The Guide for Electronic Theses and Dissertations

Introduction | Universities | Students | Technical Issues | Training | The Future

Introduction | Universities | Students | Technical Issues | Training | The Future

1. Introduction: Purpose and scope of this document, Edward Fox

The UNESCO Guide for Creating Electronic Theses and Dissertations (ETDs) aims to help all those interested in projects and programs involving ETDs. To the extent possible, it has the eventual goal of aiding all students at all universities to be able to create electronic documents and to use digital libraries. It has particular focus on the emerging genre of ETDs, which should enhance the quality, content, form, and impact of scholarly communication that involves students engaged in research. It should help universities to develop their local infrastructure, especially regarding electronic publishing and digital libraries, which in turn build upon networking, computing, multimedia, and related technologies. In so doing, it should promote the sharing of knowledge locked up in universities, and the collaboration of universities spanning across all countries and continents, from North to South, from East to West, from developing to developed, from public to private, and from large to small. The ultimate effect may be one of the world's largest sustainable programs for diffusion of knowledge, across all fields, including science, technology, and culture.

This work should be of interest to diverse audiences. Its various sections (see discussion in section 1.6) are aimed to address the needs of universities (including administrators and faculty), students (including those who wish to create ETDs as well as those who wish to make use of already-created works), and those involved in training or setting up ETD projects or programs. Ample technical details are covered to support the needs of students wishing to apply multimedia methods to enhance their ability to communicate complex results, as well as the requirements of staff building a local support infrastructure to help such students.

This work is a living document that will continue to be updated in connection with the work of the Networked Digital Library of Theses and Dissertations (NDLTD, www.ndltd.org). It was born as a result of the support provided by UNESCO in grants given to Virginia Tech and University of Montreal. It was prepared by an international team of faculty and staff; coordinated by Shalini Urs. Its organization and content are the result of the editorial labor of Joseph Moxley. Its availability in a broad range of languages is the result of teams of translators, some funded in part by UNESCO, and others volunteering their assistance.

While the development of a comprehensive worldwide program of ETDs of necessity builds upon an enormous range of knowledge and experience, it is hoped that this work will suffice as an initial Guide. We hope that those who read some or all of the sections contained herein will build upon the foundation laid out in the Introduction, so they understand the key ideas and are energized to move forward to advance the cause. In keeping with the goals of NDLTD, we hope that people around the globe who are interested in ETDs will help each other, and together transform graduate education, promote understanding, vastly expand access to new discoveries, and empower the next generation of researchers to become effective leaders of the Information Age.

1.1. What are ETDs? Edward Fox

Joining and participating in the Networked Digital Library of Theses and Dissertations, NDLTD is one of the best ways to understand the concepts 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).

There are two main types of ETDs. One type, strongly preferred since students learn (by doing) as they create them, are author-created and submitted works. In other words, these are documents that are prepared by the (student) author (as is typical in almost all cases) using some electronic tools (e.g., Microsoft Word, LaTeX.), and then are submitted in their approved and final electronic form (to their university or agent thereof). Typically, the raw form of the document (e.g., in Word's ".doc" format) is converted into a form that is easy to preserve, archive, and make accessible for future readers (e.g., that follows standards, such as PDF or XML). That form is submitted, typically over a network connection, usually with related metadata (i.e., "data about data", often cataloging information as one might find in a library catalog, including title, year, author, abstract, and descriptors). Once submitted, such ETDs can be "discovered" by those interested, as a result of searching or browsing through the metadata, or by full text searching through the full document (text, and maybe even multimedia components, like images, video, or music).

The second type of ETD is typically an electronic file that is created (usually by university or service company staff) by scanning in the pages of a paper thesis or dissertation. The resulting ETDs are much less desirable than the abovementioned type: they require much more storage space, they do not easily support full text searching, they cannot be flexibly manipulated (e.g., cannot be zoomed in on by those with poor vision), and they don't lead to the student authors learning about electronic publishing (to prepare them for electronic submission of papers, proposals, or other works now commonly required). Nevertheless, such page images can be made accessible at low cost so that those afar can print and read a facsimile of the original paper pages.

In the subsequent discussion, most of the focus is on the first type of ETD mentioned above. However, the second type is commonplace in projects where a retrospective capture of old works is desired, or where a university wishes to share its research, is willing to go to considerable expense in that regard, and is not very concerned with educating or empowering students in electronic publishing methods.

1.1.1 ETDs as a new genre of documents,

Edward Fox

With thousands of students each year preparing ETDs, the creativity of the newest generation of scholars is being continuously expressed as they work to present their research results using the most appropriate form, structure, and content. While conforming as needed to the requirements of their institution, department, and discipline, students should develop and apply skills that will best prepare them 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.

The first benefit is that new, better types of TDs may emerge as ETDs develop as a genre. Rather than being 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 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

---

1.2.1 Minimize duplication of effort, Edward Fox

One 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.

1.2.2. Improve visibility, Edward Fox

Once ETDs are collected on behalf of educational institutions, digital library technology makes it easy for works to be found. Through http://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. As the collection of available ETDs 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.

1.2.3. Accelerate workflow, Edward Fox

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.

1.2.4. Costs and benefits, Edward Fox

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.

1.3 Purpose, goals, objectives of ETD activities,

Edward Fox

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.

1.3.1 Helping Students Be Better Prepared, Edward Fox

Most documents created today are prepared with the aid of computers. Many universities have "writing across the curriculum" programs, to ensure that students can create electronic documents that convey their knowledge and understanding, and demonstrate their ability to participate in the scholarly communication process.

To function as effective knowledge workers, students must go beyond word processing skills that lead only to paper documents. They must learn to work with others, to share their findings by transmitting their results to others. This teamwork makes it feasible to collaborate, to co-author works, and thus to participate in research groups or teams, which are common throughout the research world (at the very least involving a faculty advisor and graduate student author). It also makes it pertinent for students to participate in common activities of modern researchers. Thus, they can be trained to submit a proposal electronically (e.g., as is required by the US's National Science Foundation) and to submit a paper to a conference (where papers are uploaded by authors, and downloaded by editors/reviewers, as part of the collection and selection activities).

1.3.1.1 Helping Students Be Original, Joseph M. Moxley

Originality is a defining feature of academic research. Using new media can empower you to conduct research in new ways. As suggested by the Exemplary ETDs, you can incorporate streaming video, interviews of subjects and settings, which readers can then view, which could be particularly useful for a case study or ethnographically informed research. You can provide different reading pathways for multiple audiences. For example, inexperienced lay audiences can have a more simplified version of your study, whereas more technical audiences can have more detailed analysis and citations. The technology allows you to present your work in new and different ways. For example, in your research you can include audio, video, and animations. You can add spreadsheets, databases, and simulations. You can even create virtual reality worlds.

1.3.1.2 Helping Students Network Professionally, Joseph M. Moxley

Having an ETD helps build your career. Your work is published in a timely manner, visible, and easily accessible. Timely publication makes your up-to-the-minute research instantly available. Upon publication, ETDs immediately become part of the NDLTD and are available for use by anyone having access to the Internet. Visibility allows people inside and outside of the academic arena to see and use your research. Just having an ETD can multiply the number of times your work is read. This exposure increases the possibilities that your work will be cited in others' publications, which adds to your prestige and can help your future advancement. Accessibility makes reading and using your work easy. Instead of having to request and await the arrival of a printed copy, your work immediately displays on a computer screen and can be printed on demand. By using today's communication tools wisely, you can save time and produce a more influential work. You can manage your information so you know where you’ve kept your notes; use powerful search tools to gather information and manage evaluations and revisions; and use interpretation tools for quantitative and qualitative documents.

Additional Resources

Networking on the Network by Phil Agee
This is a 52000 word essay that Phil Agee wrote with the intention of “get[ting] it into the hands of every PhD student in the world.”

1.3.2 Improving graduate education, and quality/expressiveness of ETDs,

Edward Fox

Graduate Education

Around the globe, people have diverse views toward graduate students and graduate education. Some universities have strong advocacy for graduate students, often involving a graduate school and graduate dean. Others have no focus whatsoever on graduate students as a group, with all support thereof assumed to come by way of faculty mentors and advisors. Regardless of the local context; however, few would argue that graduate students should have fundamental knowledge and skills that will allow them to be effective researchers.

ETDs in Graduate Education

Accordingly, the move toward ETDs aims to enhance graduate education in effective researching. Building on the fact that students learn best by doing, this move encourages students to create and submit their own ETDs, thus learning a bit about electronic publishing and digital libraries in the context of preparing a work that is of importance to them. In particular, they need to gain some knowledge and skill in electronic document preparation, understanding not only the superficial aspects of relevant tools, but also beginning to learn about such processes as archiving, preservation, cataloging, indexing, resource discovery, searching, and browsing.

Aids in Communication

Furthermore, once students become exposed to electronic documentation preparation processes, they often become aware of aids to help them communicate. Tools can be utilized to help with figures and drawing, to assist with analysis and graphing, or to support sharing and interactive exploration of data (e.g., through spreadsheets or database management systems). Images can be prepared as thumbnails that go into the body of a document to help the reader, as well as in varying sizes and resolutions to promote in-depth analysis.

Readership and Quality

Beyond these tools and aids, however, is the fact that authors spend more time when their likely readership is large. Students will tend to produce higher quality work, and faculty will demand better writing and clearer presentation of results, if the audience for a work numbers in the hundreds or thousands, as opposed to the common current situation where a mere handful will read the document. With ETDs leading to large numbers of accesses, many students and faculty will work hard to enhance quality beyond that commonplace prior to the advent of ETDs.

1.3.2.1 helping faculty; How do faculty benefit from ETDs?

Joseph M. Moxley

- Each student could develop a bibliography reflecting his or her work, and a collective bibliography would emerge encompassing all of a faculty member's advisees.
- A student's acquired expertise will not completely leave with that student but will remain to help bootstrap new students (and new interests of the faculty member).
- The efforts of students working with a faculty member can be known to a wider audience. This would provide publicity and enhanced visibility for the student and that student's lab and major professor.
- Students who know how to use tools, such as Microsoft Word's tracking or commenting features, are better prepared for future e-publishing; they can use these tools for future collaboration and mentoring, which should save the faculty member time with the reviews and revisions.

1.3.3 Increasing readership of ETDs, communicating research results, Edward Fox and Joseph M. Moxley

Once a student embarks upon the process of creating an ETD, which student will assured of a one or more order of magnitude increase in the number of readers. Such an audience emerges when ETDs are locatable because of searches using popular search engines, and/or are available for other students as well as more advanced researchers to draw upon. Many look for the large bibliographies and comprehensive literature reviews found in ETDs. Others look for in-depth discussion of research methods. Some seek data sets to use for follow-up studies. A growing number of faculty use ETD collections as reference sources, saving space on their shelves or time required to walk to a library. Some also refer to ETDs in classes, or in homework assignments, especially where there are important results and/or clear expressions of concepts and ideas.

Through such ETDs, students can communicate more effectively. Color figures are much easier to attain, in many situations, than those limited to black and white. Complex tables can be built, with sorting and subtotals incorporated, because of software tools. Spreadsheets or simulations help readers gain hands-on familiarity with data and analysis, promoting a deeper understanding. For those studying phenomena that could be characterized with color photos (using digital cameras or scanners), digital video sequences, audio files, medical images, or other digital representations. Thus, expanded and more effective communication of research results can be aided by usage of ETDs.

Once a student embarks upon the process of creating an ETD, that student can be assured of a one or more order of magnitude increase in the number of readers. Such an audience emerges when ETDs are locatable because of searches using popular search engines, and/or are available for other students as well as more advanced researchers to draw upon. Many look for the large bibliographies and comprehensive literature reviews found in ETDs. Others look for in-depth discussion of research methods. Some seek data sets to use for follow-up studies. A growing number of faculty use ETD collections as reference sources, saving space on their shelves or time required to walk to a library. Some also refer to ETDs in classes, or in homework assignments, especially where there are important results and/or clear expressions of concepts and ideas.

At Virginia Tech, for example, many popular theses and dissertations are available to the public electronically. In 1996, there were 25,829 requests for ETD abstracts and 4,600 requests for ETDs themselves; by 1999 (January-August), there were over 143,056 requests for abstracts and 244,987 requests for ETDs. As of October 1999, the most popular ETD at VT had been requested over 75,000 times. (See VT's download statistics at **http://scholar.lib.vt.edu/theses/data/pdatah.htm).**

1.3.4 Helping universities develop digital library services and infrastructure, Edward Fox

Many universities have little or no involvement in digital library efforts. Yet, it is clear that most universities will have digital library efforts as part of, or in addition to, conventional library efforts. Digital libraries are of particular value when distance education is involved, or when a university has a number of far flung sites. They also can promote more flexible access to information (e.g., 24 hours/day, 7 days/week, from home or office, with no blocking because another has borrowed a work).

Establishing an ETD program automatically moves a university into the digital library era. Software available through NDLTD, or from NDLTD members, is freely available that allows a university to develop its own digital library. Setting up that digital library helps the university bring together the personnel and infrastructure required for other digital library projects, too. Though the demands are small, the digital library that emerges will force consideration of almost all of the key concerns of those who work with digital libraries. Also, since various sponsors interested in digital library technology are helping in many ETD projects, there will be a continual enhancement of the digital library services that relate to improving local infrastructure.

1.3.5 Increasing sharing and collaboration among universities and students, Edward Fox

As more and more ETDs become available at universities or through their agents, more and more students will draw upon the emerging digital library of ETDs. This will make it easier for studnets to learn of the work of other students. Contact is quite feasible since many ETDs include email addresses and other contact information, of faculty helpers/advisors at least if it is not feasible to track down a graduating student author. Once contact is made, it is possible to have further discussion of references, application of methodologies, reuse of datasets, and extension of prior work, as well as investigation of remaining open problems. This may lead to friendships, collaborations, joint publications, and other benefits. Such may expand beyond student-student relationships to also include faculty, groups, laboratory teams, and other communities.

1.3.6 Enhancing access to university research, Edward Fox

Once ETDs are available, through diverse means, others may gain access. In particular, such access may be the only recourse open to those in developing countries who cannot afford to make purchases from Proquest, who cannot wait for expensive shipping of copies through interlibrary loan, who cannot attend the myriad conferences that demand the considerable expenses related to travel, or who cannot pay for expensive journals (that only may have short summaries of thesis or dissertation results).

Access may result from word of mouth, from search, from announcements, or by following links or citations. They may occur quickly after a work is made known to those whose works are referred to, or to those who employ Selective Dissemination of Information services. They may occur intermediated by digital library systems or similar mechanisms. They may result when ETDs are referred to journal articles, conference papers, reports, course notes, and other forms.

Numerically, if theses and dissertations are all released as ETDs, the number of works per year may be around the number of journal articles published by university students. The number may be 20-50% of the number of journal articles prepared by the faculty. Regarding students, in most cases, an ETD will be the only publication of an author. If the number who read ETDs is 10 or 100 times the number who read a paper thesis, then it is clear that there will be a significant increase in access to university research as a result of an ETD program.

1.3.6.1 Searching, Edward Fox

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.

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 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.

1.3.6.2 Browsing: classification systems, classification schemes used in different disciplines, Edward Fox

A key method to gain access to ETDs is through browsing. Browsing promotes serendipity, in analogous fashion to when a person looks around in library stacks, picking up and glancing at a number of works, typically ones that are relatively close to each other.

Browsing often involves a researcher in a learning process connecting with the concepts, areas, and vocabulary used in a particular field. A researchers often moves around in "concept space", seeing what concepts are broader and which are narrower, which are related, and which are examples or applications of theories or methods. Thus, in the case of medical works, browsing often encourages researchers to think about diseases, treatments, location in an organism (or human body or subsystem thereof), symptoms, and other considerations. In many fields, browsing involves exploring a taxonomic system, managing an organization chart or hierarchical structure, or moving from a term to a more focused noun phrase.

In many disciplines, there are official classification systems. Some are quite broad, in some cases covering all areas, as is the case when using the subject headings prepared by the US Library of Congress (LCSH), or the Dewey Decimal Classification system (DDC, available from Forest Press, owned by OCLC). However, for in-depth characterization of a research work in a particular field, it is more appropriate to use the category system for the field. In computing, the ACM system is popular. In medicine, MeSH or UMLS are widely used. In physics, PACS is widely used. Gradually, digital library support for ETDs will support browsing using both broad schemes like DDC, as well as narrow schemes like PACS, integrated synergistically.

1.3.6.3 Well known sites/resources for ETDs, Susan Dorbatz

NDLTD runs the Web site http://www.theses.org/ (also under the alias http://www.dissertations.org/) as a central clearinghouse for access to ETDs. This site points to various other locations 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 both electronic and microfilm TDs.

Other corporations as well as local, regional, national, and international groups associated with NDLTD have Web sites too, such as http://www.cybertheses.org/[1] for the International Francophone project or http://www.dissonline.org/ for the German Dissertation online project. 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 (http://www.openarchives.org/).

The German "Dissertation Online" project was undertaken by the Initiative of Information and Communication of the German Learned Societies (http://www.iuk-initiative.org/index.shtml.en).

This project was funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) for 21/2 years (April 1998 - October 2000). The final conference took place in Berlin from October 30-31, 2000 (http://www.dissonline.de/abschlusstagung/index.html). The project worked at an integrative level and aimed towards a German wide initiative to bring scholarly publications, such as dissertations, diploma and master theses which are usually lost in libraries, online. Through the joint work of 6 academic disciplines (mathematics, physics, chemistry, educational sciences, computer science and libraries such as the State and University Library Lower Saxony (SUB Göttingen) and the German National Library (DDB: Die Deutsche Bibliothek)) which took place at several locations (Duisburg, Oldenburg, Erlangen, Berlin Computing Centre and School of Education, Göttingen, Frankfurt) the project was highly successful in Germany and elsewhere. A tight cooperation with the Networked Digital Library of Theses and Dissertations (NDLTD[2]), set up by Edward Fox from the Virginia Polytechnic Institute and State University, USA, was established.

The developments and the movement "DissOnline.de" resulted in establishing a bureau for coordination (Koordinierungsstelle DissOnline) at the German National Library (DDB). All German efforts taken are now coordinated and all developments (tools, guidelines, etc.) are collected by this bureau.

The main tasks of the original project still reflect the problem areas that have to be taken into consideration while setting up local, national or global projects on electronic theses and dissertations:

- Legal Issues
- Document formats
- Metadata
- Retrieval
- Multimedia
- Library issues
- Archiving

The results of these different subtasks are integrated into the different sections that follow for students, universities, technical issues and trainers.

Additional Thesis Collections

1. <a href="http://elfikom.physik.uni-oldenburg.de/dissonline/PhysDis/dis_europe.html">PhysDis</a>
   a large collection of Physics Theses of Universities across Europe

2. <a href="http://www.iwi-iuk.org/dienste/TheO/">TheO</a>,
   a collection of theses of different fields of 43 Universities in Germany,
   in as much as the Theses do contain Metadata

3. <a href="http://MathNet.preprints.org/">MPRESS</a>,
   a large collection of European Mathematical Theses.
   which contains as a subset

4. <a href="http://mathdoc.ujf-grenoble.fr/Harvest/brokers/prepub/query.html#math-prepub">Index nationaux prépublications, thèses et habilitations </a>
   a collection of theses in France in Mathematics
   "

[1] For more information about Cybertheses, please read section 4.3.4.1.3 of this Guide.

[2] http://www.ndltd.org

1.4 Brief history of ETD activities: 1987-2000, Edward Fox

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. 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. 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 arrived in St. Petersburg, FL for the third annual conference.

1.5 Global cooperation in ETD activities, Edward Fox

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 an impact on the advancement 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.

1.6 Overview of rest of the guide, Edward Fox

Subsequent sections further explain ETD activities. Section 2 discusses issues for university decision makers and implementers of projects on campuses. Section 3 presents the topic for students. 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 Universities

By the end of 2001, the number of universities involved in NDLTD was well over 100. Scores of other universities were considering work with ETDs, and hundreds of universities were aware of the concept. NDLTD hopes that forthwith there will be ETD efforts in every country and then in every state/province, eventually in every leading university, soon serving every language group, and ultimately in every college and university. Since NDLTD aims to help, and there are no costs to join, it is hoped that membership will rise to closely match the number that are interested in ETDs.

It is not clear why any university should not become involved in ETD efforts. Once an ETD program on a campus has evolved to the point that ETD submission is a requirement, the effort saves money for the university as well as the students, while providing many important benefits. And, reaching such a point is not hard, if there is a local team, with effective leadership, that has a clear understanding of ETDs and what is occurring elsewhere in terms of support, cooperation, and collaborative efforts.

This section of the Guide explains why and how universities can establish ETD programs, and helps those involved in an ETD program to address concerns and problems that may be voiced. It appears on the one hand that an ETD program has the healthy effect of helping a university to engage in a rich dialog on a wide variety of issues related to scholarly communication. This is important, since we are in the midst of a revolution in these processes, and both students and players must be aware of the situation if they are to manage effectively (and economically, while working in accord with the core values of scholars). On the other hand, ETD efforts have advanced to the point that proceeding with an ETD program, while not as "sexy" as some of the more vigorously funded digital library efforts, does work well, providing real solutions to real concerns, leading to sustainable and beneficial practices. In short, launching an ETD program is a "no brainer" that can quickly advance almost any university to the happy position of having an effective and economical digital library initiative in place, which also can serve as a model for other efforts.

2.1. Why ETD's? Ana Pavani and Joseph M. Moxley

The quality of a university is reflected by the quality of its students' intellectual products. Theses and dissertations reflect an institution's ability to lead students and support original work. In time, as digital libraries of ETDs become more commonplace, students and faculty will make judgments regarding the quality of a university by reviewing its digital library. Universities that incorporate new literacy tools, such as streaming multimedia, will attract students who hope to produce innovative work.

Starting an ETD program is like starting any other project: a need for the results must exist so all those involved will be motivated and committed through all the steps to the end--the moment when ETD's have become a regular and consolidated activity in the graduate programs of the University.

ETD's are based on the joint work of graduate students, mentors, graduate deans, administrative staff, library staff and the IT team. The success of the implementation of the ETD program requires the commitment of all these players plus that of the university's higher administrative officers.

Additional Resources

Moxley, Joseph M. American Universities Should Require Electronic Theses and Dissertations. (*Educause Quarterly, No. 3 2001, pp. 61-63.)*

2.1.1. Reasons and strategies for archiving electronic theses and dissertations, Ana Pavani and Jose Canos

1. ETD's make the results of graduate programs widely known.

2. Graduate programs may be evaluated by the number of theses and dissertations (TDs) produced and by the number of accessible ETD's.

3. In many countries, when financed with public funds, it is expected that TDs will be made public. ETDs are the easiest way to accomplish this.

4. TDs are part of the assets and history of the universities.

5. TDs exist and are published on paper, so why not publish them electronically?

6. TDs are referred by examining committees, a warranty of quality to be published.

7. TDs contain bibliographical reviews.

8. TDs present the methods used during research, thus allowing these methods to be used by others.

9. TDs allow extensions to be identified and undertaken.

10. TDs hold information that will help avoid duplication of efforts.

11. To publish TDs funded with public money is a way of returning the results to society.

12. To electronically publish TDs makes the results known nationally and internationally.

13. To electronically publish TDs makes it less expensive to students, who do not have to print as many copies.

14. Electronically published TDs yield easier and faster access to information.

15. ETDs require less storage space.

16. ETDs can identify and connect national and international research groups.

17. Access to information enhances the quality of TDs.

18. An ETD program introduces digital libraries in the universities allowing other projects to bloom.

19. ETDs are way of sharing intellectual production.

20. Wide knowledge of good quality TDs strengthens the faculty, the graduate programs and the university.

21. Widely known results allow copies to be identified in an easier way.

22. Universities will be able to share knowledge on digital libraries.

23. Access to information enhances the quality of TDs.

The reasons are purposely listed in the order they were presented during the discussion and, no doubt, seem to be quite scattered.

Thus, let's imagine 3 categories of reasons: benefits to students, benefits to universities and benefits to regions/countries/society. The reasons that were listed above can be reorganized and assigned to these categories.

First, there are benefits to students which reflect on the university and on society too. Of the 23 reasons above, 15 are beneficial to students:

24. Access to information enhances the quality of TDs.

25. An ETD program introduces digital libraries in the universities allowing other projects to bloom.

26. Electronically published TDs yield easier and faster access to information.

27. ETDs are way of sharing intellectual production.

28. ETDs can identify and connect national and international research groups.

29. ETDs make the results of graduate programs widely known.

30. Graduate programs may be evaluated by the number of theses and dissertations (TDs) and by the number of accessible ETDs.

31. TDs allow extensions to be identified and undertaken.

32. TDs contain bibliographical reviews.

33. TDs hold information that will help avoid duplication of efforts.

34. TDs present the methods used during research, thus allowing these methods to be used by others.

35. To electronically publish TDs makes it less expensive to students who do not have to print as many copies.

36. To electronically publish TDs makes the results known nationally and internationally.

37. Wide knowledge of good quality TDs strengthens the faculty, the graduate programs and the university.

38. Widely known results allow copies to be identified more easily.

Next, the universities are the focus and some reasons that were listed under benefits to students will appear again:

39. Access to information enhances the quality of TDs.

40. An ETD program introduces digital libraries in the universities allowing other projects to bloom.

41. Graduate programs may be evaluated by the number of theses and dissertations (TDs) and by the number of accessible ETDs.

42. ETDs can identify and connect national and international research groups.

43. ETDs make the results of graduate programs widely known.

44. ETDs require less storage space.

45. TDs are part of the assets and of the history of the universities.

46. To electronically publish TDs makes the results known nationally and internationally.

47. Universities will be able to share knowledge in digital libraries.

48. Wide knowledge of good quality TDs strengthens the faculty, the graduate programs and the university.

49. Widely known results allow copies to be identified more easily.

Benefits that appeared under the previous 2 categories will be listed again in this last category, devoted to regions/countries/society:

50. Access to information enhances the quality of TDs.

51. ETDs can identify and connect national and international research groups.

52. In many countries, when financed with public funds, it is expected that TDs will be made public. ETDs are the easiest way to accomplish this.

53. To publish TDs funded with public money is a way of returning the results to society.

54. Universities will be able to share knowledge in digital libraries.

55. Widely known results allow copies to be identified in an easier way.

Finally, two other reasons for ETDs are:

56. TDs exist and are published on paper, so why not publish them electronically?

57. TDs are referred by examining committees, a warranty of quality to be published.

There were 23 reasons when they were first listed. When they were classified, we got 34. The explanation is simple - many reasons bring benefits to more than one category.

It is not hard to come to the conclusion that ETD's are beneficial to all, and that ETD programs are good and should be considered by universities.

2.2. How to develop an ETD program, Ana Pavani

Some points must be considered before starting an ETD program because they impact on:

- The stage where the project starts;
- The type of training to be provided;
- The technology to be adopted;
- The personnel to be hired, if needed;
- The time frame for the program to be operative;
- The total cost of implementation.

The team implementing the ETD program must be aware that decisions must be made based on legislation, culture, financial conditions, infrastruture and political aspects of their area. So, they must be prepared to analyze the important issues, suggest solutions and have them approved by the propoer authorities. Only when these aspects are considered can the program start.

Many of the following points may not apply to developed nations but can be crucial to developing countries.

The points are grouped in four categories:

- Analysis of the stage of the automation of the library system: Is there an OPAC?

- Does it support the MARC format?

- Does it support digital objects? What types? Can the contents of objects be searched?

- Does it have a Web interface?

- Is it compliant with Protocol Z39.50?

- Is there a project to automate the catalog? Does the system to be used support the MARC format? What is the time frame for the catalog to be automated?

- Can the system to be used can hold a digital library (types of digital objects, Web interface, Z39.50)?

- Number of TDs per year and to date (both on paper and on digital files):

- Per year - this number is important because it influences:

- The number of persons in the help desk;

- The decision on how to start - by some areas to proof concept or all programs;

- The planning of the growth of infrastructure/equipment.

- To date - this number is important because it influences:

- The decision to make the retrospective capture in some areas and/or dates or all TDs;

- The decision to use scanner + OCR in some areas and/or dates or all TDs (on paper);
- The planning of the types and numbers of equipment;
- The planning of the team;
- The planning of the growth of infrastructure/equipment.
- Format presentation of TDs:
- One or many formats? The number of DTDs, templates and viewing filters depends on this;
- Is an update of format needed or desired? The process must be defined - topdown, consensus, etc;
- Is there legislation to comply with?
- Authors' rights and publishing conditions:
- Is there legislation to comply with?
- Is a document of authorization required? If so, define terms and approve with legal department;
- Establish conditions? For black out period, for on campus publishing only, for partial publishing, etc.

The width of the scope of the topics above shows that an ETD program involves many areas of the universisty and all of them must be commited to it. After these points have been addressed, the project of the ETD program may begin.

An ETD program, like any other project, requires that the roles of each player be well defined. All the people involved must be aware of the importance of their work individually and of the interaction that makes up a team. There must be a commitment in all levels, from the highest administration to the graduate students.

In terms of the university, the 3 main players are:

- The Graduate Office;
- The IT/Computer Group;
- The Library.

They will lead the university in deciding upon:

- The formats for ETDs;
- The way to deal with authors' rights and publishing procedures;
- The workflow for the new dissertations;
- The strategy and workflow for retrospective capture;
- The style sheets and filters for visualization;
- The preservation format(s) and procedures;
- The digital library system to be used;
- The identification of the digital documents;

- The relation of the ETD digital library with legacy systems (library, administrative, etc.);
- The support team to be used;
- The training program.

At the beginning of the project, they must have the support of the highest administration to:

- Change the TDs' formats, if necessary;
- Change the culture of mentors, students and administrative staff;
- Solve the problems related to authors' rights;
- Assure preservation and information security of the digital collection (ETD);
- Make sure funds are provided for equipment (HW & SW), personnel and training programs.

All the above topics are fully discussed in the following sections of this Guide. A remark is necessary though, in the last bullet the term SW includes from style sheets to the digital library system.

Once the ETD program is established, the second important step is to assure it will continue and become a part of the university's regular operation. The university must support the program to make sure that:

n Results are evaluated and the necessary corrections are implemented;

n Enhancements are sought and implemented as a consequence of the evolution of technology;

n Funds are provided for equipment (HW & SW), personnel and training.

A good way to start the discussion in the university is by writing a pre-project to be submitted by the 3 players to the high administrative officers. The main topics of the pre-project should include but not be limited to:

- The objectives:
- Main objective
- Secondary objectives
- The benefits to:
- Students
- The institution
- The region/country/society
- The characteristics of the ETD program and of digital library:
- Functional characteristics
- Technical characteristics
- The results to be achieved
- A brief and focused description of the program and the project
- A description of the main steps of the project

- An estimate of the resources (human and material) needed to implement the program
- An estimate of the annual resources (human and material) needed to keep the program in operation
- An estimate of the time frame to implement each step of the project and to have the program operable
- The commitment, responsibilities and actions of each participant involved
- The relation of the program to other programs in:
- The national scenario
- The international scenario

The writing of this pre-project will help the team organize and ascertain everyone's involvement. In addition, the higher administration will be able to decide based on more reliable information, increasing the degree of commitment of all involved in the program.

2.2.1. Scenarios illustrating approaches, schedules and workflow, Edward Fox and José H. Canós

A common workflow at a campus involved in NDLTD is as follows. Each student will use a home computer, or computer in a lab, to submit his or her ETD. His or her work will have been created on the computer he or she is working with, or will have been moved there using some media format, such as diskette or CD-RW. The student will go to a well-known URL (e.g., at Virginia Tech, will start at http://etd.vt.edu and then follow links).

First, students will provide an ID and password that authenticates them. Next, they will enter in metadata about their ETD. When that is complete and checked, they will use their browser to locate each of the files that make up the ETD, so uploading can proceed. Eventually, the entire ETD will be uploaded.

Later, a person in the graduate school will find that a new ETD has been uploaded. He or she will check the submission. If there are problems, he or she will use email to contact the student so that changes/corrections can be made, and the process repeated up through this step. Once a proper ETD is submitted, a cataloger in the Library will be notified. He or she will catalog the work, adding in additional categories, and checking the metadata. Eventually he or she will release the work to allow proper access, according to the student/faculty preferences.

More refined workflow models can be applied. Let us suppose that at the Polytechnic University of Valencia, Spain the process starts from a catalog of ETD proposals, published by faculty members filling in the appropriate form; each proposal may include the title, keywords, abstract, level of expertise required, and other useful information. A student can apply for a number of proposals (specifying an order of preference) filling a form including his/her personal data. A faculty committee makes the final assignment of proposals to candidates. At this point, most of the ETD metadata have been collected, and there is no need to introduce them during the submission process.

2.2.2 The Role of the Graduate School/Graduate Program, John Eaton

This section looks at how electronic publication of theses and dissertations will enhance graduate education. Topics discussed include: improved knowledge of electronic publication technologies, greater access to scholarly information, wider distribution of an author's work, and student and faculty concerns.

Introduction:

The move by Graduate Schools to allow or even require students to submit theses and dissertations as electronic or digital documents (ETDs) creates much excitement, both positive and negative, among the students and faculty who will be affected by this initiative to digitize these important documents. These positive and negative views will no doubt be tempered by increased knowledge of the ETD process and through increased experience in creating and archiving ETDs. At this time in the development of the ETD process, I believe the importance of an open-minded approach to this new way of expressing the outcomes of masters and doctoral research is captured very well in the following statement by Jean-Claude Guédon in his work, Publications électroniques (1998):

When print emerged, universities failed to recognize its importance and almost managed to marginalize themselves into oblivion. With a new major transition upon us, such benign neglect simply will not do. Yet the challenges universities face in responding to an increasingly digitized and networked world are staggering. Universities need a vision allowing them to express their dearest values in new forms, rather than protect their present form at the expense of their most fundamental values.

The ETD initiatives now under way in universities around the world are about bringing fundamental change to our current concept of what constitutes a thesis or a dissertation. In the U.S., this concept has not changed significantly since students first began to submit paper theses and dissertations in our first research universities over 120 years ago. By moving from a paper presentation of research results to a digital presentation, we make available to the ETD author a powerful array of presentation and distribution tools. These tools allow the author to reveal to masters and doctoral committees, to other scholars, and to the world, the results of their research endeavors in ways and with a level of access never before possible.

TOP

Changes in Presentation:

I believe graduate schools and faculty, in the name of maintaining quality, have all too often inhibited the creativity of graduate students by forcing them into a mold to which they must all conform. This is nowhere more evident than in the thesis or dissertation where format restrictions abound. Some graduate schools have special paper with printed margins within which all written material must be contained. Some graduate schools still read and edit the entire text of every thesis or dissertation. Many have thesis or dissertation editors whose reputation for using fine rulers and other editorial devices for enforcing graduate school format are legendary.

I believe that the student must submit a high quality document that is legible, readable, and that conveys the results of the research or scholarship in a manner that is clear and informative to other scholars. The document does not, however, need to be narrowly confined to a specific format if it meets the above criteria. To create a high quality ETD students must be information literate. That is, they must, at a minimum, have a level of knowledge of office software that will allow them to create a document that if printed would result in a high quality paper document. This kind of properly formatted digital document thus becomes the primary construct of the author, rather than a paper document. In conducting training workshops for Virginia Tech students, a number of which are older non-traditional students, we have found that this lack of office software skills is the single greatest impediment to their being able to produce a good "vanilla" ETD—that is, an ETD that has the appearance of a paper ETD, but is submitted as a digital document.

As early 1999 about 80% of Virginia Tech's 1500 ETDs are vanilla ETDs. Accordingly, we have emphasized the development of these skills, which number less than ten and can be taught in an hour, in our student ETD workshops. Once the student has the fundamental skills to produce an ETD, they are ready, if they desire, to move on to more advanced topics for producing a visually and audibly enhanced ETD. Advanced topics include landscape pages; multimedia objects like graphs, pictures, sound, movies, simulations; and reader aids like internal and external links, thumbnail pages, and text notes. Students are not required to use these enhancement tools, but by giving them access to these tools we open creative opportunities for students to more clearly express the outcomes of their masters or doctoral research.

To maintain quality, the student's thesis or dissertation committee must actively participate as reviewers in this process and must be prepared to exercise judgment concerning the suitability of material for inclusion in the ETD. The resulting "chocolate ripple" or in some cases "macadamia nut fudge" ETDs are the forerunners of a new genre of theses and dissertations which will become commonplace in the future.

Whether tomorrow's graduate students are employed inside or outside the university environment, the ubiquitous presence and use of digital information will certainly be a major part of their future careers. For this reason efforts to increase the information literacy are certain to benefit graduate students long after they have used these skills to produce a thesis or a dissertation.

TOP

Valuable Content:

The traditional view is that the doctoral dissertation and less so the masters thesis provides a one time opportunity for the student to do an in depth study of an area of research or scholarship and to write at length about the topic, free of the restrictions on length imposed by book and journal editors. Such writings may contain extensive literature reviews and lengthy bibliographies. They may also contain results of preliminary studies or discussions of future research directions that would be very valuable to the researchers and scholars who follow. Primarily because of restrictions on the length of journal articles, such information exists only in theses and dissertations. I believe this view is correct and should be maintained in the digital thesis or dissertation.

Access and attitudes

The attitudes of students and faculty toward the value of theses and dissertations vary greatly. For the reasons given above some value them highly. Others, particularly some faculty, see them as requirements of graduate schools that have little value. These individuals consider the journal publication the primary outcome of graduate student research. I do not dispute the added value of the peer review process for journal articles and for books, yet I do firmly believe that so long as the scholar or researcher using ETDs as information sources recognizes theses and dissertations for what they are, these documents are valuable sources of information.

Indeed, these information sources have been grossly underutilized because of the difficulty in obtaining widely available, free access to them either through university libraries or through organizations like University Microfilms. If a comprehensive worldwide networked digital library of theses and dissertations existed, I believe the impact and utilization of these sources of information would rise in proportion to the increased access. This view is supported by experience at Virginia Tech in our ETD project. Research done in 1996 by the Virginia Tech library showed that the average thesis circulated about twice a year and the average dissertation about three times a year in the first four years they were in the library. These usage statistics do not include the use of copies housed in the home departments of the students or the usage of dissertations in the University Microfilms collection. Even so, the usage of the 1500 ETDs in our digital library far outpaces the use of paper documents.

Growth in usage has been steady and remarkable. For the calendar year 1998 there were over 350,000 downloads of the PDF files of the 1500 ETDs that were in the VT library. This is over 200 downloads for each ETD in the collection. The distribution of the interest in the ETDs is equally remarkable. The majority of the interest comes from the U.S with inquiries in 1998 coming from the following domains: 250,000 from .edu, 88,000 from .com, 27,000 from .net, 6,800 from .gov, and 3400 from.mil. Inquiries also come from countries around the world including the 8,100 from the United Kingdom, 4,200 from Australia, 7,300 from Germany, 3,900 from Canada, and 2,200 from South Korea. The most accessed ETDs have been accessed tens of thousands of times with many over one thousand accesses. To learn more about accesses see http://scholar.lib.vt.edu/theses/data/somefacts.html

TOP

Publication and plagiarism

When the ETD project began at Virginia Tech, some students and faculty expressed great concern that publishers would not accept derivative manuscripts or book manuscripts from ETDs. For some publishers this concern is legitimate and the ETD project has put into place a system for students and advisors to restrict access to ETDs until after journal articles appear. This system seems to satisfy faculty, students and publishers. Publishers that have discussed this matter with us usually have not expressed concern with the release of the ETD after the journal article is published. One exception may be small scholarly presses that publish books derived from ETDs. These presses view the book as having a sales life of several years after the initial date of publication. In these cases, it may be necessary to extend the period of restricted access well beyond the publication date of the book.

For the longer term, however, it is important that researchers and scholars regain control of their work by becoming more knowledgeable about their rights as original creators and as holders of the copyrights to the work. This requires universities to have active programs to educate their faculty and students about copyright. Publishers also need to be educated to be less concerned about ETDs interfering with the marketability of their journals. This can be done, in part, by an effort on the part of researchers and scholars to educate publishers of their professional journals. They need to help persuade journal editors that ETDs most often are not the same as the journal articles derived from them, and that there is a serious difference because they have not been subject to the stamp of approval that is the result of peer review. As such they should not be considered a threat to the news value or to the sales potential of the journal. It is interesting to note that a Virginia Tech survey of students who had released their ETDs worldwide showed that twenty students had published derivative manuscripts from the ETDs with no publisher resistance to accepting the manuscripts.

It is also noteworthy that the American Physical Society has a practice of sharing electronic copies of preprints of manuscripts undergoing peer review (http://xxx.lanl.gov/). Those that successfully pass peer review are published in the Society's journals. This practice is essentially the same as the practice being proposed for ETDs above.

The risk of plagiarism is next on the list of concerns of students and faculty. We do not yet have enough experience with ETDs to speak authoritatively about this issue. If one thinks a bit about it though, it seems that the risks of exposure of plagiarism will deter such activity. Most researchers and scholars still work in fields where a fairly small group of workers have detailed knowledge of their work. It follows that because of the size of the field and because of the ease of detecting plagiarized passages in electronic documents, the risks of detection will make wide spread plagiarism unlikely.

More disconcerting to me is the closely related concern of researchers and scholars that by reading their students ETDs, other researchers and scholars will achieve a competitive edge in the contest for grants and contracts. Most research in U.S. universities is done in the name of supporting the well being of the nation and is being sponsored directly or indirectly with public tax dollars. There is something wrong with a view that research and scholarship should not be shared among other researchers and scholars for the above reasons. Yet the concern is understandable in today's financially stretched research universities where the competition for promotion and tenure among young faculty is fierce. Similarly, faculty are encouraged to develop intellectual property in which the university claims a share. I'm not sure if we have gone too far down this road, but I am concerned that our obligation as scholars to make our work known to other scholars is being compromised. A result of this compromise is that the goal of scholars to advance knowledge through sharing knowledge may also be slowed.

TOP

How Virginia Tech implemented the ETD Requirement

ETD discussions with the Graduate Dean, the Library, and Ed Fox, a faculty member conducting research on digital libraries, began in 1991. At that time we were exploring the possibilities of optional submission. Shortly thereafter Adobe Acrobat® software for creating and editing Portable Document Format (PDF) files came on the market. This software for the first time provided a tool that was easy to use and allowed documents to be moved between computer operating systems and platforms while retaining the original document formatting. This was a great step forward in increasing worldwide access to information while retaining the original author's formatting style. At this time we began a pilot study to determine if the Acrobat® met our needs. We determined rather quickly that it was the most suitable product for our needs at that time. In my opinion that conclusion holds true today.

We continued discussions with the Graduate School and the Library and in the Fall of 1995 concluded that we would seek to make the submission of ETDs a requirement of the Graduate School. We took a proposal to the Commission on Graduate Studies and Policies for discussion. A degree standards subcommittee discussed the proposal amongst themselves then with ETD team members, Ed Fox from Computer Science, Gail McMillan from the Library, and John Eaton from the Graduate School. In these discussions the expressed concerns dealt with archiving and preservation, the burden to the students and the burden to the faculty and departments. After full discussion, the subcommittee recommended approval of the proposal. The commission discussed and approved the proposal, subject to the following provisions.

n That a student training process be conducted to show students how to produce an ETD.

n That necessary software (Adobe Acrobat®) be made available to students in campus computer labs.

n That the faculty not be burdened by this process.

n That a faculty/graduate student advisory committee be established to advise the Commission on Graduate Studies and Policies on the ETD project.

With these provisions agreed to, the Commission approved a one year voluntary submission period to be used for beginning the student ETD workshops, informing the university community, and development of the infrastructure needed to move to requiring ETDs, after which ETDs would become a requirement in the spring semester of 1997. All went very smoothly while the process was voluntary. Workshops were started, software was placed in campus computer labs, visits were made to departments, articles were published in the campus newspaper, and the advisory committee was formed.

Late in the spring semester of 1997, after the mandatory requirement began, a small but vocal group of faculty, mostly from the life sciences and chemistry expressed a serious concerns about compromising the publication of derivative manuscripts from ETDs made available world wide. While we had a provision for withholding release of ETDs pending publication of manuscripts, the time period of six months was thought to be short. The ETD team responded to this concern by giving the student and the advisor greater control of the access to the ETD through the ETD approval form which can be found at http://etd.vt.edu/. The modifications made to the ETD approval form seem to have satisfied faculty concerns about publication, and since that date the ETD project has operated very smoothly at Virginia Tech and is now rapidly becoming and integral part of graduate education.

TOP

Conclusion

The ETD project has provided the opportunity for fundamental change in the expression of and access to the results and scholarship done by students in research universities around the world. These tools also can easily be extended to the expression of and access to research done by faculty. As scholars, we should not let this opportunity slip by. As Jean-Claude Guédon said "Benign neglect simply will not do".

2.2.3 Role of the Library and Archives, Gail McMillan

ETDs have a very positive impact on libraries because they are an easy way to expand services and resources. With ETDs libraries can evolve into digital libraries and online libraries. Because authors create and submit the digital documents and yet others validate them (i.e., graduate school approval), all the library has to do is receive, store, and provide access. This is not a radical departure from what libraries do normally; only these documents are electronic. Today, many libraries are already handling electronic journals, so ETDs can extend the multimedia resources to the online environment and give every library something unique in their digital resources.

The library can do more, such as improving workflow, reducing the time from receipt to public access, and, of course, one ETD can have multiple simultaneous users. ETDs can be submitted directly to the library server so that as soon as they are approved, they can become available to users, eliminating the need to move the documents from the Graduate School to the Library. This change in workflow can also eliminate the time delay previously caused by bounding and cataloging them prior to providing access.

The most tenuous and highly emotional service libraries provide to ETDs is archiving. Because not enough time has elapsed to prove that digital documents can live for decades in publicly accessible digital libraries, the uncertainty of online archives causes great unease to many. Libraries must, therefore, be careful about security and back-ups.

Another role that the library plays can be to put a prototype in place. While the Graduate School is nurturing the policy evolution among the academic community, there is a model developing to meet the needs of the academy. Establishing an ETD project Web site that documents the evolving initiative, listing active participants, providing a sample submission form and potential policy statements can do this. These statements might address levels of access, copyright statements, and link to existing ETD initiatives.

When a university is adopting policies about ETDs it is helpful to have a place where its students, faculty, and administrators can see what an electronic library of digital theses and dissertations might be. Many graduate students are anxious to participate in an ETD initiative and the library’s Web site can take advantage of their enthusiasm. Invite students who have complete their TDs to submit them electronically. This will build the initial ETD database and test the submission form as well as give the Graduate School personnel opportunities to compare the old and the new processes with somewhat familiar TDs.

The library is also in a position to offer graduate students the incentive to participate. Most libraries collect binding fees so that theses and dissertations can be bound uniformly. Archiving fees can replace binding fees when ETDs replace paper TDs. However, the library may wish to offer to eliminate this fee for the first (limited number of students) who submit ETDs instead of TDs.

ARCHIVING ELECTRONIC THESES AND DISSERTATIONS

The best chance electronic information has of being preserved is when it is used online regularly and continually. As soon as it is not used, there will be trouble remembering the media that produced it and that made it accessible.

As ETDs begin to join the library’s traditional theses and dissertations, it is a good time to align the commitment and the resources to maintain these online information resources over time. A library’s Special Collections Department and/or its University Archives are often responsible for storing and preserving theses and dissertations. Document parallel standards, policies, and procedures for electronic theses and dissertations (ETDs).

The academic departments determine the quality of the work of their students, while the individual thesis/dissertation committees approve the student's work on its own merits. The Graduate School primarily oversees mechanical considerations, the purpose of which is to provide a degree of uniformity, to assure that each thesis or dissertation is in a form suitable for reading and/or viewing online and that it can be preserved. The University Archives ensures long-term preservation and access to this record of graduate students’ research.

With digital materials libraries give access and simultaneously prolong the life of the work, ensure the durability of the present through stability of the means of mediation.

Factors Effecting Archiving

1) Access

The first goal is to have all ETDs online and available all the time from a stable server. If necessary (depending on the capabilities of the server), some ETDs could be moved to a secondary server. Considerations for moving ETDs to a secondary server are usage (ETDs with fewest accesses) or age (oldest ETDs). Formats and file sizes probably would not be a factor in employing a secondary server, though extremely large ETDs may be prime candidates for separate online storage and access.

If it becomes necessary to move some ETDs to a secondary server, programs would be written to trigger migration. Currently "age" would be easier to program, but in the future "usage" (actually, lack of use or few downloads) would be preferable characteristics for migrating ETDs to a secondary server for archiving.

URNs would link migrated ETDs. URNs could be mapped to PURLs at some future date.

2) Security

Data

ETDs that been submitted but not yet approved should be frequently backed-up (e.g., hourly) if changes have occurred since the last back up; otherwise, generate a back-up programmatically every few hours. Make a weekly back up of ETDs in all directories (i.e., all levels of access). Make copies programmatically and transfer them to another server; make weekly back-ups to tape for off-line storage. Retain copies in quarterly cycles and annually archive to a CD-ROM.

Content

Authors cannot modify their ETDs once approved. Exceptions are made with proper approval. Viewers/readers cannot modify or replace any ETDs. Only in extreme circumstances would the system administrator make modifications to an ETD (e.g., when requested by the Graduate School to change the access restrictions or to activate or change email addresses).

3) Format Migration

The library should share with the university the responsibility to guarantee that ETDs will be available both within and outside the scholarly community indefinitely. To keep ETDs reader-friendly and to retain full access will mean migrating current formats to new standard formats not yet known. This will be done through the cooperative efforts of the library (who maintains the submission software, the database of ETDs, and the secure archive) and university computing expertise.

Standard formats should be the only acceptable files approved. Formats recommended for ETDs that may need to be converted to new standards in the future.

Image Formats: CGM (.cgm); GIF (.gif); JPEG (.jpg); PDF (.pdf); PhotoCD; TIFF (.tif)

Video Formats: MPEG (.mpg); QuickTime – Apple (.qt and .mov); Encapsulated Postscript (.eps)

Audio Formats**: AIF (.aif); CD-DA , CD-ROM/XA (A or B or C); MIDI (.midi); MPEG-2; SND (.snd); WAV (.wav)**

Text Formats: **ASCII (.txt); PDF (.pdf); XML/SGML according to the document type: "etd.dtd" (.etd) ETD-ML**

Authoring Formats: **Authorware, Director (MMM, PICS)**

Special Formats: AutoCAD (.dxf); Excel (.xcl)

2.3.1 Intellectual Property Rights, Gail McMillan and Edward Fox

Whether an author is creating an electronic or paper thesis or dissertation, it does not change their rights and obligations under the law. While university policies vary, it is the custom that the person who creates a work is the owner of the copyright. Therefore, the author of an electronic thesis or dissertation is the copyright holder and owns the intellectual property, their ETD. Within the United States, authors have rights protected by law particularly US Code, Title 17, especially section 106. Authors get to decide how their works will be reproduced, modified, distributed, performed in public and displayed in public. An author may use another's work with certain restrictions known as "fair use" (US Code, Title 17, sect. 107). The four factors of fair use that must be considered equally are: (1) purpose and character of use; (2) nature of the copyrighted work; amount and substantiality; and (4) effect. In the United States, libraries are also considered in the same copyright law under section 108. [For further explanations, see

http://www4.law.cornell.edu/uscode/17/ch1.html]

The owner of an ETD, as explained above usually the student, must take direct action if an ETD service is to be provided. The wording that is agreed to in writing, by student authors as well as the faculty working with them, must make clear how ETDs are handled. The wording used at Virginia Tech is one model. In the approval form used for this purpose, the following is agreed:

n The student certifies that the work submitted is the one approved by the faculty they work with.

n The university is given authority, directly or through third party agents, to archive the work and to make it accessible, in accord with any access restrictions also specified on the form. This right is in perpetuity, and in all forms and technologies that may apply.

2.3.2 Publishers, Gail McMillan

A continuing topic of discussion in the ETD community, including Graduate School administrators, research faculty, and librarians, is the question of “prior publication.” That is, whether publishers and editors of scholarly journals view electronic theses and dissertations that are available on the Internet and through convenient Web browsers as being published because they are so readily and widely available.

John Eaton, Dean at Virginia Tech’s Graduate School, surveyed graduate student alumni in 1998 and 1999 and he asked about publishing articles derived from their ETDs. One hundred percent of those who had successfully published had not had any problems getting published because their theses or dissertations were online and readily available on the Internet. By looking at the results Joan Dalton’s 1999 survey of publishers and Nan Seaman’s 2000 survey as well as at Eaton’s surveys of graduate student alumni, the ready availability of ETDs on the Internet does not deter the vast majority of publishers from publishing articles derived from graduate research already available on the Internet. [See "Do ETDs Deter Publishers? Coverage from the 4th International Symposium on ETDs," Gail McMillan. *College and Research Libraries News*, v. 62, no. 6 (June 2001): 620-621. http://scholar.lib.vt.edu/staff/gailmac/publications/pubrsETD2001.html]

Several publishers have also attested to this and these statements are available in a variety of sources such as http://www.ndltd.org/publshrs/index.html. (others?) At the Cal Tech ETD 2001 conference, Keith Jones from Elsevier stated emphatically that his company encourages its authors to link their articles in Elsevier journals to their personal Web sites and also authorizes faculty members’ departments to provide such links. Jones reported that Elsevier understands the importance of getting new authors such as graduate students to publish in his journals early in their careers because they are then likely to continue to publish with the same journal. He also pointed out the publishing in an Elsevier Science journal is an important source of validation for academics so that the subsequent availability of those articles from other non-profit and educational sources is not a threat.

2.3.3 Human Resources and Expertise Needed for an ETD Program, Gail McMillan

While each university's situation with regard to personnel will vary, it has been shown that ETD programs do not require a large contingency of expert professionals to initiate the program or to maintain it. With a portion of the time of one librarian and one programmer, the Virginia Tech library established the ETD Web site and documented the sequence of events that lead to the computer programs for each stage of acceptance, storage, and access, and implemented the procedures for the university.

One step that should not be overlooked is to involve every person that had a role in traditional theses and dissertation handling. This includes, but is not necessarily limited to.

n Graduate School personnel who, for example, receive the TDs as well as the various forms and payments, and who may be responsible for approving the final document

n Library personnel such as the University Archivist, reference librarian, cataloging librarian, binding clerk, and business services personnel who, for example, are responsible for the long term preservation and access as well as those responsible for processing the microfilming invoices

LIBRARY STAFFING: programmer, student assistant, faculty liaison

A programmer may spend one-half to one hour per day during non-peak periods on maintenance and development. During peak periods this person may spend eight hours per day on problems, development, and system improvements.

One or two student assistants are helpful. One student who knows programming may spend a maximum of two hours per week during non-peak times, and up to ten hours per during the periodic submission/approval peaks. Another student assistant would work face-to-face with graduate student authors to train them and to provide assistance with word processing and PDF software. This assistant might also maintain the training materials, handouts, and Web pages, including instructions for preparing and submitting ETDs.

A faculty member from the library would supervise staff; draft policies; prepare budgets; collaborate with system maintenance and developers; and monitor workflow. This person would also be a liaison to faculty, staff, students, departments, and colleges to help them to become familiar with the processing, accessing, and archiving of ETDs. The faculty liaison from the library would also conduct workshops, write articles, participate in graduate student seminars, prepare handouts and Web pages, as well as collaborate with other universities and libraries.

2.3.4. Sources for funding, Australian Digital Theses Program

The aim of the ADT Program is to establish a distributed database of digital versions of theses produced by the postgraduate research students at Australian universities. The theses are available worldwide via the web. The ideal behind the program is to provide access to and promote Australian research to the international community.

The ADT concept was an initiative of 7 Australian university's libraries in association with the Council of Australian University Librarians (CAUL).

The ADT model was developed by the 7 original project partners during 1998-1999. The program was then opened up to all CAUL members (all Australian universities) in July 2000. The original 7 partners will continue to guide and advise the national group in their role as the ADT Steering Committee.

The initial project was funded by an Australian Research Council (ARC) - Research Infrastructure Equipment and Facilities (RIEF) Scheme grant (1997/1998). The ARC is the peak Australian government research funding body and the grant to establish the ADT was a result of a successful application made by the group of 7 above. It was recognised at the outset that recurrent funding was going to be problematic, and that the model to be developed had to take this into consideration. The model developed is essentially self sustaining, with only a small commitment of resources required. The original funds were used to create such a self supporting distributed and collaborative system.

The software used is generic, and designed to be easily integrated at ADT member sites. Once installed, students can either self submit, or seek support free from the library. Theses are mounted on local servers and require a minimum of maintenance. The central ADT metadata repository is searchable and is created automatically from rich DC metadata generated from the submission/deposit form. The metadata is gathered automatically using a metadata gatherer robot. The idea behind the ADT is that producing research theses is normal business for universities and a commitment to include a digital copy to the ADT requires minimal resources. In fact, digital copies are much cheaper to produce that the traditional paper bound versions.

It must be noted that institutional membership and individual contributions to the ADT are voluntary, and will remain so for some time to come.

