Overview of a Guide for Electronic Theses and Dissertations

Edward A. Fox
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA – fox@vt.edu – http://fox.cs.vt.edu

Abstract
This chapter provides an overview of a guide for electronic theses and dissertations that is being prepared as requested by UNESCO to help with the expansion of ETD activities around the world. It roughly follows the outline developed through discussions involving the many partners working on that guide, coordinated by Shalini Urs. It builds upon experiences related to the evolution of the Networked Digital Library of Theses and Dissertations, a federation of groups interested in ETD programs. It introduces key concepts, explains matters according to the interests of students and universities, highlights technical issues, recommends a scheme for expanding training, and suggests likely future activities.

1. Introduction

1.1.What are ETDs?

Joining and participating in the Networked Digital Library of Theses and Dissertations, NDLTD [1] is one of the best ways to understand the concepts explained earlier in this booklet regarding digital libraries. It directly involves students pursuing graduate education by having them develop their theses or dissertations (TDs) as electronic documents, that is, as electronic theses or dissertations (ETDs).

1.1.1. ETDs as new genre of documents

With thousands of students each year preparing ETDs, the creativity of the newest generation of scholars is being expressed continuously as they work to present their research results using the most appropriate form, structure, and content. While conforming as needed to requirements of their institution, department, and discipline, students should develop and apply skills that will prepare them best for their future careers and lead to the most expressive rendering possible of their discoveries and ideas. Thus, ETDs are a new genre of documents, continuously re-defined as technology and student knowledge evolve.

1.2. Purpose, goals, objectives of ETD activities

The underlying purpose of ETD activities is to prepare the next generation of scholars to function effectively as knowledge workers in the Information Age. By institutionalizing this in a worldwide program, progress can be made toward tripartite goals of enhancing graduate education, promoting sharing of research, and supporting university collaboration. Particular objectives include:

- students knowing how to contribute to and use digital libraries;
- universities developing digital library services and infrastructure;
- enhanced sharing of university research results; and
- ETDs having higher quality and becoming more expressive of student findings.

1.3. Brief history of ETD activities: 1987-2000

The first real activity directed toward ETDs was a meeting convened by Nick Altair of UMI in Ann Arbor, Michigan during the fall of 1987 involving participants from Virginia Tech, ArborText, SoftQuad, and University of Michigan. Discussion focussed on the latest approaches to electronic publishing and the idea of applying the Standard Generalized Markup Language (SGML, an ISO standard approved in 1985) to the preparation of dissertations, possibly as an extension of the Electronic Manuscript Project. In 1988, Yuri Rubinsky of SoftQuad was funded by Virginia Tech to help develop the first Document Type Definition (DTD) to specify the structure of ETDs using SGML. Pilot studies continued using SoftQuad's AuthorEditor tool, but only with the appearance of Adobe's Acrobat software and Portable Document Format (PDF) in the early 1990s did it become clear that students could easily prepare their own ETDs. In 1992 Virginia Tech joined with the Coalition for Networked Information, the Council of Graduate Schools, and UMI, to invite ten other universities to select three representatives each, from their library, graduate school/program, and computing/information technology groups. This meeting in Washington, D.C. demonstrated the strong interest in and feasibility of ETD activities among US and Canadian universities. In 1993, the Southeastern Universities Research Association (SURA) and Southeastern Library Network (Solinet) decided to include ETD efforts in regional electronic library plans. Virginia Tech hosted another meeting involving multiple universities in Blacksburg, VA in 1994 to develop specific plans regarding ETD projects. On the technical side, the decision was made that whenever feasible, students should prepare ETDs using appropriate multimedia standards in addition to both a descriptive (e.g., SGML) and rendered (e.g., PDF) form for the main work.

Then, in 1996, the pace of ETD activities sped up. SURA funded a project led by Virginia Tech to spread the concept around the southeastern US. Starting in September 1996, the US Department of Education funded a three-year effort to spread the concept around the USA [2]. The pilot project that had proceeded at Virginia Tech led to a mandatory requirement for all theses and dissertations submitted after 1996 to be submitted (only) in electronic form. International interest spread the concept to Canada, UK, Germany, and other countries. To coordinate all these efforts, the free, voluntary federation called NDLTD (Networked Digital Library of Theses and Dissertations) was established and quickly began to expand [3]. Annual meetings began in the spring of 1998 with about 20 people gathering in Memphis, TN. In 1999 about 70 came to Blacksburg, VA while in 2000 about 225 were in St. Petersburg, FL for the third annual conference.

1.4. Global cooperation in ETD activities

There continues to be rapid growth and development of ETD activities around the world. Whether such efforts arise spontaneously or as extensions of existing efforts, it is hoped that all will proceed in cooperative fashion, so universities can help each other in a global collaboration [4], passing on lessons learned as well as useful tools and information. The mission of NDLTD is to facilitate such progress in a supportive rather than prescriptive manner. Over 100 members joined NDLTD by 2000, including over 80 universities, in addition to national and regional project efforts; international, national, and regional organizations; and interested companies and associations. The only requirement for joining NDLTD is interest in advancing ETD activities, so it is hoped this will help ensure global cooperation.

A number of groups involved in NDLTD are particularly interested in supporting efforts in developing countries. The sharing of research results through ETDs is one of the fastest ways for scholars working in developing countries to become known and have impact on the advance of knowledge. It also is one of the easiest and least costly ways for universities in developing countries to become involved in digital library activities and to become known for their astute deployment of relevant and helpful technologies. The Organization of American States, UNESCO, and other groups are playing a most supportive role in facilitating this process.

1.5. Overview of rest of the chapter

Subsequent sections explain further about ETD activities. Section 2 presents the topic for students. Section 3 discusses issues for university decision makers and implementers of projects on campuses. Section 4 deals with further technical details. Section 5 takes a broader view, raising the level to issues related to launching campus initiatives and training those who may train students. Finally, Section 6 provides a glimpse of future directions.

2. Students

Students are the most important participants in ETD activities. They are the main target of the education effort. They are the ones who learn by doing, and so promote access to the ETDs they prepare to help communicate their research results.

2.1. Why ETDs?

There are many reasons for ETDs. Indeed, if one asks "What are the reasons to not have ETDs?" it is difficult to find any convincing, forward-looking answer. Almost all TDs are produced as electronic documents, and if students know in advance about how to prepare ETDs, then creating their own ETD usually is a very simple process. In addition, there are special benefits that result from ETD creation.

- New genre

  The first benefit is that new, better types of TDs may emerge as ETDs develop as a genre. Rather than be bound by the limits of old-style typewriters, students may be freed to include color diagrams and images, dynamic constructs like spreadsheets, interactive forms such as animations, and multimedia resources including audio and video. To ensure preservation of the raw data underlying their work, promote learning from their experience, and facilitate confirmation of their findings, they may enhance their ETDs by including the key datasets that they have assembled.

  As the new genre of ETDs [5] emerges from this growing community of scholars, it is likely to build upon earlier forms. Simplest are documents that can be thought of as "electronic paper" where the underlying authoring goal is to produce a paper form, perhaps with color used in diagrams and images. Slightly richer are documents that have links, as in hypertext, at least from tables of contents, tables of figures, tables of tables, and indexes – all pointing to target locations in the body of the document. To facilitate preservation, some documents may be organized in onion-fashion, with a core mostly containing text (that thus may be printable), appendices including multimedia content following international standards, and supplemental files including data and interactive or dynamic forms that may be harder to migrate as the years pass by. Programs, applets, simulations, virtual environments, and other constructs yet to be discovered may be shared by students who aim to communicate their findings using the most suitable objects and representations.

- Minimise duplication of effort

  A second benefit of ETDs is a reduction in the needless repetition of investigations that are carried out because people are unaware of the findings of other students who have completed a TD. Except in unusual cases, masters' theses are rarely reported in databases (e.g., very few, except those from Canada, appear in UMI services like *Dissertation Abstracts*). Few dissertations prepared outside North America are reported either. With a globally accessible collection of ETDs, students can quickly search for works related to their interest from anywhere in the world, and in most cases examine and learn from those studies without incurring any cost.

