Improving the ETD Landscape
for
ETD 2014: 17th International Symposium on ETDs
Leicester, England; July 23-25, 2014
by
Edward A. Fox
Professor, Department of Computer Science
Virginia Tech, Blacksburg, VA 24061 USA
fox@vt.edu, slides in http://fox.cs.vt.edu/talks/2014
Executive Director, NDLTD, www.ndltd.org

Abstract:
The information processing community has a good deal of work to do so that access to ETDs is improved to properly meet the needs of the research community. This paper outlines some of what is possible. It also makes suggestions for actions that will aid improved access, which could be undertaken by: those who administer ETD programs, the faculty who guide the development of ETDs, and the students who author ETDs.

Essential for improved access is having suitable information to support access. There are two main types of information at issue, the ETD itself, and the metadata that relates to it. If improved access is to occur, much work is needed to improve each of these, as well as the elements of which they are made. We touch on the need for complete metadata records, along with machine learning approaches to include subject categories. We also discuss the references that typically are at the end of ETDs, and the images included in or with ETDs, that can enable citation databases and related access, or content based image retrieval.

Also needed for improved access is suitable processing and services. On the user satisfaction side, many expect full-text search on the ETD content, which could locate a particular chapter as well as a full work. Faceted search and browsing also are expected, where each metadata element can help, and where subject categories may be of particular utility (justifying machine learning support when categories are missing or inadequate). Other services should be provided, like recommendation, image retrieval, citation retrieval, and hypertext support to allow better browsing, along with complex queries. Researchers should be able to find all of the ETDs with a given adviser, or from a particular research group, or that cite a particularly influential ETD or paper.

Processing at institutional repositories, or national libraries, is needed too. With more ETDs emerging that are scanned from paper or microform works, OCR and related cleanup are needed to allow extracting both text and citation data. Extracting or acquiring key metadata about advisers, committees, departments, and acknowledgments can be facilitated through local efforts where quality control can be enforced. These sites can undertake more extensive processing to support some of the rich services mentioned above, to yield local benefits in addition to global benefits. Institutional repositories could be suitably enhanced.

We hope that this paper will lead to an action plan to gradually improve access to ETDs, so new as well as mature ETD centers can take helpful steps along an agreed upon roadmap.
1. Why, What, Who, How:

The aims of the ETD movement (why) are to:

- enhance graduate education, and
- expand global research collaboration.

The goals (what) are to:

- help students communicate more effectively,
- get ETDs for all theses/dissertations: next goal 5 million (growing from 3.7M),
- help make ETDs open, accessible, and suitably preserved.

The key players involved (who) can be understood several ones. One approach is according to the levels or roles, e.g., students, faculty, staff, and (graduate) administrators. Another perspective is according to the disciplines or professions involved, e.g., computer science, information technology, library/information science, archival science. Of particular important is the digital library (DL) community, with many regular conferences [1] as well as a rich literature [2, 11, 4, 3]. In simple terms, according to the 5S approach [2], DLs are complex systems that:

1. help satisfy information needs of users (societies),
2. provide information services (scenarios),
3. organize information in usable ways (structures),
4. present information in usable ways (spaces), and
5. communicate information with users (streams).

