

# UC Berkeley Digital Environmental Library

Robert Wilensky, Principal Investigator, Computer Science

wilensky@cs.berkeley.edu

Michael Stonebraker, co-PI, Computer Science

Richard Fateman, Jitindra Malik, David Forsyth, Computer Science

Martin Vetterli, EECS

Ray Larson, Michael Buckland, Nancy Van House, Library & Info Studies

Robert Twiss, Landscape Architecture

Kenn Gardels, Environmental Design

Cliff Lynch, Office of the President; John Kunze, IS&T

Gary Kopec, Phil Chou, Les Niles, Dan Bloomberg, XEROX PARC

John Hull, Ricoh California Research

Dragutin Petkovic et al, IBM Almaden

Project URL: <http://elib.cs.berkeley.edu>

## Overview

Advances in computer and communications technology are transforming how it is possible for people to work with information. As a result, existing library services are being challenged to re-design their modes of providing service. The challenge extends to technologists as well: It is crucial that we marshal our technological prowess in a way that will provide users of digital libraries with effective services for research, education, and other important national social goals.

To meet these challenges, we have put together a team of investigators with substantial experience in research, development, operation, and evaluation of library services, and in the crucial enabling digital technologies. Our research team includes experts from academia and industry in data base management systems, networked information protocols, document recognition, information

search and retrieval, natural language processing, computer vision, communication technology, library services, and system evaluation. Many of these investigators have already had substantial experience in working together on these and related topics.

The testbed system that we are constructing will constitute an innovative prototyping for a very large and original production digital library: the CERES system, to be implemented by the State of California. This system provides widespread online public access to the environmental information that is central to all aspects of the future development. A digital library project focussed on environmental data has special appeal: It is of unusually wide-ranging scientific, political, educational, and economic interest; it involves an exceptional range of object types (texts, images, video, numeric data, software); it brings the added dimension of a geographical information system; and it draws of our group's experience in developing related information and computing support systems. In addition, to develop our prototype, we have created a consortium of contributors who are providing large collections of diverse material, and a consortium of test users who will help evaluate our system.

The technical focus of this proposal is the development of several critical technologies needed to implement our vision of electronic libraries. In this vision, large numbers of geographically distributed users can conveniently access the entire contents of very large and diverse repositories of electronic objects. These repositories will exist in locations physically near or remote from the users, and will contain objects comprising text, images, maps, sounds, full-motion videos, merchandise catalogs, and scientific and business data sets, as well as hypertextual multimedia compositions of such elements. Users will be able to browse and retrieve information from these repositories by content; both organizations and private citizens will be able to easily add repositories of their own, which will interoperate with this global system.

One way to conceptualize what we are proposing is as a next-generation "xmosaic/World Wide Web" system. Such systems provide very convenient access to distributed information resources. However, they are limited in important ways. To overcome these limitations and to realize the full vision, we are focussing on the following technological areas:

### **Providing a coherent, content-based view of a diverse distributed collection.**

Collections of different kinds will exist on many servers. However, users are largely uninterested in this level of organization, and would like to interact with this system in terms of the content of the various collections, wherever they may be.

### **Scaling**

Digital repositories will be measured in terabytes. As such, digital library architectures and techniques must scale to a very large corpora. The need for a scalable systems imposes important constraints on the overall system design, especially in terms of distributed elements of the design.

### **Data acquisition, transfer and presentation technology.**

For years to come, many collections will be assembled by scanning in corpora. The problem of constructing and analyzing these images is a severe one. In addition, access to these and other documents, especially video impose problems of transmission and display, especially to citizens not equipped with powerful computational resources, high resolution monitors, and high bandwidth network connections. Therefore, a comprehensive and highly accessible electronic library must address these analysis, communication, and presentation issues.

We are addressing these problems by focussing on the following elements:

1. More accurate data capture
2. Scalability of information retrieval systems
3. A more effective client/server information retrieval protocol
4. Text analysis for retrieval and browsing
5. Image and video analysis for retrieval and browsing
6. Georeferencing documents
7. New user interface paradigms
8. Resource discovery and distributed search
9. Compression, communication, and resolutions enhancement
10. Ongoing, iterative user needs assessment and evaluation

To test our research ideas, we are applying them to a prototype electronic library focussed on "The California Environment". We have brought together a set of key users of this environmental information who will function as our experimental user group.

In addition, we are phasing the technologies we develop into our prototype in an evolutionary manner. That is, we are beginning by constructing an initial system that only modestly pushes the existing technological envelop; we will continually integrate the results of our research into this system as they reach the appropriate stage. Hence we plan to always have a working system which we can enhance, with which we can experiment, and from which we can obtain continual feedback from our users to help guide the subsequent stages of its evolution.

Nancy Van House, Acting Dean

School of Library and Information Studies

102 South Hall #4600

University of California, Berkeley, CA 94720-4600

(510) 642-9980 fax (510) 642-5814

---

[Berkeley Digital Environmental Library Home Page](#)