A Review of Emerging ETD Initiatives, Challenges and Future Developments

Terry Ann C. Cayabyab

Abstract—Fast, worldwide interest spurred countries and organizations of the impact of theses and dissertations in electronic format; hence, the conception ETD. Regarded valuably in the research arena for data management, ETD mushroomed swiftly in developed countries. Compared to developing countries, isolated initiatives exist in few scholarly writings demonstrating some successful attempts while some projects are in its infancy stage if not chaotic. Aside from the concerns on intellectual property rights, findings in this review demonstrated the problems that emerged as consequences of the integration and collaboration of ETD projects that were designed and developed individually. This literature review aimed to show the emerging global concepts that could be used as benchmark for ETD initiatives to developing countries such as the Philippines.

Index Terms—ETD, challenge, goals, review of literature.

I. INTRODUCTION

The significance of theses and dissertation as the bedrock of graduate education [1]-[4] and as a rich source of information [2], [4]-[7] has ignited the interest of the international community toward the development of electronic theses and dissertations (ETD).

With the increasing number of research outputs in graduate education [8]-[10], ETD is primarily sought to ease the submission, storage, dissemination and the cost of handling theses and dissertations[8]. This is the primary reason for ETD development, especially to developed countries. In the developing countries however, there are few, isolated initiatives, which are, if not in its infancy stage [7], rather failed [11], [5] to meet stakeholder requirements. Nowadays, organizations have collaborated to provide better access to ETDs. Aimed towards facilitating the research conduct and enhancing the productivity of scholars, ETD perspectives have expanded towards interoperability [12], open access [8] and global visibility. Likewise, the following information systems were consistently associated with ETD: institutional repository [13], digital library [2], [10], [4], research management system and knowledge management system. Several issues however thwarted the attainment of the said objectives.

Aside from copyright [9], [14], as a principal concern with ETDs, more unforeseen problems have emerged as consequences of diversely developed ETDs. Access privileges [10], [11], [15], metadata standardization [12], [16], system models [2], [11] and policies [1], [6], [14], were sought to mitigate the problems associated with ETD. At present however, there are still evidences of dissatisfaction among the research communities [1], [6], [9], [11].

This systematic review is vital in understanding the challenges in the realization of design goals for ETD. The author believes this will be a resilient basis for the development of useful ETD, especially in the context of developing countries like the Philippines whose research capabilities is being realized internationally [17]. However, only top-rated, metropolitan-located and financially-abled universities have access to these resources.

While there are evidences of best practices, the literatures reviewed in this paper demonstrate 1) non-adoption to any software development methodology; and 2) non-theory-based studies. These findings are contributory to the management of ETD projects for developing countries, where there is limited literature.

Manuscript received May 6, 2014; revised July 23, 2014.
Terry Ann C. Cayabyab is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: taccayabyab@yahoo.com).

II. METHODOLOGY

To address the objective of this study the model of [18] was adopted and served as a guide. The phases are as follows: 1) search for sources of quality literatures, 2) keyword search, 3) backward and forward search, and 4) processing.

In search for quality literatures, the author was constrained with the access privilege over electronic databases and the veracity of the source on the phenomenon that is highly technical in nature. Hence, the following electronic databases were sought: ACM, IEEEExplore, ProQuest, EBSCOHost and Elsevier. The keywords used to search academic and conference articles were electronic theses and dissertations or ETD. Primarily, the first 100 articles were reviewed for relevance; however, most of the databases have only its first 20 to 40 search results about ETD. ProQuest presented the most search results of up to the first 80 articles; scanned but were trimmed down for a few reasons. First, articles were not relevant; second, articles are downloadable in other databases; third, articles are restricted from being fully downloaded; and fourth, access is redirected to the institution where the article was homegrown. Having seen a similar study conducted by [14] about challenges in ETD development, I have prompted to consider studies starting 2003 onwards as the duration of the review. Backward and forward search wereessential to understand the chronology of emerging themes.