Building from such perspectives, action is needed (how). At the heart of the effort is the creation of ETDs, requiring the use of suitable authoring systems, tools, and methods. Fortunately, much of this is enabled by the ongoing development of methods for electronic publishing, upon which the ETD community can draw; see Fig. 1. Less obvious, but also well supported by other communities, is the development of techniques for creating and managing the many types of content (including data and information) that are a part of scholarship. From the library world come software and methods for creating and enhancing metadata, to effect a complete description of the products of research. From the ETD world and the broad repository world come various systems that support the general need for uploading – with its phases of submission, approval, and refinement, integrated through appropriate workflows. Typically these are addressed through content management systems, provided through local groups or by them through systems supported by consortia, disciplinary groups, national institutions, or corporate partners. Sometimes these also support sharing, disseminating, and discovering interesting research, but another model is often applied. Using the Open Archives Initiative Protocol for Metadata Harvesting [8], the local repository becomes a metadata provider that feeds into regional, national, or global institutions. These then provide a variety of services, including access, and sometimes are enhanced to include preservation, as well as other ways to add value and utility. As this scene broadens, active institutions often also digitize their old files, to extend the historical coverage beyond the even more useful recent works that were ‘born digital.’ As can be seen in Table 1, there are a wide variety of digital library services that can be added or enhanced as part of an ETD improvement action plan.
2. Quality and Improvement:
Scholarly communication builds upon quality, since the overall aim is improvement. With regard to information objects like ETDs (and accompanying metadata as well as auxiliary content), quality can be assessed [11] at all phases of the information lifecycle. Thus we extend Fig. 1 to get Fig. 2, which in the outer ring adds some quality measures. Those involved in ETD programs should study Fig. 2, work to measure quality along each of the dimensions mentioned, and then strive to improve the quality for each phase, process, workflow, step, activity, and scenario that relates to ETDs. Thus, better works and descriptions will be created, distribution will ensure that the right people have access (including over time as a result of preservation), users engaged in research or learning will be able to find the right works, ETDs will be easily utilized to solve problems and generate new knowledge, and future ETDs will more effectively build upon prior works.

Another summary perspective on quality is given in Table 2. This shows, for each of the key content-related concepts in the digital library field, some of the dimensions of quality that pertain to each of those concepts. Those involved in ETD programs should study Table 2, and for each content related aspect of their efforts, work to measure quality along each of the dimensions mentioned. They should strive to improve the quality of the content and content aggregations listed. Thus we will have better and more easily accessed ETDs, enhanced and accurate metadata descriptions, comprehensive collections, local catalogs and an NDLTD [7] Union Catalog that are as complete as possible, repositories that properly connect the collections and catalogs, and a rich set of services that can work together to ensure efficient and effective discovery and use of ETDs.

One of the aims of NDLTD and its website and programs [7] is to document all this, providing guidelines for best practices. These are concerned first with the works of graduate researchers, i.e., ETDs. Award winning ETDs not only include a PDF file, but also may include XML, raw/original representations, multimedia (e.g., images, audio, video, graphics, animations), software, simulations, websites, and various types of dynamic content. They usually are rich in various ways, e.g., including data, auxiliary information, and comprehensive (in terms of number and detail) sets of references, sometimes in annotated bibliographies. They may support reproducibility in experimental studies or analyses. The related metadata will make clear the level of the work (e.g., undergraduate, masters, or doctoral), the institution, members of the advisory and/or examining committee, acknowledgment data, subject categories in suitable classification systems, and suitable authority information on all people involved.

Best practices at the local level ensure suitable guidance from faculty and graduate school personnel, along with highly educational training and assistance. These relate not only to creating highly expressive ETDs and accurate metadata records, but also to issues like intellectual property rights (connected with patents, copyright, and broadness of access of the ETD). Ideally, ETDs will be in formats supported by international standards, so all content can be easily preserved. Local support by university libraries should include local or equivalent help with institutional/regional/national repository and archive systems. They also should connect with global services like those that leverage the NDLTD Union Catalog.
3. Improve and benefit from related movements:
Since the Information Life Cycle connects those working with ETDs with a number of
related movements, it is important to be aware of those, and to ensure suitable
connections are made between them and ETD programs. This is a “two way street” in
that ETD programs can leverage the work in those movements, and those movements
may find broad expression through actions related to ETDs. For example, the Open
Access movement may find its easiest expression by way of graduate students making
ETDs openly accessible. This is feasible since universities with ETD requirements share
large numbers of openly accessible works each year, while faculty may be reluctant to
change their behavior regarding open access publishing. Thus, many institutional
repositories have ETDs as their largest collection of works. Further, students who support
open access with their ETDs are likely later to support other open access publishing.

Another movement of interest to the ETD community is with regard to references.
Students might benefit from the work of those focused on references and citations. Thus,
sites like Zotero [13] can be very useful for students managing large sets of references.
They also may benefit from the findings of projects like Hiberlink [5], which can guide
them to avoid problems with “reference rot” so that in the future, readers of their ETDs
will be able to find each reference, in spite of changes in websites.

