

# A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations

Constantinos Phanouriou, Neill A. Kipp, Ohm Sornil, Paul Mather, and Edward A. Fox  
Department of Computer Science  
Virginia Polytechnic Institute and State University  
Blacksburg, VA 24061  
<http://www.ndltd.org/>  
{phanouri,nkipp,osornil,paul,fox}@cs.vt.edu

## ABSTRACT

The Networked Digital Library of Theses and Dissertations (NDLTD) is more than an online collection of Electronic Theses and Dissertations (ETDs). It is a scalable project that has impact on thousands of graduate students in many countries as well as diverse researchers worldwide. Its 59 official members represent 13 countries and integrate some of the world's newest research works, including ETD collections at Virginia Tech and West Virginia University, where ETD submission is now required. The number of ETDs in Virginia Tech's collection has nearly tripled in the last year, while the number of accesses to it has grown by more than half. NDLTD is committed to authors, aiming to improve graduate education for the over 100,000 students that prepare a thesis or dissertation each year. It encourage them to be more expressive by making incorporation of multimedia components into their theses easier. NDLTD activities include: applying automation methods to simplify submission of ETDs over the WWW; specifying the application of the Dublin Core to guarantee that metadata can satisfy needs of searching and browsing; selecting open standards and procedures to facilitate interoperability and preservation; and demonstrating a variety of interfaces, both 2D and 3D, along with exploring their usability.

**Keywords** digital library, user interfaces, information retrieval, usability engineering

## INTRODUCTION

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international effort that seeks to improve graduate education by encouraging all uni-

versities to require submission of Electronic Theses and Dissertations (ETDs). In the process of preparing and submitting their ETDs, student authors learn about the richness of expression that a digital medium makes possible and how to use online resources (i.e., digital libraries). It is through this process that universities can make available immediately and cost-effectively the research results of their graduate students as a contribution to the advancement of education and humanity [Fox, *et al.*, 1996, 1997, 1998].

NDLTD is a digital library in the richest of definitional senses [Borgman, 1999; Lesk, 1997; Fox, *et al.*, 1995]. It has a growing collection of ETDs that it makes available on the Internet; it is concerned with acquisition, preservation, and cataloging of ETDs; it provides useful and usable visualizations of the entire distributed collection. NDLTD is organizing universities and spreading new ideas about scholarly publishing through collaboration and sharing. As each member university joins the NDLTD, a local ETD submission process is planned—be it dictated by university governance, decided by faculty working group, or demanded by the graduate students themselves. Libraries renew their commitments to serve ever-widening scholarly communities, graduate schools sponsor training and workshops, and students and faculty become electronic document authors and publishers. With NDLTD, universities can evolve and share their own systems for collecting and making ETDs available and thereby contribute to the global educational process in exciting ways. As a result, graduate education and scholarly publishing will permanently change, with digital libraries playing a dominant role.

## NDLTD ACTIVITY

“We certainly want to be thorough and we absolutely must get it right, but this is not the sort of thing which will profit from passive study, and we have arrived at the point where we must begin to implement the project.” [WVU, 1998]

## Membership

As of May 1999, NDLTD has 59 members from 13 countries. Fifty-three (53) members are universities; the remainder are coalitions, non-profit organizations, or corporations.

## Governance

The NDLTD steering committee meets in September and April of each year, and is chaired by the initiative's director, Edward A. Fox. Membership includes representatives from Virginia Tech and other NDLTD member universities, Adobe, Association of Research Libraries, Coalition for Networked Information, Council of Graduate Schools, Dissertations Online (Germany), IBM, National Library of Canada, OCLC, UMI, and UNESCO. Topics discussed in the Fall of 1998 and Spring of 1999 included: membership, outreach, expansion programs, archiving, preservation, metadata, ETD submission and workflow processing, workshops, Web sites, ongoing evaluation, results reporting, particular implementations, development and plans, and future opportunities for funding.

## ETDs Required

While most universities in the NDLTD are implementing pilot programs, Virginia Tech and West Virginia University have made ETD submission a requirement for graduate students on their campuses.

