A Digital Library for Geographically Referenced Materials

Terence R. Smith
University of California, Santa Barbara

ADL will provide on-line public access to maps, photos, and other information referenced in geographic terms. Much of this data currently is found only at major research libraries.

The Alexandria Project's goal is to build a distributed digital library for materials that are referenced in geographic terms, such as by the names of communities or the types of geological features found in the material. The Alexandria Digital Library (ADL) will comprise a set of Internet nodes implementing combinations of the four primary ADL architecture components—collections, catalogs, interfaces, and ingest facilities (which a digital library uses to add documents and information about document cataloging and access).

The ADL will give users Internet access to and allow information extraction from broad classes of geographically referenced materials. In this case, having access means being able to browse, view, and download data and metadata. Information extraction involves the application of local or remote procedures to selected data and metadata.

ADL's holdings focus on collections of geographically referenced materials, including maps, satellite images, digitized aerial photographs, specialized textual material (such as gazetteers), and their associated metadata. We are extending these collections to more general classes of graphical and textual materials that have references to geographic objects.

Presently, geographically referenced information is largely inaccessible. Many important collections exist only on paper or film, and the larger collections are found only in major research libraries. The University of California, Santa Barbara (UCSB), Map and Imagery Laboratory collection, for example, contains 2 million historically valuable aerial photographs, along with the only negatives of many of these images. Where such data already exist in digital form, their accessibility is hindered by the size of individual holdings (satellite images commonly range from 100 Mbytes to 1 Gbyte) and by the difficulty of searching large collections.

To make geographically referenced information more accessible and usable, the ADL must provide user interfaces and on-line catalogs that support the formulation and evaluation of geographically constrained queries. We want the ADL to present multiple interfaces to accommodate users with various backgrounds and needs. For example, a schoolchild looking for a map of nearby rivers and trails for a camping trip will have different needs and expectations than a scientist looking for elevation and rainfall data sets for the development and testing of a vegetation distribution model.

GENERAL ADL STRATEGY AND ARCHITECTURE

The Alexandria Project's development emphasizes

- the digital library architecture's user-interface and catalog components,
- collections of geographically referenced materials,
- Internet accessibility for many users,

0018-9162/96/$5.00 © 1996 IEEE

- incremental and evolutionary design and implementation,
- digitally supportable extensions to traditional library functionality, and
- access to explicit and implicit digital library information.

The first ADL development cycle yielded a stand-alone rapid prototype system.[1] The second, current cycle provides a superset of the rapid prototype's functionality, called the Web prototype, via the World Wide Web (WWW). The next cycle will focus on developing a distributed catalog incorporating a general metadata model.

Figure 1 illustrates the basic ADL architecture, which derives from a traditional library's four major components. The catalog component includes metadata and search engines that let users identify holdings of interest. The storage component contains digital holdings, organized into collections. The user interface supports graphic-based and text-based access to the other ADL components and services. Librarians use the ingest component to store new holdings, extract metadata from holdings, and add metadata to the catalog component.

The rapid and Web prototypes' architectures are special cases of the general architecture, with differing languages and protocols at the component interfaces. Figure 1 illustrates the languages and protocols used in the Web prototype. The Web prototype's storage and catalog components are distributed, unlike those of the rapid prototype version.

THE CATALOG COMPONENT

A digital library's catalog component lets users map their information requirements into the library collection's most appropriate information set. While a traditional library cataloging system (based on author, title, and subject) serves as a digital library's basic catalog component model, it is inadequate for geographically referenced holdings, such as maps and images. Catalogs for such information must additionally support access to holdings in terms of their representations, their spatial footprints (the location of objects in the individual holdings), and their contents.

Digital library technology greatly increases our ability to extract, store, and search new classes of metadata about library holdings. A major thrust of ADL activity is thus to extend current catalog and metadata models. The ADL will also support catalog interoperability by using standards to represent and exchange catalog information.

To meet these criteria, we developed a rapid prototype catalog schema by using elements from the US Machine-Readable Cataloging (USMARC) standard and Federal Geographic Data Committee (FGDC) metadata standards.[1] We then expanded the Web prototype schema to include metadata supporting simple content-based queries.

Basic metadata: USMARC and FGDC standards

The basic metadata for geographically referenced information in the rapid and Web prototypes combine elements from the USMARC[2] and the FGDC[3] metadata standards.

