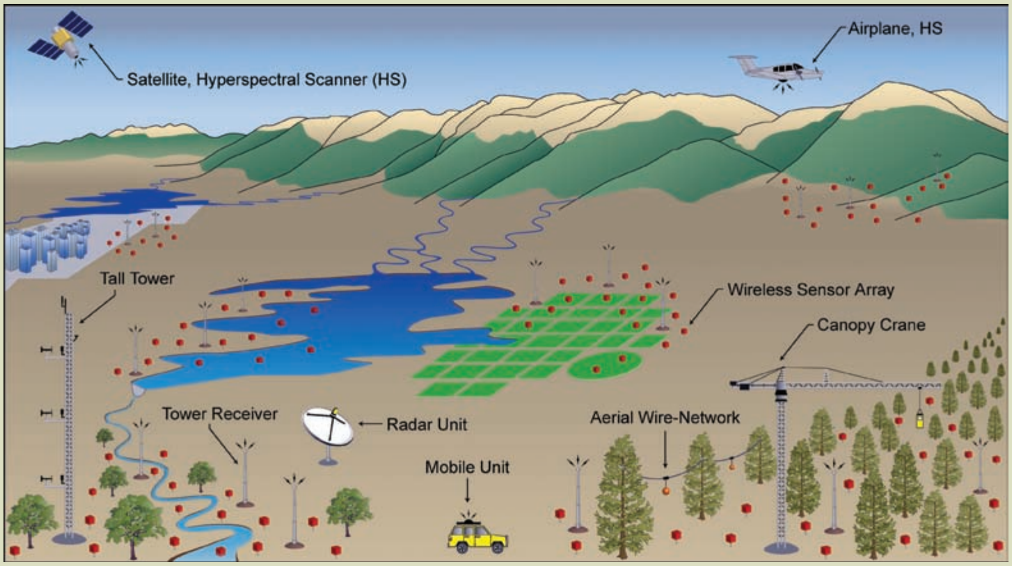
CHAPTER 3
DATA, DATA ANALYSIS, AND VISUALIZATION (2006-2010)

I. A Wealth of Scientific Opportunities Afforded by Digital Data

Science and engineering research and education have become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared and analyzed. Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavor.

New scientific opportunities are emerging from increasingly effective data organization, access and usage. Together with the growing availability and capability of tools to mine, analyze and visualize data, the emerging data cyberinfrastructure is revealing new knowledge and fundamental insights. For example, analyses of DNA sequence data are providing remarkable insights into the origins of man, revolutionizing our understanding of the major kingdoms of life, and revealing stunning and previously unknown complexity in microbial communities. Sky surveys are changing our understanding of the earliest conditions of the universe and providing comprehensive views of phenomena ranging from black holes to supernovae. Researchers are monitoring socioeconomic dynamics over space and time to advance our



An artist's conception (above) depicts fundamental NEON observatory instrumentation and systems as well as potential spatial organization of the environmental measurements made by these instruments and systems.

The image on the opposite page shows the action of the enzyme cellulase on cellulose using the CHARMM community code in a simulation carried out at SDSC. NREL will use the simulation to help develop strategies for efficient large-scale conversion of biomass into ethanol.

understanding of individual and group behavior and its relationship to social, economic and political structures. Using combinatorial methods, scientists and engineers are generating libraries of new materials and compounds for health and engineering, and environmental scientists and engineers are acquiring and analyzing streaming data from massive sensor networks to understand the dynamics of complex ecosystems.

In this dynamic research and education environment, science and engineering data are constantly being collected, created, deposited, accessed, analyzed and expanded in the pursuit of new knowledge. In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form, and to transform these data into information and knowledge aided by sophisticated data mining, integration, analysis and visualization tools.

This chapter sets forth a framework in which NSF will work with its partners in science and engineering – public and private sector organizations both foreign and domestic representing data producers, scientists, engineers, managers and users alike – to address data acquisition, access, usage, stewardship and management challenges in a comprehensive way.

