Distributed Information and Computation in Scientific and Engineering Environments

N. M. Patrikalakis[1,*] P. J. Fortier[2] Y. Ioannidis[3] C. N. Nikolaou[4]
A. R. Robinson[5] J. R. Rossignac[6] A. Vinacua[7] S. L. Abrams[1]

[1] MIT, Department of Ocean Engineering
[2] University of Massachusetts Dartmouth, Department of Electrical and Computer Engineering
[3] University of Wisconsin and University of Athens, Department of Computer Science, Greece
[4] University of Crete, Department of Computer Science, and ICS/FORTH, Greece
[5] Harvard University, Department of Earth and Planetary Sciences
[6] Georgia Institute of Technology, Graphics, Visualization, and Usability Center
[7] Polytechnic University of Catalonia, Department of Informatics, Spain

[*] Coordinating author: MIT, Department of Ocean Engineering
77 Massachusetts Avenue, Room 5-428
Cambridge, MA 02139-4307, USA
Tel: (617) 253-4555, Fax: (617) 253-8125, E-mail: nmp@mit.edu

MIT Design Laboratory Memorandum 98-7
Issued: December 3, 1998
Revised: January 14, 1999

Abstract

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

Three broad problem areas inhibit such activities: (1) Computational, e.g, insufficient infrastructure for the sharing of very large amounts of information, results, and tools; (2) Structural institutional barriers, e.g., funding, publication, and promotion policies; and (3) Social, e.g., communication barriers stemming from narrow specialization. The participants supported specific steps to address these problems: explicit support and incentives for multidisciplinary activities; the development of digitial libraries to enhance interdisciplinary communication and understanding; and development of a "virtual scientific marketplace" to disseminate tools, results, and expertise;

Keywords: distributed repositories, heterogeneous computing environment, multidisciplinary collaboration, resource discovery

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.

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" [5]) for experts, tools, and procedures, which will facilitate multidisciplinary collaboration.

More details on the workshop, including background material, position papers, and the final report are available at the workshop web site <http://deslab.mit.edu/DesignLab/dicpm/>.

2 Major Themes Discussed

2.1 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.

2.1.1 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 Ocean Observing and Prediction Systems (OOPS) will substantially accelerate progress and broaden the scope of such research. The OOPS concept involves a mix of advanced interdisciplinary sensor, measurement, modeling, and estimation methodologies, as well as major systems engineering and computer science developments [13]. The development of OOPS utilizing distributed oceanographic resources will significantly increase the 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 for scientific, operational, or management activities 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 to include. 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 then 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 historical synoptic and climatological data and acted upon by present forcings. Initial field estimates and sampling strategies are made. During the real–time operation, newly acquired data is assimilated into the models and 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) [4], 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 distributed, multidisciplinary data; preparation and fine tuning of the OOPS via extensive simulated experiments; ingression/assimilation/integration/ egression of data during operations; automatic work flow management; and widespread 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.

2.1.2 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 involves 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 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.

2.2 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 need of 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.

2.3 Research Themes

2.3.1 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 [6, 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 [3]. 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, <http://www.omg.org> [15] 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 <http://javasoft.com/> 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.

2.3.2 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 <http://purl.org/metadata/dublin_core/> and the Warwick Framework <http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrtl.cornell%2fTR96-1593> are a first step at establishing a metadata classification. They define a minimal set of optional fields that can describe an information object. XML <http://www.w3.org/TR/REC-xml> is becoming increasingly adopted as a common syntax for expressing structure in data. Moreover, the Resource Description Framework (RDF, <http://www.w3.org/RDF/Overview.html>), 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, <http://fgdc.er.usgs.gov/linkpub.html>), the biology community has defined the NBII (National Biological Information Infrastructure) Biological Metadata Standard <http://www.nbii.gov/standards/metadata.html>, 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.

2.3.3 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.

2.3.4 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.

2.3.5 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.

2.3.6 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.

2.3.7 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 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.

2.3.8 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, video-conferences, 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 [8, 11] 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 [9, 14] 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 [10].

Parallel batch processing environments such as PVM and Condor [7, 12] 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.

2.3.9 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.)

2.4 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.

3 Conclusions and Recommendations

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

Computational. Reliable and supported tools do not 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 led to the following recommendations:

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 also cooperate to establish a global distributed information registry and repository (or "virtual scientific marketplace" [5]) 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 2.3, 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 (Department of Defense, Department of Energy, Department of Commerce, National Science Foundation, etc.) to motivate standards (e.g., metadata standardization) and multidisciplinary cooperation and collaboration.

Acknowledgments

The DICPM Workshop was generously supported by the U.S. National Science Foundation under grant IIS-9812601. The authors wish to thank the following NSF Program Managers: Dr. Maria Zemankova (Information and Data Management), Dr. Howard Moraff (Robotics and Human Augmentation), Dr. Steven N. Goldstein (International Coordination), Dr. Peter Scheuermann (Operating Systems and Compilers), Dr. William Agresti (Experimental Software Systems), Dr. Ken Chong (Control, Mechanics and Materials), Dr. Art Sanderson (Integrative Systems), Dr. Jay Fein (Climate Dynamics), Dr. Henry N. Blount (Office of Multidisciplinary Activities) and Dr. Charles Myers (Arctic Research and Policy). Dr. Zemankova provided many helpful comments on this report.

Patrikalakis, Nikolaou, and Abrams have also received support for related work from NOAA and NATO, under grants NA86RG0074 and CRG971523; Nikolaou from the European Union Research on Telematics Programme under project number F0069; Robinson from ONR under grant N00014-97-1-1018; and Vinacua from a grant of the Ministry of Education of Spain.

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, NOAA, NATO, or ONR.

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

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

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

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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