**Virginia Tech.** Virginia Tech has required electronic submissions since January 1, 1997, and does not accept paper thesis and dissertation submissions. The Graduate School and University Library have collected more than 1700 ETDs. Of these, 1225 are available worldwide; the remainder are not available beyond the campus at the request of the submitting student. Most documents are in PDF, augmented by various multimedia formats (e.g., JPEG, GIF, TIFF, MPEG, WAV, HTML, VRML, QuickTime, Java applets). Most were created in Word and Word Perfect, but some were created in TeX, LaTeX, and SGML (using the ETD-ML document type definition). The Virginia Tech ETD library uses OpenText for indexing the full text of the collection.

**West Virginia University.** In August 1998, West Virginia University began to require students to submit theses and dissertations electronically [Mendels, 1998]. WVU no longer accepts paper theses and dissertations; exceptions must be approved by the Office of the Provost. WVU requires its documents to be submitted in PDF format. The West Virginia ETD collection contained 210 documents as of April 1999. The local committee for ETD implementation consists of members from its faculty, library, research centers, graduate school, and the Office of Academic Affairs.

## Other Collections

**Australian Digital Theses Project.** Seven institutions in Australia (led by the University of New South Wales, and centered in its library) are collaborating to begin accepting electronic theses from postgraduate students. They have standardized on SGML and PDF as document formats. The collection's oldest work is dated 1968.

**Dissertation.com.** Dissertation.com is part of Amazon.com and functions as a publishing agent for students. It offers electronic dissertations in PDF or paper formats for 20 to 40 US dollars. Abstracts are freely available.

**Dissertations Online.** A national project in Germany involves 4 universities, 2 large libraries, a large computing center, and 4 scholarly societies (chemistry, mathematics, physics, sociology, and education). The focus is on SGML and XML, and helps train students in their disciplines, e.g., to use the markup language for chemistry.

**Encyclopaedia Diplomica.** Encyclopaedia Diplomica is a German company acting as a selling agent for students who prepare scholarly works. Papers are in one of the following formats: Word, PDF, or PostScript. Abstracts and full tables of contents are available for free. Prices for the full documents are 150 to 300 US dollars. The collection offers approximately 20 titles. Most of the documents are in German; the rest are in English or French.

**North Carolina State University.** NCSU has about 30 ETDs in its online collection, which is sponsored by the NCSU Libraries, Graduate School, and Information Technology division. At NCSU, ETD submission is not yet required. Submissions are in PDF format. The Graduate School holds monthly thesis preparation workshops for its students.

**Rhodes University of South Africa.** The Rhodes University of South Africa has begun an ETD pilot project. They request both paper and digital submissions.

**University of Tennessee, Memphis.** The University of Tennessee, Memphis has three documents in its collection. Of these, all are in PDF, but one is also in HTML.

**University of Michigan.** While not an official member of NDLTD, the University of Michigan has begun a thesis pilot program. Instead of PDF, they have four ETDs in SGML, conforming to the Text Encoding Initiative Document Type Definition [Sperberg-McQueen and Burnard, 1994].

**University of Virginia.** The University of Virginia has adopted an ETD pilot; it accepts electronic theses from Engineering bachelor's students. The university plans

to require Master's and PhD's at a later time.

**University of Waterloo, Ontario.** The University of Waterloo in Ontario, Canada is the center of a three-institution cooperative and has sixteen documents online, in PDF with paper and PostScript sources, including one dated 1964. The site is sponsored by the Electronic Thesis Project Team and the University of Waterloo Library. They provide documents for free, but request the name, affiliation, and "reason-why" from the patron before permitting the thesis to be downloaded. The site uses OpenText for searching the full text of the collection.

#### **VIRGINIA TECH INITIATIVE (VT-ETD)**

The Virginia Tech ETD (VT-ETD) initiative has developed software and practices adopted by a number of other NDLTD members.

#### **Authoring and Training**

ETD authors are typically graduate students with above-average knowledge about computers. In a survey of graduate students after their submission, it was found that almost all of them used the Web to find information while doing their research. This makes them aware of what can be published electronically. Keeping with the goal of NDLTD that students should be able to author, submit, and maintain (with annotations and reviews) their work electronically, VT-ETD educates them on how to contribute their own work to the online community with workshops and a comprehensive and informative Web site (<http://etd.vt.edu/>).

