf.gov
  URL: http://www.cise.nsf.gov/iis/idm_pdhome.html

# 13 Background Material

1. *The Digital Earth: Understanding Our Planet in the 21st Century*, a speech by Vice-President Al Gore, California Science Center, Los Angeles, California, January 31, 1998 (http://www.opengis.org/info/pubaffairs/ALGORE.htm).
2. M. L. Dertouzos, *What Will Be: How the New World of Information Will Change Our Lives* (San Francisco: Harper, 1997).
3. *Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata (CSDGM)*, Version 2 (Draft), April 7, 1997 (http://www.fgdc.gov/Standards/Documents/Standards/Metadata/csdgmv2-0.{pdf,wpd}).
4. *NSF Workshop on Interfaces to Scientific Data Archives*, Pasadena, California, March 25-27, 1998 (http://www.cacr.caltech.edu/isda/).
5. *NSF Workshop on Scientific Databases*, Charlottesville, VA, March 1990. Technical Report CS-90-21, Department of Computer Science, University of Virginia, August 1990 (ftp://ftp.cs.virgina.edu/pub/techreports/CS-90-21.{ps,txt}.Z, ftp://ftp.cs.virgina.edu/pub/techreports/CS-90-22.{ps,txt}.Z).
6. *NSF Workshop on Workflow and Process Automation in Information Systems*, Athens, Georgia, May 8-10, 1996 (http://lsdis.cs.uga.edu/activities/NSF-workflow/).
7. President's Committe of Advisors for Science and Technology (PCAST), *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*, PCAST Panel on Biodiversity and Ecosystems, March 1998 (http://www.whitehouse.gov/WH/EOP/OSTP/Environment/html/teamingcover.html).
8. *Second European Conference on Research and Advanced Technology for Digital Libraries*, September 21-23, 1998, Heraklion, Crete, Greece (http://www.csi.forth.gr/2EuroDL/).
9. *Second IEEE Metadata Conference*, Silver Spring, Maryland, September 16-17, 1997 (http://www.llnl.gov/liv_comp/metadata/md_97.html).