- Improve visibility

  Once ETDs are collected on behalf of educational institutions, digital library technology makes it easy for works to be found. Through www.theses.org, NDLTD directly makes ETDs available, and points to other services that facilitate such discovery. As a result, hundreds or thousands of accesses per year per work are logged, for example, according to reports from the Virginia Tech library regarding the ETDs it makes publicly accessible [5, 6]. As the collection of ETDs available grows and reaches critical mass, it is likely that it will be frequently consulted by the millions of researchers and graduate students interested in such detailed studies, expositions of new methodologies, reviews of the literature on specialized topics, extensive bibliographies, illustrative figures and tables, and highly expressive multimedia supplements. Thus, students and student works will become more visible, facilitating advances in scholarship and leading to increased collaboration, each made possible by electronic communication, across space and time [4].

- Accelerate workflow

  ETDs can be managed through automated procedures honed to take advantage of modern networked information systems. Since the shift to ETDs requires policy and process discussion among campus stakeholders, it is possible to streamline workflow and save time and labor. Checking of submissions and cataloging is sped up, moving and handling of paper copies is eliminated, and delays for binding are removed. The time between submission and graduation can be reduced, and ETDs can be made available for access within days or weeks rather than months.

- Save money

  ETD submission over networks has zero cost, which compares favorably with the charges of hundreds or thousands of dollars otherwise required to print, copy, or publish TDs using paper or other media forms. In many institutions, the networking, computing, and software resources available to students suffice so that students preparing ETDs need make no additional expenditure. Similarly, on many campuses, assistance is available to answer questions and train students regarding word processing and other skills valuable for authors of electronic documents and users of digital libraries. If students elect to use personal computers and acquire their own software to use in ETD creation, these will later be useful in other research and development work, for both professional and personal needs, with low marginal expense specifically required for ETDs. Thus, it is typical that the pros far outweigh the cons regarding students preparing ETDs.

2.2.How to access ETDs

Since in most cases it is in the interest of students and universities to maximize the visibility of their research results, the general approach of NDLTD is to encourage all parties interested to facilitate access to ETDs.

2.2.1. Well known sites/resources for ETDs

Thus, NDLTD runs the Web site www.theses.org, which also has alias www.dissertations.org, as a central clearinghouse for access to ETDs. This site points to various others that support portions of the worldwide holdings of ETDs. For example, the largest corporate archive, with over 1.5 million entries, is managed by UMI, and has most doctoral dissertations from USA and Canada, as well as most masters' theses from Canada, in microfilm form, with metadata available as a searchable collection through *Dissertation Abstracts*. Since 1997 UMI has scanned new submissions (originally from microfilm, later directly from paper) and made the page images available through PDF files. With over 100,000 ETDs accessible through subscription or direct payment mechanisms, UMI hosts the largest single collection of electronic TDs as well as of microfilm TDs.

Other corporations as well as local, regional, national, and international groups associated with NDLTD have Web sites too, such as www.cybertheses.org for the internationl Francophone project or www.dissonline.org. In addition, a number of WWW search engines have indexed some of the ETD collections available so this genre is included in general Web searches.

Some other schemes allow access to ETD collections. Using Z39.50, the "information retrieval protocol", for example, the Virginia Tech ETD collection can be accessed through suitable clients or from some library catalog systems. OCLC's WorldCat service, with over 20 million catalog records, has an estimated 3.5 million entries for TDs. Perhaps most promising is that the global as well as regional and local metadata information about ETDs may become widely accessible through the Open Archives Initiative [7].

2.2.2. Search

As part of the education component of NDLTD, it is hoped that graduate students will become facile with searching through electronic collections, especially those in digital libraries. If we regard managing information as a basic human need, ensuring that the next generation of scholars has such skill seems an appropriate minimal objective. Most specifically, since graduate research often builds upon prior results from other graduate researchers, it seems sensible for all ETD authors to be able to search through available ETD holdings. NDLTD encourages that online resources, self-study materials, individual assistance, as well as group training activities be provided so that graduate students become knowledgeable about resource discovery, searching, query construction, query refinement, citation services, and other processes – both for ETDs and for content in their discipline.

2.2.3. Classification systems and schemes

Considering further the educational mission of NDLTD, it is hoped that students will learn other concepts from the fields of library and information science. As emerging scholars, they should grasp the entire information life cycle that is now being supported through digital libraries [8]. Some of those aspects are considered below. Here we note that manual or automatic schemes are often deployed to categorize or classify documents so they can later be found by referring to an appropriate category. Indeed, when people browse through a collection, they often navigate through a suitable classification system or "concept space" to find likely portions to examine.

There are general classification schemes, such as the Library of Congress Subject Headings, Dewey Decimal Classification, and simpler schemes prepared by UMI and UNESCO. The US National Library of Medicine has MeSH as well as the more extensive UMLS scheme, while for computing ACM maintains the Computing Classification System. Many other services are offered for other disciplines.

2.3.How to learn about ETDs?

Since education is the core of NDLTD efforts, it is important to ensure that a wide variety of mechanisms are in place, for students, with their varying learning styles, to be aided. First, learning by example is facilitated because thousands of ETDs are available that can be consulted, including many in ones own discipline, as well as exemplary or notable works such as those highlighted from www.theses.org. Second, participants in NDLTD typically have online training resources available, such as the Virginia Tech site at http://etd.vt.edu, where general information as well as specific local requirements are addressed. Third, most universities in NDLTD periodically offer workshops to explain about ETD preparation, often tailored to both novice and expert groups. Some of these involve presentations, while others involve hands-on activities. The latter may occur in special classrooms or laboratories, sometimes with scanners and other multimedia devices, to serve specialized as well as common needs. Typically, a campus will have a small cadre of helpers who are knowledgeable about the ETD process, and can resolve unusual problems or address special needs. Though such services are seldom needed at sites where comprehensive computer and information literacy programs are in place, it is appropriate that when ETD submission becomes a mandatory requirement, those who face difficulties should be quickly aided.

2.4.How to prepare an ETD?

Since students learn best by doing, developing their own ETD is the most effective way for the next generation of scholars to be prepared regarding electronic document production. Though details will vary over the years, this practice will ensure that students at any point in time have relevant knowledge and skills appropriate for the available technology.

2.4.1. Overview

Students preparing their ETDs should learn about the entire information life cycle, and work so their research results can be accessible to all interested parties, into the foreseeable future. This objective means that they must consider a variety of concepts and practices, related to document preparation and representation, as well as preservation and access, sketched briefly in the following subsections.

2.4.2. Writing in word processing systems

Most authors today use word processing systems. The most popular is Microsoft Word. Corel WordPerfect, in earlier years more popular, is also widely used. For those working frequently with mathematical expressions, the TeX and LaTeX family of tools (including BibTeX for bibliographies) has replaced the earlier-used UNIX suite of troff, tbl, eqn, refer, and other routines.

Office systems, developed for document preparation and high quality typesetting services, also are appropriate for long and complex works such as ETDs, when authors have requisite knowledge and skills. FrameMaker, PageMaker, Staroffice, and other packages are among the popular solutions.

Because ETDs often are complex documents, that may be developed over the years required to complete a graduate research program, it is essential that students master more than the superficial word processing skills required to produce letters and short reports. They should understand key concepts related to fonts, tables, figures, styles, and document structuring. They should be able to migrate files between versions of software, from one machine to another, to differing types of platforms, and through varying media and networks – while maintaining the message behind their content.

Since ETDs should be usable across time and space, it is imperative, however, that access to them be through suitable interchange formats, rather than transient, unpublished representations produced by particular versions of word processing systems. Accordingly, ETD initiatives have recommended widely used interchange formats like PDF, SGML, XML, and the various schemes preferred for particular types of multimedia content. As was mentioned in Section 1.3, it is preferred to have both a rendered form, like PDF, and a descriptive form, like SGML or XML. However, when that is not feasible, it is better to have one of these forms than to delay implementing an ETD initiative.

