

UPDATE ON THE NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS (NDLTD)

Edward A. Fox

INTRODUCTION

The Networked Digital Library of Theses and Dissertations (NDLTD) (<http://www.ndltd.org>) aims to ensure that all future researchers and scholars have understanding and skills in electronic publishing and have used and submitted a work of their own to a digital library. If all graduate students who are interested in research prepare an electronic thesis or dissertation (ETD) and upload it to NDLTD then electronic publishing and digital libraries will be firmly established in the world of the academy and scholarship. If NDLTD receives proper support, it will become a vast, heterogeneous, federated, multilingual, multimedia digital library with more than 200,000 new works added yearly.

NDLTD is a model digital library project with "something for everyone." Students learn important skills, save money by not having to prepare multiple paper copies, and reach a much larger audience with their theses or dissertations. Universities become involved in work with digital libraries by way of a modest investment of time and resources, which in the long run saves money relative to conventional practices, prepares them for other digital library and electronic publishing projects, and makes their research results more visible and accessible worldwide. Scholars benefit by having a large new collection available with detailed information about university investigations, with known quality, large bibliographies, and careful literature reviews. Digital library enthusiasts have a new collection to experiment with regarding problems in the field (Fox & Sornil, 1999). In subsequent sections, we explore recent progress. For additional information, see the online file of papers and publications at <http://www.ndltd.org/pubs/> (NDLTD 1998d).

VISION, BENEFITS, APPROACH, POSSIBILITIES

Graduate education is an important part of the world of higher education in the United States and around the globe (Berelson, 1960). The Council of Graduate Schools (1991) has pinpointed dissertations as playing a key role in the training of researchers; more than 50,000 are completed each year in North America. There are over 350,000 master's degrees awarded yearly in the United States many of which lead to a thesis or substantive report. Hence we adopt the goal for the Networked Digital Library of Theses and Dissertations of collecting at least 200,000 electronic theses or dissertations per year to help improve graduate education (Eaton, Fox, & McMillan, 1997) and extend knowledge sharing on a global level. This will play a role in the move toward a worldwide digital library (Fox & Marchionini, 1998).

The vision behind the Networked Digital Library of Theses and Dissertations is to accomplish multiple objectives through one collaborative effort that will scale to include every university and every graduate student:

- prepare all future scholars for the Information Age (Fox, Hall, & Kipp, 1997) so that they understand important concepts and practices related to electronic publishing, digital libraries, and intellectual property rights, and so the scholarly community will adopt these practices more quickly, smoothly, and effectively;
- help universities to enhance graduate education and to develop expertise regarding digital libraries; and
- build an important digital library that supports scholarship and that improves as a result of new research in the field.

Scalability allows the Networked Digital Library of Theses and Dissertations to expand rapidly and results from:

- having each student, as part of ensuring their education is effective for the Information Age, create their own electronic thesis or dissertation, provide required metadata for it, decide about intellectual property and access rights, and upload it into the digital library;
- having each university run its own part of NDLTD as part of its support for graduate education and its delving into digital libraries;
- applying automation to improve workflow and streamline processing of ETDs so a more efficient scheme results that is cheaper than the old way; and
- federating the various collections and servers in NDLTD so users may view the entire collection as a whole or obtain views based on geography, level, topic, or other criteria.

Other key ideas include:

- demonstrating that, for many purposes, people really can switch fully to electronic documents;
- mandating submission to NDLTD once a university is ready for that step, since then the benefits and collections increase more quickly;
- adopting standards to facilitate preservation and federated access; and
- encouraging universities to work together and share their tools and information, in this domain where competition is minimal while mutual benefit is additive (Fox, 1997, 1998; Fox, Hall, Kipp, Eaton, McMillan, & Mather, 1997).

Since each university can have its own part of the Networked Digital Library of Theses and Dissertations, it is simple to adopt local policies and procedures. In addition, local groups interested in any of the myriad aspects of digital library research can experiment locally to perfect their methods. Thus, at Virginia Tech there are a number of development efforts (Fox, Kipp, & Mather, 1998). These involve applying diverse technologies including OpenText's search system, IBM's digital library product, OCLC's SiteSearch software, a locally developed federated search system (Powell, 1998), and even an experimental system extracting the images in electronic theses and dissertations for display in a virtual reality environment (Bayraktar et al., 1998).