Since the 1960s, USMARC has been a national standard for library holdings' database descriptions. It includes fields for cataloging analog geographic data and open-ended local-use fields that can accommodate digital data. The USMARC standard contains fields like those in the FGDC standard, as well as a thesaurus-based field that permits references to specific thesauri, which are used to find terms that can be used to conduct data searches.

USMARC stores a given holding's metadata in one record with four components—a leader, a record directory, control fields, and variable fields.[2] This "flat" structure, while not optimal for a relational database, is useful for specifying metadata I/O functions and for exchanging metadata records between digital libraries.

The FGDC promotes the coordinated development, use, and dissemination of surveys, maps, and related spatial data.[3] All US federal agencies are required to use the FGDC's digital geospatial data metadata standard.[4]

The FGDC standard provides definitions for relatively few fields, along with their relations within a hierarchical structure. While these fields are adequate for cataloging digital geospatial data, they do not accommodate analog spatial materials. Moreover, the FGDC standard does not specify a metadata representation format or structure, which results in a variety of implementations and a lack of generic import/export functions.

By combining the FGDC and USMARC standards, the ADL has been able to catalog all forms of spatial data thus far encountered, including remote-sensing imagery, digitized maps, digital raster and vector data sets, text, videos, and remote WWW servers. The Web prototype's metadata schema has about 350 fields, including all FGDC fields and selected USMARC fields. To create the schema, we converted the FGDC production rules and the USMARC record hierarchy into one normalized entity-relationship data model, from which CASE tools automatically generate the physical database schemata.

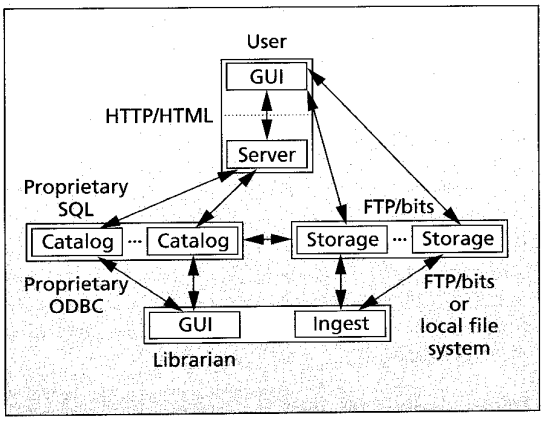

Figure 1. The main Web prototype components. This basic ADL architecture derives from a traditional library's four major components—the catalog of holdings; the storage area, organized into collections; the user interface, for access to library services; and the ingest facility, for storing and processing data from new holdings. GUI means graphical user interface. ODBC means Open DataBase Connectivity.

The ADL gazetteer

The Web prototype catalog incorporates two major extensions of the combined FGDC-USMARC metadata model, both supporting content-based search forms. The first extension allows searches of digital image holdings for occurrences of preselected image features, such as textures.

The second extension allows retrieval of ADL holdings based on the relationship between the footprints of holdings and the footprints of named geographic features, such as cities, rivers, and mountains. Lists of such features and their footprints are commonly called gazetteers. Gazetteers also include a brief description of each holding's geographic feature type, such as "populated place," often organized as a class hierarchy.

The ADL gazetteer is a union of names and features from two large standard gazetteers, as well as an intersection of their feature classes. One of the gazetteers is maintained by the US Geological Survey's Geographic Names Information System, the other by the US Board of Geographic Names. The Geological Survey gazetteer contains the names of about 1.8 million features, organized hierarchically into 15 feature classes. The Board of Geographic Names gazetteer contains the names of about 4.5 million land and undersea features.

As the ADL catalog grows, spatial indexing methods play an increasingly important role in supporting footprint queries.

The ADL gazetteer is maintained in the ADL catalog database but is also available for external search by the Excalibur semantic network text-retrieval engine. We have found the gazetteer's Excalibur version useful for fuzzy searches, where a user may not know the precise spelling or name of a feature, such as an airport.

ADL gazetteer use entails two significant research issues. First, different gazetteers use different terminologies and hierarchies to describe the same features. So far, we have been able to construct "crosswalks" between gazetteers by matching their reference documents' fields and definitions.