Graduate students can start early in learning how to have author IDs, build author
profiles, look up works of others who have author profiles, and thus help advance the
Semantic Web as well as services provided by libraries. They can learn of standards for
name identifiers [6] and rich services that build upon author IDs to help connect research
and researchers [9].

More broadly, the ETD movement can benefit from work on digital libraries, and vice
versa. There are many digital library conferences addressing research, development,
practice, and education, including a joint one in 2014 that brings together groups
connected with professional societies (e.g., ACM, IEEE) and communities that began
with a European focus (TPDL) [1]. It can help that there is a good theoretical foundation
for the DL field [2], that it can help with quality as well as integrating information across
distributed sites [11], and that there are strong advances in related technologies [4] and
applications [3].

4. Related problems and possible technical contributions:
With regard to technology [4], there are many improvements possible that will help with
ETD activities. We need new and better systems, studies to assess usage and user
satisfaction, efforts to improve or add services, and enhanced related practices and
training.

One key area is with regard to searching for ETDs. Fig. 3 shows a simplified view of the
process. Most search systems, even Google, have limited effectiveness in supporting
searching, especially of ETDs, when a comprehensive effort is desired so that all works
that are relevant are found. The usual requirement of scholars for high recall (the
percentage of what is relevant that is found) is generally poorly supported by search engines, which instead focus on high precision (the percentage of what is found that is relevant), which is desired regarding other types of needs than scholarly exploration.

One problem with searching ETDs is that there are few search systems that support effective full-text search, which is a difficult problem when each PDF file is made up of hundreds of pages. Further, many (parts of) ETDs are not accessible to search engines, so cannot be found using them. Accordingly, many search systems only use the metadata attached to an ETD, and often that is incomplete. It is common that metadata records don’t point to the actual ETD, so that it cannot be downloaded or crawled. It is often the case that there is no subject category assigned to an ETD, or the category given is something vague like “educational material.” Often the language of the work is not reported in the metadata record, and if the work is not in English (and has no abstract in English), a search system may not handle searching in the actual language employed. Sometimes, names of faculty connected with an ETD are not given in the metadata record. Often the level of the work, the department(s) of the student, the university name, and other important details are missing. Usually if a date is given it is unclear if it concerns the date of the final defense of the work, the date the student completed it, or the date that the work was put into some repository. All of these metadata quality or consistency problems cause faceted search systems to be viewed as having low effectiveness.

Related to these concerns is the lack of support for searching across languages or for searching on the content of ETDs. Technical contributions are needed to apply to ETDs methods like cross language information retrieval, or content based information retrieval [3, ch. 1] (that allows searching for works that contain something like an exemplary photo or sound or table or figure). Better methods for full-text search of large text files also are needed.

To aid discovery and exploration, browsing methods also should be enhanced with regard to ETDs. We should have recommender systems to suggest ETDs that are like one that a researcher has found. This should be extended so that researchers will be notified when a new ETD has been added to the Union Catalog that relates to works they have found to be of interest in the past. Notification or alert or RSS feeds, based on suitable profiles or filters, might be of particular help to scholars interested in new graduate research related to their work.

Hypertext/hypermedia technologies also could be applied. One aspect of this relates to connecting ETDs with authors and other people involved. As mentioned above, work with identifiers like ISNIs [6] or ORCIDs [9] can be helpful. Additional linking can result from advances discussed in the next two sections.

5. Topic and category determination
For linking or browsing, it is helpful to assign ETDs to suitable categories (in a broad framework like the Dewey Decimal System or the Library of Congress Subject Headings, or in discipline specific systems like UMLS or MeSH for medicine, or the ACM
Category System for computing). This requires a suitable taxonomy or ontology [4, ch. 3] as well as machine learning for automatic classification [4, ch. 4]. Ideally that should be done in ways tailored toward ETDs.

