

# The Digital Libraries Initiative: Update and Discussion

by Edward A. Fox  
Guest Editor

This special section of the *Bulletin of the American Society for Information Science* on the Digital Libraries Initiative begins with an article by the guest editor that provides an overview of the initiative to-date. In the two subsequent articles Michael Lesk gives perspectives on the field, while Stephen Griffin provides important data, including abstracts, of a number of recently funded digital library research projects. Drs. Lesk and Griffin are with the National Science Foundation's Information and Intelligent Systems (IIS) Division, in which Lesk serves as director (on rotation) and Griffin as program officer. The Lesk and Griffin articles are reprinted from *D-Lib Magazine*, v. 25, no. 7/8 (July/August 1999) with the permission of the authors and the Corporation for National Research Initiatives.

## Digital Libraries Initiative (DLI) Projects 1994-1999

by Edward A. Fox

Edward Fox is professor in the Department of Computer Science and Director of the Digital Research Laboratory at Virginia Tech. He directs the Networked Digital Library of Theses and Dissertations (<http://www.ndltd.org>). He also directs the Internet Technology Innovation Center at Virginia Tech (<http://fox.cs.vt.edu/itic/>). He can be reached there by mail at 660 McBryde Hall, M/C 0106, Blacksburg, VA 24061; by phone at 540/231-5113; on the Web at <http://fox.cs.vt.edu>; or by e-mail at [fox@vt.edu](mailto:fox@vt.edu)

Since 1993, the National Science Foundation (NSF) has played a lead role in an interagency federal program called the Digital Libraries Initiative (DLI). DLI emerged after several years of discussion in which a number of researchers, such as Michael Lesk (then at Bellcore), made recommendations through the reports of a series of NSF-sponsored planning workshops (see summary at <http://fox.cs.vt.edu/DLSB.html>). Thus, throughout the 1990s NSF support has been a critical factor in establishing the digital libraries field as an important area for research, development, application and practice. Though total investment around the globe – involving such institutions as libraries, universities, associations, corporations, foundations and other governments – amounts to hundreds of millions of dollars, the single most visible effort is the DLI program, which is the focus of all of the articles in this special section.

### DLI Funding

In the United States, over \$68 million in federal research awards were made through DLI over the period 1994-1999. \$24 million was awarded in 1994 by NSF, DARPA and NASA, split evenly among six "DLI-1 teams." Three were in California: two went to campuses of the University of

California (one to Berkeley and one to Santa Barbara) and the third to Stanford University. Two were in the middle of the country, to the University of Illinois at Urbana-Champaign (UIUC) and the University of Michigan. Carnegie-Mellon University (CMU) received the only East Coast award, leveraging prior work on text, image and speech processing.

Roughly \$44 million, allocated in somewhat different fashion, has already been awarded by NSF, DARPA, National Library of Medicine, Library of Congress, National Endowment for the Humanities, NASA and the FBI (in partnership with National Archives and Records Administration, Smithsonian Institution and Institute of Museum and Library Sciences) in a second phase, the "DLI-2" program (<http://www.dli2.nsf.gov>). A terse summary of these awards is shown in Table 1. Recent commitments to the three California groups in DLI-1, including sub-awards involving other partners in California (University of California, Irvine; University of California, Los Angeles; University of California, San Diego; California Digital Library) and at the University of Georgia, plus an undergraduate education award to Berkeley, account for over \$15 million. CMU also received \$4 million further support, as

well as a separate but related \$450,000 grant. The six other large grants (each for \$1 million or more) went to Columbia University, Cornell University, Harvard University, Michigan State University, Tufts University and the University of South Carolina.

Over \$500,000 was allocated to three awards from 1988 with an undergraduate emphasis (see top section of Table 1). There were six awards focused on international collaboration (see bottom section of Table 1), for a total of about \$2.3

million. The main DLI-2 program (see middle section of Table 1) involved over \$41 million through 21 awards. Of these 21, 10 were large, accounting for over \$35 million, while the remaining 11 account for about \$5.5 million. Please see the accompanying article by Stephen Griffin that provides short summaries of DLI-2 projects announced through August 1999. Other details and newer information can be found at the DLI-2 Web site or set in a broader context as part of the self-study course materials on digital libraries at Virginia Tech (see specifically <http://ei.cs.vt.edu/~dlib/projects.htm>).