Creating a document electronically is simple; enriching it with multimedia, aligning with standards, and making it interactive can be challenging. VT-ETD attempts to make interactive multimedia easier for students by providing the necessary tools and help on how to use them. Usage of multimedia in ETDs is increasing, perhaps due to regular training workshops sponsored by the Graduate School. For example, a dissertation from Chemistry contains 3D VRML models of molecules, a thesis from Animal Science contains audio clips of parrot sounds, and a thesis from Architecture contains video clips from a Turkish coffeehouse.

#### **Acquisition and Collection Management**

To improve upon ETD submission scripts that were written as prototype software, VT-ETD recently developed customized, database-driven ETD management software. As a result, students have more control over their ETD during the entire authoring and submission process. Furthermore, storing ETD metadata in a database enables the VT-ETD to provide better acquisition, searching, and browsing services. It also enables the development of multiple user interfaces to the collection.

VT-ETD encourages students to treat their theses or dissertations as electronic documents from the begin-

ning. With the new submission software students can register their documents early, set up their metadata, and upload their ETD drafts in pieces. This benefits the students in two ways. Primarily, students' work is stored in a secure location, with frequent and reliable backups; thus they are less likely to lose their work if their home or office computer fails. Furthermore, drafts are available for their committee to see and review electronically. Students can restrict public access to their work until they defend. Todd Miller's honor thesis work has developed and tested software to provide online annotation services to the ETDs.

VT-ETD provides students with four options on making their work accessible. The first option is *unrestricted*: release the entire work immediately for access worldwide. The second option is *restricted*: release the entire work for Virginia Tech access only. The third option is *withheld*: secure the entire work for patent and/or proprietary purposes for a period of one year. With the new submission software, VT-ETD now provides students with a fourth option, *mixed*, where they can break down their work and use any of the above options for each part individually. For example, they can make their abstract and introduction worldwide accessible, but restrict the main body of their work for Virginia Tech access only. Table 1 shows the distribution of their choices.

#### **Cataloging**

To facilitate collection sharing, NDLTD members are asked to freely share any MARC records available. Also, due to the variety of heterogeneous DL implementations it became obvious that a standard set of metadata elements should be identified for ETDs (which itself would aid the development of a canonical representation for ETDs). The metadata should be broad enough to permit crosswalks between many popular metadata standards and frameworks such as MARC and Dublin Core, but focussed on the domain of ETDs. VT-ETD began with the standard Dublin Core elements, and then enhanced these to tailor them to the ETD domain.

VT-ETD metadata is designed for practical application across a broad set of information storage and retrieval applications and settings, such as Web, IBM Digital Library, OCLC SiteSearch, and OpenText LiveLink. It is intended to provide a functional medium between static resource description and periodic record-keeping (low-level authority information). As such, part of the metadata is tied to (and derived from) ETD workflow and graduate school policies. Most, however, is static information that is supplied by the ETD author or by a cataloger.

The metadata framework is not rigid. The elements are guidelines, with varying degrees of recommendation for

employment in a local institutional ETD project setting. All NDLTD members are expected to collect all metadata elements marked as “mandatory,” so that there may be a minimal basis for searching across all NDLTD collections.

Each ETD has metadata describing the ETD as a whole, and then each separate part (file) of the ETD has its own metadata. URN identifiers within the metadata parts are used to tie together the metadata parts of an ETD in a parent/child structure. Implicit inheritance is used to minimize the repetition in elements for child items of an ETD.

### Preservation

The Virginia Tech Graduate School requires a specific form for the submission of ETDs to maintain the consistency of these complex documents. The formal statement of these guidelines serves graduate students submitting ETDs, professors with whom they work, and scholars who study the submitted ETDs. VT-ETD defined ETD-ML, a Document Type Definition (DTD) in both SGML and the Extensible Markup Language (XML) for the representation of ETDs. XML is a logical choice for encoding and archiving complex electronic documents [Bray, *et al.*, 1998]. To build ETD-ML, VT-ETD analyzed constructs in existing theses and dissertations and studied the rules for their submission. Software is available to NDLTD members that converts ETD-ML ETDs into HTML for Web accessibility.

### Search and Retrieval

**IBM DL.** VT-ETD developed a preliminary interface to the ETD collection using the IBM Digital Library (IBM DL) product. The Net.Data dynamic page builder component of IBM DL provides a Web front-end to the contents inside the IBM DL and allows users to search the VT-ETD collection. The full-text of the ETD PDF files, as well as abstracts, are indexed by the IBM DL text search server.