First, when an ETD is classified, it would help to have a ranked list of suitable categories rather than a single one, since many works are interdisciplinary or cross disciplinary. Since there is no requirement for shelving a work in a single place, having more than one category assigned is appropriate. Second, since ETDs often are large, it would help to have categories assigned at the chapter level too, so that individual chapters of interest could be discovered. Finally, since comprehensive category systems have many levels in their hierarchy, it would be useful for researchers browsing in an ETD collection to have classification be done only to a moderate level of detail, e.g., down to the third or fourth level as opposed to down to the fifth or sixth level. There is work underway to address these needs [4, ch. 4], but many technical challenges exist, especially with regard to obtaining suitable training data, working with complicated hierarchies, and validation.

Also helpful for browsing, in addition to categorization, is topic tagging. In other words, for a given ETD, software should extract or generate a generalized description of the topics discussed. Such is appropriate when a new area of research is emerging, and ETDs don’t fit well into an existing category system. It also helps if authors have not assigned keywords, or if the keywords given by the author are at a level of specificity different from what a searcher understands (i.e., is too specific or too broad/general). In such cases, automatic identification of topics for a text can be applied, using LDA or newer knowledge-based methods [12], though such methods may need to be tailored to ETDs to become really practical and helpful.

6. Reference Extraction and Databasing:
Browsing and linking can be further enhanced when the references found in ETDs are identified and suitably processed. Text extraction methods can be applied to find the reference section and its entries [4, ch. 5]. Advanced methods are needed to parse or analyze a particular entry in the reference section, so as to determine the author(s), type of work, date, title(s), pages, and other details. When a DOI or similar identifier is included, this task can be easy, though with URLs and “reference rot” [5], the identification may not be so straightforward. If there is a comprehensive database that can be examined, or if the ETD author supplies a well-structured reference database (e.g., using a tool like EndNote or BibTeX), the task can be simplified. However, in the general case, e.g., in a collection of ETDs that covers all disciplines, there are many reference styles that are utilized, in addition to ad hoc references that some authors use if they are not using a popular tool, or are drawing upon multiple sources with different formats. In such cases, more complicated approaches are required [10]. Applying these to large ETD collections is likely to require additional work, including knowledge engineering so that the different journals and conferences in different domains are described, with their full names as well as variations and acronyms.

If references are properly analyzed, and converted to some canonical form suitable for a bibliographic database, then it will be feasible to add links to ETDs so a reader can click
on a reference and jump right to the cited work. Other links could be followed, if the
authors and editors are connected with authority records, so their other works can be
found too. Thus, hypertext and browsing support could be greatly expanded to aid those
interested in graduate research and related studies.

7. Summary:
As explained above, there are many opportunities to improve and expand ETD efforts.
Different parties can be involved, as was explained in Section 1, which also discusses the
why, what, and how aspects of improvement. There is considerable room for
improvement, which could lead to enhanced quality of content and services. Some of this
can leverage work in related movements, including open access, repositories, and digital
libraries.

Diverse technical contributions could be applied, aiding searching, browsing, linking, and
other services. ETDs could be classified as to subject, at the level of the work as well as
the chapter. Topics could be identified, to also aid with browsing. References could be
identified and converted to easier to use forms, with DOIs or other identifiers connected,
so readers could jump directly to related works.

Accordingly, all those interested in ETD programs are encouraged to work with students,
faculty, librarians, and technologists so that ETDs can be suitably supported to better
serve the needs of learners and researchers.

8. Acknowledgments:
I thank my family, mentors, teachers, and students. In particular, this paper mentions the
doctoral research of Sung Hee Park, Venkat Srinivasan, and Seungwon Yang, the second
of which has made use of data provided by OCLC. Some of the work covered was funded
by the US National Science Foundation through grants IIS-0535057, 0916733, and
1319578. More broadly I thank all those working with ETDs, especially those connected
with NDLTD, including its Members, Board, Committees, and Working Groups.