**Table 1. Details of DLI-2 Awards by September 1999**

AWARD ID	PI NAME	INSTITUTION	Mos.	\$K
<b>DLI-2 Undergraduate Emphasis</b>				
9817406	Agogino, Alice	UC-Berkeley	12	200
9816026	Maly, Kurt	Old Dominion Univ.	12	80
9816644	Kappelman, John	UT-Austin	24	287
<b>Subtotal</b>				<b>567</b>
<b>DLI-2</b>				
9817485	Kornbluh, Mark	Michigan State	60	3,600
9817484	Crane, Gregory	Tufts	60	2,758
9817434	McKeown, Kathleen	Columbia University	60	5,002
9817496	Wactlar, Howard D.	CMU	48	4,000
9817432	Smith, Terrence	UC-Santa Barbara	60	5,800
9817799	Garcia-Molina, Hector	Stanford University	60	4,300
9817353	Wilensky, Robert	UC-Berkeley	60	5,000
9874747	Verba, Sidney	Harvard University	36	1,800
9817416	Lagoze, Carl	Cornell University	48	2,268
9874759	Etzioni, Oren	Univ. of Washington	36	598
9817492	Gorman, Paul	Oregon Health Sciences	36	650
9817511	Weiderhold, Gio	Stanford University	36	520
9817430	Choudhury, Sayeed	Johns Hopkins	36	530
9874771	Armistead, Samuel G.	UC-Davis	36	497
9817483	Seales, W. Brent	Univ. of Kentucky	36	500
9817444	Buneman, Peter	Univ. of Pennsylvania	36	505
9874781	Rowe, Timothy	UT-Austin	36	500
9817527	Myers, Brad	CMU	36	450
9817473	Chen, HC	Univ. of Arizona	36	501
9817572	Palakal, M.	Indiana Univ.	36	316
9817518	Willer, D.	Univ. of South Carolina	48	1,199
<b>Subtotal</b>				<b>41,294</b>
<b>DL International</b>				
9975164	Larson, Ray	UC-Berkeley	36	305
9905842	Byrd, Donald	Univ. of Mass	36	494
9905935	Hedstrom, Margaret	Univ. of Michigan	36	488
9906025	Calcari, Susan	UW-Madison	36	480
9907892	Lagoze, Carl	Cornell Univ./ePrint	36	292
9905955	Lagoze, Carl	Cornell Univ./ILRT	36	240
<b>Subtotal</b>				<b>2,299</b>
<b>Grand Total</b>				<b>44,160</b>

## Research Coverage of DLI

DLI-1 focused on research, and the six projects were led by individuals with strong backgrounds in technical fields, largely computer and information sciences. An inspection of the available information shows that DLI-2 has greatly expanded the support of different disciplines working in the digital libraries field. Table 2 lists in alphabetical order many of the home departments of investigators funded through DLI-2.

Another illustration of the breadth of coverage in DLI-2 can be seen in Table 3, which deals with the types of content, media or formats being studied. To aid the reader interested in particular topics, universities focusing on them also are listed.

Even with respect to technologies considered, DLI-2 is considerably broader than DLI-1. Table 4 summarizes the technical areas studied along with universities involved in each. The reader is invited to make up a list independently of areas closely related to digital libraries and compare that list with the one given. Alternatively, one might look at lists in other introductions to the field, like that in the April 1995 special section of *Communications of the ACM*. There are areas likely to be on many people's lists that were not much of a focus in DLI-2, such as abstracting, browsing, ethnography, hypertext, indexing, interaction, sociology, storage and virtual reality.

Furthermore, though there are some projects dealing with key issues of information retrieval (IR) (e.g., the Berkeley international effort) or human-computer interaction (HCI) (e.g., the CMU separate project on video editing), these topics seem to play a relatively minor role in the overall initiative. But extensive experimentation in these areas is necessary for the field to mature. Such work on IR and HCI will require readily available test-beds, usability tests involving large numbers

**Table 2. Discipline Coverage of DLI-2  
(selected home departments of investigators)**

Anthropology	Biomedical Information	Classics
Computer Science	Economics	English
Fine Arts	Geography	Geological Sciences
Government	Electrical Engineering	Environmental Science
History	Information Management	Information Studies
Language Technology	Library & Information Science	Linguistics
Management Info. Systems	Medical Informatics	Political Science
Psychology	Religious Studies	Robotics
Sociology	Spanish	Teacher Education

of users, careful comparative experiments and other related studies.