A second issue involves the exact nature of a feature's footprint. For example, is the footprint of "Santa Barbara" the city limits, City Hall, or the county boundary? Existing gazetteers often give the location of each feature, even those that cover large areas, only as a point on a map. It is often unclear how the points are chosen and whether they are centroids, corners, or arbitrarily chosen points. Features with only fuzzy footprints, such as Southern California or the Sierra Nevada mountains, complicate matters further. A person's notion of the spatial extent of these features is inherently fuzzy, so they are particularly difficult to specify.

Other catalog issues

As the ADL catalog grows, spatial indexing methods play an increasingly important role in supporting footprint queries. We are investigating various methods for indexing multidimensional hierarchical data,[5] such as footprints. In particular, we have extended Balanced-trees to Interval B-trees, which accommodate objects that span a range of values (intervals) rather than single values (points) in the data space. IB-trees decompose data objects with a given number of dimensions into the same number of intervals and then index the intervals on each dimension separately.

Although ADL's primary external interface is the WWW, the ADL catalog also supports a Z39.50 interface, which is the traditional library catalog's standard on-line protocol and the National Spatial Data Infrastructure's current standard search protocol. The FGDC coordinates the National Spatial Data Infrastructure as a collection of Z39.50 servers supporting queries against FGDC-compliant metadata.

THE USER INTERFACE

The ADL's user interface lets users

- compose spatial search queries,
- display geographically referenced materials in raster and vector formats,
- browse search results,
- employ user-configurable defaults and options, and
- retrieve data holdings in various native formats.

The rapid prototype's user interface, based on the ArcView geographic information system software package,[6] supports the first three functions.

User interface issues

The ADL Web prototype must operate within the following WWW limitations:

- Current WWW Hypertext Markup Language (HTML) interpretations lack mechanisms for presenting vector data and barely support the entry and display of geographically referenced information.
- The WWW's Hypertext Transfer Protocol (HTTP) is stateless and is designed for small, fast transactions.
- Current WWW browsers are insufficiently interactive. Helper applications are only a short-term substitute for better browser-helper communications and/or programmable browsers.

We know of no WWW browser that supports vector data display, nor does HTML make any explicit provision for vector data. This presents us with a serious challenge, given the large amount of vector data in the ADL collections. It is very difficult to input vector data, such as a geographic search region's definition. It would be natural to draw a polygon on a base map by using a mouse to either click on multiple points or click and drag over a desired region. However, these actions are not supported by current WWW browsers, which immediately send an HTTP request after a user-input event, such as a mouse click.

HTTP's statelessness hinders browsing and searching. By default, once a server responds to a client's HTTP request, neither the client nor the server retains any state or memory of the transaction, other than perhaps a log of the URL involved. This makes it difficult to implement such essential features as per-user configurations and iteratively refined searches. To simulate a stateful connection, such as a session, information must be explicitly maintained by either the client (in parameters stored in the URL or in hidden variables in the HTML form) or the server (in unique user identifiers and a session database).

A user interface should be user customizable and capable of saving a particular configuration for future use. Additionally, a user must be able to retrieve a particular data item or metadata record. Since the WWW is part of the Internet, simple file-transfer protocol bulk retrieval is straightforward. However, the ADL holdings are often extremely large, so users also require methods that let them extract and progressively transfer relatively small data-holding increments.

User interface implementation

Conceptually, the Web prototype's user interface is a collection of HTML "pages" that implement control/configuration and help/glossary links, as well as three major search capabilities—map browsing, gazetteer queries, and general catalog queries.

The user interface is designed around a state-transition model with each state representing a WWW form or page, including some that include partial or complete query results. About 25,000 lines of Tcl code running in a NaviServer HTTP server dynamically generate the HTML code for the Web prototype's user interface.

The primary function of the map browser and the gazetteer pages is to let the user define spatial extents or regions for catalog searches. The map browser defines these search regions explicitly, by zooming and panning a base map, while the gazetteer defines them implicitly as the footprints corresponding to place names and feature types. Figure 2 shows a map browser.

The visible portion of the map browser's base map (the display window) is the default search footprint (the query window). However, this relationship can be modified. For example, the user may specify a display window subset or may direct that the display window be ignored. The base map is also the background on which the gazetteer and catalog query result footprints are drawn. The base map images are dynamically generated by a Common Gateway Interface application based on the Xerox PARC Map Viewer (see URL http://mapweb.parc. xerox.com/map/), which we have modified to support generic labeling, fast panning, and graphic overlay production.