With literatures available, the next step is called processing. This step is demonstrated by the activities such as comprehending, applying, analyzing and synthesizing of the literatures. During comprehension, emerging themes were identified. Afterwards, application is conducted by categorizing the literatures. The next step is the content analysis which involves relating, comparing and explaining
the information of importance. Finally, synthesis entails coming up with generalizations. In this study, generalizations are drawn to show future research concepts.

III. RESULTS AND DISCUSSION

The studies that were reviewed are presented in this section basing on the emerging themes as follows: description of ETD; history; challenges; approaches and the implication of findings.

A. Electronic Theses and Dissertations

Empirically, even if ETD was the keyword used to search for journal articles, other information systems such as digital library, institutional repository, research management, research information system and knowledge discovery were consistently associated. Digital library is defined by [19] as “an informal collection of information, stored in digital formats and accessible over a network, together with associated services”. On the other hand, an institutional repository (IR) is defined by [20] as a “digital archive of an academic institution’s intellectual output, encompassing research articles, preprints and postprints, theses and dissertations, and administrative documents, course notes or learning objects, conference presentations and proceedings, technical reports, working papers, and multimedia material” [21]. Reference [22] further describes IR as an alternative for the management of academic information generated by universities and research centers. Theses and dissertations are contents either in the library or repository.

Electronic theses and dissertations (ETD) was introduced by [23] to delimit the broad coverage of IR contents. ETD is a subset of digital library, which is also a subset of IR. ETD is defined as a master’s thesis or doctoral dissertation in electronic, rather than paper format. “ETD refers to a master’s thesis or doctoral dissertation that is archived and circulated electronically rather than in print and bound format” [9], [14]. ETDs have gained huge attention of institutions and organizations worldwide [2], [5], [16], [24], [25] as a technological tool for the graduate research outputs to be communicated globally [9]. On the other hand, a research management system (RMS) is broader in concept compared to IR for the research processes that are integrated. The Current Research Information System (CRIS) encompasses the collection, preservation and dissemination of data about institutions, researchers, research projects, equipment, published results and other relevant data on scientific research activity [12].

B. History

The concept of ETD was initiated by UMI in 1987, during a meeting hosted by UMI in Ann Arbor, Michigan, with the participation of representatives from Virginia Polytechnic Institute and State University (Virginia Tech) and the University of Michigan [1], [7], [9], [14], [24]. UMI is now Proquest’s UMI Dissertation Publishing. In 1995, Virginia Tech started accepting ETDs; then every year thereafter, requests for ETDs increased rapidly. The project caught the attention of the Organization of American States and the United Nations Educational, Scientific and Cultural Organization, making them members of the Networked Digital Library of Theses and Dissertations. It is a collaborative ETD project that promotes to improve graduate education, and reduce submission cost while facilitating the availability of graduate student researches [9]. [24]. Thereafter, several members worldwide have joined the advocacy as shown in http://www.ndltd.org/about/membership. Likewise, affiliation to world-known organizations like NDLTD became a trend in aiming for wider visibility, example: China NDLTD [2].

<table>
<thead>
<tr>
<th>Design Goals</th>
<th>Challenges</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open access</td>
<td>- management issues (submission, storage, access &amp; dissemination)</td>
<td>[1], [6], [8], [9], [26]</td>
</tr>
<tr>
<td></td>
<td>- plagiarism; intellectual property rights</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- social &amp; philosophical issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- incompatibility of metadata with controlled vocabularies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- full-text search limitation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- migration issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- accessibility issues (redirection, indexing, proprietary format)</td>
<td></td>
</tr>
<tr>
<td>Accessibility and interoperability</td>
<td>- institutional ETDs in different structures</td>
<td>[1], [2], [5]-[9], [11], [12], [27]</td>
</tr>
<tr>
<td></td>
<td>- lack of provision for mandatory submission; limited participation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- opposition concerns (purloined ideas; substandard graduate researches; Internet access rejections; confidentiality of research topic)</td>
<td></td>
</tr>
<tr>
<td>Visibility</td>
<td>- mismatch between keywords provided by author and controlled vocabularies</td>
<td>[14], [24], [26] [28]</td>
</tr>
<tr>
<td></td>
<td>- permanent and secure preservation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- cultural attitudes on IPR</td>
<td></td>
</tr>
<tr>
<td>Sustainability</td>
<td>- limited technical skills in incorporating multimedia and interactive perspectives</td>
<td>[4], [16]</td>
</tr>
<tr>
<td></td>
<td>- costly and erroneous software</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- limited participation, fund and technical skills in incorporating multimedia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- legal issues: copyright, IPR and political implications</td>
<td></td>
</tr>
<tr>
<td>Centralization</td>
<td>- unmatched keywords generated by author and cataloger</td>
<td>[29]</td>
</tr>
<tr>
<td></td>
<td>- copyright; access fees; absence of policy &amp; infrastructure; full access</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- extensive and lengthy manuscript</td>
<td>[30]</td>
</tr>
</tbody>
</table>