II. Definitions

A. Data, Metadata and Ontologies

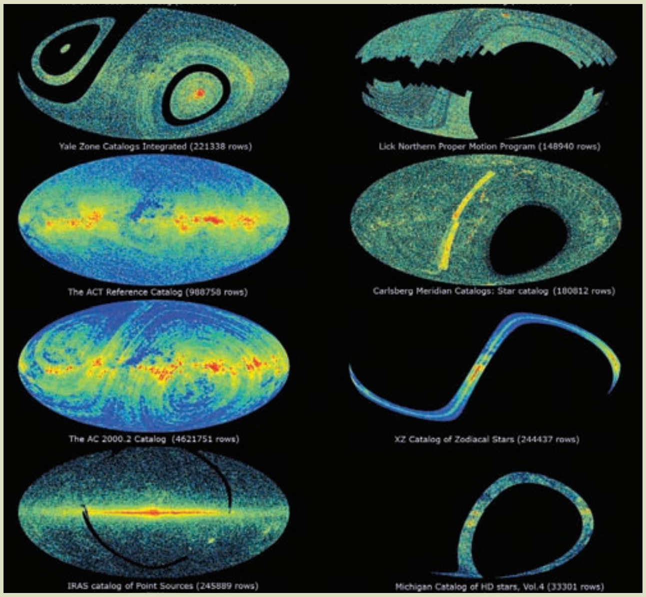
In this document, "data" and "digital data" are used interchangeably to refer to data and information stored in digital form and accessed electronically.

- ***Data.*** For the purposes of this document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.
- ***Metadata.*** Metadata are a subset of data, and are data about data. Metadata summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.
- ***Ontology.*** An ontology is the systematic description of a given phenomenon. It often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse.

B. Data Collections

This document adopts the definition of data collection types provided in the NSB report on Long-Lived Digital Data Collections, where data collections are characterized as being one of three functional types:

- ***Research Collections.*** Authors are individual investigators and investigator teams. Research collections are usually maintained to serve immediate group participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards.
- ***Resource Collections.*** Resource collections are authored by a community of investigators, often within a domain of science or engineering, and are often developed with community-level standards. Budgets are often intermediate in size. Lifetime is between the mid- and long-term.


Yale Zone Catalogs Integrated (221338 rows)
Lick Northern Proper Motion Program (148940 rows)
The ACT Reference Catalog (988758 rows)
Carlsberg Meridian Catalogs: Star catalog (180812 rows)
The AC 2000.2 Catalog (4621751 rows)
XZ Catalog of Zodiacal Stars (244437 rows)
IRAS catalog of Point Sources (245889 rows)
Michigan Catalog of HD stars, Vol.4 (33301 rows)

The National Virtual Observatory's Sky Statistics Survey allows astronomers to get a fast inventory of astronomical objects from various catalogs.

- ***Reference Collections.*** Reference collections are authored by and serve large segments of the science and engineering community and conform to robust, well-established and comprehensive standards, which often lead to a universal standard. Budgets are large and are often derived from diverse sources with a view to indefinite support.

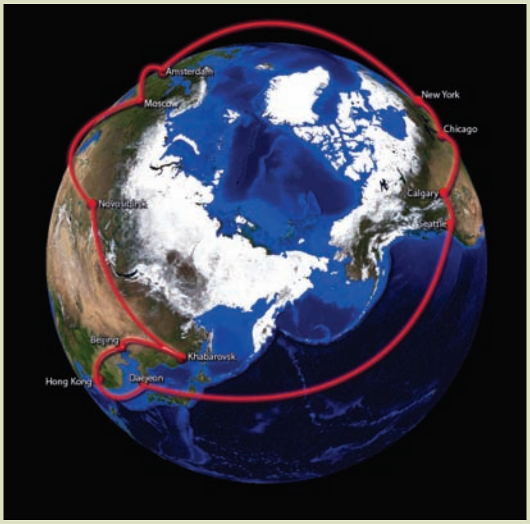
Boundaries between the types are not rigid, and collections originally established as research collections may evolve over time to become resource and/or reference collections. In this document, the term data collection is construed to include one or more databases and their relevant technological implementation. Data collections are managed by organizations and individuals with the necessary expertise to structure them and to support their effective use.