Gazetteer queries may interact with the map browser. For example, if a map browser query window contains the USA but not Europe, then a gazetteer query for Paris will return Paris, Texas, but not Paris, France. The map browser, in turn, may be directed to reset the query window to the smallest geographic rectangle that bounds the gazetteer query result.

Query windows resulting from map browser-gazetteer interactions are ultimately passed to the catalog page for incorporation into catalog queries. The catalog page lets the user search against geographic footprints and any metadata field (such as theme, time, or author) expressed as textual or numeric values in the ADL catalog.

Catalog queries are assembled from user input into a generic conjunctive normal form representation and are then translated to the specific query language (currently the Structured Query Language) of the catalog database-management system. (To support the catalog, we are evaluating the Illustra, O2, Oracle, and Sybase database-management systems.) Query results are converted to HTML tables with hyperlinks to browse images and online holdings. Query results are presented incrementally, with a metadata field subset displayed initially and complete fields subsequently displayed for user-selected hold-

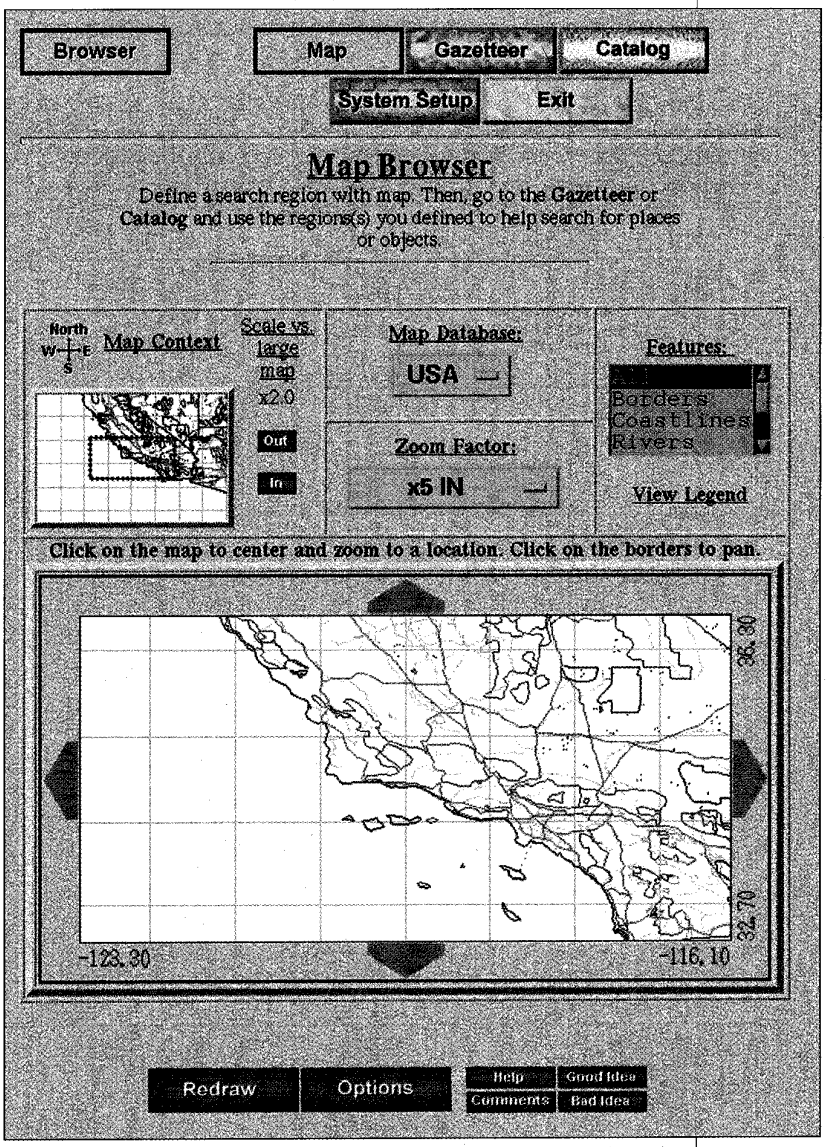

Figure 2. The ADL's map browser component. The browser lets users define the geographic regions they want to search for in image catalogs. The search is done by zooming and panning a base map.

ings. The format and fields used in the query results are user-configurable.