C. Design Goals and Challenges

In the study of [24], it was stated that studies about ETD focused more on “why” and not of “how”. The first focus is on the advantages and benefits of having an ETD, whereas the second pertains to the processes related to the
development. In both cases, the challenges encountered in promoting ETD are tackled. In contrast however, there is 25% on the description of “why” ETD and 75% focused on “how” but is a combination of descriptive and developmental study. This only shows that, although several global ETD initiatives already exist, the need for more studies on ETD development is imperative because of the scenarios implying that goals for developing ETDs are not yet met. Failure of the ETD project include the following evidences: 1) there is a minimal access to ETD [1], [9]; 2) minimal contents added to the database since the time the system went live [7] and 3) no longer used system [11], thus requiring migration of data.

The emerging goals and challenges associated with the development and evaluation of ETD are presented in Table I.

1) Open access

Data management is one of the initial objectives of ETD development. First of which is to facilitate the concern on library space occupied by hardbound theses and dissertations manuscripts that are increasingly produced [8]. Second, theses and dissertations in electronic format reduce submission and handling costs [9]. Last, ETD aids in the achievement of library productivity goals. Included therein is to assist scholars in their research needs, which is made possible by making research materials available, disseminated and easily accessible to scholars. This way, theses and dissertations become more usable [24].

Several institutional initiatives emerged for the said goals; however, there were hesitations for patronizing ETDs. First, the printed manuscripts are considered more permanent as compared to an electronic copy stored in CD-ROM. The secured preservation is questioned because the storage device is said to last within 15 to 20 years [14]. As observed, none of the literatures reviewed discuss the feasibility of adopting ‘cloud computing’ for this purpose. Second issue pertains to copyright infringement, plagiarism [1], [2], [8], [14], [16], [29] and ownership rights to the dissertation of graduate students [14]. The worries of students are associated with their research topics from being “stolen” prior to publication [11], [14]. Reference [1] tried to resolve the issue with the fair use policy; however, this is still debatable up to the present [1].

a) Policy and file format

In terms of data management, specifically the submission process, policy and file format could affect the implementation of ETD. In the study of [14], the successful implementation of ETD by NDLTD members is associated with policies on training and submission process. At Bar-Ilan and Haifa University, deposit of document is mandatory [11]. While in India, the Vidyanidhi project has failed due to the absence of provision for mandatory submission [5]. Motivated to pursue ETD in India, another project called Shodhganga was initiated. In line with this, a survey was conducted to evaluate the participation of the 20 members in line with the memorandum of understanding that encourages submission of ETDs. Findings showed however, that the project is yet at the infancy stage [7]. These studies show that the imposition of a policy plays an important role in the implementation of a project.