2.4.3. Preparing a PDF document

The most popular page representation scheme, a published de facto standard developed by Adobe, now being considered as an international standard, is the Portable Document Format, PDF. Adobe has promised to provide a Reader free of charge into the foreseeable future, which will read current as well as previous versions of PDF, so that archives of documents will remain easily usable. Adobe also provides tools for creating, annotating, and manipulating PDF documents, through its own word processing software, printer drivers, and distilling from PostScript. In addition, some public domain tools work on the published PDF format, such as ghostview.

Adobe's Acrobat software, installed on a Windows, Macintosh, or UNIX platform, allows most suitable documents to be converted to PDF in moments. From word processors such as Word, WordPerfect, and Framemaker, each document portion can be "printed" to the Distiller printer driver, yielding a PDF file. The Distiller converts PostScript files to PDF files. Acrobat software allows multiple PDF files to be assembled into larger PDF files by inserting documents or deleting pages in an existing PDF file.

To avoid problems for future readers, authors should embed all fonts in their documents (when that is allowed). Otherwise, software displaying or printing PDF content will attempt to find a similar font and extrapolate from it, which may cause serious problems. Similarly, authors should use so-called "outline" fonts as opposed to bitmap fonts, so that display and printing can proceed to scale characters as required. Thus, when using TeX or LaTeX, the bitmap fonts commonly found in a standard installation should **not** be used. Instructions at http://etd.vt.edu, for example, explain how publicly available outline fonts can be obtained and substituted. Related problems occur when bitmap images are included in documents and scaled. Vector graphics, special outline font symbols, or object-based image tools should be used instead when possible so that rendering in PDF conveys the correct message. Most problems can be avoided by: planning in advance, following the advice of knowledgeable authors, and testing samples of all types of content that will be in the final ETD.

2.4.4. Preparing for conversion to SGML/XML

Converting from word processing forms to SGML or XML requires more planning in advance, different tools, and broader learning about document processing concepts than does working with PDF. In addition, the end result is a representation that is easier to preserve, more reusable, and supportive of more powerful and effective schemes for searching and browsing. All of these advantages, however, must be weighed against the facts that there are fewer people knowledgeable about these matters, that often tools to help are more expensive and less mature, and that the process may be complicated, difficult, and time consuming. In 2000, there are tens of thousands of ETDs created by scanning (mostly by UMI, but also at sites like MIT and the National Document Center in Greece), thousands converted from word processors into PDF, and hundreds in SGML or XML – illustrating the relative effort required of students to prepare ETDs in each of these forms.

SGML and XML are markup languages (Standard Generalized Markup Language and extensible Markup Language, respectively). Both use tags, normally shown in between "<" and ">" symbols, with names or labels inside, around sections of documents that are thus "marked" or "bracketed". Technically, structures describable this way conform to labelled bracketed grammars. This means that parts are nested within parts, just as subsections are contained within sections. The grammar or structure scheme for a type or class of documents – e.g., book, article, poem, musical score, or dictionary – is specified by a Document Type Definition (DTD). SGML requires a DTD and so is used with well-understood documents while XML, being more extensible while at the same time having stricter rules about closing tags, employs DTDs optionally.

Simple word processing emphasizes layout or what-you-see-is-what-you-get (WYSIWYG) editing. Emphasizing what documents look like is quite distinct from focusing on the logical structure, for which markup schemes are best. Shifting from word processing representations to XML, requires a different way of thinking, a different approach. The problem is harder than producing HTML by exporting from a word processor, since instead of just having a document that looks like the original, it is necessary that the marked-up version itself is correctly tagged.

Some word processors have been extended to facilitate such an approach. Microsoft produced *SGML Author for Word* as an add-on package for Word 95, and new versions of WordPerfect can export content according to markup schemes. Eventually it is likely that most popular word processors will export to XML. Clearly, the resulting markup can surround document sections, headings, paragraphs, lists, figures, tables, citations, footnotes, hyperlinks, and other obvious constructs. In addition, regions with the same style can be tagged. Thus, to allow easy conversion from word processing to markup schemes requires choosing a target DTD and then consistently using document objects and styles so that there is a clear mapping from them to tags.

Conversion from LaTeX is slightly simpler since the TeX approach involves using formatting commands that can be mapped to tags in XML. However, LaTeX does not require strict nesting of commands, so it may not be clear where to place end-tags. Further, LaTeX users may not consistently use the same sequences to designate changes in structure, making translation more complex. Finally, LaTeX coding of mathematical expressions is very difficult to translate to markup schemes for mathematics, like MathML.

Because of the inherent complexity of converting from word processing schemes to markup representations, it is necessary to include steps for checking and correcting converted forms. Parsers can ensure syntactic correctness, so detecting problems is often simple. To ensure semantic correctness, however, manual inspection may be required. A further test would involve rendering the marked-up document, for example to a printed or PDF form, and ensuring that the result suitably matches the output resulting from the original word processing version. In any case, human labor is likely to be needed to correct conversion errors, and presupposes that students understand enough about the process and desired output to accomplish this with facility.

2.4.5. Writing directly in SGML/XML

Since having an ETD encoded using SGML or XML is a desirable result, it also is appropriate to use special word processors or other tools developed for directly producing marked up documents. This is somewhat analogous to the process of directly producing HTML, and no doubt a broad range of tools like those available for HTML will eventually be suitable for XML authoring.

One approach, suitable for experts, is to prepare a text document using a text processing tool or editor like notepad, vi, or emacs. Then all tags must be manually entered, and document structure specified by hand. Alternatively, structure editors designed specifically for XML can be employed. Since the demand for such is smaller than for conventional word processors, currently available tools either are expensive, limited, or not very mature. Further, it is necessary that a syntax checker or parser either be built into the editor, or used in coordination with it, so that errors are quickly corrected.

2.4.6. Integrating multimedia elements

While most training related to word processors covers conventional text documents, perhaps along with simple drawings and inserted pictures, handling of multimedia portions of an ETD is often best managed through separate processes. Tools and special hardware exist for entering and editing complex graphics, images, sound, music, animations, video, and interactive multimedia productions. On most campuses, special laboratories or offices exist that have suitable facilities along with experts who can train seriously interested authors.

However, the learning curve for such is often steep, and students should not lightly choose to include multimedia content unless it really helps them express their research results and/or will lead to skills they desire for the future.

Once produced, multimedia content should be saved in a suitable standard form. International standards like JPEG for images or MPEG for audio and video should be employed so that in future years it will be easy to understand such content. Since such conversions, however, may lead to some losses due to translation and compression, authors may wish to include both the original multimedia content as well as the standard version.

Similarly, as an aid to those interested in reading an ETD, multimedia content may be included in a number of forms. Thus, if a reader wants to view a video but only has moderate bandwidth available to download the ETD, they may be satisfied with a much smaller low-resolution version of a video. At the same time, another reader with a faster connection may prefer to view a high-resolution version. Finally, a reader with a very low bandwidth connection may want to see only a small set of images that are key frames summarizing the video.

Ultimately, multimedia content must be connected to the rest of an ETD. Usually the multimedia information is stored in separate files. These may be referred to or even linked (through hypermedia constructs) to the text or other multimedia constructs. One often appropriate scheme is to have a thumbnail image in the body of the document, which, when selected, links to a corresponding much higher resolution image, and/or video.

2.4.7. Providing metadata – inside, outside documents

In addition to multimedia, documents are often supplemented with metadata (i.e., data about data), typically catalog information. Through a series of meetings scheduled through January 2001, a metadata specification conforming to the Dublin Core [9] standard and tailored to describe ETDs has been under development. The aim is that eventually every ETD will have an associated metadata description following that specification.

Such metadata can be included inside an ETD, making it a self-describing document, especially when XML is used. It is straightforward to encode Dublin Core based metadata in XML, and that can be included near the beginning or in a header portion of an XML ETD. This is similar to the practice with documents encoded in SGML according to the TEI or TEI-lite standards, developed through the Text Encoding Initiative.