Each NDLTD member has to apply to its own national, regional and community funding agencies. However, recurrent funding will probably be an ongoing issue unless a commitment is made at the government level to support such initiatives. In the broader context, ongoing funding for an international community body, which the NDLTD now is, will be difficult to achieve. A possible solution would be for the NDLTD to seek Non-Government Organization [NGO] status and thus secure ongoing funding from UN instrumentalities such as UNESCO. Such funding is critical to ensure the good work already achieved by the NDLTD in bringing together a large and disparate number of institutions from across the globe bound by an ideal to promote, support and facilitate open unrestricted access to worldwide research contained in theses, and to share without cost/commercial barriers experiences, support and guidelines with the world community in an open, transparent way.

2.3.5 Costs, Gail McMillan, Guylaine Beaudry, Susanne Dobratz, Viviane Boulétreau

Evaluation of Costs at Virginia Tech

Libraries usually have some of the infrastructure already in place to handle ETDs, especially if they are already providing access to electronic journals or digital images. However, many administrators like to have data about costs associated with an initial budget, and for this reason the following costs for personnel, hardware, and software were established by Virginia Tech’s Digital Library and Archives. Keep in mind that existing personnel and hardware can be commandeered for the ETD prototype and that frequently software is available on the Internet as shareware. This is how the Virginia Tech initiative began in 1995 and continued into the first year of required ETDs, 1997.

[taken from http://scholar.lib.vt.edu/theses/data/setup.html]

These estimates assume that the university/library would adapt existing programs, scripts, Web pages, software, etc. that has been developed by Virginia Tech’s Digital Library and Archives (http://scholar.lib.vt.edu/theses). No additional equipment, software, or staff was necessary for the VT library to begin or to maintain the ETD system and services for the first few years of its initiative. The VT library, however, was not initially responsible for any aspect of training (not the Web site and not the face-to-face instruction), so these costs are not included.

$24,000	STAFF
$36,000	EQUIPMENT
$15,000	SOFTWARE
$65,000	*Total*

The estimated costs below show that for about $65,000 a library could replicate the Digital Library and Archives' ETD initiative if it adapts what VT has already developed for the NDLTD.

STAFFING: $24,000

Programmer: .25 FTE at $6000-$6,600/yr

$26,602/annual; $12.79/hr

Student assistant: .25 FTE at $2900.00-$4500.00/yr

- knowledge of PDF and html desirable; train to use ftp and telnet if necessary
- desirable: programming experience.
- salary depending on skill level at VT would be $6.00 - $10.00/hr

Faculty liaison: .25 FTE @ $50,000 = $12,500

EQUIPMENT--SERVING ETDS: $26,000

server: $11,000

Virginia Tech did not purchase equipment for its ETD initiative; instead the library incorporated this additional responsibility into the original server, a NeXt3.3 running HP. In Sept. 1997 VT purchased a Sun Netra:

RAID disk space: $5,000

tape drive for back-ups : $3500

9 Gb disk drives : $250

Each ETD requires an average of 2.5Mb

CD-ROM recorder : $1000

2 workstations : $5,000

SOFTWARE--SERVING ETDS: $250-$15,000

$3000 Netscape Commerce Server (comes at no charge with server)

$10,000 OpenText LiveLink search engine

VT used freeWAIS for first three years of its ETD initiative.

$1700/yr OpenText 1998 subscription fee for Customer Assistance Program

forward compatible versions of software, monthly product and corp. info updates, access to Product Specialists for technical assistance.

$40 x 2 Adobe Acrobat

$50 x 2 Microsoft Office: including word processor

$150 x 2 Adobe Photoshop

At VT the computer laboratory called the New Media Center had staff, software, and equipment that helped students with every stage of ETD preparation and submission. Without such a facility:

STUDENT SUPPORT: EQUIPMENT

$1000 CD-ROM recorder

$5000	scanner
$3000	2 printers: LaserJet and Color Printer
$500	digital camera
$1200	VCR, DVD
$600	drawing tablet

STUDENT SUPPORT: SOFTWARE

$40 x 2	Acrobat
$50 x 2	Word Processor
$150 x 2	Photoshop

STUDENT SUPPORT: STAFF

- Training and assistance for equipment and software
- Maintenance of training materials: Web pages: instructions and handouts

Cataloging Costs

These may not change from the costs of cataloging traditional theses and dissertations. One advantage is that for the same costs, more information can easily be added to the bibliographic record because of the ease of copy-and-paste features of word processors. For example, include the abstract in online catalog records for ETDs and index this MARC field to enhance findings through keyword searching. Another advantage of ETDs is that there are no longer the fees associated with binding, security stripping, labeling, shelving ($.10/vol. estimated), circulating ($.07/vol. estimated), etc. The Virginia Tech Library saved about 66% of the processing costs because of the greatly reduced handling of ETDs; costs dropped from $12 per TD to $3.20 per ETD.

Evaluation of the HR needed for the ETD Program at Humboldt-University

The estimates are for approx. 15 doctoral dissertations per month written in WinWord, with a usual amount for graphics, tables, literature and nearly no multimedia. These estimations are for rendering XML ETD.

Duty	Personal estimates
Accepting Format Control of Word Applying Digital Signatures Courses for authors one day per month WebPages Production of Information material	1 Computing Center staff + 1 student worker with 20 hrs/week
Conversion to SGML/XML	All in all 1 person
Producing Printing Copies from the PDFs delivered by the students (German law says there have to be 4 high quality paper copies for archiving)	Library staff (20% of time)
Metadata	1 librarian (20% of time)

Management	Approx. 20% of a management person in the library

Evaluation of the HR needed for the ETD Program at Lyon 2

The estimates are for approx. 130 theses a year for rendering XML ETD. We need one full time person for the deposit (verification, edition and printing), conversion and metadata production. A student is necessary on some particular periods (just before the dead-line for academic job appliance).

The training takes us one week a year. The management, communication, development of web pages and pedagogical tools can be estimated as .5 FTE of a professional.

Evaluation of the HR needed for the ETD Program at Université de Montréal

The estimates are for approx. 350 doctoral theses per year written in WinWord, with a usual amount for graphics, tables, literature and nearly no multimedia. These estimations are for rendering XML ETD with HTML version for distribution on the Web.

Duty	Personal estimates

Accepting Format control of Word and LaTeX files Workshop for students WebPages Conversion to SGML/XML	1 FTE technician staff
Developpement and administration of the system Resolving problems of conversion Conception of training tools for workshop	1 FTE professionnal
Metadata (MARC and Dublin Core)	0.25 FTE library technician
Management	0.25 FTE of a manager

2.3.5.1 Processing charges, Edward Fox

Many universities that have paper submissions of theses and dissertations collect funds for processing these works. Usually the funds are collected from students. In some cases they might be collected from grants, or from a sponsor (e.g., from the National Library, with regard to works in Canada).

These funds commonly are called "binding fees" or are given other names designating their use for processing. A typical fee might be $20 or $30 per work.

In the case of transition to ETDs at Virginia Tech, there were changes to these fees. First, in 1996, when submission of ETDs was encouraged but voluntary, the processing fee was waived for those students submitting an ETD instead of a paper document. This provided a monetary incentive to move toward ETDs.

However, in 1997, with a requirement in place, the processing fee was re-instituted. However, it was renamed "archiving fee". Thus, funds were collected from students:

- to help defray the costs of the ETD program;

- to provide a pool of funds to archive and preserve works throughout their lifetime, allowing for migration to new media types (e.g., automatic copying to newer types of storage media, such as from one online disk to a more modern disk), as well as conversion to new formats (e.g., moving from SGML to XML, or from one version of PDF to a newer version).

2.3.5.2 Budgets

Guylaine Beaudry (guylaine.beaudry@umontreal.ca)

The estimated cost of a project to electronically produce and distribute theses varies based on a number of factors such as: the expertise and competence of your team; the technology used; the volume of documents to process; and the cost of living in your country (human resources, computer equipment, communications, etc.). In these circumstances, it is impossible to provide any budgetary estimates, at least without knowing the operating conditions and elements of a particular situation. Giving an estimate, even in the broad sense of an order of magnitude, would be of no use. Nonetheless, this section will allow you to determine your expenditure budget by providing a list of budgetary items that must be foreseen.

The expenditure budget of an electronic theses project involves four principal modules. These are: the start-up; the implementation of thesis production and distribution services,; communications; and student training.

Start-up

Start-up costs are related to what you require to begin a project to electronically produce and distribute theses. Besides the personnel, who constitute the most important element in any electronic thesis distribution project, the start-up costs principally relate to infrastructure and training. Since the network infrastructure is usually provided by the host institution, this expense is omitted from the following table.

Start-up
Human Resources

Professionals	• Adoption/creation of a procedure for processing theses • Development of tools, programmes and scripts • The number of professionals required is linked to the volume of theses to be processed and to the variety of disciplines taught at your institution
Technicians	• Production of theses, assisting students, and routine tasks • The number of technicians required is linked to the volume of theses to be processed
Management	• Supervision • Communication with the university's administration, and with external partners • Day-to-day management
Infrastructure	
Server	• Materials (the actual server, UPS, back-up unit, etc.) • Software (operating system, search tool, etc.)
Server site	Ø This is usually provided by the University
Computer equipment for processing and for management	• Materials (PC, printer, etc.) for all members of the team • A workstation (PC) to test different software and to receive students' files (by ftp) in a secure manner • Software chosen to suit the chosen technology and assembly line (Office suite, Adobe, XML software, HTML editor, etc.)
Training	
Training team members	• Organized training or self-instruction • Manuals and documentation

Table 1– Start-up for an electronic thesis production and distribution project

Implementation of production and distribution services

The resources required to create an assembly line for producing theses are related to the technologies used as well as the expertise and experience of team members. For instance, the choice to create valid XML documents from the files submitted by students, even if many useful XML tools are now available at reasonable prices, necessarily implies larger investments. Nonetheless, it is important to remember that more costly technological solutions may provide significant benefits to other electronic publication projects within the same institution. For instance, they might be useful for publishing electronic journals or for digitizing other types of collections. Table XXX lists the general stages needing attention when establishing a budget, regardless of the technologies chosen.

Implementing production and distribution services	
Production	• Analysis of needs • Choice of a metadata model (including a permanent referencing system) • Integration of the metadata creation or generation process • Testing different software • Planning and testing of the production assembly line • Creation or adoption of a follow-up tool for production work • Formal implementation of the service

Distribution	• Install and set parameters for the search tool (full-text and metadata) • Create and manage the Web site o Create the site’s architecture o Compose the site’s information and home page o Conceive and create the site’s graphic signature o Create the navigation interface • Ensure Web referencing (Web search tools, indexes, other ETD sites, etc.). • Create mechanisms for managing access • Install a tool to measure visits to, and usage of, the site

Table 2 – Implementing production and distribution services

Communication

The communications plan often makes the difference between successful projects and aborted ones. A communications plan must be drafted when the technical processes are in place and the university’s governors approve the project. Worktime and other resources must be budgeted for the creation and implementation of a communications plan. The principal tasks usually required to this end are listed in Table XXX.

Communication Plan

Human Resources	• Drafting the Communications plan[1] • Organizing and holding information sessions with the university's professors, researchers and students • Writing the content of information documents • Meeting with journalists from the university or from the city's newspapers.
Tools for communicating and promoting the project	• Posters, brochures and other documents • E-mails to professors and students • Putting information on-line on the project's Website

Table 3 – Communication Plan

Training Students

Distributing theses on the Web is in itself an effective means of valuing an institution's students and research. Nevertheless, emphasis must be placed on training students so that they can master the tools of document creation that allow them to circulate their research results. The use of new information technologies has become essential for university studies. Whether it involves searching databases and the Internet, or using software to help manage and assess data and present findings, the ability to work easily with these tools will be of daily use for students throughout their future careers. Many universities undertaking electronic theses projects have uncovered serious skills gaps in terms of creating tables, of integrating figures and images, and of using functions that automatically generate tables of contents or lists. It is clear that basic training must be offered to better prepare students for the process of writing important documents like theses.

Student training must be offered through many forms such as workshops, on-line tutorials and personal consultations. One can decide to offer one or another or all of these forms of training. In many institutions, the preparation and offer of training to students results from the collaboration of many different units: the faculty of graduate studies, the libraries and the information technologies services.

Training Students	
Planning and drafting	• Determining the students' current state of knowledge/skills • Analyzing needs • Planning the various training modules • Preparing and drafting the pedagogical tools
Workshops	• Preparing workshops, examples and exercises • Organizing the workshops (selecting classrooms, installing software, etc.)[2] • Hiring assistants (eventually students) for the workshop's "hands-on" period
On-line tutorials	• Adapting workshop content and materials for the Web
Personalized consultation service	• Time taken by the professionals and technicians to answer student questions (by e-mail, telephone, and eventually in person).

Table 4 – Training Students

[1] It is often useful at this stage to get help from a communications agent. There are often agents in our universities, for instance in the Department of Communications.

[2] Many formulas exist concerning the type of personnel who should lead the workshops for students. The leaders can be students, contract teachers, or people from the electronic thesis team (professionals or technicians). This last approach seems to give the best results. However, hiring students for the workshop's "hand-on" period provides much-needed assistance.

Plagiarism, Jean-Claude Guédon

The issue of plagiarism often arises among the arguments used to express skepticism with regard to putting theses online. In short, many people tend to think that because a digitized thesis is easily copied in part or in whole, it can be easily plagiarized. Consequently -so goes the reasoning - it is better to keep theses offline.

The argument is largely false and can be refuted fairly easily. To begin with, it is easy to recall that the invention of the **Philosophical Transactions** (1665) by Henry Oldenburg, the Secretary to the Royal Society in London, was motivated by the issue of intellectual property. Oldenburg reasoned that if the research results of Scientist X. were printed in a journal (after being certified as being of good quality and original) and that journal was made widely available through the multiplication of copies, then Scientist. X would have a better chance to lie ownership claims than if he/she held back these results. By apparently giving away the results of his/her work, a scientist ensures his/her intellectual property most effectively. The ability to compare new results to already published work makes plagiarism a very risky business at best.

Theses are not so well protected at present. Widely dispersed across many institutions in many countries (and languages), they are so poorly catalogued on a national or international basis that they often disappear from sight. This means that someone taking the time to read a thesis in a remote university in a country where the cataloguing is poorly organized may well be able simply to use that thesis and make it pass for one's own. Occasionally, such cases emerge in the literature, even in the United States despite the fact that the cataloguing of theses is most advanced in that country.

The paradox of placing theses on line, especially if these theses can be harvested through some technique that involves full text searching can help identify analogous texts rather easily. As a result, far from placing the digitized theses at risk, putting them on line in a manner that optimizes their access, irretrievability and, therefore, visibility, offers a very efficient way to protect intellectual property and prevent plagiarism. In fact, it would probably be relatively easy to design software that could make periodic sweeps through inter-operable theses collections according to ever more sophisticated algorithms in order to ferret out such possible forms of plagiarism. With many languages involved, it is clear that no perfect solution will ever appear; however, those theses available online will be more protected than theses that remain poorly catalogued and are not readily available outside the institution from which they are issued.

In effect, putting theses on-line amounts to rediscovering Oldenburg's wisdom when it comes to scientific intellectual property. It provides what could arguably turn out to be the best deterrent to plagiarism, wherever it may arise. The more theses appear on line, the fewer will be the chances of carrying on successful plagiarism.

2.4.1 An Introduction to Assessment and Measurement, Joan Lippincott

I shall consider assessment to include the gathering of information concerning the functioning of students, staff, and institutions of higher education. The information may or may not be in numerical form, but the basic motive for gathering it is to improve the functioning of the institution and its people. I used functioning to refer to the broad social purposes of a college or university: to facilitate student learning and development, to advance the frontiers of knowledge, and to contribute to the community, and the society.

(Alexander W. Astin, *Assessment for Excellence*, 1991, p. 2)

Electronic theses and dissertations are not only products of student research, but also marks of the students' preparation to become scholars in the information society. In pursuit of this broader social purpose, this section of the *Guide* focuses on two separate but related topics: *assessment* and *measurement.* Both are important components of an institutional ETD program. The role of assessment in an ETD program is to understand whether the *goals* and *objectives* of the program are being met, while issues of measurement focus on the *production* and *use* of an institution's ETDs.

Assessment of a program's goals and objectives yields information that may be of great value to policymakers and administrators. Higher education institutions are increasingly asked by their boards, by state legislators, and by federal government agencies to demonstrate the effectiveness of their programs, which has led to the development of many assessment programs on campuses. Assessment is often used to provide trend data or to assist in resource allocation, with data helping to make a case for the viability of an ETD program or to determine which parts of the program require the most (or least) resources. Ideally, an ETD assessment program will have ties to overall campus assessment activities.

However, whether or not accountability to administration or government is a driving factor in developing assessment for an ETD program, any new effort can benefit from a systematic way of measuring what it is achieving. Such measures as the number of ETDs produced at an institution each year or the number of times a dissertation has been downloaded from the institution’s web site are frequently cited to demonstrate the success of an ETD program.

The goals of this section of the *Guide* are:

1. to encourage those involved in developing and implementing ETD programs to think at early stages about broad questions of assessment, and

2. to familiarize individuals working on measurement of ETDs with developing national and international initiatives that are developing standard ways of measuring the use of electronic information resources.

2.4.2 Types of Assessment

Joan Lippincott

Assessment takes two basic forms, based on the point at which the evaluation is done:

- *Formative evaluation* takes place during the development of a program. It is used to help refocus activities, revise or fine tune a program, or to validate the success of particular activities. Formative assessment in an ETD program may help diagnose whether training sessions are useful to students, whether the submission process worked efficiently, whether the students learned about electronic scholarly communication issues, or whether the students find what they learned in order to produce an ETD helpful in their employment.

- *Summative evaluation* is used to examine the program at its completion or at the end of a stage. Since an ETD program is ongoing, summative evaluation can be used on an annual basis to yield statistics that can be compared from year to year.

2.4.3 The Assessment and Measurement Process, Joan Lippincott

This chapter of the *Guide* is intended for those who have responsibility for implementing an institutional or departmental ETD program. The process of creating a successful assessment and measurement requires *planning, creating goals and objectives, choosing types of measures,* and *deciding what type of data to collect.*

Planning

Some campuses form an ETD committee or team, consisting of representatives from the Graduate School, the faculty, the library, the computing center, and other relevant units on campus. As part of their overall planning for the development of an ETD program, the committee should make explicit their goals for assessment and measurement of the program and put in place mechanisms to collect data related to the goals. The data collection could be gathered from web statistics packages, from student surveys, or from interviews, depending on what type of information meets the assessment's goals.

Several units of the university—including academic departments, the graduate school, information technology services, and the library—are involved in any ETD program. In developing an assessment plan, it is useful to think convergent. Are there particular things that would be useful for several of these constituents to know? This will leverage the value of the assessment and may allow for joint funding and implementation, thereby spreading the costs and the work.

During the planning process, the ETD committee may focus on a variety of issues, including:

- the impact that the ETD program is having on the institution's reputation;
- the degree to which the ETD program is assisting the institution in developing a digital library; and
- the benefits to and concerns of students and advisors participating in the ETD program.

Once the focus of the assessment and measurement activities is identified, the ETD committee should assign responsibility for development and implementation to another team. This team should include assessment and measurement experts from institutional units such as an institutional planning office, a survey research institute, or an instructional assessment office.

Creating Goals and Objectives

- A necessary prerequisite to assessment is a clear understanding of the ETD project's goals. Each institution's assessment plan should match the goals and objectives of the institutional ETD program, which may have a broader scope than the simple production of electronic content.
- If an ETD program does not have clearly defined goals, an excellent resource is the NDLTD web site. This site includes the goals of the NDLTD, which can be adapted to local needs. These goals include:

- Improving graduate education
- Increasing availability of student research

- Lowering costs of submission and handling of theses and dissertations
- Empowering students
- Empowering universities
- Advancing digital library technology

Within each of these broad areas, many types of measures can be developed to help evaluate whether the ETD program is succeeding.

Choosing Types of Measures

Institutions have many choices in what they measure in an ETD program. After aligning the assessment goals with the institution's goals, the assessment plan must describe what types of measures are needed for various aspects of the program. McClure describes a number of categories of measures, including those that focus on extensiveness, efficiency, effectiveness, service quality, impact, and usefulness. (Charles R. McClure and Cynthia L. Lopata, *Assessing the Academic Networked Environment*, 1996, p. 6)

An *extensiveness measure* collects data on such questions as how many departments within the university are requiring ETDs or how many ETD submissions are made each year. This type of data lends itself to comparison, both as trend data for the individual institution and in comparisons to peer institutions.

An *efficiency measure* collects data to compare how life cycle costs of ETDs compare to those of print theses and dissertations. For example, Virginia Tech includes such a comparison on its Electronic Thesis and Dissertation Initiative web site.

Effectiveness measures examine the degree to which the objectives of a program have been met. For example, if an objective of the institution's ETD program is to empower students to convey a richer message through the use of multimedia and hypermedia, data can be collected that displays the proportion and number of ETDs employing such techniques by year. If an objective of the program is to improve students' understanding of electronic publishing issues, the institution can measure such understanding prior to and after the student produces an ETD.

Measures of service quality examine whether students are receiving the training and follow-up assistance they need.

Deciding What Data to Collect

Frequently, discussions of how to assess electronic information resources are limited to defining ways of counting such things as searches, downloads, and hits. These measures are certainly useful, but they provide a limited view of the overall value of electronic information resources.

Many vital questions cannot be answered with statistics about searches, downloads, and hits: Can users access more information than in the past due to availability of information resources online? Has the availability of electronic information resources improved individuals' productivity and quality of research? Has the availability saved them time?

Collection of data for assessment should be designed to answer some of these questions, to address the educational goals embedded in an ETD program, and to gauge whether or not those goals have been achieved.

Decisions about data collection are also informed by the institutional mission and goals. For example, if the institution is interested in increased visibility both nationally and internationally, then statistics on downloads of ETDs by country and institutional IP address could be useful. Examining who is using an institution’s ETDs by country and by amount of use would also be a valuable gauge of impact. As recommended systems develop, this area may grow in importance.

There are a number of questions related to users that are possible targets for assessment. These include:

n Are students achieving the objectives of the ETD program?

n Are students using tools such as Acrobat appropriately and efficiently?

n Has the availability of student work increased?

n Do students have an increased understanding of publishing issues, such as intellectual property concerns?

In addition, a number of questions related to student satisfaction could be addressed in data collection plans. These may include:

n Were students satisfied with the training or guidance they received to assist them with producing an ETD?

n Did the availability of their dissertation on the web assist them in getting a job?

n Are they using the technology and electronic authoring skills they learned in their current work?

Usefulness to students may be a factor of the availability of their ETD on the web, assisting employers in gauging their area of research and the quality of their output. Availability may also lead employers to contact students for openings that require a particular skill set. And students may find that the skills of preparing an ETD and the framework of issues associated with the ETD, such as intellectual property issues, is useful in their places of employment after graduation.

The usefulness of an ETD program to students, faculty, and others may be an important factor to measure in order to gather data that can be conveyed to administrators and funding agencies. This data might best be collected six months to a year after the completion of the ETD.

Finally, the ability for faculty and students around the world to easily examine the dissertation and thesis output of a particular department may provide a new dimension to rankings and ratings of graduate departments. Monitoring national ratings in the years pre- and post-implementation of an ETD program could be useful, although may be only one factor in any change in ranking or rating.

2.4.4 Useful Models for Measuring Production and Use of ETDs, Joan Lippincott and Jose H. Canos Cerda

In addition to assessing some of the programmatic goals of the ETD program, institutions will want to have some basic assessment measures in place to document the production and use of their ETDs. The work done at Virginia Tech's Electronic Thesis and Dissertation Initiative can provide a model for other institutions. Using web statistics reporting software, Virginia Tech monitors a number of measures for their ETDs, including:

- the availability of campus ETDs
- multimedia in ETDs
- which domains within the United States and abroad are requesting ETDs
- requests for PDF and HTML files
- distinct files requested
- distinct hosts served

- average data transferred daily

In compiling its counts, Virginia Tech eliminates everyone working on ETDs at the institution, including the Graduate School. They try to eliminate repetitive activity from robots and other sources of that type as well. Virginia Tech is also working on the compilation of an international count of ETDs produced in universities.

Institutions must decide whether they will report their ETD collections and usage separately, in conjunction with other campus web site usage, or in conjunction with other electronic resources managed by the library. Now, when practices and standards for gathering these statistics are evolving, institutions may need to collect the information and report it in conjunction with more than one type of related collection. In any case, institutions should keep abreast of national and international initiatives that are seeking to define and standardize statistical reporting of the number and use of electronic information resources.

Several projects currently focus on collection of statistics related to information resources:

The Association of Research Libraries (ARL) collects statistics from libraries of large research universities in North America and has been working on several initiatives that explore data collection related to digital information. In particular, one of its New Measures initiatives, the E-Metrics Project, is recommending a particular set of measures and defining their collection.

The International Coalition of Library Consortia's (ICOLC) has created Guidelines for Statistical Measures of Usage of Web-Based Indexed, Abstracted, and Full-Text Resources. They encourage vendors of electronic information products to build into their software statistical report generation that will meet the ICOLC Guidelines, promoting comparability of products.

In Europe, the EQUINOX Project, funded under the Telematics for Libraries Programme of the European Commission, addresses the need to develop methods for measuring performance in the electronic environment within a framework of quality management. Their products include a consolidated list of data sets and definitions of terms.

Library Statistics and Measures, a web site maintained by Joe Ryan of Syracuse University, also provides a useful set of links to resources on library statistics and measures.

Another important type of post-processing is the extraction of statistical information from metadata sets. For administrative purposes, institutions may be interested in the number of ETDs supervised by each professor, the keywords most used , the month(s) in which more ETDs are submitted, etc. Usually, relevant metadata are extracted from the ETD database and processed using specialized tools like Microsoft Excel. The access to the database can be done using either ODBC drivers or specialized middleware utilities.

2.4.5 Statistics and Usage, Gail McMillan

Gathering data about ETDs can be done through online surveys and log file analysis. Online surveys can also be used to gather data from graduate student authors submitting ETDs and from ETD readers accessing them. An easy to use online survey system is available to NDLTD members at http://lumiere.lib.vt.edu/surveys/ Ask graduate student ETD authors a variety of questions about the process as well as about support services and use this data to improve resources and services. Query readers to discover their reasons for accessing ETDs, the results of their use of ETDs, as well as to improve digital library resources and services.

Questions for student authors may be designed to improve the process of preparing and submitting electronic documents. Sample questions might include:

1. While preparing your ETD, where did you find answers to your questions?

2. If you consulted the VT ETD Web site, please indicate if the site was useful.

3. If you attended an ETD workshop, please indicate if you found the workshop useful.

4. If you used a [particular] computer lab, please indicate if the staff was helpful.

5. How many ETDs did you consult while preparing your ETD?

6. Compared to what you expected, how difficult was it to create a PDF file?

7. My computer is a [PC, Mac, Unix, other].

8. Where were you when you submitted your ETD?

9. Compared to what you expected, how difficult was it to submit your thesis/dissertation electronically?

10. Within the next 1-2 years, what do you intend to publish from your ETD?

11. If you restricted access to your VT ETD, on what did you base your decision?

12. Please include any comments or questions that you have about ETDs.

Usage

In the online environment, usage is similar to but not equal to library circulation and re shelving statistics. Nevertheless, to report usage, institutions should also report numbers of available ETDs so that downloads can be viewed in the context of what was available at a given point in time.

It would be beneficial if all libraries providing access to ETDs would capture and report similar data about usage, but at this time we are not doing this. Gail McMillan has periodically surveyed and reported numbers of ETDs available from the NDLTD. Her data is based on institutions reporting numbers of ETDs available through their institutions, but the NDLTD has not begun to report institutional use of or access to their ETD collections.

See http://scholar.lib.vt.edu/NDLTD/

2.4.6 Measurement in Related Contexts, Joan Lippincott and Jose H. Canos Cerda

Another important type of post-processing is the extraction of statistical information from metadata sets. For administrative purposes, institutions may be interested in the number of ETDs supervised by each professor, the keywords most used, the month(s) in which more ETDs are submitted, etc. Usually, relevant metadata are extracted from the ETD database and processed using specialized tools like Microsoft Excel. The access to the database can be done using either ODBC drivers or specialized middleware utilities.

For a broad view of *counting information*, two projects are widely regarded as providing interesting models and data.

n Peter Lyman and Hal R. Varian's How Much Information project at the University of California, Berkeley is an attempt to measure how much information is produced in the world each year.

n The OCLC Web Characterization Project conducts an annual web sample of publicly available web sites to analyze trends in the size and content of the web.

Some programs that provide guidance or models for *collecting institutional data in higher education* are also available. These projects can provide definitions for data, survey questions, and descriptions of data collection that can be adapted for one's own institution.

n K. C. Green has been conducting the Campus Computing Project since 1990. His work charts the increasing use of technology on campuses.

n The TLT Group’s Flashlight Program, under the direction of Steve Ehrmann, has developed a subscription-based tool kit that provides a large, structured set of assessment techniques and data collection models that can be adapted by individual campuses that want to study and improve the educational uses of technology. The Flashlight Program web site also includes valuable overviews of assessment issues and provides advice on deciding what to assess and how to develop questions.

n The Coalition for Networked Information’s Assessing the Academic Networked Environment project provides case studies of campuses that implemented assessment projects.

n One of the participants in the CNI project, the University of Washington, has a rich set of assessment instruments and reports on its web site, UW Libraries Assessment.

2.4.7 Guidelines for Implementing an Assessment Program for ETDs, Joan Lippincott

The Coalition for Networked Information (CNI) sponsored a project, “Assessing the Academic Networked Environment,” in which institutional teams developed and implemented assessment projects related to a variety of areas, including teaching and learning, electronic reserves, computer skills, and electronic library resources. From the project reports and informal feedback from the participants, CNI developed a set of guidelines for institutions engaging in assessment activities related to networks or networked information. The guidelines focus on the process of doing assessment in higher education. The suggestions can be applied directly to assessment projects for ETDs.

n Bring together an assessment team of individuals from various units on campus that can add useful perspectives and expertise; include, if possible, someone who specializes in assessment.

n Align the overall goals of the assessment initiative with the institution’s goals and priorities.

n Gain support from the administration at as many levels as possible.

n Make a realistic determination of the resources (staff, time, equipment, and money) that are available for the assessment.

n Choose a manageable portion of the assessment project as the first implementation. Do not attempt to do a comprehensive assessment of campus networking on the first try.

n Consider using more than one assessment technique to measure the aspect of networking that you have chosen; particularly consider combining quantitative and qualitative approaches as complementary techniques.

n Identify carefully who are the audiences for the assessment reports.

n Examine what you might do with the information you collect, including improving services, seeking additional funding and determine whether your data will provide what you need for that objective.

n Refine assessment instruments on a periodic basis and incrementally add new components.

n Monitor the work of national groups such as ARL, EDUCAUSE, CNI, and the Flashlight Project to see whether materials they develop and guidelines they produce can provide a framework for your project.

(Joan K. Lippincott, "Assessing the Academic Networked Environment," *Information Technology in Higher Education: Assessing Its Impact and Planning for the Future*, ed., Richard N. Katz and Julia A. Rudy. Jossey-Bass, 1999, pp. 21-35.)

2.4.8 Student Comments, Gail McMillan

In addition to gathering availability and usage data, online surveys can be used to gather information from readers to voluntarily agree to be surveyed by answering questionnaires. Some of the useful information that can be gathered includes. Sample questions for ETD readers might include:

1) Where do you work/study?

2) What do you do?

3) What type of computer are you using?

4) What is the speed/type of connection are you using?

5) Are you familiar with Adobe PDF?

6) Are you familiar with online databases?

7) If you are from a university, does your institution accept electronic theses and dissertations (ETDs)?

8) If your institution does not accept ETDs, do you think it should?

9) Have you ever submitted an ETD? (to find out if your readers are ETD authors past, present or future)

10) For what purpose are you using this digital library?

11) Did you download any ETDs?

12) If you downloaded any ETDs, how did you find them?

13) If you downloaded any ETDs, how easy was it to find what you were looking for?

14) If you searched for an ETD, how fast was the response to your search request?

15) How often do you plan to use Virginia Tech's ETD library?

16) How often do you plan to use other ETD libraries?

17) Comments from survey respondents

2.4.9 List of Resources, Joan Lippincott

General Information:

Association of Research Libraries. ARL Statistics and Measurement Program.

http://www.arl.org/stats/index.html

Association of Research Libraries. Supplementary Statistics 1998-99. Washington, DC: ARL, 2000.

http://www.arl.org/stats/sup/SUP99.pdf

Astin, Alexander W. Assessment for Excellence

The Philosophy and Practice of Assessment and Evaluation In Higher Education.

American Council on Education/Oryx Press. 1991.

Developing National Library Network Statistics & Performance Measures.

http://www.albany.edu/~imlsstat/#Working

EQUINOX: Library Performance Measurement and Quality Management System.

http://equinox.dcu.ie/

International Coalition of Library Consortia (ICOLC). Guidelines for Statistical Measures of Usage of Web-Based Indexed, Abstracted, and Full Text Resources. November 1998.

http://www.library.yale.edu/consortia/webstats.html

McClure, Charles R. and Cynthia L. Lopata. Assessing the Academic Networked Environment. Washington, DC: Coalition for Networked Information, 1996.

NDLTD: Networked Digital Library of Theses and Dissertations.

http://www.ndltd.org/

Ryan, Joe. Library Statistics & Measures.

http://web.syr.edu/~jryan/infopro/stats.html

Technology in Higher Education and Web Studies:

The Campus Computing Project.

http://www.campuscomputing.net/

Coalition for Networked Information. Assessing the Academic Networked Environment.

http://www.cni.org/projects/assessing/

Katz, Richard N. and Julia A. Rudy. Information Technology in Higher Education: Assessing Its Impact and Planning for the Future. New Directions for Institutional Research No. 102. San Francisco: Jossey-Bass, 1999.

Lyman, Peter and Hal R. Varian. How Much Information?

http://www.sims.berkeley.edu/how-much-info/

Online Computer Library Center, Inc. Web Characterization Project.

http://wcp.oclc.org/

TLT Group. Flashlight Program.

http://www.tltgroup.org/programs/flashlight.html

University of Washington. UW Libraries Assessment.

http://www.lib.washington.edu/surveys/

Virginia Polytechnic Institute and State University. Electronic Thesis and Dissertation Initiative

http://etd.vt.edu/

2.5 Policy Initiatives: National, Regional and Local; Discipline Specific; Language Specific

Susanne Dobratz

National

For universities, it seems most practical to participate in national ETD initiatives. For those initiatives it is advisable that the National Library, which is often in charge of archiving the country's literature, takes a leading role. We see those approaches in Germany within the Dissertation Online Initiative or in Canada. The national library as a central point, organising not only the archive structure but also the cooperation between universities may serve as a central entry point to the ETD initiative for interested parties.

Pricipal tasks of a so called central coordination bureau for national ETD initiatives are:

- providing a coordination structure for the cooperation of universities for political, organisational, technological and educational issues and developments
- providing an organisational concept for funding local or regional initiatives, therefore negociating and cooperating with national funding agencies
- defining special interest and working groups in which representatives from universities can participate
- organising workshops for participating universities covering special topics in order to discuss and solve particular problems
- cooperating with the international initiatives, e.g. NDLTD, Cybertheses, MathDissInternational or PhysDiss.

In France, a national program for ETDs has been initiated by the Ministry of Education. The electronic deposit shall be compulsory within the end of the year. The organisational scheme adopted is defined as follows :

- each university will be in charge of the conversion of its theses and dissertation into an archiving format (SGML/XML).
- associations of institutions may allow the mutualization of human and technical resources.
- a national institution, the Association des Bibliothèques de l'Enseignement Supérieur (ABES) (approximative translation: Association of Universitary Libraries) has been designated as the national central bureau.

The following text was taken from Prof. Peter Diepold (Sourcebook for ETDs)

In 1996 four German learned societies - comprising the fields of chemistry, informatics, mathematics, and physics - signed a formal agreement to collaborate in developing and using digital information and communication technologies (ICT) for their members, scientific authors and readers. The objectives of this collaboration were

- **on a local level**: to bring together the activities of individual - and often isolated - university researchers and teachers in the various academic fields;
- **nation wide**: to join forces in voicing the interests and needs of scientific authors and readers toward the educational administration, granting agencies, research libraries, documenting agencies, publishing houses and media enterprises;
- **globally**: to use the widespread international contacts of the learned societies to exchange concepts, development and solutions and adopt them to the specific needs within one's own field.

The initiative soon caught public attention, leading to the enlargement of the group. Since then, the learned societies in the fields of education, sociology, psychology, biology, and electronic engineering have also committed themselves to the advancement of the goals of the ``IuK-Initiative''. (IuK stands for ``Information und Kommunikation'' in German).

Funds were granted for three years by the Federal Ministry of Education and Research, `Bundesministerium für Bildung und Forschung (BMBF) (www.bmbf.de).

One major project within this initiative was the Dissertation Online project, undertaken from April 1998 until October 2000. The activities of one of the workgroups led to a proposal to the German Research Foundation (DFG) to fund an interdisciplinary project to present dissertations online on the Internet, involving five universities (Berlin, Duisburg, Erlangen, Karlsruhe, and Oldenburg), and five academic fields, chemistry, education, informatics, mathematics, and physics.. DFG stands for ``Deutsche Forschungsgemeinschaft'' (http://www.dfg.de/) and is Germany's National Science Foundation. Funding was initially restricted to one year.

The first phase started in the spring of 1998 and was terminated in March 1999 with a conference held in Jena, Germany, provoking much attention among librarians and academics. Though an infrastructure had been set up and a number of problems were solved, much remained to be done. Therefore a subsequent proposal to DFG was drafted. DFG funds were awarded for a second year, this time with a heavy emphasis on the collaboration with libraries and university computing centers. The project's research and development extended from May 1999 to October 2000. The overall volume of both grants was some US \$ 700,000. New participants in the second proposal are computing centers and German National Library, ``Die Deutsche Bibliothek'' (DDB) (http://www.ddb.de/) . The project was directed by Prof. Peter Diepold, professor of computer uses in education at Humboldt University, Berlin. (Email:peter@diepold.de)

Later, this project was made into a national initiative with the German National Library as leading partner, who established a buereau for coordination.

Contact address:
Dr. Nikola Korb
Die Deutsche Bibliothek
DissOnline
Adickesallee 1
60322 Frankfurt / Main
Germany
Email:korb@dbf.ddb.de

Regional

South America???

Discipline specific

Discipline specific initiatives focus on bringing researchers and scholars from single research fields together. In the past this has been a very succesful way to establish active communities. These initiatives give members of those communities the benefit of making problems such as those that arise within global or national or generally spoken interdisciplinary approaches or even very small discipline specific problems easier to solve. Therefore discipline specific approaches may reach results faster and more easily than broader projects.

PhysNet

The PhysNet Service exists within the Physnet initiative. PhysNet - the worldwide Network of Physics Departments and Documents - provides a set of information services for physicists. PhysNet is a distributed information service. It uses the information found on the web-servers of the worldwide distributed physics institutions and departments of universities as a distributed database. The restriction to those professional institutions which are accepted by the learned societies ensure the quality and relevance of the offered information. PhysNet serves only professional specific information posted by the scientists themselves. Therefore PhysNet complements the services of commercial providers. All information of PhysNet is kept, stored and maintained by its creators at their local institution’s server or individual homepage. The creators retain all rights to their data. PhysNet only gathers and processes the locally available information of physics institutions to make them globally accessible. PhysNet is a noncommercial service. The access to information offered by PhysNet is free to anyone. The aim of PhysNet is to provide a longtime stable and distributed information service for physics with the collaboration of many national and international societies and physics organisations.

The PhysNet-Services are:

- **PhysDep** offers a set of lists of links to nearly all Physics Institutions worldwide ordered by continent, country and town. In addition, it offers a HARVEST-based search engine to search across the listed Institutions.
- **PhysNet** provides lists of links to document sources of the distributed Physics Institutions. Such document sources are, for example, preprints, research reports, annual reports, and lists of publications of local research groups and individual scientists. The service is also completed with a HARVEST-based search engine.
- **Journals** lists Physics-related Journals, which are available with free fulltext on the web. A list of 'EPS Recognized Journals' is given as well.
- **Conferences, Workshops and Summer Schools** offers a list of servers with Conference Lists in different fields of Physics.
- **PhysJobs** offers a list of links to various physics-related job sites on the web. It also implements a search facility to search for information on these sites.
- **Education** provides online educational resources for physics (e.g. Lecture Notes, Seminar Talks, Visualization and Demonstration Applets), listed by subject areas.
- **Links** lists futher sources of physics information on the web and information services of other fields and disciplines.

- **Services** provides various related tools e.g. to enrich and improve homepages and document sources by adding correct MetaData according the international Dublin-Core standard.
-

More specifically on the subspace PhysDoc of PhysNet: it is serving documents distributed around the world at Physics University departments. <a href="http://physnet.uni-oldenburg.de/PhysNet/physdoc.html">PhysDoc</a>.
It comes with link lists sorted by country and town/state and institution. The present content is about 100.000 documents or document lists (publication lists).

A search engine is attached <a href="http://physnet.uni-oldenburg.de/PhysNet/physdoc_carmen.html</a> which will go into full operation in early 2002, which allows to search for metadata of the documents (author, title, fulltext, keyword). A specific tool of it is to search both in PhysDoc and in MPRESS, the respective distributed document system of Mathematics of the International
Mathematical Union. The special appeal of Physdoc is that it serves a ranking, and allows to find the match of a keyword to PACS classification numbers and its respective counterpart in MSC, the respective mathematical classification scheme. This matching is not by trying to find the same words but by serving articles of the mathematics, which a physicist working in the respective field would search for.
This is a major outcome of the research programme CARMEN.

Theses and Dissertations of Europe in Physics are served by the search engine <a href="http://elfikom.physik.uni-oldenburg.de/dissonline/PhysDis/dis_europe.html">PhysDis</a>.
The link lists of lists of local University physics theses are served sorted by country, town and University. It is operated by Kerstin Zimmermann. PhysDis is part of a server <a href="http://www.dissonline.org">DissOnline</a> which serves a list of

University Libraries in Germany serving theses and dissertations, most of them using DC-metadata by now, and a good set of tools http://www.dissonline.org/tools.html with guides for the authors for writing LaTeX files, upload tools to create DC metadata by the author, by the library, and an installation guide to set up a HARVEST gatherer and broker to join thedistributed PhysDis Service.

Responsible for the server of this initiative is:
Prof. Eberhard Hilf
Carl von Ossietzky University Oldenburg
Department for Physics
Institute for Science Networking
Ammerländer Heerstraße 121
26129 Oldenburg
Germany
Phone: +49 (0)441 798 2742
Fax.: +49 (0)441 798 3201
Email: info@isn-oldenburg.de
WWW: www.isn-oldenburg.de

MathDissInternational

The MathDissInternational project developed as a service within the MathNet initiative, set up in Germany at the Konrad Zuse institute for iinformation technology. Within the scope of the project MathDiss International, a permanent international online full-text document server for mathematical dissertations will be established. In this connection, questions concerning online presentation of the documents and the problems of long-term archiving (from TeX resp. LaTeX documents) will be considered. They include the question of how to homogenize such files in order to enable their later conversion into programming languages following XML. Furthermore, the expansion of research possibilities using online documents is being planned. Providing access to the tables of contents, lists of tables and illustrations and bibliographies on the LaTeX level is of top priority. Because of the structure of mathematical documents written in LaTeX we have a lot of high quality information which gathers dust in the archives without being used for the retrieval of scientific documents. This situation should and could be changed because LaTeX has become a widely accepted tool in mathematical literature.

The project MathDiss International will be sponsored by the DFG: Deutsche Forschungsgemeinschaft (the German Research Foundation) for one year. At the end of that year, the results will be turned over to the State and University Library of Lower Saxony in Göttingen which has offered to provide the long-term support for the new Math-Net Services.

The Project Program includes:

- Complete inclusion of the dissertations through metadata and the expansion of the research possibilities using the source code.
- Standardization of the input files in consideration of mathematical contextual structuring.
- Integration of the services in the Math-Net by the adaptation of the project results to the standards used there.
- The creation of forms of organization for the long-term safeguarding of the service.
- International marketing for the worldwide opening of the service.
- Tests on the conversion of mathematical dis-sertations into new mark-up languages, e.g. MathML.

Responsible for this project are : Prof. Dr. Günter Törner and Thorsten Bahne at the mathematics department of the University of Duibsurg.
Address:
Gerhard-Mercator-University Duisburg
Department 11 - Mathematics
Lotharstr. 65
47057 Duisburg

Germany
Fon.: +49 (0)2 03 379 26 67 / 68
Fax: +49 (0)2 03 379 25 28
E-Mail: toerner@math.uni-duisburg.de / bahne@math.uni-duisburg.de

CogPrints

CogPrints is an electronic archive for papers in any area of Psychology, Neuroscience, and Linguistics, and many areas of Computer Science and Biology, which uses the self-archiving software of eprints.org The CogPrints project is funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as part of its Electronic Libraries (eLib) Programme.

ArXiv
The arXiv.org e-Print archive (formerly xxx.lanl.gov) is a fully automated electronic archive and distribution server for research papers. Covered areas include physics and related disciplines, mathematics, nonlinear sciences, computational linguistics, and neuroscience. A Service of the Los Alamos National Laboratory and the U.S. Department of Energy.

Language specific

2.5.1 Policy Initiatives: National, Regional and Local, Discipline Specific: the case of France, Jean Paul Ducasse

Introduction

In France, policies concerning the electronic distribution of theses derive from two different sources:

n the public power, represented by the Ministry of Higher Education which, through the intermediaries of the Research Division (Direction de la Recherche) and the Library Division (Direction des Bibliothèques), has to date held the responsibility for the description and physical archiving of theses.

n the university establishments possessing the ability to grant doctorates, be they universities or the major schools.

The regulatory framework was set by decree on 25 September, 1985. The decree laid out the procedures for the submission, description and reproduction of theses or other works presented to obtain a doctorate.

The current initiatives for the electronic distribution of theses are numerous and extremely varied, as much in their technical details as in their political ones. Over the past several months, there has been a tendency towards the coordination and grouping of initiatives. This has involved the creation of linkages and networks around establishments that have instituted technical archiving and distribution solutions.

The French system is currently structured on three levels:

n A local level, based on the initiatives of certain producer institutions

n A regional level, where a number of research institutions have aggregated at a regional or thematic level

n A national level, based on the creation of a ministerial working group, following a circulated letter written by the Minister in October 2000. This working group must define a new regulatory framework.

The Local Level: the policy for the electronic distribution of theses at the Université Lumière Lyon2

Introduction

The basic principle of the programme for the electronic distribution of theses, which was developed by the Université Lyon2 in partnership with the Université de Montréal, is to employ standardized formats which provide the conditions for perennial archiving and which permit distribution that guarantees effective and total interoperability. In most cases this involves formatting, so as to structure all the documents. This formatting is performed with computerized tools, and is necessary for widescale distribution over the internet.

The Université Lumière Lyon2 changed the conditions for the electronic production, submission and distribution of theses when it introduced its Thesis Charter. This Charter spells out official conditions doctoral students must accept. In return, the students can receive supervision and training in the use of text processing tools within their research group.

This document guarantees the University's commitment to support young scholars, so that they benefit from the best conditions for creating, archiving and distributing their work when defending their thesis and after.

The UNIVERSITÉ DE LUMIÈRE LYON2's THESIS CHARTER

The Université de Lumière Lyon2's Thesis Charter is proposed in conformity with the decree of 3 September 1998. Its preamble restates the appendix of this decree (Model Charter) and fills out the text with the following conditions:

1) Every thesis is prepared within a research group linked to a Doctoral School. The doctoral school's particular role is:

a) to give a student applying for admission to the doctorate all pertinent information on the following two subjects: first, the functioning of doctoral studies, the research supervisors and pertinent teams for his/her project as well as the follow-up he/she can expect; and second, the possibility for student aid, scholarships, bursaries, and partnerships with a company likely to provide him/her with the means to completer his/her project;

b) to approve the agreement between the student and the supervisor for the preparation of a thesis;

c) to coordinate the academic training dispensed in order to obtain the DEA as well as during the years of doctoral studies;

d) to offer doctoral students a larger academic environment than their thesis specialization, for instance by organizing transversal seminars, methodology courses, exchanges with other laboratories (especially European ones), by favoring discussion forums, and by making all useful documents available on a server, etc...;

e) to consult annually with the thesis supervisor about the progress of work and to give an assessment of exemptions from extensions. The thesis supervisor’s follow-up, like that of the doctoral school’s director, can be conducted by e-mail when the student in undertaking activities far from the university;

f) to inform doctoral students about the question of professional insertion throughout their doctoral studies, notably by organizing internships and information sessions with professionals, but also by organizing where possible a course of study involving rotation, with periods in firms, in national education etc.;

g) to track this insertion after the thesis defense or after a post-doctoral period in an external laboratory.

2) A doctoral thesis is a contribution to the knowledge produced within a research team. In order to ensure the best possible distribution of these contributions within the scientific community, the University recommends that doctoral students prepare their thesis using a computer. To this end, it provides training courses for doctoral students to help them with the composition of their thesis: word processing, use of style sheets, software tools necessary for their project, formal structuration etc. The research teams and/or the Research Division at the SIR (Computer Research Support) provide doctoral students with workstations. Except in exceptional cases, they must submit their thesis in computerized form, from which they have produced printed copies. The SCD (Common Documentation Service) ensures the respect for the norms of indexation and more generally for the recommendations of the ABES (Higher Education Bibliographic Agency).

3) A thesis is accessible:

a) at the SCD, under the conditions set out by the decree of 25 September 1985 that applies Law no. 84-52 of 26 January 1984;

b) from the University’s server, once the author and the jury have agreed to its electronic distribution.

The rules of usage concerning the necessity of the jury’s authorization before distribution and the confidentiality clauses that can apply to parts of the thesis, apply to the electronic version in the same manner as they do to the paper version.

4) Doctoral students are represented on the scientific committee of each doctoral school. In the case of conflict between a student and his/her thesis supervisor, the Director of the thesis school plays the role of the first mediator foreseen by the Model Charter. If the conflict persists, the President must ask the Scientific Council to nominate a mediator from outside the doctoral school and/or the Establishment. The mediator reports to the President, who is the arbiter of last instance. In the case of joint supervision, the Director of the doctoral school plays the role of mediator for all conflicts between the student and the French thesis supervisor, or between the two thesis supervisors. If the conflict persists, the Presidents of the two universities will decide in the last instance. When the Director of the doctoral school is also the thesis supervisor, he or she is replaced in the role of mediator by the vice president in charge of research.

5) The doctoral schools have the responsibility for distributing this charter to the DEA and doctoral students under their supervision.

Approved unanimously by the Scientific Council of 7 December 1998

The Edition Electronique de SeNTIERS cell is responsible for the electronic archiving and distribution of theses defended at Lyon 2. In this role, the cell is involved in the legal deposit of the thesis. With time, the legal deposit of paper copies, as in current practice, is likely to fall out of use. This will imply relatively significant changes in the mode of producing documents. As well, in order to allow doctoral students to pursue their work without too many perturbations, two systems for submission coexist at Lyon 2: **electronic submission** and **mixed submission**.

Organization of the thesis' administrative circuit

The establishment of an electronic archiving and distribution system assumes that the administration possesses an electronic version of the research work; this translates into the creation of a mode of thesis submission that includes an electronic submission.

Electronic submission should be organized in an automatic fashion, on a server able to support anywhere between a mailbox-type submission system to a more complete system permitting the recording of the submission, of metadata, etc.

This submission will naturally be integrated within the tradtional administrative circuit of the thesis and can take several forms (independently of hte software used or the discipline in question):

- the mixed submission,
- -the electronic submission.

The mixed submission

This form of submission should be seen as a step towards the completely electronic submission, but it is also a satisfactory solution for establishments lacking the infrastructure and personnel needed for electronic submission.

The organization of the system

It is organized in serveral stages:

1) The student submits paper copies and an electronic copy at the thesis service, and attests to the conformity of the two versions. At this point, he/she authorizes (or refuses) the distribution of this work on the Internet.

2) The electronic publishing service validates the submission. It is a question of verifying the readability of the files, the presence or absence of particular characters, as well as the presence in electronic or paper form of all the non-textual elements (images, sounds,...).

3) After the defense, one of the following cases is likely:

n -the jury authorized distribution of the thesis in its current state. In this case, the doctoral student has one month to make minor corrections (spelling) to the thesis, in the form of erratum. There is no new submission. The thesis is converted into SGML (archiving format) and then into HTML and XML (distribution formats). The thesis is place on the University's intranet (equivalent to consultation in the library) and, if authorized by the student, on the Internet. The description of the thesis is broadcast.

n -the jury demanded corrections. In this case, a new electronic submission and validation is necessary. Once the president of the jury validates the corrections, the thesis follows the normal treatment process.

n -Non-corrected theses, or those subject to a confidentiality clause, are archived in SGML and are only distributed on the intranet, according to the case. Their description is broadcast, but with mention of their confidentiality.

Electronic submission

This second system removes the risk of non-conformity inherent in the mixed submission. The legal deposit of the thesis thus consists of an electronic submission. The printed copies of the thesis required for the defense and for deposit in the library are printed from the electronic submission.

If this system is more satisfactory, it nevertheless requires a supplementary infrastructure. In order to guarantee that the printed copies are truly drawn from the electronic version submitted, the University must offer the authors the means to effectuate a complete electronic submission and must undertake the task of printing the copies needed for the defense.

This system thus requires the support of many different types of structures. These include the traditional thesis service, a relatively heavy computer infrastructure made available to doctoral students (digitization of illustrations, video acquisition, sound, user assistance) and a copy service capable of providing irregular output on short notice (the frequency of thesis defenses varying in function of the university calendar).

Starting from the files produced by the student, a “single file” printing (PostScript and/or PDF) will be undertaken by the University, under the student’s control (lay-out, rendering of illustrations, etc...).

n -the source files as well as the PostScript file will be engraved and transmitted to the electronic publishing service for archiving while awaiting the defense.

n -Only the PostScript file will be transferred to the copy service’s print station.

The resources necessary for the implementation of a thesis treatment platform in an establishment or group of university establishements

Two functioning architectures can be considered:

n **-a client-server system**. Here, a server site ensures the management of computing resources while client sites ensure production (from a distance) and distribution (locally). A web interface manages the communication between the different client sites and the server site.

Associations of institutions can develop as a result of geographic or disciplinary proximity, by creating synergies and by combining skills and competencies.

n **-an autonomous system**. Here, each production site undertakes, in an autonomous manner, the entire operation of processing, archiving and distribution. This would apply to sites possessing the necessary human resources, or treating a volume of documents justifying such a system.

Human resources:

The existing system requires diverse skills, both in computer engineering and in engineering documents, which correspond to different tasks:

-managing information systems: updating programs, managing both archiving and putting information on-line, user assistance, linking users to the group(s) in charge of developing and updating the system;

-processing the documents: formatting documents for processing, digitizing illustrations, sounds, video, etc., verification of the processing and capture of metadata.

The presence of the two types of human resources is not necessary in each production center.

-processing the documents: 1 full time equivalent **d'IE (?)** for an annual volume of 100 theses per year.

-information management: Needs vary depending on whether the local configuration or the client-server configuration is used.

Training in the use and administration of the software platform for processing theses

The training activities are targeted at two types of agents with specific skills:

-recognized computer skills allowing the agents in question to act on several levels. At the local level, in order to assist those using the software platform, these skills are the maintenance and adaptation of the software platform. They must also be given skills at the network level, where they will serve as relays between users and the group piloting the evolution of the platform, as well as participate in the activities of this piloting group (computing decisions and developments).

-electronic publishing skills: electronic submission of theses; training doctoral students in the use of generic document forms; preparation and formatting of documents according to the rules determined by the chosen software platform. This is a new domain, and defining its profile with exiting qualifications is a delicate process: mastery of electronic document production tools, knowledge of the different types of academic production, and pedagogical ability to transfer this knowledge to a public composed of doctoral students and researchers.

The role of Doctoral Schools and the training of doctoral students

The implementation of these new processes requires the participation of the principal actors involved.: doctoral students and their supervisors. The training of authors, thesis supervisors, and research structures must occur within the Doctoral schools. These school constitute the ideal frame of reference for several reasons:

-the training provided to the student will be adapted to their discipline and thus to the specific computing tools that they may have need to use;

-the supervisors will be implicated in the process;

-the sharing of this knowledge between the different actors of a Doctoral school will provide a supplementary factor of internal cohesion.

With time, the Doctoral Schools should become support centers for the use of the new tools of scientific production and publication. In the course of thesis writing, doctoral students will receive the tools and the help required to write their document in an effective manner, and will acquire all the skills needed for its processing.