Following along these lines, and possibly of particular interest to ASIS members, is consideration of the ties to information science that are visible in DLI-2. Geographical information and medical informatics are the focus of several efforts. Christine Borgman of UCLA's Graduate School of Education and Information Studies is a co-principal investigator playing a role in the University of California, Santa Barbara project, while Javed Mostafa of the School of Library and Information Science at Indiana is a co-principal investigator in their project. Librarians are co-investigators on several projects. In the international program, two of the projects are run from schools of information (i.e., at Berkeley, Michigan). But overall, few funded DLI-2 projects are run out of library or information science departments or schools. In general most project direction is by computer rather than information scientists.

### Continuing DLI-2

It is clear from the funding for DLI-2 that reviewers and agencies involved largely felt that DLI-1 activities should be continued. While UIUC was not supported, its key partner in DLI-1, University of Arizona, is supported in DLI-2, continuing in particular the work on automatic classification, aiming to consolidate results by scaling up and comparing algorithms. Though the University of Michigan did not receive a follow-on award per se, Margaret Hedstrom in their School of Information is leading a project funded at almost \$500,000 on the topic of preservation (using emulation). Further, work on agents that is rather similar to that at University of Michigan (but somewhat more focused) is being supported at Indiana, Bloomington (for personalized information filtering) and at Washington (to aid retrieval from the WWW). One successful supplement to the project at Michigan was the Joint NSF-European Union (EU) Working Groups on Future Directions of Digital Libraries Research (<http://www.dli2.nsf.gov/workgroups.html>) that stimulated extensive international discussion. Also, the JSTOR effort

(<http://www.jstor.org/>) launched at Michigan has become a serious commercial venture involving digitization of important old journals.

All of the other DLI-1 projects are continuing earlier work with a relatively high level of funding. Consolidation is in evidence too, with coordination of the three California efforts. All three will develop testbeds and foster interoperability, a strong point of the prior work at Stanford. Each will carry out evaluations. All three have efforts on user interfaces, regarding presentation, and on analysis of collection data. In addition, the San Diego Supercomputer Center will act as collection clearinghouse and the California Digital Library will facilitate statewide collaborative knowledge creation and dissemination.

The Santa Barbara effort is focused on building the Alexandria Digital Earth Prototype as a digital earth modeling system made up of Information Landscapes. That effort extends prior work through a broader vision, with many goals for further technical development and with user testing involving UCLA and other partners.

The Berkeley proposal discusses a very large number of

**Table 3. Types of Content and DLI-2 Sites Where They Are Studied**

Types	Universities
Bibliographic Records	Arizona
Engineering Education	UC-Berkeley
EPrints	Cornell (intl ePrint)
Folk Literature	UC-Davis
Geo-referenced Info.	UC-Santa Barbara
Health Care	Oregon Health Sciences
Humanities	Tufts; Kentucky
Library Reference	Washington
Medical Images	Stanford
Mixtures of Media	UC-Berkeley (intl); Cornell (intl ILRT)
Patient Records	Columbia
Sheet Music	Johns Hopkins; UM-Amherst (intl)
Skeletons	UT-Austin
Simulations	South Carolina
Social Science Data	Harvard
Speech	Michigan State
Video	Carnegie Mellon
Web	Arizona; Pennsylvania; Washington
X-ray CT Scans	UT-Austin

research topics around the theme “Re-inventing Scholarly Information, Dissemination and Use.” But the proposal body does not appear to connect this motivating theme to the lively self-publishing efforts expanding around the globe (e.g., e-prints, reports, dissertations, courseware, biomedical

information). Rather, in the tradition of Berkeley UNIX they propose to build general tools to help digital library users do more on their own and also to study models and conduct user studies on dissemination and use.