Another concern with regards to data management is the file format. For an ETD, the standard file formats are PDF, HTM, XML [1], doc, docx and odt [12]. The first is considered the dominant format for three reasons: first, it is compatible and portable across other programs and platforms [8]; second, it is widely used in electronic publishing since it is easily created with any word processor; and third, it has security and accessibility controls [2]. However, compared with the other formats, PDF does not support full text search [1]. Moreover, PDF is being challenged for being a proprietary format [1]. In systems development, deciding on the file format is a factor for deciding the hardware and software specifications in order to sustain the storage and access capabilities of ETDs. Also, the format will also depend on the objective set (e.g. XML for full-text search).

b) Storage platforms

There are two set-ups for storing the contents of an ETD [1], either centralized [27] or decentralized [29]. Collaboration of ETD systems where contents are stored in one repository is centralized; whereas, ETDs that are stored in multiple repositories is decentralized [1]. Germany’s Deutsche Bibliothek and Africa’s Database of African Theses and Dissertations have a centralized database [11] while the China Networked Digital Library of Theses and Dissertations is decentralized [2]. None of the studies reviewed presented the advantage or disadvantage of the two set-ups, nor any guideline for the appropriateness of any set-up.

c) Accessibility issues

One of the failure evidences of ETD project is the underutilization of contents [6], [8], despite the growing number of studies produced yearly [1], [9], [10]. Nowadays, ETDs are accessible on the Internet via popular search engines (e.g. Google, Dogpile et al.) [15] or scholarly databases (e.g. ProQuest, ScienceDirect, Google Scholar, IEEE et al) [31]. The first option however, is being questioned in terms of quality of service, structure, consistency and provenance [1] as compared to the second option that follows certain formats and standards [15]. The latter is more preferred source of materials however, there are also some challenges being encountered by scholars.

Earlier, one of the first challenges of ETD development that was mentioned is copyright. In the study of [11], a resolution was successfully handled by a legal adviser ascribing the ownership to the author of the work. Consequently, student-author may opt to define access preference for his work. The access categories are summarized as follows:

- open, fee-based, restricted and mixed
- free, home-bound, time-bound and section-specific

For open or free access databases, access and download privileges on the database contents are allowed without any cost, examples are the WorldCat or Open WorldCat and Australasian Digital Theses. However, the second category is accessible to its members only. NDLTDOn the other hand is fee-based; that is, upon payment of institutional membership fee, libraries of universities and their clients are given free access and download capabilities to electronic database contents. Restricted documents are not allowed to be accessed at all [9]. Mixed category allows free search for
citations, access to abstracts and the first 24 pages of any TnD. However, in the case of the ProQuest Database (formerly UMI), payment details are required for a presumed order of the complete TnD [9]. Home-bound refers to graduate UMl), payment details are required for a presumed order of the complete TnD [9]. Time-bound means that documents are redirected to the university’s library where the student graduated [9]. There are also instances when the full-text TnD is not yet available; hence, researchers are required to order the document [9] (e.g. first 128,000 characters) [26] or first 24 pages of the document only [14].

Reference [32] stated that, access restrictions to these materials hinder the “research results to be communicated to the outside world”. Moreover, instead of the supposedly open access and global visibility of researches [9], there is a growing evidence of proprietors [1]. Aside from lack of knowledge, nurses also lament on the access fee resulting to the underutilization of ETDs to their community. These circumstances altogether, further deprive scholars from developing countries from accessing these materials openly [4]. In this connection, a survey may be needed to prove that only a few well-known universities are capable of providing this service to their institutions. This may support the need for differential price structures stated by [14] for developing countries.

2) Visibility, interoperability and centralization

Even with existing ETDs and journal databases that either store or provide a link to access theses and dissertations, there are still evidences of user dissatisfaction. The most common problems are: a) search results not really needed; and b) difficulty in looking for the right material. During the earlier years, ETDs had been developed institutionally and with diverse structures. Later was it realized that achieving national global visibility would require interoperability or centralization mechanisms to enable transfer of data among digital repositories. The two approaches are a) creation of a portal for centralization purposes, and b) definition of metadata standard for interoperability.

a) Portal

In the study of [2] and [11], the development of China Academic Library and Information System (CALIS) was described. It is a federation of academic libraries in China with more than 152 members that have their institutional ETDs in different structures. The local system is for the submission, annotation, approval, archival, distribution and access to their local ETDs. A central portal for the CNNDLTD union collection was sought to support the academic education system based on China’s academic education digital library. Another project called Shodhganga repository, is a collaborative, national project. It hosts, maintains and makes the digital repository of the Indian ETD. The INFLIBNET Centre provides a platform for students to deposit their TnDs and make them available to scholars. Likewise, it provides link to other Indian and journal databases.