Queries may also return ADL holdings' footprints, which can be displayed on the map browser base map. Unfortunately, it is common for a catalog query to return many more footprints than the map browser's small display can show legibly. When multiple data holdings' footprints are displayed on the same map, it is difficult to distinguish which footprint is associated with which item. We continue to experiment with heuristics and visual aids, such as clustering and labeling, to eliminate this confusion. In Figure 3 we show examples of the browse graphics that may be the partial result of a query.

The Web prototype's user interface stores all user-configuration parameters, query statements, and current query result sets in a database that is separate from the catalog and that is maintained by the NaviServer HTTP server. This state information may also be stored on the client side in "hidden" HTML form variables. This lets a user save an ADL session by using the browser's save-page feature. The user may restore a session by reloading the saved page. Otherwise, the server handles state maintenance using a minimal opaque client-side handle to identify the current session.

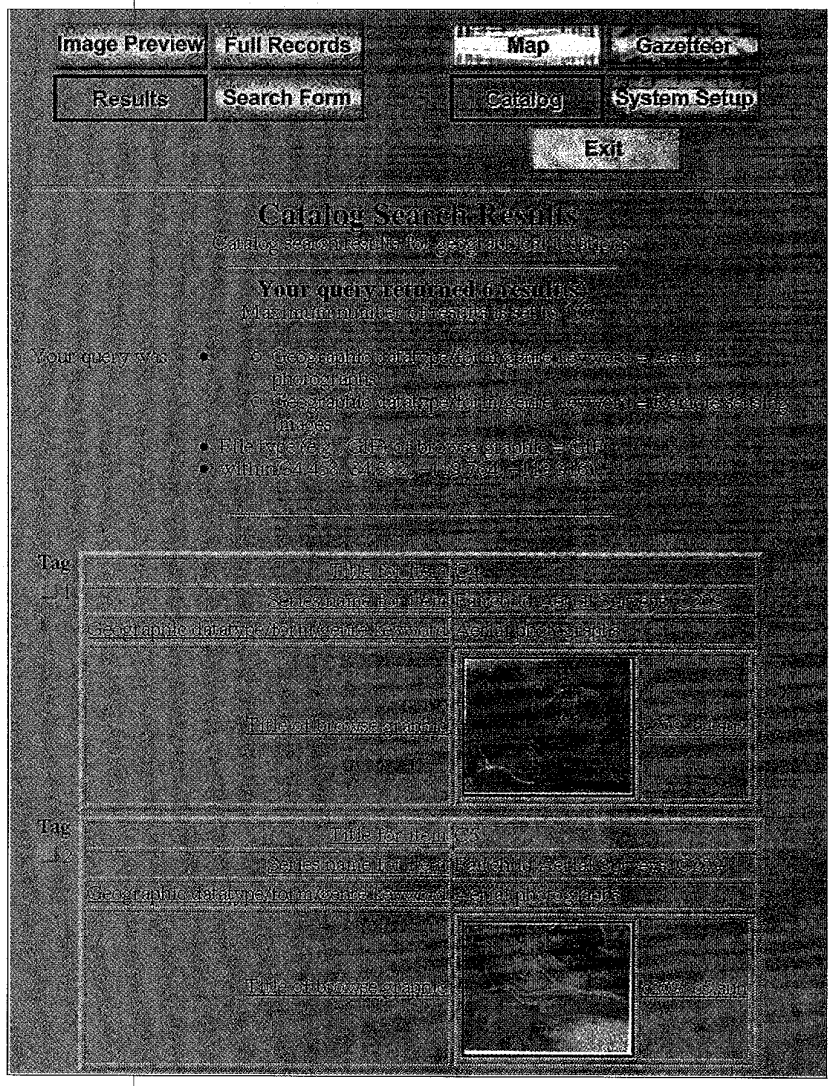

Figure 3. The display of catalog search results shows a response to an image catalog query. The requested images are shown along with pertinent information.

IMAGE PROCESSING AND PARALLEL PROCESSING

We are applying image-processing and parallel-processing technologies to a range of ADL issues. Image processing has implications for efficient storage, access, and retrieval of digital-library holdings. Parallel processing is important for ensuring adequate performance by heavily used digital libraries.