The Stanford proposal adopts a different approach, emphasizing a comprehensive problem analysis of four barriers to effective digital libraries. One barrier is that contents and systems are highly diverse and heterogeneous. The other barriers are needs for which no solution now exists: filtering mechanisms, portable interfaces and an economic infrastructure that guarantees privacy. Like at Berkeley, the Stanford team will develop software. It will be for value filtering, for portable devices, for extending their earlier InfoBus into the InterServ suite of models and protocols and for economic modeling.

A smaller project at Berkeley (run in connection with the engineering education coalition, NEEDS) is part of the DLI-2 undergraduate emphasis (<http://www.dli2.nsf.gov/addendum.html>), leading toward a national digital library for Science, Mathematics, Engineering and Technology Education (SMETE-lib). Expansion of this effort in upcoming years is likely to go beyond planning and pilot grants to large-scale efforts. Thus it is important that there be closer coordination with other DLI efforts than has occurred to-date.

### Outside Activities

As Michael Lesk indicates in the following article, a great deal of work on digital libraries has proceeded quite independently from DLI. For example, OCLC, the Online Computer Library Center in Dublin Ohio (<http://www.oclc.org>), has led the way on the Dublin Core ([http://purl.org/metadata/dublin\\_core](http://purl.org/metadata/dublin_core)) workshop series, the most important metadata standards activity for the field (though there are others emerging from IMS and IEEE, focused on education). OCLC also has helped run W3C-sponsored work on the Resource Description Framework (RDF) and coordinates CORC (Cooperative Online Resource Catalog), the worldwide cooperative library venture to catalog the WWW, that benefits from a variety of tools developed at OCLC. Another important tool from OCLC is the SiteSearch retrieval system (essentially the same as that used for FirstSearch), recently converted to Java. On the production side of things, OCLC owns one subsidiary (Forest Press) responsible for work on the Dewey Decimal Classification and so is exploring its use in digital libraries and knowledge management. Another OCLC subsidiary handles preservation and digitization; internally there is support as well for electronic journals and their permanent availability.

Commercially, there are many digital library efforts. IBM sells a shrink-wrapped software system called Digital Library. In Japan, several companies involved in library automation sell and adapt digital library software to leading universities. Internationally, thanks to significant funding and other support, digital libraries are under development in many countries, especially in Europe and Asia (see April

**Table 4. Technical Areas and DLI-2 Sites Where They Are Studied**

Types	Universities
3-D Modeling	UC-Santa Barbara; UT-Austin
Access Control	UC-Berkeley
Agents	Indiana-Bloomington; Washington
Archiving/Preservation	South Carolina; Univ. of Michigan (intl)
Audio Retrieval	Johns Hopkins; Michigan State; UM-Amherst (intl)
Classification, Clustering	Arizona
Data (Access) Services	Harvard
Digital Video	CMU
Economic Models	UC-Berkeley; Stanford
Electronic Notebooks	UC-Berkeley
Federation	UC-Berkeley (intl); Cornell; UW-Madison (intl)
Geographic Info. Systems	UC-Santa Barbara
Images	UC-Berkeley; UC-Santa Barbara; Kentucky; Stanford; UT-Austin
Information Filtering	Indiana; Stanford
Information Visualization	CMU
Learning Contexts	UC-Santa Barbara
Linking	Cornell (intl – ePrint)
Log (Trace) Analysis	Oregon Health Sciences
Mobile Computing	Stanford
Multimedia Fusion	CMU; Columbia
Natural Language Processing	Columbia
OCR	UC-Berkeley; Johns Hopkins
Parallel Processing	Arizona
Protocols	Stanford
Personalization	Columbia
Provenance	Penn.
Restoring Manuscripts	Kentucky
Speech Processing	UC-Davis; Michigan State
Summarization	CMU; Columbia
Text Analysis	Tufts
Video Editing	CMU



1998 special section of *Communications of the ACM*). ACM has run international conferences for the field since 1996. Other conferences and workshops have occurred or are planned in Australia, Croatia, France, Germany, Hong Kong, India, Japan, Portugal, Singapore, Taiwan, United Kingdom, etc. Many include reports on or are closely connected with DLI (see <http://www.dli2.nsf.gov/workshops.html>). Two workshops have focused on international cooperation for the field of digital libraries (see <http://www.ks.com/idla/>).