The current IBM DL search interface allows users to search the VT-ETD collection in two ways: through the metadata or the full-text index. A user can perform either type of search separately or use both types simultaneously in the same search. Thus, for example, a user can search the collection for each time the phrase “digital library” appears in an ETD, meanwhile specifying that each retrieved ETD must be a thesis from computer science submitted prior to 1997.

**SIFT.** The SIFT filtering software from Stanford has been adapted by Zhambo Sun to work for ETDs. This allows interested parties to specify information needs through email or a WWW interface. When integrated into the rest of the workflow, this should lead to an email notification whenever a new submission matches

any stored user profile.

### NEW DIRECTIONS IN VISUALIZATION

VT-ETD is trying to enhance the information retrieval process for DLs by developing richer browsing interfaces to its ETD collection. In particular, VT-ETD is experimenting with 3D interfaces on desktop machines and on immersive virtual reality devices.

#### 3DL

3DL presents the ETD collection as a 3D VRML model through which users can navigate. It mimics a traditional library: including lobbies, elevators, floors, signs, displays, windows, artworks, doors, rooms, bookcases, and books. Doors are hyperlinks to rooms and books are labeled hyperlinks to items in the ETD collection [Kipp, 1997]. In addition to the usual library components, 3DL uses images extracted from the collection and presents them as hyperlinks in a “virtual art gallery” [Bayraktar, *et al.*, 1998]. VT-ETD developed 3DL as an alternative interface to the ETD collection.

#### CAVE-ETD

CAVE-ETD extends the 3DL project from the desktop to an immersive virtual reality environment. CAVE-ETD runs in the Cave Automated Virtual Environment (CAVE), a 10x10x10-foot room, with stereoscopic projections on three walls and the floor, wherein the user may interact with the world through tracking devices, eyeglasses, and a wand. In CAVE-ETD, “Books” are organized on “shelves,” shelves are laid out in “aisles” in a “room,” and rooms are labeled and arranged in a logical sequence. Books can be browsed on the shelves by navigating through the room and reading the titles on the book spines. Real-time clustering methods are being investigated to determine their utility. Although it is unlikely that we will all have a 3D CAVE in our office, it is more likely that we will have a miniature version on our desk.

Both the CAVE-ETD and the 3DL rely on a user’s prior knowledge and experience in a traditional library. This conforms to the usability principle of familiarity [Hix and Hartson, 1993] which aligns with the results of the usability trials of both the 3DL and the CAVE-ETD.

### QUANTITATIVE EVALUATION

**Collection Size.** VT-ETD began collecting ETDs in 1995. By the end of 1998, Virginia Tech had 1546 electronic documents (theses, dissertations, and other documents) in its ETD collection (Table 2).

ETD authors are encouraged to include various multimedia components. Most PDF files include color images or figures. Ninety-three percent (93%) of the files in the collection are PDF and text files, while nearly 7% of the files are supplemental images, sounds, and movies (Table 3).

**Access Statistics.** As the collection grows and gains popularity and more institutions join NDLTD, the number of accesses to the system goes up (Table 4).

The monthly access graph is shown in Figure 1. We can see that number of accesses tends to increase each year. However, there were fewer accesses during the summer break when universities are not in session.

Among US domestic domains, educational institutions contributed to the largest number of requests. Half of these accesses were from users at Virginia Tech, affirming that local researchers and authors are using their own collection. Commercial interest is next, followed by other organizations, while government domains continued to show high interest (Table 5).

Each of the top-five accessing countries has increasing number of accesses every year (Table 6). The United Kingdom and Germany dominated the accesses from outside the US. This trend corresponds to the advancements in network facilities in those countries.

## USABILITY EVALUATION

**3DL.** Human interfaces to VRML browsers are notoriously bad [Carey and Bell, 1997]. VT-ETD usability trials support this conclusion. Even with high-resolution, high-speed desktop displays, refresh rate was poor and navigation was clumsy. Users said that looking at the rooms interface was “nice” but that a plain list of titles would be more useful. VT-ETD is working to improve the speed and usability of the 3DL interface before further evaluation.