III. Developing a Coherent Data Cyberinfrastructure in a Complex Global Context

Since data and data collections are owned or managed by a wide range of communities, organizations and individuals around the world, NSF must work in an evolving environment constantly being shaped by developing international and national policies and treaties, community-specific policies and approaches, institutional-level programs and initiatives, individual practices, and continually advancing technological capabilities.

At the international level, a number of nations and international organizations have already recognized the broad societal, economic, and scientific benefits that result from open access to science and engineering digital data. In 2004, more than 30 nations, including the United States, declared their joint commitment to work toward the establishment of common access regimes for digital research data generated through public funding. Since the international exchange of scientific data, information and knowledge promises to significantly increase the scope and scale of research and its corresponding impact, these nations are working together to define the implementation steps necessary to enable the global science and engineering system.

The U.S. community is engaged through the Committee on Data for Science and Technology (CODATA). The U.S. National Committee for CODATA (USNC/CODATA) is working with international CODATA partners, including the International Council for Science (ICSU), the International Council for Scientific and Technical Information (ICSTI), the World Data Centers (WDCs) and others, to accelerate the development of a global open-access scientific data and information resource, through the construction of an online "open access knowledge environment," as well as through targeted projects. The Global Information Commons for Science is a multi-stakeholder initiative arising out of the second phase of the World Summit on the Information Society that can provide important opportunities for international coordination and cooperation. The goals of this initiative include improving understanding of the benefits of access to scientific data and information, promoting successful institutional and legal models for providing sustainable access, and enhancing coordination among the many science and engineering stakeholders around the world.



The GLORIAD network, an optical network ring around the northern hemisphere, promotes new opportunities for cooperation and understanding for scientists, educators and students.

A number of international science and engineering communities have also been developing data management and curation approaches for reference and resource collections. For example, the international Consultative Committee for Space Data Standards (CCSDS) defined an archive reference model and service categories for the intermediate and long-term storage of digital data relevant to space missions. This effort produced the Open Archival Information System (OAIS), now adopted as the "de facto" standard for building digital archives, and provided evidence that a community-focused activity can have much broader impact than originally intended. In another example, the Inter-University Consortium for Political and Social Research (ICPSR) - a membership-based organization with over 500 member colleges and universities around the world - maintains and provides access to a vast archive of social science data. ICPSR serves as a content management organization, preserving relevant social science data and migrating them to new storage media as technology changes, and also provides user support services. ICPSR recently announced plans to establish an international standard for social science documentation. Similar activities in other communities are also underway. Clearly, NSF must maintain a presence in, support, and add value to these ongoing international discussions and activities.

Activities on an international scale are complemented by activities within nation states. In the United States, a number of organizations and communities of practice are exploring mechanisms to establish common approaches to digital data access, management and curation. For example, the Research Library Group (RLG – a not-for-profit membership organization representing libraries, archives and museums) and the U.S. National Archives and Records Administration (NARA – a sister agency whose mission is to provide direction and assistance to federal agencies on records management) are producing certification requirements for establishing and selecting reliable digital information repositories. RLG and NARA intend their results to be standardized via the International Organization of Standardization (ISO) Archiving Series, and may impact all data collections types. The National Institutes of Health (NIH) National Center for Biotechnology Information plays an important role in the management of genome data at the national level, supporting public databases, developing software tools for analyzing data, and disseminating biomedical information.

At the institutional level, colleges and universities are developing approaches to digital data archiving, curation and analysis. They are sharing best practices to develop digital libraries that collect, preserve, index and share research and education material produced by faculty and other individuals within their organizations. The technological implementations of these systems are often open-source and support interoperability among their adopters. University-based research libraries and research librarians are positioned to make significant contributions in this area, where standard mechanisms for access and maintenance of scientific digital data may be derived from existing library standards developed for print material. These efforts are particularly important to NSF as the agency considers the implications of not only making all data generated with NSF funding broadly accessible, but of also promoting the responsible organization and management of these data so that they are widely usable.