Tools such as generic document forms and technical specifications will be provided by the Ministry. The

Doctoral Schools will be charged with adapting them to the particular practices of a discipline or of the establishment (in the case of the style sheet). A generic training program in using generic document forms will be available and adaptable according to local needs. In light of past experience, six hours of training per doctoral student seems sufficient for training students in the advanced use of text processing. One might expect that, within several years, students will be trained in text processing before starting the doctorate and that training costs during the doctorate will be increasingly minimal.

A permanent link should be established between the electronic production and distribution services and the Doctoral Schools so as to create synergies between the users and the electronic publishing specialists. This relationship should be based on permanent cooperation, and initiated in the training sessions through a presentation of the service and of its results.

Regional and Disciplinary Policies:

The implementation and publication of the Université Lumière Lyon2's policy, and the computerized means of processing that the University established, have allowed for the creation and structuration of a network of university and scientific establishments. The universities of the Rhône-Alpes region, united within the context of a University Conference, have decided to implement the principle recommendations enunciated by the Université Lyon2. Subsequently, at the level of the French territory, as well around its borders, one notes a tendency to regroup, on the one hand universities (Lyon, Marne la Vallée, Paris-Sud, Grenoble, Genève), and on the other hand scientific establishments (CNRS, Inra, Inserm). One equally finds disciplinary groups like the "Mathdoc" network which unites the mathematical laboratories, and which is in turn linked to a European network. This network's work involves collecting the metadata produced by each researcher which are necessary for the description and announcement of theses. These metadata are then archived and made accessible either on the site of the laboratory or on the researcher's own site. The approach is less institutional, and thus less enduring, in the sense that its conservation is left to the appraisal of the researchers and of the laboratories.

Not all of these institutions have adopted the Lyon2 model (RTF=>SGML/XML). Some have preferred the generic document form produced by Adobe, while others produce documents using Latex and distribute them in “postscript” format. Nevertheless, reflection about these choices is carried out in public, thanks notably to the work undertaken by the Ministerial group.

French Public Policy: The Minister’s circular

To conclude the work led by a reflection group, the Minister responsible for Higher Education released a circular in October 2000 which defined the major axes of the futre policy concerning the electronic archiving and distribution of theses.

Electronic Distribution of Theses

Text addressed to University presidents, and to presidents or directors of institutions of higher education

Theses defended in universities and other institutions of higher education constitute documents of the highest value. Looking after their promotion serves both the interests of the young doctors and the institutions, as well as the end of increasing the international visibility of French research.

The deep transformations which have for some time characterized information technology have clearly rendered the existing system for valorizing theses (defined by the decree of 25 September 1985 dealing with the submission, description, reproduction and distribution of theses) obsolete.

It is on the basis of this observation and in recognition that:

-theses are henceforth produced “naturally” in computerized form,

-the equipment and networks in institutions of higher education have been greatly developed,

-the majority of universities are currently positioning themselves as producers and distributors of electronic information,

that a group associating the services of the Ministry of National Education and the Ministry of Research, the conference of university presidents, the association of university library directors, and numerous experts having undertaken experiments in this domain, submitted a report to me on the electronic distribution of theses. This report, whose principal conclusions I have validated, can be consulted on the Ministry's server at the following address:

http://www.sup.adc.education.fr/bib/

The proposed new system foresees the distribution of theses on the Internet once a certain number of conditions are met:

-authorization of the head of the institution, following the advice of the jury and the authorization of the author, while respecting intellectual property regulations,

-respect by the doctoral student of minimum technical specifications,

-conversion of the thesis, using automated assembly lines, into adequate archiving and distribution formats for storage and for placing on-line.

The intervention of numerous actors above and beyond the doctoral student will be required:

-that of the establishment where the thesis is defended, by way of:

.the doctoral schools, who are responsible for providing the student with training and technical assistance,

.the common documentation services, responsible for describing the thesis and providing the document's electronic address in collective and local catalogues,

.the service charged with converting and putting theses on-line, using the software provided to them.

-that of the state or of a national operator, by means of

.elaborating technical specifications and training supports,

*.**guaranteeing** or providing processing chains,*

.providing secure archiving.

On these bases, in agreement with the Minister of Research, and after consultation with the CPU, I have decided:

-to put a project group in place,

-to elaborate a new decree concerning the submission, description, archiving and distribution of theses,

-to organize training activities for the institutions, or the groups of institutions, who wish to rapidly enter into this new system,

-to put into place the collective functions necessary to give overall coherence to the process, taking into account the acquired skills of the National Thesis Reproduction Workshops [atelier nationaux de reproduction des thèses] (ANRT), the Higher Education Bibliographic Workshop [atelier bibliographique de l'enseignement supérieur] (ABES), and the National Computing Centre for Higher Education [centre informatique national de l'enseignement supérieur] (CINES).

This new scheme will obviously take time to put into place, progressing as the institutions introduce adequate assembly lines. It goes without saying that the old system defined by the decree of 1985 will continue to apply to the theses defended in institutions which have not yet taken the corresponding measures.

The Minister of National Education

Jack LANG

A certain number of principles underpin this political interest in the electronic distribution of theses:

To increase the visibility and to valorize French scientific production at the international level

To favor a new approach to the thesis so that it is considered as a dynamic database rather than a static body of knowledge. The electronic version of the thesis becomes a genuine work instrument which meets the demands of users, starting from the user’s own agenda of research and investigation.

To develop training for student-researchers in the use of information technology and electronic publishing. Student-researcher will find themselves in the role of scientific information producers since they will be able to master these mechanisms and thus increase their autonomy. They will personally acquire the status of being information producers and distributors.

This should gather together the elements permitting scientists to produce, archive and distribute all their research work by themselves. Let us remember that France’s approach to the electronic distribution of theses is integrated into a vaster francophone project of putting the results of public research online.

2.6 E-Commerce: fee based methods, Australian Digital Theses Program

The ADT Program has developed a simple e-commerce option for members to use should they choose. The file format for the ADT is to always include a *front.pdf* file. This file contains author, title, abstract, contents pages, acknowledgements, etc. It is designed to give an expanded view of the theses' contents before deciding to view the whole thesis. Viewing the whole thesis can then either be unrestricted and free or at cost using the local institutions online payment system. Currently no members of the ADT have used this option.

For a description of the Online Payment System currently in use at The University of New South Wales please refer to:

http://www.library.unsw.edu.au/thesis/adt-ADT/info/ecommerce.html

[The University of New South Wales Library is the leader of the ADT Program]

3. Students, Edward Fox

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.

This section of the Guide highlights issues of interest to students who write or refer to ETDs. It provides guidance on how they can proceed in these activities, as well as assistance to help them in understanding the context in which these activities take place.

A wide variety of technologies and approaches are discussed. These should be considered in light of local policies and practices. In some cases students will want a simpler effort. In others they may wish to engage in more elaborate preparation of an ETD than is locally common, in order to learn, and/or in order to be more expressive in their scholarly communication. It is hoped that the content in section 3, as well as the more detailed technical information in section 4, is helpful.

3.1 How to learn about ETDs? Edward Fox and Joseph M. Moxley

See ETD Resources for helpful external links: Tools for Writers; Web Design; Bibliography Tools; Career Resources; Plagiarism; E-journals; E-books; Citation Manuals; Libraries; Multimedia Examples; ETD Pilot Projects; Indexes, Abstracts; Evaluating Resources; Search Engines.

For model ETDs, see our Exemplary ETD Database.

Some students who prepare an ETD undertake this effort as an independent. That is, they prepare a paper document as required by their university and turn that in, but are not allowed to submit an ETD to their degree granting institution. If they understand about ETDs and realize the benefits of such works, if they know about creation and submission, they may elect to submit an ETD, such as to a service for independent ETD authors which are afforded by NDLTD or other groups. In such cases, students are usually "on their own" and will follow instructions in this Guide or at other online sites, such as those hosted by Virginia Tech (http://etd.vt.edu).

In addition to learning about ETDs from this guide or other online sources (see also from www.theses.org), students may be helped by local staff. Typically staff at institutions that are members of NDLTD will have a project in place, with personnel identified who can assist. Often there are two types of helpers. One type involves trainers. These typically run workshops or other training events or programs to help students to create ETDs. In some cases there are both basic and advanced training efforts, depending on the level of interest of the student and/or depending on how "rich" (in terms of use of multimedia content, for example) an ETD will be.

The other type of helper is a person who provides assistance on demand. This may be a person who works in a campus New Media Center (see, for example, http://www.nmc.vt.edu). These centers may have a wide variety of multimedia devices (to capture images, audio, or video - as well as to help one create such resources from scratch). Alternatively, there may be people in one's own laboratory or department with particular expertise in creating certain types of digital works that might be an important part of an ETD. In short, students should seek out such help where appropriate and work with trainers and mentors as desired to hone their knowledge and skills, as well as to prepare the very best ETDs.

3.1.1 Importance of satisfying local requirements,

Edward Fox

When students prepare ETDs, they must be sure to undertake this effort in accordance with all requirements. They should be aware and considerate of the interests, time, and abilities of the faculty who advise and examine them. They must be sure that those in charge of the graduate program and library, as applicable, are satisfied with their submission. Typically, there will be written or online instructions, such as are available for Virginia Tech students (http://etd.vt.edu).

If their university does not accept ETDs, students can submit as an "independent", according to rules and procedures being developed by NDLTD. Instructions may be updated from time to time and will be found at www.theses.org.

3.1.2 Learning from other ETDs, Edward Fox

Documents are often prepared in a similar form to that of prior documents. Students are encouraged to go to www.theses.org and follow links found there to other ETDs. Of particular interest should be the set of Notable ETDs found at that site. Modeling one's work after these may be an effective strategy as long as local requirements are satisfied (see above).

In its Fall 2001 meeting, the NDLTD Steering Committee decided to pursue this strategy, giving awards for the best masters and for the best doctoral ETDs. Once these are identified, they will be linked in with the above-mentioned Notable ETD section.

3.2 How to prepare an ETD? (approaches),

Edward Fox

ETDs are prepared to facilitate scholarly communication. They are vehicles for transmitting the research results of a student, in the most effective way, to each person with interest. Truly effective communication requires students to become facile with tools and methods of expression so their ideas and findings can be clearly conveyed. At the same time, since communication takes place across gaps in space and time, it is important that the form chosen for an ETD be understood in different places and in future times, as typically occurs when standard representation schemes are used.

Thus, at a 1994 meeting of ten universities discussing ETDs in Blacksburg, Virginia, USA, it was recommended that ETDs be prepared in both a rendered and a descriptive form, like PDF and SGML. The former ensures that a reader sees things the way the author desires, which may be of particular importance with regard to mathematics or artistic works. The latter instead emphasizes the logical structure and content (as is done with HTML pages) and makes it easier to precisely specify the target of a search.

However, no authors write using tools that store their works only in PDF, and few authors directly create SGML documents. Rather, they employ tools they select based on cost, availability, popularity, familiarity, efficiency, or other criteria. For many, the choice is a package like Microsoft Word that may have been bundled with their computer. For others, like mathematicians who need to work with proofs and equations, the main decision may be what representation to use (e.g., LaTeX), with subsidiary decision regarding what editor or other special tool can best manipulate LaTeX files.

Such choices may be discipline specific, may be based on what is commonly used by faculty and students who are in a particular group, or who use a particular computing environment. If there is no clear choice imposed, then a decision could be based on the information provided in subsequent sections of this Guide. These cover, in particular:

- 3.2.2 – using word processing systems
- 3.2.3 – using an editor (or converter) to prepare SGML/XML
- 3.2.4 – preparing PDF
- 3.2.5 – rationale/overview for SGML/XML - and converting to them
- 3.2.6 – multimedia
- 3.2.7 – metadata considerations
- 3.2.8 – intellectual property considerations

3.2.1 Overview: writing with word processors and structured editors, Edward Fox

Details regarding a variety of approaches to use in preparing ETDs are found in subsequent parts of section 3.2. In all cases, there is a writing phase and a conversion phase, so that an archivable form is produced. The writing phase should allow authors to efficiently and effectively capture their research results in a form that will be understandable by others. The conversion phase should take the written work and use it to create PDF and/or SGML/XML.

Authors preparing an ETD must learn about electronic publishing to succeed in their task. In the future, costs may decrease on structured editors and other software that directly yields SGML/XML. Meanwhile, a small number may invest (their money and time) in such as SoftQuad's Author/Editor. But today, most authors learn about word processing, since it is commonly used for other writing tasks, such as letters, reports, papers, articles, and books. Word is the most common word processor. Word Perfect is somewhat popular. FrameMaker is a more costly package, harder to learn, but quite powerful, and used in professional writing efforts. LaTeX is popular with mathematicians, scientists, and engineers who deal with mathematical notation, proofs, equations, etc.

From word processing systems such as these, it is relatively easy to prepare PDF files. It also is possible to prepare SGML or XML. However, this latter typically requires a good deal of pre-planning, work on conversion, and final editing/checking to make sure the archival form correctly conveys the author's intent.

Some of these efforts yield metadata for ETDs as a byproduct, either directly or in connection with a conversion effort. Thus, if "TEI-lite" (drawn from work on the Text Encoding Initiative) is employed, metadata can be produced from the document by analyzing the header portion. Generally, however, authors prepare a metadata record for their ETD by filling in some type of form, entering in suitable information into a database from which the desired representation can be generated as needed. The metadata, as well as other forms, may carry a description regarding copyright and other intellectual property right management issues.

In addition to writing, authors may convey their results using multimedia devices. Special tools are often employed to prepare graphic aids, animations, or musical compositions. Other tools support conversion of photographs, video, music, and other formats. In some cases, multimedia content can be included with the rest of the work, as when images are included in a PDF file. Irregardless, international standards for the various media types, as well as their combinations, should be followed so archivable works result; otherwise these parts of an ETD may become lost to interested readers in the distant future. We aim to avoid such losses whenever possible, through training, following best practices, and building upon recent work on ETDs, as documented below.

3.2.2 Writing in word processing systems,

Edward Fox

The following subsections explain in detail how to work with Word, Word Perfect, LaTeX, and FrameMaker. Note, however, that many tools rapidly become obsolete. Authors should work with current, supported versions of tools. Authors should refer to the content below as appropriate, but should seek other aid from their local institution as needed to prepare their ETD.

3.2.2.1. Microsoft Word and Office 2000,

Joseph M. Moxley

Inspired by the NDLTD, a group of faculty and staff joined together to create the Digital Media Institute at the University of South Florida. We created our community back in 1999 with the following goals in mind:

n Improving the quality of e-documents, particularly theses and dissertations.

n Researching ways authoring tools, particularly Microsoft Office 2000, can be used to facilitate electronic theses and dissertations.

n Researching how technologies alter graduate education, including mentoring relationships, topic selection, intellectual property, writing processes, and publishing practices

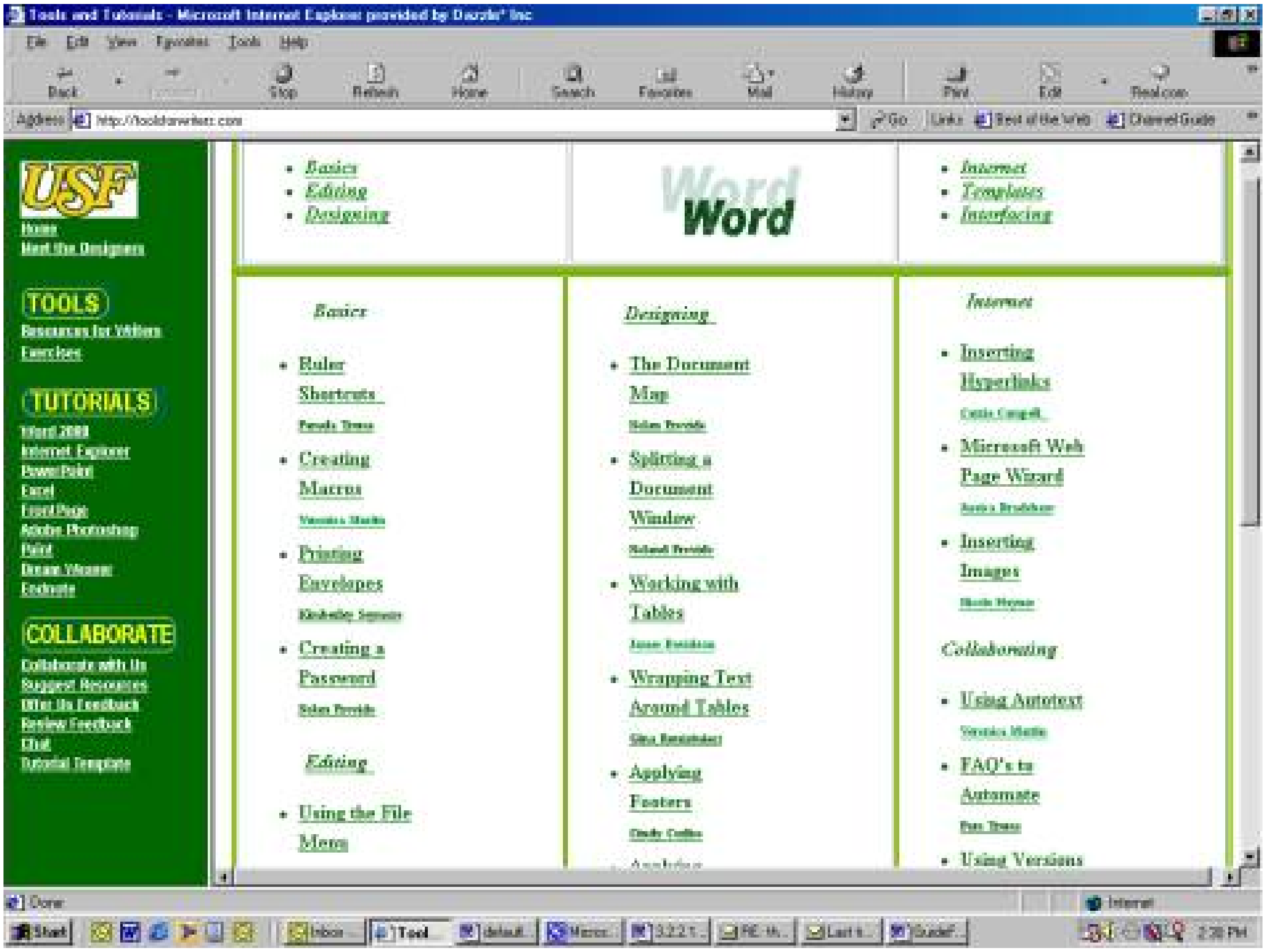
n Working with tool developers to keep abreast of new tools for researchers and writers

n Providing an open forum for the exchange of ideas regarding the evolutions of new media scholarship.

n Providing training workshops, reference materials, and support for graduate students interested in contributing to the University of South Florida's digital library of electronic theses and dissertations.

To achieve these goals, we wrote a proposal that Microsoft subsequently published; see MSFT proposal. We also secured funding from Time Warner Communications to provide 33 graduate students with high-speed Internet access for one year. And then we crafted training workshops where we attempted to explore what the Office 2000 tools could do for us as mentors of graduate students and as writers. We quickly came to focus on exploring the collaborative features of Word 2000.

At the undergraduate level, we have involved technical writing students in creating over 50 tutorials on Office 2000 features that ETD authors will find useful; see http://toolsforwriters.com.



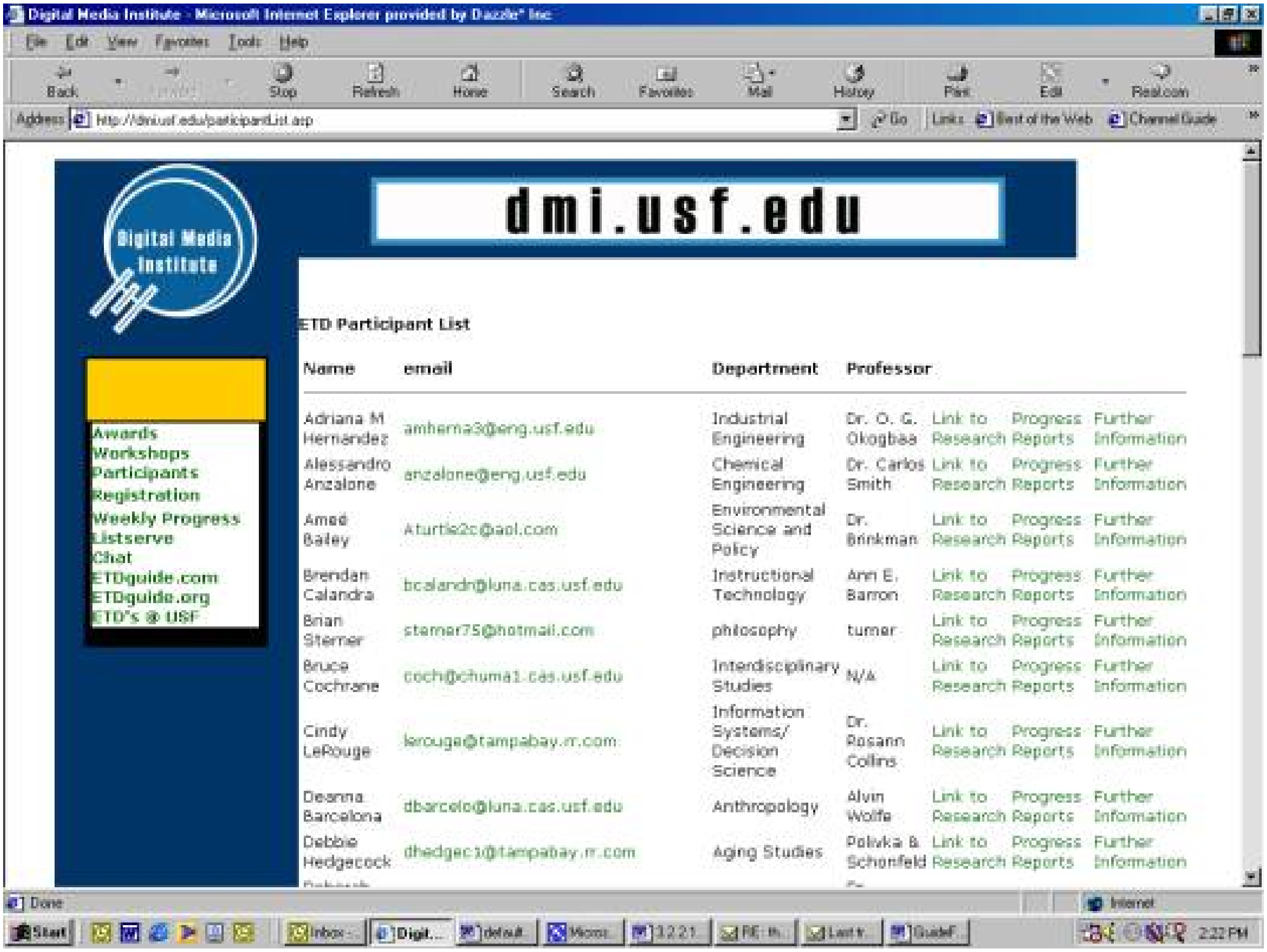
While we are excited about engaging our undergraduate students in this research and support endeavor, our greatest efforts have occurred at the graduate level. Since creating our community three years ago, my colleagues and I have analyzed how Microsoft's Office 2000 can be used to better support students' needs as writers of multimedia scholarship and faculty members' needs as mentors of electronic theses and dissertations. Using case study and ethnographic methodologies, we have researched how communication technologies can improve graduate education, particularly academic scholarship. Following Walter Ong, who theorized "Technologies are not mere exterior aids but also interior transformations of consciousness" (82), we are researching how technologies alter graduate education, including mentoring relationships, topic selection, intellectual property, writing processes, and publishing practices.

In the preliminary stages of our investigation, we focused on examining the Office 2000 suite, yet we expect to investigate related tools for writers, including bibliography, and quantitative and qualitative data analysis tools. We chose to focus initially on Office 2000 because it is used by so many other members of the NDLTD. Office 2000 includes all of the necessary components (word processor, database, spreadsheet, presentation graphics, electronic mail) necessary to author a thesis or dissertation, and all of these components can be used to produce HTML code, as well as native-format documents. In addition, Office 2000 has powerful features for collaboration and multimedia authoring. Outlook--Microsoft's email and calendaring tool--serves as a framework for document workflow, calendaring, sharing and exchange. For example, regardless of their locations in time and space, faculty and students can use Outlook to provide students with an integrated set of reviews and links to grammar and punctuation references. From any document in Office 2000, faculty and students can use NetMeeting to synchronously discuss documents, including audio/video-based discussions.   They can invite scholars outside the committee to respond to drafts. Numerical data, as well as graphical representations of it, can be published out of Excel in such a fashion as to permit limited manipulation and re-analysis from a Web Browser. More extensive analyses can be formed by “roundtripping” the data back into Excel.

Throughout our research, as we work with Office 2000 tools in proposal preparation, research, and thesis/dissertation writing, we are asking "What tools are really useful? What motivates or dissuades innovative use of tools?” Some graduate students are maintaining a Case Study Journal where they reflect on how use of software tools influences our research, writing, and relationships with mentors. In turn, some faculty are reflecting on ways the tools influence mentoring, scholarship, and teaching and learning. Jude Edminster, a doctoral student in Rhetoric and Composition, is conducting an ethnographic investigation of our project; see http://dmi.usf.edu/edminster/ETDProposal/.

Ultimately, we expect our research will reveal ways faculty and graduate students can use software tools and plug-ins to critique and develop theses and dissertations, including insights into necessary training and resources. We believe this work is an important first step toward transforming our graduate programs so they better prepare students for the Knowledge Age.

Results of our research can be viewed at our project home: http://dmi.usf.edu.


Digital Media Institute - Microsoft Internet Explorer provided by Dazzle* Inc
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Mail History Print Edit Real.com
Address http://dmi.usf.edu/participantList.asp
Go Links Best of the Web Channel Guide
Digital Media Institute
dmi.usf.edu
Awards
Workshops
Participants
Registration
Weekly Progress
Listserve
Chat
ETDguide.com
ETDguide.org
ETD's @ USF
Done
Internet
Start
Inbox
Digit...
default
Micros
3.2.2.1
RE: th
Last t
GuideF
2:22 PM

ETD Participant List

Name	email	Department	Professor			
Adriana M Hernandez	amherna3@eng.usf.edu	Industrial Engineering	Dr. O. G. Okogbaa	Link to Research	Progress Reports	Further Information
Alessandro Anzalone	anzalone@eng.usf.edu	Chemical Engineering	Dr. Carlos Smith	Link to Research	Progress Reports	Further Information
Amed Bailey	Aturtle2c@aol.com	Environmental Science and Policy	Dr. Brinkman	Link to Research	Progress Reports	Further Information
Brendan Calandra	bcalandr@luna.cas.usf.edu	Instructional Technology	Ann E. Barron	Link to Research	Progress Reports	Further Information
Brian Sterner	sterner75@hotmail.com	philosophy	turner	Link to Research	Progress Reports	Further Information
Bruce Cochrane	coch@chuma1.cas.usf.edu	Interdisciplinary Studies	N/A	Link to Research	Progress Reports	Further Information
Cindy LeRouge	lerouge@tampabay.rr.com	Information Systems/ Decision Science	Dr. Rosann Collins	Link to Research	Progress Reports	Further Information
Deanna Barcelona	dbarcelo@luna.cas.usf.edu	Anthropology	Alvin Wolfe	Link to Research	Progress Reports	Further Information
Debbie Hedgecock	dhedgec1@tampabay.rr.com	Aging Studies	Polivka & Schonfeld	Link to Research	Progress Reports	Further Information

Additional Readings

Moxley, Joseph M. "New Media Scholarship: A Call for Research." Change: The Magazine of Higher Learning (Scheduled Publication Date: November/December 2001)

Moxley, Joseph M. "Dissertating in a Digital Age: the Future of Composition Scholarship" Invited Chapter. *Reinventing the Discipline in Composition*

and Rhetoric and a Site for Change. Edited by Sheila Carter-Tod, Catherine G. Latterell, Cindy Moore, and Nancy Welch.

Moxley, Joseph M. American Universities Should Require Electronic Theses and Dissertations. (*Educause Quarterly, No. 3 2001, pp. 61-63.)*

Edminster, Jude and Joseph M. Moxley. Electronic Theses and Dissertations: An International Perspective on Best Practices. *Computers and Composition*.

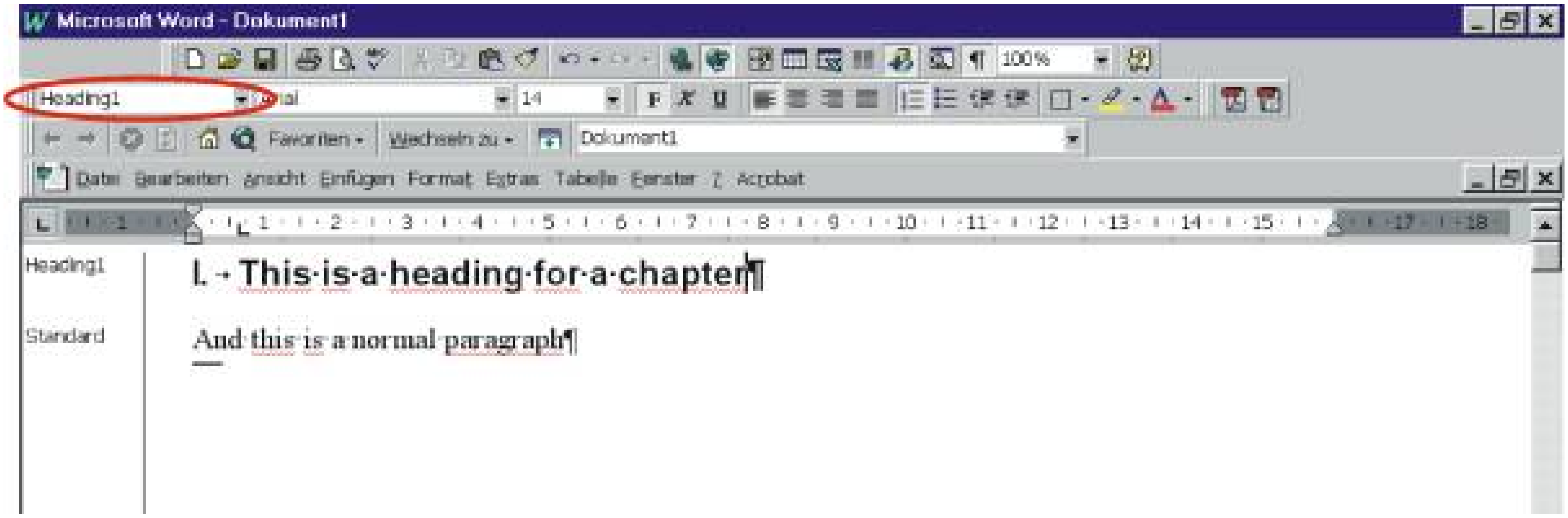
Moxley, Joseph M. The Role of Compositionists in Creating the Networked Digital Library of Theses and Dissertations. Texts and Technology. Janice Walker, Editor. Hampton Press.

3.2.2.1.1. Using Style Sheets, Suzanne Dolbratz

Preparing a word document to be converted into an archivable form using the SGML or XML standard first of all means using word stylesheets. Usually the university provides these style sheets, as they contain university specific structuring and formatting.

In Word it is possible to distinguish between the information a document holds and the structure in which it is written. Style sheets provide the structure. They help you, for example, if you want to format all your headings for chapters with the same style.

If you use the style sheet *heading1* (see picture below), than it may be possible to associate a certain formatting like *text height: 14pt, text-font Arial, paragraph settings: left bound, leave 12 pt space after a heading, number the headings automatically with roman numbers*, etc. with the structure element *heading1*.

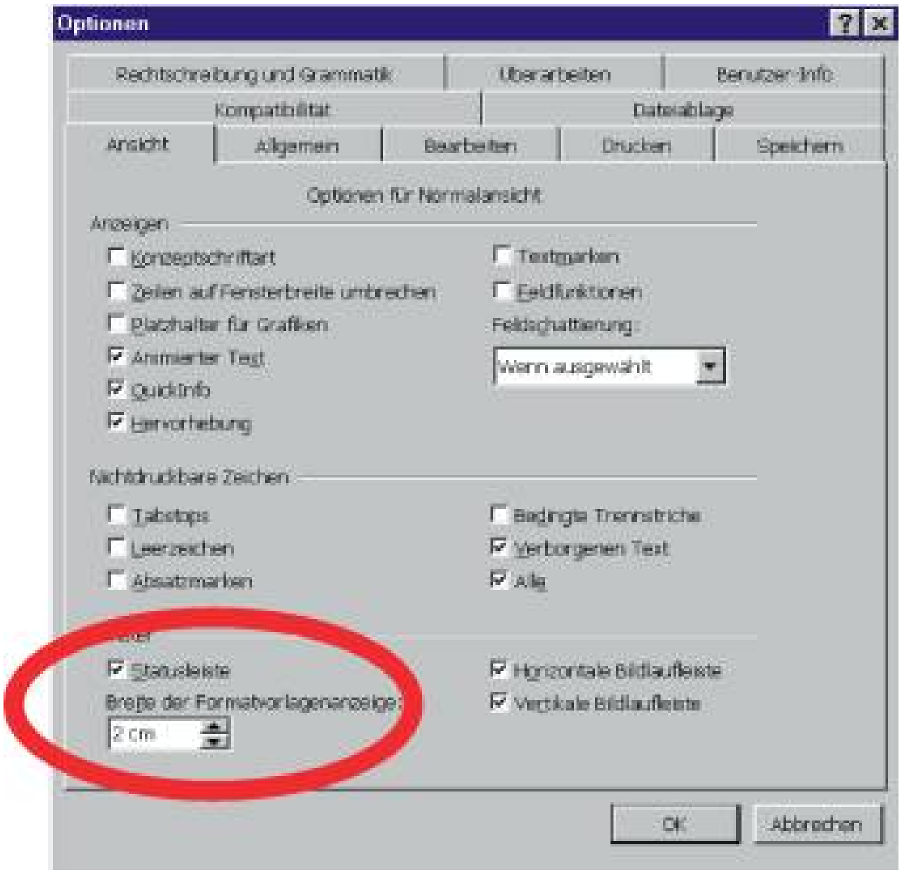

Microsoft Word - Dokument1
100%
Favoriten
Wechseln zu
Dokument1
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle Fenster ? Acrobat

Heading1 Arial 14

Heading1 I. → This·is·a·heading·for·a·chapter¶

Standard And this is a normal paragraph¶

You can see the left row within the word window. This is the so-called structured view, which you can get by choosing the Normal option under View. The names of the paragraph structures are displayed if you choose the point Options under Extras. Within the popup menu there is a possibility to set the width of the style sheet view. (see below). This is how it is displayed in German Word97, but is similar in other language word systems.



Optionen für Normalansicht

Anzeigen

- ☐ Konzeptschriftart
- ☐ Zeilen auf Fensterbreite umbrechen
- ☐ Platzhalter für Grafiken
- ☑ Animierter Text
- ☑ QuickInfo
- ☑ Hervorhebung
- ☐ Textmarken
- ☐ Feldfunktionen
- Feldschattierung: Wenn ausgewählt

Nichtdruckbare Zeichen

- ☐ Tabstops
- ☐ Leerzeichen
- ☐ Absatzmarken
- ☐ Bedingte Trennstriche
- ☑ Verborgenen Text
- ☑ Alle

- ☑ Statusleiste
- Breite der Formatvorlagenanzeige: 2 cm
- ☑ Horizontale Bildlaufleiste
- ☑ Vertikale Bildlaufleiste

3.2.2.1.2. Useful Plug-ins for Microsoft Word,

Jose H. Canos Cerda

Some word processing systems are extensible in different ways. One of them is the definition of macros written in some programming language. MS Word, for instance, uses Visual Basic for Applications as macro language*)*. Another way to extend MS Word is by *add-ins*, applications built in some programming language that are inserted into Word's runtime (like plug-ins in web browsers). This feature permits the development of added-value tools for writers. For instance, BibWord (http://mariachi.dsic.upv.es/bibword) is a bibliography management tool for MS Word 2000. The current version allows Word users to insert references from a bibliographic database and automatically generates the reference list at the end of a document. Macros also could be used to attach a form to the ETD document, the fulfillment of which would provide the ETD metadata from the word processing environment.

3.2.2.2. Corel WordPerfect, Susanne Dobratz

The Corel WordPerfect Suite provides another possibility to produce large documents. It is a system like Microsoft Word, but sold by another vendor: Corel Inc. (see http://www.corel.com). It contains nearly the same features and capabilities as Microsoft Word. For writing a complex document like a thesis or dissertation using WordPerfect, it is advisable to use style sheets that are provided by the system itself or by the university. Those style sheets are called WordPerfect templates (WPT). They allow the users to structure their documents using structure components like headings, tables, lists, etc. Default style sheets can be used by clicking on File / New / and than choose one of the templates.

Usually there has been a native WordPerfect template produced in order to help you with additional drop-down menus, etc. To use the provided template for WordPerfect 8, you have to choose File / New / Options / Add Project / Add New Document / Call it "Digital Dissertation" / Search for dissertation.WPT. This allows you to use the WordPerfect style sheet by choosing under the main window: File / New / Choose "Digital Dissertation" and Create. This enables the functions of the drop-down menu as in the following figure.

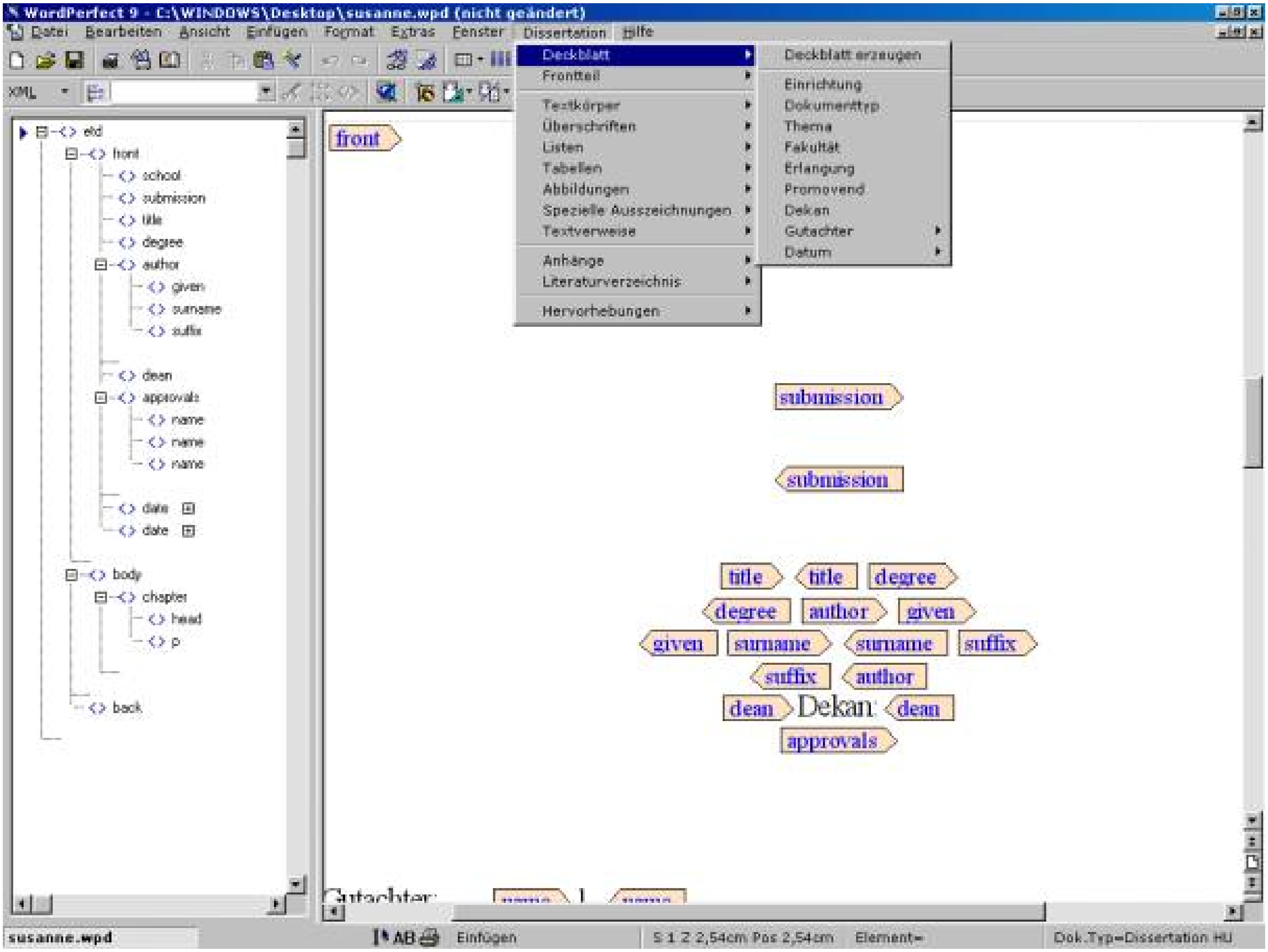


Figure 1: Using a WordPerfect style sheet for digital dissertations (used with version 9 of WordPerfect suite)

3.2.2.3 LaTeX, Susanne Dobratz

Scientists within the natural and engineering sciences have special needs for mathematics and algorithmic graphics. The text formatting system LaTeX has been used for decades to mark up scientific documents. Even today, there is no viewable alternative to print texts containing a lot of mathematics without using LaTeX. This system uses a kind of semantic or typographic markup for rendering formulas, graphs, and so on. Within some disciplines LaTeX is nearly exclusively used to render complex documents.

TeX and LaTeX

(words taken from Anthony Atkins latex-page)

TeX is a document formatting language (and the program that processes it) written by Donald Knuth for the professional preparation of complex publications. It excels particularly at formatting mathematical equations and for managing two-dimensional presentations of data (tabular and otherwise). LaTeX is a set of macros written by Leslie Lamport as a "front-end" to TeX that makes articles, reports, theses, dissertations, and books easy to create and manage.

How to get LaTeX

LaTeX is free to download from any CTAN archive (http://www.ctan.org), and works on Macintosh, MS-DOS, Unix, and Windows 3.1/95/NT (though some commands may vary on some architectures). To convert your electronic thesis or dissertation in LaTeX, you must first type your document completely into the ASCII editor using the LaTeX macros appropriately, then use a certain chain of commands that produce a layout and printable version of the document.

LaTeX under UNIX / LINUX systems

To create LaTeX files, all you need is an ASCII-based editor, like Emacs, Vi. Writing a dissertation just means typing the contents and the LaTeX-commands directly in an ASCII-based file and save this as *.tex.

To compile a LaTeX file and produce a printable version of the document, you have to follow the following steps:

1. Run latex "latex mydissertation.tex" This produces the following files: mydissertation.dvi / mydissertation.aux, etc.

2. Run dvips "dvips mydissertation.dvi" This produces a file dissertation.ps that is printable on a printer, or convertible into PDF.

While writing your thesis in LaTeX, please keep the following rules in mind:

As document style we advise to choose report or book, because both start with chapter as the highest order for section structuring. The preamble of the latex file could look like in the following example:

```
\documentclass[12pt,a4,titlepage]{book}

\usepackage{babel}

\usepackage{longtable}

\usepackage[dvips]{epsfig}
```

With usepackage we import additional styles that are needed, e.g. for tables, mathematics, figures etc. In order to get archivable form of the latex dissertations we advise not to use or to program complex macros. Simple \newcommand or \renewcommand may be used, e.g.

\newcommand{\begin{itemize}}{bi}

Headings can be separated using the following commands:

Document Structure	Level
\part{Heading Part I }	-1
\chapter{Heading Chapter 1}	0
\section{Heading Subchapter 1.1}	1
\subsection{Heading Section 1.1.1}	2
\subsubsection{...}	3
\paragraph{...}	4
\subparagraph{...}	5

Levels -1 to 2 appear in the table of contents. Part is used to split the whole document into several parts. The chapters numbering are constantly growing. Within the document than a single page is displayed, that contains: Part I Introduction or Part II Method and so on.

Chapters are numbered without taking the parts into account. The numbering is standardized: Chapter 1 Mathematics.

Sections are subunits of chapters and numbered: Basic Algorithms.

Sections are numbered as follows: 1.1.1 Decision Tree Algorithm A.

For those parts like acknowledgements, dedication, and curriculum vita where authors usually don't want to use numbering, the following style can be used:

\chapter*{Thank You} . The asterisk prevents the numbering.

Appendices are included using the \appendix command. Please use commands as in the following example if your appendix consists of several chapters:

\appendix or

\appendix* not numbered headings of the appendices

```
\chapter{Program Source}

\chapter*{Curriculum Vita}
```

Using graphics: figures and pictures should be included in LaTeX documents using the eps (encapsulated postscript) format. Before including them, one has to use a certain style package in the preamble:

```
\usepackage[dvips]{epsfig}
```

The parameter [h] positions the figure at the current position.

Keep in mind, always to use a caption-environment to put the figure captions below the picture:

```
\begin{figure}[h]

\begin{center}

\epsffile{didi.eps}

\end{center}

\caption[short description for the table of figures]{Long description for the text}

\end{figure}
```

The title page is the most complicated part. Most universities supply own templates for the title page and the whole dissertation. There is no best practice available. In order to separate the several items on a title page in order to be able to reuse those information pieces e.g. if the whole dissertation is converted to HTML or SGML/XML, we advise to use \newcommands as simplest method to apply a pseudo structure to a LaTeX title page. Usually LaTeX provides the following standards item for a title page:

```
\date{}

\author{}

\title{}
```

But as this is not enough for a thesis, most universities provide own style sheets or templates.

Tables should be used as follows: authors are advised to use the table-environment because it provides the possibility to include table captions in a structured way.

```
\begin{table}

\caption{Tabellenbeispiel}

\begin{center}

\begin{tabular}{ccc}

x & 1 & 2 \\ \hline

1 & 1 & 2 \\

2 & 2 & 4 \\ \hline

\end{tabular}

\end{center}

\end{table}
```

Citations can be used as their own structured items as follows:

1. Using the citation-environment. This is used for inline citations.

```
\begin{citation}{label1}

Contents

\end{citation}
```

2. Using the quotation-environment. This is used to structure whole paragraphs as citations. Those citations use an indent like usual paragraphs.

```
\begin{quotation}

contents

\end{quotation}
```

3. another method is the use of the quote-environment. This environment is used for whole paragraph citations, but those paragraphs don't have an indent.

```
\begin{quote}

content

\end{quote}
```

Numbered lists are typeset using the enumerate-environment. By integration new enumerate-environment in existing one a hierarchically nested sublist is built.

```
\begin{enumerate}
\item {Testitem1}
\begin{enumerate}
\item {Ebene 2 Testitem1}
\item  {Ebene 2 Testitem2}
\end{enumerate}
\item Testitem2
\end{enumerate}
```

Bulleted lists are typeset using the itemize-environment. Here a hierarchical nesting is also possible.

```
\begin{itemize}
\item Testitem1
\begin{itemize}
\item Ebene 2 Testitem1
\item Ebene 2 Testitem2
\end{itemize}
\item Testitem2
\end{itemize}
```

Definition lists contain a definition term and a definition text.

```
\begin{description}

\item[Definition term] Explanation of the definition term

\item[Element2] Explanation  2

\end{description}
```

If an author wants to include source code this is best done using the \verbatim-environment.

```
\begin{verbatim}

   #!/usr/bin/perl  -w

   #+----------------------------------+

   #| this script has been written 1998 by

   #+----------------------------------+

\end{verbatim}
```

Anchors, references and cross-references are typeset using the \label command, which links a key to the specified item of a document.

```
\begin{verbatim}

\label{keyword}

\end{verbatim}
```

References to these parts have to use the command \ref or \pageref in order to produce a reference to the object or to the page.

ref{keyword}

pageref{keyword}

A very important part of a dissertation is the bibliography. We advise all authors to use the bibtex-system and graphical front ends, e.g. bibview under LINUX or UNIX systems to manage bibliographic records and entries. References to bibliographic entries that are held in a bibtex-database are written as in the example:

\cite{schluessel}

The bibtex-database can be included into the LaTeX file by the following command, where a predefined style like alpha, plain, apalike can be used to layout the entries:

\bibliography{file name without .bib}

\bibliographystyle{style, e.g. alpha, plain, apalike, etc.}

Within the BibTeX-system database entries can be done using a plain ASCII editor. like emacs. There are several types of literature predefined:

Article in Conference Proceedings

Article in a Journal

Article in a Collection

Chapter or pages in a book

Conference Proceedings

Book

Booklet, but no Publisher, Institution

PHD Thesis

Masters Thesis

Technical Report

Technical Manual

Unpublished

The following example shows how a BiBTeX-entry has to be written:

```
%  Article in a Journal

@Article{shortkey2,

    author =      {Name, Firstname},

    title =       {Title No. 2},

    journal =     {Journal for ETDs},

    year =        {1999},

    OPTkey =      {},

    OPTvolume =   {},

    OPTnumber =   {},

    OPTpages =    {},

    OPTmonth =    {},

    OPTnote =     {},

    OPTannote =   {}
```

The following table shows those items have to be used for certain bibliographic entry types:

Category	Author	Title	Book title	Journal	Publisher	Year	Chapter	School	Institution
In Proceedings	X	X	X			X			
Article	X	X		X		X			
In Collection	X	X	X		X	X			
In Book	X	X			X	X	X		
Proceedings		X				X			
Book	X	X			X	X			
Booklet		X							
PhD Thesis	X	X				X		X	
Master Thesis	X	X				X		X	
Tec Report	X	X				X			X
Manual		X							
Unpublished	X	X							
Miscellaneous	X	X				X	Month	Note	How published

Table 1: Compulsory fields for bibliographic entry types

Category	**Optional**

In Proceedings	editor, volume, series, pages, address, edition, month, organization, publisher, note
Article	key, volume, number, pages, month. note, an note
In Collection	editor, volume, series, type, chapter, pages, address, edition, month, note
In Book	volume, series, type, address, edition, month, note
Proceedings	editor, series, address, edition, month, note
Book	volume, series, type, address, edition, month, note
Booklet	author, how published, address, month, year, note
PhD Thesis	type, address, month, note
Master Thesis	type, address, month, note
Tech Report	type, number, address, month, note
Manual	author, organization, address, edition, month, year, note
Unpublished	month, year
Miscellaneous	

Table 2: Optional fields for bibliographic entry types

To process a latex and a bibtex-file under UNIX system you have to type the following command sequence:

```
latex mydissertation.tex

bibtex mydissertation.aux

latex mydissertation.tex
```

This produces the following files: mydissertation.dvi / mydissertation.aux /mydissertation.bbl / mydissertation.blg, etc.

Run dvips "dvips mydissertation.dvi" This produces a file dissertation.ps that is printable on a printer, or convertible into PDF.

LaTeX under Windows operating systems

Using LaTeX under MS Windows requires the TeY System, a DVI-Viewer, Ghostscript and Ghostview. There are several LaTeX distributions: MikTeX , a highly regarded setup for Windows 95/NT (http://www.miktex.org/)and emTeX , the classic DOS and OS/2 TeX setup by Eberhard Mattes (ftp://ctan.tug.org/tex-archive/systems/msdos/emtex/).

There are front ends available for Latex that provide a WYSIWIG view to the user. One of the most used ones is Scientific Workplace by McKichan Software Inc. (http://www.mackichan.com/products/swp30.html). The disadvantage is that it is quite costly for single users.

3.2.2.4.2. Writing in Word processing systems: FrameMaker, Humboldt-University Berlin

Why should one use FrameMaker instead of MS Word?

Framemaker provides in comparison to Microsoft Word a much more sophisticated tool for electronic publishing and for a cross-media publishing.

It allows:

n Production of real and large structured text (not just address cards) (and is stable handling those).

n Production of semantically structured text (not just text that uses style sheets).

n Provision of a WYSIWYG user interface to edit XML documents, instead of an XML tree editor à la Spy or Xeena.

n Capabilities of producing a good-looking paper version of a document with all layout features that are professionally used at printing companies.

Framemaker+SGML especially combines the power of an excellent word processor (better than MS Word) with a good structure editor.

In order to produce structures documents and some sort of style sheets, one has to learn FrameMaker's EDD language.

As FrameMaker belong to the product family of Adobe, it provides an add-on to produce a high quality PDF, as digital preprint copy of a written document.

Using FrameMaker+SGML6.0 for a conversion of MS Word documents into SGML instances.

Editing or converting using FrameMaker is much more complex than the previously described methods. FrameMaker is able to import formatted Word documents keeping the stylesheet information and exporting the document via an internal FrameMaker format as SGML or XML documents.

In order to proceed with a conversion using FrameMaker, you will need the following configuration files: a conversion table that contains the list of the Word styles and the corresponding elements within the FrameMaker internal format. This table is saved within the FrameMaker internal format (*.frm).

A document type definition will be saved within FrameMaker internally as EDD (Element Definition). It is saved within the FrameMaker internal document format (*.edd).

FrameMaker uses layout rules for the internal layouting of documents. Within this layout definition, the layout of documents is described just like it is within MS Word documents: single formats and their appearances like text height, etc., are defined. This file is also stored as (*.frm file).

The Read-Write Rules contain rules that define which FrameMaker format will be exported in which SGML / XML element.

The SGML- or XML DTD has to be used as well, including Catalog- or Entity files, as well as Sub DTDs, like CALS for tables.

To process a conversion, a new SGML application has to be defined within FrameMaker+SGML. This application links all files that are needed for a conversion as described above. It enables FrameMaker to parse the output file when exporting a document to SGML or XML.

A workflow and a technology for conversion to ETD using FrameMaker+SGML6.0 was first developed at the Technical University Helsinki, within the HUTPubl project (1997-2000), see http://www.hut.fi/Yksikot/Kirjasto/HUTpubl.

You can find more information about using FrameMaker+SGML for an XML Authoring at http://tecfa.unige.ch/guides/xml/frame-sgml/html/quick-fm-xml-guide.html. (See also Danny R. Vint "SGML at Work", Prentice Hall, New Jersey, 1999.)

3.2.3. Writing directly in SGML/XML, Susanne Dobratz

The desired situation for retrieving archivable ETDs would be the one, that authors write in an XML-editor, according to a Main-DTD, and choose those parts of the DTD, that are inevitable for their thesis or dissertation.

Some desktop publishing systems today provide an opportunity to save as SGML or XML. Investigations as to whether those tools are useable for such complex documents as a thesis or a dissertation led to the following conclusion:

n Writing in WordPerfect or FrameMaker+SGML enforces the author to learn new writing habits. While writing, they have to think about the structure of their documents, e.g., which part is a heading, which part is a definition list; or they have to think to add certain parts immediately to the document, like, references, table and figure captions, etc.

n While writing according to a specified DTD, the desktop publishing system often internally checks syntax correctness by using an XML parser. Some of those internal parsers are still not stable enough and may cause the system to crash, as experienced with WordPerfect 9.0. Those parsing procedures in between the active scientific thinking and writing often disturbs the authors.

n Most of the pure XML editors could not produce an appropriate and layouted printed copy or PDF file that would satisfy the approval of readers of the printed version of a document. Some of those editors simply fail to process large and complex documents.

n Most of the tools are not ready yet, especially in a sense that would allow to use user or domain specific DTDs. Staroffice and other tools support their own vendor specific DTD only.

Although the world of desktop publishing systems is actually changing, there are still too few tools that are sufficient in:

- The support and appearance of their graphical user interface,
- The provision of a certain amount of features normal word processors have, like automated numbering, colors, table management, link management, style sheets.
- The platform independence or cross-platform availability,
- The support of user specific DTDs, and standard DTDs, like TEI, Docbook, etc.,
- The export quality of produced XML: tables, tags,
- The stability of usage,
- Their commercial availability and price.

The following systems are able to export into SGML or XML:

Export with user specific DTD

- WordPerfect Version 7.0 (Corel) http://www.corel.com
- FrameMaker+SGML6.0 (Adobe) http://www.adobe.com

Export with system specific (vendor) DTD

- Openoffice (SUN / open source) http://www.openoffice.org
- AbiWord (AbiWord / open source) http://www.abisource.com
- Kword (KOffice, KDE Project / open source) http://www.kde.org

Conversion Tools

- Omnimark (Omnimark) (http://www.omnimark.com)
- MarkupKit (Schema) http://www.schema.de
- Majix (Tetrasix) http://www.tetrasix.com
- TuSTEP (RZ Uni Tübingen) http://www.uni-tuebingen.de/zdv/tustep/index.html

(More information about SGML/XML tools can be found at: http://www.w3.org/XML/#software.)

3.2.4 Preparing a PDF Document, Edward Fox, Susanne Dobratz

A popular page representation scheme, a published de facto standard developed by Adobe is the Portable Document Format, PDF. Adobe provides the Acrobat Reader free of charge (and promised it into the foreseeable future), which will read current as well as previous versions of PDF. It is downloadable at http://www.adobe.com/products/acrobat/readstep.html. Adobe also provides tools for creating, annotating, and manipulating PDF documents, through its own word processing software, printer drivers, and distilling from PostScript. The whole suite is called Adobe Acrobat and is actually available with version 5. Adobe's Acrobat software, installed on a Windows or Macintosh 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.

