THE ALEXANDRIA DIGITAL LIBRARY PROJECT: DISTRIBUTED LIBRARY SERVICES FOR SPATIALLY REFERENCED DATA

Barbara P. Buttenfield
Department of Geography
University of Colorado, Boulder, CO 80309
Email: babs@colorado.edu
and
Michael F. Goodchild
Department of Geography
University of California, Santa Barbara, CA 93106-4060
Email: good@ncgia.ucsb.edu.

ABSTRACT

This paper provides an update on the Alexandria Digital Library project (ADL), centered at University of California-Santa Barbara with a satellite site at University of Colorado-Boulder. We discuss the components of the publicly accessible Web implementation. We describe the origins of ADL, its objectives of providing access to the services of a map and imagery library over the Internet, and of merging maps and images into the library information mainstream. We describe the development of the ADL prototypes, and focus on the features of the current implementation that distinguish ADL from other digital library efforts. The paper ends with an overview of outstanding research issues raised by ADL and other related projects, and of the impact such developments are likely to have on the accessibility of spatial data.

ORIGINS OF THE ALEXANDRIA DIGITAL LIBRARY

To service those who need digital data, new products appear with increasing frequency, and one can access increasing quantities of digital data on the Internet. Federal agencies that produce and distribute datasets are converting physical distribution mechanisms to electronic form. Data enhancement is increasingly outsourced to private companies who add value to federal products, repackage and redistribute them on the Internet. Scientists who previously ordered data on magnetic tape or CD-ROM from agencies or companies can now access data products directly via the Internet. The challenge for those wishing to access electronic data sources is to navigate the ever-increasing volume of information on the Internet, to locate data appropriate to an application, and to download them. This requires a new set of skills for the scientist and also requires provision of new tools for generalized and specialized data delivery. A major challenge for the coming decade is to enhance the accessibility to all types of digital data, including but not limited to geographically referenced environmental data.

These issues challenge many branches of science, commerce and technology. Organization of and access to digital data via the Internet as been identified as a "National Challenge" in the Information Infrastructure Technology Applications component of the U.S. High Performance Computing and Communications Program (HPCC). National Challenges are fundamental applications that have broad and direct impact on the Nation's competitiveness and the well-being of its citizens, and that can benefit from the application of HPCC technology and resources. (NSF, 1996a; Tosta, 1994) In the Fall of 1993, a solicitation for proposals responding to a National Research Initiative on Digital Libraries was issued with joint sponsorship from the National Science Foundation (NSF), the National Atmospheric and Space Administration (NASA), and the Advanced Research Projects Agency (ARPA). "One goal of this [Digital Libraries] Initiative is to establish better linkages between fundamental science and technology development upon which key aspects of the National Information Infrastructure depends. ... The projects' focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks — all in user-friendly ways." (NSF 1996b).

Six awards for a period of four years were issued from a pool of seventy-two submissions:

- Carnegie Mellon University: "Informedia: Integrated Speech, Image and Language Understanding for Creation and Exploration of Digital Video Libraries"
- University of California, Berkeley,: "The Environmental Electronic Library: A Prototype of a Scalable, Intelligent, Distributed Electronic Library"
- University of California, Santa Barbara: "The Alexandria Project: Towards a Distributed Digital Library with Comprehensive Services for Images and Spatially Referenced Information"
- University of Illinois: "Building the Interspace: Digital Library Infrastructure for aUniversity Engineering Community"
- University of Michigan: "The University of Michigan Digital Libraries Research Project"
- Stanford University: "The Stanford Integrated Digital Library Project"

Home pages for all these projects may be found through NSF (1996a).

"We see these projects as taking the next step — and a very large one — in our ability to make available vast stores of knowledge and innovative information services based on High Performance Computing and Communications technologies to researchers, students, educators and the general public." Young, 1994). Each of the six awards has focused on a unique library issue, ranging from digital video made available to public schools (Carnegie-Melon), to digital versions of science and engineering journals made available to a university campus (Illinois). One project (Stanford) has undertaken to deliver high performance digital linkages between the other five. Three of the six awards focus on environmental data. The Alexandria Project is one of these.