Alternatively, and clearly required for previously prepared SGML or XML documents, or documents represented in PDF, metadata can be a separate XML file that is associated or linked with the ETD. Varying approaches to packaging data and metadata together are possible. Note, however, that when metadata is separate, it is then possible for it to be replicated, distributed, and harvested so that ETDs can be more easily discovered without requiring that the actual ETD be examined. Indeed, to allow such processing, even when metadata is included inside an ETD, it is recommended that routines be prepared that can extract the metadata portion to allow separate use.

2.4.8. Protecting intellectual property / how to deal with plagiarism

Though in most cases it is beneficial to share research results, so that others can learn from student studies and give credit to them through citations, it is necessary to provide various types of protection when desired by authors or to deal with potential abuses. Automated schemes can help, such as watermarks, digital signatures, and checksums; these are discussed further in Section 4.2. Programs to detect plagiarism also can be used to compare a new ETD with already available ETDs, ensuring that blocks of identical or similar text are not copied. Further, education, training regarding ethical and professional behavior, and suitable policies can support the guidance of faculty and university staff to promote the spirit of scholarly investigation and collaboration.

2.5.Naming standards

To maximize portability, students should name the various parts of an ETD using the lowest-common-denominator standard for file names, typically the "8.3" form used in old systems like DOS, where a name of no more than 8 alphabetic characters is followed by a period and an alphabetic file type (e.g., pdf, jpg, mpg, txt, xml, sgm). If possible, complex directory structures should be avoided and a simple flat list used, also to ensure portability. Further, references to those names should be relative, rather than absolute, e.g., as etd.pdf rather than c:\documents\etd.pdf or /usr/student/thesis/etd.pdf.

Clearly, each file should have a unique name. Similarly, each ETD in a collection should have a unique and permanent identifier. Since each degree-granting institution can use a unique identifier for their archive, every ETD in the world can have a unique overall identifier made by composing the archive and ETD identifiers.

2.6 How to submit your ETD?

Once a student has prepared an ETD, in most institutions involved in NDLTD, they can submit their work over the internet to a local or regional site for further processing. Following local policies, procedures, and instructions, delivered through training sessions or explained on a Web site, they will typically invoke a Web browser on the computer where their ETD resides. The workflow usually involves them entering a password or other authentication of their identity, filling in a form that provides needed metadata information, and uploading each of the files in the ETD “package”. Since they will supply their email address during this process, they can be notified, by those enforcing quality control standards in the graduate program and library, regarding any corrections or missing data they must supply, as well as when key stages in the approval process are achieved.

2.7 Becoming a researcher in the electronic age

In addition to learning about word processing, electronic document processing, and key concepts related to digital libraries, students also must gain other skills in order to be prepared to be researchers and scholars. They must be ready to meet future challenges of the electronic age, where technology continues to advance, often leading to changes in common practices that may save time or improve accuracy. Caution regarding unproven technology is sensible, but straightforward advances like increases in computing and networking speeds, or decreases in prices of experimental equipment, may be unwise to ignore. Further, innovations may lead to tools dramatically aiding their investigations. Thus, learning to deal with change is part of the wisdom that scholars must develop to survive in the complex modern world.

At the same time, scholars must remained anchored by core values such as honesty, integrity, curiosity, ingenuity, generosity, friendship, diligence, perseverance, and responsibility. They must follow the dictates of society and ethics as well as reason and truth. They should give credit as due to those who have helped them or advanced knowledge in ways related to their work.

With the aid of faculty and colleagues, following departmental and other local and discipline-specific practices, they must choose what type of access is appropriate to the various parts of their ETD, and when. For many the decision will be simple, allowing universal access to the entire work. If they must limit access, it is recommended that they do so for as short a time as possible and for as few parts of the ETD as is necessary, to maximize the amount and duration of access. In general, scholars are rewarded most by sharing their discoveries as widely as possible, but in today’s entrepreneurial world they may seek patent protection in order to have time to commercialize their work, if it involves one of the small number of inventions that are ready for technology transfer. If publishing is appropriate, on the other hand, they should seek to ensure that their ETD is available as well as any related prior or derivative works released in the form of articles or books. In some cases they may be required to delay access to (part of) the ETD (for some period of time). What is most important in all this, however, is that students and faculty honestly confront their responsibilities as scholars, learn key concepts related to intellectual property rights, respect laws and policies, follow contracts and agreements with sponsors and publishers, and strive to achieve balance among the many conflicting opportunities and demands they face. All in all, preparing an ETD should greatly expand the learning experience of graduate researchers, thus helping better prepare the next generation of scholars for the Information Age.

3. Universities

3.1 Why ETDs?

For universities, an ETD program has numerous advantages, in addition to the grand one of helping build a worldwide collection of millions of graduate research reports. First, it is a way for the research carried out in connection with their graduate programs to become visible to large numbers of interested parties around the world. High quality ETDs may add not only to the reputation of the students preparing them, but also to the faculty, research groups, laboratories, centers, departments, colleges, and universities involved. Even on a single campus, other students engaged in research, as well as instructors seeking interesting examples for classes, may benefit from each ETD added to the local collection.

Second, an ETD program may save time, labor, and funds that would be devoted to more conventional processing of paper TDs. If a campus switches from paper to electronic submission, there are savings in library shelf space, binding, shelving, hauling and shipping, and reductions in the costs associated with checking and cataloguing. Based on experience at Virginia Tech, the value exceeds $10,000 per year.

Third, an ETD program helps lead to improvements at universities regarding digital library infrastructure. Though the number of works and accesses are only moderate relative to larger digital libraries and online collections, a full implementation of an ETD initiative constitutes a complete digital library application. Indeed, the planning, training, implementation, and operation of an ETD program can be thought of as a complete digital library case study [10]. It should be easy afterward to undertake other digital library projects. Conversely, if a campus has digital library efforts underway, adding ETD services should be a relatively easy enhancement.

Fourth, and most importantly, ETD programs may raise the understanding on a campus of key concepts. There may be increased awareness of the value of multimedia methods to express research results. There may be more understanding of digital libraries, more support for digital preservation programs, more willingness for authors to submit their works into open archives, and more emphasis on the development of skills related to searching, accessing, and re-using knowledge resources. There may be increased discussion and understanding of issues related to intellectual property rights, publishers, the value of university research, and the various ways in which research results can be disseminated. There may be increased valuation of information literacy, and expanded support for graduate programs.

3.2. How to develop an ETD program?

Campuses interested in ETD programs engage therein when there is sufficient leadership and initiative. If a concerted effort is made, the entire process may be completed in less than half a year, though some campuses may gradually shift toward ETDs over several years.

Typically, an ETD program must be developed as a team effort involving those involved in graduate education, library and archive operations, and computing / information technology support. The relative roles of these three groups, and others involved as per campus situations, depend on local policies, procedures, resources, skills, and initiative. While a particular campus can learn from the experience of active institutions in NDLTD, or work in concert with neighboring or peer institutions as part of cooperative programs, local action is nevertheless needed for this effort that deals with student education and campus infrastructure.

Some universities have a strong graduate program, in some cases run from a graduate school or as part of a division of research and graduate studies. Others have a commission responsible for graduate activities, or control such efforts through a faculty senate or other governance group. In some cases, separate discipline or profession oriented schools or colleges (e.g., a College of Engineering or a Law School) control graduate efforts and manage all activities related to TDs. Accordingly, decisions to engage in ETD programs may be decentralized, and a part of a campus may support ETDs before other groups, or a representative group may deliberate regarding any campus-wide projects. In any case, from the graduate program area the key contributions are to expand graduate education to support the initiative, and to change policies to allow ETDs in addition to, and eventually instead of, paper TDs.

Libraries often are the active party in launching an ETD initiative since they usually receive TDs, catalog them, and make them accessible to local readers or to others through interlibrary loan services. Many libraries also assist students in learning to use digital libraries. They may provide archival services, or there may be a separate campus archive – in any case digital preservation is often of concern.