IV. The Next Five Years: Towards a National Digital Data Framework

Motivated by a vision in which science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved, NSF's five-year goal is twofold:

- To catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable; and
- To support development of a new generation of tools and services facilitating data mining,

Images produced by Montage on SDSC TeraGrid from the 2MASS all-sky survey, provide astronomers with new insights into the large-scale structure of the Milky Way.

integration, analysis, and visualization essential to turning data into new knowledge and understanding.

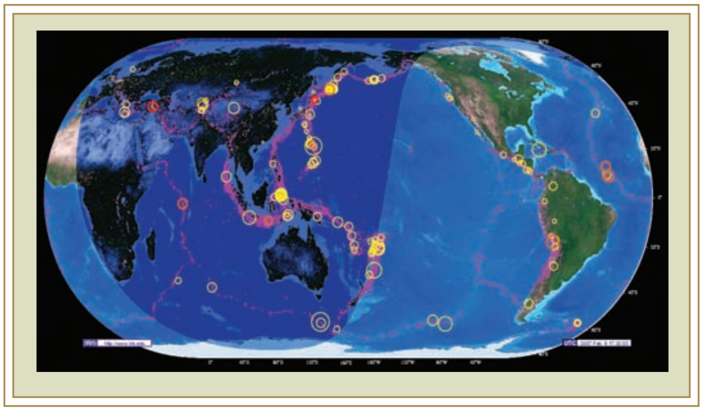
The resulting national digital data framework will be an integral component in the national cyberinfrastructure framework described in this document. It will consist of a range of data collections and managing organizations, networked together in a flexible technical architecture using standard, open protocols and interfaces, and designed to contribute to the emerging global information commons. It will be simultaneously local, regional, national and global in nature, and will evolve as science and engineering research and education needs change and as new science and engineering opportunities arise. Widely accessible tools and services will permit scientists and engineers to access and manipulate these data to advance the science and engineering frontier.

In print form, the preservation process is handled through a system of libraries and other repositories throughout the country and around the globe. Two features of this print-based system make it robust. First, the diversity of business models deriving support from a variety of sources means that no single entity bears sole responsibility for preservation, and the system is resilient to changes in any particular sector. Second, there is overlap in the collections, and redundancy of content reduces the potential for catastrophic loss of information.

The national data framework is envisioned to provide an equally robust and diverse system for digital data management and access. It will promote interoperability between data collections supported and managed by a range of organizations and organization types; provide for appropriate protection and reliable long-term preservation of digital data; deliver computational performance, data reliability and movement through shared tools, technologies and services; and accommodate individual community preferences. NSF will also develop a suite of coherent data policies that emphasize open access and effective organization and management of digital data, while respecting the data needs and requirements within science and engineering domains.

The following principles will guide the agency's FY 2006 through FY 2010 investments:

- Science and engineering research and education opportunities and priorities will motivate NSF investments in data cyberinfrastructure.
- Science and engineering data generated with NSF funding will be readily accessible and easily usable, and will be appropriately, responsibly and reliably preserved.
- Broad community engagement is essential to the prioritization and evaluation of the utility of scientific data collections, including the possible evolution from research to resource and reference collection types.
- Continual exploitation of data in the creation of new knowledge requires that investigators have access to the tools and services necessary to locate and access relevant data, and understand its structure sufficiently to be able to interpret and (re)analyze what they find.
- The establishment of strong, reciprocal, international, interagency and public-private partnerships is essential to ensure all stakeholders are engaged in the stewardship of valuable data assets. Transition plans, addressing issues such as media, stewardship and standards, will be developed for valuable data assets, to protect data and assure minimal disruption to the community during transition periods.
- Mechanisms will be created to share data stewardship best practices between nations, communities, organizations and individuals.
- In light of legal, ethical and national security concerns associated with certain types of data, mechanisms essential to the development of both statistical and technical ways to protect privacy and confidentiality will be supported.