It is also possible to produce PDF documents on UNIX systems. However, the latest version of Acrobat Distiller that was available for certain UNIX platforms such as Solaris or HP-UX was version 3.1. Authors writing in LaTeX can use ghostscript to produce PDF files. But in order to obtain readable PDF documents, issues of used fonts, used conversion scripts, etc. have to be considered.

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.

PDF-Tools:

- Acrobat (http://www.adobe.com)
- Ghostscript (http://archiv.leo.org/pub/comp/general/typesetting/tex/support/ghostscript/ or http://www.cs.wisc.edu/~ghost) and Viewer Tool ghostview (http://archiv.leo.org/pub/comp/general/typesetting/tex/support/ghostscript/gnu/ghostview/)
- NikNak (http://www.niknak.de/is/5dorder.htm)
- XPDF (http://www.footlabs.com/xpdf)
- Magellan/ Drake (http://bcl-computers.com)
- Gemini (http://www.iceni.com)
- Omnipage (http://www.scansoft.com)

3.2.4.2 From LaTeX, Susanne Dobratz

Generally speaking, there are several possibilities for producing PDF document from a LaTeX document.

1. Using Postscript and scalable fonts for PDF

"One of the most confusing issues in both Postscript and PDF is the handling of different types of fonts. A PDF-producing application can deal with a font in one of three ways: First it can take the entire font and embed it in the file; second it can make a subset; or third it can simply embed some summary details about the font (such as its name, metrics, its encoding, its type - sans serif, symbol, for example - and clues about its design) and rely on the display application to show something plausible. This last strategy is preferred for documents that are to be delivered on the Web, since it creates the smallest files. The display application can work again in several ways. It can try to find the named fonts on the local system; it can simply substitute fonts as intelligently as possible; or it can use Multiple Master fonts to mimic the appearance of the original font." (from Goosens; Rahtz: The LaTeX Web Companion, page 29)

The default installation of dvips uses fonts with a fixed resolution (.pk fonts) encoded as 300dpi (dots per inch) bitmaps. This is unnoticeable for printing; however, the resulting PDF files are barely legible when scaled down to today's screen resolutions (typically 72dpi). These fonts are embedded in Postscript Output as Type 3 fonts. Acrobat Distiller cannot handle those fonts, because there are no font descriptors available. It leaves them embedded in PDF files and renders them very badly, although printing those documents doesn't make too many differences, if the original resolution was high enough.

Therefore it is necessary to install Postscript Type 1 fonts (True Type) for the dvips program. Many commonly used fonts have been converted to Type 1 fonts, e.g.: All Computer Modern family fonts, all fonts from the American Mathematical Society, the St. Mary's Road symbol fonts, the RSFS script fonts, the TIPA phonetic fonts and the XY-pic fonts.

The Type 1 Computer Modern fonts are provided by Virginia Tech and part of this guide (cmps.tgz /

cmps.tar.gz). These files are about 5 MB.

To install the fonts you have to…

On standard LINUX systems they are already installed:

1. Copy all files which are in the gz-archive under the directory pfb into the directory, in which dvips looks for fonts, e.g. /usr/local/teTex/texmf/fonts.

2. In the directory e.g. /usr/local/teTex/texmf/dvips/misc there is a file psfonts.map. Please add the content of the files cmfonts.map, cyrfonts.map, eufonts.map,and lafonts.map to that file. They are provided with this cmps.tgz archive.

3. The config.ps file is usually used for defining the resolution. This is irrelevant, because dvips now uses the scalable fonts instead of the bitmapped pk fonts.

4. The afm und pfm directory in the archive is not used by dvips.

To obtain a ps-file which uses Postscript fonts and is convertible into PDF you have to run the following command sequence:

1. latex mydissertation.tex.

2. bibtex mydissertation.aux if bibtex is used.

3. latex mydissertation.tex.

4. dvips -P cmz mydissertation.dvi: This produces a file dissertation.ps that is printable on a printer, or convertible into PDF.

5. If Acrobat Distiller is installed on the system "distill mydissertation.ps" which produces a PDF file: mydissertation.pdf.

2. Producing Rich PDF

Producing a WWW-readable PDF is just the first part of a PDF production. It is more sophisticated to produce a PDF file that takes advantage of the hypertext features of PDF and adds links and cross-references to a PDF file.

You can use the Adobe Exchange software under Windows/Macintosh to add links, etc., to a ready produced PDF file, or you can produce those features directly from LaTeX using the Hyperref-package. This package has been developed by Sebastian Rahtz and uses the outcome of the Hypertex project.

This package extends the capabilities of the LaTeX cross-referencing commands (TOC,bibliographies, etc.) to produce \special commands that a driver can turn into hypertext links. It also defines new commands for LaTeX.

For using hyperref a global option can be used within the LaTeX file:

```
\documentclass[dvips]{article}

\usepackage{hyperref}
```

In order to produce PDF-information, it is possible to insert title and author information that are then displayed in the PDF file as follows:

In LaTeX:

```
\documentclass[dvips]{article}

\usepackage[

\usepackage[

pdfauthor={Susanne Dobratz},

pdftitle={ Test of the  pdftex Package },

pdfcreator={pdftex},

pdfsubject={electronic publishing in LaTeX},

pdfkeywords={keyword1,keyword2}
]{hyperref}
```

This looks in PDF like this:

```
%PDF-1.2

%âãÏÓ

1 0 obj

<<

/CreationDate (D:191010522170228)

/Keywords (keyword1,keyword2)

/Creator (pdftex)

/Title (Test of the  pdftex Package)

/Producer (dvips + Distiller)

/Author (Susanne Dobratz)

/Subject (electronic publishing in LaTeX)

>>
```

The usual \label and \autoref commands are used to produce hyperlinks. The \autoref-command replaces therefore the usual \ref-command in LaTeX. So the following document structures are automatically referenced, if a \label has been applied. This also automatically produces Adobe-PDF bookmarks and hyperlinks to chapters, sections, etc. if the LaTeX command \tableofcontents is used.

Within the LaTeX file there are some additional user macros available to produce hyperlinks:

\href{url}{text}	The *text* is used a hyperlink to the *url* . This URL must be a full URL (like http://www.cybertheses.org)
\hyperbaseurl{url}	A base url is established, prepended to other specified URLs to make it easier to write PDF documents.
\hyperimage{image url}	The image referenced by the image url is inserted.

\hyperdef{category}
{name}{text}
	A target area of the document (text) is marked and given the name category.name

\hyperref{url}{category}
{name}{text}
	The text is made into a link to url#category.name

\hyperref[label]{text}	The text becomes a link point to a point established with a \label command (using the symbolic name label).

It is even possible to use Acrobat specific commands, e.g.menu options to navigate etc., like in this example from Sebastian Rahtz:

```
\usepackage{fancyhdr}

\pagestyle{fancy}

\cfoot{\NavigationBar}

\newcommand{\Navigationbar}{%

\Acrobatmenu{PrevPage}{previous}~

\Acrobatmenu{NextPage}{next}~

\Acrobatmenu{FirstPage}{first}~
\Acrobatmenu{LastPage}{last}~

\Acrobatmenu{GoBack}{back}~

\Acrobatmenu{Quit}{quit}%}
```

For further information and help, we recommend the book by Goosens/ Rahtz: The LaTeX Web Companion.

The \special commands that are added by using the LaTeX macros have to be interpreted by DVI drivers or viewers in order to produce PDF links.

The following DVI drivers are supported by the hyperref package:

hypertex

dvips
writes \special commands to Postscript tailored for dvips

dvipsone
writes \special commands to Postscript tailored for dvipsone

pdftex
writes commands for pdftex, and produces PDF directly

dvipdfm
writes \special commands to be used for Mark Wicks' DVI to PDF driver dvipdfm

dviwindo
writes \special commands to be used for Y&Y's Windows previewer. It interprets them as jumps within the previewer

vtex
writes \special commands, which are interpreted as hypertext jumps for MicroPress'HTML and PDF producing TeX variants

3. Using PDFTeX

PDFTex is a variant of Tex that produces directly a PDF output. Usually a Latex or Tex system produces a DVI output. PDFTex can also produce DVI output.

You may use pdfTex instead of LaTex using macro packages as context or hyperref or others to write the actual document.

"When producing DVI output, for which one can use pdfTex as well as any other Tex, part of the job is delegated to the DVI postprocessor, either by directly providing this program with commands, or by means of \specials. Because pdfTex directly produces the final format, it has to do everything itself, from handling color, graphics, hyperlink support, font-inclusion, up to page imposition and page manipulation. As a direct result, when on uses a high level macro package, the macros that take care of these features have to be set up properly.

Currently all mainstream macro packages offer pdfTex support in on way or the other. When using such a package, it makes sense to turn on this support in the appropriate way, otherwise one cannot be sure if things are set up right." (from the pdfTex User manual at http://www.tug.org/applications/pdftex/pdftexman.pdf).

The following main macro packages support pdfTex:

for LaTeX users the hyperref package by Sebastian Rahtz

the standard LaTeX graphics and color packages have pdfTex options

the ConTeXt macro package by Hans Hagen has extended support for pdfTex

Literature and Sources:

http://www.tug.org/applications/pdftex/

Michael Goosens; Sebastian Rahtz: The LaTeX Web Companion, Addison-Wesley, 1999: ISBN: 0-201-43311-7

3.2.5 Preparing for Conversion to SGML/XML,

Susanne Dobratz

Section 4.3.5.3 defines SGML and XML.

The concept of Document Type Definitions (DTDs)

A **document type definition (DTD)**, in the sense of XML, defines rules or templates, which are used to produce similarly, structured documents. A DTD describes the content model of a class of documents. It consists of:

n An **element declaration**, which is the main part of a DTD and the structural definition. Elements can contain other elements, characters or nothing. Element declarations define the name of the element and the logical content (sub elements) of an element. (See [**10**].) An important part of the element declaration is the content model. It is here that the document architect indicates the order and occurrence of other element or character data.

n A **notation declaration**, which defines a notation for external formats, e.g., for graphics (gif, jpeg), mathematics (TeX, LaTeX), 3D objects (VRML) and other formats, that cannot be coded directly in XML.

n An **entity declaration**, which defines character, sets and replacement objects for characters. Everything from a single character on up can be defined with a single entity. There are two basic types of entities: general and parameter. Parameter entities are only allowed in declarations, and are usually used to make a DTD more readable or to control processing. General entities are used in the document instance; the documents build upon the DTD.

n An **attribute list declaration**, where attributes and their values for the different element types defined in the element type declaration is listed.

To define a DTD, a special syntax is needed, which does not conform to the usual XML syntax where a document contains elements which are enclosed in "tags:" a start tag (e.g. <author>) and an end tag (e.g. </author>), producing code like this: <author> Joe Miller </author>

DTDs for electronic dissertations used worldwide

The fact that currently available authoring systems for XML still have not won wide recognition has led to different strategies at different universities regarding XML documents. Most of these projects were started between 1995 and 1997, in a time when XML was alive, but where no tools or standardized DTDs were available. A view of those projects from today’s perspective illustrates the demand for a rethinking and redesign of those approaches in order to come to a standardization.

All the presented DTDs are built upon similar principles. A classical dissertation (which can be seen as monograph) consists of 3 main components: an extensible **title page** with abstracts, declarations, etc., the **dissertation corpus**, which includes text, pictures, audio, video, tables and so on, as well as the **appendices**, which contain data sheets, bibliographies, acknowledgements and others.

The following DTDs are currently in use at different institutions:

n **ETD-ML.DTD**: Virginia Polytechnic Institute and State University (Virginia Tech)

n **DiML.DTD:** German Dissertationen Online Projectes

- **TDM.DTD**: University of Iowa

- **HutPubl.DTD**: Technical University Helsinki

- **TEI-Light.DTD**: Ann Arbor und Lyon

- **ISOBook.DTD:** University of Oslo

- **TEI-based DTD with extensions for natural sciences**: Swedish University of Agricultural Sciences Uppsala

All those Document Type Definitions are so-called author-DTDs. This means that they are primarily used to support the authoring and the conversion process and do not first of all address document archiving and preservation issues. One may ask why all those different DTDs have prevailed. This is mainly because the scientific orientation of the mentioned universities is quite varied. Lyon, Oslo and Michigan, which use TEI-Light.dtd, mainly serve students in the arts and humanities. Problems using TEI.DTD or DocBook.DTD are recognized at universities, which support a strong natural science community, such as Berlin, Helsinki or Uppsala. Often a dissertation is a cumulative work, e.g., in Lyon or Helsinki.

Preparing for Conversion

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.

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.

Bibliography

[1] http://lcweb.loc.gov/cds/lcsh.html#lcsh20

[2] http://www.bibliothek.uni-regensburg.de/rvko/rvko.php3

[3] http://purl.org/DC/

[4] http://www.w3.org/rdf

[5] Edward Fox: Networked Digital Library of Theses and Dissertations, Web matters, Aug., 12th 1999, http://helix.nature.com/webmatters/library/library.html

[6] Website of the standards committee of NDLTD: http://www.ndltd.org/standards/

[7] http://dochost.rz.hu-berlin.de/epdiss/dtd-workshop/index.html

[8] Tad Lane, Scalable Vector Graphics - Web Graphics with Original-Quality Artwork, in: BITS, November 1999, http://lanl.gov/orgs/cic/cic6/bits/november_99/novbits1.html

[9] Neill Kipp: Beyond the Paper Paradigm: XML and the Case for Markup; in: Part II "Guideline for Writing and Designing ETDs" ETD Sourcebook, Weisser, Moxley and Fox editors, 1999

[10] B. Travis, D. Waldt: The SGML Implementation Guide, Springer, Berlin-Heidelberg-New York, 1995

[11] Ed Dumbill: The State of XML, June, 16th, 2000 in XML.com, http://www.xml.com/pub/2000/06/xmleurope/keynote.html

3.2.5.1 Preparing for Conversion to SGML/XML from MS Word, Susanne Dobratz, Viviane Bouletreau

Performing a conversion from MS Word documents into instances of a specified SGML or XML DTD is a very complex task. What you will need for that is:

n A SGML or XML document type definition (DTD) that serves as structure model for the output. One says that the output SGML document is valid to the specifies DTD, or it is an instance of this DTD:

n A Word style sheet that holds paragraph and character styles according to the structures in the DTD. So if in a DTD you have defined a structure for Author (e.g. expressed in the output file as):

```
<author>
<title>Dr.</title><firstname>Peter</firstname><surname>Fox</surname>
</author>
```

You have to find expression in Word:
paragraph styles: author
character styles (just to be used within an author-paragraph): firstname, surname, title

n You will need some kind of a configuration file that allows the mapping of the DTD elements into Word elements and vice versa.

n You will need an SGML or XML parser to check the output SGML/ XML document against the DTD.

Often a conversion is done by using a plug in to MS Word directly. But other options use the Microsoft internal exchange format RTF (Rich Text Format) for conversion. Those tools can interpreted the RTF file with the MS Word style that are still coded in this RTF file and export it into an SGML document. This process mostly happens within batch mode without using much graphical user interfaces.

Within the following paragraphs we describe several approaches:

1. Approche of the Université de Montréal, Université de Lyon 2, Universidad de Chile

2. Humboldt-University Berlin and Germanwide Dissertation Online project

There are other approaches in development as well, especially within Scandinavia and the University of Oslo/ Norway. We don't refer to their solution yet.

Conversion method of the Cyber theses project

The process line for converting Word files into SGML documents developed within the CyberThèses project uses scripts written with the Omnimark language.

The input of the process line is an RTF file with a "structuring style sheet" and the output is an SGML document encoded according to the TEI Lite DTD (see the TEI web site at http://etext.virginia.edu/TEI.html).

The conversion process is constituted of three main steps :

n a first one converts the RTF file into a flat XML file encoded according to DTD of RTF. The produced file is a linear sequence of paragraph elements having each one an explicit "style name" attribute corresponding to the RTF style names.

n the second step consists in the re-generation of the hierarchical and logical structure of the document based on the analysis of style name attribute.

n last, a SGML parser allows validating the conformity of the produced SGML document with the TEI Lite DTD.

Some supplementary scripts then allow the export of the SGML document towards other formats (HTML, XML).

Most of the scripts are available from the CyberTheses web site : http://www.cybertheses.org (well, actually there will be soon…)

This system is devoted to a particular DTD, but its generalization to other document models shall not raise any difficulty.

Using SGML Author for Word (Humboldt-University Berlin)

Why did we use the SGML Author for Word?

The "Dissertation Online" project implemented and refined a conversion strategy that allows to convert documents written in MS word with a special style sheet (dissertation.dot) into an SGML instance of the DiM.dtd.

We used this product from Microsoft, the SGML Author for Word, due to several reasons:

1. SGML Author is quite easy to configure

2. It is easy to use.

3. It is less expensive than other software producing SGML files with the same quality.

4. It supports an international standard for tables: CALS.

5. As it is a Word-Add-On it handles documents in MS-Word doc- format better than other tools.

6. As we started using this technology in 1997, it supported from the very beginning Word97, the version of word, which was the actual one that time.

Unfortunately, Microsoft didn't continue the development of this tool. So there are new versions available for Office 2000 or Office XP. But the internal document format from MS Word 97, MS Word 2000 and Office XP are the same in the sense of the conversion into SGML. This means documents written in Word 2000 or Office XP can be imported into Word97 and therefore a conversion can be done.

Preconditions:

For a successful conversion from a word document into a DiML document you will need:

n The DiML-document type definition (diml20.dtd, calstb.dtd)

n the SGML-Author for Word97 (may not available at Microsoft Shops any more, but NDLTD esp. Prof. Dr. Edward Fox may provide English versions of it that work with English Word)

n The Association file for the Microsoft SGML-Author for Word (diml20.dta)

n The converter style sheet, which consists of several macros programmed to make the preconversion process easier.

n The perl programming language (free Software)

- The nsgmls-Parser (free Software)

- Several perl scripts to correct the transformation of tables.

You must have the following software installed at you computer:

- SP (NSGMLS) (Parser for SGML-Files by James Clark). (new version are available at http://openjade.sourceforge.net/doc-1.4/index.htm, but we haven't tested that)

- Run SP (A WYSIWYG tool for SP by Richard Light). http://www.light.demon.co.uk/runsp/index.htm

- Perl (a scripting language for using the perl scripts).

The converter style sheet and the author's style sheet can be obtained from the following website: http://dochost.rz.hu-berlin.de/epdiss/vorlage.html

Converter scripts and perlscripts can be obtained from http://www.educat.hu-berlin.de/diss_online/software/tools.exe (Perl scriptc, DTD and converter file for MS SGML-Author for Word - KonverterDiML2_0.dta)

Conversions

The conversion from a Microsoft Word document into a SGML document, which is an instance of the DiML.dtd that is used at Humboldt-University, takes several steps:

1. Step: Preparing the conversion without using *the converter Microsoft SGML Author for Word* directly

Check the correct usage

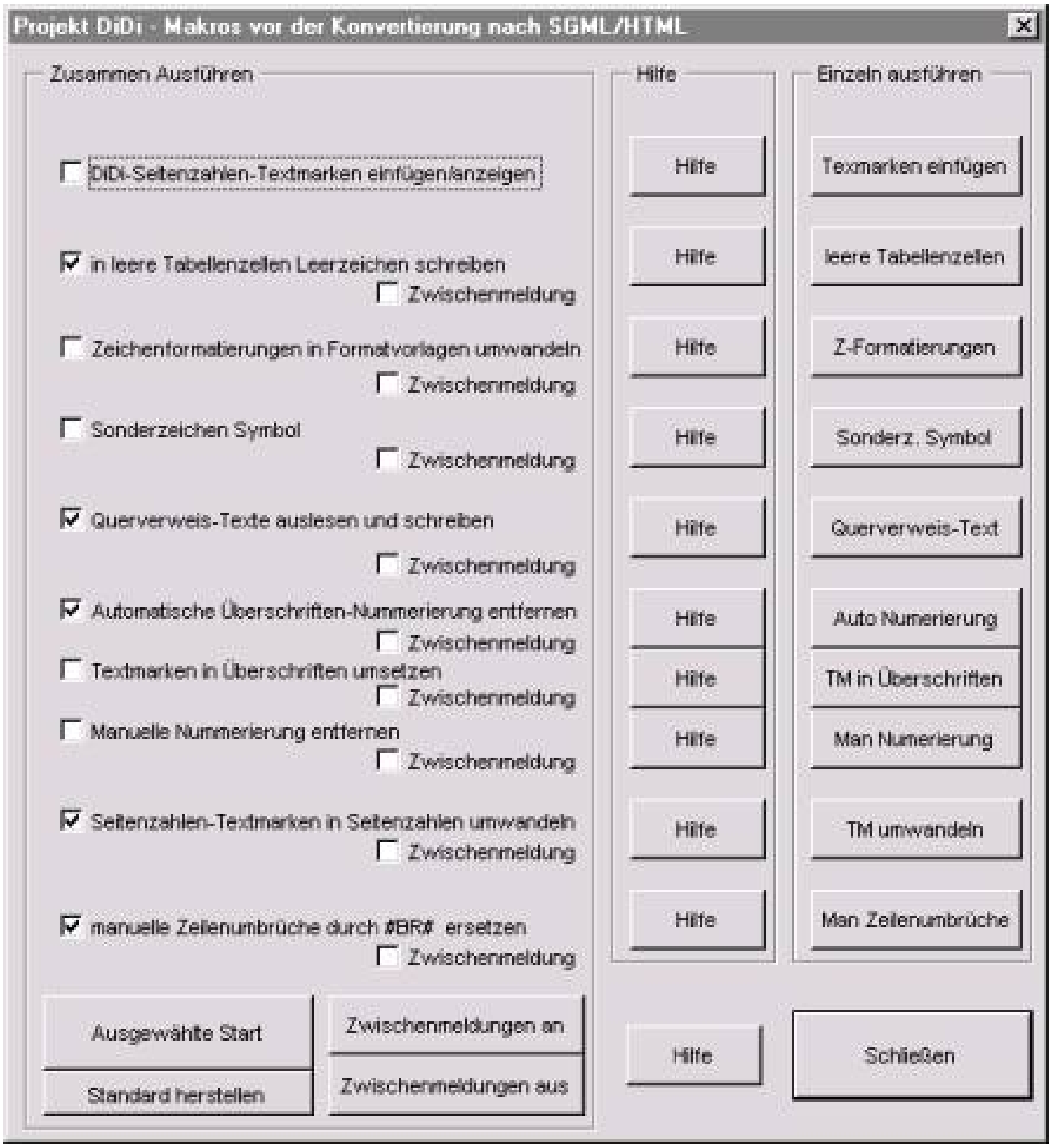
Load the style sheet for conversion (NOT the one for the authors) see, see figure below.

There is a special feature to get the page numbers out of the Word document by using certain word specific text anchors. Those have to be converted into hard coded information using a page number style sheet.

Formatting that has been applied by the author without using style sheets have to be replaced by the correct style sheets.

In order to get a correct display of tables later on by using CSS style sheets within common browsers, empty table cell have to be filled up with a single space (letter).

Soft coded line breaks have to be preserved for the conversion. This is done by inserting special characters #BR# to that. This will be used to insert later a special SGML tag for soft line breaks <br/>.



Projekt DiDi - Makros vor der Konvertierung nach SGML/HTML

Zusammen Ausführen

☐ DiDi-Seitenzahlen-Textmarken einfügen/anzeigen

☑ in leere Tabellenzellen Leerzeichen schreiben
☐ Zwischenmeldung

☐ Zeichenformatierungen in Formatvorlagen umwandeln
☐ Zwischenmeldung

☐ Sonderzeichen Symbol
☐ Zwischenmeldung

☑ Querverweis-Texte auslesen und schreiben
☐ Zwischenmeldung

☑ Automatische Überschriften-Nummerierung entfernen
☐ Zwischenmeldung

☐ Textmarken in Überschriften umsetzen
☐ Zwischenmeldung

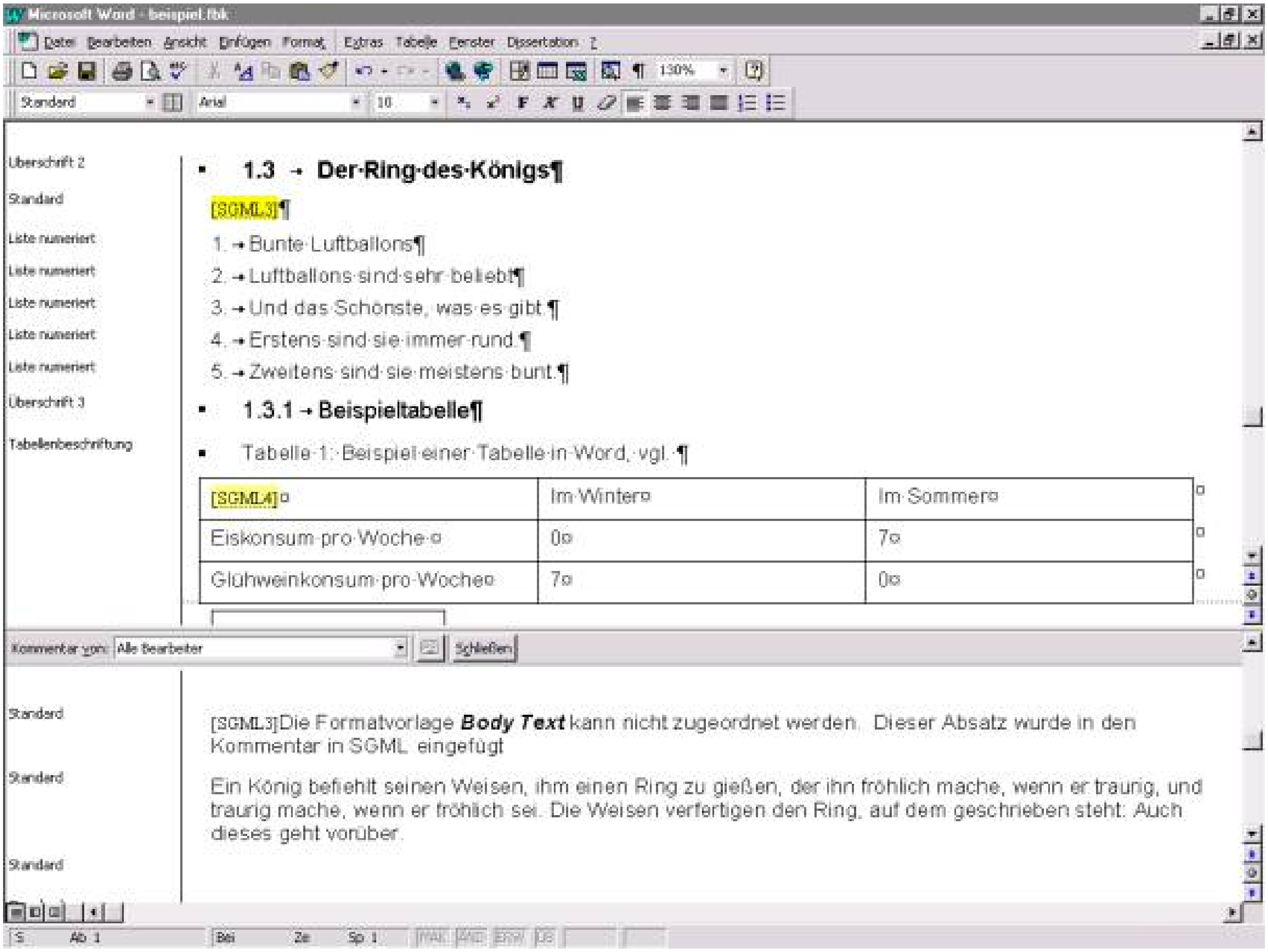
☐ Manuelle Nummerierung entfernen
☐ Zwischenmeldung

☑ Seitenzahlen-Textmarken in Seitenzahlen umwandeln
☐ Zwischenmeldung

☑ manuelle Zeilenumbrüche durch #BR# ersetzen
☐ Zwischenmeldung

2. Step: Converting with Microsoft SGML Author for Word

- Press the button "Save as SGML" within the FILE menu.
- Load the converter file KonverterDiML2_0.DTA
- Check the XML/SGML output using the feedback file (fbk) see figure below.



1.3 → Der Ring des Königs¶

[SGML3]¶

1. → Bunte Luftballons¶
2. → Luftballons sind sehr beliebt¶
3. → Und das Schönste, was es gibt.¶
4. → Erstens sind sie immer rund.¶
5. → Zweitens sind sie meistens bunt.¶

1.3.1 → Beispieltabelle¶

Tabelle 1: Beispiel einer Tabelle in Word, vgl. ¶

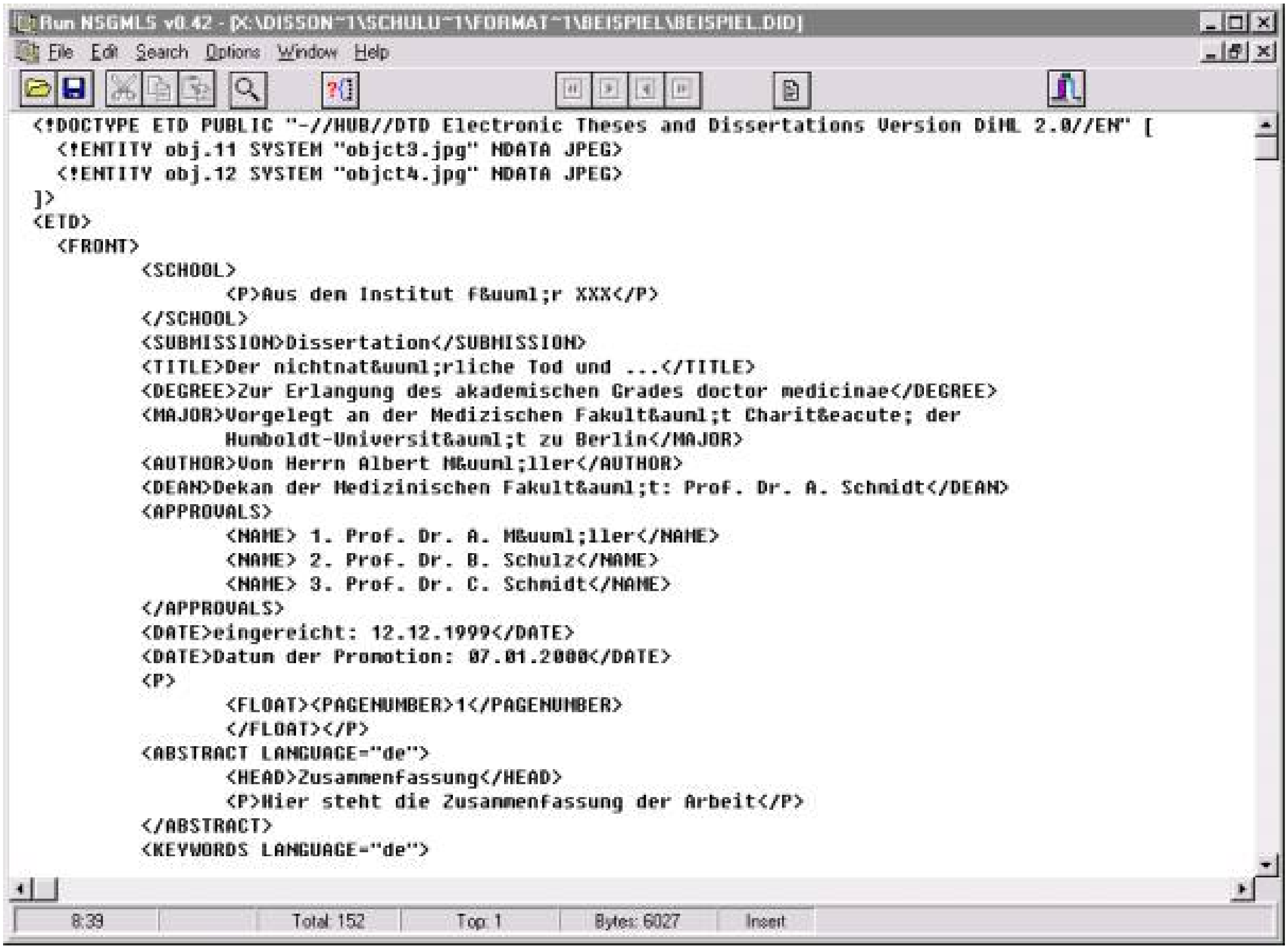
[SGML4]	Im Winter	Im Sommer
Eiskonsum pro Woche	0	7
Glühweinkonsum pro Woche	7	0

[SGML3]Die Formatvorlage ***Body Text*** kann nicht zugeordnet werden. Dieser Absatz wurde in den Kommentar in SGML eingefügt

Ein König befiehlt seinen Weisen, ihm einen Ring zu gießen, der ihn fröhlich mache, wenn er traurig, und traurig mache, wenn er fröhlich sei. Die Weisen verfertigen den Ring, auf dem geschrieben steht: Auch dieses geht vorüber.

3. Step: Work through the output file (output according to the DiML.dtd)automatically.

- Load the perlskripts using the batch file preprocessor.bat
- Parse the DiML file
- Errors have to be wiped out manually



Run NSGMLS v0.42 - [X:\DISSON~1\SCHULU~1\FORMAT~1\BEISPIEL\BEISPIEL.DID]
File
Edit
Search
Options
Window
Help
8:39
Total: 152
Top: 1
Bytes: 6027
Insert

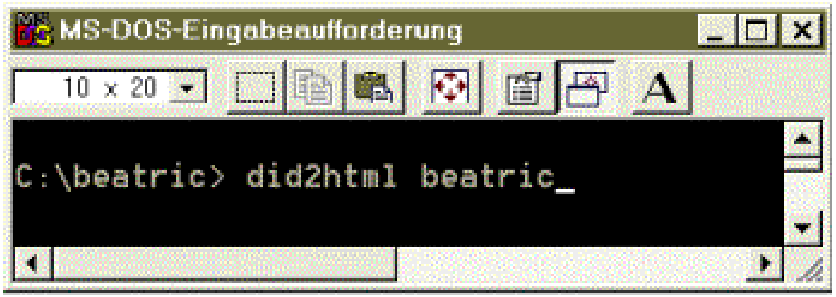
```
<!DOCTYPE ETD PUBLIC "-//HUB//DTD Electronic Theses and Dissertations Version DiML 2.0//EN" [
  <!ENTITY obj.11 SYSTEM "objct3.jpg" NDATA JPEG>
  <!ENTITY obj.12 SYSTEM "objct4.jpg" NDATA JPEG>
]>
<ETD>
  <FRONT>
         <SCHOOL>
                <P>Aus dem Institut f&uuml;r XXX</P>
         </SCHOOL>
         <SUBMISSION>Dissertation</SUBMISSION>
         <TITLE>Der nichtnat&uuml;rliche Tod und ...</TITLE>
         <DEGREE>Zur Erlangung des akademischen Grades doctor medicinae</DEGREE>
         <MAJOR>Vorgelegt an der Medizischen Fakult&auml;t Charit&eacute; der
                Humboldt-Universit&auml;t zu Berlin</MAJOR>
         <AUTHOR>Von Herrn Albert M&uuml;ller</AUTHOR>
         <DEAN>Dekan der Medizinischen Fakult&auml;t: Prof. Dr. A. Schmidt</DEAN>
         <APPROVALS>
                <NAME> 1. Prof. Dr. A. M&uuml;ller</NAME>
                <NAME> 2. Prof. Dr. B. Schulz</NAME>
                <NAME> 3. Prof. Dr. C. Schmidt</NAME>
         </APPROVALS>
         <DATE>eingereicht: 12.12.1999</DATE>
         <DATE>Datum der Promotion: 07.01.2000</DATE>
         <P>
                <FLOAT><PAGENUMBER>1</PAGENUMBER>
                </FLOAT></P>
         <ABSTRACT LANGUAGE="de">
                <HEAD>Zusammenfassung</HEAD>
                <P>Hier steht die Zusammenfassung der Arbeit</P>
         </ABSTRACT>
         <KEYWORDS LANGUAGE="de">
```

4. Step: Transforming the DiML file into a HTML file

n Load the perl scripts by using the batch file did2html.bat

n Check the HTML Output.

n Correct possible errors manually within the SGML file and repeat the transformation.



```
C:\beatric> did2html beatric_
```

A demonstration QuickTime video may be found at the ETD-Guide server as well. (see http://www.educat.hu-berlin.de/diss_online/software/didi.mov)

Other Tools

Text editors, Desktop Publishing Systems that can export SGML/XML documents

Tools that export using a user specified [1] DTD:

WordPerfect since Version 7.0 (Corel http://www.corel.com)

FrameMaker+SGML6.0 (Adobe) (http://www.adobe.com)

Tools that exports using their own native[2] DTD:

Openoffice (SUN/open source) (http://www.openoffice.org)

AbiWord (AbiWord/ open source) (http://www.abisource.com)

Kword (KOffice, KDE Project/ open source) (http://www.kde.org)

Converter Tools:

Omnimark (Omnimark) (http://www.omnimark.com)

MarkupKit (Schema) (http://www.schema.de)

Majix (Tetrasix) (http://www.tetrasix.com)

TuSTEP (RZ Uni Tübingen) (http://www.uni-tuebingen.de/zdv/tustep/index.html)

Literature

[1] Bollenbach, Markus; Rüppel, Thomas, Rocker, Andreas: FrameMaker+SGML5.5. Bonn; Reading, Mass., Addison-Wesley Longman, 1999, ISBN 3 8273 1508 5

[2] St. Laurent; Biggar, Robert: Inside XML DTDs. New York, McGraw Hill, 1999, ISBN 0 07 134621 X

[3] Ducharme, Bob: SGML CD. New Jersey, Prentice Hall, 1997, ISBN 0 13 475740 8

[4] Smith, Norman E.: Practical Guide to SGML/XML Filters. Plano. Texas, Wordware Publishing Inc., 1998, ISBN 1 55622 587 3

[5] Goldfarb, Charles; Prescod, Paul: XML Handbuch. München, Prentice Hall, 1999, ISBN 3 8277 9575 0

[1] Mit diesen Programmen kann die DiML.dtd des Projektes Dissertationen Online verwendet werden.

[2] Mit diesen Programmen ist kein Export in eine DiML.dtd -konforme Instanz möglich. Es werden nur herstellereigene DTDs verwendet.

3.2.5.2 In WordPerfect, Susanne Dobratz

WordPerfect supports structured writing since version 7. Most of the following text has been taken from a Whitepaper by Corel Inc. that appeared for WordPerfect 9 in June 1999.

Writing XML using WordPerfect means that the author is forced to use specific structures. The software will parse the file by an underlying XML parser and check correctness of the written file while it is being written by the author. This may cause the system to slow down sometimes because the parsing process may take a lot of the system's resources during that time. Generally, the writing behavior is different than the one for systems like Microsoft Word, where no parsing during runtime is being performed. As the author, you will have to think about your document structure, on how the information pieces are put together and follow each other (e.g. if something is supposed to be a heading or the beginning of a new chapter).

Standard WordPerfect 9 templates allow users to embed XML components. These WordPerfect templates include the following XML components:

1. Document Type Definition (DTD)
2. Compiled version of the DTD, a Logic file (LGC)
3. Layout Specification Instance (LSI)
4. Alias File (LNM)

The **Document Type Definition (DTD)** defines the elements and the structured relationship between the elements, entities and attributes. The DTD defines all valid elements; the order in which they can be used and how many times a particular element can appear in a document. When compiled as binary logic file (LGC) by the Word Perfect DTD Compiler, the DTD is integrated to a WordPerfect Template along with the Layout Specification Instance File and the Alias File.

The **Layout Specification Instance (LSI)** specifies formatting information, such as bold, underline, italics etc., for the start and end tags. Layout files can associate elements with their respective WordPerfect styles. It is possible to insert a specific WordPerfect formatting command or a text string before, after or in place of a specified XML element. The .lsi file can run macros when users insert a specific element, so template designers can build dialog boxes, prompts and other help to the authors in order to support them in writing correct XML documents. It is possible to use several .lsi files with the same DTD. This is useful if multiple output formats have to be produced from the file, e.g. one for printing and one for online publishing. When compiled the .lsi file integrates to a WordPerfect template.

An **Alias** file (LNM) specifies the descriptive names for elements in the DTD. This is useful when the tag names defined in the DTD are not appropriate for the end user. This may be because templates are created for end users who speak a different language or are created for a different audience. For instance non-technical users may relate to different tag names than technical users.

WordPerfect also provides standards templates to the user and allows the creator to customize the user interface with menu items, toolbars and view of specific documents. All those elements may include and run macros.

Example:

An example of a DTD, a LSI file and a WordPerfect template for digital dissertations used in Germany (Humboldt-University Berlin) is provided to the users of the guide. If you want to write a dissertation in XML using WordPerfect you have to work with the SGML/XML functionality. Please consult the information center of the installation CD-Rom of the WordPerfect suite with the keyword SGML. A detailed installation and usage guide will be shown.

You will need the following files:

n	diml1_0.dtd	
n	diml1_0.lgc	Logic file
n	diml.lsi	Layout file
n	diml.lnm	Alias file
n	dissertation.wpt	Dissertation style sheet

WordPerfect provides the user also with the following programs:

n a XML File Wizard,

n a XML Project Designer,

n a XML DTD Compiler,

n a structured Tree View,

In order to compile a DTD into a Logic file, you have to follow the steps:

1. copy the following files into the corel\suite8\programs\mapfiles directory: diml1_0.dtd; cals_tbl.dtd;hubspec.ent

2. Edit the iso8879.map file in a text editor:

   Add the lines:

```
PUBLIC    "-//HUBspec//ENTITIES Special Symbols//EN" "hubspec.ent"
PUBLIC    "-//HUB//DTD Cals-Table-Model//EN" "cals_tbl.dtd"
```

Press Compile. Now the LGC file has been produced and is useable for writing an SGML/XML document.

You can now open WordPerfect and choose Tools / SGML / Document types. Here point to the created diml1_0.LGC file and if you want to use a standard layout to the diml.LSI file.

Your own layout can be produced using the Layout Designer program that is in the WordPerfect suite. Just choose open diml1_0.lsi or open a new file and start creating or manipulating your own layout style.

Start writing a dissertation directly with SGML/XML by choosing under WordPerfect the option Tools /

SGML / Document types.

Usually there has been a native WordPerfect template produced in order to help you with additional drop down menus etc. To use the provided template for WordPerfect 8 you have to choose File / New / Options / Add Project / Add new document / Call it "Digital Dissertation" / Search for dissertation.WPT. This allows you now to use the WordPerfect style sheet by choosing under the main window: File / New / Choose "Digital Dissertation" and Create. This enables the functions of the drop down menu as in the following figure.

3.2.5.4 Preparing for Conversion to SGML/XML in LaTeX, Susanne Dobratz

The Problem

If ETDs should be archived for he next 20-50 years and still be readable and useable, it in necessary, that equally to the approach for MS Word, we use predefined style sheets for LaTeX. Only by standardising the usage of Latex, a quick and sustainable solution for a conversion into XML can be designed. As LaTeX is mostly used within the natural sciences and mathematics, the encoding of complex mathematical symbols, formulas and expressions is one of the major problems for such a conversion. As there are XML document type definitions or schematics for mathematics, MathML (see http://www.w3.org/math) and most math software, like Maple, Mathematica, etc supports an export into MathML, this standards has to be used as an output from LaTeX as well.

The LaTeX format should enable an easier conversion to XML, because of its structured approach to text processing. But the usage habits of LaTeX users, which tend to program complex macro packages in order to style a sophisticated print layout make it much more difficult to receive homogeneously structured documents in most cases. Also does the not parseability for structural and syntactical correctness complicates a conversion.

Converting mathematical expressions into XML can be done using 3 different strategies:

§ 1. Convert them into graphics that are easily interpreted and presented by common Internet browsers. Here a search within formulas or a further usage is excluded.

§ 2. To convert them into MathML,

§ 3. To leave them in a LaTeX encoding within the XML file. Then Plugins like IBM

Techexplorer, or Math Viewer are able to interpret the LaTeX code and produce an on-the –fly rendering of formulas and mathematical expressions.

There are semantic differences in LaTeX between the encoding of formulas. So authors have to be aware of the differences of LaTeX –tags or commands that are on a semantic level and on those, which are on a layout, level.

Example:

Pi represents the mathematical constant, which is the ratio of a circle's circumference to its diameter, approximately 3.141592653.

Encoding this in MathML <pi>

```
<apply>

<approx/>

<pi/>

    <cn type = "rational">22<sep/>7</cn>

</apply>
```

This will be rendered as follows: $\pi \approx 22/7$

Instead of coding it simply as letter pi, which may stay as a name for a variable:

```
<apply>

   <approx/>
```

```
    <pi/>

    <cn type = "rational">22<sep/>7</cn>

</apply>
```

This would be rendered as: **pi ≈ 22 / 7**

Software and Tools:

In order to produce an XML document out of a LaTeX document there are several possibilities:

n TeX4ht is a highly configurable TeX-based authoring system for producing hypertext. It interacts with TeX-based applications through style files and postprocessors, leaving the processing of the source files to the native TeX compiler. Consequently, TeX4ht can handle the features of TeX-based systems in general, and of the LaTeX and AMS style files in particular. (http://www.cis.ohio-state.edu/~gurari/TeX4ht/mn.html)

n WebEQ : a Java-based collection of tools for authoring and rendering MathML, including a visual editor, a WebTeX to MathML translator, and a rendering applet for interactive mathematics on Web pages. WebEQ also provides Java Programmers with API documentation and libraries for other MathML aware applications. (http://www.dessci.com/de/features/win/default.stm#TeX or http://www.dessci.com/features/win/default.stm#TeX)

For further information on the usage of different tools, please use:

Michael Goosens; Sebastian Rahtz: The LaTeX Web Companion, Addison-Wesley, 1999: ISBN: 0-201-43311-7

3.2.5.5 Checking and Correcting, Susanne Dobratz

A. Important Parts for a checking procedure

After an author has written his or her ETD, the service institutions, like the university library or computing and media center, have to check whether the ETD is complete, readable and correct. Therefore, a checklist is useful. The checklist should consider the following parts of a document and should also be open to the authors for self-checking:

1. Organizational questions
2. The WinWord, WordPerfect or LaTeX document
3. The PDF-Version of a digital dissertation
4. Metadata

Checking organisational questions

After receiving an ETD from an author, several organisational issues have to be proofed:

n Does the student belong to the university?

n Has he passed his exams?

n Has he passed the approvals?

- Has he missed deadlines or not?

Checking the WinWord / WordPerfect document

For the application of style sheets the following parts of a word document are mostly critical for further usage:

- Can the document as a whole been opened within the WinWord or WordPerfect systems at the service institutions?

- Has the student used all style sheets correctly?

- Has the students used the heading feature of WinWord or WordPerfect?

- Is the title page fully styled and all information filled in?

- Do figures and tables have own captions, and has the insert caption feature of the text formatting system been used?

- Are lists styled as lists?

- Are Tables produced using the table features from the text formatting system? (Sometimes authors use the tabulator to build tables)

- Has the authors used a reference managing system for the references? This makes the automatic formatting of the bibliography much easier and enables a linking into the text parts.

- Did the author use the automatic spell-checking features instead of applying hard coded – letters in the text?

Checking the LaTeX document

- Have the guidelines and rules been followed by the author?

- Has the template or styles file been used?

- Has the author used bibtex to collect the references?

- Have all figures been provided in an EPS (encapsulated postscript) format?

- Have all additional styles been provided by the author?

n Is the whole ETD processible at a computer of the library or the computing centres?

Checking the PDF document

n Is the PDF document readable? Can it be opened within the actual version of the Adobe Acrobat Reader?

n Does the PDF file contain all text pars?

n Does it contain hyperlinks for multimedia additions? Do the hyperlinks work within an actual Internet browser?

Checking the metadata

n Has the author provided all Dublin Core metadata requested about himself and his thesis?

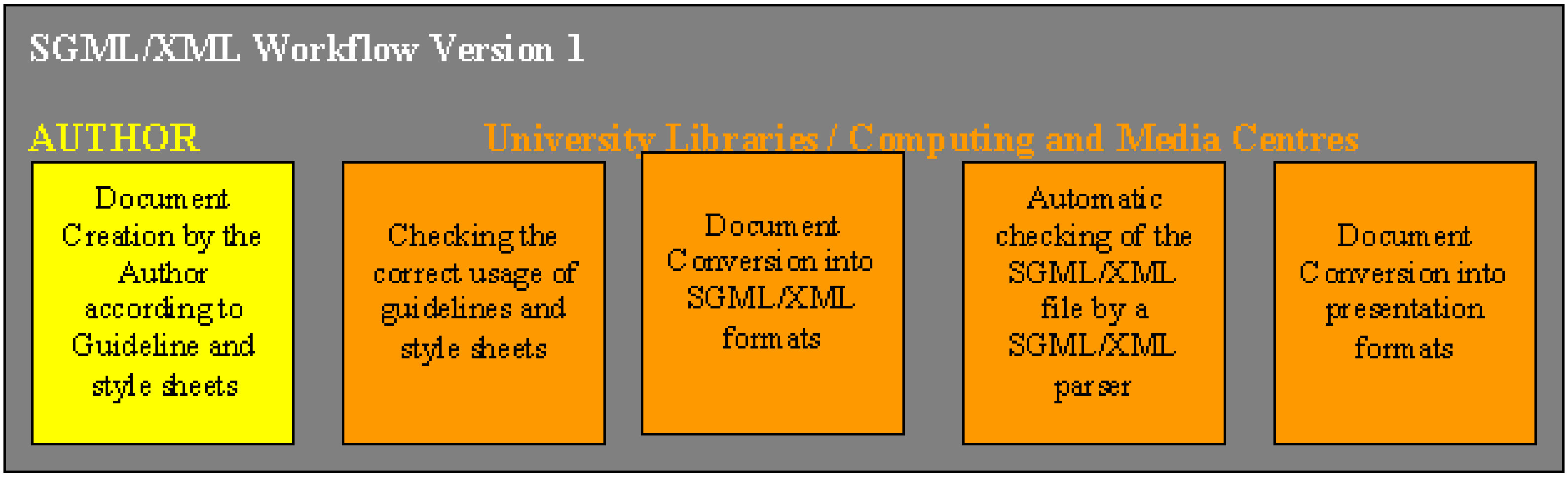
n Are there keywords in different languages, according to different classification schemas?

n Are there abstracts in different languages?

n In which format has the metadata information been provided?

B. Checking for an SGML/XML-based publication and archiving workflow

An SGML/XML-based archiving strategy today consist on a conversion workflow, as shown:

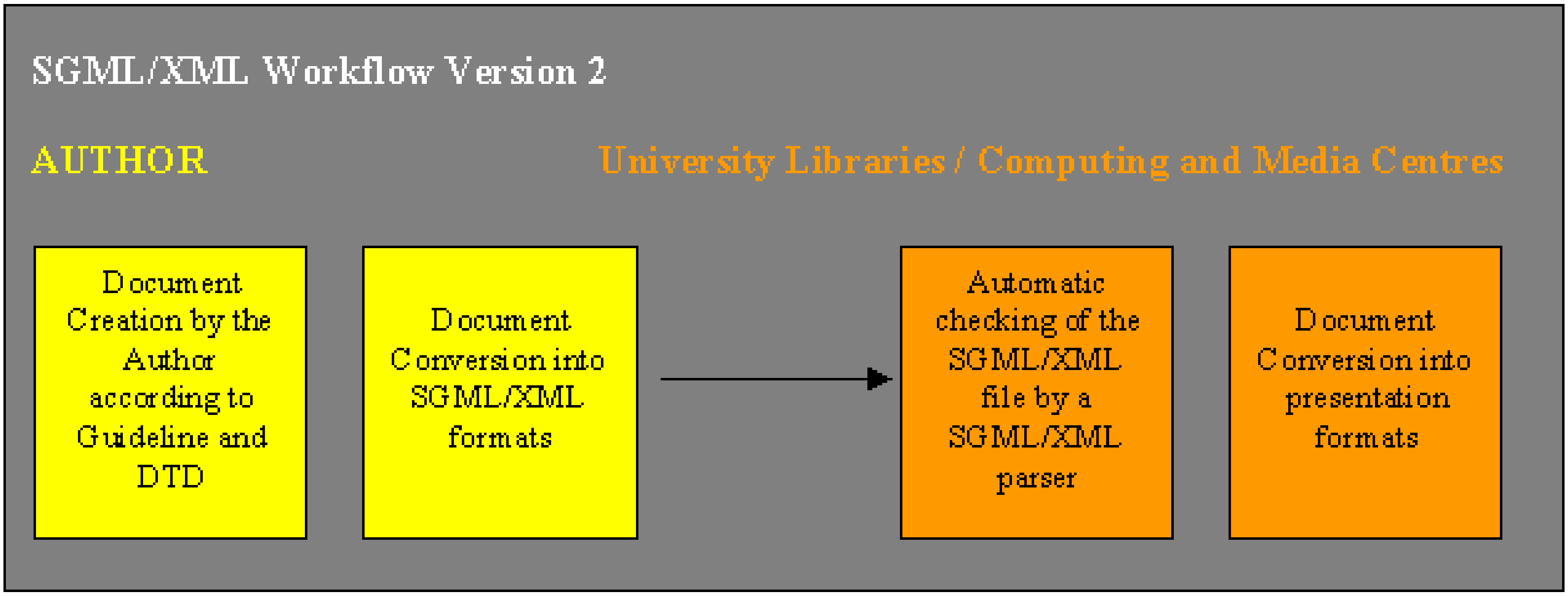

SGML/XML Workflow Version 1
AUTHOR
University Libraries / Computing and Media Centres
Document Creation by the Author according to Guideline and style sheets
Checking the correct usage of guidelines and style sheets
Document Conversion into SGML/XML formats
Automatic checking of the SGML/XML file by a SGML/XML parser
Document Conversion into presentation formats

Workflow Version 1: The Conversion from native text formatting formats (like doc WinWord) into SGML/XML-compatible documents is done by a service of the university. Here checking and correcting basically consists of the checking of the adequate usage of the style sheets and the guidelines provided by the university. This procedure is in practice at Humboldt-University Berlin, at Université Lyon 2, Université de Montréal, …

This checking has to be done within the word processing systems used, e.g. WinWord and can e.g. partially automated by Macros defined within the word processing system's macro language, e.g. Visual Basic. Then, the checking person of the staff runs this checking macros and can find out, whether special styles have been used or not. This checking procedure has to be added by a manual checking of the correct usage of the styles applied to the document by the author. For this reason it is very helpful to have a checklist for each document. This will ensure the level of strictness for the style control equivalent for every document.

Workflow Version 2: The conversion from native text formats (like doc WinWord) into SGML/XML-compatible documents is done by the author himself or the author writes directly in SGML/XML. Here checking and correcting basically consists of the checking of the adequate usage of SGML/XML by an SGML/XML parser and concentrates especially on the correct usage of the document type definition (DTD) provided by the university and the guidelines.

This procedure is e.g. in practice at the University of Iowa, Iowa City.


SGML/XML Workflow Version 2
AUTHOR
University Libraries / Computing and Media Centres
Document Creation by the Author according to Guideline and DTD
Document Conversion into SGML/XML formats
Automatic checking of the SGML/XML file by a SGML/XML parser
Document Conversion into presentation formats

This can completely automated by an SGML/XML-parser and checking scripts that produce an error list for the checking staff and can save a lot of time in comparison to the previously described workflow. The disadvantage of this model from today's perspective is that a very comprehensive author's support has to be designed and carried out, in order to enable authors either to perform an initial conversion from any text formatting system into an SGML/XML compliant document or to use an SGML/XML editor in a way that allows the author to understand and interpret messages.

3.2.6. Integrating multimedia elements, Simon Pockley

Inaccessible *dead media* has little value. Students who are incorporating moving images, audio and live data streams into their ETDs should not underestimate the work involved in managing these resources. How these resources are created, and the form and format they are created in, will determine how your ETD can be managed, used, preserved, and even re-used in the future.

When your ETD enters a networked electronic environment, it does not exist in isolation. It becomes part of a boundless resource space in which descriptions of the work and its components (metadata) need to be in an internationally recognized form if the work is to be accessible.

Inter-operable standards that will allow for this metadata to be understood by everyone (even machines) are now available. However, the application of these standards requires a new form of collaborative relationship between you, the creator, and indexing initiatives such as the NDLTD.

This means that you, the creator, have to take responsibility for describing each layer or 'object' of content as an independent entity capable of being accessed and manipulated in its own right. From its conception, a compound digital resource should be seen as a composition of objects in an encoding architecture that expresses the spatial and temporal relationships between these objects.

- These objects may be audio (mono or stereo) and visual (2D or 3D) or text.
- They may also be composed from several sources.
- They may be simultaneously acquired, processed, transmitted and used in real time.

If, in the future, the encoding of these objects is to be decoded, the metadata must include format information. In the interests of interoperability, it is good practice to select formats from the list of **Internet Media Types** (MIME values) whenever possible (this list is a registry where there is a procedure for adding new types, if necessary).