**CAVE-ETD.** In trial runs in the CAVE-ETD, we noticed that new users have difficulty adjusting to the interface. Although the interface corresponds most closely with that used in most 3-D computer games, using the CAVE “wand” input device is not natural. “Sideways stepping” would also make browsing through books on shelves more useful, say users. Providing a useful amount of text (e.g., author, title, year) on the book spines substantially slows the CAVE display. The usability study produced many qualitative hints for designers of three-dimensional interactive library interfaces.

## CONCLUSIONS

Digital libraries are more than organizations of information. They are systems by which societies cope with their information problems and through which societies provide information services to users. NDLTD contains a document repository, indeed, but it also consists of the system and society by which that document repository is grown, accessed, maintained, and preserved.

NDLTD is a live test of a new economic model for digital libraries, whereby automation and federation, plus coupling to normal practices and use of standards, lower the

costs sufficiently so that in the normal course of work by authors, graduate schools, and university libraries, a sustainable worldwide digital library can be built, leading to unprecedented sharing of research results. Ongoing research and development work, at Virginia Tech and by other NDLTD members, should expand and improve the services and benefits of this initiative.

## ACKNOWLEDGMENTS

The authors acknowledge the efforts of the NDLTD team, particularly John L. Eaton and Gail McMillan. We also thank Bharadwaja Vadapalli, Prashant Choudhary, Jay Rathi, Nirav Kamdar, Murat Bayraktar, Chang Zhang, for their work on 3DL. For ongoing development of VT-ETD and his careful collection of statistics, we thank Anthony Atkins. For their contribution to the ETD CAVE we thank Kevin Curry, Fernando Das Neves, and Hussein Suleman.

**Funding.** As of September 1, 1996, the U.S. Department of Education Fund for the Improvement of Post-secondary Education (FIPSE) provided grant support for a three-year project, “Improving Graduate Education with the National Digital Library of Theses and Dissertations (NDLTD).” This follows earlier support from the Southeastern Universities Research Association (SURA) for the “Development and Beta Testing of the Monticello Electronic Library Thesis and Dissertation Program.”

## REFERENCES

- Bayraktar, Murat, Chang Zhang, Bharadwaj Vadapalli, Neill A. Kipp, Edward A. Fox, “A Web Art Gallery,” Proceedings of Digital Libraries '98, the Third ACM Conference on Digital Libraries, Pittsburgh, June 1998.
- Borgman, Christine L., “What are Digital Libraries: Competing Visions,” *Information Processing and Management, Special Issue for Digital Libraries*, Gary Marchionini and Edward A. Fox, issue editors, 1999.
- Bray, Tim, Jean Paoli, and C. M. Sperberg-McQueen, “Extensible Markup Language (XML) 1.0,” W3C Recommendation, <http://www.w3.org/TR/REC-xml>, February, 1998.
- Carey, Rikk, and Gavin Bell, *The Annotated VRML 2.0 Reference Manual*. Addison-Wesley, 1997.
- Fox, Edward A., and James Powell, “Multilingual Federated Searching Across Heterogeneous Collections,” *D-lib Magazine*, September, 1998.
- Fox, Edward A., Brian DeVane, John L. Eaton, Neill A. Kipp, Paul Mather, Tim McGonigle, Gail McMillan, William Schweiker, “Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources,” *D-lib Magazine*, September,

**Table 1: Accessibility for first 1729 VT ETDs**

Accessibility type	Number of Documents	Percent
Withheld	338	19.6
Unrestricted	820	47.4
Restricted	542	31.3
Mixed	29	1.7
Total	1729	100.0

**Table 2: ETD collection size through 1998.**

ETD Types	% of ETDs	Pre-1996	1996	1997	1998	Total	%96-97	%97-98
Dissertations	46.0	4	35	167	505	711	377	202
Theses	52.8	14	49	232	522	817	373	125
Others	1.2		1	4	13	18	300	225
Totals		18	85	406	1040	1546	374	158
% of all ETDs		1.16	5.5	26.1	67.3			

**Table 3: Separate multimedia files in first 1454 VT ETDs**

File type	Number of Documents	Percent
PDF, text	5334	93.3
Image	322	5.6
Movie	45	0.8
Sound	18	0.3
Total	5719	100.0