The Alexandria Project goal is to deliver comprehensive library services for *distributed data* archives of *geographically referenced* information stored on map sheets and series, photographic and satellite images, atlases, and other geographic media. "Distributed data" means the library's components may be spread across the Internet, as well as coexisting on a single desktop. "Geographically referenced" means that items are associated with one or more regions (``footprints") on the surface of the Earth. Geographically referenced information has been traditionally treated as a 'separate' problem by librarians and data archivists, due to complexities of spatial ordering, layering, and spatial and temporal autocorrelation. Our intention is to eliminate the traditional distinctions made in libraries between general collections of books and text with special collections such as maps and photos. The project includes assessment of user needs, basic research to address technical impediments, software development, and a rigorous program of evaluation and quality control. These requirements are met through a collaboration between disciplines at several institutions, the private sector, and segments of the geospatial data user community, including earth scientists, librarians and information archivists, and students and teachers in K-12 classrooms.

IMPLEMENTING ALEXANDRIA ON THE WEB

The centerpiece of the Alexandria Project is the Alexandria Digital Library (ADL), an online information system inspired by the Map and Imagery Laboratory in the Davidson Library at the University of California, Santa Barbara. The ADL currently provides access over the World Wide Web to a subset of the Map and Imagery Library holdings, with other geographic datasets coming online on a continuing basis. The ADL is publicly available on the World Wide Web, and includes tutorials, general reference information about spatial data and digital spatial data sources, and functions for browsing and retrieving actual maps, images and data. These functions currently include a catalog, a gazetteer, and a geographic browser that also displays geographic footprints of data sets that one can retrieve, view metadata, or download.

Users can browse ADL holdings electronically and search by spatial or temporal location or by metadata content. Spatial searches by placename or by spatial footprint can be refined according to specific time periods, data resolution, data category (satellite image, topographic map, geologic map, etc.) Efforts are underway to implement browsing tools based on collections maintenance criteria (map sheets having multiple editions, e.g..) or based on information content (to initiate a search for a Spot image containing a hydroelectric dam, e.g.).

Multiple versions of the Library are currently under development. Reasons for this include the need to have a stable Library system in place to support user evaluation studies, and the need to demonstrate that individual software modules are operational before they are added to the general testbed, while still providing a system design platform for experimentation and benchmarking. One advantage of the multiple version approach is that the system design can be 'frozen' at specific stages of development. These phases mark a chronology of design, planning, and results of user evaluation studies.

The first phase produced a rapid prototype running commercial off-the-shelf software (ArcView) on a UNIX platform. This version was based on a multi-window environment that is common to anyone who has worked with GIS software packages. The rapid prototype was completed in Spring, 1995, and served as an early platform for user interface evaluation efforts. A subset of the rapid prototype was ported to a Windows platform, and burned onto CD-ROM. Twenty-five hundred copies of the CD-ROM were distributed along with a questionnaire to solicit community feedback. Unfortunately, Microsoft released its Windows-95 operating system almost concurrently with distribution of the CD-ROM, and incompatibility has limited the potential high volume of feedback we had hoped for. However, the CD-ROM version served to make the Alexandria Project visible in many working environments where UNIX is not available, and it continues to be used in selected schools and libraries in North America.

The current phase of system design extends rapid prototype functions in a World-Wide Web environment. System architecture includes a storage component, a catalog component, an ingest component, and an interface component. The storage component is designed to accommodate very large collections of very large digital objects. Environmental data is alternatively characterized by high resolution, multispectral raster data, and overlaid themes of vector data compiled at multiple map scales. Storage requirements are large. For example, an analog air photograph scanned at 600 dots-per-inch commonly requires 30 MB (90 MB for color) per archived image (Andreson et al, 1996). A single collection of historical photography containing hundreds or thousands of images could require storage on the order of single terabytes at the point of archival. Distributed storage provides the only feasible architecture for multiple datasets, and Internet protocols (e.g.., Z39.50) are being implemented to handle delivery and transfers. Current system holdings focus on the southern California region.

The catalog component is a special emphasis for current system development efforts (Smith, 1996). The catalog systematizes all types of information by which the Library holdings may be organized. By implication, the catalog contents form the basis for user browsing. An archive may be searched only on items which are organized in its catalog. (One reason the Web is difficult to navigate is that it lacks a catalog.) The Alexandria catalog allows browsing by placename, by data theme, by spatial footprint, by date of compilation, or by metadata as defined by FGDC/USMARC standards. Placenames are provided by the Geographic Names Information System (GNIS) gazetteer, which includes 1.8M names of US features/15 classes, and by the Board of Geographic Names (BGN) gazetteer , including 4.5M names of land/undersea features. The catalog is stored in a central relational database (Sybase) housed in Santa Barbara. Metadata records are stored in a single centralized archive using Microsoft Access.