Available [on-line]

http://www.isi.edu/in-notes/iana/assignments/media-types/media-types/

An important (proposed) standard that multimedia content providers should investigate is the **Synchronized Multimedia Integration Language (SMIL)**. This standard will allow hypertext creators to define and synchronize multimedia elements (video, sound, still images) for web presentation and interaction.

Available [on-line]

http://www.w3.org/AudioVideo/

3.2.7. Providing metadata – inside, outside documents, Simon Pockley

There is nothing new about the concept of metadata. Metadata is resource description; the kind of information found in a library catalogue. What is new in the digital world is the essential role that you, the creator, now play in providing this information. Good quality metadata is easy to provide at the point of creation but usually difficult, expensive or impossible to discover retrospectively.

At one level, this is because all digital resources are in some way dependent on electronic mediation by computers and software and it is only at the point of creation that a record of these dependencies and descriptions can be recorded. At another level, it is the sheer volume of creation that alters the role of the librarian or custodian from cataloguer to metadata repository manager.

In an ideal world, all digital material would be created independent of proprietary hardware and software. In other words, everything would run on commonly available hardware using freely available (public domain) software such as a web browser.

In the real world, many content creators will be producing work on-line or off-line that is either hardware or software dependent (or both). Unfortunately, the costs of emulation, migration and licensing increase if resources are generated in proprietary or platform dependent formats. If possible, try to use commonly available open source formats.

Metadata is information about these applications and formats, which allows for licensed versions to be archived so that the material can be displayed or accessed. In order to be able to provide long-term access to a digital resource, the NDLTD **needs** the following metadata:

- Information about the content creator (rights, contributors, publisher,);

- Information about the content that will help it to be found or discovered (coverage, description, title, subject, relationships);

- Information about the resource (formats, system requirements, date, identification).

Storing metadata

Metadata can be stored in:

1. The object or document being described.

There are a growing number of audiovisual formats that allow for metadata to be embedded in the file itself. For example, a text format like HTML allows you to embed metadata in the header of the file and recent versions of image formats such as MPEG include space for metadata. This has the advantage that the information is self-contained and is truly transportable across systems. The major disadvantage is that systems accessing the object will have trouble catering for multiple views or meanings.

2. A separate file that can be externally accessed but is linked to the object or document.

This has the advantage that different communities can gather the metadata for different purposes. It has the disadvantage of being open to misinterpretation through syntax error or unrecognised schema.

3. A separate file stored in a database.

The NDLTD model encourages students to submit their metadata to a central repository for indexing in a database. The database will then point to the object/document. This also allows for multiple instances of the metadata for one document. It also provides for enhanced administrative tools (as are normally provided by database systems). Advanced database systems could provide a very sophisticated management system. This is the most expensive method to implement but it has the advantage of being significantly more flexible and provides administrative support from the outset.

See: **European Projects such as Metadata Observatory.** The aim of the Observatory is to maintain and promote a knowledge base for metadata for multimedia information to continually assess relationships between Dublin Core and other initiatives, especially undertaken in Europe, in order to assist evolution of standardised metadata schemes.

Available [on-line]

http://www.cenorm.be/isss/Workshop/metadata-observatory/Home%20Page.htm

See: **The Dublin Core Metadata Initiative.** This is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

Available [on-line] http://dublincore.org/

3.2.8 Protecting intellectual property / how to deal with plagiarism, Edward Fox

Protecting against plagiarism may involve specific steps as explained below, in addition to action in accord with the discussion above in section 2.3.6.

First, there is software available to detect plagiarism, or, in simple terms, similar copies. Stanford's SCAM software, and other software developed in Australia and other locations, discussed in the digital library literature and elsewhere, is available. This software can compare two documents, finding sections that are identical or similar (e.g., up to within simple substitution changes). If such software were to be widely run, it could ascertain for each new work if there were an ETD previously submitted that is very "close". While determined authors might defeat such software, more refined software could be developed to highlight cases where even such protection efforts were involved.

Second, it should be noted, as stated in section 2.3.6, that if a work in made widely available, it is more likely that someone, seeing a new ETD that involves plagiarism, would detect that situation. In effect, as more and more ETDs are made accessible, there is a larger community monitoring abuses.

Third, since it is known that ETDs are often read by many more people than would read a paper dissertation, there is a strong psychological pressure to discourage plagiarism. One aspect of this is that students are aware of dire penalties if plagiarism is detected - they may be removed from their degree program and forever disgraced. Another aspect is that faculty working with students on ETDs, being aware that many might read the work, are likely to be more diligent than with paper works with regard to checking the validity and quality of results reported. In other words, it becomes more likely that faculty will carefully study ETDs that have their name on them as advisors or examiners. In particular, it is exceedingly unlikely that a student would be able to write about work they had not done, without that being detected.

In summary, detection by machines or people, and the threat of severe penalties, are likely to discourage students from even considering plagiarism with regard to ETDs.

3.3 Naming Standards: File Names and Unique Ids, Australian Digital Theses Program

The ADT Program uses two simple protocols to describe – the PDF files associated with a thesis, and the generation of a unique ID for each thesis.

ADT Program filename standard

To facilitate access to ADT theses it is important that the following filename standards are followed for naming the PDF files prior to uploading them through the Deposit form:

Please note - ADT filenames should only contain:

n Letters from the alphabet (upper or lower case). Example: **05appendix.pdf; 05APPENDIX.pdf; 05Appendix.pdf**

n Or numbers **(0-9)**

n Or any appropriate alphanumeric combination. **02Chapter1.pdf; 09Chapter10.pdf**

n The only symbols permitted are - (hyphen) or _ (underscore). Example: **06Appendix-Images.pdf;**

06Appendix_Images.pdf

§ Create one PDF file containing title/author information; abstract; acknowledgments; table of contents; introduction; preface and any other introductory text that is not part of the main body of the thesis. Call this file **- 01front.pdf**

§ Create another file containing the whole of the thesis (including what is contained in the front.pdf file) and call this **02whole.pdf**
…alternatively…

§ Break the thesis up into smaller files maintaining a relevant filename structure such as **02chapter1.pdf**; **03chapter2.pdf**; **04appendix1.pdf**; **05bibliography.pdf**; etc.

§ In order to maintain the desired file sequence the files need to be numbered using the standard 01-99 (or 001-999 if individual files exceed more than 99 for an individual thesis). Numbering theses is required no matter which filename standard is adopted.

Example 1	Example 2
01front.pdf **02whole.pdf**	**01front.pdf** **02chapter1.pdf** **03chapter2.pdf** **04appendix.pdf** **05bibliography.pdf**

The 01front.pdf file is compulsory. The official ADT v1.0 or v1.1 software will create an empty front file automatically if a real one is not uploaded during the deposit/submission process. This is to serve as a memory prompt for administrators. While it is difficult to recommend exact file sizes the project team suggests that the whole.pdf file should be within the 2-4MB range. For larger, and hence multiple file theses (necessary for scanned ones) it is suggested that the file size should be in the 4-10MB range.

This standard will allow viewers to easily identify the relevant parts of the ADT theses via a common filename structure. It will also allow viewers to quickly look at the table of contents and other introductory information without having to wait for the whole thesis to download, or to better determine if they want to see the whole thesis in the first place. The standard is also an effective and simple way to present an introductory view of a thesis at the outset. This is similar in a way to the 20-25-page view that

UMI/Bell&Howell have introduced for the DAI database. Similarly, this ADT standard will facilitate e-commerce transactions if participants choose this option. This would work just as the UMI/DAI model - the metadata/abstract and front.pdf file would be available freely, with charges kicking in to access the rest of the theses' pdf files.

ADT Program unique Ids

The combination used by all ADT members is as follows: the word "**adt-**" immediately followed by the unique national **institution code** as per the *Australian Interlibrary Resource Sharing Directory*, immediately followed by the **year**, **month**, **day**, **hour**, **minute** & **second/s** the thesis was deposited to the local server. As the ADT Program is a distributed national model, it is critical that the URLs are unique within that context. The combination above ensures this.

Examples of the other institutions would be:
../adt-ANU20010223.162256/ (Australian National University)
../adt-WCU19991112.125812/ (Curtin University of Technology)

../ adt-NUN20010510.153038/ (The University of New South Wales)

….and so on.

3.4 How to submit your ETD, Gail McMillan

Universities requiring or accepting ETDs should provide students with basic levels of support such as hardware (e.g., computer workstations), software (including programs for creating the works such as word processors, as well as other software such as sound and image programs), and opportunities for training in how best to use the hardware and software and how to submit their ETDs.

3.4.1. Local support, Australian Digital Theses Program

The ADT Program is an Australian university libraries initiative. The library at the ADT member institution provides local support. Support to academics and students is varied and depends on the levels of expertise of the individuals involved in the submission process. The levels of support offered include:

n provision of workstation/s, including software, in the library dedicated to the submission process plus library technical support as appropriate to aid students to submit and convert files into PDF

n provision of basic instructions and information for students wanting to submit theses independently. See example of this at The University of New South Wales: http://www.library.unsw.edu.au/theses.html#dep

n library technical expertise to entirely do the submission & conversion process on behalf of students

n provision of information via the library's website, pamphlets, presentations, formal letters & information for faculties, schools, graduate committees, academic staff and students about the local ADT Programs. The sharing of this information by the ADT member institutions as well

n broader support at the institutional level - sharing expertise, knowledge and information as appropriate

All library support is seen as part of the normal university scholarly process and is provided free of any charge.

3.4.2. Typical workflow, local policies and procedures, Gail McMillan

Apart from the ADT Program support processes described above, local workflow and policies within the ADT membership will vary. Typically, what is most common across all members is that the libraries are driving the ADT Program through their respective institutions, providing all levels of support and provision of information, and that only research theses are acceptable - i.e. PhD or Masters by research. What is less common is the local procedures insofar as how theses are deposited to the local server, who has the authority to approve/make public and administer the program and how the ADT theses are integrated within the local IT infrastructure. Access to theses also varies between institutions. While most encourage and actively support free and unrestricted access, some take a much more conservative approach and restrict all theses to the local domain. Obviously, all institutions will apply some temporary restrictions to access if required, for reasons of patents pending, copyright and upcoming publications.

n An example of workflow, policies & procedures at The University of New South Wales [UNSW] Library is as follows:

n UNSW Library provide support as required, see 3.4.1. above.

n When theses are deposited, auto alerts are sent to the student, supervisor, the ADT coordinator and the cataloguers

n The cataloguers are responsible for checking the theses are original and have been awarded. They edit the metadata to include appropriate thesaurus terms/subject headings. The cataloguers are also responsible for approving the theses and making them public. They also catalogue the theses into the local OPAC and the National Bibliographic Database. These ETDs are then available via the national distributed ADT database, the local view of the ADT database [i.e. UNSW-ETDs only], the local OPAC and the National Bibliographic database. Most are available unrestricted with a few restricted to the local domain due to copyright reasons

n Two separate units are responsible for the ADT Program at UNSW Library. The General Services Department's Learning Support Unit provides all support, promotion and associated information. The cataloguing department is responsible for administration of the ETDs when deposited to the server. The ADT coordinator at UNSW is also the overall coordinator and is responsible for liaising and guiding the program at the national level.

3.5 Becoming a researcher in the electronic age: responsibilities, opportunities, issues of access, IPR, benefits to others and to oneself, Edward Fox

In 2001, the National Research Council of the USA produced a booklet identifying many of the responsibilities, opportunities, and other issues faced by young faculty as well as graduate students completing their theses or dissertations. Similar groups have prepared other works. In particular, it is clear that there have been dramatic changes in scholarship that have resulted from the availability of computer tools, the shift to digital libraries, and the tremendous increase in resources available to young researchers (with respect to computing, networking, and content). Such researchers should learn about electronic publishing, should use digital libraries, should be aware of intellectual property rights, and should leverage new opportunities made available through the enhancement of related technologies and infrastructure. These all encourage involvement in ETD activities, and lifelong participation in the world of scholarly communication that in many cases was first made visible to them in connection with their ETDs.

4. Technical issues, Ed Fox

To efficiently and effectively implement an ETD program, involved institutions need to develop a suitable technical infrastructure – a side benefit and related goal of the initiative. This section outlines the key aspects of the technical portion of an ETD effort, covering production, dissemination, and access.

4.1.1. Contexts: local, regional, national and global, **Ana Pavani**

World Diversity

Our world is a very diverse place. We have many geographies, climates, races, cultures, and languages. Another characteristic that differentiates places of the world is the level of development. This creates various levels of access to:

n Food and water

n Housing, electricity and sanitation

n Medical care

n Education

n Jobs

n Information and knowledge

World Infrastructure

Infrastructure is different in different regions of the world too.

Third world nations have characteristic social/economic gaps among groups of persons and/or regions within the countries. In one nation, developed areas, where people have access to all the items listed above and where infrastruture is good, coexist with very poor regions where living conditions are bad.

This coexistence leads to a varied range of infrastructure levels in general and in universities too. There are universities with very good campuses and with the infrastructure comparable to those in developed nations. Others have bad installation, lack equipment and do not have good networks and Internet connections. This bad fortune reflects on students, faculty and staff who may not be proficient with information technology tools.

When an ETD program is considered, some aspects of the infrastructure must be examined. They can be grouped in 3 categories, as follows.

Local (in the university) infrastructure

n The level of automation of the library in terms of cataloging of the collection, library system, equipment for the staff and for end users, etc

n The number and quality of machines available to students

n The number and quality of machines available to the administrative staff

n The network conditions - connection to all university buildings, speed, reliability and support

n The level of computer literacy of students, faculty, library staff and administrative staff

n The number of machines connected to the Internet

n The connection of the university network to WAN's and the Internet - speed, reliability and support

Regional and national infrastructure

n The network connections - speed, reliability and support;

n The existence of other ETD and/or digital library projects to develop the culture and to seek/provide support

n The possibility of funding, from agencies and/or private companies, to ETD programs

Global

n The network connections - speed, reliability and support;

n The dissemination of information on ETD and/or digital library projects to provide support to those who want to start programs

n The agreement on minimum standards for systems, technology and metadata to allow interoperability and seamless access n The discussion and agreement on languages to identify ETD's so that they can be searched, retrieved and used

4.1.2 Networking: hardware, software, Edward Fox

Networking and ETDs

With the spread of the Internet and the WWW, and the emergence of local area networks as well as wide area networks (LANs and WANs), network facilities have enormously expanded in many of the educational institutions of the world. To support the needs for access, and the related processes that deal with making content accessible, many universities have made networked computers available to their graduate students, as well as to faculty, staff, and undergraduate students. The ETD initiative can build on those investments, employing them to support submission and downloading of ETDs. By focusing on networked access, there can be considerable savings that result from elimination of manual handling and physical distribution. It is recommended that, except for ETDs that have enormous amounts (e.g., gigabytes) of multimedia content, authors, as well as university staff, avoid procedures that require transfer of content using diskettes, CDs, or other physical media.

Network Traffic and Hardware

Universities should consider the amount of traffic on their networks, making sure that networking hardware (Internet connection, routers, and cable plants) accommodates well the demand for uploading and downloading ETDs). This accommodation generally will not be a problem if adequate support for email, rapid access to WWW, and other types of usage is provided.

Software

Regarding software for accessing ETDs, there is not much of a special nature that is required. Web browsers, support for Java applets, and multimedia presentation tools are typically sufficient. Special aids may be needed for PDF, that Adobe provides for free at Adobe at www.adobe.com, or SGML/XML (becoming more widely available). The real need for special software is to help manage the ETD submission process, handle local workflow, make ETDs accessible, and facilitate search. NDLTD and its various parts have prepared software to help with all of these.

4.1.3 Seamless access: Open Archives Initiative, federated search,

Edward Fox

Access to ETDs that are produced by students around the globe requires some mechanism for connecting with the many computers that house those ETDs. There are two basic approaches.

Federated Research

In federated search, a user's information need, expressed in the form of a query, is sent by the federated search system to all the sites that support the searching over local ETD collections. Then, when the sites have completed their searching and generated results, either the user can view each site that might have some relevant content (see Powell & Fox, 1998), or some type of fusion of results leads to a single merged list (as with Dienst, see Lagoze & Davis, 1995). While federated search yields up-to-the-moment results, such currency is usually not of high priority in the ETD world (where daily updates should suffice). At the same time, federated search may involve complex timeout and backup site management, if some remote sites are down or slow to respond. At best, federated search is often slow (due to network delays) and suffers from having to manage a wide diversity of representations of data at remote sites, leading in some cases to low data quality. Nevertheless, see such a service for ETDs from www.theses.org.

Harvesting

Harvesting is the second basic approach. As is explained in section 4.3.4.1.6, the Harvest system first clearly demonstrated this solution, and is still in use in Germany and other locations. However, this is being superceded by the Open Archives Initiative (www.openarchives.org, see Lagoze & Van de Sompel 2001). NDLTD has developed a harvesting-based OAI access scheme for handling the global collection of ETDs; see Suleman et al., 2001 (parts 1 and 2). The basic outline of the approach is as follows:

- Each ETD is described (with metadata) using MARC21 or ETD-MS (http://www.ndltd.org/standards/metadata/ETD-ms-v1.00.html).
- Each ETD site runs an open archive, which responds to OAI requests for metadata by providing

Dublin Core records, as well as either (or both) MARC21 and ETD-MS. For example, the software for ETD management developed by Virginia Tech has such a capability (http://www.dlib.vt.edu/projects/OAI/software/ndltd/ndltd.html).

n State, provincial, national, regional or other organizations may harvest from these sites, and run their own open archives and related services.

n Virginia Tech harvests from all sites (or group sites that have harvested for university sites) to develop a union collection (http://oai.dlib.vt.edu/~etdunion).

n VTLS Inc., as a service to NDLTD, provides search access to the union collection (http://www.vtls.com/ndltd).

References

Lagoze, C. and J. R. Davis. 1995. “Dienst - An Architecture for Distributed Document Libraries”, in Communications of the ACM, Vol. 38, No. 4, p. 47, ACM, 1995.

Lagoze, Carl and Herbert Van de Sompel. 2001. The Open Archives Initiative Protocol for Metadata Harvesting, Open Archives Initiative, January 2001. Available http://www.openarchives.org/OAI/openarchivesprotocol.htm

J. Powell and E. Fox. Multilingual Federated Searching Across Heterogeneous Collections, D-Lib Magazine, Sep. 1998 http://www.dlib.org/dlib/september98/powell/09powell.html

Hussein Suleman, Anthony Atkins, Marcos A. Gonçalves, Robert K. France, and Edward A. Fox, Virginia Tech; Vinod Chachra and Murray Crowder, VTLS, Inc.; and Jeff Young, OCLC. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html

Hussein Suleman, Anthony Atkins, Marcos A. Gonçalves, Robert K. France, and Edward A. Fox, Virginia Tech; Vinod Chachra and Murray Crowder, VTLS, Inc.; and Jeff Young, OCLC. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-

pt2.html

4.2 Production of ETDs, Edward Fox

Electronic publishing technology and infrastructure is needed to support the production and archiving of ETDs. Section 4.2.1 provides an overview of this matter. The next two sections explain key issues related to the two main approaches: PDF and SGML/XML. Remaining sections give further information of specific matters such as metadata and post processing.

4.2.1 Overview: hardware, software, multimedia, scripts, encoding, document representations/conversions, Edward Fox

Preparing an ETD is somewhat like preparing a book to be given to a publisher, and then distributed electronically (and perhaps on paper). There are many aspects to this process, briefly summarized in the next paragraphs.

Hardware Perspective

First, there is the hardware perspective.

Computers

Authors of ETDs nowadays almost always use a computer for this activity. In many cases this means a personal computer, though in some cases a terminal, PDA, or another device might be chosen. With the continuing increase in performance and functionality, given a particular price, it is becoming more feasible for authors to have their own computers. However, in some cases, an office or laboratory or shared computer might be used, at least from time to time.

Other Devices

In addition to a computer, authors may use other special devices to prepare parts of the ETD. In particular, in the case of multimedia content, parts might result from using a scanner, digital camera, digital camcorder, slide scanner, microphone, sound card, MIDI device, or other special equipment. Special systems might be used for audio or image or video editing, though in some cases such editing can be done on a PC.

Software

Second, consider software.

Software Editors

Software that can help with ETD production may be specialized by role. Text editors like Wordpad, vi, and emacs allow creation of files of characters, often encoded in ASCII. More powerful word processors may allow handling of more extensive sets of characters, like UNICODE, and entry of character codes from large sets (e.g., for Chinese, Japanese, or Korean texts).

Separate Editors

Separate editors may support multimedia content. Photoshop handles photos and other images, for example, while Sound Forge allows manipulation of audio files. Premiere is a tool for video editing that also can process audio, animations, and other related components. Editing tools may handle conversions as well, or accept the results of conversion. Ultimately, one may think of authoring, capture, conversion, and editing tools for all media types as all having the objective of amassing a pool of components that go into the ETD.

Integration

Large documents thus are made of pools of content objects, and students can use special software to integrate the content into a coherent whole. A simple integration is to have a linear structure, like a book, where all the components are ordered, as in a sequence of pages, manipulated by a word processor. It is simple for images to be included in such a work, as long as sizing is adjusted. However, large images, and other multimedia content (e.g., audio), must be integrated in a different fashion.

Such integration may involve linking (as in hypermedia, found in many Web sites). Additional interactivity may require a tailored hypertext system (e.g., Toolbook, Guide), or a multimedia integration package (e.g., Director or AuthorWare). Such a system will support synchronization (e.g., using the WWW standard, SMIL), or complex performances. Ultimately, the most sophisticated integration of content with reader interaction requires a programming language. In the case of multimedia languages, this typically is called a scripting language. However, students can also use general purpose languages like Java for sophisticated handling of content along with interaction.

Representation

Third, consider representation. Each content object must have a representation scheme, depending on its form. Text components are characterized by character encoding, as well as supplemental information to support presentation (e.g., font, size, style). Multimedia components are encoded in suitable forms, depending on resolution or other measure of level of detail, compression, and other attributes. For the purpose of archival preservation, the representations used for content objects (e.g., UNICODE, MPEG), as well as those used to describe the organization and rendering of content (e.g., SMIL), will be in standard forms.

Archival Forms and Related Software Support

Fourth, for an ETD or other large work, one can choose from archival forms, and related software support.. In particular, for more detail about software, see the next subsection. For details about PDF, SGML/XML, and other matters, see the next sections.

4.2.2 Page Description Languages (PostScript and PDF)

Melonie Warfel, Adobe Systems, Inc.

Evolution of Networks and the World Wide Web

The evolution of networks and the World Wide Web have profoundly affected the content and quality of communications. The complexity and visual content of documents is no longer limited by the artistic skills of the designer. Itranets and Web sites are being tapped into for creative content that is much more complex than could have been imagined. As creative expression expands to new levels, the quality of the printed output must also rise to the challenge.

The Internet and the Printing Workflow Process

The Internet has changed not only the dynamics of the creative process but the entire printing workflow process. The Internet is used to transmit information electronically while CD-ROMs and servers are used for storing and accessing data. One often finds difficulty in predicting what will be printed because documents can come from many sources. And it’s equally difficult to predict where they will be printed, because documents can be distributed around the world electronically and printed locally. The printing workflow has evolved from creating, printing, copying, and distributing hard-copy documents to creating, electronically publishing, and printing documents on demand. It was clearly time to advance the imaging standard.

Adobe Postscript

With the introduction of PostScript in 1985, Adobe Systems Incorporated sparked a revolution in how we communicate on the printed page. Since its introduction, Adobe PostScript has become the printing and imaging technology of choice for corporations, publishers, and government agencies throughout the world. In fact, 75 percent of commercial publications are printed on Adobe PostScript devices, including black-and white printers, color printers, imagesetters, platesetters, and direct digital printing systems. Adobe PostScript is also the display imaging system in some of the most advanced workstations on the market today.

Adobe PostScript 3 takes the PostScript standard beyond a page description language into a fully optimized printing system that addresses the broad range of new requirements in today's increasingly complex and distributed printing environments.

Adobe Postscript and PDF

By fully integrating Portable Document Format (PDF)—an open file format that preserves the visual fidelity of documents across applications and platforms—into Adobe PostScript 3, documents can be delivered electronically and printed directly. A PDF file delivers the single "digital master" for use in electronic, printed, and mixed workflows, ensuring the highest fidelity across all media.

Adobe Postscript Advanced Features

With Adobe PostScript 3, Adobe delivers advanced features for the new digital document. Today, document creators draw on a variety of sources and increasingly rely on color to convey their messages. And tomorrow, these documents will be delivered and printed on a virtually unlimited range of devices.

4.2.3 Markup languages: SGML, XML,

Edward Fox

Supporting Works on Computers and Paper

Since the mid-1980s, the electronic publishing community has faced the issue of supporting works targeted for use on computers, or on computers and paper (i.e., for dual publishing). To achieve maximal flexibility, it is desirable to separate the description of document structure from the rendering of that structure into some paper or screen form. SGML and, later, XML, were devised with this goal in mind.

SGML and XML

SGML and XML are markup languages. In particular, they are meta-languages. One can provide a document type definition (DTD) that conforms to the rules of SGML, and then be able to create documents of that type. The same applies to XML. However, with XML, it is possible instead to use a schema to specify the type of document considered, or to just have markup without a DTD, demonstrating that this indeed is an eXtensible Markup Language.

Other Markup Languages

In the SGML/XML family are other markup schemes, or applications of these. HyTime extends SGML with architectural forms to allow handling of rich multimedia and hypermedia content. RDL, developed by e-Numerate Solutions, Inc., allows handling of numeric content in an application of XML, while XBRL applies XML to business reporting. For many ETDs, tailored markup schemes (e.g., MathML for mathematics, or chemical markup language), also are used for particular types of content expression.

The following subsections explain further how ETDs can be produced using markup languages.

4.2.3.1 Software

Guylaine Beaudry

Susanne Dobratz

Viviane Bouletreau

Since XML is software independent, many types of software can be used for the production of ETDs in XML. The next table gives the software you can use by type of tools. This is a short list of tools we used in our projects. Nevertheless, many others can suit you requirements, among them, freeware. http://www.garshol.priv.no/download/xmltools/ is a good place to find free XML tools and software. The XML Cover Page (http://www.oasis-open.org/cover/sgml-xml.html) is the reference.

XML Editor		
XML Spy	http://www.xmlspy.com	This tool integrates more than an XML editor.
Xmetal	http://www.xmetal.com/top_frame.sq	
Corel WordPerfect Suite	www.corel.com	See section 3.2.2.2 Can also be used as regular wordprocessing software.
Word processing software		
MS Word	www.microsoft.com	See section 3.2.2.1

StarOffice	Version 6.0 Beta http://www.sun.com/software/star/staroffice/6.0beta/	A good alternative to MS Word, exports some kinds of XML. Supports import and export of MathML files.
OpenOffice.org	http://www.openoffice.org	Exports XML but not to (user) variable DTDs
AbiWord	http://www.abisource.com	open source
Kword	http://www.kde.org	(KOffice, KDE Project/ open source)
Amaya	http://www.w3.org/amaya	Open Source, World wide web Consortium For Mathematical Formulas in MathML
Hancom	HancomWord 5.2 (included in HancomOffice 1.5 and more) http://www.hancom.com	Linux based Office Package fully compatible with MS-word file and templates.
Conversion		
Any scripting language	PERL, JAVA, etc.	

Omnimark	http://www.omnimark.com	Very interesting, but no more free, and no more significant discount for universities (about 10.000 US$ for the standard release)
Avenue.Quark	http://www.quark.com/products/avenue/	Not really interesting
Save as XML plug-in in PDF	http://www.adobe.com/support/downloads/89a2.htm	
Tetrasix	http://www.tetrasix.com	Only partially useable Good for the Majix prepared DTDs
WorX SE	http://www.hvltd.com/	Plugin für Word
Schema MarkupKit	http://www.schema.de	
I4I	http://www.i4i.com	Add-On for Word
XML Formatter for Printsolutions		
3B2	**http://www.3b2.com**	Typesetting from SGML
PowerPublisher	http://www.mai-kg.de	Mai KG, Germany
XML XSL-Formatter		
NeXt Publisher	http://www.nextsolution.co.jp/English/index.html	NeXt Solution

FOT	http://www.pro-image.de	Pro-Image.de
FOP	http://www.apache.org	Apache XML Project / Addon for Cocoon

4.2.3.2 DTDs for ETDs

Susanne Dobratz

(This section was taken from an article by P. Potter, P. Strabala, D.Dobratz, M. Schulz about ETDs, that is due to appear in “The Internet and Higher education” 4/2001)

XML Authoring Systems

The fact that currently available authoring systems for XML still have not won wide recognition has led to different strategies at different universities regarding XML documents. Most of these projects were started between 1995 and 1997, in a time when XML was alive, but where tools or standardized DTDs were barely available. A view of those projects from today’s perspective illustrates the demand for a rethinking and redesign of those approaches in order to come to a standardization.

DTDs

All the presented DTDs are built upon similar principles. A classical dissertation (which can be seen as monograph) consists of 3 main components: an extensible **titlepage** with abstracts, declarations, etc., the **dissertation corpus**, which includes text, pictures, audio, video, tables and so on, and **appendices**, which contain data sheets, bibliographies, acknowledgements and others.

The following DTDs are currently in use at different institutions:

- **ETD-ML.DTD**: Virginia Polytechnic Institute and State University (Virginia Tech)
- **DiML.DTD:** German Dissertationen Online Projectes
- **UIowa2K.DTD**: University of Iowa
- **HutPubl.DTD**: Technical University Helsinki
- **TEI-Light.DTD**: Ann Arbor und Lyon
- **ISOBook.DTD:** University of Oslo
- **TEI-based DTD with extensions for natural sciences**: Swedish University of Agricultural Sciences Uppsala

Author-DTDs

All these Document Type Definitions are so-called author-DTDs. This means that they are primarily used to support the authoring and the conversion process and do not primarily address document archiving and preservation. One may ask why all those different DTDs have prevailed. This is mainly because the scientific orientation of the mentioned universities is quite varied. Lyon, Oslo and Michigan, which use TEI-Light.dtd, mainly serve students in the arts and humanities. Problems using TEI.DTD or DocBook.DTD are recognized at universities that support a strong natural science community, such as Berlin, Helsinki or Uppsala. Often a dissertation is a cumulative work, e.g., in Lyon or Helsinki.

Université Laval, in collaboration with the Université de Montréal, is working during 2001-2002 on the modelisation of a new DTD for ETD. The DTD and its documentation will be post at www.theses.umontreal.ca.

DTDs for multimedia content

"Structured data," such as mathematical or chemical formulas, spreadsheets, address books, configuration parameters, financial transactions, technical drawings, etc., are usually published on the Web using layout programs such as Postscript or PDF, or by putting them into graphic formats like gif, jpeg, png, vrml, and so on. Programs that produce such data often also store it on disk, using either a binary or text format. Therefore, if someone wants to look at the data, he usually needs the program that produced it. With XML, data could be stored in text format, which allows the user to read the file without having the original program. XML can be thought of as a set of rules, guidelines, or conventions, for designing text formats for data in a way that produces files that are easy to generate and read (by a computer).

In addition to the older standard SGML, there are several emerging standards that use XML encoding to overcome the disadvantages common to web publishing in HTML. The following sections give an overview of standards that have been established during the last few years or which are still works in progress, but widely recognized.

XML DTDs and Schemas

For standardized knowledge management this variety of XML DTDs and Schemas seems confusing. A closer look, though, gives another perspective: every scientific subject defines and uses its own standards. The following document type definitions can roughly be classified in:

1) Schemata that use semantic tags to mark real content items, e.g., MathML or CML.

2) Schemata that are used for visualisation and layout purposes and to control the browser synchronization, e.g., HTML, SVG (Scalable Vector Graphics), SMIL(Synchronized Multimedia Integration Language).

3) Schemata that are principally designed to perform the exchange of data with huge databases, e.g., cXML(commercial XML).

Electronic Publishing

Within the field of "Electronic Publishing," these developments have led to new opportunities to structure scientific information, not just text-based but also so-called active contents and multimedia elements. This brings the whole field to a new level of information processing or knowledge management.

The different approaches for electronic publishing at universities creates a very heterogeneous environment. The following tables show how difficult it might be to subsume all those different models under one concept in order to achieve valuable and searchable information systems based on XML. Crosswalks between all those DTDs have to be defined in order to build a distributed retrieval engine, capable of searching within internal document structures "throughout the world." Not only different DTDs are used, but also different strategies to perform a conversion from usual text formatting systems into highly structured documents in SGML or XML.

Partner	**DTD**	**Conversion to SGML/XML**	**Conversion from SGML/ XML into HTML, PDF**
Humboldt-University Berlin	DiML	SGML-Author for Word	Perl-script, DSSL
Virginia Polytechnic Institute and State University	ETD	SGML-Author	Perl-script
University of Iowa	Uiowa2K	Majix	CSS
University of Montreal/ Université de Lyon 2	TEI-Light	Omnimark rtf2sgml	XSL
Technical University Helsinki	HutPubl	Frame-Maker+ SGML	DSSSL, Frame-Maker
University of Michigan	TEI-Light	Omnimark rtf2sgml	--
University of Oslo	ISO-Book	Balise	--

Table 1: Different Approaches to DTDs and Conversion worldwide

4.2.3.2.1 Berlin DTD-Workshop, Susanne Dobratz

Scope of the workshop

The Idea behind this workshop was to bring together experts and developers working on SGML or XML-based Document Type Definitions for electronic theses and dissertations. Within the NDLTD-initiative several document type definitions (DTDs) for dissertations have been developed. To come to a standardized or generalized searching and archiving structure, it is absolutely necessary to summarize all those developments and research and to work out correspondences between all of them. If SGML and XML are the preferred future document formats for theses and dissertations, the DTD will play a major role within the whole initiative. The following questions were discussed:

- How to get to a generalized DTD?
- Is there just one DTD necessary or is it possible to have more than one DTD, e.g., for different subjects and sciences?
- Where exactly are the current differences between the existing document type definitions?
- Which approach shall we follow to unify our DTDs?
- How to come to an aggreed Dublin-Core metadata element set?
- How many markup should be done by the author, how much by machine translation?
- Evaluating several tools and ways to come to an SGML/XML-based document

Outcome

This expert workshop covering the topic of using XML for publishing theses and dissertations electronically in universities was held in May 2000 at Humboldt-University Berlin, Germany. It focused on the ways in which XML can be used in university libraries in order to deliver, process and archive scholarly high quality electronic publications. The recognition that a worldwide range of various approaches for SGML/XML-based publishing concepts exists, led to the conclusion that interoperability on an international level is inevitable. Experts from the USA, Finland, Norway, Sweden, France, Portugal, the United Kingdom and Germany discussed commonalities and differences among their models, approaches and document type definitions (DTDs).

An Objective of the Workshop

The objective of the workshop was also to exchange experiences in document conversion and author's support therein, to agree on a Dublin Core metadata set worldwide and to share tools. The potential support of authors was recognized as a very essential part within the electronic publishing workflow. Attempts to convince authors to leave their native word processing systems for systems that support structured writing and/or export to SGML or XML usually tend to fail, for several reasons. First, authors are reluctant to switch from the system they are used to. This is exacerbated by the fact that publications usually have to be written within a very short time frame, leaving little or no time for authors to learn a new system, even if they were so inclined. Until XML authoring and editing tools become as simple to use as a word processor, this approach will still be viewed as an added burden to the student. Second, current SGML/XML support systems are usually far more expensive than common ones.

Prearchival Format

Many universities follow another strategy and accept documents in a pre-archival format. This necessitates conversion of the documents by the university libraries or media center. The advantage is that authors are able to prepare their documents within their native word processing systems using style sheets especially designed for that conversion, e.g., in Microsoft Word. Some institutions that have a high focus on format are reluctant to approve this method because the SGML or XML publication may be altered from the word processing document that was deposited by the student. Even though this method seems to exact a high resource cost, it is a step toward the ideal, where authors write in XML directly, using their native word processors. We are not far from this ideal.

ETD Projects with XML-based Approach

The following electronic theses and dissertation (ETD) projects have an XML- based approach already in place or are presently in a pilot phase:

n · Swedish University of Agricultural Sciences (SLU) Libraries, SWEDEN,

n · Virginia Polytechnic Institute and State University, University Libraries, USA

n · Sentiers, Université Lumiére, Lyon 2, (Cytbertheses.org project), FRANCE n · University of Montreal, CANADA

n · Universidad de Chilé (Santiago de Chilé), CHILE

n · Humboldt-University Berlin, "Dissertation Online", GERMANY

n · University of Oslo, Center for Information Technology Services, NORWAY

n · University of Iowa, Graduate College, USA

n · University of Michigan at Ann Arbor, Library, USA

n · Helsinki University of Technology, Library, FINLAND

For further information, please visit the workshop-website at: http://dochost.rz.hu-berlin.de/epdiss/dtd-workshop/index.html

With the workshops there has been a collection of tool, that are used fpor SGML/XML publishing at different universities. This collection can be found at: http://dochost.rz.hu-berlin.de/epdiss/dtd-workshop/cdrom/index.htm .

4.2.3.3 Support for students to write directly in XML

Dilshad Akhter and Edward A. Fox

This topic is discussed briefly in 3.2.1 and in more detail in 3.2.3.

The following extends that discussion by focusing on tools to help students to write directly in XML.

XML Tools

To author and view documents written in XML, different kinds of tools are required. They are:

- Editor
- Parser/Validator
- Browser

Editor

There are quite a large number of XML editors, for different platforms are available. Most of them costs less than $100. A lot of them have free evaluation version that can be downloaded. the following list gives link to some such editors.

- **XML Spy**
- **XMLWriter**
- **XED**
- **Xeena**

- **Morphon XML Editor**
- **Emile for MAC**

Parser/Validator

When a XML document is created, it is parsed/validated to see if it is syntactically correct and well formed. To accomplish this task, we need a XML parser/validator. There are a lot of XML parser available. Most of the XML editors mentioned above is also a XML parser. For example, XML Spy provides a IDE for XML, i.e. it is both editor and parser.

view a XML documents, we need a browser that supports XML. Internet explorer 5.0 supports it. There are some other browser like **Mozilla**, **InDelv**, etc. also supports XML. Netscape 5.0 is in beta. It will be able to integrate with the **DocZilla** plug in, which will support a wide range of XML functionality. **HyBrick** is an SGML/XML browser from Fujitsu. It is widely expected that these and other next generation browsers will support XML in a way that lets it fully integrated with HTML. A stylesheet is required to view the XML documents. For ETD a cascading stylesheet has been developed that. That stylesheet can be downloaded from here.

4.2.3.4 Conversions from Word to SGML/XML

Susanne Dobratz

Performing a conversion from MS Word documents into instances of a specified SGML or XML DTD is a very complex task. What you will need for that is:

§ A SGML or XML document type definition (DTD) that serves as structure model for the output. One says that the output SGML document is valid to the specifies DTD, or it is an instance of this DTD:

§ A Word stylesheet that holds paragraph and character styles according to the structures in the DTD. So if in a DTD you have defined a structure for Author:
e.g. expressed in the output file as
<author>
<title>Dr.</title><firstname>Peter</firstname><surname>Fox</surname>
</author>
You have to find expression in Word:
paragraph styles: author
character styles (just to be used within an author-paragraph): first name, surname, title

§ You will need some kind of a configuration file that allows the mapping of the DTD elements into Word elements and vice versa.

§ You will need an SGML or XML parser to check the output SGML/ XML document against the DTD.

Conversion Methods

Often a conversion is done by using a plug into MS Word directly, but other options use the Microsoft internal exchange format RTF (Rich Text Format) for conversion. Those tools can interpret the RTF file with the MS Word style that is still coded in this RTF file and export it into an SGML document. This process mostly happens within batch mode without using many graphical user interfaces.

Within the following paragraphs we describe several approaches:

1) Approach of the Université de Montréal, Université de Lyon 2, Universidad de Chile

2) Humboldt-University Berlin and Germanwide Dissertation Online project

There are other approaches in development as well, especially within Scandinavia and the University of Oslo/ Norway. We don't refer to their solution yet.

Conversion method of the Cybertheses project

Proposition of section (Vivi)

The process line for converting Word files into SGML documents developed within the CyberThèses project uses scripts written with the Omnimark language.

Coversion Process

The input of the process line is an RTF file with a "structuring style sheet" and the output is an SGML document encoded according to the TEI Lite DTD (see the TEI web site at http://etext.virginia.edu/TEI.html).

The conversion process is made up of three main steps :

§ first, one converts the RTF file into a flat XML file encoded according to DTD of RTF. The produced file is a linear sequence of paragraph elements each one having an explicit "stylename" attribute correponding to the RTF style names.

§ the second step consists of the re-generation of the hierarchical and logical structure of the document based on the analysis of stylename attribute.

§ last, a SGML parser allows one to validate the conformity of the produced SGML document with the TEI Lite DTD.

Some supplementary scripts then allow the export of the SGML document towards other formats (HTML, XML).

Most of the scripts will soon be available from the CyberTheses web site : http://www.cybertheses.org
This system is devoted to a particular DTD, but its generalization to other document models shall not pose any difficulty.

Using SGML Author for Word (Humboldt-University Berlin)

Why did we use the SGML Author for Word?

The "Dissertation Online" project implemented and refined a conversion strategy that allows writers to convert documents written in MS word with a special stylesheet (dissertation.dot) into an SGML instance of the DiM.dtd.

We used this product from Microsoft, the SGML Author for Word, for several reasons:

1) SGML Author is quite easy to configure

2) It is easy to use.

3) It is less expensive than other software producing SGML files with the same quality.

4) It support an international standard for tables: CALS.

5) As it is a Word-Add-On it handles docuents in MS-Word doc- format better than other tools.

6) As we started using this technology in 1997, it supported from the very beginning Word97, the version of word which was the actual one that time.

Unfortunately, Microsoft didn't continue the development of this tool. So there are no new versions available for Office 2000 or Office XP. However, the internal document format from MS Word 97, MS Word 2000 and Office XP are the same in the sense of the conversion into SGML. This means documents written in Word 2000 or Office XP can be imported into Word97 and therefore a conversion can be done.

Preconditions

For a succesful conversion from a word document into a DiML document you will need:

§ The DiML-document type definition (diml20.dtd, calstb.dtd)

§ the SGML-Author for Word97 (may not available at Microsoft Shops any more, but NDLTD esp. Prof. Dr. Edward Fox may provide English versions of it that work with english Word)

§ The Association file for the Microsoft SGML-Author for Word (diml20.dta)

§ The converter stylesheet, which consists of several macros programmed to make the preconversion process easier.

§ The perl programming language (free Software)

§ The nsgmls-Parser (free Software)

§ Several perl scripts to correct the transformation of tables.

Software

You must have the following software installed at you computer:

§ SP (NSGMLS) (Parser for SGML-Files by James Clark). (new versions are availabe at http://openjade.sourceforge.net/doc-1.4/index.htm, but we haven’t tested that)

§ Run SP (A WYSIWYG tool for SP by Richard Light). http://www.light.demon.co.uk/runsp/index.htm

§ Perl (a scripting language for using the perl scripts).

The converter stylesheet and the authors stylesheet can be obtained from the following website: http://dochost.rz.hu-berlin.de/epdiss/vorlage.html

Converter scripts and perlscripts can be optained from http://www.educat.hu-berlin.de/diss_online/software/tools.exe (Perl scriptc, DTD and converter file for MS SGML-Author for Word - KonverterDiML2_0.dta)

Conversions

The conversion from a Microsoft Word document into a SGML document, which is an instance of the DiML.dtd that is used at Humboldt-University, takes several steps:

Step 1

Preparing the conversion without using *the converter Microsoft SGML Author for Word* directly

§ Check the correct usage

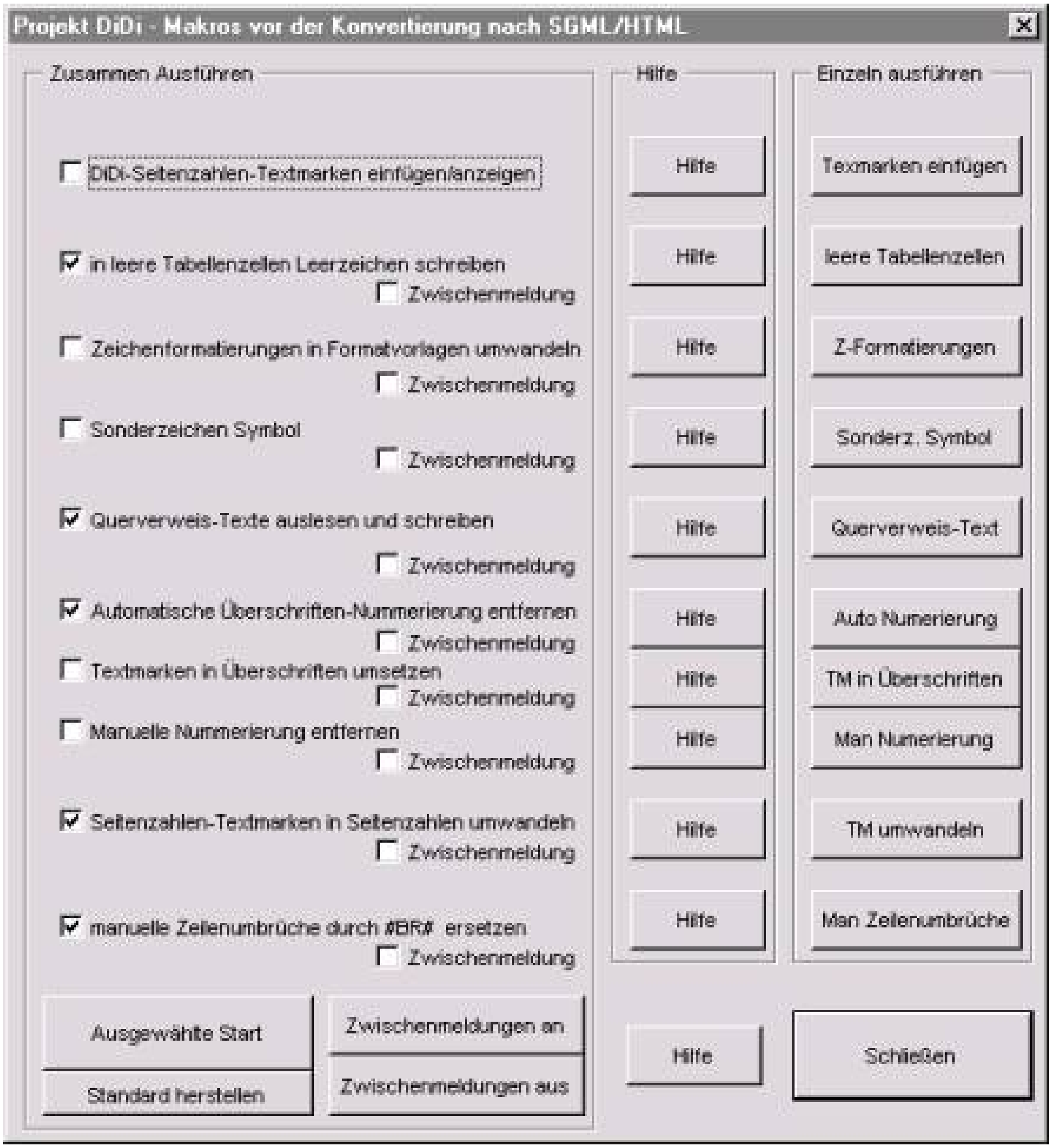
§ Load the stylesheet for conversion (NOT the one for the authors), see figure below.

§ There is a special feature to get the page numbers out of the Word document by using certain word specific text anchors. Those have to be converted into hard coded information using a page numer stylesheet.

§ Formattings that have been applied by the author without using style sheets have to be replaced by the correct style sheets.

§ In order to get a correct display of tables later on by using CSS stylesheets within common browsers, empty table cell have to be filled up with a single space (letter).

§ Soft coded line breaks have to be preserved for the conversion. This is done by inserting special characters #BR# to that. This will be used to insert later a a special SGML tag for soft line breaks <br/>.



Projekt DiDi - Makros vor der Konvertierung nach SGML/HTML

Zusammen Ausführen

☐ DiDi-Seitenzahlen-Textmarken einfügen/anzeigen

☑ in leere Tabellenzellen Leerzeichen schreiben
☐ Zwischenmeldung

☐ Zeichenformatierungen in Formatvorlagen umwandeln
☐ Zwischenmeldung

☐ Sonderzeichen Symbol
☐ Zwischenmeldung

☑ Querverweis-Texte auslesen und schreiben
☐ Zwischenmeldung

☑ Automatische Überschriften-Nummerierung entfernen
☐ Zwischenmeldung

☐ Textmarken in Überschriften umsetzen
☐ Zwischenmeldung

☐ Manuelle Nummerierung entfernen
☐ Zwischenmeldung

☑ Seitenzahlen-Textmarken in Seitenzahlen umwandeln
☐ Zwischenmeldung

☑ manuelle Zeilenumbrüche durch #BR# ersetzen
☐ Zwischenmeldung

Step 2

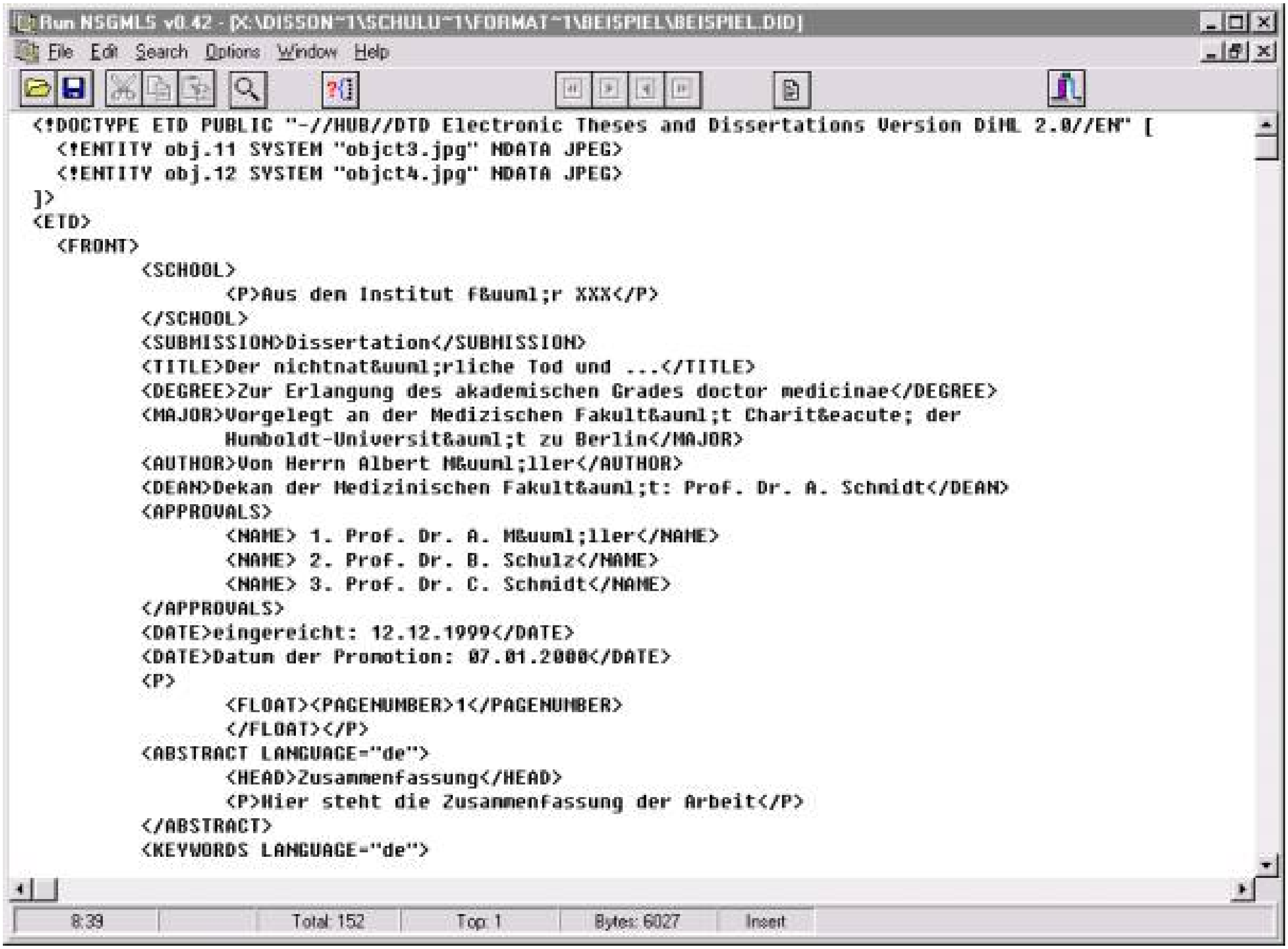
Converting with Microsoft SGML Author for Word

- Press the button "Save as SGML" within the FILE menu.
- Load the converter file KonverterDiML2_0.DTA
- Check the XML/SGML output using the feedback file (fbk) see figure below.

Step 3

Work through the output file (output according to the DiML.dtd) automatically.

- Load the perlskripts using the batch file preprocessor.bat
- Parse the DiML file


Run NSGMLS v0.42 - [X:\DISSON~1\SCHULU~1\FORMAT~1\BEISPIEL\BEISPIEL.DID]
File
Edit
Search
Options
Window
Help
8:39
Total: 152
Top: 1
Bytes: 6027
Insert

```
<!DOCTYPE ETD PUBLIC "-//HUB//DTD Electronic Theses and Dissertations Version DiML 2.0//EN" [
  <!ENTITY obj.11 SYSTEM "objct3.jpg" NDATA JPEG>
  <!ENTITY obj.12 SYSTEM "objct4.jpg" NDATA JPEG>
]>
<ETD>
  <FRONT>
        <SCHOOL>
              <P>Aus dem Institut f&uuml;r XXX</P>
        </SCHOOL>
        <SUBMISSION>Dissertation</SUBMISSION>
        <TITLE>Der nichtnat&uuml;rliche Tod und ...</TITLE>
        <DEGREE>Zur Erlangung des akademischen Grades doctor medicinae</DEGREE>
        <MAJOR>Vorgelegt an der Medizischen Fakult&auml;t Charit&eacute; der
              Humboldt-Universit&auml;t zu Berlin</MAJOR>
        <AUTHOR>Von Herrn Albert M&uuml;ller</AUTHOR>
        <DEAN>Dekan der Medizinischen Fakult&auml;t: Prof. Dr. A. Schmidt</DEAN>
        <APPROVALS>
              <NAME> 1. Prof. Dr. A. M&uuml;ller</NAME>
              <NAME> 2. Prof. Dr. B. Schulz</NAME>
              <NAME> 3. Prof. Dr. C. Schmidt</NAME>
        </APPROVALS>
        <DATE>eingereicht: 12.12.1999</DATE>
        <DATE>Datum der Promotion: 07.01.2000</DATE>
        <P>
              <FLOAT><PAGENUMBER>1</PAGENUMBER>
              </FLOAT></P>
        <ABSTRACT LANGUAGE="de">
              <HEAD>Zusammenfassung</HEAD>
              <P>Hier steht die Zusammenfassung der Arbeit</P>
        </ABSTRACT>
        <KEYWORDS LANGUAGE="de">
```

Step 4

Transforming the DiML file into a HTML file

- § Load the perl scripts by using the batch file did2html.bat
- § Check the HTML Output.
- § Correct possible errors manually within the SGML file and repeat the transformation.

A demonstration quicktime video may be found at the ETD-Guide server as well. (see http://www.educat.hu-berlin.de/diss_online/software/didi.mov)

Using FrameMaker+SGML6.0 for a conversion of MS Word documents into SGML instances.

Editing or converting using FrameMaker is much more complex than the previously described methods. FrameMaker is able to import formatted Word documents keeping the stylesheet information and exporting the document via an internal FrameMaker format as SGML or XML documents.

In order to proceed with a conversion using FrameMaker you will need the following configuration files.

1) A conversion table. This contains the list of the Word styles and the corresponding elements within the FrameMaker internal format. This table is saved within the FrameMaker internal format (*.frm).

2) A document type definition will be saved within FrameMaker internally as EDD (Element Definition) It is saved within the FrameMaker internal document format (*.edd)

3) FrameMaker uses layout rules for the internal layout of documents. Within this layout definition the layout of documents is described just like it is within MS Word documents: single formats and their appearances like text height, etc. are defined. This file is also stored as (*.frm file).

4) The Read-Write Rules contain rules that define which FrameMaker format will be exported in which SGML / XML element.

5) The SGML- or XML DTD has to be used as well, including Catalog- or Entity files, as well as Sub DTDs, like CALS for tables.

