NCSA
NASA
• PROJECT •
HORIZON
Test Applications and Digital Library Technologies in Support of Public Access to Earth and Space Science Data

HORIZON: A Digital Library Project for Earth and Space Data Serving the Public

R. Wilhelmson[1,2], M. Folk[1], M. Ramamurthy[2], B. Schatz[1,3,5], N. Yeager[1], R. Crutcher[1,4],
and M. Winslett[3]
[1]National Center for Supercomputing Applications
[2]Department of Atmospheric Sciences
[3]Department of Computer Science
[4]Astronomy Department
[5]Graduate School of Library and Information Science

University of Illinois at Champaign-Urbana (bw@ncsa.uiuc.edu)

1. Introduction

Project Horizon is a multifaceted project that seeks to enhance digital library technology, specifically World Wide Web (WWW) client and server technology, in support of easy to use and scalable public access for locating, moving, and analyzing both earth and space science data. The project is funded by NASA and centered at the National Center for Supercomputing Applications (NCSA), the developer of NCSA Mosaic and the Hierarchical Data Format (HDF) now being used in the EOS project. Activities integrate and leverage off of existing research, development, and discipline-specific activities at NCSA and at the University of Illinois Departments of Astronomy, Atmospheric Sciences, Computer Science, and Library and Information Science. Client side efforts include the implementation of the Interactive Image Client Environment (IICE) providing enhanced functionality on the client (user) machine. Server side developments involve HDF (and netCDF) file browsing capabilities and file format conversions. Research and development areas include efficient access to large data sets, scalable server technologies, and next-generation information systems. Finally, there are two WWW testbed servers, one for earth science (The Daily Planet) and one for space science (Astronomy Digital Image Library), that provide working real world applications to thoroughly test and demonstrate the above technologies.

Project information is available on the WWW at http://www.atmos.uiuc.edu/horizon/ . A brief summary of some of these activities is included in this preprint along with additional WWW addresses**.**

2. Interactive Images on the WWW

Interactive Image Client Environment (IICE) extends the interactive capabilities of traditional Web environments using client hardware capabilities. This enables the generation of customized interactive products on the client machine that would be too costly to produce in mass on the server. This approach scales with increasing users, each with their own computer, whereas scaling a server environment to meet demand requires the addition of new server machines.

IICE is being developed using the Java from Sun Microsystems. It consists of an object-oriented C++ - like Java language, a Java interpreter and a WWW browser called HotJava that has the interpreter imbedded in it. Hot Java is being developed for many of the common platforms including UNIX, Mac, and Windows. Use of this capability means that applications (called applets) can be written once in Java and then executed on any of these clients rather than writing the application for each platform. The interface/output then appears directly inside an HTML document.

Initially applets are being written to interpret images through translation of pixel values into physical quantities (e.g. a radiance level into temperature), to overlay vector plots/images, and to allow animation with direction and speed under user control (Wojtowicz and Wilhelmson, 1996).

3. Scientific Data Server

3.1 An HDF Browser

Several efforts are underway to provide additional capabilities related to the use of HDF and the WWW. A server side HDF browser has been developed for displaying HDF browse versions of images, tables, arrays, attributes, and text. One prototype version of this scientific data server/browser (SDS) provides access to "objects" stored in HDF (http://hdf.ncsa.uiuc.edu:4321/). By using the new subsetting and subsampling capabilities of HDF 4.0, the server also allows users to extract subsets and subsamples from HDF arrays. Since the resulting HTML documents contain tables of ascii values, these numbers can be easily copied and pasted into other applications, such as spreadsheets, for further analysis and computation. A simple search engine has also been constructed to retrieve data based on keyword matching of metadata parameters and full-text search of annotations and attribute values.

Another prototype version is currently being used for Coastal Zone Color Scanner (CZCS) files at Goddard Space Flight Center's DAAC (http://daac.gsfc.nasa.gov/WORKINPROGRESS/ OCDST/czcs_data.html). Currently there are two versions of the HDF browser/server configured to facilitate browsing for level 2 and level 3 CZCS products. Level 2 products include the North Atlantic Region (1981), classic images of interest, and the Chesapeake Bay Region. Level 3 products include regional and global composites such as the (1X1) chlorophyll mean. In these versions, images are converted into GIF format and embedded within HTML (Hypertext Markup Language) pages. This conversion is done on the server but will be extended to the client in the future, provided that the client has a working version of HDF installed.