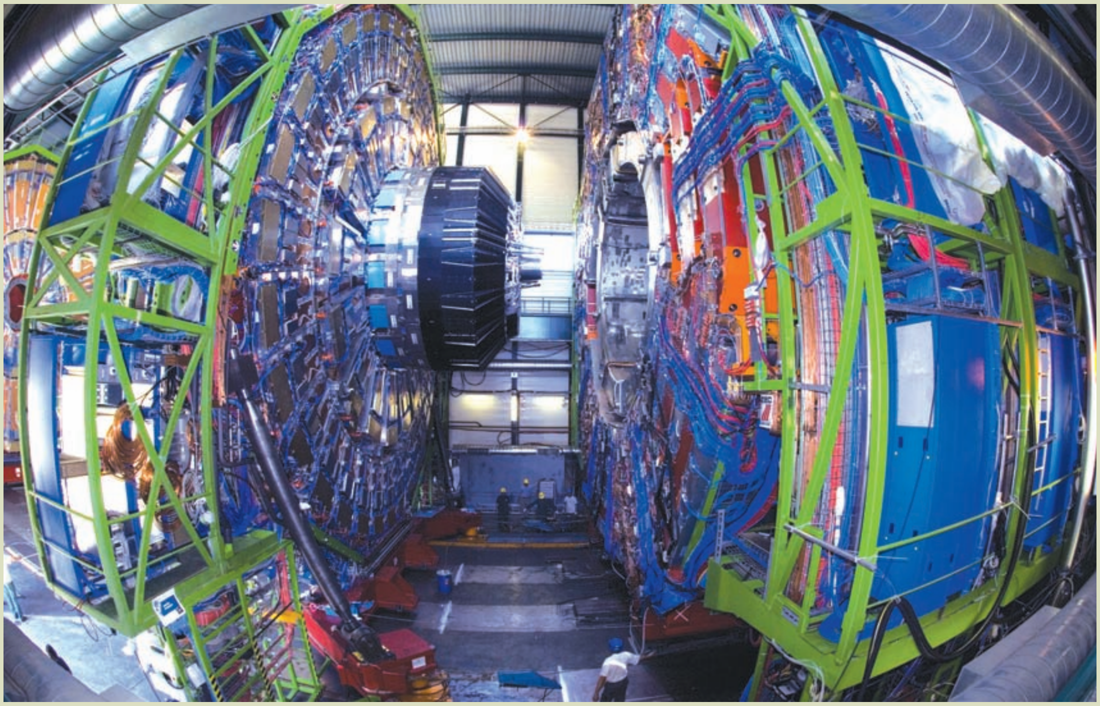
The IRIS Seismic Monitor System allows scientists and others to monitor global earthquakes in near real-time, visit seismic stations world-wide, and search the web for earthquake information.

A. A Coherent Organizational Framework - Data Collections and Managing Organizations

To date, challenges associated with effective stewardship and preservation of scientific data have been more tractable when addressed through communities of practice that may derive support from a range of sources. For example, NSF supports the Incorporated Research Institutions for Seismology (IRIS) consortium to manage seismology data. Jointly with NIH and DOE, the agency supports the Protein Data Bank to manage data on the three-dimensional structures of proteins and nucleic acids. Multiple agencies support the University Consortium for Atmospheric Research, an organization that has provided access to atmospheric and oceanographic data sets, simulations, and outcomes extending back to the 1930s through the National Center for Atmospheric Research.

Existing collections and managing organization models reflect differences in culture and practice within the science and engineering community. As community proxies, data collections and their managing organizations can provide a focus for the development and dissemination of appropriate standards for data and metadata content and format, guided by an appropriate community-defined governance approach. This is not a static process, as new disciplinary fields and approaches, data types, organizational models and information strategies inexorably emerge. This is discussed in detail in the Long-Lived Digital Data Collections report of the National Science Board.