6) To process a conversion a new SGML application has to be defined within FrameMaker+SGML. This application links all files that are needed for a conversion as described above. It enables FrameMaker to parse the output file when exporting a document to SGML or XML:

A workflow and a technology for conversion for ETD using FrameMaker+SGML6.0 was first developed at the Technical University Helsinki, within the HUTPubl project (1997-2000), see http://www.hut.fi/Yksikot/Kirjasto/HUTpubl

Other Tools

Text editors, Desktop Publishing Systems that can export SGML/XML documents

Tools that export using a user specified [1][1] DTD:

WordPerfect since Version 7.0 (Corel http://www.corel.com)

FrameMaker+SGML6.0 (Adobe) (http://www.adobe.com)

Tools that export using their own native[2][2] DTD:

Openoffice (SUN/open source) (http://www.openoffice.org)

AbiWord (AbiWord/ open source) (http://www.abisource.com)

Kword (KOffice, KDE Project/ open source) (http://www.kde.org)

Converter Tools:

Omnimark (Omnimark) (http://www.omnimark.com)

MarkupKit (Schema) (http://www.schema.de)

Majix (Tetrasix) (http://www.tetrasix.com)

TuSTEP (RZ Uni Tübingen) (http://www.uni-tuebingen.de/zdv/tustep/index.html)

References

[1] Bollenbach, Markus; Rüppel, Thomas, Rocker, Andreas: FrameMaker+SGML5.5. Bonn; Reading, Mass., Addison-Wesley Longman, 1999, ISBN 3 8273 1508 5

[2] St. Laurent; Biggar, Robert: Inside XML DTDs. New York, Mc Graw Hill, 1999, ISBN 0 07 134621 X

[3] Ducharme, Bob: SGML CD. New Jersey, Prentice Hall, 1997, ISBN 0 13 475740 8

[4] Smith, Norman E.: Practical Guide to SGML/XML Filters. Plano. Texas, Wordware Publishing Inc., 1998, ISBN 1 55622 587 3

[5] Goldfarb, Charles; Prescod, Paul: XML Handbuch. München, Prentice Hall, 1999, ISBN 3 8277 9575 0

[1][1] Mit diesen Programmen kann die DiML.dtd des Projektes Dissertationen Online verwendet werden.

[2][2] Mit diesen Programmen ist kein Export in eine DiML.dtd -konforme Instanz möglich. Es werden nur herstellereigene DTDs verwendet.

4.2.3.4.2 Conversion to SGML/XML from LaTeX,

Susanne Dobratz

Looking at this problem at first gives the impression that due to the fact, that writing in LaTeX also supports a kind of structured writing, a conversion into SGML or XML compatible documents can quite easily been done.

Often people use as software Onmimark, Balise or simple Perl scripts for conversion.

Problems

Problem 1

The LaTeX format itself enables due to the structural approach an easier conversion into SGML/XML. But often the pure LaTeX approach goes hand in hand with sophisticated macro programming, so that in many cases the pure structure of a LaTeX document will be destroyed by macros or the conversion will be much harder to perform. Also the lack of a parser that checks the correct usage of structure elements like chapter, section, subsection make a conversion more complicated.

Problem 2

If an author wants to define a mathematical formula in LaTeX, he has 2 basic opportunities:

- Producing the mathematical formula as picture

- Defining them with the appropriate mathematical LaTeX features as text formulas.

First and Second Versions

The first version prevents from any secondary usage of the formula. The second version allows the reusability of a formula in different contexts. So it makes it easy to prove correctness of a statement by importing the LaTeX formula into a mathematical software package like Maple or Mathematica. Formulas coded in LaTeX can be displayed in a rendered form in an browser by software or plug-ins like IBM Techexplorer, Math Viewer . As LaTeX coded formulas still have the disadvantage, that they are not encoded using so called sematic tags, the usage of MathML is highly advised MathML is an XML document type definition for mathematics developed by the W3C.

The Letter e

In LaTeX authors often don't distinguish between the letter 'e', that may stand for a variable and the Euler constant. In MathML there is a huge difference whether something is encoded as variable e or as the Euler e (2,718. . .). Therefor, the usage of layout definition in LaTeX for mathematics complicats the conversion into MathML and therefore into any SGML/XML format.

Possible Solutions

In order to prepare mathematical formulas in LaTeX for a conversion, many universities and the TeX User Groups around the world are working soon the definition of certain macros that can be transformed into the appropriate MathML definitions.

See University of Montréal at www.theses.umontreal.ca or University of the Bundeswehr in Munich.

Rendering style-sheets, Viviane Bouletreau and Susanne Dobratz

The problem

In the case of conversion of documents into SGML/XML, the original file must be written using a structuring style sheet (template). But students often give a lot of importance to the layout of the final document and make many personalizations and adaptations on the original structuring style sheet. It is our duty to also convert also this aspect of their work. The conversion tools devoted to the document content have to be completed by conversion tools that may generate rendering style-sheets : typically XSL files for XML documents.

Layout Demand

As SGML or XML cannot be directly read by users using today's browsers (Opera, Netscape, Internet Explorer), it is necessary either to provide layout information in the form of a stylesheet or to transform those highly structured documents into layout or printable version in HTML, PDF,PS, which can be more easily used by users to read the documents.

Generally, universities have 2 possibilities to cope with the demand for layout:

> **1)** Try to preserve all layout information the author (student) has given to the MS Word or other document. This strategy involves the author directly by the personalization of a standard style sheet for his/her work.
>
> **2)** Universities could decide to develop one single style sheet or a choice of certain style sheets useable for all ETDs. This would support a corporate design and solve the layout question on a more general level, leaving it to the ETD production department.

Style Sheets for SGML or XML documents

The development of such tools is being achieved in France (collaboration between Lyon 2 and Marne-la-Vallée) for documens produced in MS Word and other RTF compatible authoring tools. Theirs are based on the analysis and the extraction of the typographic characteristics associated to each used style.

Equivalent Tools

The equivalent tools shall be easy to develop for document produced with LateX, as this kind of language natively uses the notion of rendering style-sheets.

Style sheet languages for XML are:

§ Cascading Style Sheets (CSS)

§ Extensible Style Language (XSL) .

As CSS are not powerful enough to handle the complexity and demand for large XML documents as theses and dissertations, this is not advisable.

Substandards within XSL Standard

Within the XSL standard, one distinguishes between several sub-standards:

§ Extensible Style Sheet Transformation (XSLT). This part allows the user to produce stylesheets that act like small programs. They transform the original document, that is always valid against a specified DTD, into a document that either follows another DTD (which allows an easier rendering within browsers, as HTML.dtd) or allows a transformation of the document into other document description languages such as Rich Text Format (RTF), LaTeX, PDF. From those formats the production of a printed version may be possible.

§ XPath allows to authors to build expressions that link to other XML documents, but not only to the whole document. The citation of a certain subsection and the citation of e.g. section 3 to 5 might be possible once common browsers support this linking technology.

§ XSL:Fo (Formatting vocabulary, that can be applied to the nodes of an XML document)

Example for a Printing On Demand Service (POD) based on the use of stylesheets

Digital Libraries which use their document servers as long term electronic archives will not make printed information dispensable. On the contrary: for users of these information systems the desire for printed documents is increasing. In most cases this desire often focuses not on the whole document as such, but on particular parts of it like chapters, citations and so on. For that reason the Humboldt-University Berlin's printing on demand project aims toward the development of a technology which allows the users to print the only the desired part of a certain document.

Printing on Demand with XML

For the printing on demand component with XML the usage of Apache/Cocoon was chosen. This software uses an XSLT-engine to produce an HTML or PDF-Version on the fly. "The Cocoon Project is an Open Source volunteer project under the auspices of the Apache Software Foundation (ASF), and, in harmony with the Apache webserver itself, it is released under a very open license. Even if the most common use of Cocoon is the automatic creation of HTML through the processing of statically or dynamically generated XML files, Cocoon is also able to perform more sophisticated formatting, such as XSL:FO rendering to PDF files, client-dependent transformations such as WML formatting for WAP-enabled devices, or direct XML serving to XML and XSL aware clients. "[1][1]

Cocoon

As Cocoon does not consist of a printing on demand component (especially a selection feature) a small workaround using different XSLT-stylesheets had to be created. The users view, containing an HTML-view onto the actual document includes check boxes which the user can use to select parts of a specific document. This view is produced by the XSLT-Broker stylesheet which calls a default stylesheet, that produces HTML (XSLT-Stylesheet with option document.xml?format=html). If the user selects certain parts of the document by clicking in the checkboxes, by clicking on the "OK" button a perl script (PHP-Choise) is called. This script selects the desired chapters and sections of the document by using XPath-expressions (http://dochost.rz.hu-berlin.de/proprint/bsp/slides.xml?CHAPTER=3 and http://dochost.rz.hu-berlin.de/proprint/bsp/slides.xml?CHAPTER=4) cuts those parts out of the document and holds them in the main memory. This procedure is carried out by the XSLT-Broker-stylesheet that has now been called with the XML-option (document.xml?format=xml). These parts are added to one single XML-document (all in the main memory!) and processed by the XSLT Broker-stylesheet either with the print option or the HTML option (document.xml?format=pdf or document.xml?format=html)

n

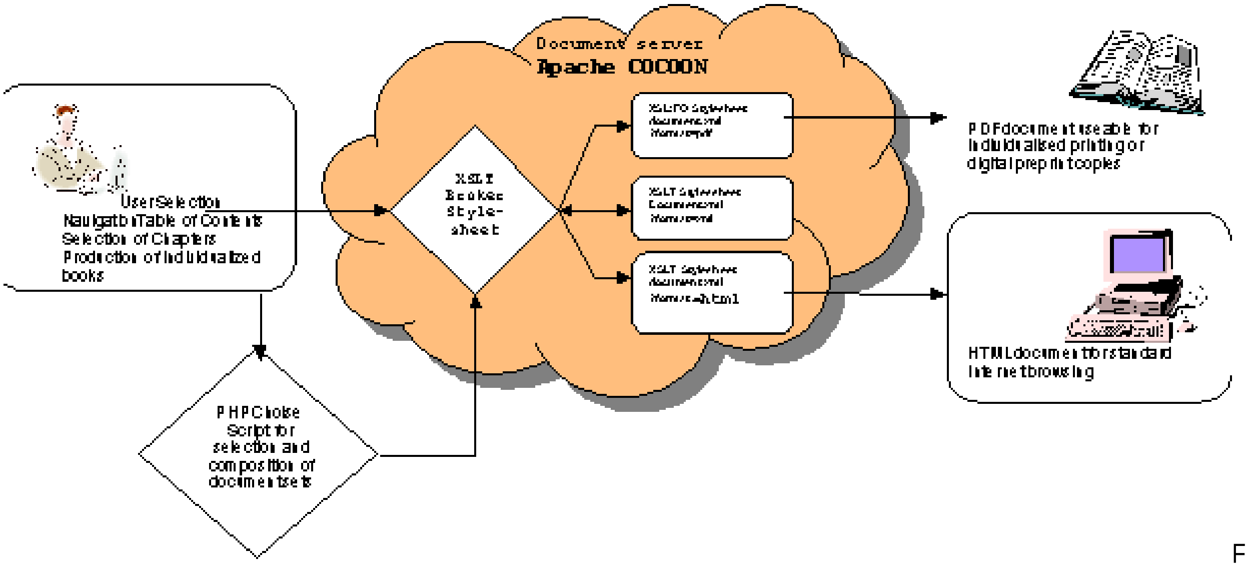

Document server
Apache COCOON
UserSelection
NavigationTable of Contents
Selection of Chapters
Production of individualized books
XSLT Broker Style-sheet
XSLFO Stylesheet document.xml ?format=pdf
XSLT Stylesheet Document.xml ?format=xml
XSLT Stylesheet document.xml ?format=html
PDFdocument useable for individualised printing or digital preprint copies
HTMLdocument for standard Internetbrowsing
PHPChoise Script for selection and composition of documentsets

Figure 1: Usage of Apache/Cocoon for Printing On Demand

---

[1][1] http://xml.apache.org/cocoon/index.html

4.2.4 Metadata, Crosswalks, Gail McMillan

With the advent of ETDs, traditional catalogers can learn about cataloging online resources, or they can apply what they already know about cataloging online resources to cataloging theses. Catalogers may need to adapt existing policies and procedures, but their workstations may already have the necessary software (e.g., word processor, PDF reader, etc.). This may also be an opportunity to implement cataloging policy changes, for example, allowing authors to assign keywords in addition to or instead of continuing to assign a controlled vocabulary such as the *Library of Congress Subject Headings*. Of course, catalogers will add the necessary MARC fields required for computer files. More information can easily be added to the bibliographic record because of the ease of copy-and-paste features of word processors, so include the abstract in online catalog records and index this field to enhance findings through keyword searching.

The following table includes MARC fields included in bibliographic records for ETDs, with additional information about Dublin Core metadata, etc. For the latest information about ETD metadata, see http://www.ndltd.org/standards/metadata/current.html

METADATA CROSSWALK http://www.dlib.vt.edu/~paul/ndltd/scm0998_imp-dev.htm (link to fuller document)

Usage	**Dublin Core Metadata elements**	**USMARC tag**	**MARC notes**	**Metadata notes**
M	1.0 DC.Title	245 $a	title	name the author assigned to the work
	1.1 DC.Title.X-Notes	500	source of title note	

M	2.0 DC.Creator.PersonalName	100 $a	author, personal	person primarily responsible for creating the intellectual content of the work
O	2.2 DC.Creator.Address			
M	2.3 DC.Creator.X-Institution	710 $a	institution's full name	name
R	2.4 DC.Creator.X-Major			
O	2.5 DC.Creator.X-College	710 $b	college's official name	
R	2.6 DC.Creator.X-Dept	710 $b	department's official name	
O	3.0 DC.Subject	653 $a	uncontrolled keywords	keywords or phrases describing the subject or content of the work
		650 $a	LCSH	
		690 $a	LCSH	
R	4.0 DC.Description.Abstract	520 $a	author's summary	textual description of the content of the work
R	5.0 DC.Publisher.CorporateName	260 $b	entity responsible for making the work available in its present form	
R	6.0 DC.Contributor.X-Chair. PersonalName	700 $a	added personal name person(s) or organization(s) who made significant contributions to the work but that contribution is secondary to the author	

R	6.1 DC.Contributor.X-Committee. PersonalName	700 $a	added personal name	
O	7.0 DC.Date.Valid			date committee approved the work in final form(s)
M	7.1 DC.Date.X-Approved	260 $c	date Graduate School approved the work	
M	8.0 DC.Type	655$2	local index term--genre/form	category of the work:Text.Thesis.Doctoral or Text.Thesis.Masters + (see DC enumer'd list)
M	9.0 DC.Format	856$q	locat.+access/file trans. mode	work's data format; ID software (+hardware?) to display/operate the work (see DC list)
M	10.0 DC.Identifier	856$u	URL (or PURL, hndl, URN)	unique ID of the work(s)
		856$b	If IP address: (Access number)	
	10.1 DC.Indentifier.X-CallNumber.LC	090$a		
O	11.0 DC.Source	786$n 0 __	Data Source Entry/Title	metadata describing original source from which the work was derived

O	12.0 DC.Language	546$a	language note	language of intellectual content of the work
		041$a	language code	
M	13.0 DC.Relation	787$n	nonspecific relationship note describes how ETD's parts relate to entire work	
		787$o	nonspecific relationship ID URL	
O	14.0 DC.Coverage	500$a	general note spatial or temporal characteristics of intellectual content of the work	
		255 $c	If spatial: cartographic statement of coordinates	
		513 $b	If temporal:period covered note	
	15.0 DC.Rights	540$a	terms governing use/repro copyright statement	
		856$u	If URL: (with $3=rights)	
M	15.1 DC.Rights.X-Availability			Should be "accessibility?" none/some/full access
O	15.2 DC.Rights.X-Availability.Notify			date to assess accessibility
O	15.3 DC.Rights.X-Proxy.PersonalName		contact re accessibility if author unreachable	
O	15.4 DC.Rights.X-Proxy.Address			

O	15.5 DC.Rights.X-Checksum.MD5	Reveals if file is corrupted. Appropriate here?
O	15.6 DC.Rights.X-Signature.PGP	Should be accessibility? none/some/full access

	M = mandatory
	O = optional
	R = highly recommended
			MARC 502 derived from type, creator's institution, date valid	

			MARC 538 = format	
	Paul Mather and Gail McMillan, Virginia Polytechnic Institute and State University, Nov. 18, 1998		MARC does not equal Metadata: missing from Metadata=file size; several notes (vita, abstract), 949/040 (more?) Leader and 008--some redundant info from variable fields	

4.2.5 Naming standards, Ed Fox

This topic is discussed in 3.3. The focus there is on naming of files and on unique identifiers. Here we consider those matters briefly, but also discuss URNs and OAI naming.

At Virginia Tech, naming of the parts of an ETD is left to authors. Many have a single document, called etd.pdf, but there are many other selections. The entire ETD is referred to by way of a summary page. This has an ID like:

etd-12598-10640 or (in earlier cases) etd-5941513972900

These two correspond to full URLs, respectively:

- http://scholar.lib.vt.edu/theses/public/etd-12598-10640/etd-title.html
- http://scholar.lib.vt.edu/theses/public/etd-5941513972900/etd-title.html

Other but similar schemes are in use elsewhere – recall 3.3.

The key point is to have an ID for each work, and a way to resolve that to the actual document. In some cases, a URN is given, such as a handle or PURL, which in turn is resolved to a URL. (See 4.3.1.) That allows users to remember the URN, and be assured that over time, as documents are moved about, there still be automatic linking from the URN to the designated document.

IDs as discussed above also are important with regard to OAI. Assuming that, as is required, each archive has a unique identifier, then as long as each record in the archive, as required, has a unique internal ID, then the combination of the two will uniquely determine a record in an archive.

4.2.6 Encryption and Watermarking,

Charles Myers

Encryption and Digital Signature Overview: Using digital signatures

Digital signatures act like conventional signatures — allowing you to “sign off” on anything that requires an approval. You can simply attach your “signature” to the document. In addition, a signature stores information, like the date and time, and allows you to track document versions and validate their authenticity.

To create a digital signature profile:

1. Choose Tools > Self-Sign Signatures > Log In.

2. In the Acrobat Self-Sign Signatures - Log In dialog box, click New Profile.

3. In the User Attributes area of the Acrobat Self-Sign Signatures - Create New User dialog box, enter your name and whatever other information you want to include in the three optional fields.

4. In the Profile File area of the Acrobat Self-Sign Signatures - Create New User dialog box, enter the path name for the folder in which you want to store your signature profile or click Browse and choose a folder. Enter a password of at least six characters in the User Password and Confirm Password fields and then click OK.

To add a digital signature to a document:

1. Click on the Digital Signature tool in the Tool bar and then click and drag where you want to place your signature.

2. In the Acrobat Self-Sign Signatures- Sign Document dialog box you can select an option from the Reason for Signing Document pop-up menu or enter a reason in the field, and you can enter a location in the Location, e.g. City Name field.
Note: If you're using a third-party signature handler, follow the instructions displayed on screen. You may be prompted to log in to the handler or enter required information.

3. Enter your password in the Confirm User Password field and then click Save Document.

4. If this is the first signature added to the document, the Save As dialog box is displayed. Enter a name and choose a location for the file and then click OK.
Note: If the Save As dialog box is displayed when you add a digital signature, you end up with two copies of the document: one unsigned and one signed.

5. From this point on, you should use the signed version.

6. To display a list of a document's signatures, click the Signatures tab in the Navigation pane. The Signatures palette's pop-up menu contains several commands for working with digital signatures; the Properties command lets you see the attributes of a digital signature.

4.2.7. Packaging, Tony Cargnelutti

The Australian Digital Theses (ADT) Program software is a modified version of the original Virginia Tech ETD software and is in its second release. The ADT software modifications were to make it generic, flexible and customisable for easy integration within the local IT infrastructure. This is critical as the ADT Program is a distributed and collaborative system involving a large number of Australian universities.

The software is distributed to all ADT members free by ftp download as a ***.tar*** file. The reason for using a ***.tar*** file is that it keeps related files together, and thus facilitates the transfer of multiple files between computers. The files in a .tar archive must be extracted before they can be used.

Extracting the distribution .tar file will reveal a directory (adt) containing 2 further directories (cgi-gin & public_html) plus installation instructions in a *readme.txt* file plus an empty adt-ADT admin site to help identify the structure of the admin side of the software.

Also included in the release package is a test site for members to look at and use to familiarise themselves on the look, structure and functionality.

Three dummy theses are used as examples of how theses can look, and how they fit into the admin structure. General software details, as well as all other information regarding the ADT Program is publicly available on the ADT Information page @: http://adt.caul.edu.au/

Responsibility for the ADT software and initial setup support for new members is taken by the lead institution - The University of New South Wales Library.

Overview of the ADT Program software.

1. Deposit Form

- generic look and feel

- includes both Copyright & Authenticity statements n completely revised help screens

n revised, as well as new alerts for errors etc..

n does not allow non compliance with core ADT standards - eg filenames; illegal symbols

n all fields & processes compulsory unless otherwise indicated on the form

2. Administration pages

n possible to edit html as well as change restrictions

n with new update function it is now possible to ADD, DELETE, RENAME files. Another feature of the update function is to be able to move files to a NOACCESS directory, as well as and to make them accessible again. This is designed primarily for certain parts of theses that cannot be made public for reasons of copyright, patents, legal reasons, and other sensitivities

n now possible to make thesis available without any restriction (ideal), restrict to campus only, restrict whole thesis for approved caveat period, totally restrict parts of thesis. Any combination is possible with the choice, or choices made, being reflected in the local view of the thesis. That is, if thesis is restricted to campus only this is now obvious, similarly if part of thesis is restricted (noaccess) this is obvious too. Knowing if restrictions apply at the outset will not frustrate those searching the database

n now possible to easily un-make a deposit. That is, to remove a thesis from public/restricted view and/or to take this back to the deposit directory where any editing and changes can be made before re-approving and making accessible again

n refined URI structure using date (yyyymmdd) and time (hhmmss)

n revised and new help & alert pages

3. Metadata

n completely revised and updated according to latest Dublin Core Qualifiers document released 11th July 2000.

n to aid search functionality keywords/phrases (ie DC.Subject) are each repeated as separate element strings. This has resulted in a revised look of the HotMeta database view. The brief record (default) now shows Title, Author, Date, Institution/School, with the expanded view showing all the metadata.

4.2.8.1 Backups, Mirrors, Susan Dobratz

An archival infrastructure for ETD should not only consider document format or the use of digital signatures, but also a consequently run concept for mirroring and backups.

Mirrors ensure, that archives run stable, regardless of usage, bandwidth and location.

A common server concept for an ETD archive is the following one:

1. Production server
2. Archive Server (Storage Unit, like IBMs Tivoli Storage Management System)
3. Public Archive server
4. Archive Mirror at different geographical location.

The **production server** is the server that communicates with the users, especially the authors. Here the uploads and metadata processing is done.

The **archive server** ensures that a permanent incremental backup of the public archive server is done. Only system administrators have access to the archive server. This secures the access to the ETDs and their digital signatures. It prevents users from manipulating ETDs and their authenticity and integrity.

The **public archive server** itself is the server that is known as the document server from the outside. Here the ETDs can be access by users, the retrieval is available. We advise to secure this server with a special RAID (Redundant Array of Independent Disks) system, that allows a security in case of a hardware failure of the disk, that might be caused by extreme usage of the server. A RAID system holds internally 2 equal copies of the server distributed on independent hard disks. So if one copy cannot be accessed due to a head crash or other hardware inconsistencies, the second copy will take over functionality and operate as the first copy. The user itself will not realize, which copy is actually in use.

Additionally to the archive server there should be another storage management system server, an **archive mirror** that mirrors the original archive. This is important if due to environmental accidents e.g. fire, earthquake or anything else destroys the original archive server. So a data loss, or the loss of the full archive can be prevented. This backup archive should be located at another geographical position, even at another continent.

4.3 Dissemination of ETDs

Ultimately, ETDs are designed to be disseminated, to at least some audience. Aspects of this activity are considered in the next subsections.

First, each ETD must have an ID, and there must be a link between the ID and the actual work. Second, each ETD must have a metadata record attached, that can be used for resource discovery and other purposes. Third, there must be a link between the ID, the metadata, and the actual body of the ETD. Fourth, the works must be made available, typically through a Web server. Sixth, there may be summaries or full works that are provided for discovery, or for indexing that is designed to lead to discovery. For example, summary pages may be indexed by Web search engines so that they may be found as a result of a Web search.

Note that it is encouraged for all NDLTD sites to keep a log regarding dissemination of the local ETDs, as well as other related information, so statistical and other reports can be prepared both for individual sites and for sites that have aggregate information (such as NDLTD).

4.3.1. Identifying URN, PURL, Guylaine Beaudry

Resources distributed on the Internet are accessible by means of a syntax which corresponds to their physical location. This syntax is defined by the RFC 1738 and is known as a Uniform Resource Locator (URL). This way of doing things creates certain problems which we must often confront. Who has not encountered the famous HTTP error 404 Not Found, which indicates that the server cannot find the location of the requested resource? This does not mean that the resource is no longer on the server, because it may simply have been moved to another location. URLs have no means of being automatically updated when a resource is moved to another place, such that we often run up against that famous HTTP error.

While the URL identifies the address of a resource, the Uniform Resource Name (URN) identifies the actual resource, the unit of information, much like the ISBN does for books. To draw a parallel, the URL corresponds to a users' postal address while the URN corresponds to users' social insurance or social security number. The URN is thus attached to a resource and not to a physical address. By knowing this identifier, it is possible to find this resource even if its physical address changes. The URN ensures an institutional commitment to the preservation of access to a resource on the Internet.

In the framework of the Université de Montréal's digital thesis pilot project, undertaken in 1999-2000, we implemented a system for producing URNs based on the model proposed by the CNRI.[1] A global server based at the CNRI manages "Naming authorities" which refer to publisher's numbers. A local server installed at the thesis distribution station in turn houses a database which manages the associations between URNs and URLs. All of this bears close resemblance to NetworkSolution's system for managing the DNS which regulate the IP addresses of computers linked to the Internet, except that in our case it is documents being given addresses as opposed to computers.

The model proposed by CNRI is the Handle system. This system is also the cornerstone for the DOI[2] Foundation's system. The construction of the Handle falls in two parts. The URN's prefix corresponds to the publisher's number (the Université de Montréal's publisher's number is 1012). This number is unique and cannot be used by any other organization. "Sub-names" can be added following this number in order to subdivide it into more precise units. This sequence is followed by a slash ("/") and a freely chosen alphanumeric sequence. Thus, a Handle-type URN for theses reads as follows:

hdl:1012.Theses/1999-Albert.Mathieu(1959)-[HTML]

We chose the year of the thesis defense, the author's name, his/her date of birth and the format of the file as the constitutive elements of a thesis' URN identifier. Please note that one must first download the CNRI's plugin in order to use the Handle system. This system has the advantage of being fairly much in conformity with the requirement of RFC 1737 concerning the framework regulating a URN system. Its application is nevertheless fastidious since one absolutely requires the plugin to be able to resolve the links. After experimentation with the CNRI's system, the Université de Montréal intends to use another system for our ongoing electronic theses project.

Another interesting avenue is the PURL[3] system created by OCLC. Let us note that a document attached to a PURL can be modified, contrary to other norms or applications for the use of a URN. The PURL system largely follows the same principle as the Handle system except that the URNs are resolved using a URL address. This solution has the advantage of not requiring the use of a plugin. In fact, a PURL is a URL. Rather than pointing directly to an Internet resource, a PURL points to an intermediary resolution service. This service associates the PURL with the active URL, which is then provided to the client. The client then normally gives access to the resource.[4] It is possible to register PURLs with an intermediary service (such as the OCLC's[5]) or to install the service on one's own server.

[1]< http://www.handle.net/ >

[2]< http://www.doi.org/ >

[3]< http://www.purl.org/ >

[4]for more details see < http://purl.oclc.org/OCLC/PURL/INET96 > and < http://purl.oclc.org/OCLC/PURL/FAQ >

[5]< http://purl.oclc.org/ >

4.3.2. Metadata models for ETD's, Ana Pavani

One of the objectives of an ETD program is to yield easy access to TDs. Since we are dealing with digital libraries, we are implicitly dealing with libraries. One of the actions performed on a library catalog is that of *search and retrieve*. This is the first step towards accessing the contents of a library item; the second step is the use (read, listen, view, etc) of the item.

In order to be efficient in the *search and retrieve* action, the user must search a catalog in which the items were properly identified, besides using good search functions.

This section is about the identification of ETD's, which is a very important step towards their dissemination. The identification will be accomplished through the use of the metadata elements whose set is named the metadata model of the digital library of TDs.

Before we address metadata models for ETDs, it is important that some ideas are brought to the discussion. These ideas are related to the choice of a model to be considered later on. These models must be rich and versatile to contain information of different natures and to be searched by users from all over the world.

It is obvious that the richer and more versatile the metadata model is, the more time and effort it takes to capture (collect and record) the information into the digital library. The decision on which model to use will have to take this into consideration. In some situations it may be necessary to adopt the simplest possible model in order to make the metadata capture viable. Later in this chapter the Dublin Core

Metadata Element Set will be introduced. It seems that it is the consensus of the minimum identification to be used for ETD's.

The ideas for us to think about are:

Many languages in one world

Our world is a very diverse linguistic place. Those who work with information and are involved in international projects know English. This is the language they use to communicate, to access the Internet, to read technical literature, etc.

At the same time, not only many other languages exist but some of them have large numbers of native speakers. The 100 most spoken languages of the world, when first language speakers are counted, can be found in http://www.sil.org/ethnologue/top100.html. In descending order, the first 10 are Chinese (Mandarin), Spanish, English, Bengali, Hindi, Portuguese, Russian, Japanese, German (Standard) and Chinese (Wu).

If only the other 9 languages are considered, it is not hard to imagine the numbers of texts that are written and published every year. The same happens with TDs. The number of TDs published in languages other than English must be very big.

ETD's to be read all over the world

One of the purposes and benefits of an ETD program is to yield easy access to the results presented in TDs, no matter where the reader is and where the dissertation was written.

We assume that ETD digital libraries are to be connected to the Internet so that their contents can be shared worldwide, to make sure this benefit is accomplished.

Metadata

Metadata are data about data or information about information.

The metadata elements are the attributes used to describe a digital library item just like the ones used to catalog items in a traditional library.

Many of these attributes are language dependent, as for example titles, abstracts, subjects, keywords, etc. Others obviously are not, as for example authors' names, digital format, number of bytes of the file, etc.

Since some metadata elements are language dependent and TDs are written in many languages, we can expect that most probably the metadata will use the language of the work. This can pose a problem for search and retrieve activities since most of us are not fluent in as many languages as we would like to be.

Contents and instances

The items of a digital library may be identified in 2 different levels; the same way the items of a traditional library are. The first level is the *content* which is equivalent to a title of a traditional library and the second is the *instance* which is equivalent to a volume.

A *content* is the logical definition of an item of the digital library and it is identified by a set of attributes. An *instance* is the physical realization of a content or title. It is a digital object and is identified by a set of attributes too.

The use of contents and instances allows contents to have multiple instances either in different formats or due to physical partitions. This will yield a one to many relationship among contents and instances.

The use of contents and instances also allows the access control to be performed on the partitions instead of on the content. This makes the digital library more flexible in terms of dealing with intellectual property rights.

Therefore, we can conclude that there are attributes that are particular to contents and others that refer to instances. The metadata model must contain both.

Contents, instances and metadata

Some metadata elements are common to all contents, as for example title, abstract, type, etc, while others are common to all instances, as for example electronic format, access level, etc.

On the other hand, some metadata elements are specific to some contents, as for example translation control - original content, translator, etc, and others are specific to some instances, as for example special equipment, expiration date, remote location, etc.

From this comment, we can see that the metadata model must be versatile to contain attributes that are common to all contents and to all instances and also the specific ones, in order to accommodate specialization of the digital library items.

Contents, instances and languages

Contents may be language dependent. The language of the content is the one in which it is written, spoken or sung.

Other languages may be associated with a content - the ones in which it is catalogued. It is possible to describe a content written/spoken/sung in one language in other language(s). This way, there is one catalog entry in each of the languages to be used.

The use of multilingual cataloguing yields points of access in different languages if the search is performed in all of them. This topic will be addresed in section 4.3.4 Database and information retrieval.

Metadata models and languages

It is possible to define the digital library to hold more than one language. A good choice would be, at least, the language(s) of the nation where TDs are developed and English.

If this is the case, the metadata model can have all attributes that are language dependent written in each language to be used in the digital library and the language code must be a part of the primary key in the database.

Attributes that are language independent would have only one representation in the database.

Metadata schemes

There are quite a few metadata schemes. Some are strictly related to library items while others have a broader scope, as for example the ones devoted to digital objects to be used in Web Based Education. Some schemes are well known and should be mentioned:

- DCMES - Dublin Core Metada Element Set
http://purl.org/dc/documents/rec-dces-19990702.htm

Under the responsibility of the DCMI - Dublin Core Metadata Initiative http://www.purl.oclc.org/metadata/dublin_core/ http://purl.org/dc/ This metadata element set will be presented in section 4.3.3. Cataloguing: MARC, DC, RDF

- n IMS Project - Instructional Management System Project http://www.imsproject.org/ The metadata element set defined by the IMS Project has the objective of identifying digital objects used in Web based Education. It contains all the elements of the DCMES and many more.

- n LOM - Learning Objects Metadata of the Learning Technology Standards Committee of the Institute of Electrical and Electronics Engineers (LTSC/IEEE) http://ltsc.ieee.org/doc/wg12/LOM_WD4.htm/ The metadata element set defined by the LTSC/IEEE (http://ltsc.ieee.org/) has the objective of identifying digital objects used in Web based Education. It contains all the elements of the DCMES and many more.

- n LoC - Core Metadata Elements of the Library of Congress http://lcweb.loc.gov/standards/metadata.html

The second and the third are used when WBE is under. Since they contain the DCMES, no conflict exists to the general digital library identification.

Specialization of the metadata models for TDs

Besides the usual data contained in general purpose metadata schemes, there are some types of information related to TDs that may be of interest to the university. For this reason, it may be useful to consider adding extra metadata elements to the traditional metadata schemes.

The additional elements can be separated in 3 groups:

n Administrative information - department, date of presentation, date of acceptance, financial support, etc.

n Academic information - level, mentor, examining committee, etc.

n Traditional library information - university, library system, control number, call number, etc.

These may be useful to yield information concerning the graduate programs of the university.

Conclusion - metadata models for ETD's

The definition of the metadata model for an ETD digital library must combine:

n The needs for proper identification of ETD's for the goals of access to be achieved (national access? international access?)

n The administrative needs of the university

At the same time, the restrictions imposed by budget or operation time frames must be to taken into consideration. There is a balance between what is desired and what is possible. Some comments concerning this balance are made:

n For international access, the use of English besides the original language(s) is mandatory. This means that titles and abstracts must be translated, and that subjects headings, keywords, etc will be multilingual catalogs to be maintained.

n For the ETD digital library to be a part of the international community, the minimum requirements in terms of ETD identification must be met. This means that at least the DCMES must be used.

n For the university to have good control of the intellectual property, the use of content / instance concept allows access specifications to be established on the digital objects. Thus, some objects may be made public while others may have different types of restrictions due to format or to intellectual content.

n In the definition of the workflow to operate the ETD program, attention must be given to the capture of the metadata elements. If non-librarians are involved in the process, there must be a good training program and a careful review process so that the attributes are catalogued right.

The choice to the metadata model is very important and the team in charge of the implementation of the ETD program must study the possibilities before making the decision. Minimum standards must be met.

4.3.3 Cataloguing: MARC, DC, RDF, Ana Pavani, Gail McMillan

Many traditional library systems exchange and store records using the MARC Format (MAchine Readable Catalog), one of the realizations of the ISO 2709 Standard. The MARC Format has approximately 1,000 fields, many with subfields which can be repeated. The use of this format allows a very detailed description of the items. This format has a specific field (856) to identify of electronic objects associated with the intellectual item and its other physical instances.

The DCMES (Dublin Core Metadata Element Set - http://dublincore.org/documents/dces/) is a set of 15 attributes divided into 3 groups: content, intellectual property and instanciation. Associated to them there are the Dublin Core Qualifiers (http://dublincore.org/documents/dcmes-qualifiers/) that enhance the identification of the items.

There is a relation between the MARC Format and the DCMES since there is an intersection between the 2 sets of attributes.

The RDF (Resource Description Framework - http://www.w3.org/TR/1999/rec-rdf-syntax-19990222/) is a foundation for processing metadata. It specifies a representation for metadata as well as the syntax for encoding and transporting this metadata. The objective is to yield interoperability of Web servers and clients, and to facilitate automation of processing of resources. It can be used to describe Web pages, sites or digital libraries.

The DCMES can be used with the RDF representation.

Considering the MARC Record and the Cataloging Department Work Flow

MARC Bibliographic Records

Catalogers may want to focus initially on what fields are currently included in the MARC bibliographic record for theses and how these would be the same or different for ETDs. The MARC record for theses is not very robust and often has a local twist, presenting valuable information in a unique format that can be seen only at the originating institution. Optimally, author, title, abstract, and other relevant bibliographic information would be programmatically adapted to the appropriate MARC fields. The extent of *AACR2r* compliance was another complicating factor. For example, would programming change upper-case letters to lower case?

Requiring authors of theses (in all formats) to provide keywords for use in the bibliographic record may enhance search results. Assigning Library of Congress subject headings (or other controlled vocabulary) is very time consuming, so having the authors assign the uncontrolled subject headings may be an appropriate alternative. MARC tag 653 would be appropriate for author-assigned terms. Without LC subject headings, however, these MARC records would be considered "minimal level," rather than full level, cataloging. This seemed particularly unjust to because the ETD bibliographic records would actually be more robust than previous theses cataloging because additional information is included.

The catalogers on the ad hoc task force suggested including tables of contents (MARC tag 505) and abstracts (MARC tag 520) since the standard copy-and-paste features of today's word processors would make this a relatively easy process. The table of contents for dissertations, however, proved to be quite generic, usually containing only the standard dissertation topics (e.g., literature review, methodology, findings, etc.), and, therefore, not an enhancement to the information available about the work in the OPAC. The abstract, however, contains valuable information and provides valuable information about the research topic. The 520 is also an indexed field in our online catalog and, therefore, a word-searchable field for OPAC users. Adding the abstract (250-350 words) can, however, add tremendously to the length of the MARC record. *See figure 2 and figure 3.*

Cataloging conventions have not generally included the name of the thesis author's department as a standard feature of the bibliographic record. This is an opportunity to modify local cataloging practices

Taking advantage of the opportunity to incorporate changes in theses cataloging, consider using MARC tag 502, the dissertation note field, to include the degree, institution, and year the degree was granted, expanding institution to include the name of the department. The new note would follow this example:

502 Thesis (Ph. D. in Mechanical Engineering)--Virginia Polytechnic Institute and State University, 1955.

Evaluating the potential value of an e-thesis bibliographic record provides the opportunity to propose a substantially enhanced record of real and continuing value to OPAC users. *AACR2r* compliance may be an issue. In reviewing her *Cataloging Internet Resources: A Manual and Practical Guide,* Nancy Olson states that when cataloging Internet-accessible documents, consider them to be published documents. Therefore, publisher information belongs in field 260 of an e-thesis record. [Coming from a serials background it seems reasonable to add a 710 for this corporate body tracing.] Additional fields required for cataloging computer files include tag 256 for computer file characteristics, tag 538 for notes of system details, and tag 856 for formatted electronic location and access information. These additional fields (505, 260, 710, etc.), however, also increase the length of the record and, therefore, should be carefully considered as should the usefulness of the information provided in meeting the combined needs of OPAC users and computerized access and retrieval systems.

--

MARC bibliographic record for an ETD

--

VT University Libraries - - - - - - - - ADDISON - - - - MARC BIBLIOGRAPHIC RECORD

[OCLC fixed field tags]

Local lvl: 0 Analyzed: 0 Operator: 0000 Edit: Type cntl:

CNTL: Rec stat: n Entrd: 010608 Used: 010706

Type: a Bib lvl: m Govt pub: s Lang: eng Source: d Illus: a

Repr: Enc lvl: K Conf pub: 0 Ctry: vau Dat tp: s M/F/B: 0

Indx: 0 Mod rec: Festschr: 0 Cont: b

Desc: a Int lvl: Dates: 2001,

[003-049 system assigned fields and information]

```
1. 001    ocm47092981 010608
 2. 003    OCoLC
 3. 005    20010608092044.0
 4. 006    m        d s
 5. 007    c \b r \d u \e n \f u
 6. 035    1475-05560
 7. 040    VPI \c VPI
 8. 049    VPII
1. 099    Electronic Thesis 2001 Alvarez
 2. 100 1  Alvarez, Leticia, \d 1973-
 3. 245 14 The influence of the Mexican muralists in the United States
          \h [computer file] : \b from the new deal to the abstract
          expressionism / \c Leticia Alvarez.
 4. 256    Computer data (1 file)
 5. 260    [Blacksburg, Va. : \b University Libraries, Virginia
          Polytechnic Institute and State University, \c 2001]
 6. 440  0 VPI & SU. History. M.A. 2001
 7. 500    Title from electronic submission form.
 8. 500    Vita.
```

9. 500 Abstract.

10. 502 Thesis (M.A.)--Virginia Polytechnic Institute and State
University, 2001.

1. 504 Includes bibliographical references.

2. 520 3 This thesis proposes to investigate the influence of the Mexican muralists in the United States, from the Depression to the Cold War. This thesis begins with the origins of the Mexican mural movement, which will provide the background to understand the artists2 ideologies and their relationship and conflicts with the Mexican government. Then, I will discuss the presence of Mexican artists in the United States, their repercussions, and the interaction between censorship and freedom of expression as well as the controversies that arose from their murals. This thesis will explore the influence that the Mexican mural movement had in the United States in the creation of a government-sponsored program for the arts (The New Deal, Works Progress Administration). During the 1930s, sociological factors caused that not only the art, but also the political

ideologies of the Mexican artists to spread across the United States. The Depression provided the environment for a public art of social content, as well as a context that allowed some American artists to accept and follow the Marxist ideologies of the Mexican artists. This influence of radical politics will be also described. Later, I will examine the repercussions of the Mexican artists2 work on the Abstract Expressionist movement of the 1940s. Finally I will also examine the iconography of certain murals by Mexican and American artists to appreciate the reaction of their audience, their acceptance among a circle of artists, and the historical context that allowed those murals to be created.

1. 538 System requirements: PC, World Wide Web browser and PDF reader.
2. 538 Available electronically via Internet.
3. 653 mural painting \a WPA \a abstract expressionism
4. 856 40 \u http://scholar.lib.vt.edu/theses/available/etd-05092001-130514

5. 945 NBJun2001

6. 949 dpm/tm 06/07/01

7. 994 E0 \b VPI

--

Figure 3:

OPAC display for an electronic thesis from the Virginia Tech VTLS opac

--

VT University Libraries - - - - - - - - ADDISON- - - - - - - - - - - -FULL RECORD

CALL NUMBER: Electronic Thesis 2001 Alvarez

Author: Alvarez, Leticia, 1973-

Title: The influence of the Mexican muralists in the United States

[computer file] : from the new deal to the abstract expressionism / Leticia Alvarez.

File Type: Computer data (1 file)

Imprint: [Blacksburg, Va. : University Libraries, Virginia Polytechnic Institute and State

University, 2001]

Series: VPI & SU. History. M.A. 2001

Note: System requirements: PC, World Wide Web browser and PDF reader.

Note: Available electronically via Internet.

Remote Acc.: http://scholar.lib.vt.edu/theses/available/etd-05092001-130514

Note: Title from electronic submission form.

Note: Vita.

Note: Abstract.

Note: Thesis (M.A.)--Virginia Polytechnic Institute and State University, 2001.

Note: Includes bibliographical references.

Abstract: This thesis proposes to investigate the influence of the

Mexican muralists in the United States, from the Depression

to the Cold War. This thesis begins with the origins of the

Mexican mural movement, which will provide the background to

understand the artists2 ideologies and their relationship

and conflicts with the Mexican government. Then, I will

discuss the presence of Mexican artists in the United

States, their repercussions, and the interaction between censorship and freedom of expression as well as the controversies that arose from their murals. This thesis will explore the influence that the Mexican mural movement had in the United States in the creation of a government-sponsored program for the arts (The New Deal, Works Progress Administration). During the 1930s, sociological factors caused that not only the art, but also the political ideologies of the Mexican artists to spread across the United States. The Depression provided the environment for a public art of social content, as well as a context that allowed some American artists to accept and follow the Marxist ideologies of the Mexican artists. This influence of radical politics will be also described. Later, I will examine the repercussions of the Mexican artists2 work on the Abstract Expressionist movement of the 1940s. Finally I will also examine the iconography of certain murals by Mexican and American artists to appreciate the reaction of their audience, their acceptance among a circle of artists, and the historical context that allowed those murals to be

created.

Key Words: -- mural painting

In terms of the broader topic of bibliographic control of electronic publications, focus on adding to current cataloging practices those fields that would enhance the OPAC users' access and conform to *AACR2r*. So many of the fields describing computer files appear to be redundant; 245 \h, 256, 516, and 538, for example; which tell the OPAC user over and over that the item is a computer file. To stay within the stringent restrictions of full-level cataloging, the members of the task force saw no way to avoid requiring catalogers to use most of the available fields. Concentrate on the MARC tags that would provide information about access. The principal fields include: 256 (computer file characteristics), 506 (restrictions on access note), 516 (type of computer file or data note), 530 (other formats available), 538 (system details note), 556 (accompanying documentation), and 856 (electronic location and access).

Current OPACs, in addition to the limitations of hardware and workstations, however, still prevent most users from accessing electronic texts or images directly and smoothly from one menu or even from a single, multi-function workstation. However, workstations are gradually becoming available that permit users to copy the URL from the bibliographic record and paste it into a World Wide Web browser for accessing an e-text from a single terminal. Knowing this was possible include MARC tag 856.

Another issue that must be addressed is using subfield u or splitting the URL into the multiple subfields. We went for simplicity and decided to format the 856 subfield u so that it could be copied and pasted into a World Wide Web browser. Again, we were not willing to wait for the programming that would be necessary to combine the separate subfields into a clickable URL.

Cataloging has greatly benefited from advances in library automation and the cataloging of e-texts is ripe for further automation. It is now possible to derive MARC cataloging from text mark-up languages, subsets of XL, SGML such as TEI (Text Encoding Initiative) headers, and possibly even HTML (hypertext markup language) tags.

As one way of getting from the e-text to the MARC record, consider programmatically tagging or having authors tag the basic record including author, title, publisher, file size, file type, abstract, formatted contents notes, and the like. However, it will be more timely to use the submission form, begun by the author, added to by the Graduate School, and then the library, as the basis for the cataloging record. The submission form asks the author to supply the following information, to which I have added the MARC tags.

ETD Submission Form

Name: [MARC tag 100]

Title: [MARC tag 245]

Document Type (check one):

Abstract: [MARC tag 520]

Keywords: [MARC tag 653]

1.

2.

3.

4.

5.

6.

Department: [MARC tag 502]

Degree: [MARC tag 502]

Filename(s), size(s): [MARC tag 256]

1.

2.

3.

4.

In addition to considering the MARC record and the cataloging department work flow, consider a procedure for getting the files from the Graduate School (approving unit), the mechanics of making an ETD available to a cataloger (from the secure and private environment of the server) and for moving an e-thesis into public access. Have the cataloger forward a copy as each ETD is processed to a server at UMI. If UMI would prefer batch processing, files could be accumulated (i.e., stored in a directory on the e-theses server) for batched file transfer, or perhaps a UMI-access point could be established on the ETD server from which its staff could retrieve them.

With input from the University Archivist and addressing a concern of the Graduate School's, long term preservation and access of ETDs should also be factored into the procedures. A plan may include periodically writing ETDs to CD-ROMs for security back-ups and possibly longer term preservation. While this is may not be the final answer, an alternative has not been brought forward; how frequently this would be done has also not been determined.

Processing Electronic Theses: a possible scenario

----

1. Graduate School

* Electronically transfers approved file to library theses server

E-mails Thesis Transmission Form to library thesis coordinator

2. Library/Cataloger

* Downloads ED from closed server to her workstation

* Prepares cataloging (see new features above in figure 4)

* Adds a screen to the file that includes the call number and property "stamp" (using "memos" feature of Acrobat)

* Move file to server for public access

* Electronically send file to UMI or move to UMI holding file

3. Library Theses Server Administrator

* Indexes text for word searchability

* Integrates new index with existing index

* Maintains server, including weekly back-ups (stored on site) and monthly tapes (stored off site)

* Removes files from closed server following completed processing

*   Makes CD-ROMs

4.   Special Collections Department/University Archives

*   Retains CD-ROMs

*   Works with Theses Server Administrator as necessary to ensure that archival files are accessible

----

Theses and dissertations as electronic files may be the first major source of electronic texts that many libraries encounter regularly. Seize this opportunity to enhance the OPAC users search results by expanding current theses cataloging and taking advantage of online information prepared by authors. Since authors will probably not be adding TEI, MARC, or other element tags to their documents to help cataloging in the near term, catalogers could use the information available in a variety of online sources including the document itself or from the online submission form to provide the basic descriptive MARC fields. Whether programmatic changes can be made or standard copy-and-paste features of word processors are incorporated, enhancing the ETD bibliographic record does not require a lot of extra work.

See also "Electronic Theses and Dissertations: Merging Perspectives," chapter in *Electronic Resources: Selection and Bibliographic Control*, Pattie, Ling-yuh W. (Miko), and Bonnie Jean Cox, eds. New York: Haworth, 1997 (105-125). [Simultaneously published in *Cataloging and Classification Quarterly*, 22(3/4)]

4.3.4. Database and information retrieval, Ana Pavani

The next step after identification of the items of the digital library, the ETDs, is to address storage of the cataloguing attributes and the action of searching and retrieving. Remember that the quality of retrieval is dependent both on the programming of the search and retrieve functions but, as important as this, on the quality of the information used to catalog the items of the collection.

Databases are common and suitable tools to store, search and retrieve information. Besides this, they can also be very helpful in the process of capturing the attributes since they have the general function of managing information.

Before implementing the database, the database model must be created. This will happen only after the metadata model has been defined and related to other existing identification procedures such as traditional cataloguing on an automated library system.

If the traditional OPAC is to be maintained during the ETD program, it is desirable to avoid duplicated information. Thus, the attributes that are present in the OPAC should not be repeated and a link between the OPAC record and the ETD metadata is be created.

No matter where information is stored, the user should be able to perform the types of search that are standard in library systems: author, title, keywords, subjects, ISBN, etc.

As mentioned in section 4.3.2., some information that is used to identify the items are language dependent (title, keywords, subjects, etc). If the database holds only one language per record, search procedures are to be performed using the arguments in this language. If a multilingual database is modelled, it is recommended that search be language independent, i.e., that the argument be checked against all languages. In this last situation, after a record is found as athe result of s search, all its language instances should be displayed to the user so that he/she can choose the language for retrieval.

Database extenders (text, image) may be considered to increase the number of points of access by performing the searches in the ETD's not only on the cataloguing attributes.

The relation with the legacy systems databases must be examined since information concerning the TDs may be stored on them. Some examples are the graduate program, the mentor, the examining committee, etc.

The next sections address specific aspects of this topic.

4.3.4.1 Packaged solutions

Since 1995, there have been plans to share work developed in connection with ETDs so as to help others. The first such effort was funded by the Southeastern Universities Research Association (SURA), to spread the work around the Southeast of the USA. Virginia Tech acted as the agent for this effort, creating software, documentation, training materials, and other resources. In particular, it became clear that software was needed to support student submission, staff checking, library processing, and support for access. The Virginia Tech resources were packaged in connection with these efforts and the follow-on work funded by the USA’s Department of Education. These were made available starting in 1996 through local and NDLTD sites, and updated regularly since.

The following subsections give details of the Virginia Tech solutions, as well as extensions to it, and alternatives. NDLTD offers to assist such sharing by providing a clearinghouse for whatever seems appropriate and useful to share.

4.3.4.1.1DiTeD – Digital Theses and Dissertations, Jose Luis Borbinha, Nuno Freire

Theses and dissertations are traditionally covered by the legal deposit law in Portugal[1]. Nowadays, almost all the thesis and dissertations are created using word processors, just confirming the fact that science and technology became one of the first areas for digital publishing.

In this context, the deposit of theses and dissertations emerged as an ideal case study for a scenario concerned with a specific genre. For that, the National Library of Portugal promoted the project **DiTeD – Digital Thesis and Dissertations** [1] from which a software package with the same name originated.

Requirements

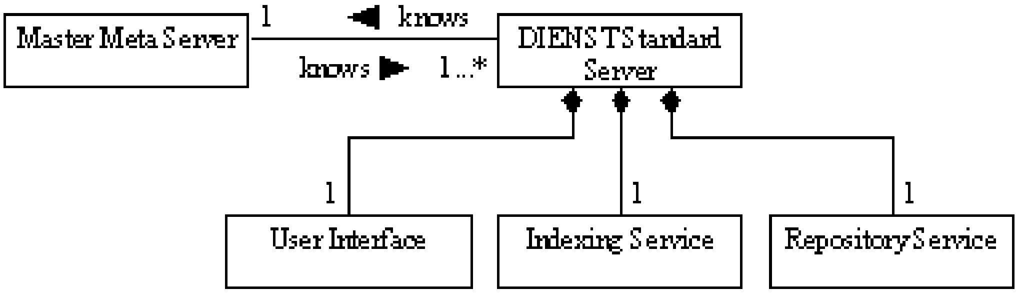
Theses and dissertations carry special requirements for registration and access, since their contents are usually used to produce other genres, such as books and papers, or they can include sensitive material related with, for example. patents. This requires the management system functionality to make it possible to the authors to declare special requirements for access, which have to be registered and respected.

Universities have a long tradition of independence in their organization, culture and procedures. As a consequence, soon it was learned that it would be impossible to reach, in the short and medium term, any kind of overall agreement for common formats or standard procedures with the different administrative services. Therefore, the main objective defined for DiTeD was the development, on the top of the Internet, of a framework that would connect the National Library to the local university libraries and would make it possible to support a full digital circuit for the deposit of theses and dissertations.

In this sense, the technical framework should de designed as a distributed and asynchronous architecture, composed by local modules that would implement autonomous local digital libraries, and a central module at the National Library for the formal deposit. Theses and dissertations would be uploaded in the local modules using the local networks, and arrive at the National Library by the Internet after the successful execution of a specific workflow. This workflow should be implemented locally at the university libraries, or centrally at the National Library, or it could be even a combination of both the cases.

Architecture

A solution for this framework was found in the DIENST technology [3], which provides a good set of core services. DIENST also has an open architecture that can be used with great flexibility, making it possible to extend its services and build new functionalities. The basic entities of this architecture are shown in Figure 1, as a class diagram in UML – Unified Modeling Language.



1. Figure 1: Entities of the DIENST architecture.

Master Server

The Master Meta Server provides the centralized services, including a directory of all the local servers members of the system. Only one of these servers must exist in each system.

In DiTeD this server exists at the National Library. It was renamed Master Server, and differs substantially from the original versions developed for DIENST. The original server was designed to manage only metadata, while now it is necessary to manage also the contents of the theses or dissertations and give support to the workflow for its submission and deposit.

DIENST Standard Server

The DIENST Standard Server is the server installed at the university libraries. This server was modified in DiTeD, and renamed Local Server. The following core modules compose it:

- **Repository Service**: This is where the documents are stored. It manages metadata structures and multiple content formats for the same document, functions that were substantially extended in DiTeD (to support a specific metadata format, as also to recognize a thesis or dissertation as possibly composed by several files). It is also possible to define and manage different collections in the same server.

- **Index Service**: This service is responsible for indexing the metadata and responding to queries. Small adjustments were made in DiTeD to support diacritics in the indexes and queries, a requirement in the Portuguese writing.

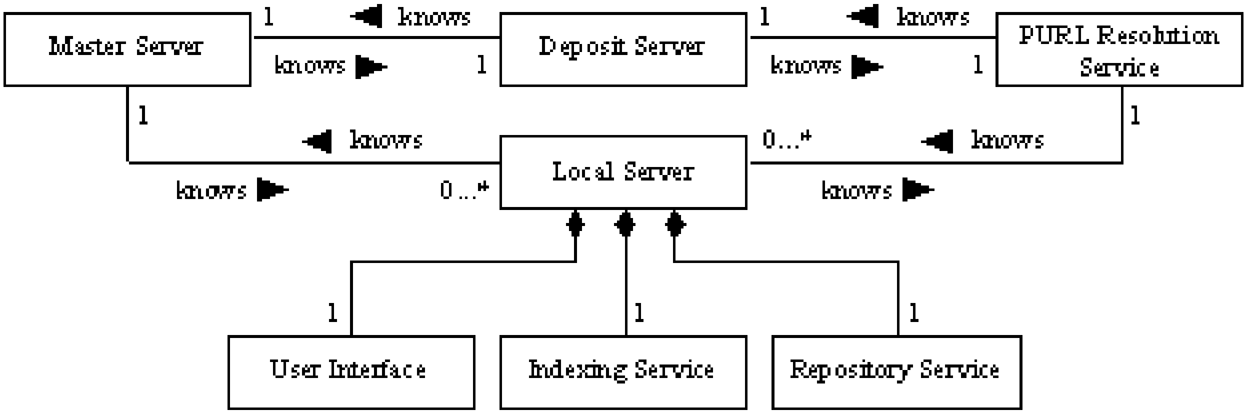
- **User Interface**: This service is responsible for the interaction with the user. It was extended in DiTeD to support a flexible multilingual interface and a workflow for submissions using HTTP.