Computing or information technology groups may run digital library systems, or may support such efforts in the library. Through offsite storage and backup services they may help manage digital preservation activities.

Any of the three groups may run education or training programs so that students understand the local ETD program and develop skills for creating and submitting ETDs. Special support for multimedia is most often provided through computing or information technology groups, though that may be through a special media center or in the library. Control of the overall process often is in the hands of the graduate program, though it may be managed in the library.

By way of example it may be of interest to consider the situation at Virginia Tech. The Graduate School runs the program, setting policies. The Computing Center hosts some of the computers and Web sites involved, though most are in the Library. The New Media Center runs training workshops and supports walk-in students needing help. Students upload their works to a Library computer, running locally developed workflow and database management software (freely available for other campuses to adapt), which allows access by both Graduate School and Library personnel who review and approve submissions for subsequent cataloging. The accessible digital library is run by the Library, which also assumes responsibility for long-term preservation, collecting a $20 archiving fee for this purpose. In the case of doctoral dissertations, UMI is paid with student funds for works to be uploaded into the UMI collection as well. Though there have been minor shifts in responsibility since the time this workflow was put in place in 1996, the whole operation proceeds smoothly, and regular surveys not only support tuning but also show general satisfaction with the program.

3.3. What are the key concerns and their resolution?

Since an ETD program calls for change, there are inevitable complaints and concerns that arise. However, based on the experiences of NDLTD members, there are reasonable solutions for all problems raised [3].

First, there are concerns regarding ownership of intellectual property rights related to ETDs. In most institutions, ownership of rights for an ETD rests with the author. However, in some institutions, the institution itself may claim or request assignment of such rights. When research results reported in an ETD arise through funding by a particular sponsor, conditions agreed to when that funding was accepted may have an effect on the rights related to the ETD. Eventually, though, it will be clear what party or parties own the rights on the ETD, and it will be known if there are any special constraints that must be met. In addition, it should be known who are the stakeholders who will advise about rights management issues, for example, legal counsel, intellectual property rights offices, faculty supervising the research, or colleagues involved in related research.

Second, there is the matter of what access is allowed to an ETD. Such a decision is of concern to the abovementioned stakeholders. They may decide differently for any part of an ETD, since digital library technology can allow separate access controls to be in effect as appropriate for different portions (e.g., a chapter that covers information that appeared earlier in a journal, a chapter submitted for possible appearance in another journal, an image provided for scholarly study and criticism by a third party, or a literature review that discloses no new methods but instead is likely to be of interest to the general public). One decision, promoting scholarly communication, is to make content freely available. Another decision, satisfying desires to limit access to the local campus, may be to restrict access to the university community and its library patrons. Strictest control, such as when patent protection is sought, is to avoid disclosure except to those supervising or reviewing the ETD as required for approval. Note, however, that in the interest of facilitating access, at least in the long term, any of the schemes for control may have a time limit, though possibly allowing renewal.

Third, there is the question of how ETDs relate to publishers. For most students, there are no publications involved, so this is a non-issue. For students in the humanities or social sciences, for example, where advancement often hinges upon publishing a book, usually involving a limited print run, discussion with prospective publishers should proceed prior to deciding about ETD access. Available data suggests that it is very rare for a student to publish a book that is at all similar to their TD, and that there is little evidence that public access to an ETD will hurt future sales of an eventual published book that relates. Nevertheless, students working on a book may decide to limit access to the university community for a reasonable period if so advised by a publisher. On the other hand, when a student works in other fields, such as the hard sciences, they may consult with the publisher of a journal to determine if there is a problem regarding making their ETD publicly available. If their ETD has similar content to an already published article, they should secure permission from the copyright holder for the article, and typically will add an acknowledgement. If they hope that their ETD will lead in the future to a journal article, they may find that publishers have no concern with the ETD being available, or else may be required (for a time) to limit access, typically to the university community. Eventually it is hoped that as NDLTD expands, and ETD programs become better understood, then all publishers (not just those on a list that have notified NDLTD) will see how different the genre are, and will allow free access to ETDs.

Fourth, there is the issue of plagiarism. It is true that if ETDs are readily available then people may copy from them and claim others' works as part of their own. However, search technology makes it possible to detect such copying (even more so than is possible today, where so many theses available only on paper remain unknown to most scholars). Further, TDs are supervised by groups of faculty, who should be knowledgeable about their students' research, and who often carry the authority of honor codes and other strict rules. Thus, students who commit plagiarism may run a terrible risk of detection and severe punishment.

Fifth, there is the matter of cost. Running an ETD program involves personnel to propose, publicize, initiate, refine, and institutionalize the activities. If lessons are learned from those already engaged in successful ETD activities, startup costs can be reduced, and smooth operation can soon occur. If a campus is committed to having knowledgeable graduate students able to prepare electronic documents, who are well prepared to be scholars in the electronic age, there is little extra load needed for implementing an ETD program. Indeed, as was mentioned in Section 3.1, when ETDs instead of paper TDs are required, there should be net savings relative to old processing methods. However, if a paper form is managed in addition to an ETD, or if ETD preparation is by university staff instead of by students, there will be small additional work incurred. Typically, any extra work can be carried by existing staff in connection with their normal duties, and certainly involves no more than the effort of a part-time employee.

3.4 Evaluation

Implementing an ETD program should be accompanied by formative evaluation efforts to ensure that needed improvements and refinements are made as soon as possible. At Virginia Tech, data is collected whenever feasible at workshops, when ETDs are submitted, when people wish to access the ETD collection, and periodically from students after varying lengths of time following graduation. No student has yet reported a problem with a publisher resulting from their submitting an ETD.

Generally, quantitative and qualitative results have been quite positive. Most ETDs are accessed hundreds or thousands of times as opposed to the normal case of TDs that are accessed much less than ten times per year. Most students are in favor of the program. Some have made new contacts or been pleased that their works have been of interest to or impressed others favorably. Workshops (usually for beginners, though sometimes for those interested in advanced topics) are generally found to be helpful.

A very small number of students, typically those with little facility in electronic publishing, are unhappy with the initiative. They argue that they should not be required to submit an ETD, and complain about extra work involved. It is likely, however, that they would oppose any effort making computer and information literacy mandatory.

3.5. Policy Initiatives

University ETD programs must fit into the general schemes of local, regional, and national initiatives for education and scholarly communication. Many of those, such as the NCSTRL project for computing to provide access to technical reports [11-13], function as federations supported by distributed processing technology. NDLTD similarly assumes that the overall collection is composed of a number of repositories, that can be harvested from, or can participate in a federated search service [14]. The organizing principle behind each repository may vary as needed.

Most NDLTD members are individual universities that have elected to join and participate as an institution. Some begin that process by way of a pilot effort in a particular campus sub-unit that is ready to support student submissions before campus-wide infrastructure and policies are in place. On the other hand, some groups of universities join together, building upon related initiatives or practices for collaboration, to develop ETD programs as shared efforts. For example, OhioLINK supports ETD efforts for all interested institutions in the state of Ohio.

In Catalunya, a consortium involving universities and libraries agreed to manage the regional and language-related group of interested institutions.

University Lyon 2 in France and University of Montreal in Canada are cooperating in a Francophone effort to encourage ETD activities in the French speaking world. This is analogous to efforts involving ISTEC and OAS to support efforts in Latin America and Ibero-America. In all these cases, special support by interested organizations, in most cases involving small amounts of funding for programs, has facilitated workshops and training. However, the vast majority of the costs of shifting to ETD programs is carried by individual universities and their staff involved in that work.

At the national level, small amounts of funding have supported launching ETD activities. As was discussed in Section 1.3, regional support by SURA and national funding by the Department of Education led to the initial spread of the concept in the Southeast and then to the rest of the USA. Funding also has supported national programs in Germany, Australia, India, and most recently, through the Mellon Foundation, in South Africa. Generally, such funding is limited in duration, since, as is discussed in Sections 3.1 and 3.2, mature programs are self-sustaining.