The current HDF server/browser supports netCDF data (not image) files. In addition, two-way conversion between FITS, a commonly used format in observational astronomy, and HDF extends the usefulness of the browser for ADIL (http://hdf.ncsa.uiuc.edu:8001/fits/).

3.2 Panda Software

Large multidimensional arrays are a common data type in high-performance scientific applications. In addition, large arrays will become increasingly available through the WWW. Without special techniques for handling access to these arrays, i/o can easily become a large fraction of execution time for scientific and information applications using these arrays, especially on parallel platforms. The Panda (Persistence and Arrays - http://bunny.cs.uiuc.edu/CADR/panda.html) software library provides high-level abstract interfaces that free the application developer from the need to consider low-level physical storage details to reach acceptable i/o performance levels. It also provides advanced i/o support through efficient layout alternatives (appropriate chunking of the array) on disk and in main memory for large multidimensional arrays in addition to support for high performance array i/o operations. The high level interfaces for the array i/o operations provide ease of use and enhance the portability of an application.

Panda is built on top of existing commodity file systems such as AIX. The high-level interfaces provide ease of use, application portability, and, most importantly, allow plenty of flexibility for an efficient underlying implementation. The Panda i/o library exhibits excellent performance on the massively parallel SP2, attaining 83-98% of peak AIX performance on each i/o node in test experiments. It shows excellent scalability with data size and increased number of processors, and it provides very high throughput compared to ordinary AIX file system performance. These results can be traced to Panda's use of 'server-directed i/o', the high-level user interface, and built-in facilities to rearrange arrays from one physical schema to another during i/o. There is also a Sun version and ports to other systems are underway.

Research within the HORIZON project is aimed at improving access time for subsampling large arrays. This is expected to be a major bottleneck for accessing large remotely-sensed images on servers and in parallel environments.

4. Information Systems Research

The Net of the Twenty-First Century will have a very different character than at present. There will be billions of repositories, both large and small, maintained and indexed by many communities. These repositories are organized collections of information, each with indices and a search engine. For example, one repository might include journal articles with a sophisticated search engine that can look for words and word patterns in the main text, in figure captions, or in tables. Another could include large scientific data sets such as those kept at NASA DAAC sites along with a pattern recognition search engine and notes from scientists studying the data. These capabilities go well beyond what current WWW browsing and server capabilities provide. Because of the large number of these "distributed" repositories, substantial organization and cross-correlation of information is needed in order for users to most effectively make use of them. A typical session with the Net in the future will be like a reference session in a large library where a person moves through interlinked resources seeking information. In a digital library, the user will navigate themselves through multiple interlinked network-based resources around the world to explore answers to their questions.

The development of repositories is underway. Then, once it becomes easy for users to publish their own documents and maintain their own repositories, issues of navigation and analysis outweigh those of access and organization. There are several efforts partially supported by the HORIZON project that are aimed at scalable access, organization, and analysis.

4.1 Scalable servers

One of the key issues in providing a scalable and distributed repository environment is the naming of WWW documents. After investigating several strategies, the Corporation for National Research Initiatives' (CNRI) "Handle" server is being adopted. A handle or Uniform Resource Name (URN) is a unique, location-independent, and permanent document identifier. Documents are referred to by URNs, not URLs as currently done. When the location of the document is changed, the embedded links (URNs) that point to that document do not break. Rather, the document's URL is changed in the URN database. These URNs are intended to be used with the WWW in the short term and with larger non-WWW based systems in the future.

The handle management system includes a handle generator and a handle server. The server consists of a caching server and a global set of base handle servers and includes secure tools for administration of handles. The server is designed to resolve very large numbers of handles rapidly, and to scale without limit. URNs will be deployed in several testbeds to investigate their reliability and eventually the CNRI "global" handle will be employed in the NCSA production environment for use by the general community.