In addition to the goals listed above, it is hoped that the Networked Digital Library of Theses and Dissertations will improve the culture of scholarship. By having a common environment wherein new research is freely shared, there can be greater interchange among students, faculty, and other scholars. Since this environment can be shared worldwide, there can be more understanding internationally and development of larger invisible colleges—all aided through the many tools provided for computer-supported cooperative work. Since universities will support NDLTD locally, it will be possible for undergraduates to connect with research efforts and even to submit their own theses and major papers for inclusion. Finally, all of this may have a profound effect on the world of publishing as universities become more directly involved in the collection and dissemination of research results.

CONCERNS, PROBLEMS, OPPOSITION

On the technical side, there are many digital library problems that relate to improving NDLTD. Since solutions exist or are under development for these, most are discussed in the next section. Here, however, it is appropriate to mention three examples. First, there is the problem of facing the incredibly diverse and fragmented world of electronic publishing, where a myriad of tools, packages, and versions are employed. Some-

how one must translate from these to a sensible number of standard forms that can be effectively preserved (Datta & Fox, 1993). This problem, couched in terms of using SGML (Standard Generalized Markup Language, an ISO standard), led to the first workshop about ETDs hosted by UMI in November 1987. Approaching this problem is at the heart of projects that have been called electronic archives (Fox, 1990), electronic libraries and, most recently, digital libraries (Fox, Akscyn, Furuta, & Leggett, 1995).

Second, there are problems related to cataloging and classifying the contents of digital libraries (McMillan, 1996). One aspect of this concerns standards, too. The Machine-Readable Bibliographic Information Committee (1996) approach leads to the MARC family adopted by many who do cataloging. Recent meetings and activities have led to refinement of a simpler alternative, the Dublin Core, and to crosswalks between it and MARC (Library of Congress, 1997).

Third, there are problems related to searching and accessing a large distributed collection like NDLTD. A closely related problem is that of handling technical reports. This has been examined in the computer science field (Fox, 1995), first through projects like WATERS (French, Fox, Maly, & Selman, 1995), and recently through the Networked Computer Science Technical Report Library (NCSTRL) (Lagoze, 1998). At present, federated searching for ETDs can be handled with a simple system developed at Virginia Tech (NDLTD, 1998a), but more advanced systems will be warranted as the collection expands.

On the social side, there are even more problems that relate to large distributed digital library projects (Borgman, 1996). Since NDLTD calls for widespread change, uses advanced technologies, relates to the Internet and WWW, involves universities and publishers in new relationships, and is emerging at a time of turmoil in the field of scholarly publishing, it has been quite visible in the news (NDLTD, 1998b). Universities are concerned about their expenditures for journals and seek to play a more active role in the process of scholarly publishing (Association of Research Libraries et al., 1998). NDLTD aims to help in this arena, including assisting students, faculty, and universities to understand the issues more thoroughly. Thus it must deal with diverse policies regarding prior publication, derivative works, inclusion of copyrighted materials in ETDs, and tradeoffs among modes of publication (e.g., NDLTD versus journal versus university press). In addition, there are tradeoffs for ETDs regarding collecting, archiving, and accessing—what are the roles of corporations (e.g., UMI), organizations run by libraries (e.g., OCLC), universities, university consortia, and national projects/libraries?

Ultimately, these social issues lead to actions by individuals and universities. Though members of NDLTD (1998c) include supporting organizations and consortia as well as some divisions of universities, most that

have joined represent the entire university, so we focus on that group. First, these organizations must identify those interested in the technical, policy, educational, research, and social issues. Volunteers, people assigned, or representatives of the various constituencies (especially the graduate school and library) must join in the local ETD team. Second, there must be discussion on campus so that the general aims of NDLTD can be specialized for the campus. Finally, these ongoing activities must lead to an implementation project that will evolve over the years.

SOLUTIONS, IMPLEMENTATION

With origins dating back to 1987, and with funded research and development led by the NDLTD team at Virginia Tech since 1996, many solutions have emerged. At Virginia Tech, where ETDs have been required since the start of 1997, over 1,100 were in the collection by mid-1998. By that point, there were over 230,000 downloads from around the world of the HTML pages that cover the content of the cover sheet and abstract and almost 200,000 downloads of the full documents.