While almost all NDLTD-related universities allow free access to works, at MIT a different financial arrangement is involved. Some students prepare ETDs, while others still submit paper TDs that are scanned, yielding PDF files containing page images. Access to the metadata for the entire MIT collection is free, as is display of PDF files on screen, but MIT collects payment through an e-commerce scheme for printing of TDs from its repository. Requests for old TDs not in the electronic collection lead to scanning of those works so they are added to the collection, resulting in a partial retrospective conversion of in-demand work. Of course commercial organizations like UMI, Diplomica.com, Dissertation.com, and others also must have business plans to allow them to provide services to students and universities related to TDs. It must be remembered, however, that the essence of NDLTD is to support education of students, sharing of research results, building university infrastructure, and other causes that only relate indirectly to whatever and however many other access schemes arise with regard to the ETDs that students learn how to produce.

4. Technical issues

Fully implementing an ETD initiative on a campus requires application of the latest technology, since the overall aim is to prepare students and universities to function effectively in the Information Age. In the following subsections a high level portrait is painted of some of the key technical issues.

4.1. Infrastructure

Digital libraries focus on the content dimension of modern information technology that also depends on two other key dimensions: computing and communication. They are made possible, and can operate on the global scale needed for NDLTD, in large part because other forces, such as the growth of the Internet, and the requirements of research and education, lead to sufficient processing and bandwidth.

On many campuses, graduate students have their own computers, or gain access to computers in their research groups, in their departments, in college or campus computing laboratories, in media centers, or in library resource rooms. Most campuses have wireless networks for laptops, or wired local area networks. Student residences may have network connections served by the campus, an ISP, or modems allowing access to a wide variety of local or commercial services. Local networks are connected to regional or national networks or high-speed backbones. Countries continuously increase the bandwidth of their connections to the rest of the global information infrastructure, leading to further improvements in services for students.

Since a typical ETD only requires about a megabyte of storage, it can be managed with inexpensive systems and networks. Only if large multimedia works, such as videos, are included, is it necessary to utilize more significant amounts of storage or bandwidth. Even a large video (e.g., the several gigabytes required for a full movie compressed according to the MPEG-2 standard), though, is not expensive to store. Storage costs now are under $5 per gigabyte, and will continue to shrink by roughly half each year into the foreseeable future. Thus, if as at Virginia Tech students pay roughly $20 archiving fee when submitting an ETD, they will more than cover the storage expense even when submitting extensive multimedia materials. With most campuses collecting less than a thousand ETDs per year, even if the average ETD size increases from 1 to 100 megabytes, the total yearly storage requirement can be managed easily on a PC or small workstation. Similarly, transmitting ETDs over networks only requires comparable resources to downloading a software package over the Web.

More demanding than hardware or software, however, is providing services to the local campus and to other groups involved in NDLTD. Today, a federated search service is available at www.theses.org [14], which provides a moderate level of support by routing queries to the currently small number of sites that allow searching of local collections. Fortunately, it is relatively easy through the Open Archives Initiative [7] for a local campus to make works for which public access is allowed easily accessible through a harvesting protocol. A tiny amount of additional software suffices for a Web server to support harvesting from those locations, so that www.theses.org or other sites can collect all available metadata.

Even at the global scale, if say metadata for ten million ETDs (each probably requiring less than 1 kilobyte of storage) were aggregated, the total storage involved would only be on the order of 10 gigabytes. Thus, providing a centralized search service building upon harvesting from eventually thousands of universities is not infeasible. Further, such size would allow replication at a number of regional sites, increasing reliability and improving performance.

4.2. Production of ETDs

Production of ETDs in NDLTD should be the job of students, supported by university infrastructure. Here we consider some further details that extend the discussion of Section 2.

4.2.1. Overview

Preparing an ETD typically requires common hardware and software readily available to graduate students. Only if multimedia content is included is it necessary to use scanners, audio or video capture devices, or other special input units when converting from analog to digital data. For such content, it also may be necessary to employ special software packages, as might be available and supported in a media center. Further, after producing a desired rendering of key research concepts, it may be necessary to convert to archival standards (e.g., JPEG, MPEG) in order to ensure future use.

To be usable with computers, content must be encoded using some type of representation scheme. Fundamentally that is what happens using any software system that allows manipulation of digital content. To shift from one representation to another it is necessary to import into one form and export into another, or to employ a conversion or translation tool. If large numbers of conversions are involved, or if the translation process is complex, scripts may be used to help automate the process. If space is an issue, conversion may involve compression, to reduce storage or networking transfer costs, followed by eventual decompression, such as when rendering occurs to final display, sound, or print forms. In any case, standards should be followed as much as possible, to facilitate interchange and preservation.

Generally, standards exist for common types of content. Only in the case of unusual, or highly interactive multimedia content, is it likely to be the case that no suitable standards have yet been developed. For example, with packages like HyperCard, AuthorWare, or Director, when special programs or scripts are involved, the only recourse may be to provide a vendor-specific, secret, proprietary file. In such cases it is recommended that to help allow partial preservation into the future, a sequence of screen dumps, exports of the text of scripts or routines, and other partial views or extracts should also be produced and retained.

The bottom line in all this is for students to understand key concepts of content, storage, manipulation, interchange, and reuse so as to be prepared for future work with digital information.

4.2.2. Page Description Languages

The most popular representation of ETDs is inside word processing systems. However, these forms typically involve vendor-specific, secret, proprietary schemes. Accordingly for interchange and preservation conversion is needed to a more standard form. In this subsection we explore further the use of PDF, while in the next subsection XML is considered.

Many modern printers receive data ready to be printed in the PostScript language, developed in the 1980s by Adobe. To increase portability and functionality, Adobe developed PDF in the 1990s. Their Distiller will convert from PostScript to PDF, which is a file format that includes a section containing page image descriptions. Other parts of a PDF file may include hypertext links, images, thumbnail versions of pages, digital signatures, a table of contents or bookmark structure, and other information. PDF, a published standard that has been used by other software companies as well, should become an international standard too.

One noteworthy feature of PDF is that it is scalable, so that those with limited visual abilities may enlarge parts of a document as needed to enhance perception. Further, it supports annotation, so that draft ETDs can have notes added by reviewers to pass on corrections and suggestions. A digital signature feature allows the work to be secured so as to ensure authenticity. Watermarking allows ownership to be asserted so subsequent unauthorized use can be detected. Other tools may allow searching inside a PDF file for particular words or phrases. Doubtless additional capabilities and enhancements will extend its utility, probably helping position it to facilitate some of the operations now feasible with XML.

4.2.3. Markup Languages

ETDs can be interchanged and preserved using SGML or XML. Given current trends, it is most likely that XML will be used, so the following discussion focuses on that scheme; working with SGML would be similar except in some details.

One use of XML is to encode metadata about ETDs. That concept was explored in connection with applying Dublin Core to ETDs at the fall 1999 DC-7 Conference in Frankfurt. Further discussion proceeded at a May 2000 Berlin meeting and at a short meeting at ECDL'2000 in Lisbon in September 2000. It is hoped that consensus will be reached on this matter at a January 2001 meeting to be hosted by OCLC in Dublin, Ohio.

XML also can encode entire ETDs, typically according to a structuring standard or DTD (recall Section 2.4.4). In 1988, the first SGML DTD for

ETDs was developed for Virginia Tech, by SoftQuad. Neill Kipp developed a newer version in 1997. XML versions were later developed at Virginia Tech, University of Iowa, University of Michigan, University of Montreal, and other locations. Any of these structures allows an ETD to be prepared and later searched, displayed, printed, or reused in part. Further, it may be possible to convert most if not all of a work between the structuring described by one DTD and that of another DTD, so at least some portability is ensured. It is hoped that this matter will be explored further by the NDLTD standards committee, which aims to support as much standardization as is feasible given the many requirements involved in allowing graduate students to participate in all disciplines, countries, language groups, and educational settings.

