

Coalition of
Academic
Supercomputing
Centers

C A S C



CREATING DIGITAL LIBRARIES FOR THE 21ST CENTURY

Alabama Supercomputing Authority
Huntsville, Alabama

Arctic Region Supercomputing Center
Fairbanks, Alaska

Arizona State University
Tempe, Arizona

Boston University Center
for Computational Science
Boston, Massachusetts

Center for Advanced Computing Research
Caltech
Pasadena, California

Center for Computational Sciences
Lexington, Kentucky

ORNL Center for Computational Sciences
Oak Ridge, Tennessee

Center for Innovative Computer
Applications at Indiana University
Bloomington, Indiana

Center for Research on
Parallel Computation
Houston, Texas

Cornell Theory Center
Ithaca, New York

High Performance Computing
Education and Research
Albuquerque, New Mexico

National Center for
Atmospheric Research
Boulder, Colorado

National Center for Supercomputing
Applications at UIUC
Champaign, Illinois

National Energy Research
Scientific Computing Center
Berkeley, California

National Supercomputer Center
for Energy and Environment
Las Vegas, Nevada

North Carolina Supercomputing
Center at MCNC
Research Triangle Park, North Carolina

Ohio Supercomputer Center
Columbus, Ohio

Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania

Purdue University
West Lafayette, Indiana

San Diego Supercomputer Center
San Diego, California

Supercomputer Computations
Research Institute
Tallahassee, Florida

Texas A&M University
Supercomputer Center
College Station, Texas

Texas Advanced Computing Center
Austin, Texas

The Pennsylvania State University
University Park, Pennsylvania

University of Florida
Gainesville, Florida

University of Maryland
College Park, Maryland

University of Southern California
Information Sciences Institute
Marina del Rey, California

University of Utah, Center
for High Performance Computing
Salt Lake City, Utah

University of Wisconsin
Madison, Wisconsin

The restructuring of the nation's library system into a high capacity digital archive of books, periodicals, drawing, manuscripts, documents and photographs that can be transmitted at high speed to computer screens around the world in original format, is rapidly becoming a reality.

This emerging "meta-library" is likely to have the broadest impact of any outgrowth of the high tech revolution to date, with millions of students, faculty and researchers making use of it via the Internet and access becoming essential for any enterprise or educational institution wishing to remain current and competitive.

In the fall of 1994 the Library of Congress announced its goal of converting the most important materials in its collections and from the collections of all public research libraries in America into digital form by the year 2000. This will provide unprecedented access to a vast repository of information and resources, including rare books, historic documents and photos, in a manner that protects the originals from overuse and vandalism. It also will resolve massive storage problems with materials that would take up countless shelves to be stored on a single disk.

It must be remembered that this awesome undertaking would not be possible without technology developed with federal support by researchers at academically based high performance computing centers and national research labs. These centers and labs are playing a key role in current efforts to digitalize all of the nation's libraries, developing software and hardware to improve storage, indexing, search and retrieval of major bodies of information. These are tasks that require the kind of fundamental R & D skills available primarily within the academic high performance computing community.

Ohio's Electronic Library -- A partnership between the Ohio Supercomputer Center (OSC) and the Ohio Library and Information Network (OhioLINK) is providing the state's public and private colleges and universities with instantaneous electronic access to a growing array of previously unavailable or difficult to obtain materials. This includes complex satellite images, paintings, maps and rare texts in original format, as well as national and international demographic and geographic information and other resources normally not affordable or manageable by individual libraries. OhioLINK is acquiring and expanding the Library's offerings, while OSC is providing large-scale storage and high speed network access for participating institutions. Together they are creating centralized access and organizing systems. OSC has multi-terabyte storage capacity, with each terabyte representing approximately 138 million typed pages.

Alabama Virtual Library -- The Alabama Supercomputing Authority is providing the technology for a statewide virtual library that will offer all students and teachers equal access to the collections of the entire system from computers at each branch location. In addition to connecting all public libraries, the goal is to include access to the nine million volume holdings of the Network of Alabama Academic Libraries. While most Alabama libraries could not possibly afford the approximately 3,000 journals judged essential to support a core education, these can be made available on-line. This is just one example of the benefits of such a resource.

Large-Scale Databases -- At the San Diego Supercomputer Center (SDSC), computer scientists are creating the digital library architecture to facilitate storage, retrieval, integration and analysis of major databases in fields like astronomy and medicine. Neurobiologists, for example, produce large data sets from three-dimensional CAT, PET and MRI scans, electronic microscope tomography and confocal light microscopes. University-based repositories, already packed with data, expect to exceed several terabytes within the next five years, with each terabyte translating to approximately 138 million printed pages. SDSC is participating in several collaborations to create a digital "atlas" of brain imaging that connects such databases and enables neuroscientists to advance their research by linking high performance computing visualization and 3-D data acquisition.

Another example is in astronomy, where high performance computing and new data handling resources are being combined with large-scale digital sky surveys. By integrating all sky surveys in optical, infrared and radio wavelength, astronomers will have the unprecedented capability of performing detailed studies across the entire collection.

"Biology Workbench" -- While biologists increasingly rely on high performance computing for their research, information collected by thousands of independent labs is being generated in formats, vocabularies and interfaces that are not compatible with one another. In effect they do not speak the same computational language. The National Center for Supercomputing Applications has developed a new software tool to act as a sort of universal interpreter, making information from more than 100 public databases accessible to the entire biomedical community. As easily as surfing the Web, they can use this tool to search all major molecular and structural biology databases available on the Internet. *Biology Workbench* functions like a simple extension of the user's hard drive, and may well be a model for the future of Web-based computing in many fields.

Improved Search Engines -- As its part of the federally funded Digital Library Initiative (DLI), the National Center for Supercomputing Applications (NCSA) is designing a more intelligent search engine for exploring documents and publications on the Internet. This technique called "semantic extraction" involves looking for words and phrases in their semantic context, rather than the current approach, which focuses solely on individual words in phrases without regard to their particular meaning to the researcher. The new engine first looks for common word and phrases in a document, then considers how often and in what context they frequently appear. Finally, it searches for related documents that are similar in content, noting those that have the same phrases appearing in the same context.

With the DLI project still in its infancy, these "semantic extractions" are being run with relatively small amounts of data -- at least by supercomputing standards. An application designed by NCSA's DLI researchers recently completed a semantic sort of nearly 2,000 full-text journal articles. The trial considered more than two million unique phrases, yet took only about 24 hours on NCSA's most advanced supercomputer.