Image processing

Bandwidth and/or storage limitations often make it impractical to retrieve a large image from a digital library as a single item. Furthermore, different users may want different image resolution levels. Maintaining hierarchical, multiscale representations of image data generally solves these problems. We employ wavelet transforms.[7]

Wavelets have been widely used in many image-processing applications, including compression, enhancement, reconstruction, and image analysis. Fast algorithms exist for computing the forward and inverse wavelet transforms, and users can easily reconstruct desired intermediate levels. In addition, the transformed images (wavelet coefficients) map naturally into hierarchical storage structures.

We are also applying image-processing techniques to achieve content-based access to digital-library holdings. Our current implementation uses texture to describe and catalog a library of images' content.

BROWSING AND PROGRESSIVE DELIVERY. In wavelet decomposition, the lowest-resolution components may be used conveniently as thumbnail images for browsing. Experience with thumbnail images in the rapid prototype convinced us that they are invaluable for rapidly evaluating a large number of images. With wavelets we can support a richer browsing model in which users may zoom in on a given region until they reach an acceptable level of detail. Wavelet transformations support the rapid delivery of the low-resolution browse images and the incremental higher-resolution components.

Current WWW browsers cannot display wavelet data directly. The Web prototype avoids this restriction with a customized helper application invoked by the client browser when it receives an image of a Multipurpose Internet Mail Extension-type "wavelet." The helper application retains

Figure 4. The image browsing tool for large aerial photos. The reduced-resolution version (left) of a large aerial photograph is searched for housing developments, which are then shown in thumbnail images (right).

the previously downloaded components so that the Web prototypes' user interface only has to transmit the next component in response to a request for higher-resolution data.

The helper application is not our preferred long-term wavelet display solution because it requires us to make a locally developed executable program available for all possible ADL client hardware/software environments. We are pursuing development of an inverse wavelet transform as an "applet" in a portable language, such as Java, that can be downloaded into a standard WWW browser, such as Netscape.

TEXTURE-BASED RETRIEVAL. Content-based retrieval is critical for accessing large digital image collections. The ADL project team is investigating the use of texture as a basis for content-based search,[8] initially by adding catalog indices based on image texture features. Texture information is extracted from images as they are ingested, using banks of Gabor (modulated Gaussian) filters. This is roughly equivalent to extracting lines, edges, and bars from the images at different scales and orientations. We then use simple statistical features of the filtered outputs, such as mean and standard deviation, to match and index images.

The Web prototype catalog includes a texture-template database that can be matched with textures extracted from ADL collection holdings. Initiating a search by choosing an image region is just one access class enabled by this information. We will use the region's texture to retrieve matching texture templates, which will refer us to the ADL holdings in which they occur.

Figure 4 shows an example of browsing large aerial photographs by using reduced-resolution versions and even smaller thumbnail images, which can be searched for particular geographic feature types. In the figure, the larger image, which is a reduced-resolution version of an aerial photograph, was searched for housing developments, which are shown in the thumbnail images. The original photograph is 5,000 pixels by 5,000 pixels, the reduced-resolution version is 512 pixels by 512 pixels, and the thumbnail images are 64 pixels by 64 pixels.

Parallel processing

The Alexandria Project team is investigating parallel computation[9, 10] to address various performance issues, including multiprocessor servers, parallel I/O, and parallel wavelet transforms, both forward (for image ingest) and inverse (for efficient multiscale image browsing).

We have developed a prototype parallel HTTP server containing a set of collaborative processing units, each capable of handling a user request. The server's distinguishing feature is resource optimization based on close collaboration of multiple processing units. Each unit is a workstation (for example, a Sun Sparc or a Meiko CS-2 node) linked to a local disk. The disks are mounted at a network file system to all processing units. Resource constraints affecting server performance are

- processing unit speed and memory size,
- the background load that is imposed by nonserver processes,

- I/O bandwidth between the processing unit and its local disk,
- network latency and bandwidth between a processing unit and a remote disk, and
- disk contention when multiple I/O requests are accessing the same disk.

We actively monitor the system resource units' CPU, disk I/O, and network loads and then dynamically schedule incoming HTTP requests to the appropriate node. This keeps the server's performance relatively insensitive to request load while allowing it to scale upward with additional resources. In simulations, response time improved significantly by using multiple processing units and did not change significantly when the request rate increased, even up to 30 million per week.