Preparing XML can be done through conversion from word processing systems (e.g., Word or WordPerfect) or formatting schemes (e.g., LaTeX). From a word processor, some well-known interchange form, such as RTF, that can carry style and other information as well as textual content, is usually the export target. Translators that have been trained to convert from particular RTF sequences to XML constructs then prepare an XML document that can be checked with an XML parser and then refined with an XML editor.

XML editors also can be used directly by authors to prepare the entire ETD. This style of authoring may be particularly appropriate for some types of research where many media objects carry the content. For example, this was done with a chemistry ETD prepared at Virginia Tech that used SGML tools to prepare the document skeleton, which referred to scores of VRML and other special files that used virtual reality and other representations to carry the bulk of the message. However, until training about XML and support for it with powerful tools expands, such an approach is likely to require either extensive knowledge or a good deal of assistance by campus personnel.

The final stage of working with XML involves rendering or presenting of research results. Standards like XSL and corresponding tools, along with definitions of how to present each XML construct, allow content to be portrayed in human-readable forms.

4.2.4. Metadata, cross walks, packaging, naming standards

Today, most ETDs are catalogued in a local library. Typically, the data is represented using a MARC (Machine Readable Catalog) scheme, such as USMARC, UKMARC, or UNIMARC. “Crosswalks” or conversion routines exist to convert from one such form to another, or from MARC to XML, or vice versa. For example, Robert France at Virginia Tech developed a MARC to XML converter so that Open Archives sites can export MARC-encoded metadata through XML.

ETDs sometimes are more than a single document supplemented with metadata. When there are multiple parts it is common to store them as separate files that are in a single directory. It is simple to upload each of these files, and for readers to download some or all as desired. Packaging with schemes like tar or zip are a bit risky to employ since they are not highly standard or portable. In the future, digital library packaging schemes may emerge, however, that are more suitable.

Naming of ETDs is another realm for standardization. OCLC's PURL and CNRI's handle schemes allow URN (uniform resource name) methods to attach persistent names to ETDs so that they can be located using them, now and into the foreseeable future.

4.2.5. Post processing

The final stages of production of digital content are usually referred to as "post processing." On occasion, university staff may undertake some conversion to standard forms (usually then saving both the "raw" and converted forms). Typically, though, these stages proceed after all checking and correction is completed, and a final version is received. In the case of ETDs, this involves student submission of the approved version. Only in rare cases will some important correction or addendum be allowed thereafter, which can be handled through typical version control schemes, with suitable approvals recorded.

Protecting ETDs involves several types of special processing. Authenticating an ETD calls for ensuring that it remains unchanged relative to the original submission. By computing a number of mathematical functions over an ETD file, such as parity, checksum, or hash codes, a record can be produced that can be compared with the results of the same computations made over what is assumed to be a proper copy. This type of process is used with digital signatures, which also include certification that a trusted party vouches for the signatures. In the case of watermarks, some image can be overlaid with an image chosen by the property right owner, so that the source and customer of the distribution of a digital object can be proven. In steganography, where data is hidden inside a digital object, arbitrary information may be recorded for later use in prosecuting thieves, in ways that are hard to remove in spite of subsequent analysis or compression. All of these schemes may be deployed when desired by authors, or as standard practice at individual institutions, as needed to ensure the integrity of policies regarding preservation, protection, and rights management.

Further protection is required to account for physical damage, disasters, or other attacks on ETD archives. Copies should be made using various media forms, such as CD-ROM that may have long shelf life and may be immune to electromagnetic forces. Stronger security results from having copies at multiple locations, preferably distant from the master copy. Backups, off-site storage, and mirroring methods provide safety and in the latter case also may help improve access from remote users.

4.3. Dissemination of ETDs

Though providing access to ETDs is not the most crucial part of NDLTD activities, supporting dissemination is an important responsibility.

The first aspect of this involves identifying ETDs. As mentioned, some URN scheme is needed so that a permanent identifier can be given to graduating students that will ensure persistent access thereafter. An ETD can be assigned an ISBN (e.g., as is done by UMI, which considers that it is thus publishing a book) or a DOI (i.e., digital object identifier, often given by publishers). URNs like PURLs or handles can be used (and, indeed, DOIs build upon handle technology). If a university participates in the Open Archives Initiative, then each work is assigned a unique identifier in that archive, and the archive in turn has a unique identifier in the OAI registry.

A second support for dissemination is to have a metadata record for each ETD, which carries any classification and cataloging data available. Whether some type of MARC-based scheme or Dublin Core form is used, some standard interchange mechanism, like MARC transport format or XML, also is required. When possible, the metadata should follow standards developed by NDLTD to support global resource discovery. Typically, in addition to title and abstract, there should be author-assigned keywords, entries according to a discipline-specific classification system, and entries made following more general schemes such as: Library of Congress Subject Headings, Dewey Decimal Classification, UNESCO or UMI categories, etc.

Finally, the metadata records about ETDs, possibly supplemented with the actual ETD content itself, should be used to support resource discovery and access. Typical approaches are explained in the following two subsections.

4.3.1. Databases and information retrieval systems

Managing submission of ETDs and supporting subsequent access can be aided by database management technology. Anthony Atkins at Virginia Tech has developed a number of versions of such software, has refined that and made it portable, and supports its use by many NDLTD members. This has in turn been adapted to multilingual use in Spain and other countries, and to large projects as in the Australian initiative. At MIT, the Dienst software [15] has been adapted instead, while at sites like University of Montreal, Canada and Humboldt University, Germany, other software has been developed.

In addition to managing submission, workflow, and metadata fields with database tools, information retrieval systems are often used to support searching and browsing in ETD collections. In most cases this is done with software used on a campus in connection with other types of searching efforts or in connection with library automation services.

One generous offering by an NDLTD member, VTLS Inc., is to use its powerful library automation software system, Virtua, to support the worldwide initiative, free of charge. VTLS is happy to receive either MARC-type or XML formatted metadata for all ETDs created worldwide, in any language, and to provide a centralized union catalog search service through Virtua. Since the metadata provided should have a unique identifier for each ETD described, this mechanism should provide valuable support for discovering and accessing ETDs.

4.3.2. Searching

Since students learning about ETDs should gain proficiency in searching through digital libraries, it is important that they develop suitable skills. They should understand about data and metadata, and be able to work with metadata records that eventually lead to ETDs. In particular, they should understand the 15 elements in the Dublin Core [9] and how searches can be built using one or more of those. They should understand about classification and categorization schemes, how to browse through thesauri, how to narrow or broaden, how to navigate to related concepts, how to combine elements of a description, and the principles behind set-based or ranking-based retrieval systems. They should understand about full-text searching as well as content-based multimedia retrieval (e.g., of images, sounds, or videos). Further, they should feel comfortable with varying styles of interfaces, searching using queries or examples, schemes involving relevance feedback, and information summarization and visualization mechanisms aimed to enhance their capabilities for finding relevant information. For all this to be possible, universities and others supporting NDLTD should provide powerful services, and ensure that students gain requisite skills with them suitable for effective functioning in the Information Age.

5. Training the trainers

For ETD programs to spread to every graduate student, a vast expansion of NDLTD and the efforts of its many partner groups is required. At this point, a broad program of training those who can train others is needed.

5.1 Motivating universities to participate

Though there are many reasons for ETDs (see Sections 1.1, 2.1, and 3.1), awareness of this situation has not been spread widely. Many universities are unaware of the notion, or have incomplete or inaccurate knowledge. Further, since launching an ETD effort typically requires participation of a number of stakeholders on a campus, the first step in spreading the idea to a university usually involves assembling a sufficiently large group of interested parties and decision-makers, explaining about NDLTD, and clarifying the many misunderstandings that may exist. Once there is understanding, a number of stakeholders are usually motivated to proceed, and if there is suitable leadership and resolve, an ETD program will emerge.