Identifiers

Two Local Servers are running at the National Library. One, named Deposit Server, is used to locally store the deposited theses and dissertations coming from all universities (the deposit will consist in a copy, so in the end each thesis or dissertation will exist in two places, the Local Server and the Deposit Server). A second Local Server is used as a virtual system for those university libraries that do not have the necessary technical resources or skills to maintain their own server.

Each thesis or dissertation deposited in DiTeD automatically receives a URN [4], which will be registered and managed by a namespace and resolution service. This is in fact a simple implementation of the concept of PURL – Persistent URL [2], with the particular property that it resolves any PURL by returning its real URL in the original Local Server, unless it is not available anymore. In this case, it resolves it by returning its URL in the Deposit Server. The entities of this final DiTeD architecture are shown in Figure 2.

The prefix of the URN has the form “HTTP://PURL.PT/DITED”, while the suffix is formed by an identifier of the university library (the “publisher”) and by a specific identifier of the work itself, automatically assigned locally.



2. Figure 2: Entities of the DiTeD architecture.

Workflow

The workflow comprises two main steps: submission and deposit.

Submission

The submission process comprises the following steps:

Delivery: The process starts with the submission by the student of the thesis or dissertation to a local server. In this step the student fills a metadata form, where it is recorded the bibliographic information and the access conditions. All of this information is hold in a pending status, until it is checked.

Verification: In a second step a librarian checks the quality of the submission (a login in the local server gives access to all the pendent submissions). This task is supposed to be assured by a local librarian, but it can be also assured remotely, such as by a professional from the National Library (in a first phase of the project, this task will be assured by the National Library, especially to assure uniformity in the criteria and test and tune the procedures).

Registration: If everything is correct (metadata and contents), the thesis or dissertation is stored in the local repository, and the student receives a confirmation. Otherwise, the student is contacted to solve any problem, and the submission remains in the pending status.

Deposit

The deposit consists in the copy of the thesis or dissertation, as also of its metadata, from the Local Server to the Deposit Server. This is done in the following steps:

What’s new: Periodically, the Master Server contacts the repository of a Local Server to check if there are new submissions. The Local Server replies, giving a list of the identifiers of the new submissions.

Delivery: For each new submission, the Master Server sends a request to the Local Server to deposit it in the Deposit Server. Because this Deposit Server is also a Local Server, this deposit works just like a normal local submission.

Verification: A librarian in the National Library checks the deposit. This double checking is important, especially in the first times of the project, to reassess the procedures and test the automatic transfer of files over the Internet –not always a reliable process).

Registration: If everything is correct, the thesis or dissertation is stored in the deposit repository, the final URN (a PURL) is assigned, and both the student and the local librarian receive a confirmation. The metadata is also reused to produce a standard UNIMARC record, for the national catalogue. If it detected any problem, the local librarian is contacted and the deposit remains in the pending status.

One can argue that, if the Deposit Server is really also a Local Server, than the first step would be excused and the Local Server could perform the delivery automatically after a successful submission. This can be a future optimization, but for now the reason for this extra step is to preserve the requirement of an asynchronous system, making it possible for the Master Server, for example, to better control the moment of the deposit (such as to give preference for the night periods).

Metadata

DiTeD utilizes a metadata structure for theses and dissertations defined by the National Library and coded in XML. This structure contains descriptive bibliographic information about the work and the author, as well as information about the advisers and jury members, access conditions, etc. This metadata structure is configurable at installation time, making the software flexible for use in other countries, or even with other publication genres. Metadata may also be accessed and exported in other formats, like UNIMARC and Dublin Core.

Multilingual interface

DiTeD’s user interface has multilingual capabilities, allowing the users to switch between the available languages at any time. The base configuration includes English and Portuguese.

Software availability

The software is maintained by the National Library of Portugal, and distributed freely for non-commercial use.

Access to the software package may be requested by email to dited@bn.pt.

References

[1] <http://dited.bn.pt>

[2] <http://purl.org>

[3] <http://www.cs.cornell.edu/cdlrg/dienst/software/DienstSoftware.htm>

[4] Sollins, K; Masinter, L. (1994). Functional Requirements for Uniform Resource Names. RFC 1737.

[1] The legal deposit law corresponds to a deposit system legally enforced, whereby authors, publishers or other agents must deliver one or more copies of every publication to the deposit institution (this is the case of the National Library of Portugal, for example).

4.3.4.1.2. ADT, Tony Cargnelutti

Australian Digital Theses Program

http://adt.caul.edu.au/]

ADT Program is a collaborative Australia-wide university libraries initiative. Membership is open to all Australian universities and is voluntary. National coordination of the ADT is via the Council of Australian University Librarians [CAUL; an umbrella group representing all Australian university libraries]. The ADT Steering Committee is currently investigating a proposal to further extend ADT Program & software to the Australasian region.

ADT software was designed to be transportable and to be flexible enough to be 'plugged in' at each member institution with minimum modification. The ADT software was also designed to automatically generate simple core metadata which is gathered automatically to form national distributed 'metadata' database of ADT-ETDs

ADT software is based on the original Virginia Tech [VT] software. The original ADT v1.0/1999 released in April 1999, with upgrade v1.1/2000 released in October 2000.

ADT software basics :

- perl scripts extended by the library cgi.pm, which uses objects to create web forms on the fly and parse their content

- facilitates file uploading and form handling by generation of html via function call and passing of the form state n extension of variable use to make scripts more generic to facilitate local institutional setup and style

n developed a standard for generation of unique addresses [URLs] for deposited documents

n automatic generation of DC metadata from the deposited information

The distribution software package requires modification of a list of variables and some webserver-dependent adjustments so it runs in standard way for each of the local institutions. Local search software, which may be institution-mandated, is not included in the package. It is expected that each institution will use their own security accordingly.

ADT software overview:

n Deposit Form : generic look and feel; includes copyright and authenticity statements; complete set up help screens; quality control using alerts when errors are made and does not allow non compliance with core ADT standards

n Administration pages: possible to edit html and change document restrictions; also possible to add, delete, rename files as well as moving files to no access directory to restrict files temporarily for copyright reasons; varying levels of restrictions possible; easy to 'un make' deposit

n Metadata: revised and updated according to latest DC Qualifiers document; generated and gathered automatically, used to create central national searchable 'metadata' database

n ADT Standards: few, simple but necessary for success of collaborative program. Core standards are - research theses only; PDF document format; PDF filename convention; unique URL; Metadata standard.

Summary:

n the ADT software & model is fully transportable and designed to be easily installed locally by any participating institution regardless of local IT infrastructure and architecture

n the ADT software facilitates loading digital versions of theses to the local institutions' servers where the PDF files will be housed permanently

n the theses can be fully integrated into the local access infrastructure and searched using any local database, and/or the local web based catalogue

n all ADT Program theses can also be searched nationally via the ADT Program metadata database. This database is constructed from DC metadata generated automatically during the deposit process. The metadata gathered creates rich records that allow highly flexible and specific searching, with links back to the local institutions' servers where the full digital theses are housed

n the ADT Program is a collaborative effort across the whole Australian university community and a proven model for creating a national dataset of digitised theses

n the ADT software and model is relatively inexpensive to install, integrate within local requirements and process. Once this is achieved it is virtually maintenance free, is sustainable, scaleable, and very cost effective for both the institution and the student/author

n the ADT metadata and other standards conform to current internationalstandards and therefore have potential to integrate with other international open archive initiatives

n theses can be deposited from anywhere, and similarly, the metadata can be gathered from anywhere

n full details and information available from the ADT homepage

Please see accompanying diagram - ADT Architecture@aglance

4.3.4.1.3. Cybertheses

The www.cybertheses.org site is the result of cooperative project that started at first between l'Université de Montréal "http://www.theses.umontreal.ca/" et l'Université Lumière, Lyon 2 "http://www.univ-lyon2.fr/, supported by the Fonds Francophone des Inforoutes "http://www.francophonie.org/fonds/", and dealing with the theme of the electronic publication and distribution of theses on the Web.

At first, this cooperation dealt with the conception and creation of a production line for electronic documents, using the SGML norm. It also had the objective of setting up a server to be shared by the different participating establishments so as to allow their theses to be indexed.

In a desire for openness, we decided to enlarge participation on this server to all establishments of higher education distributing full-text versions of their thesis on the Internet, without constraints based on the language used or on the chosen format of distribution.

The www.cybertheses.org site allows theses to be indexed on line using a common metadata model. Its implementation provides a structure for this growing cooperative effort by basing itself on the Internet's own modes of functioning and of distributing skills. Our wish is that it can quickly ensure a better distribution of the research work conducted within the participating establishments and serve as an effective tool for the entire community of researchers.

From the conception of the project, the partners wanted to contribute to the distribution of software tools created for use in the university environment. These tools are available to the partners in the Cybertheses network. They permit all the partners to participate, on equal footing, in the construction of an electronic university library that functions in a dispersed manner and applies the concept of distributed intelligence.

One of the principles that motivated the implementation of this program is to favor, as much as possible, the re-appropriation of research work by the researchers themselves. Our objective is to eventually help create a new political economy of knowledge, which makes researchers the masters of their publications, beyond any economic constraints. Many aspects of our project respond to this goal :

n Placing theses freely and completely on-line on the Internet, and thus widening their distribution, means that the thesis will no longer be considered as solely the result of a research project. It will instead become a genuine work instrument integrated to a much larger system in order to satisfy the user's demand.

n The creation of the Cybertheses database containing the metadata of the participating institutions' theses. Cybertheses provides an efficient indexation system and rapid searching, even while significantly increasing the visibility and the distribution of the theses.

n The use of free or freely distributed software is favored at each stage of the production and distribution process.

n The sharing of research, self-created software programs, and documentation between the Cybertheses partners, thereby creating a "toolbox" permitting participation in the program.

The Cybertheses partners are concerned with developing solutions that can be used by all actors, be they of the North, South or East. Our procedure in based on appropriating the skills and techniques related to the electronic production and distribution of the university community's research results.

For more information, consult: http://www.cybertheses.org/

4.3.4.1.4 VT DB and other tools,

Anthony Atkins

This page describe the hardware and software requirements involved in setting up your own ETD database.

Hardware

To use this software, you must have a web server available. At Virginia Tech we use a UNIX-based server platform. You should allocate enough disk space on your machine for at least a year's worth of submissions. Our site averages 2.5 Megabytes per submission. Keep in mind that it's better to have more space early on, as the scripts are not designed to deal with a collection which spans multiple drives. You should also have enough memory to handle the web server, the database server, and whatever other tasks you have in mind. As an example, our site uses a dual-processor Sun Enterprise 250 with 384 Mb of RAM, running Solaris 2.7. Our machine has an 18Gb drive allocated solely for the ETD collection.

Software

Before you can make use of the scripts provided here, you must have the following installed:

Mysql

Mysql is a database server and client which partially implements the SQL 9.2 standard. It is many ways similar to other SQL databases, such as Oracle, Postgres, and miniSQL. UNIX versions of Mysql are made available without charge to education institutions at http://www.mysql.com/. A version of

Mysql for Windows NT is available as well for an additional charge, although the scripts are not designed for use with Windows NT.

Perl

All of the scripts included with this distribution are written using perl. Perl is also freely available (under the GNU public license) from http://www.perl.com/CPAN/. It is recommended that you download, install, and test the latest version available for your operating environment.

CGI.pm

The CGI module for perl is one of the most widely used and best supported libraries of CGI oriented routines in existence. Virtually all of the query handling performed in these scripts relies heavily on CGI.pm. CGI.pm is available from http://www.perl.com/CPAN-local/modules/by-module/CGI/.

The DBI and DBD:Mysql modules for perl

The DBI module for perl is a generic set of database calls designed to interface with a wide range of different database technologies in a powerful, reliable, and easy to understand way.

To make use of the DBI module, you need a DBD module for the particular database you intend to use. The DBD:Mysql module (also known as the DBD:Msql module) allows you to easily perform all types of database operations on a Mysql database from within perl.

Both modules are available from http://www.perl.com/CPAN/modules/dbperl/.

The Tie-IxHash module for perl

The Tie-IxHash module is a very small add-in that allows you to reliably output hashes in the order they are defined in. Without this module, none of the global hashes that contain department names, degree information, etc. would appear in the order we'd like them to. This module is available at http://www.perl.com/CPAN-local/modules/by-module/Tie/

Web Server Software

The perl scripts provided are designed for the most part to be used through a CGI interface, meaning that you must have a compatible web server installed. The freely available Apache Web Server is our recommendation, although any web server capable of seamlessly handling html output from perl scripts should be acceptable. Apache is available from http://www.apache.org/.

Once you have all the preceding items installed and tested, you should be ready to download and install the scripts

4.3.4.1.5 Library automation: OPACs, VTLS software, Edward Fox

ETDs are intimately connected with libraries. The first key automation of libraries, launched in the 1980s, involved computerization of the card catalogs. Online Public Access Catalogs (OPACs) were developed, and supported search through records describing library holdings.

Many library catalogs have their holdings encoded according to the MARC (machine readable cataloging) standard. Managed by the US Library of Congress, the MARC standard in the USA has been applied in many other contexts, leading to broader standards like UNIMARC. The latest standard as of 2001 is MARC 21.

VTLS, Inc. (www.vtls.com), is one company that sells software (e.g., their Virtual system, which works with MARC) and services to support catalog search and other related library requirements. OPACs in many cases have evolved into sophisticated digital libraries, and can be used to help manage ETD collections. Indeed, ETDs at most universities can be found by searching in the local catalog system, as long as one can distinguish this genre or type of work from other holdings (e.g., books, journals, maps).

VTLS has volunteered equipment, software, staff, and services to help NDLTD. From www.vtls.com/ndltd one can gain access to a system affording search and browse support to the union catalog of all ETDs that can be harvested using the mechanisms supported in the Open Archives Initiative.

4.3.4.1.6 Harvest Usage in Germany, Susanne Dobratz

The HARVEST system is often used for a fulltext search within the ETD archives. In Germany, most of the university libraries are using this particular software.

Preconditions for using Harvest

Before the installation, one should check whether the following technical preconditions are fulfilled.

Hardware:

fast processor(e.g. Sparc5...)

fast I/O

enough RAM (> 64 MB) – and 1-2 GB free disk space (sources 25 MB)

Operating Systems supported

DEC OSF/1 ab 2.0

SunOS ab 4.1.x

SunSolaris ab 2.3

HPUX

AIX ab 3.x

Linux alle Kernel ab 1999 on

alle Unix-Platformen

WindowsNT

Following additional software is needed to use Harvest:

Perl v4.0 and higher (v5.0)

gzip

tar

HTTP-Server (with remote machine)

GNU gcc v2.5.8 and higher

flex v2.4.7

bison v1.22

Harvest Components

The Harvest system consists of two major components:

The Harvest Gatherer

The Harvest Broker.

This allows establishing a distributed retrieval and search model.

Installation procedure (ftp://ftp.tardis.ed.ac.uk/pub/harvest/develop/snapshots/)

Gatherer

This program part is responsible for collecting the full text and metadata of the dissertations. The Gatherer visits several sites regularly, sometimes daily or weekly and builds an incremental index area database. The collected index data are held in a special format, called SOIF-Format (Summary Object Interchange Format). The Gatherer can be extended so that it can interpret different formats.

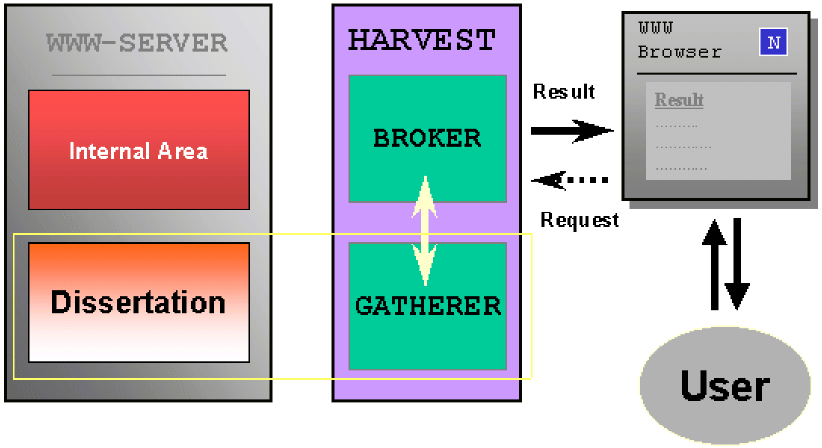
Broker

This part of the software is responsible to provide the indexed information by using a search interface. The broker operates as Query-Manager and does the real indexing of the data.

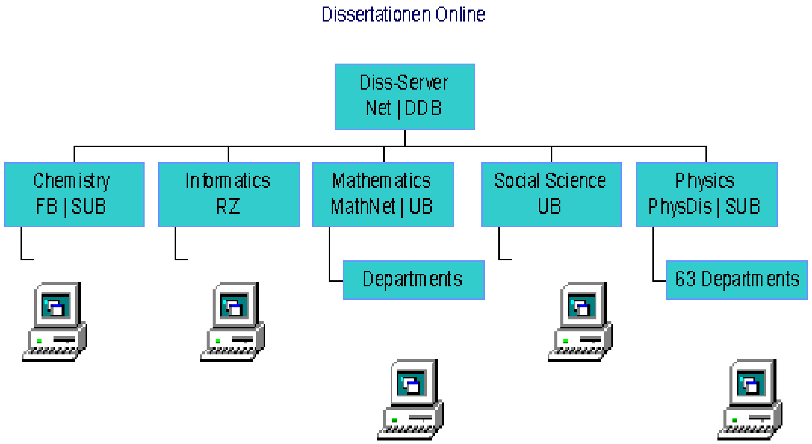
Using a Web based interface he can reach several Gatherer and Broker simultaneously and perform several search requests.

In Germany there has been established a Germany-wide retrieval interface on the basis of the Harvest software, called The (Theses Online Broker), which is accessible via: http://www.iuk-initiative.org/iwi/TheO

Within NDLTD a special Broker has been set up to add the German sites using Harvest to the international search.



- Figure 1: Harvest System Architecture (from Kerstin Zimmermann)



- Figure 2: Architecture for a distributed Harvest Network (by Kerstin Zimmermann)

Harvest is able to do a search within the following document formats:

C, Cheader	Extract procedure names, included file names, and comments
Dvi	Invoke the Text summarizer on extracted ASCII text
FAQ, FullText,README	Extract all words in file
Framemaker	Up-convert to SGML and pass through SGML summarizer
HTML	Up-convert to SGML and pass through SGML summarizer
LaTex	Parse selected LaTex fields (author, title, etc.)
Makefile	Extract comments and target names
ManPage	Extract synopsis, author, title, etc., based on `-man"
News	Extract certain header fields
Patch	Extract patched file names
Perl	Extract procedure names and comments
PostScript	Extract text in word processor-specific fashion, and pass through Text summarizer.
RTF	Up-convert to SGML and pass through SGML summarizer
SGML	Extract fields named in extraction table
SourceDistribution	Extract full text of README file and comments for Makefile and source code files, and summarize any manual pages
Tex	Invoke the Text summarizer on extracted ASCII text
Text	Extract first 100 lines plus first sentence of each remaining paragraph
Troff	Extract author, title, etc., based on `-man", `-ms", `-me" macro packages, or extract section headers and topic sentences
Unrecognized	Extract file name, owner, and date created

Configuration for PDF files:

Before the Harvest-Gatherer can collect PDF documents and transform into SOIF format it has to be configured.

Using only the standards configuration ignores the format. In order to make a format known to the Gatherer a summarizer for PDF has to be build:

Delete the following line in the file /lib/gatherer/byname.cf:
Pdf ^.\.(pdf|PDF)$*

Configuire the PDF summarizer. Use Acrobat to transfer PDF documents into PS documents, that are used by the summarizer. A better choice provides the xpdf packages by Derek B. Noonburg (http://www.foolabs.com/xpdf). It contains a PDF-to text converter (pdftotext), that can be integrated into, the summerizer Pdf.sum:

```
    /usr/local/bin/pdftotext $1
    /tmp/$$.txt Text.sum
    /tmp/$$.txt rm /tmp/$$.txt
```

Configuring the Gatherers for HTML-Metadata

The Harvest-Gatherer is by standard configured to map every HTML metatag into an SOIF attribute, e.g. <META NAME="DC.Title" CONTENT="Test"> into an own SOIF attribute, that is equal to the NAME attribute of the metatag.

The configuration can be found at:

```
    <harvest home>/lib/gatherer/sgmls-lib/HTML/HTML.sum.tbl
```

The summarizer table has entries like this:

```
    <META:CONTENT> $NAME
```

If a retrieval should be done only in the HTML metatags, meaning within certain SOIF attributes, those attributes have to put in front of the search request and put into the retrieval forms provided to the user, e.g.

```
   DC.Title: Test
```

Searching Metadata encoded in HTML

As in Germany there is a nationwide agreed metadata set for ETDs, those are searchable within the German wide Harvest network.

The following example shows, how those Dublin Core metadata (only a small part is displayed) is encoded within the HTML front pages for ETDs:

```
<META NAME="DC.Type"                          CONTENT="Text.PhDThesis">

<META NAME="DC.Title" LANG="ger"        CONTENT="Titelseite: Ergebnisse der CT- Angiographie bei der Diagnostik von Nierenarterienstenosen">

<META NAME="DC.Creator.PersonalName"         CONTENT="Ludewig, Stefan">

<META NAME="DC.Contributor.Referee"    CONTENT="Prof. Dr. med. K.- J. Wolf">

<META NAME="DC.Contributor.Referee"    CONTENT="Prof. Dr. med. B. Hamm">

<META NAME="DC.Contributor.Referee"    CONTENT="PD Dr. med. S. Mutze">
```

For this agreed metadata set, there has been formulated a suggestion, on how the metadata can be produced and operated at the university libraries. The following schema shows the details:

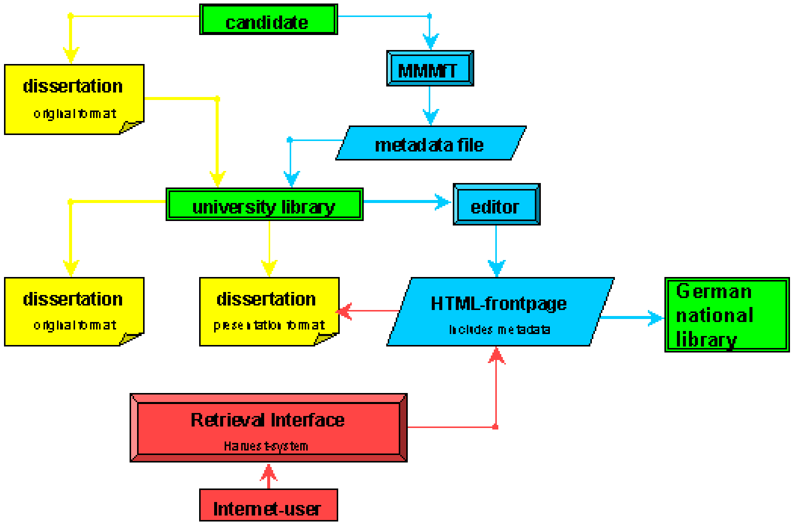
The doctoral candidate uploads his ETD document to the library. He does this while filling in an HTML form, that collects internally metadata in Dublin Core format.

The university libraries check the metadata of correctness and the ETD of readability and correct usage of style sheets.

The university library adds some descriptive metadata to the metadata set and put a presentation version of the ETD on its server. During this procedure an HTML format page containing the Dublin Core metadata encoded as HTML metatags will be produced.

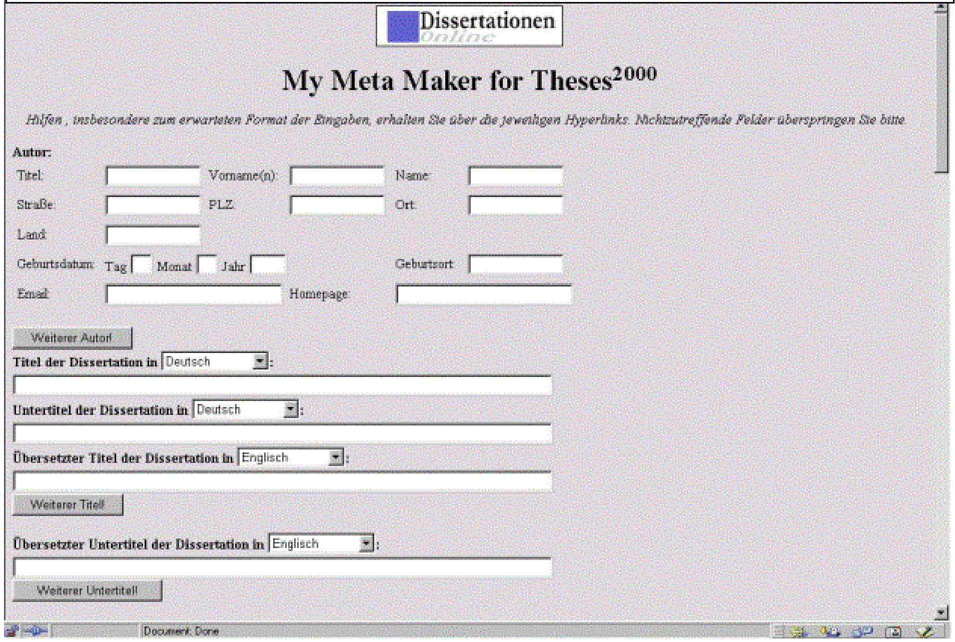
At last submits the university library the metadata to the National library, which is in charge of archiving all German language literature.

The national library copies the ETD and metadata to its own internal system.



- Figure 3: Germanywide proposed metadata process: schematical view (by Thorsten Bahne)

Figure 1: Metadata Tool for Authors and Library (see http://zoe.mathematik.uni-osnabrueck.de/cgi-bin/MMMfT2000/MMMft.cgi)



Dissertationen Online

My Meta Maker for Theses^{2000}

Hilfen , insbesondere zum erwarteten Format der Eingaben, erhalten Sie über die jeweiligen Hyperlinks. Nichtzutreffende Felder überspringen Sie bitte.

Autor:

Titel: ___ Vorname(n): ___ Name: ___

Straße: ___ PLZ: ___ Ort: ___

Land: ___

Geburtsdatum: Tag ___ Monat ___ Jahr ___ Geburtsort: ___

Email: ___ Homepage: ___

Weiterer Autor!

Titel der Dissertation in Deutsch:

Untertitel der Dissertation in Deutsch:

Übersetzter Titel der Dissertation in Englisch:

Weiterer Titel!

Übersetzter Untertitel der Dissertation in Englisch:

Weiterer Untertitel!

Figure 1: Theses Online Broker (TheO): the wide Search Interface for ETDs using Core Metadata Tags coded in HTML (at http://www.iuk-initiative.org/iwi/TheO)

Searching SGML/XML documents

Harvest also allows a search within SGML/XML DTD (document type definition) elements.

All that has to be done to configure the Gatherer component according to the following rules:

Within the home of the Harvest software (written as <harvest-home>) in /lib/gatherer/byname.cf a line has to be added: DIML ^.*\.did$. (DiML is the DTD used at Humboldt-University, did is the file names of the SGML documents accorning to the DIML-DTD). This says the Harvest-Gatherer, which Summerizer should be used, if documents ending with .did are found.

Now, the summarizer has to be build and saved as DIML.sum within the filesystem at <harvest-home>/lib/gatherer: (The summarizer contains the following line*: #!/bin/sh exec SGML.sum ETD $**

Within the catalog file <harvest-home>/lib/gatherer/sgmls-lib/catalog the following entries have to be done: (They point to the public identifiers of the DIML and from DIML used DTVs)
DOCTYPE ETD DIML/diml2_0.dtd
PUBLIC "-//HUB//DTD Electronic Thesis and Dissertations Version DiML 2.0//EN" DIML/diml2_0.dtd
PUBLIC "-//HUB//DTD Cals-Table-Model//EN" DIML/cals_tbl.dtd
PUBLIC "-//HUBspec//ENTITIES Special Symbols//EN" DIML/dimlspec.ent

Now <harvest-home>/lib/gatherer/lib/sgmls-lib/DIML can be created (mkdir <path>) and four files copied into the path:
diml2_0.dtd, cals_tbl.dtd, dimlspec.ent und diml2_0.sum.tbl (DTD, entity file and sumarizer table). The file diml2_0.sum.tbl consists of the DTD tags that should be searcheable and the appropriate SOIF attributes

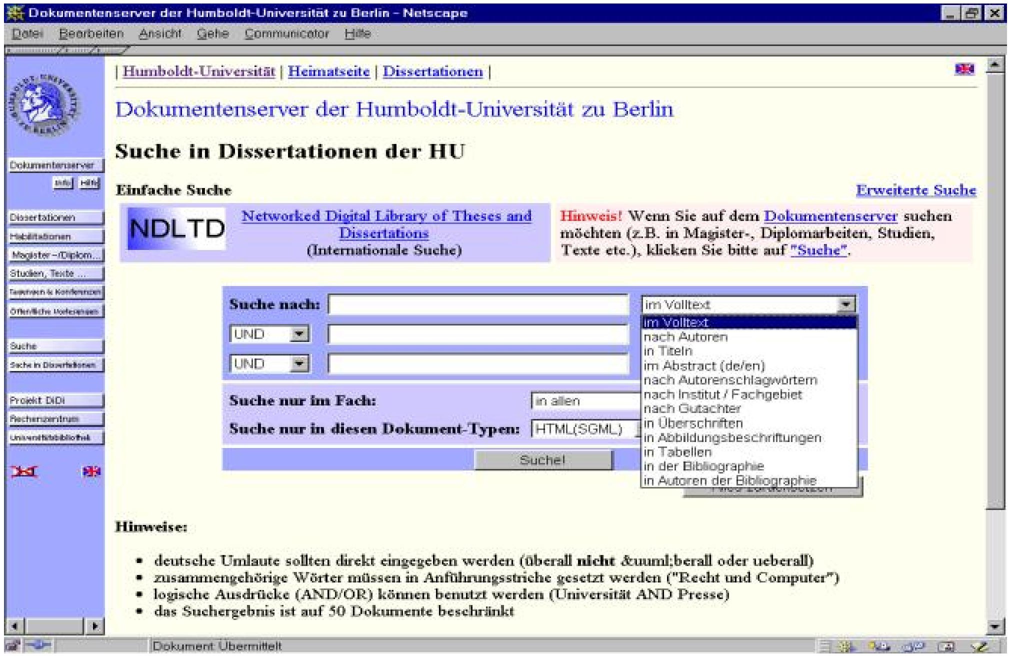
The Gatherer can be launched now.

In order to saerch within certain SOIF tags, the name of the SOIF attribute has to be put in front of the search term, e.g. *searching for "title: Hallo"means, searching within the SOIF-attribute 'title' for the search term 'Hallo'.*

At Humboldt-University Berlin, there has been installed a prototype that allows a retrieval within documents structures, so a user may search within the following parts of a document and therefore specialize the search in order to retrieve only the wanted information and hits:

- Fulltext (im Volltext)
- For authors (nach Autoren)
- In titles (in Titen)
- In abstracts (Im Abstract)
- Wthin authors keywords (in Autorenschlagwörtern)
- For Institutes/ Subjects (nach Institut/ Fachgebiet)
- For approvals (nach Gutachtern)
- Headings of chapters (Überschriften)
- Captions of figures (in Abbildungsbeschriftungen)
- In Tables (in Tabellen)
- Within the bibliography (in der Bibliographie)
- For Autor names within bibliography (nach Autoren in der Bibliographie)

Figure 1: Search INterface of Humboldt-University, enables a search within SGML/XML structures of collection (at http://edoc.hu-berlin.de)



| Humboldt-Universität | Heimatseite | Dissertationen |

Dokumentenserver der Humboldt-Universität zu Berlin

Suche in Dissertationen der HU

Einfache Suche

Erweiterte Suche

NDLTD Networked Digital Library of Theses and Dissertations (Internationale Suche)

Hinweis! Wenn Sie auf dem Dokumentenserver suchen möchten (z.B. in Magister-, Diplomarbeiten, Studien, Texte etc.), klicken Sie bitte auf "Suche".

Suche nach: im Volltext

UND

UND

im Volltext
nach Autoren
in Titeln
im Abstract (de/en)
nach Autorenschlagwörtern
nach Institut / Fachgebiet
nach Gutachter
in Überschriften
in Abbildungsbeschriftungen
in Tabellen
in der Bibliographie
in Autoren der Bibliographie

Suche nur im Fach: in allen

Suche nur in diesen Dokument-Typen: HTML(SGML)

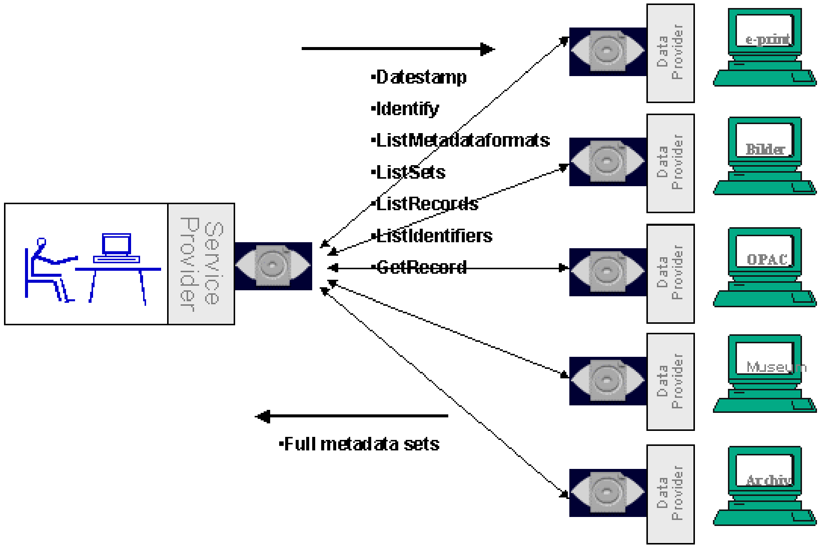
Suche!

Hinweise:

- deutsche Umlaute sollten direkt eingegeben werden (überall **nicht** &uuml;berall oder ueberall)
- zusammengehörige Wörter müssen in Anführungsstriche gesetzt werden ("Recht und Computer")
- logische Ausdrücke (AND/OR) können benutzt werden (Universität AND Presse)
- das Suchergebnis ist auf 50 Dokumente beschränkt

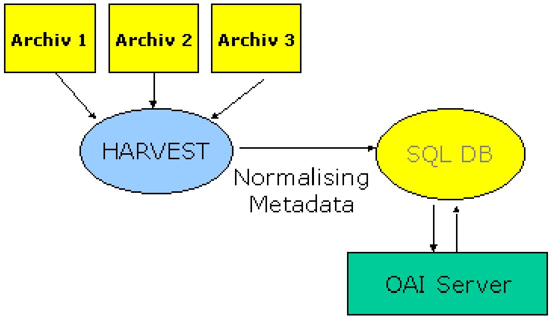
Harvest and OAI

With the growing enthusiasm for the approach of the open archives initiative, where a special OAI software protocol allows to send out requests to document archives and receive standardized metadata sets as answers, there are ideas on how this could be connected with the Harvest-based infrastructure that has been set up in Germany.



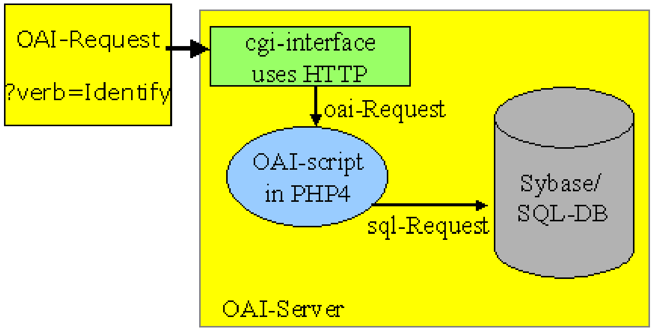
- Figure 7: Schematic View: the OAI Retrieval process

Making a Harvest archive OAI compliant means, that the information, that the Gatherer holds, has to be normalized (same metadata usage) and that the index has temporarily saved within a database. The Institute for Science Networking at the Department of Physics at the University of Oldenburg, Germany developed the following software. This software, written in php4, uses an SQL database to perform the OAI protocol requests. The SQL database holds the normalized data from the Harvest Gatherer.



- Figure 8: How to configure Harvest to be used with the OAI specification (by Heinrich Stamerjohanns)

Other university document servers, like the one at Humboldt-University additionally hold the Dublin core Metadata within an SQL-Database (Sybase) anyway. There a php4 script operating at the cgi-interface reads the OAI protocol requests that is transported via the HTTP protocol and puts them into SQL statements, which are than used as requests for the SQL-database. The database respond is given in an SQL syntax as well, which is than transformed into OAI protocol syntax using XML and Dublin Core.(see http://edoc.hu-berlin.de/oai)



- Figure 9: Providing a OAI interoperability for archives using a database model for their metadata

4.3.4.1.7 The NDLTD Union Catalog (*www.vtls.com/ndltd*)

The Networked Digital Library of Theses and Dissertations is an interconnected system of digital archives. After NDLTD became a reality, the logical next step was to give users the means to search and browse the entire collection of theses and dissertations. Much research went in to deciding how best to do this, and a couple of proto-solutions were examined. In the end, by adopting the Metadata Harvesting Protocol of the Open Archives Initiative to gather metadata in the ETDMS format, NDLTD was able to make the collection of ETDs accessible via a central portal. VTLS Inc., using their Virtua--Integrated Library System (Virtua ILS), is hosting this portal, providing a Web interface to the ETD Union Catalog. This portal gives users a simple and intuitive way to search and browse through the merged collection of theses and dissertations. Once relevant items are discovered via a browse, keyword, or Boolean search, users can follow links that go directly to the source archives or to an alphabetical index that allows for further searching.

VTLS' flagship product, Virtua ILS, is especially suited to the needs of NDLTD. For starters, it is inherently a distributed system and adheres to emerging standards, such as the Unicode® standard, for encoding metadata. Secondly, VTLS' Web portal, called Chameleon Gateway to emphasize its changeability, is not only straightforward and easy to use, but highly customizable and translatable, giving global users the ability to retrieve information *and* enter searches in a multitude of languages. Currently, versions of the NDLTD portal interface exist in 14 languages, including Arabic, Korean, and Russian, and plans call for support of all languages used by NDLTD members.

4.3.5.1. Metadata (searching), Simon Pockley

In addition to creating an ETD, students now need to become their own cataloguers. The tools being developed by the NDLTD will assist in the submission of a basic description of the ETD but there are skills to be developed and issues to be aware of.

Students will have already discovered that one of the greatest barriers to finding information is the difficulty of coming up with the right terminology. Lists of standardized subject heading terms, structured thesauri, and fielded searching have been created to remedy this problem.

Accurate metadata improves precision and increases the recall of the content of ETDs by using the same standardized terms or elements. However, even if common metadata elements are used, there is no guarantee that the vocabularies, the content of the elements, will be compatible between communities of interest. Students and researchers working within specialized areas sometimes forget that language and terms often have particular and precise meanings. Outside this field of interest, global searches can return too much of the wrong information.

Generating accurate metadata requires some of the basic skills of resource description and good practice in avoiding three language problems that cause poor precision:

n Polysemy: words with multiple meanings. For example, if we are searching for an article that discusses types of springs and their uses, we might retrieve articles on freshwater springs or on the season of spring, as well as on leaf springs, flat springs, or coil springs.

n Synonymy: words representing the same concept, although they may do it with different shades of meaning. Take the words 'ball,' 'sphere,' and 'orb,' or 'scuba diving' versus 'skin diving.' If we look for scuba diving, but the term used is skin diving, we will miss materials we might otherwise find. Good metadata should draw these materials together, despite their use of different synonyms.

- Ambiguity: If we return to our example of springs, we can see what differentiates these meanings is their context. It is unlikely an article on coil springs will also discuss water quality. The other words used in the article, and the processes described, will be entirely different. A search engine must understand the meaning, not just be able to match the spelling of a word, if it is going to differentiate between different meanings of the same word. A possible solution to this difficulty lies in the recent development of the notion of Application Profiles. Application Profiles provide a model for the way in which metadata might be aggregated in 'packages' in order to combine different element sets relating to one resource. This is a way of making sense of the differing relationship that implementers and namespace managers have towards metadata schema, and the different ways they use and develop schema. Students should investigate these developments within their community of interest.

Resources

See: Metadata: Cataloging by Any Other Name ... by Jessica Milstead and Susan Feldman ONLINE, January 1999

Available [on-line] http://www.onlineinc.com/onlinemag/OL1999/milstead1.html

See: Application profiles: mixing and matching metadata schemas Rachel Heery and Manjula Patel

Available [on-line] http://www.ariadne.ac.uk/issue25/app-profiles/

4.3.5.2 Fulltext, Edward Fox

When all of the text of an ETD is available for searching, a digital library system is said to support fulltext searching. Users can submit queries that call for documents that have particular phrases, words, categories, or word stems appearing anywhere in the text (e.g., in the middle of a paragraph, or as part of the caption of a figure).

In fulltext searching it often is possible to specify that query terms appear in the same paragraph, same sentence, or within n words of each other. These refinements may work together with support for exact or approximate phrase and/or word matching.

For fulltext searching to work, the entire document must be analyzed, and used to build an index that will speed up searching. This may require a good deal of space for the index, often around 30% of the size of the texts themselves. Further, such searching may lead to decreased precision, since a document may be located that only makes casual mention of a topic, when the bulk of the document is about other topics. On the other hand, fulltext searching may improve recall, since works can be found that are not classified to be about a certain topic. Further, fulltext searching often yields passages in a document, so one can find a possibly relevant paragraph, rather than just a pointer to a document that then must be scanned to ascertain relevance.

4.3.5.3. SGML/XML Overview, Susanne Dobratz

SGML/XML is a multiple-targeted strategy (see [9]). "It allows librarians to ensure longevity of digital dissertations. Modern hardware and redundancy can keep all the bits of an electronic thesis or dissertation (ETD) intact. But electronic archives must be modernized continually as new document formats become popular." As librarians always tend to think in decades, document formats like TIFF, Postscript or PDF do not meet their requirements. If PDF is replaced by another de facto (industry, not ISO-like) standard, preserving digital documents would mean converting thousends of documents. XML can help overcome those difficulties. "XML is the new ASCII" [**Error! Reference source not found.**]." If an electronic document is to be of ‘archival quality, it should be liberated from the page metaphor." (See [**Error! Reference source not found.**].)

A second reason for using SGML/XML is that it ensures reusability of documents by preserving raw data and content-based structuring of information pieces. Preserving data for statistics and formulas in mathematics and chemistry could allow reasearchers to reuse and repeat simulations, calculations and experiments, deriving the needed data directly from an archive.

Third, using structured information allows the reuse of the same information or documents in different contexts, i.e., the same digital dissertation can be used to produce an online or print version, and to produce additional information products, like monthly proceedings containing the abstracts of all dissertations produced within the university during the last month, or a citation index. Additionally, the dissertation can be displaysd for different media, so a Braille reader or an automatic voice synthesizer could be used as a back-end machine.

Another reason for using markup for encoding documents is that a wider, more qualified retrieval could be provided to the the users of an archive. As university libraries are more and more challenged by the problem of handling, converting, archiving and providing electronic publications, one of the major tasks is providing a new quality for retrieval within the user interface. Using an SGML/XML-based publishing concept enables a new quality in the distribution of scientific contents via specific information and knowledge management.

What does SGML/XML mean?

The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web. The current W3C Recommendations are XML 1.0, Feb '98, Namespaces, Jan '99, and Associating Stylesheets, Jun '99, and XSLT/XPath, Nov '99.(http://www.w3.org/XML) The development of XML started in 1996 and it is a W3C standard since February 1998, which may make you suspect that this is rather immature technology. But in fact the technology isn't very new.

Before XML there was the Standard Generalized Markup Language (SGML), developed in the early '80s, an ISO standard since 1986, and widely used for large documentation projects. And of course HTML, whose development started in 1990. The designers of XML simply took the best parts of SGML, guided by the experience with HTML, and produced something that is no less powerful than SGML, but vastly more regular and simpler to use. While SGML was mostly used for technical documentation and much less for other kinds of data, with XML it is the opposite.

"Structured data", such as mathematical or chemical formulas, spreadsheets, address books, configuration parameters, financial transactions, technical drawings, etc. are usually put on the Web using the output of layout programs as Postscript or PDF or by putting them into graphic formats like gif, jpeg, png, vrml, and so on. Programs that produce such data often also store it on disk, for which they can use either a binary format or a text format. So, if soemebody wants to look at the data, he usually needs the program that produced it. With XML those data could be stored in a text format, which allows the user reading the file without having the original program. XML is a set of rules, guidelines, conventions, whatever you want to call them, for designing text formats for such data, in a way that produces files that are easy to generate and read (by a computer).

The eXtensible Markup Language (**XML**) is a markup or structuring language for documents, a so-called metalanguage, that defines rules for the structural markup of documents independently from any output media. XML is a "reduced" version of the Structured Generalized Markup Language (**SGML**), which has been an ISO-certified standard since 1986. In the field of internet publishing, it never achieved wide success due to the complexity of the standard and the high cost of the tools. It prevailed only in certain areas, such as technical documentation in large enterprizes (Boeing, patent information). The main philosophy of SGML and XML is the strict separation of content, structure and layout of documents. Most ETD projects use either the SGML standard (ISO 8879 with Korregendum K vom 4.12.1997) or the definition of the World Wide Web Consortium (W3C) XML 1.0 (10.02.1998, revised 6.10.2000). The crux of all those projects was always the document type definition (DTD).

4.3.5.3 SGML/XML, Tejas Patel and Edward A. Fox

SGML (Standard Generalized Markup Language) and **XML** (eXtensible Markup Language) are markup languages, which use tags (“<” and “>”) with names of labels inside around the sections of the documents that are thus marked or bracketed. Document Type Definition (DTD) specifies the grammar or structure for a type or a class of documents. SGML requires a DTD while XML employs DTD optionally. But given current trends it seems that XML is most likely to be used due to the following reasons.

n XML is a method for putting structured data in a text file for "structured data" think of such things as spreadsheets, address books, configuration parameters, financial transactions, technical drawings, etc. Programs that produce such data often also store it on disk, for which they can use either a binary format or a text format. The latter allows you, if necessary, to look at the data without the program that produced it. XML is a set of rules, guidelines, conventions, whatever you want to call them, for designing text formats for such data, in a way that produces files that are easy to generate and read (by a computer), that are unambiguous, and that avoid common pitfalls, such as lack of extensibility, lack of support for internationalization/localization, and platform-dependency.

n **XML looks a bit like HTML but isn't HTML**
Like HTML, XML makes use of *tags* and *attributes* (of the form name="value"), but while HTML specifies what each tag and attribute means (and often how the text between them will look in a browser), XML uses the tags only to delimit pieces of data, and leaves the interpretation of the data completely to the application that reads it. In other words, if you see "<p>" in an XML file, don't assume it is a paragraph. Depending on the context, it may be a price, a parameter, a person. In short it allows you to develop

your own mark up language specific to a particular domain.

n **XML documents can be preserved for a long time.**
XML is, at a basic level an incredibly simple data format. It can write in 100 percent pure ASCII text as well as in a few other well-defined formats. ASCII text is reasonably resistant to corruption. Also XML is very well documented. The W3C’s XML 1.0 specification tells us exactly how to read XML data.

n **XML is license-free, platform-independent and well-supported.**
By choosing XML as the basis for some project, you buy into a large and growing community of tools (one of which may already do what you need!) and engineers experienced in the technology. Opting for XML is a bit like choosing SQL for databases: you still have to build your own database and your own programs/procedures that manipulate it, but there are many tools available and many people that can help you. And since XML, as a W3C technology, is license-free, you can build your own software around it without paying anybody anything. The large and growing support means that you are also not tied to a single vendor. XML isn't always the best solution, but it is always worth considering.

n **XML is a family of technologies.**
There is XML 1.0, the specification that defines what "tags" and "attributes" are, but around XML 1.0, there is a growing set of optional modules that provide sets of tags & attributes, or guidelines for specific tasks. There is, e.g., *Xlink* which describes a standard way to add hyperlinks to an XML file. *XPointer* & *XFragments* are syntaxes for pointing to parts of an XML document. (An Xpointer is a bit like a URL, but instead of pointing to documents on the Web, it points to pieces of data inside an XML file.) CSS, the style sheet language, is applicable to XML as it is to HTML. *XSL* is the advanced language for expressing style sheets. The *DOM* is a standard set of function calls for manipulating XML (and HTML) files from a programming language. *XML Namespaces* is a specification that describes how you can associate a URL with every single tag and attribute in an XML document. What that URL is used for is up to the application that reads the URL, though. *XML Schemas* help developers to precisely define their own XML-based formats. There are several more modules and tools available or under development.

n **XML provides Structured and Integrated Data**
XML is ideal for large and complex data like ETD’s because data is structured. It not only lets you specify a vocabulary that defines the elements in the document; but it also allows you to specify relations between the elements.

XML can encode metadata about DTD’s.
Documents are often supplemented with metadata (that is data about data). If such metadata were included inside an ETD then it would make ETD self-describing. XML can encode such metadata. However on the downside XML comes with its own bag of discomforts.

1) Conversion from word processing forms to XML requires more planning is advance, different tools and broader learning about processing concepts than it is required for PDF.

2) There are many fewer people knowledgeable about these matters and tools that support this conversion are less mature and expensive. Also process of converting may be complicated, difficult and time consuming.

3) Writing directly in XML by using XML authoring tools requires some prior knowledge of XML.

4) Also XML is very strict regarding the naming and ordering of tags. It is also case sensitive illustrating the relative effort required by students to prepare ETD’s in this form.

Process of Creating an XML document

XML documents have four-stage life cycle.

XML documents are mostly created using an editor. It may be a basic text editor like notepad. or .vi. editor. We may even use WYSIWYG editors. The XML parser reads the document and converts it into a tree of elements. The parser passes the tree to the browser that displays it. It is important to know that all this processes are independent and decoupled from each other.

Putting XML to work for ETD’s

Before we jump into the XML details for ETD.s we should make certain things clear, since we would be using them on a regular basis now onwards.

DTD (Document Type Definition):

An XML document primarily consists of a strictly nested hierarchy of elements with a single root. Elements can contain character data, child elements, or a mixture of both. The *structure* of the XML document is described in the DTD. There are different kinds of documents like letter, poem, book, thesis, etc. Each of the documents has its own structure. This specific structure is defined in a separate document called Document Type Definition (DTD).

DTD used is based on XML and it covers most of the basic HTML formatting tags and also some specific tags from the Dublin core metadata. A DTD has been developed for ETD. The developed DTD is too generic. If someone wants to use mathematical equation or incorporate some chemical equation, it won't be sufficient. For that we can incorporate **MathML** (Mathematical Markup Language) and/or **CML** (Chemical Markup Language). There are defined DTDs for these languages that we also have to use for our documents. But research of incorporating more that one DTD for different parts of the documents is still going on.

CSS (Cascaded Style Sheets):

CSS is a flexible, cross-platform, standards-based language used to suggest stylistic or presentational features applied throughout entire websites or web pages. In their most elegant forms, CSS are specified in a separate file and called from within the XML or HTML header area when documents loads into the CSS-enabled browser. Users can always turn off the author's styles and apply their own or mix their important styles with the authors. This points to the "cascading" aspect of CSS.

CSS is based on rules and style sheets. A **rule** is a statement about one stylistic aspect of one or more elements. A **style sheet** is one or more rules that apply to a markup document.

An example of a simple style sheet is a sheet that consists of one rule. In the following example, we add a color to all first-level headings (H1). Here's the line of code - the rule - that we add:

```
H1 {color: red}
```

XSL (the eXtensible Stylesheet Language):

XSL is a language for expressing stylesheets. It consists of two parts:

1. A language for transforming XML documents, and
2. An XML vocabulary for specifying formatting semantics.

If you don't understand the meaning of this, think of XSL as a language that can transform XML into HTML, a language that can filter and sort XML data, a language that can address parts of an XML document, a language that can format XML data based on the data value, like displaying negative numbers in red, and a language that can output XML data to different devices, like screen, paper or voice. XSL is developed by the W3C XSL Working Group whose charter is to develop the next version of XSL.

Because XML does not use predefined tags (we can use any tags we want), the meanings of these tags are not understood: <table> could mean an HTML table or maybe a piece of furniture. Because of the nature of XML, the browser does not know how to display an XML document.

In order to display XML documents, it is necessary to have a mechanism to describe how the document should be displayed. One of these mechanisms is CSS as discussed above, but XSL is the preferred style sheet language of XML, and XSL is far more sophisticated and powerful than the CSS used by HTML.

XML Namespaces

The purpose of XML namespaces is to distinguish between duplicate element type and attribute names. Such duplication might occur, for example, in an XSLT stylesheet or in a document that contains element types and attributes from two different DTDs.

An XML namespace is a collection of element type and attribute names. The namespace is identified by a unique name, which is a URI. Thus, any element type or attribute name in an XML namespace can be uniquely identified by a two-part name: the name of its XML namespace and its local name. This two part naming system is the only function of XML namespaces.

- XML namespaces are declared with an xmlns attribute, which can associate a prefix with the namespace. The declaration is in scope for the element containing the attribute and all its descendants. For example code below declares two XML namespaces. Their scope is the A and B elements:

```
<A xmlns:foo="http://www.foo.org/" xmlns="http://www.bar.org/">

      <B>abcd</B>

</A>
```

n   If an XML namespace declaration contains a prefix, you refer to element type and attribute names in that namespace with the prefix. For example code below declare A and B in http://www.foo.org namespace, which is associated with the *foo* prefix:

```
<foo:A xmlns:foo="http://www.foo.org/">

<foo:B>abcd</foo:B>

</foo:A>
```

n   If an XML namespace declaration does not contain a prefix, the namespace is the default XML namespace and you refer to element type names in that namespace without a prefix. For example, code below is same as previous example but uses a default namespace instead of foo prefix:

```
<A xmlns="http://www.foo.org/">

      <B>abcd<B>

</A>
```

Glossary

attribute

XML structural construct. A name-value pair within a tagged element that modifies certain features of the element. For XML, all values must be enclosed in quotation marks.

cascading style sheets (CSS)

Formatting descriptions that provide augmented control over presentation and layout of HTML and XML elements. CSS can be used for describing the formatting behavior of simply structured XML documents, but does not provide a display structure that deviates from the structure of the source data.

CDATA section

XML structural construct. CDATA sections can be used to mark tags or reserved characters with quotation marks and thus prevent them from being interpreted. For this reason, the CDATA section is especially useful for escaping markup and script. The syntax for CDATA sections in XML is <![CDATA[...]]>.

character data

XML structural construct. The text content of an element or attribute. XML differentiates this plain text from markup.

character set

A mapping of a set of characters to their numeric values. For example, Unicode is a 16-bit character set capable of encoding all known characters; it is used as a worldwide character-encoding standard.

component

An object that encapsulates both data and code, and provides a well-specified set of publicly available services.

data type

The type of content that an element contains: a number, a date, and so on. In XML, an author can specify an element's data type, for example, with a tokenized attribute type. Microsoft is working with the W3C to define a set of standard types that anyone can freely use.

document element

The top-level element of an XML document; only one top-level element is allowed. The document element is a child of the document root.

Document Object Model (DOM)

The standard maintained by the W3C that specifies how the content, structure, and appearance of Web documents can be updated programmatically with scripts or other programs. The proposed object model for XML matches the Document Object Model for HTML so that script writers can easily learn XML programming. The XML DOM will provide a simple means of reading and writing data to and from an XML tree structure.

document root

The top-level node of an XML document; its descendants branch out from it to form the XML tree for that document. The document root contains the document element and can also contain a set of processing instructions and comments.

document type declaration

XML structural construct. A production within an XML document that contains or points to markup declarations that provide a grammar for a class of documents. This grammar is known as a Document Type Definition. The document type declaration can point to an external subset (a special kind of external entity) containing markup declarations, or can contain the markup declarations directly in an internal subset, or both. The DTD for a document consists of both subsets taken together. The syntax of the document type declaration is <!DOCTYPE *content* >.

Document Type Definition (DTD)

The markup declarations that describe a grammar for a class of documents. The DTD is declared within the document type declaration production of the XML file. The markup declarations can be in an external subset (a special kind of external entity), in an internal subset directly within the XML file, or both. The DTD for a document consists of both subsets taken together.