Since data are held by many federal agencies, commercial and non-profit organizations, and international entities, NSF will foster the establishment of interagency, public-private and international consortia charged with providing stewardship for digital data collections to promote interoperability across data collections. The agency will work with the broad community of science and engineering producers, managers, scientists and users to develop a common conceptual framework. A full range of mechanisms will be used to identify and build upon common ground across domain communities and managing organizations, engaging all stakeholders. Activities will include: the support of new projects; development and implementation of evaluation and assessment criteria that, among other things, reveal lessons learned across communities; support of



Researchers check functionality and performance of the Compact Muon Solenoid detector at CERN before its closure. Built on the Large Hadron Collider, it provides a magnetic field of 4T.

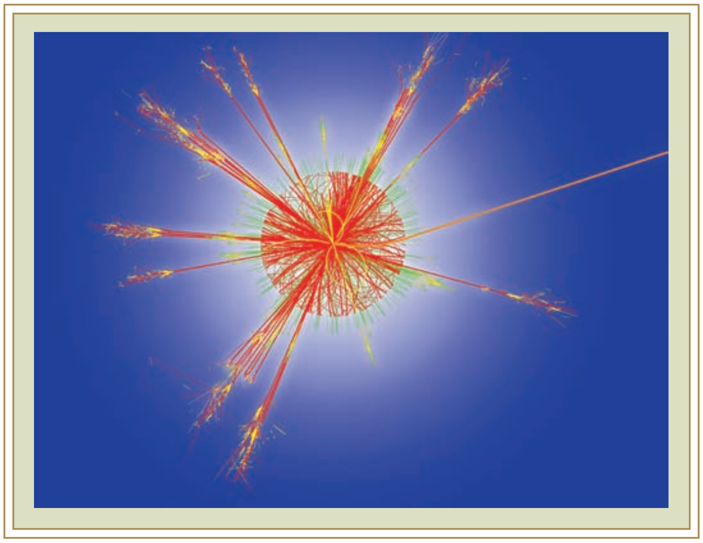
community and intercommunity workshops; and the development of strong partnerships with other stakeholder organizations. Stakeholders in these activities include data authors, data managers, data scientists and engineers, and data users representing a diverse range of communities and organizations, including universities and research libraries, government agencies, content management organizations and data centers, and industry.

To identify and promote lessons learned across managing organizations, NSF will continue to promote the coalescence of appropriate collections with overlapping interests, approaches and services. This reduces data-driven fragmentation of science and engineering domains. Progress is already being made in some areas. For example, NSF has been working with the environmental science and engineering community to promote collaboration across disciplines ranging from ecology and hydrology to environmental engineering. This has resulted in the emergence of common cyberinfrastructure elements and new interdisciplinary science and engineering opportunities.

B. Developing A Flexible Technological Architecture

From a technological perspective, the national data framework must provide for reliable preservation, access, analysis, interoperability, and data movement, possibly using a web or grid services distributed environment. The architecture must use standard open protocols and interfaces to enable the broadest use by multiple communities. It must facilitate user access, analysis and visualization of data, addressing issues such as authentication, authorization and other security concerns, and data acquisition, mining, integration, analysis and visualization. It must also support complex workflows enabling data discovery. Such an architecture can be visualized as a number of layers providing different capabilities to the user, including data management, analysis, collaboration tools, and community portals. The connections among these layers must be transparent to the end user, and services must be available as modular units responsive to individual or community needs. The system is likely to be implemented as a series of distributed applications and operations supported by a number of organizations and institutions distributed throughout the country. It must provide for the replication of data resources to reduce the potential for catastrophic loss of digital information through repeated cycles of systems migration and all other causes since, unlike printed records, the media on which digital data are stored and the structures of the data are relatively fragile.