Handling ETDs is established practice there, supported by staff in the graduate school, library, computing laboratories, and departments. Students carry out their investigations, present their results to their committee in a final defense, complete their ETD, and submit it through a tailored WWW form that captures relevant metadata. The graduate school checks the submission and requests revision until an acceptable form is provided. The library makes the work available through the digital library NDLTD (1998a), also adding cataloging information for the MARC record (which in the 856 field directs people to the online full-text copy) that goes into the campus catalog system.

There is a WWW site for students to help them understand policies and technical issues (NDLTD, 1998f). There are detailed instructions on how to prepare an ETD, explanations of standards and practices related to multimedia and hypermedia, files of questions and answers, templates, and checklists. Most students provide one or more Portable Document Format (PDF) files, while some supply SGML or XML instead; multimedia content is included in supplementary files.

There is an approval form (NDLTD, 1998g) that is printed and goes on file for each ETD to handle the concerns of publishers. It records the decision of the author and faculty committee regarding the breadth of access (e.g., locked up for patent purposes, restricted to campus while awaiting publication in a journal, or open for worldwide downloading). The WWW site includes letters from publishers that allow worldwide ETD access as well as inclusion of derivative articles in their journals. However, students completing the approval form and thinking about other publishers of their results must consult their policies. It is hoped that NDLTD

files will record these policies to reduce the need for future inquiries, and more publishers will endorse worldwide access through NDLTD.

NDLTD received funding in September 1996 as a national project (Fox, Eaton, McMillan, Kipp, Weiss, Arce, & Guyer, 1996). Within a year it expanded to an international effort with rapid growth (Fox, Eaton, McMillan, Kipp, Mather, McGonigle, Schweiker, & DeVane, 1997). By summer 1998, thirty-five universities had joined NDLTD. Ongoing efforts to involve more universities will continue to require discussion, visits to campuses, and presentations. Reaching all universities worldwide will take years but should accelerate as larger and more prestigious universities join and a critical mass is achieved.

On the international scene there are promising signs for growth. NDLTD members are on all continents. Groups of universities are joining including three in Canada. Australia has a funded national project involving seven universities. St. Petersburg State Technical University has submitted several proposals for funding to expand efforts in Russia. The National Library of Portugal has joined, and can represent all the universities in the nation. Several universities in Korea and Singapore are members with particular interest in supporting multilingual collections.

Since most graduate studies are carried out in local languages, it is essential that NDLTD include works in all languages and support all fonts. Some groups supporting NDLTD have active research programs in this area (Cao, Lu & Low, 1998; Leong, Cao, & Lu, 1998). These could lead to refinement of the existing federated search system, which uses translation of English query terms through a dictionary database for current multilingual searching (Powell, 1998).

To expand the capabilities of the NDLTD software, a requirements study has been underway in connection with work on an honors thesis. In addition to focus groups and brainstorming in a decision support facility, the effort includes developing and refining a number of scenarios (Carroll, 1995). By spring 1999, there should be a prototype to illustrate the advanced capabilities possible in a digital library focused on aiding graduate students who find, use, and produce electronic theses or dissertations.

CONCLUSION

It is now clear that, eventually, graduate education will shift from more traditional forms to electronic theses or dissertations. Various related projects have emerged (NDLTD, 1998e) and it is hoped will all fit into the broad umbrella of NDLTD. More and more institutions are a part of this federation, and as each develops its own local program, the collection size will grow rapidly.

Issues regarding widespread access and harmonizing with the practices of publishers will remain, since NDLTD aims to effect both change

and to extend understanding of policies and practices. Other social concerns will be dealt with by local institutions, the NDLTD Steering Committee, and the implementation team at Virginia Tech as they continue to arise in the normal course of the evolution of this expanding digital library initiative.

On the technical side, ensuring interoperability and preservation, handling a large multilingual collection, and expanding services are key goals. Several proposals for research support have been submitted, and work will continue to deal with the challenges of this model digital library project.

In conclusion, NDLTD has had considerable growth throughout its short history, has dealt with many diverse problems, and has ambitious plans for future activity. It is hoped that new universities will join and contribute not only their ETDs but also share their findings, information, and developments so all can benefit from this broad university collaboration.