4.2 Secure Object Repositories

Another issue in the development of repositories is the storage of secure objects that include documents, data, and code for manipulating them. Staff at NCSA, CNRI, and Cornell are designing and implemented a secure object repository infrastructure that will ensure contractual "Terms and Conditions" including copyrights and payment. In addition, a prototype certificate authority will be implemented to provide needed authentication for users of the repository.

4.3 Interspace

The Interspace (http://csl.ncsa.uiuc.edu/IS.html) is a future system that seeks to unify disparate distributed information resources in one coherent model. The Interspace, as a entity, is a collection of interlinked information spaces where each component space encodes the knowledge of a community or a subject domain. An information space is a collection of interlinked objects. With the Interspace, users can cross-correlate information in multiple ways from multiple sources. Standard services include inter-object linking, remote execution, object caching, and support for compound objects (usually referred to as compound documents).

Ultimately, the Interspace system is an attempt to represent all types of data/objects in one flexible, cohesive, and scalable system. Navigating information paths and grouping related items is a fundamental operation in the Interspace. Semantic retrieval and community classification, with interactive support for vocabulary switching across domains and subject indexing for amateur classifiers, is provided.

The architecture of the Interspace has been defined and implementation of a prototype is now underway. The Interspace architecture is an application environment for interconnecting spaces for manipulating information, much as the Internet is a protocol environment for interconnecting networks for transmitting data. Effectively implementing this environment will require in the future building operating systems incorporating both computer science research on distributed objects and information science research on semantic retrieval.

The prototype system being implemented will be tested with NASA datasets (Earth & space science images). It will also be linked to the NSF/ARPA/NASA funded Digital Library Initiative testbed at the University of Illinois that is focusing on engineering and science journals (http://www.grainger.uiuc.edu/dli/ and Schatz et al., 1996).

5. Testbed Servers

There are two testbed servers for the HORIZON project, one for astronomy images called ADIL (Astronomy Digital Image Library) and one for atmospheric and environmental science information called TDP (The Daily Planet™).

5.1 ADIL and AipsView

ADIL is being built for use by astronomers and the public to access astronomy images. These two user communities will use their own appropriately designed HTML interface for locating and displaying the image data. The public will use more popular terminology to locate historical and current images (to be implemented) while an astronomer will browse the library "card" index, searching on object names, regions of the sky, object types (galaxies, molecular clouds, etc.), type of data (continuum map, spectral-line datacube, etc.), frequency and/or spectral-line transition, etc. (see http://monet.ncsa.uiuc.edu). Both interfaces are provided through a network browser (e.g., NCSA Mosaic or Netscape) which queries a relational database. When a set of files of interest has been located, a postage stamp representation of the image or a short movie of a data cube is displayed. The full image can then be acquired by simply clicking on the data transfer button.

There are several ways in which the digital library will be of benefit. Observers planning a project will have access to work previously carried out on the same object, perhaps at different frequencies or in different spectral lines. Visualization and analysis tools can be used to re-bin and re-register image data sets, so astronomers will be able to use the archive for easy and straightforward overlay of existing images for comparison and analysis. Full resolution color images can then be acquired with useful metadata provided for astronomers and interpreted for the public.

ADIL has now been linked to the NASA Astrophysical Data Systems (ADS - http://adswww.harvard.edu/) Abstract Service. Image Preview pages in the Library contain direct links to related abstracts in the ADS abstract database. Similarly, ADS users who locate abstracts related to images in the Library may now access the images from the ADS interface. Currently work is underway to combine ADS abstract searches with ADIL searches in order to improve the location of images.

AipsView is being developed to aid in (http://monet.ncsa.uiuc.edu/AipsView/av.html) visualization of ADIL images. Currently, it is a tool for two-dimensional visualization and relies on Motif and Xlib for its user interface and drawing capabilities. A companion tool for three-dimensional visualization, AipsView3, requires OpenInventor, and is in the early stages of development.

AipsView has an easy to use graphical user interface and can read FITS image files and single-SDS files written in HDF format. Images are displayed in 2D including orthogonal slices from 3D data cubes. Image scaling, animation, simultaneous display of multiple images, synchronized animated display through multiple data cubes, and other image standard functionality is provided. AipsView requires a reasonably sized machine, with "reasonable" amounts of memory and swap space. It has been tested on SUN (SunOS and Solaris), SGI, HP, IBM RS6000, and DEC Alpha, and as a client of a MacExodus server.