The ingest component currently provides for input of data, metadata, and catalog information. Data ingest can be accomplished in a number of ways: by scanning analog material, by transfer of created metadata records from Microsoft Access, or transfer from other sources (e.g.., frame-level records from air photo databases, sheet-level records for indexed map series , and USMARC catalogued records for single maps). Following data ingest, new data items are 'added' to Alexandria by creating new catalog and metadata records with pointers (presently in the form of URLs) to the ingested files. When the metadata and catalog records are placed online, the data files become available.

The interface component is most visible to users. To some, there may appear to be no difference between the interface and the Library. Interface functions include tools for indexing, retrieval, and data browsing, tools to formulate queries by location, time, metadata, and content. Interface utilities to guide image fusion, compression, and filtering are under construction that will speed data delivery and facilitate exploration of distributed archives.

RESEARCH ISSUES

The Web testbed presents major challenges for system designers. Models guide the design of some but not all components of a digital library, and issues related to the nature of geographically referenced data complicate the situation. For some components (the catalog, the digital objects), models that exist are only partially sufficient. For other components (particularly the gazetteer and map browser) important gaps in use protocols must be remedied to enable fully operational library functions on geospatial data.

Models for data catalogs include the U.S. MARC (Machine-Readable Cataloging) record system in use by the Library of Congress. However, the MARC record system does not include standard protocols for geographic referencing. Protocols for exchanging data and cataloging metadata have been established in the Spatial Data Transfer Standard (FGDC, 1992), and in its successor, the Metadata Content Standard (FGDC, 1995) but these remain simultaneously incomplete and cumbersome to apply (Goodchild, 1995a). Extending the classes of catalog queries supportable by ADL to incorporate content-based searching will require significant extensions to the current cataloging model. Extending the metadata in the catalog component and the associated search procedures to support the extended classes of queries provokes a similar challenge (Goodchild, 1995b).

The storage component of ADL contains the collection of digital objects. For the purposes of an operable digital library, a digital object must include the binary representation of the information of interest (the "data"), procedures for interpreting/retrieving the data, and a universal object identifier (called an *oid*). An important issue in distributed Internet applications is that there is currently no accepted standard for *oid's*. There are a number of alternative suggestions relating, for example, to URx's of various forms (where the ``x" is an identifier, locator, or name) (Andreson et al, 1996).

Another problem is that the digital objects in ADL are typically very large. For satellite images, a size of 150 MB is not uncommon and may exceed 2 GB. Spatial image collections are also large. The UCSB Map and Imagery Library has a collection of over two million air photos in analog form which, when scanned at 600 dpi, each require between 25MB and 100MB of storage space, resulting in a collection whose size exceeds 100 TB. These are preliminary reasons why collections of such items must inevitably be distributed. Eventually, as the library contents increase, the catalog itself should be distributed as well. Considerations for distributing a catalog include the difficulties faced by users in finding an appropriate item; the cost of examining or downloading large items over bandwidth-limited channels; and the provision of access to distributed sets of storage locations (Smith et al, 1996).

The Web environment lacks protocols for map browsing. In particular, the ADL must operate within three significant limitations. First, the standard Web page construction language (hypertext mark-up language, or HTML) lacks mechanisms for presenting vector data or entering spatially-indexed information. This is apparent when attempting to define a spatial search region on a browse map. For example, no Web browsers currently available provide drag-and-click "lasso-ing" functions. Such actions are not currently supported by Web browsers, which immediately send an HTTP request after a single mouse click. Second, HTTP is a stateless protocol, designed for small, short transactions. By default, after a user completes an HTTP request, neither the client nor the server maintains any state or "memory" of the transaction. Each request appears to the server to be completely new. This statelessness prevents library-significant activities such as setting user-defined environment parameters, retaining a query history, or performing iteratively refined searches. Finally, no Web browser that we know of supports vector data display. This is a serious issue, since a significant and important portion of spatially-indexed information collections involve items represented in vector format (Andreson et al 1996).

Lastly, we lack a model of a digital library user. Models of traditional (physical) library users do not translate directly to a digital library environment, which is categorically and progressively more than a physical library in electronic form. One can manipulate information in a digital library, process and re-process it, even re-ingest the newly formed information that may result from such processing. Many aspects of digital library use have never occurred to potential users, and this makes it difficult to articulate information needs and requirements. Metadata needs and requirements are similarly difficult to identify (Bretherton and Singley, 1994). One can build up a profile of ADL users over time, through transaction logging, videotaping, focus groups, and semi-structured interviewing, and these types of data are being collected and are described elsewhere (Buttenfield, 1995; Buttenfield and Kumler, 1996). The challenge is that as ADL changes in appearance and in functionality, user evaluation becomes a process of aiming at a moving target. This is not to say that assessment of user needs is impossible, nor that user evaluation cannot be accomplished, only that the customary paradigms are not completely adequate.