ACKNOWLEDGMENTS

Thanks go to the NDLTD Steering Committee, Technical Committee, and Virginia Tech management team (with co-principal investigators John Eaton and Gail McMillan and graduate research assistants Robert Hall, Neill Kipp, Paul Mather, Tim McGonigle). Others helping at Virginia Tech include Emilio Arce, Anthony Atkins, Brian DeVane, Scott Guyer, James Powell, Bill Schweiker, and Laura Weiss. Thanks also go to all the members of NDLTD and their faculty, staff, and students.

Funding has been provided by the Southeastern Universities Research Association and the U.S. Department of Education (FIPSE). Donations have been given by IBM, Adobe, Microsoft, OCLC, and others. Support and in-kind contributions have been given by many groups including the Coalition for Networked Information, Conference of Southern Graduate Schools, Council of Graduate Schools, and SOLINET (SOutheastern Library NETwork).

REFERENCES

- Association of Research Libraries, Association of American Universities, & Pew Higher Education Roundtable. (1998). To publish and perish. *Policy Perspectives* [special issue], 7(4), 1-11.
- Bayraktar, M.; Zhang, C.; Vadapalli, B.; Kipp, N.; & Fox, E. A. (1998). A Web art gallery. In *Proceedings of the Third ACM Conference on Digital Libraries, DL '98* (Pittsburgh, PA, June 23-26, 1998) (pp. 277-278). New York: ACM Press.
- Berelson, B. (1960). *Graduate education in the United States*. New York: McGraw-Hill.
- Borgman, C. (Ed.). (1996). *Social aspects of digital libraries* (A workshop hosted by UCLA, sponsored by NSF, February 16-17, 1996, UCLA, CA). Retrieved September 17, 1999 from the World Wide Web: <http://www-lis.gseis.ucla.edu/DL/>.
- Cao, L.; Leong, M. K.; Lu, Y.; & Low, H. B. (1998). Searching heterogeneous multilingual bibliographic sources. *Computer Networks and ISDN Systems*, 30(1-7), 612-615.