**Table 4: Access Statistics through 1998**

	1996	1997	1998	%96-97	%97-98
Total successful HTTP requests	37,171	247,573	379,742	566	53
Average successful requests per day	102	678	1040	665	153
Distinct hosts served	9015	22,725	36,724	152	62
Total data transferred (Gb)	3.229	25.9	50.0	704	93
Average data transferred per day (kb)	9.038	73.6	136.9	814	186

**Table 5: Accesses from domestic domains**

Domain	96	97	98	%96-97	%97-98
US Education (.edu)	15,314	112,876	254,268	637	125
US Commercial (.com)	5,309	48,540	88,169	814	82
Networks (.net)	2,522	14,026	27,972	456	99
Other Organizations (.org)	375	3,132	1,434	735	-54
US Government (.gov)	282	1,362	6,885	383	406

**Table 6: International Accesses (selected)**

Countries	1996	1997	1998	%96-97	%97-98
United Kingdom	850	2922	8170	244	180
Germany	346	2378	7373	587	210
Australia	608	2501	4223	311	69
France	463	1161	4431	151	282
Canada	713	2367	3970	232	68

1997.

Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, Scott Guyer, "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources" *D-lib Magazine*, September, 1996.

Fox, Edward. A, Robert M. Akscyn, Richard K. Furuta, and John J. Leggett, "Digital Libraries," *Communications of the ACM*, 38(4), pp. 22-28, April, 1995.

Hix, Deborah, and H. Rex Hartson, *Developing User Interfaces*, John Wiley and Sons, 1993.

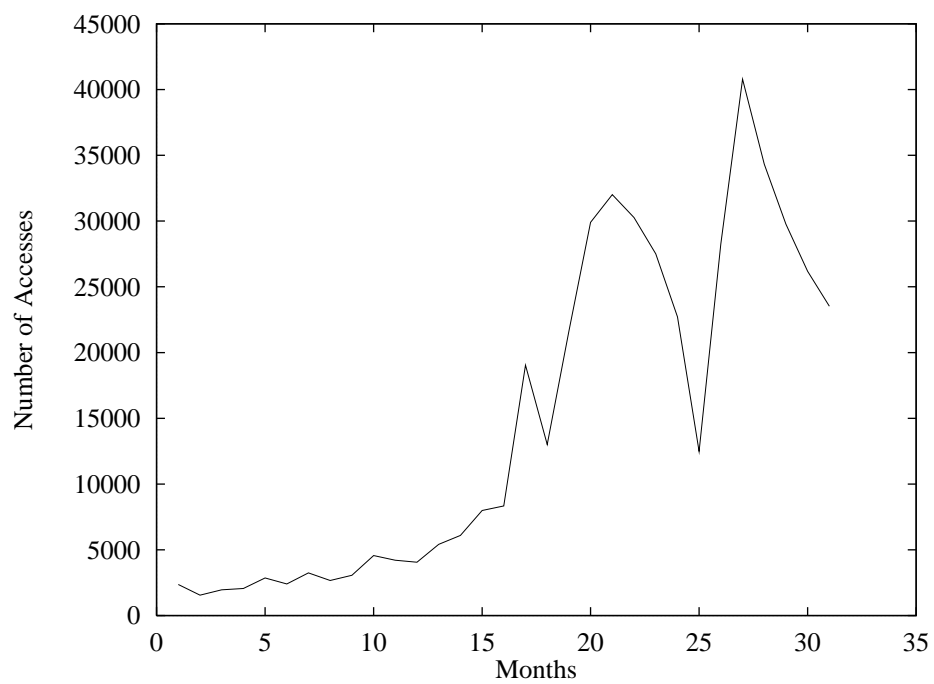
Kipp, Neill A., "Case Study: Digital Libraries with a Spatial Metaphor." SGML/XML '97 Conference Proceedings. Graphic Communications Association, Alexandria, VA, December, 1997.

Lesk, Michael. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan-Kaufmann, 1997.

Mendels, Pamela. "Paper-Bound Thesis Dusted Off, Digitally," *New York Times*, September 5, 1998.

Sperberg-McQueen, C. M., and Lou Burnard, editors, *Guidelines for Electronic Text Encoding and Interchange: TEI P3*, Text Encoding Initiative, Chicago, 1994.

West Virginia University, "Frequently Asked Questions on ETDs," <http://www.wvu.edu/~thesis/etd-faq.html>, July, 1998.



**Figure 1: Monthly accesses (January 1996—July 1998)**