Two of the many other related efforts are especially notable. One is the ongoing series of TREC (Text REtrieval Conference) meetings and competitions. Covering information retrieval and filtering, this National Institute of Standards and Technology (NIST) effort has expanded to handle multiple languages, to deal with interactive sessions and to start to cover media beyond text. The other is the D-Lib activity (<http://www.dlib.org>). Most visible in that category is *D-Lib Magazine*, but also important are the working groups. One has dealt with the Networked Computer Science Technical Reference Library (NCSTRL, <http://www.ncstrl.org>). Another has dealt with metrics. It is likely that others will emerge.

### Assessment and Conclusion

With work on DLI since 1994, and a new round of funding allowing a broad range of projects to proceed, it seems timely to assess the progress and promise of the Digital Libraries Initiative. That is a difficult task, requiring a book or books, since there have been many hundreds of publications that should be covered (<http://www.dli2.nsf.gov/publications.html>). Furthermore, it is difficult to gauge how many related studies were motivated by DLI efforts or simply parallel the DLI efforts. The comments below reflect this larger scene and provide one person's viewpoint of overall progress.

First, we see ongoing progress and adoption of the work in the information retrieval field. TREC has shown that methods studied before the 1990s scale up to larger collections. A number of projects have demonstrated success with broadening to diverse languages and media forms. While more work is needed, there has been quite a lot done already on image retrieval, and a growing effort on retrieval from speech, music and video. It is time for controlled experimentation and comparative studies, as well as trials with wavelets and other technologies. Much work is needed regarding information visualization, which is really just in its infancy. In that case, as well as in clustering and classification, we are only in the early days of applying the storage and processing capabilities now readily available – to make a significant difference for common users.

Second, we see widespread acceptance of the broadening of the digital libraries field; not only libraries but also museums and archives are within scope. Data collections, Web pages, educational materials, experimental data, simulations and the whole province of electronic publishing are also

under consideration. Collection development is proceeding in novel ways, whether from digitization, the work of dedicated curators, feeds from publishers, user annotations, traces of expert users or through self-archiving. Users are not only scholars and researchers, but also teachers and students, as well as special groups devoted to particularly interesting collections. We are just beginning to see some commercialization, aided by the realization that digital libraries are the high-end of information systems.

Third, we see a coupling of this initiative with attempts to organize the WWW. There will continue to be interplay between work on digital libraries and efforts such as those involving OCLC, in particular Dublin Core, RDF and CORC, that will have Web-wide impact. There will be related advances in searching using various languages as well as media forms. More and more objects will have metadata associated, and (semi) automatic systems will aid in cataloging as well as browsing. As large numbers of collections emerge and are called “digital libraries,” advances will occur to search many together, leading to a second-generation federated search system that allows users to “slice and dice” whatever is available into any convenient organization desired.

Fourth, there is support of undergraduate education that depends on collections or repositories of curriculum or courseware resources. For example, NSF's Division of Undergraduate Education (DUE) funded 16 projects during the period 1993-1998 on collection building within specific disciplines/curricula. One is the Computer Science Teaching Center, available at <http://www.cstc.org>. A number of 1999 awards related to digital libraries are expected to be funded by DUE, in several cases also supported by other parts of NSF.

That leads to the final key point, regarding users. Personalization will indeed become feasible, starting with pilot efforts but ultimately becoming more common. Tailored systems are being built at the level of the organizational library (e.g., through virtual libraries devised by staff for a university community, or with technologies like SFX from University of Ghent and the Los Alamos National Laboratory – see <http://lib-www.lanl.gov/~hvds/sfx/htmls/sfxhome.html>). Content will be aggregated in various ways, community ratings will be considered and user actions will be analyzed, either by client or agent software. Though DLI efforts in this regard continue to operate at the exploratory level and are not a major focus in the initiative, several projects espouse personalization and enough others are working in this area that we can expect some real progress within a few years.

In conclusion, readers are urged to study the next two articles about DLI and to be in touch with the staff of projects that are of interest. The information science field needs closer ties with the important emerging area of “digital libraries” in which the next generation of high-end information systems is gestating and in which a large number of well-supported interesting collections are developing.