- Carroll, J. M. (Ed.). (1995). *Scenario-based design: Envisioning work and technology in system development*. New York: John Wiley & Sons.
- Council of Graduate Schools. (1991). *The role and nature of the doctoral dissertation: A policy statement*. Washington, DC: Council of Graduate Schools.
- Dalal, K., & Fox, E. A. (1993). *Document translation: Dissertations and technical reports*. Unpublished Technical Report TR-93-31, Department of Computer Science, Virginia Tech, Blacksburg, VA.
- Eaton, J.; Fox, E. A.; & McMillan, G. (1997). Electronic Theses and Dissertations (ETDs) and their contribution to graduate education. In *Proceedings of the 53rd Annual Meeting Midwestern Association of Graduate Schools, MAGS* (pp. 73-78). Chicago: MAGS.
- Fox, E. A. (1990). How to proceed toward electronic archives and publishing. *Psychological Science*, 1(6), 355-358.
- Fox, E. A. (1995). World-Wide Web and computer science reports. *Communications of the ACM*, 38(4), 43-44.
- Fox, E. A. (1997). Networked Digital Library of Theses and Dissertations: An international collaboration promoting scholarship. *ICSTI Forum, Quarterly Newsletter of the International Council for Scientific and Technical Information* (No. 26) 8-9 November. Retrieved September 17, 1999 from the World Wide Web: <http://www.icsti.nrc.ca/icsti/>.
- Fox, E. A. (1998). Networked digital library of theses and dissertations: A framework for east-west collaboration. In *Proceedings of Asian Digital Library Workshop, "East Meets West"* (5-7 August 1998, Hong Kong, hosted by University of Hong Kong). Retrieved September 17, 1999 from the World Wide Web: <http://www.ndltd.org/pubs/Asia98.htm>.
- Fox, E. A.; Aksyn, R.; Furuta, R.; & Leggett, J. (1995). Guest editors' introduction to digital libraries. *Communications of the ACM*, 38(4), 22-28.
- Fox, E. A.; Eaton, J.; McMillan, G.; Kipp, N.; Weiss, L.; Arce, E.; & Guyer, S. (1996). National Digital Library of Theses and Dissertations: A scalable and sustainable approach to unlock university resources. *D-Lib Magazine*, September. Retrieved September 17, 1999 from the World Wide Web: <http://www.dlib.org/dlib/september96/theses/09fox.html>.
- Fox, E. A.; Eaton, J.; McMillan, G.; Kipp, N.; Mather, P.; McGonigle, T.; Schweiker, W.; & DeVane, B. (1997). Networked Digital Library of Theses and Dissertations: An international effort unlocking university resources. *D-Lib Magazine*, September. Retrieved September 17, 1999 from the World Wide Web: <http://www.dlib.org/dlib/september97/theses/09fox.html>.
- Fox, E. A.; Hall, R.; & Kipp, N. (1997). NDLTD: Preparing the next generation of scholars for the information age. *New Review of Information Networking (NRIN)*, 3, 59-76.
- Fox, E. A.; Hall, R.; Kipp, N.; Eaton, J. L.; McMillan, G.; & Mather, P. (1997). NDLTD: Encouraging international collaboration in the academy. *DESIDOC Bulletin of Information Technology*, 17(6), 45-56.
- Fox, E. A.; Kipp, N.; & Mather, P. (1998). How digital libraries will save civilization. *Database Programming & Design*, 12(8).
- Fox, E. A., & Marchionini, G. (1998). Toward a worldwide digital library: Introduction. *Communications of the ACM*, 41(4), 28-32.
- Fox, E. A., & Sornil, O. (1999). Digital libraries. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 415-432). New York: ACM Press.
- French, J.; Fox, E. A.; Maly, K.; & Selman, A. (1995). Wide area technical report service—technical reports online. *Communications of the ACM*, 38(4), 45.
- Lagoze, C. (1998). *NCSTRL Home Page*. Retrieved September 17, 1999 from the World Wide Web: <http://www.ncstrl.org>.
- Leong, M. K.; Cao, L.; & Lu, Y. (1998). Distributed Chinese bibliographic searching. *Communications of the ACM*, 41(4), 66-68.
- Library of Congress. (1997). Metadata, Dublin Core and USMARC: A review of current efforts (MARBI Discussion Paper no. 99, January). Retrieved September 17, 1999 from the World Wide Web: <gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc>.
- Machine-Readable Bibliographic Information Committee. (1996). *The USMARC Formats: Background and principles* (Machine-Readable Bibliographic Information Committee in conjunction with Network Development and MARC Standards Office, November).

- Retrieved September 17, 1999 from the World Wide Web: <http://lcweb.loc.gov/marc/96principl.html>.
- McMillan, G. (1996). Electronic theses and dissertations: Merging perspectives. *Cataloging and Classification Quarterly*, 22(3/4), 105-125.
- NDLTD. (1998a). *ETD Digital Library*. Retrieved September 17, 1999 from the World Wide Web: <http://www.theses.org>.
- NDLTD. (1998b). *NDLTD in the News*. Retrieved September 17, 1999 from the World Wide Web: <http://www.ndltd.org/news/>
- NDLTD. (1998c). *NDLTD Official Members*. Retrieved September 17, 1999 from the World Wide Web: <http://www.ndltd.org/contacts/>.
- NDLTD. (1998d). *NDLTD Papers and Publications*. Retrieved September 17, 1999 from the World Wide Web: <http://www.ndltd.org/pubs/>.
- NDLTD. (1998e). *NDLTD Related Projects*. Retrieved September 17, 1999 from the World Wide Web: <http://www.ndltd.org/related/projects.htm>.
- NDLTD. (1998f). *Virginia Tech Electronic Thesis and Dissertation home page*. Retrieved September 17, 1999 from the World Wide Web: <http://etd.vt.edu/>.
- NDLTD. (1998g). *Virginia Tech Graduate School Electronic Submission Approval Form*. Retrieved September 17, 1999 from the World Wide Web: <http://etd.vt.edu/submit/approval.htm>.
- Powell, J. (1998). *Virginia Tech federated searcher*. Retrieved September 17, 1999 from the World Wide Web: <http://jin.dis.vt.edu/fedsearch/>. Also used for *Networked Digital Library of Theses and Dissertations: Federated search*. Retrieved September 17, 1999 from the World Wide Web: <http://www.theses.org>.