High quality metadata, which summarize data content, context, structure, interrelationships, and provenance (information on history and origins), are critical to successful information management, annotation, integration and analysis. Metadata take on an increasingly important role when addressing issues associated with the combination of data from experiments, observations and simulations. In these cases, product data sets require metadata that describe, for example, relevant collection techniques, simulation codes or pointers to archived copies of simulation codes, and codes used to process, aggregate or transform data. These metadata are essential to create new knowledge and to meet the reproducibility imperative of modern science. Metadata are often associated with data via markup languages, representing a consensus around a controlled vocabulary to describe phenomena of interest to the community, and allowing detailed annotations of data to be embedded within a data set. Because there is often little awareness of markup language development activities within science and engineering communities, effort is expended reinventing what could be adopted or adapted from elsewhere. Scientists and engineers therefore need access to tools and services that help ensure that metadata are automatically captured or created in real-time.



A simulated event of the collision of two protons in the ATLAS experiment. The colors of the tracks emanating from the center show the different types of particles emerging from the collision.

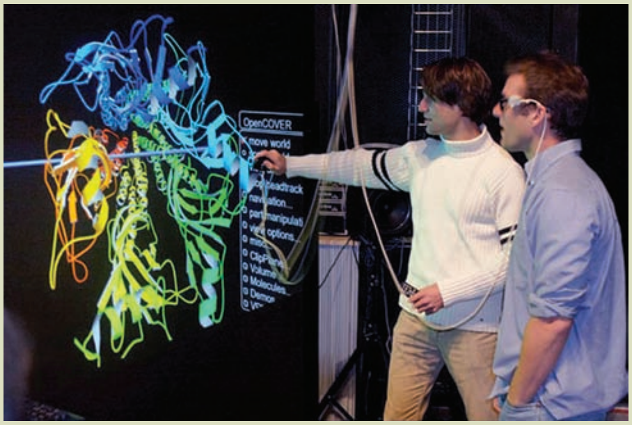
Effective data analysis tools apply computational techniques to extract new knowledge through a better understanding of the data and its redundancies and relationships by filtering extraneous information and by revealing previously unseen patterns. For example, the Large Hadron Collider at CERN generates such massive data sets that the detection of both expected events, such as the Higgs boson, and unexpected phenomena require the development of new algorithms, both to manage data and to analyze it. Algorithms and their implementations must be developed for statistical sampling, for visualization, to enable the storage, movement and preservation of enormous quantities of data, and to address other unforeseen problems certain to arise.

Scientific visualization, including not just static images but also animation and interaction, leads to better analysis and enhanced understanding. Currently, many visualization systems are domain or application-specific and require a certain commitment to understanding or learning to use them. Making visualization services more transparent to the user lowers the threshold of usability and accessibility, and makes it possible for a wider range of users to explore a data collection. Analysis of data streams also introduces problems in data visualization and may require new approaches for representing massive, heterogeneous data streams.

Deriving knowledge from large data sets presents specific scaling problems due to the sheer number of items, dimensions, sources, users, and disparate user communities. The human ability to process visual information can augment analysis, especially when analytic results are presented in iterative and interactive ways. Visual analytics, the science of analytical reasoning enabled by interactive visual interfaces, can be used to synthesize the information content and derive insight from massive, dynamic, ambiguous, and even conflicting data. Suitable fully interactive visualizations help absorb vast amounts of data directly, to enhance one's ability to interpret and analyze otherwise overwhelming data. Researchers can thus detect the expected and discover the unexpected, uncovering hidden associations and deriving knowledge from information. As an added benefit, their insights are more easily and effectively communicated to others.

Creating and deploying visualization services requires new frameworks for distributed applications. In common with other cyberinfrastructure components, visualization requires easy-to-use, modular, extensible applications that capitalize on the reuse of existing technology. Today's successful analysis and visualization applications use a pipeline, component-based system on a single machine or across a small number of machines. Extending to the broader distributed, heterogeneous cyberinfrastructure system will require new interfaces and work in fundamental graphics and visualization algorithms that can be run across remote and distributed settings.