9. References:
   (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries

2. Edward A. Fox, Marcos Andre Goncalves, and Rao Shen. Theoretical Foundations for
   Morgan & Claypool, 2012, 180 p., http://dx.doi.org/10.2200/S00434ED1V01Y201207ICR022,
   supplementary website https://sites.google.com/a/morganclaypool.com/dlibrary/


10. Sung Hee Park, "Discipline-Independent Text Information Extraction from Heterogeneous Styled References Using Knowledge from the Web", June 2013, VT CS Ph.D. dissertation


Table 1: Digital library services taxonomy

<table>
<thead>
<tr>
<th>Infrastructure Services</th>
<th>Add Value</th>
<th>Information Satisfaction Services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repository-Building</strong></td>
<td><strong>Add</strong></td>
<td><strong>Satisfaction</strong></td>
</tr>
<tr>
<td>Creational</td>
<td>Value</td>
<td>Services</td>
</tr>
<tr>
<td>Cataloging</td>
<td>Conservation</td>
<td></td>
</tr>
<tr>
<td>Crawling (focused)</td>
<td>Converting</td>
<td>Classifying</td>
</tr>
<tr>
<td>Describing</td>
<td>Copying/Replicating</td>
<td>Clustering</td>
</tr>
<tr>
<td>Digitizing</td>
<td>Emulating</td>
<td>Evaluating</td>
</tr>
<tr>
<td>Federating</td>
<td>Renewing</td>
<td>Extracting</td>
</tr>
<tr>
<td>Harvesting</td>
<td>Translating (format)</td>
<td>Indexing</td>
</tr>
<tr>
<td>Purchasing</td>
<td></td>
<td>Measuring</td>
</tr>
<tr>
<td>Submitting</td>
<td></td>
<td>Rating</td>
</tr>
<tr>
<td>Acquiring</td>
<td></td>
<td>Reviewing (peer)</td>
</tr>
<tr>
<td>Cataloging</td>
<td></td>
<td>Surveying</td>
</tr>
<tr>
<td>Crawling (focused)</td>
<td></td>
<td>Translating (language)</td>
</tr>
<tr>
<td>Describing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1: Information lifecycle

Borgman et al. 1996
http://is.gseis.ucla.edu/research/dig_libraries/
Fig. 2: Quality and the information lifecycle

Table 2: Quality dimensions

<table>
<thead>
<tr>
<th>DL Concept</th>
<th>Dimensions of Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital object</td>
<td>Accessibility</td>
</tr>
<tr>
<td></td>
<td>Pertinence</td>
</tr>
<tr>
<td></td>
<td>Preservability</td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
</tr>
<tr>
<td></td>
<td>Similarity</td>
</tr>
<tr>
<td></td>
<td>Significance</td>
</tr>
<tr>
<td></td>
<td>Timeliness</td>
</tr>
<tr>
<td>Metadata specification</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Conformance</td>
</tr>
<tr>
<td>Collection</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Impact Factor</td>
</tr>
<tr>
<td>Catalog, Repository</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
</tr>
<tr>
<td>Services</td>
<td>Composability</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
</tr>
<tr>
<td></td>
<td>Effectiveness</td>
</tr>
<tr>
<td></td>
<td>Extensibility</td>
</tr>
<tr>
<td></td>
<td>Reusability</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
</tr>
</tbody>
</table>
Fig. 3: Searching overview

Table 3: Summary of action plan

<table>
<thead>
<tr>
<th>Activity/Phase/Store</th>
<th>Some Aspects/Portions/People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilization</td>
<td>Search, browse, recommend, CBIR, CLIR, Annotate: comment/rate/tag</td>
</tr>
<tr>
<td>Linked Data</td>
<td>Metadata, ETDs, Multimedia, Data, References, ISNs, MetaArchives</td>
</tr>
<tr>
<td>Adding Value</td>
<td>Content mining, Classifying, Topic spotting, Reference extraction/ databasing, Migrating</td>
</tr>
<tr>
<td>Repositories</td>
<td>IR, ETD collection, Data collections, References collection, ORCIDs</td>
</tr>
<tr>
<td>Creation by</td>
<td>Students, faculty</td>
</tr>
</tbody>
</table>