5.2 The Daily Planet™

The Daily Planet™ (http://www.atmos.uiuc.edu/) is a WWW-based environmental information server developed in the Department of Atmospheric Sciences that provides access to meteorological, climatological, hydrological, and Earth Observing System (EOS) databases, multimedia educational modules, distributed archives of data sets (both real-time and retrospective), and other Internet-based resources. It has a large national user base and experiences up to 80,000-180,000 requests per day. It is used for education at all levels including teachers and students associated with the CoVis Project (http://www.covis.nwu.edu/).

The goal of CoVis is to bring together scientists, teachers and high school students through the use of high-speed computing and communication technologies to carry out projects as part of a high school student's learning process. Through video-teleconferences, students at their schools can interact with mentors at universities and can have access to daily and historical weather information over the Internet. Instructional modules have been and are being developed with text and graphics augmented by sound, animation, and video to aid the student in learning about atmospheric science.

Recently through CoVis and HORIZON funding, the Weather Visualizer was announced (Ramamurthy et al., 1996). It allows users to generate customized weather images "on the fly" from real-time weather data. This enables the user to display only information of interest to them. The introductory document of the Weather Visualizer consists of a graphical panel with six weather categories: Surface Observations, Upper Air Observations, Upper Air Soundings, Radar Summary, Satellite Imagery, and Forecast Images.

For each of these categories there is an HTML document that presents choices and solicits input as to which meteorological parameters to display. When selection of the parameters is completed, the form is processed on the server, resulting in the return of the image, plot, or textual data requested. Typical end products would be a map of US surface observations, radar echo summary, and frontal analysis superimposed on an infrared satellite image background; a table of forecast model output statistics; or a Stuve thermodynamic diagram.

An additional feature is the generous use of “helper sections”, which employ the hypertext functionality of HTML to explain what the various parameters and map items mean, and how the items or products are typically used in interpreting the data. For example, context sensitive help and explanations make complex imagery understandable by novices. The range of analysis options available will make this tool valuable to advanced users as well.

The increasing use of this software over the WWW creates an important testbed environment for the HORIZON project. The server acts as both a document source and as a compute engine creating the weather displays. This places a significant new burden on the server. In order to explore the impact of increased usage, five new HP715/100 workstations and 45 Gbytes of disk will be coupled with the current HP715/75 and HP720 server and preprocessed product generation machines. An additional new HP715/64 workstation will act as a router to distribute incoming requests between the machines.

6. Acknowledgments

The HORIZON Project is funded through NASA IITA CAN-OA-94-1. The Digital Library Technology (DLT) project and the Public Use of Remote Sensing Data (RSD) project are two related elements of the Information Infrastructure Technology and Applications (IITA) component of NASA's High Performance Computing and Communications Program (HPCC). Further information can be found on the WWW at http://dlt.gsfc.nasa.gov/

7. References

Ramamurthy, M. K., R. B. Wilhelmson, J. G. Kemp, S. E. Hall, M. Sridhar, W. L. Chapman, B. Fishman, D. Gordin, R. Pea, and L. Gomez, 1996: CoVis Geosciences Web Server: An internet-based resource for the K-12 community. Preprints, 4th Symposium on Education, Atlanta, Georgia, AMS.

Ramamurthy, M. K., R. B. Wilhelmson, R. D. Pea, L. M. Gomez, and D. C. Edelson, 1995: CoVis: A National Science Education Collaboratory. Fourth Symposium on Education, January 16-17, Dallas, Texas.

Schatz, B., . Mischo, T. Cole, J. Hardin, L. Jackson, A. Bishop, L. Star, P. Cochrane, and H. Chen,1996: Digital Library Infrastructure for a University Engineering Community: Towards Search in the Net via Structure and Semantics. IEEE Computer special issue on Large-Scale Digital Libraries, 12pp. (to appear).

Wojtowicz, D. P., and R. B. Wilhelmson, 1996: The Interactive Image Markup Language (IIML) for web use. Preprints, 12th International Conference on Interactive Information Processing Systems for Meteorology, Oceanography and Hydrology, Atlanta, Georgia, AMS.