To address this range of needs for data tools and services, NSF will work with the broad community to identify and prioritize needs. In making investments, NSF will complement private sector efforts, for example, those producing sophisticated indexing and search tools and packaging them as data services. NSF will support projects to conduct applied research and development of promising, interoperable data tools and services; perform scalability/reliability tests to explore tool viability; develop, harden and maintain software tools and services where necessary; and harvest promising research outcomes to facilitate the transition of commercially-viable software into the private sector. Data tools created and distributed through these projects will include not only access and ease-of-use tools, but also tools to assist with data input, tools that maintain or enforce formatting standards, and tools that make it easy to include or create metadata in real time. Clearinghouses and registries from which all metadata, ontology, and



CAVE software, released by RCSB PDB and CalIT2, provides a new way of visualizing 3D macromolecular structures in an immersive, virtual reality environment. The CAVE allows users to move through and around a structure projected in the CAVE.

markup language standards are provided, publicized, and disseminated must be developed and supported, together with the tools for their implementation. Data accessibility and usability will also be improved with the support of means for automating cross-ontology translation. Collectively, these projects will be responsible for ensuring software interoperability with other components of the cyberinfrastructure, such as those generated to provide High Performance Computing capabilities and to enable the creation of Networked Resources and Virtual Organizations.

The user community will work with tool providers as active collaborators to determine requirements and to serve as early users. Scientists, educators, students and other end users think of ways to use data and tools that the developers did not consider, finding problems and testing recovery paths by triggering unanticipated behavior. Most important, an engaged set of users and testers will also demonstrate the scientific value of data collections. The value of repositories and their standards-based input and output tools arises from the way in which they enable discoveries. Testing and feedback are necessary to meet the challenges presented by current datasets that will only increase in size, often by orders of magnitude, in the future.

Finally, in addition to promoting the use of standards, tool and service developers will also promote the stability of standards. Continual changes to structure, access methods, and user interfaces mitigate against ease of use and against interoperability. Instead of altering a standard for a current need, developers will adjust their implementation of that need to fit within the standard. This is especially important for resource-limited research and education communities.

C. Developing and Implementing Coherent Data Policies

In setting priorities and making funding decisions, NSF is in a powerful position to influence data policy and management at research institutions. NSF's policy position on data is straightforward: all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved. Through a suite of coherent policies designed to recognize different data needs and requirements within communities, NSF will promote open access to well-managed data, recognizing that this is essential to continued U.S. leadership in science and engineering.

In addition to addressing the technological challenges inherent in the creation of a national data framework, NSF's data policies will be redesigned as necessary to mitigate existing sociological and cultural barriers to data sharing and access, and to bring them into accord across programs and ensure coherence. This will lead to the development of a suite of harmonized policy statements supporting data open access and usability. NSF's actions will promote a change in culture such that the collection and deposition of all appropriate digital data and associated metadata become a matter of routine for investigators in all fields. This change will be encouraged through an NSF-wide requirement for data management plans in all proposals. These plans will be considered in the merit review process and will be actively monitored post-award.

Policy and management issues in data handling occur at every level, and there is an urgent need for rational agency, national and international strategies for sustainable access, organization and use. Discussions at the interagency level on issues associated with data policies and practices will be supported by a new interagency working group on digital data formed under the auspices of the Committee on Science of the National Science and Technology Council. This group will consider not only data challenges and opportunities discussed throughout this chapter, but especially the issues of cross-agency funding and access, the provision and preservation of data to and for other agencies, and monitoring agreements as agency imperatives change with time. Formal policies must be developed to address data quality and security, ethical and legal requirements, and technical and semantic interoperability issues that will arise throughout the complete process from collection and generation to analysis and dissemination.

As already noted, many large science and engineering projects are international in scope, and thus national laws and international agreements directly affect data access and sharing practices. Differences arise over privacy and confidentiality, from cultural attitudes to ownership and use, in attitudes to intellectual property protection and its limits and exceptions, and because of national security concerns. Means by which to find common ground within the international community must continue to be explored.