We observed similar response speedups using a multinode server while varying the size of the retrieved image files, which are typical ADL holdings. Since ADL requests' computational and I/O demands vary dramatically for large images and complex metadata, the load-balancing approach offers a 20 to 50 percent performance improvement over a simple round-robin approach.

ADL IS BEING BETA-TESTED by numerous government agencies (including the US Geological Survey and the Library of Congress), universities (including several University of California campuses, Stanford University, and the University of Colorado), and corporations (including Sun Microsystems and Digital Equipment Corp.).

Currently, users must have passwords to access ADL. However, we plan to "go public" in July 1996 by eliminating the password requirement. By that time, we expect to have a sufficiently large data collection and sufficiently powered servers to make ADL useful and accessible.

Government agencies, schools, corporations, and even individuals trying to find, for example, elevation data for their back yards will find ADL helpful. Users will be able to look up the information they need and, if necessary, download the data. Meanwhile, we plan to regularly update and expand our collection with data from throughout the world. In this way, ADL will become increasingly useful. ∎

Acknowledgments

The Alexandria Project is a consortium of universities, public institutions, and private corporations headed by UCSB and supported by NSF, ARPA, and NASA under cooperative agreement NSF IRI94-11330.

Alexandria Project members who contributed directly to the writing of this article and to the research work involved are D. Andresen, L. Carver, R. Dolin, C. Fischer, J. Frew, M. Goodchild, O. Ibarra, R. Kemp, R. Kothuri, M. Larsgaard, B. Manjunath, D. Nebert, J. Simpson, A. Wells, T. Yang, and Q. Zheng.

References

1. C. Fischer et al., "Alexandria Digital Library: Rapid Prototype and Metadata Schema," *Proc. Forum on Research and Technology Advances in Digital Libraries*, Springer-Verlag, Secaucus, N.J., 1995.
2. Library of Congress MARC Development Office, *Maps: A MARC Format*, Library of Congress Information Systems Office, Washington, D.C., 1976.
3. Federal Geographic Data Committee Newsletter, No. 1, Spring 1991, Federal Geographic Data Committee, Reston, Va.
4. US Federal Geospatial Data Committee, *Content Standards for Digital Geospatial Metadata*, US Geological Survey, Reston, Va., 1994.
5. R. Kothuri and A.K. Singh, "Indexing Hierarchical Data," Tech. Report TR95-14, Computer Science Dept., Univ. of California, Santa Barbara, 1995.
6. Environmental Systems Research Inst., *ArcView 2.0c Software, Alpha/OSF1 Version*, Environmental Systems Research Inst., Redlands, Calif., 1978.
7. M. Vitterli and C. Herley, "Wavelets and Filter Banks: Theory and Design," *IEEE Trans. Signal Processing*, Vol. 40, No. 9, Sept. 1992, pp. 2,207-2,232.
8. B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," Tech. Report CIPR-TR-95-06, Electrical and Computer Eng. Dept., Univ. of California, Santa Barbara, 1995.
9. D. Andresen et al., "SWEB: Towards a Scalable World Wide Web Server on Multicomputers," *Proc. 10th Int'l. Parallel Processing Symp.*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07255, 1996.
10. D. Andresen et al., "Scalability Issues for High-Performance Digital Libraries on the World Wide Web," *Proc. Forum on Research and Technology Advances in Digital Libraries*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07402, 1996.

Terence R. Smith *is a professor of geography and computer science at UCSB and the Alexandria Digital Library Project's director. At UCSB, he was the Department of Computer Science's chair from 1986 to 1990 and the National Center for Geographic Information and Analysis' associate director from 1988 to 1990. His research interests include the design, construction, and use of digital libraries; the provision of computational support for modeling science and engineering activities; and the development of river system evolution theories. He received an undergraduate degree in geography in 1965 from Cambridge University and a PhD in environmental engineering in 1971 from Johns Hopkins University. He has published over 90 research articles in various disciplines, including geography and computer science.*

Readers can contact Smith at the Department of Computer Science or the Department of Geography, University of California, Santa Barbara, CA 93106; e-mail smithtr@cs.ucsb.edu. Readers can obtain more information from the Alexandria Project's Web site at http://alexandria.sdc.ucsb.edu.