Electronic Data Interchange (EDI)

An existing format used to exchange data and support transactions. EDI transactions can be conducted only between sites that have been specifically set up with compatible systems.

element

XML structural construct. An XML element consists of a start tag, and end tag, and the information between the tags, which is often referred to as the contents. Elements used in an XML file are described by a DTD or schema, either of which can provide a description of the structure of the data.

entity

XML structural construct. A character sequence or well-formed XML hierarchy associated with a name. The entity can be referred to by an entity reference to insert the entity's contents into the tree at that point. The function of an XML entity is similar to that of a macro definition. Entity declarations occur in the DTD.

entity reference

XML structural construct. Refers to the content of a named entity. The name is delimited by the ampersand and semicolon characters; for example, &bookname; and &#x3C;. It is used in much the same way as a macro.

Extensible Linking Language (XLL)

An XML vocabulary that provides links in XML similar to those in HTML but with more functionality. Linking could be multidirectional, and links could exist at the object level rather than just at a page level.

Extensible Markup Language (XML)

A subset of SGML that provides a uniform method for describing and exchanging structured data in an open, text-based format, and delivers this data by use of the standard HTTP protocol. At the time of this writing, XML 1.0 is a World Wide Web Consortium Recommendation, which means that it is in the final stage of the approval process.

Extensible Stylesheet Language (XSL)

A language used to transform XML-based data into HTML or other presentation formats, for display in a Web browser. Differs from cascading style sheets in that it can present information in an order different from that in which it was received. XSL will also be able to generate CSS along with HTML. XSL consists of two parts, a vocabulary for transformation and the XSL Formatting Objects.

ID

A special attribute type within the XML language. The ID attribute on the XML element provides a unique name, enabling links to that element using the IDREF attribute type. The value associated with the ID attribute must be unique within that XML document. IDs are currently declared with a DTD or schema.

markup

XML structural construct. Text in an XML document that does not represent character data: start tags, end tags, empty-element tags, entity references, character references, comments, CDATA section delimiters, DTDs, and processing instructions.

mixed content

XML structural construct. An element type has mixed content when elements of that type can contain character data, optionally interspersed with child elements. In this case, the types of the child elements can be constrained, but not their order or their number of occurrences.

namespace

A mechanism to resolve naming conflicts between elements in an XML document when each comes from a different vocabulary; it allows the commingling of like tag names from different namespaces. A namespace identifies an XML vocabulary defined within a URN. An attribute on an element, attribute, or entity reference associates a short name with the URN that defines the namespace; that short name is then used as a prefix to the element, attribute, or entity reference name to uniquely identify the namespace. Namespace references have scope. All child nodes beneath the node that specifies the namespace inherit that namespace. This allows nonqualified names to use the default namespace.

NDATA

The literal string "NDATA" is used as part of a notation declaration. See also notation.

notation

Usually refers to a data format, such as BMP. A notation identifies by name the format of unparsed entities, the format of elements that bear a notation attribute, or the application to which a processing instruction is addressed.

notation declaration

A notation declaration provides a name and an external identifier for a notation. The name is used in entity and attribute-list declarations and in attribute specifications. The external identifier is used for the notation, which can allow an XML processor or its client application to locate a helper application capable of processing data in the given notation.

processing instruction (PI)

XML structural construct. Instructions that are passed through to the application. The target is specified as part of the PI. The syntax for a PI is <?pi-name *content*?>.

Resource Definition Framework (RDF)

An object model similar in function to an application programming interface (API), RDF can be used by developers to access the logical meaning of designated content in XML documents.

root element

Sometimes this term is used to refer to the document element but this is misleading, since the top-level element and the document root are not the same. Because of this ambiguity, use of the term "root element" is discouraged.

schema

A formal specification of element names that indicates which elements are allowed in an XML document, and in which combinations. A schema is functionally equivalent to a DTD, but is written in XML; a schema also provides for extended functionality such as data typing, inheritance, and presentation rules.

Standard Generalized Markup Language (SGML)

The international standard for defining descriptions of structure and content of electronic documents. XML is a subset of SGML designed to deliver SGML-type information over the Web.

target

The application to which a processing instruction is directed. The target names beginning with "XML" and "xml" are reserved. The target appears as the first token in the PI. For example, in the XML declaration <?xml version="1.0"?>, the target is "xml".

text markup

Inserting tags into the middle of an element's text flow, to mark certain parts of the element with additional meta-information.

tokenized attribute type

Each attribute has an attribute type. Seven attribute types are characterized as tokenized: ID, IDREF, IDREFS, ENTITY, ENTITIES, NMTOKEN, and NMTOKENS.

Uniform Resource Identifier (URI)

The generic set of all names and addresses that refer to resources, including URLs and URNs. Defined in Berners-Lee, T., R. Fielding, and L. Masinter, Uniform Resource Identifiers (URI): Generic Syntax and Semantics. 1997. See updates to the W3C document RFC1738. The Layman-Bray proposal for namespaces makes every element name subordinate to a URI, which would ensure that element names are always unambiguous.

Uniform Resource Locator (URL)

The set of URI schemes that have explicit instructions on how to access the resource on the Internet.

Uniform Resource Name (URN)

A Uniform Resource Name identifies a persistent Internet resource.

valid XML

XML that conforms to the vocabulary specified in a DTD or schema.

W3C

World Wide Web Consortium

well-formed XML

XML that meets the requirements listed in the W3C Recommendation for XML 1.0: It contains one or more elements; it has a single document element, with any other elements properly nested under it; each of the parsed entities referenced directly or indirectly within the document is well-formed. A well-formed XML document does not necessarily include a DTD.

World Wide Web Consortium (W3C)

The international consortium founded in 1994 to develop standards for the Web. See

XLL

Extensible Linking Language

XML

Extensible Markup Language

XML declaration

The first line of an XML file can optionally contain the "xml" processing instruction, which is known as the XML declaration. The XML declaration can contain pseudoattributes to indicate the XML language version, the character set, and whether the document can be used as a standalone entity.

XML document

A data object that is well-formed, according to the XML recommendation, and that might (or might not) be valid. The XML document has a logical structure (composed of declarations, elements, comments, character references, and processing instructions) and a physical structure (composed of entities, starting with the root, or document entity).

XML parser

A generalized XML parser reads XML files and generates a hierarchically structured tree, then hands off data to viewers and other applications for processing. A validating XML parser also checks the XML syntax and reports errors.

4.3.5.4 Multimedia, Edward Fox

Searching through works with multimedia content requires extra support beyond what is usually provided (e.g., searching through metadata or fulltext). Content based image retrieval (CBIR), searching in files of spoken language, and searching in video collections are all supported with special methods and systems.

While most ETD collections today allow search for multimedia content based on descriptions of such content (i.e., metadata), in the future, as collections grow and have richer assortments of multimedia elements, it is likely that multimedia database and multimedia content search software will be more widely deployed. Indeed, Virginia Tech carried out some preliminary work in this regard with its ETD collection, using tools provided by IBM (e.g., QBIC, VideoCharger). Other vendors and systems exist, and can be used for searching local, regional, national, and international collections of ETDs.

4.3.6 Interfaces, Edward Fox

Information retrieval (IR) systems have been improving since the 1950s. One of the most important areas of advance in the 1990s and in the 21st century benefits from the rapid enhancement of human-computer interaction (HCI) that results from new types of interfaces, new methods of interaction, and new integration of those activities with IR methods. Information visualization, for example, allows faster analysis of the contents of a collection of ETDs, or of the result set from a search.

5. Training the trainers, Edward Fox

We believe that all universities should have ETD programs. Clearly that will be the case, when one considers the situation in the long term (e.g., 10 years). Why not join now, so that students over the next decade can benefit? Why not participate, so that the research of universities is more widely shared? Why not develop local infrastructure, so that students are properly prepared for the Information Age, save money, and save money for the university?

Training efforts should aim to make clear the benefits of working with ETDs. They should assuage concerns, and make obvious how changes can be made in a smooth transition from current practice.

5.1 Training the Trainers: Initiatives to support Electronic Theses and Dissertation projects in Latin America, Johann van Reenen

Abstract: This section focuses on the outreach work of the Ibero-American Science & Technology Education Consortium (ISTEC) and selected other organizations in developing EDT projects in Latin America. Training for librarians and for EDT trainers are described.

Many Latin American universities have digital library projects, some of which include electronic theses. However, standards and consistency may be lacking in local EDT initiatives. The Ibero-American Science & Technology Education Consortium (ISTEC) and its partners have been creating learning opportunities and instigating local projects in digital libraries and EDT's. This section describes ISTEC's outreach process and progress in regards to EDT projects for the science and technology libraries that are members of the organization.

Overview of ISTEC

ISTEC is a non-profit organization comprised of educational, research, and industrial institutions throughout the Americas and the Iberian Peninsula. The Consortium was established in September 1990 to foster scientific, engineering, and technology education, joint international research and development efforts among its members and to provide a cost-effective vehicle for the application of technology.

With start-up funding from the State of New Mexico and selected IT companies, the ISTEC board created four initiatives to address obstacles to IT developments and to encourage IT manpower development. These are:

1. The *ACE Initiative* champions continuing engineering and computer sciences education projects. The most important goals are to upgrade human resources and curriculum development through training and non-traditional exchange programs. The methodology involves on-site training, web-based education, video courses, satellite delivery, and "sandwich" graduate programs. The latter brings graduate students from Ibero-America together with experts from ISTEC member organizations to ensure excellence.

2. The *Research and Development (R&D) Initiative* focuses on the development and enhancement of laboratory infrastructure at member organizations. The major goal is the design and installation of modular, flexible, and expandable laboratory facilities for education, training, and R&D with links to the private sector.

3. The *Los Libertadores Initiative* champions networks of excellence in the region.The main goal is to network Centers of Excellence equipped with the latest telecommunications and computer technology to provide real-time access to a world-wide system of expertise and knowledge. This requires partnerships among industries and governments to create an Ibero-American academic

and R&D Internet backbone.

4. The *Library Linkages Initiative (LibLINK)* is ISTEC’s information creation, management and sharing project. Below is a description of LibLINK efforts in developing digital library projects in Latin America, especially in the area of EDT’s.

Overview of the Library Linkages (LibLINK) *project of ISTEC*

The major goal of *LibLINK* is to design and implement innovative, international Science and Technology (S&T) information-sharing services. The annual compound growth rate of the Rapid Document Delivery (RDD) project has been hovering around 200% since 1995. Over 27 libraries in 19 countries are connected in real-time and documents are provided using the *Ariel*® software. The RDD project, although the most popular service, is a foundation for the more important digital library initiatives which were started in 1998. The projects within *LibLINK* can be categorized as follows:

- Connecting libraries for information transfer. This is accomplished through opening S&T library collections - especially Latin American collections - for scholars through regional networks created to compliment the *LiBLINK* document delivery services. Currently these include *LigDoc* in Brazil, *PrEBi* in Argentina, *REBIDIMEX* in Mexico, and most recently, a cooperative group of libraries in Colombia.
- Training librarians and researchers in digital library concepts.
- Working with the Networked Dissertation/Thesis Library (NDTL) initiative at Virginia Tech to expand the concept in Latin America. The LibLink initiative seeks to promote easy access to scientific information in the region, especially to thesis and dissertations of master's and doctoral candidates as this is our member organizations’ most important intellectual property.
- Advancing and piloting new types of scholarly communication by

actively supporting new publishing efforts such as the NDLTD and the Open Archives initiatives.

*LibLIN*K volunteers plan and carry out workshops and mini-conferences to facilitate the above. Funding generally come from grants provided by organizations such as the US National Science Foundation (NSF) and other national science councils such as CONACyT in Mexico, and regional organizations such as the OAS and UNESCO.

LibLINK *and EDT's in Latin America*

We have refined a process for involving librarians and computer scientists in digital library projects that has proven successful. The principles on which ISTEC and *LibLINK* base their outreach efforts are:

- to establish the capacity of libraries and library staff for participating in digital projects.
- Site visits Participation in regional or local IT and computer science workshops to identify computer scientists working on digital library projects or components thereof. We are especially interested in initiatives created in isolation from each other and from their local libraries.
- In this way digital library initiatives and researchers are identified and a DL group can be established from the above findings that consist of librarians and computer scientists/engineers. Outcome: *Critical mass of computer scientists and librarians linked to each other and to ISTEC.*
- With this groundwork done, we plan and find funding sources for a Digital Library Workshop that generally have two major aims:
- To share information about current DL initiatives in that specific country or region
- To provide training in EDT's as the preferred first DL project
- In some cases, this is also an opportunity to create strategic plans for coordinated national projects

This process has resulted in the following EDT and DL workshops:

- A NSF/CONACyT digital library workshop that included NDLTD training by Ed Fox, in Albuquerque, NM. Funding was obtained from various sources, mainly the National Science Foundation, CONACyT (the Mexican Science Council) and the Organization of American States. July 7-9, 1999
- A successful DL conference in Costa Rica for Central American countries (Seminario / Taller Subregional sobre Bibliotecas Digitales). A full day EDT workshop was delivered by Ed Fox, followed by a day of digital leadership training and planning for local EDT projects. Funding was provided by the Organization of American States(OAS), the US Ambassador to Costa Rica and ISTEC. San Jose, Costa Rica, November 1999 (more detail below).
- 1st Course for the Training for Project Directors for Electronic Thesis and Dissertation Projects. The course was organized by UNESCO, the VII CYTED, Universidad de los Andes, and ISTEC's Library Linkages Initiative and held in conjunction with the *VII Jornadas Iberoamericanas de Informatica*, Cartagena de Indias, Colombia, from August 30 to September 1, 2000. This is an example of how EDT training can be piggy-backed on a significant regional event that is synergistic and that provide more value for the money spent to attend.
- A 2nd Training Course for Directors of EDT Projects was funded by UNESCO (Montevideo), the Asociación de Universidades Grupo Montevideo (AUGM), and ISTEC. This "Train the Trainer" course was held in Montevideo, December 7-9, 2000. The main goals for this series of courses are to create a group of specialists responsible for the dissemination and management of electronic dissertations and theses. The trainer for both these sessions was Ana Pavani from the Pontificia

Universidade Catolica do Rio de Janeiro.

- REBIDIMEX is the Mexican operations for ISTEC Library Linkages. This group has had a number of meetings to develop coordinated digital library and EDT projects in Mexico. The digital theses project at the library of the Universidad de las Américas-Puebla is an excellent example, http://biblio.udlap.mx/tesis/, and forms the basis for a national Mexican EDT project.

Case Study:

The *Primer Seminario-Taller Subregional sobre Bibliotecas Digitales*, sponsored by the OAS and ISTEC at the Universidad de Costa Rica, San Jose, Costa Rica, mentioned above, provides a good case study of EDT outreach events. Each participating Central American country was asked to identify universities with sufficient technological infrastructure to support a digital library/ EDT project. Then each organization was funded to send representatives from each of their computer systems and library groups. The agenda focussed on providing one whole day of basic training by Ed Fox (a co-author of this *Guide*) in digital library and EDT concepts, followed by another day of leadership training for digital environments and a planning session.

During this portion the groups identified a project that all could participate in. They chose the digitization of their organization’s theses and dissertations and making it available through the Open Archives system using the processes developed by Virginia Tech and others described in other sections of the *Guide*. Regional working groups were assigned. The most important outcomes, however, were the creation of a network of librarians and computer scientists that understand EDT technological, operational and political issues and that now have contacts for joint projects in the region.

What next?

The model of creating synergism and connections between librarians and computer scientists and focusing their energies on basic digital library/EDT projects will continue to be replicated in other parts of Latin America. Ana Pavani (a section author of the *Guide)* continues to deliver EDT courses with the help of ISTEC, such as one in Pernambucu, Brasil, in the spring of 2001. Members from ISTEC's regional and executive offices regularly speak at conferences regarding EDT's and are available to help arrange EDT events. Organizing training events and developing joint funding arrangements takes a lot of time, expertise, local contacts, and effort, but is critical for creating opportunities in under-served countries.

ISTEC and its partner organizations, the OAS, International Development Bank, UNESCO, etc, continue to work together to offer regional digital library workshops in Latin America. UNESCO is formulating an international strategy for creating and disseminating electronic theses and dissertations that will support Latin American outreach efforts (UNESCO, 1999). As well, we are assisting governments to draft suitable policies to improve access to information, especially in making their universities' intellectual property (theses and dissertations) widely available to publicize the universities' research strengths. ISTEC is also sponsoring the Spanish translation of the *Guide* and will disseminate it through the ISTEC Science & Technology Portal.

Other Latin American Projects

The *Latin American Network Information Center (LANIC)* at the University of Texas at Austin is the most comprehensive resource for academic and economic information on Latin America (http://lanic.utexas.edu/), but not specifically for EDT projects. Some beginning and mature EDT projects can be found at:

n The Library System of the Universidad Católica de Valparaíso (Chile). http://biblioteca.ucv.cl/tesis_digitales/

n The Digital Technology Research and Development Center (CITEDI) of the Instituto Politecnico Nacional in Tijuana, México. http://www.citedi.net/docs/tesis.htm

n The UNESP site at the Universidade Estadual Paulista in Brazil. http://www.cgb.unesp.br/e-theses/

n The digital theses project at the library of the Universidad de las Américas-Puebla, Mexico. http://biblio.udlap.mx/tesis/

n University of Antioquia (Medellin, Colombia)

n University of Sao Paulo (Brazil). http://www.usp.teses.br/

n In Chile the Universidad de Chile has been developing Cyberthesis, since November 1999. This is an electronic theses production project with support from Unesco and the cooperation of the Universit de Lyon and the Universit de Montreal. The Information Service and Library System, SISIB, is coordinating the Electronic Theses and Dissertations Project (Cybertesis), applying the production process developed by the Universite de Montreal, based on the conversion of texts to SGML/XML. In 2002, the Universidad de Chile will organize training workshops in the production of ETD (structured documents) for other Chilean universities.

Associations

The Transborder Library Forum/FORO Transfronterizo de Bibliotecas share many of the aims of ISTEC's *Library Linkages* initiative. Their meetings have been held annually since 1991 to work on ways to improve communications relating to border issues and to foster professional networking among librarians from Mexico and the United States. Recently, Canadians and representatives from Latin American libraries also interested in NAFTA and border issues began to have a presence. At the 10th Transborder Library FORO held in Albuquerque, New Mexico in March 23-25, 2000, attendance of several representatives from Latin American libraries were sponsored and ISTEC's LibLink project provided a training session and talks about the REBIDIMEX initiative in Mexico.

The Association of Latin American and Caribbean Academic Libraries (Bibliotecas Universitarias da America Latina e do Caribe)) sponsored a LibLINK workshop that included EDT discussions and talks at the annual meeting in Florianopolis, Brazil, in April 2000. ISTEC continues to have a presence at their conferences. And are involved with the Brazilian EDT initiatives.

The impact of Bandwidth and Infrastructure issues on EDT outreach in Latin America

Bandwidth and IT infrastructure are important factors for digital EDT project developers in Latin America. IT policy development is another. More aggressive action is needed in both the governmental and industrial sectors. ISTEC emphasize these issues at the "IT Challenge" conferences by bringing industrial and academic members together with regional decision makers, such as ministers of education and technology and representatives from national science councils.

The first high-performance Internet link between North and South America for research and education was inaugurated in Santiago, Chile on September 12, 2000. Chili and the USA connected their respective high performance networks, REUNA* and Internet2, enabling collaboration among researchers and educators at universities in the two countries. Such high-performance network links are critical to ensure the bandwidth required for future format-enriched EDT projects. ISTEC and its partners are strongly committed to advance this cause.

* Reuna is a collaboration between the National Universities of Chili that introduced the Internet in Chile in 1992. Reuna´s high-speed network, REUNA2, is an ATM network of 155 Mbps. across the country. The National University Network is a non-profit consortium of 19 leading Chilean universities plus the National Commission for Science and

Technology. Its mission is the creation and development of networks and services in IT aimed at supporting participation in the Information Society.

Conclusion:

The Library Linkage initiative of ISTEC has found a methodology for encouraging and supporting EDT developments in Latin America that has proven successful. The most important step is to identify local players in digital library initiatives in both libraries and computer science and computer engineering departments. The next step then brings these players together at events that provide opportunities for training, information sharing and national/regional EDT project planning. ISTEC monitors subsequent developments and provide support to keep projects going as appropriate. The most important strategic outcome we are aiming for is to create an open archive structure that will provide access to all science and technology theses and dissertations of member organizations through the ISTEC Portal. We believe that this will be a rich source of innovation, manpower identification and development, and an opportunity to highlight the intellectual property of our member universities.

REFERENCES:

UNESCO (1999). Workshop on an international project of electronic dissemination of thesis and dissertations, UNESCO, Paris, September 27-28, 1999 http://www.unesco.org/webworld/etd/

NOTES:

Ibero-American Science & Technology Education Consortium. For more information see www.istec.org

REUNA2. For more information see: http://www.reuna.cl

5.2 Tool kits for trainers, Luc Grondin

This section of the Guide outlines the approaches to student training used by certain universities. For the most part, links are made to sites where the tools themselves can be consulted and eventually used or adapted for local use.

5.2.1 Université de Montréal, Luc Grondin

The training offered to doctoral students registered at the Université de Montréal, entitled "Helpful tools for writing a thesis" ("*Outils d'aide à la rédaction d'une thèse*"). is part of that institution's programme for electronically distributing theses. The training seeks to respond to two sorts of needs: 1- the needs related to the processing of theses (greatly aided by the adequate use of a normalized document template); 2- the needs of the doctoral students (who want to increase their capabilities in using writing tools, and, in the process, their level of productivity and the quality of their production).

The implementation of a programme of electronic thesis distribution affects the entire institution. At the Université de Montréal, three units are actively involved: the Faculty of Graduate Studies, the Library , and the Information technology services (*Direction générale des technologies de l'information et de la communication* -- DGTIC, which holds the mandate to process and distribute theses in electronic formats). The training of doctoral students is planned and provided by individuals drawn from these three units. In terms of the sessions held in September 2001, four people acted as trainers (one from Graduate Studies, one from the Libraries and two from the DGTIC).

The content of the training is as follows:

1. Welcome and general presentation of the Programme for the electronic publication and distribution of theses (DGTIC)

2. Presentation of the Université de Montréal's *Style Guide* for masters and doctoral theses (Graduate Studies).

3. On-campus services offered to thesis-writing students: equipment rentals, self-service digitizer, digital camera, etc. (DGTIC).

4. Presentation of the capabilities of the EndNote software for managing bibliographic referencing (Library)

5. Practical exercises with the Word document template to be used by the students.

Points 1 to 4 in this general plan involved presentations given by trainers, while point five directly involved participants in concrete exercises, accompanied by several demonstrations. For these exercises, the students used a working text (a Word document containing the principal editorial elements of a thesis, but without a proper layout), a Word document template developed for theses, and a list of instructions.. The exercises essentially consisted of applying the template's styles to the working text, of producing and inserting an image captured on the screen into the text, and of automatically producing a table of contents and an index of tables.

As well, several documents were provided to participants at the sessions: the instructions, the document template (Word style sheets) as well as training guides and informative documents of interest to the students. All these documents are available on-line at www.theses.umontreal.ca.

The workshops are offered to all doctoral students, whether they are at the beginning of their research or near the end of their writing. To reach the largest possible number of students, we used a multi-dimensional communications plans: posters, advertising, messages to student associations, and letters from the Dean of Graduate studies to the heads of departments and research centres. The number of seats in the training laboratory is limited to 15 to 20 students per session, depending on the laboratory used. Students are invited to register in advance, either by telephoning one of the DGTIC's administrative assistants, or, in an autonomous manner, by using an interactive Web form managed using a CGI scripts. This latter form tells students how many participants have already registered for each session and the maximum number of seats available. When the maximum number of registrations is reached, the script no longer allows registrations. Nevertheless, an invitation to register on a "waiting list" allows students to signal their interest and to be rapidly informed when new workshops are held. This also allowed us to note the pertinence of the training: the available spaces were quickly filled and the waiting list allowed us to reach interested students for the next sessions.

5.3. Demonstrations, Explanations,

Gabriela Ortuzar

Many technological choices such as encoding formats, computing tools, dissemination tools, and copyrights, have an important impact on the ETD project's success. It is necessary an overview of the different aspects of producing, disseminating, archiving theses in electronic formats, implementation and management of ETD projects.

At the present time several universities have initiated individual or cooperative projects for publishing and distributing theses, with different strategies approach : NDLTD (Networked Digital Library of Theses and Dissertations) with 103 member universities, Cyberthèses (Université de Montrèal, Université de Lyon, Universidad de Chile, Université de Genève with about 10 member universities) , Digital Australian Theses Program (8 australian universities) and Dissertationen Online (5 german universities).

5.3.1 Initiatives & Projects, Gabriela Ortuzar

n Collectives:

n • NDLTD

n Networked Digital Library of Theses and Dissertations

n • Cyberthèses

n [Université de Montréal] [Université de Lyon] [Universidad de Chile] [Université de Genève]

n • ADT

n Australian Digital Theses Program

n • Dissertationen Online

n Humboldt University of Berlin, Duisburg University , Oldenburg University, Erlangen University, Göttingen University.

n • The Joint Electronic Theses & Dissertations Project

n [University of Toronto] [York University]

n

n Individual university projects:

n

n • Humboldt-Universität zu Berlin [Germany]

n Projekt Digitale Dissertationen der Humboldt-Universität zu Berlin

n • Virginia Polytechnic Institute and State University (Virginia Tech) [United States]

n Electronic Theses and Dissertations Initiative

n • North Carolina State University [United States]

n ETD Electronic Theses & Dissertations

n • West Virginia University [United States]

n Electronic Theses and Dissertations

n

n • Massachussets Institute of Technology [United States]

n MIT Theses Online

n • University of South Florida [United States]

n Electronic Theses and Dissertations

n • Université de Montréal [Canada]

n Cyberthèses

n • Université de Lyon [France]

n Cyberthèses

n
- Universidad de Chile [Chile]

n Cybertesis

n
- Université de Genève [Switzerland]

n Cyberdocuments

n

add for Collectives:

1.

<a href="http://elfikom.physik.uni-oldenburg.de/dissonline/PhysDis/dis_europe.html">PhysDis</a>

a large collection of Physics Theses of Universities across Europe

2.

<a href="http://www.iwi-iuk.org/dienste/TheO/">TheO</a>,

a collection of theses of different fields of 43 Universities in Germany,

in as much as the Theses do contain Metadata,

3.

<a href="http://MathNet.preprints.org/">MPRESS</a>, a large collection of European Mathematical Theses.

which contains as a subset

4.

<a href="http://mathdoc.ujf-grenoble.fr/Harvest/brokers/prepub/query.html#math-prepub">Index nationaux prépublications, thèses et habilitations </a>

a collection of theses in France in Mathematics

5.3.2 Guidelines and Tutorials for ETDs, Gabriela Ortuzar

The inclusion of courses and tutorials with online demonstrations can facilitate a better understanding of all the production and submission process. With a comprehensive guide for ETD creation, the students can learn how to create ETD in word processing software, how to convert to PDF format and the submission procedure.

- Requirements and Guidelines for the Preparation of Master's and Doctoral Theses Penn State University
- Thesis and Dissertation Guide North Carolina State University
- Informationen für Autoren Humboldt-Universität zu Berlin
- Ressources Cyberthèses [Université de Montréal] [Université de Lyon]
- ETD Workshop Outline North Carolina State University

5.3.2.1 Specific guidelines, Gabriela Ortuzar

This section includes a series of tools intended to support the work of production and submission of ETD.

Writing in word processing systems: ETD Formatting

- Humboldt University of Berlin Word-style (for German Word Versions only)

http://dochost.rz.hu-berlin.de/epdiss/vorlage.html

- Université de Lyon

http://www.univ-lyon2.fr/sentiers/edition/theses/ressources.html

http://www.univ-lyon2.fr/sentiers/edition/theses/cours/ModeEmploi.pdf

http://mirror-fr.cybertheses.org/ressources.html

- Université de Montréal
http://www.cybertheses.org/cybertheses/ressources.html#style

- University of South Florida http://www.lib.usf.edu/virtual/etd/format.html

- Virginia Tech http://etd.vt.edu/guidelines/index.html

ETD Submission

- Virginia Tech : http://scholar.lib.vt.edu/ETD-db/help/

- ADT : http://www.library.unsw.edu.au/thesis/adt-ADT/info/example.html

- University of South Florida: http://www.lib.usf.edu/virtual/etd/submit.html

- Humboldt-University Berlin:

Automated Upload-Tool: http://dochost.rz.hu-berlin.de/cgi/dokupload/dokupload.cgi

Explanation at http://dochost.rz.hu-berlin.de/epdiss/abgabe.html

- *in TeX or LaTeX*

- Humboldt University of Berlin: LaTeX-Richtlinien / LaTeX-Tool-Seite / FAQs

and http://dochost.rz.hu-berlin.de/epdiss/latex/latex.html

- Virginia Tech : http://etd.vt.edu/howto/slides/latex/latex.html

Prepare a PDF document

- Virginia Polytechnic Institute and State University (Virginia Tech)

Create PDFs from Word :

Create PDFs from WordPerfect for Mac :

Create PDFs from WordPerfect for Windows

- Humboldt-University Berlin

Section on PDF and PDF from LaTeX :

Writing SGML/XML

Sites with SGML/XML-Theses and Dissertations

- HelsinkiUniversity of Technology (http://www.hut.fi/Yksikot/Kirjasto/HUTpubl/)

- Humboldt-University Berlin (http://dochost.rz.hu-berlin.de/epdiss/)

- Université de Montreal (http://www.cybertheses.org/)

- Université de Lyon 2 http://www.univ-lyon2.fr/sentiers/edition/

- University of Iowa (http://www.uiowa.edu/~gradcoll/etd.html)

- University of Michigan at Ann Arbor (http://www.umdl.umich.edu/um_diss_study.html)

- University of Oslo (http://www.digbib.uio.no/)

- Swedish University of Agricultural Sciences Uppsala (http://www.bib.slu.se/stp/pub2000/)

Virginia Polytechnic Institute and State University (Virginia Tech) (http://www.uiowa.edu/~gradcoll/etd.html)

- Universidad de Chile http://www.cybertesis.cl/

Metadata

- NDLTD : http://www.ndltd.org/standards/metadata/current.html

- Université de Montréal : http://www.pum.umontreal.ca/theses/metadata/1.0/

- Dissertationen Online : http://www.ub.uni-duisburg.de/dissonline/eindex.html

- ADT: http://www.library.unsw.edu.au/thesis/adt-ADT/info/metadata.html

Copyright

n NDLTD : http://www.ndltd.org/cpright/index.htm

n Virginia Tech: http://etd.vt.edu/howto/copyright.html

n In Germany theses are to obey the German Urheberrecht, which is different from the copyright in anglosaxon countries. see http://elfikom.physik.uni-oldenburg.de/dissonline/urheber.html, Urheberrecht theses in Germany

5.4. Creating an online database of problem solving solutions, Viviane Bouletreau

A quite important document must be at the users' disposal together with the various tools implemented for the electronic edition of scientific documents:

§ User’s guide for the preparation of the documents,

§ User’s guide for the conversion itself,

§ Pedagogical toolkit for the training of authors,...

This first edition, as complete as it may be, is nevertheless insufficient. It must be completed by the online diffusion of a database listing all the problems usually encountered and their solutions (FAQ) eventually completed with links to specific documents with examples.

Most of the time, such a database cannot be a priori completely defined and will have to be enriched as the number of users grows and their mastery of the various tools increases.

The collection of the questions and problems are done very efficiently thanks to the creation of discussion lists or forums opened to the users. It is recommendable to group the frequent questions or problems by subject, to offer an easier and friendly consultation.

Then, two methods considered for the construction of the database of solutions:

§ Synthesis, based on the discussion list, prepared by some “experts” used to supply the database with questions and answers. The information available for the users is validated.

§ The content of the forum indexed directly; the information is not validated, and it is thus the users themselves who will have to make a sort in the list of answers they get for their request. If this second method is lighter in terms of administration and management, its efficiency depends on the quality of the posted messages (questions and answers): precise and clear subject, good presentation of the problem, argumentation of the answer, etc.

Whichever the adopted solution may be, its implementation may rest on usual free software classical list and database servers.

The online diffusion of the so constructed FAQ must still have the name and e-mail of a member of the technical staff or a visible link to the discussion list in order to answer the questions not taken into account in the database.

As an example, CyberThèses proposes a forum construct with a MySql database and interfaced with Php, and will develop a second database from validated information.

5.5 Help develop a broad local team,

Australian Digital Theses Program

The core of the ADT Program was developed at The University of New South Wales Library (UNSW Library) as the lead institution. The team at UNSW Library included the overall coordinator & designer; technical manager & programmer; metadata consultant as well as the web coordinator & designer. This team developed the model from the conceptual to reality. The most important thing was not to lose focus, to keep the model as close to the original project description proposal as possible and to not overly complicate processes - in fact keep them as simple as possible. It was seen as critical to develop a workable model for the project partners to test and further refine.

During the development and testing process, all 7 original partners were consulted and had input at all stages. Two workshops for all 7 members were held approximately 12 months apart. These were used to discuss all aspects of the ADT model as well as to agree on the standards and protocols used. The agreed standards are at the core of the distributed model of the ADT Program. Without the involvement of an effective team to lead the process, and the effective input of the broader team, arriving at the desired outcome would have been very difficult.

The ADT Program is now effectively working across all 7 original member sites. Membership to the ADT has now been opened up to all Australian universities. It is anticipated that all Australian universities will become members and thus form a comprehensive national program. The benefits of a broad membership team are many: shared infrastructure, shared software development, shared metadata, shared documentation and training as well the shared satisfaction that comes with effective collaboration for the common good. The membership to the ADT may in time also include others in the region such as New Zealand and

5.6 Standards, cooperation and collaboration, Australian Digital Theses Program

While each institution will have differences as to the way in which its procedures are implemented and ETDs presented to satisfy local needs, for the potential of optimal global dissemination of ETDs to be achieved, adherence to basic standards is essential. These standards relate to document format and settings, filename protocols and metadata. Agreement on use of a small set of standard metadata elements can facilitate harvesting for creation of collaborative databases. While individual institutions can apply additional metadata including subject or format schemas as desired, if a minimum level of metadata is established, particularly if this metadata can be automatically generated, then any institution can access the document regardless of its resources and expertise. This provides an entry point for retrieval of an ETD from any institution without compromising the ability of other institutions to provide very rich metadata for their ETDs.

Cooperation and collaboration between institutions in the creation and dissemination of ETDs has a number of benefits. From the point of view of creating ETDs, the benefits include the utilization of software and procedures developed and tested by others, sharing generic training tools, sharing of expertise in problem solving and developmental work and the provision of mutual support. In the dissemination of information about research findings contained in ETDs, collaborative approaches can be particularly effective in small or developing countries where the total volume of ETDs may not be large. In these situations a national or regional approach can provide increased visibility, economies of scale and sharing of resources required to mount and maintain ETDs.

There are a number of models for cooperation and collaboration in creation, dissemination and preservation of ETDs. While the models can vary in detail, major elements are:

- Shared infrastructure

In this model, a central agency provides the infrastructure for publication, dissemination, maintenance and preservation of ETDs. The central agency provides the server and the network access to a central repository of ETDs. Other institutions cooperate with the central agency by supplying copies of theses or dissertations. These copies could be supplied in digital form, conforming to the basic standards, or the central agency can be responsible for ensuring a suitable digital version is created and made available. The central agency has the responsibility for maintaining archival versions. This model could operate at a supranational, national or regional level, or could be used by a group of universities with similar interests.

- ## Shared software development

Sharing of generic software, which is easily installed and maintained, can be a cost effective way of establishing an ETD program. This can be of particular benefit for institutions where staff with highly developed IT skills are in short supply. Collaborative development or modification of new versions of software can also be very cost-effective.

- ## Shared metadata

In this model, institutions publish and maintain digital theses on their own institutional servers but the metadata is harvested to produce a central database of details of the theses from the collaborating institutions. The metadata is linked to the full text of the ETD wherever it resides. The home institution retains the responsibility for the preservation of the ETD.

- ## Shared documentation and training tools

Sharing of detailed documentation on all aspects of operating an ETD program is a very cost effective method of collaboration. The development of generic procedures and training programs, which can be customized for local conditions, can facilitate the participation of institutions in an ETD program. Sharing of documentation is also likely to reinforce the use of standards, which will ensure the ETDs are readily discoverable

5.7.1 Developing Centres of Expertise,

Australian Digital Theses Program

An effective mechanism for providing support is to identify centres of expertise which are prepared to assist others in setting up their ETD system. These centres of expertise can be regional, national or within an international geographic or linguistic region. The benefits flowing from these centres of expertise include assistance with designing processes, assistance with implementing software, especially if a common software product is being used, assistance in use of metadata and continuing support as systems evolve. Another very important function of a centre of expertise is provision of a focus around which a group can develop which can offer a self-support function and an environment taking into account local or regional conditions for continuing discussion and development of ETD programs as the technology continues to evolve. This latter is essential if ETD programs are to prosper.

The centre of expertise concept also has the great advantage of reducing significantly the learning curve in starting up an ETD program and consequently the timeframe and cost of new programs.

5.7.1 Developing Centers of Expertise, Australian Digital Theses Program

An effective mechanism for providing support is to identify centers of expertise, which are prepared to assist others in setting up their ETD system. These centers of expertise can be regional, national or within an international geographic or linguistic region. The benefits flowing from these centers of expertise include assistance with designing processes, assistance with implementing software, especially if a common software product is used, assistance in use of metadata and continuing support as systems evolve. Another very important function of a center of expertise is provision of a focus around which a group can develop which can offer a self-support function and an environment taking into account local or regional conditions for continuing discussion and development of ETD programs as the technology continues to evolve. This latter is essential if ETD programs are to prosper.

The centre of expertise concept also has the great advantage of reducing significantly the learning curve in starting up an ETD program and consequently the timeframe and cost of new programs.

6.1 Expanding ETD initiatives, Edward Fox

ETD initiatives will spread in many ways. Word of mouth, from universities with successful programs, is one of the most effective means. Word of mouth, from graduates of universities with ETD programs, will spread the idea further.

Further, as the worldwide collection of ETDs grows, more and more researchers, including graduate students, will make use of the valuable content. As use expands, others will be convinced to include their works in the collection, so it approaches full coverage.

6.2 Transforming Graduate Education,

Joseph M. Moxley

Academics stand on a precipice separating our past, when genres of communication evolved slowly, and our future, when new genres emerge overnight. Our concepts of research, the authority of knowledge, and the shape of content are being radically challenged. We have difficulty imagining what dissertations or academic digital libraries will look like ten years from now. The shape of a dissertation is evolving from the first six-page, handwritten thesis at Yale University in 1860 into a form we cannot yet predict. Today's researchers and scholars are challenging the conventions of linear texts, one-inch margins, and texts written for extremely narrow audiences. They are integrating video, audio, animation, and graphics into their works. They are creating interactive elements, including real-time video, pivot tables, and online writing spaces.

The power of ETDs is rooted in access. No longer are theses and dissertations just an academic hurdle, a last step in the arduous process of graduate education. Instead, ETDs are a meaningful connection with significant readers. Collaborative author tools enable faculty to serve on dissertation committees at universities distant from their home campuses and using tools such as NetMeeting to mentor students from a distance. Rather than accepting their research and scholarship will be read only by a select few (i.e., their committees), graduate students can now expect many readers.

Predicting the future of academic scholarship is a little like predicting the stock market: both are volatile and unpredictable. Given this fact, however, it appears that there are a number of emerging trends that will affect our enterprise:

- Dissertations will matter more than they have in the past. Thanks to digital libraries, which increase access from one or two readers to sometimes more than 60,000, students and universities will pay greater attention to the quality of students' research and writing.

- Given this increased access, both students and universities may begin to pay greater attention to the quality of scholarly writing.

- Progressive universities will use their digital libraries of ETDs to market their programs, and universities will provide the resources students need to write multimedia research.

- Multimedia documents will transform author-reader relations. Authors will interact synchronously with readers, create different reading paths for different readers, and use visuals, animation, and pivot tables.

- Students will increasingly search the worldwide digital libraries of ETDs, resulting in research that is more collaborative and more current.

- Across disciplines, students will provide links that clarify the significance, methodology, and findings of their work to a broader range of readers, including lay audiences, thereby helping the general public better understand the value of academic scholarship. As an example, students in the social sciences can incorporate video of cultures and primary subjects; they can create polyvocal case studies and ethnographies—that is, studies with alternative interpretations.

- Faculty members will work more collaboratively with students, resulting in more complete bibliographies and saved time.

6.3. Managing technology changes, Simon Pockley

The ease of distributed access accompanying the digital revolution has a darker side. Technological obsolescence and ephemeral formats have left little firm ground upon which to build the infrastructure necessary for the effective management and preservation of digital resources. An infrastructure supporting long-term access needs to be able to accommodate the continuous evolution of the technology as well as continuous streams of data. This means that a durable record of such work now includes a flexible and adaptable approach to maintaining live access.

Devices and processes used to record, store and retrieve digital information now have a life cycle of between 2 - 5 years—far less than some of the most fragile physical materials they are seeking to preserve. The practice known as 'refreshing' digital information by copying it onto new media is particularly vulnerable to problems of `backward compatibility' and `interoperability'. Software capable of emulating obsolete systems currently relies on the goodwill and enthusiasm of talented programmers but has little basis in economic reality.

The foundation of effective management of an ETD as digital resource occurs at the point of creation. How a digital resource is created and the form and format it is created in will determine how it can be managed, used, preserved and re-used at some future date. Authors/creators and migrators of all digital resources need to be aware of their importance to the life span or term of access of a digital resource.

Simple but practical strategies that might increase the life span of an ETD are as follows:

- do not assume any stability in hardware or software and try to record all the dependencies that support your ETD.

- Use non-proprietary formats such as (ASCII, Unicode or plain text) as much as possible (e.g. Notepad or Simple Text)

- Use widely available formats such as Word or PDF that can be easily converted into plain text

- Use widely available non-proprietary image formats such as JPEG or PNG rather than GIF (proprietary format)

- if you must use custom or proprietary software, make such you keep a record of the version number (e.g. Real Media 7) and all the hardware dependencies (e.g. MAC system 7).

This information can be stored within the metadata associated with your ETD. For example, you could use the Dublin Core metadata element, 'DC.Relation.Requires'. As technologies evolve, the use of preservation metadata to record this kind of information is increasing. Check with your library to see what standard

6.4 Interoperability, Susanne Dobratz , Guylaine Beaudry

At our various locations, document servers for electronic theses and dissertations have been set up independently from each other. The aim is to build a network of sites, which would allow for worldwide retrieval within a heterogeneous knowledge base, independently from the physical location of the data provided. The users should not have to navigate and search the various servers separately. They should be provided with one retrieval interface that can link to all the different nodes of the network of ETDs sites. This is the horizontal level on which retrieval could take place.

Another level, which we call the vertical level of the information portal, would be the one, which configures the retrieval interface in a manner allowing the user to retrieve only the relevant and desired information rather than receiving all of the information that can be possibly provided. We wish to avoid the "Altavista effect" of information overload. The user should be able to search within specific subjects and for specific information structures. For example, they should be able to undertake searches just within the author field or the title field, for certain keywords or institutions or just within the abstracts field. A highly sophisticated retrieval facility would allow a worldwide search within certain internal document structures, such as the bibliography.

For the scientific use of theses in the humanities and social sciences, as well as in the natural and technical sciences, it is necessary to offer not only bibliographical metadata and full text but also structural information for retrieval purposes, such as:

- the table of contents;
- captions of tables and graphs;
- special index terms such as name or person indexes or location indexes etc.);

- references (links) to external sources (printed resources as well as Web sources);
- the bibliography;
- references or footnotes within the work;
- definitions;
- mathematical / chemical formulas;
- theses / hypotheses

These structural metadata are an integral part of the document and have to be defined by the author. At present, this predominantly takes place while formatting the text (e.g. headings, footnotes etc.). In order to also use these structural data for retrieval, they must be tagged as such by the author, by using either a structured language like LaTeX, or "style sheets" as with WinWord.

What could be the lowest common denominator for interoperability?

The first step within the above mentioned development is to reach agreement on a common metadata set for theses and dissertations and to formulate guidelines on how to use it for ETD projects. See http://www.ndltd.org/standards/metadata/ for the Dublin Core metadata set proposed by the NDLTD. Those guidelines could be supported by additional free software tools, which would allow library staff to create the necessary metadata set without needing a technical knowledge of the actual encoding in HTML or XML/RDF. Such a "metamaker" has been developed for the German ETD projects and could be translated into English, French, Spanish and Portuguese. MySQL or other free software can be used as the underlying database system.

The Open Archives Specification: A chance for metadata interoperability

During the last 2 years one initiative has effected the discussions about interoperability in digital libraries and the digital lbrary community enormously.

The development of

- a protocol, that can easily be implemented at archive servers and

- a metadata set based upon the Dublin Core metadata set

Allow archives, like ETD servers, preprint archives as well as museums and other institutions to provide their local catalogues to a worldwide community without having to implement specialised and complicated interfaces.

So the Open Archives Framework (see http://www.openarchives.org) allow an interoperability f heterogenious and distributed ETD archives and servers in a very low interoperability level.

For the ETD initiatives and projects the OAI compliance has to be seen as chance to connect ETD servers worlwide.

6.5 A vision for the future, Australian Digital Theses Program

Theses and dissertations represent a global source of information resulting from cutting edge research. While a proportion of this information is published in other forms much of the detail is not, and research emanating from lesser-known institutions, particularly in the developing countries, may be less likely to be published in mainstream journals. Creation of this information in electronic form making it readily accessible via the Web through standard, ubiquitous and free software programs provides the key to dissemination of this information independent of the source of the research. The ETD initiatives to date have proven that electronic theses and dissertations can be created using relatively low technology at a cost, which would be within the reach of most institutions. Portable packages have been developed which eliminate the majority of the developmental work required. The point has been reached where all research institutions can technically establish their own ETD program.

Traditionally, theses and dissertations have been extremely underutilized sources of information due to their lack of physical availability. The development of ETDs provides the opportunity for theses and dissertations to be recognized as a basic channel for the dissemination of research findings and an essential resource in the discovery process. Therefore, the focus for the future needs to be to ensure optimal access to ETDs by information seekers. This in essence means ensuring that ETD metadata records are accessible through as many channels as possible and are retrieved as integral components of searches without the researcher necessarily specifying an ETD. Some ways of achieving this could be

- Creation of a virtual union catalogue of ETD metadata through frequent regular harvesting of data from ETD sites which could be individual, regional or national

- Integration of ETD metadata into general metadata repositories for electronic scholarly information

- Ensuring search engines being established for other electronic scholarly information initiatives such as the Open Archives Initiative also search the ETD metadata repositories

- Inclusion of ETD metadata in local library catalogues

- Inclusion of ETD metadata in subject or form oriented databases

The development of ETD programs worldwide and the implementation of access structures have the potential to significantly enhance the opportunity for all researchers, independent of geographic and economic constraints, to make their contribution to the global research effort.

Work in all of these areas continues under development by NDLTD. Acting as agent for NDLTD, Virginia Tech is running a union catalog, drawing upon sites that support OAI. The result is accessible through Virginia Tech sofware as well as VTLS's Virtua software.

With support from the USA's National Science Foundation, Virginia Tech also is engaged in a number of research activities related to ETDs. These include matching efforts funded by DFG in Germany (in collection with Oldenburg U.) and by CONACyT in Mexico (with Puebla and Monterrey). These aim to promote mirroring (of metadata as well as regular data), high performance access, effective searching and browsing, visualization of results and of sites, and other advanced schemes.

It is hoped that all involved in the ETD efforts will assist, with a global perspective, so that all universities become involved, and ultimately all students submit an ETD, thus becoming better prepared to be a leader in the Information Age.

Exemplary Electronic Theses and Dissertations

Our goal here is identify "technologically innovative" theses and dissertations. We want to provide models of new media scholarship for the next generation of scholars and researchers.

Do you know of any theses or dissertations that are crafted in new ways, perhaps using streaming multimedia, interactive features (chats, listservs, response questionnaires, three-dimensional models, animation? If so, please take a moment to add the ETD to our collection: http://etdguide.org/ETDs.asp

By celebrating the innovative work of our graduate students and graduate faculty, we can inspire future research.

Exemplary ETDs in Our Database

View the ETDs that have been submitted to our database as exemplary models for future researchers and writers

Exemplary ETD Database

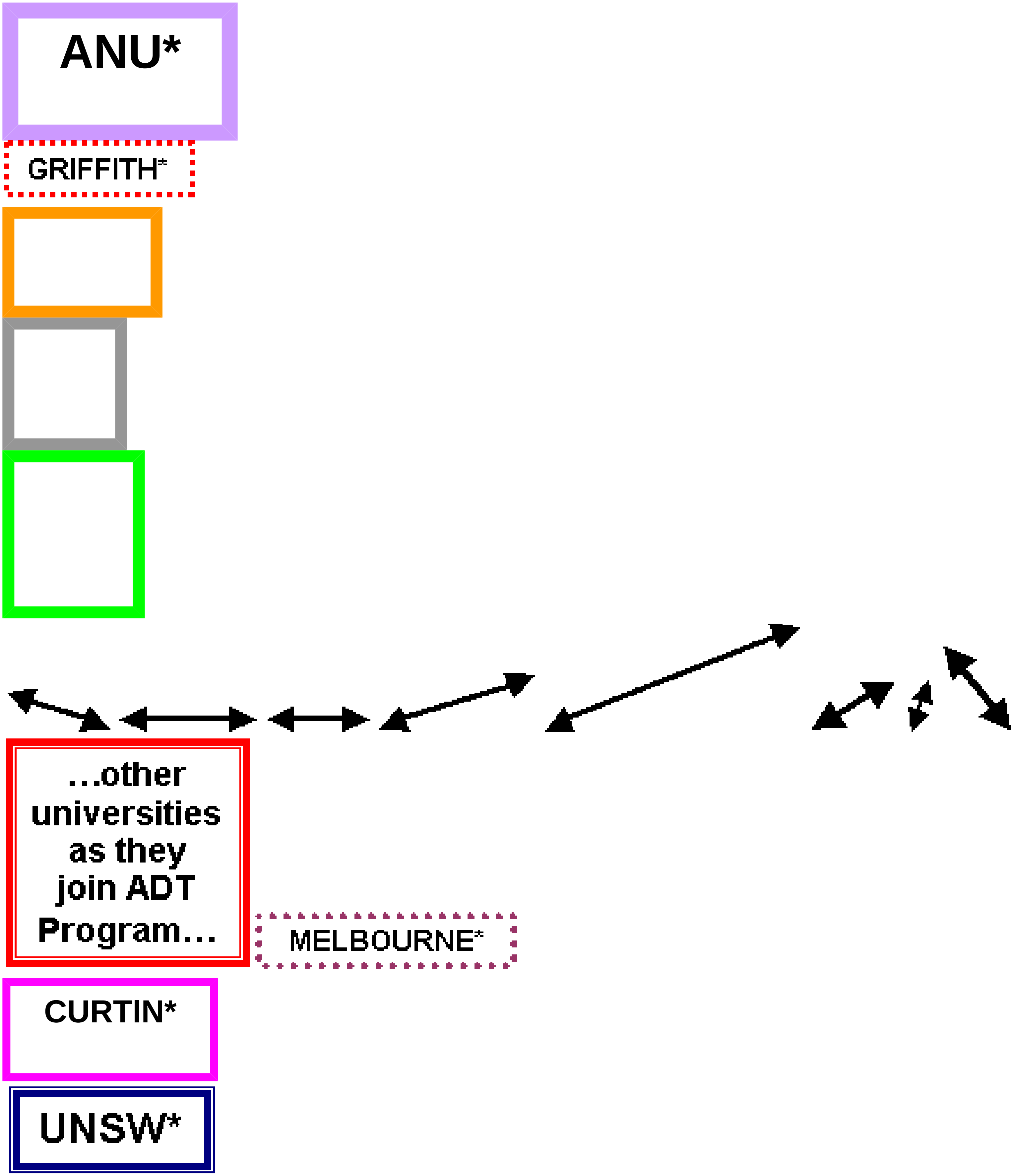
Submit an exemplary ETD in our database. Please be sure to clarify under "special features" what distinguishes the ETD from a "digital perspective." For example, does the ETD incorporate hyperlinks, multimedia, visuals, animation, or other interactive features?

[ETD Guide] [ETD Models] [ETD Resources] [Collaborate] [About this Site] [Chat] [NDLTD] [About the Authors]

adt architecture@glance

- each member institution installs ADT software on local servers as well as integrating process into local IT infrastructure. PDF files are housed on local institutions' severs.
- DC metadata automatically generated with each local thesis deposit. This metadata is gathered automatically into one central repository - the ADT national 'metadata' database.
- National 'metadata' database is searchable [via all ADT metadata elements], search results provide links back to local institution where the PDF files [full ETDs] are housed.
- **[*** UNSW, Curtin, Griffith, ANU, Sydney, Queensland & Melbourne
universities original ADT project group. UNSW lead institution **]**

QUEENSLAND* SYDNEY*



ANU*

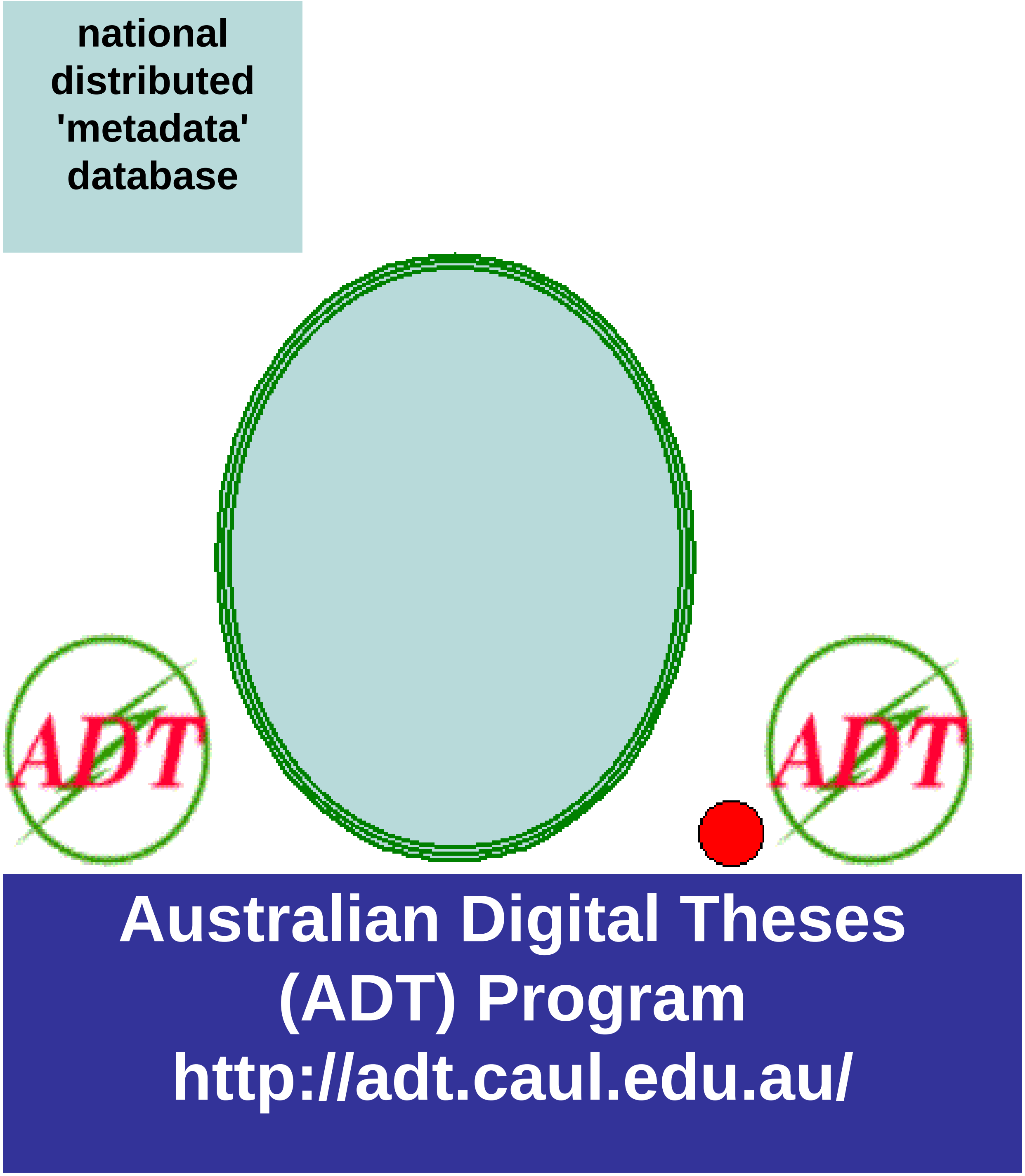
GRIFFITH*

…other universities as they join ADT Program…

MELBOURNE*

CURTIN*

UNSW*



national distributed 'metadata' database

Australian Digital Theses (ADT) Program http://adt.caul.edu.au/

4.3.4.1.2. ADT Architecture@aglance