SUMMARY

The goal of the Alexandria Project is to develop a user-friendly digital library system that provides a comprehensive range of services to collections of maps, images, and spatially-referenced information. This paper reports on a project to design, develop and test a distributed, high-performance digital library, available on the Internet, in which collections of spatially-indexed information in digital form is dispersed geographically. The program of research and development represents a major step towards the evolution of a distributed digital library supporting both textual and geographically referenced sources of information. We intend to create library services that are scalable to the national level. While various technical issues relating to the storage and content-based access and retrieval of spatial data remain, our long-term goal is to remove the mainstream library distinction between text and special materials such as maps and images.

Readers interested in visiting the Alexandria Digital Library or reading more about the Project may start at the following Web page: http://alexandria.sdc.ucsb.edu.

ACKNOWLEDGMENTS

This paper forms a portion of the Alexandria Digital Library Project, jointly sponsored by NSF, NASA, and ARPA. Funding by the National Science Foundation (NSF IRI-94-11330) is greatly appreciated. Matching funds from the University of California and the University of Colorado are also acknowledged.

REFERENCES

Andresen, D., Carver, L., Dolin, R., Fischer, C., Frew, J., Goodchild, M., Ibarra, O., Kothuri, R., Larsgaard, M., Manjunath, B., Nebert, D., Simpson, J. Smith, T., Yang, T., .Zheng, Q. 1995 The WWW Prototype of the Alexandria Digital Library, **Proceedings** of the International Symposium on Digital Libraries, Tsukuba, Japan, 1995. On the World Wide Web at http://alexandria.sdc.ucsb.edu/public-documents/papers/japan-paper/

Bretherton, F. and Singley 1994 Metadata: A User's View. **Proceedings**, 7th SSDM, Charlottesville, VA: 166-174.

Buttenfield, B.P. 1995 User Evaluation for the Alexandria Digital Library Project. **Proceedings**, Workshop "How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum." Allerton, Illinois: National Science Foundation and the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. On the World Wide Web at http://edfu.lis.uiuc.edu/allerton/95/s2/buttenfield.html

Buttenfield, B.P. and Kumler, M.K. 1996 Tools for Browsing Environmental Data: The Alexandria Digital Library Interface. **Proceedings** Third International Conference on Integrating Geographic Information Systems and Environmental Modeling". Santa Fe, New Mexico, January 21-25, 1996 (no page numbers).

FGDC 1992 Federal Geographic Data Committee. **Spatial Data Transfer Standard.** Washington DC: Federal Information Processing Standard 173.

FGDC 1995 Federal Geographic Data Committee. **Content Standards for Digital Geospatial Metadata.**

Goodchild, M. F. 1995a Sharing imperfect data. In H. J. Onsrud and G. Rushton (Eds), **Sharing Geographic Information**. New Brunswick, NJ: Rutgers University Press: 413-425.

Goodchild, M.F. 1995b **Alexandria Digital Library :Report on a Workshop on Metadata**, held in Santa Barbara, California, November 8, 1995. On the World Wide Web at http://alexandria.sdc.ucsb.edu/public-documents/metadata/metadata_ws.html.

NSF 1996a National Science Foundation **NSF Digital Libraries Awards Announcement.** On the World Wide Web at http://www.cise.nsf.gov/iris/DLAnnounce.html.

NSF 1996b National Science Foundation **NSF/ARPA/NASA Digital Libraries Initiative Home Page**. On the World Wide Web at http://www.cise.nsf.gov/iris/DLHome.html.

Smith, T.R. 1996 **The Meta-Information Environment of Digital Libraries.** On the World Wide Web at http://alexandria.sdc.ucsb.edu/public-documents/dlib/

Smith, T.R., Geffner, S. and Gottsegen, J 1996. **A General Framework for the Meta-Information and Catalogs in Digital Libraries**. On the World Wide Web at http://alexandria.sdc.ucsb.edu/public-documents/ieee/

Tosta, N. 1994 Continuing Evolution of the National Spatial Data Infrastructure. **Proceedings**, GIS/LIS, Phoenix Arizona: 769-777. On the World Wide Web at http://wwwsgi.ursus.maine.edu/gisweb/spatdb/gis-lis/gi94096.html

Young, P. 1994 Assistant Director of the NSF directorate for Computer and Information Science and Engineering, quoted in announcing the Digital Libraries Awards. On the World Wide Web at http://www.cise.nsf.gov/iris/DLAnnounce.html.