5.2 Tool kits for trainers

Trainers require tools to carry out their work. They should have knowledge and experience from involvement in an ETD program, so they may refer to their own knowledge and have examples at hand to explain concepts and practices. They should study the many resources available through NDLTD Web sites, and use PowerPoint slide shows, papers, news releases, and other materials as needed.

They may wish to load a handy set of tools and files onto a laptop computer that they bring to training sessions. If they will be explaining PDF, they should have Acrobat software and may wish to demonstrate not only accessing a notable ETD, but also may show how to go from a word processing form to PDF. If they are explaining XML, they should have a notable ETD developed using XML, an XML editor, an XML parser, and a tool to render that builds upon XML and XSL. They should have an XML DTD to show, and style files that work with XSL. If they are showing conversion from word processor to XML, they should show the original files with styles, the intermediate (e.g., rtf) file that results, and the output from conversion to XML.

To help address questions and concerns, trainers should develop a set of question-answer pairs, as can be found in Frequently Asked Question (FAQ) files. They may refer to these, and also encourage interested parties to consult them online. Part of that collection may be a set of policies and procedures to follow.

Most important among these are those related to access options and standards. Forms developed at Virginia Tech and other locations to summarize the access options are very helpful for students and faculty to examine, and for policy makers to review and adapt, to account for local needs and attitudes. Similarly, lists of standards to follow, that determine what is supported for preservation and what is covered in training, are very important.

5.3 Teamwork, cooperation, and collaboration

For trainers to be effective, they must leverage their efforts. On a particular campus this means that a local team must be developed. An effective team will have people who reinforce each other, represent campus constituencies, and involve all key stakeholders. There must be effective leadership, and a positive attitude backed with sufficient energy/enthusiasm to ensure progress. The different groups involved must be willing to cooperate, solve problems, adapt solutions, and assuage concerns. As needed, they should draw on others to help, including seeking advice from others involved in NDLTD.

Three schemes exist for providing assistance. First, there is an annual ETD conference at which time hundreds of interested parties share their solutions and learn about advances in technology, training, tools, and techniques. Second, there are sites that have established ETD efforts and offer assistance. Those leading national programs, for example, may serve as centers of excellence, and can be visited or may send representatives to help with onsite training. Finally, there are numerous electronic services that afford assistance. Web sites (e.g., run by NDLTD), listservs (e.g., etd-l@listserv.vt.edu for general discussion, or special lists for particular committee or focus efforts), email, and other mechanisms can be consulted. All in all, cooperation and collaboration allow groups to benefit from the accomplishments and knowledge of others.

6 The future

We conclude this chapter with a brief view of the future of NDLTD.

Since the establishment of NDLTD in 1997, there has been a steady growth in membership. This is likely to continue, or perhaps accelerate. Referring back to Sections 1.1, 2.1, and 3.1 one might expect rapid progress. Indeed, especially given the joining of large groups in the year 2000, such as the efforts in Catalunya, Ohio, and South Africa, the future shows promise.

Yet, technology transfer is slow, and change at universities often slower. Further, since the effects of ETD programs will change the whole future of scholarship, there is likely to be opposition, or at least a considerable amount of resistance from inertia. There may be confusion as corporations enter the scene to profit from the results of sharing led by students and universities. There may be confusion as publishers and students grapple with the many changes in policies and economics that will result from ongoing changes in scholarly communication and library practices. Yet, the ETD program has a clear foundation and strives to prepare students and universities for such changes; as one of the most constructive efforts in that sphere it is hoped that it will engender strong support into the future.

Since NDLTD is primarily an educational program it must necessarily adjust to advances in technology, especially related to electronic publishing, digital libraries, scholarly communication, and dissemination of research. The initiative as a whole, and each university involved, must learn to deal with change, which is one of the key goals. Such change must be balanced with what is feasible for students to learn, what universities can economically support, what will ensure portability, and what will enable preservation. Since NDLTD operates as a federation, now supporting federated search, and in the future enabling harvesting through the Open Archives Initiative, there must be agreement among members to allow interoperation.

Following suitable standards, especially regarding metadata, and providing at least minimal services, such as those called for in OAI, will allow very low cost global services to support local and regional efforts.

In the future, NDLTD plans to offer an increased set of services – not just search but also browsing, annotation, and selective dissemination of information (i.e., routing according to profiles). Searching against millions of works will need to be supported by tools for handling full-text, multimedia content-based matching, query by example, and other approaches. Additional mechanisms for preservation, agreements to enhance performance through mirroring, and flexible handling of works in many of the world's languages will all be needed. Continual evaluation and refinement of services, tailored training and education, and increased sharing and collaboration should help ensure ongoing improvement and eventual fulfillment of the many goals and objectives of ETD programs.

We invite you to learn, participate, and contribute to this cooperative venture!

References:

[1] E. Fox, "NDLTD: Networked Digital Library of Theses and Dissertations", 2000. http://www.ndltd.org

[2] E. A. Fox, J. Eaton, G. McMillan, N. Kipp, L. Weiss, E. Arce, and S. Guyer, "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources," *D-Lib Magazine*, vol. 2, 1996. http://www.dlib.org/dlib/september96/theses/09fox.html

[3] E. A. Fox, J. L. Eaton, G. McMillan, N. Kipp, P. Mather, T. McGonigle, W. Schweiker, and B. DeVane, "Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources," *D-Lib Magazine*, vol. 3, 1997. http://www.dlib.org/dlib/september97/theses/09fox.html

[4] E. A. Fox, R. Hall, N. A. Kipp, J. L. Eaton, G. McMillan, and P. Mather, "NDLTD: Encouraging International Collaboration in the Academy," *Special Issue on Digital Libraries of DESIDOC Bulletin of Information Technology (DBIT)*, vol. 17, pp. 45-56, 1997. http://www.ndltd.org/pubs/dbit.pdf

[5] E. A. Fox, G. McMillan, and J. Eaton, "The Evolving Genre of Electronic Theses and Dissertations," presented at Digital Documents Track of HICSS-32, Thirty-second Annual Hawaii International Conference on Systems Sciences (HICSS), Maui, HI, 1999. http://scholar.lib.vt.edu/theses/presentations/Hawaii/ETDgenreALL.pdf

[6] J. L. Eaton, E. A. Fox, and G. McMillan, "The Role of Electronic Theses and Dissertations in Graduate Education," *The Council of Graduate Schools Communicator*, vol. 31, pp. 1, 1998.

[7] H. Van de Sompel, "Open Archives Initiative". WWW site. U. Ghent: OAI Group, 2000. http://www.openarchives.org

[8] C. Borgman, "Social Aspects of Digital Libraries," UCLA, Los Angeles, NSF Workshop Report, Feb. 16-17, 1996, 1996. http://www-lis.gseis.ucla.edu/DL/

[9] Dublin Core Community, "Dublin Core Metadata Initiative". WWW site. Dublin, Ohio: OCLC, 1999. http://purl.org/dc

[10] E. A. Fox, "The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC," in *Proceedings NIT'99*. Taipei, Taiwan, 1999. http://www.ndltd.org/pubs/nit99fox.doc

[11] J. R. Davis and C. Lagoze, "NCSTRL: Design and Deployment of a Globally Distributed Digital Library," *J. American Society for Information Science*, vol. 51, pp. 273-280, 2000.

[12] C. Lagoze, "NCSTRL: Networked Computer Science Technical Reference Library", Cornell University, 1999. http://www.ncstrl.org

[13] B. M. Leiner, “The NCSTRL Approach to Open Architecture for the Confederated Digital Library,” *D-Lib Magazine*, vol. 4, 1998. http://www.dlib.org/dlib/december98/leiner/12leiner.html

[14] J. Powell and E. Fox, “Multilingual Federated Searching Across Heterogeneous Collections,” *D-Lib Magazine*, vol. 4, 1998. http://www.dlib.org/dlib/september98/powell/09powell.html

[15] C. Lagoze and J. R. Davis, “Dienst: An Architecture for Distributed Document Libraries,” *Communications of the ACM*, vol. 38, pp. 47, 1995.