Both the CALIS and INFLIBNET Center provide access to TnDs. Their set-up differs in terms of the storage mechanism. CALIS is decentralized, while INFLIBNET is centralized. In both projects however, the reason for adopting a centralized or decentralized platform was not presented, which could have guided ETD developers.

b) Metadata

Metadata refers to the set of fields used to store and access information about an ETD. Brought about by the heterogeneous structure that ETDs had been created with, metadata elements that are supplied by author should be systematically processed. Implementing these steps of evaluating the compatibility of data models could potentially increase the availability of research outputs [12]. The emergence of ETD studies on metadata demonstrates the viability of goal attainment, as shown in the succeeding studies.

In the study of [12], the integration of ETD with research management system involves the following steps: 1) analysis of the metadata of various systems; 2) definition of a comprehensive metadata set that is compatible across other systems; 3) extension of data model to store the comprehensively defined metadata; and 4) expansion of the system with a module for storing metadata elements. Similarly, [2] demonstrated the importance of metadata standard for an interoperable environment hence, a standard based on Dublin Core (DC) element was also defined to cater the needs of the Chinese community of scholars for the China NDLTD. Aimed towards open and wider accessibility to ETDs, the University of Louisville’s has adopted DC-based metadata in the migration of database to CONTENTdm [26]. Reference [8] likewise described the University of Waterloo’s E-thesis project towards open access. In partnership with Theses Canada and the NDLTD, the adoption of OAI-compliant metadata is one of the factors that contributed to the attainment of the objective.

Empirically, the standardization of metadata is needed for interoperability and/or integration [12], [16]. Interoperability is used to describe different models of exchanging data among information systems; while, integration is the set of methods for facilitating data from among interoperable systems [12]. There are two metadata formats adopted in the projects that were described in the studies reviewed: ETD Metadata Standard (ETD-MS) and Dublin Core (DC) Element Set. Presented in Table II are the projects and the metadata format used.

<table>
<thead>
<tr>
<th>Metadata Format</th>
<th>Project Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETD-MS</td>
<td>NDLTD</td>
<td>[2], [12], [24]</td>
</tr>
<tr>
<td></td>
<td>OhioLINK ETD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNNDLTD</td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td>DART-Europe</td>
<td>[12], [26]</td>
</tr>
<tr>
<td></td>
<td>E-Thesis Portal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRIVER</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CONTENTdm</td>
<td></td>
</tr>
</tbody>
</table>

Aside from metadata, the protocols (e.g. Open Archives Initiative for Metadata Harvesting Protocol, MARC21 and OAI-ORE) and controlled vocabularies are additional mechanisms sought. The protocol is used to harvest metadata elements, modify and enhance the data; afterwards, added to a library catalog (e.g. WorldCat or Open WorldCat). The role
of the protocol likewise, improves the accessibility of ETD [24], and possibly free of charge [9]. In the survey of 266 ETD repositories conducted by Ref. [4], majority utilized the OAI-PMH despite the more competitive capability of the OAI-ORE. A comparative analysis of the protocols are recommended in the said study to evaluate the key characteristics, sustainability of standard, level of dependency on different platforms and their deposit methods. Finally, controlled vocabularies were compared with author-assigned keywords for use in online catalogs and databases and made available in the Library of Congress Subject Headings (LCSH). Findings showed that there were a large percentage of terms that are not covered by cataloger-assigned LCSH. Hence, cataloging mechanism that considers the controlled vocabularies is vital for global visibility of ETDs across diverse platforms [28].

3) Sustainability

In building an information system such as ETD, a systematic process is required which is applied, not only during the development phase but even after the system has been implemented. This is to ensure that system continues to meet the needs of stakeholders. One of the challenges encountered during the development phase is the cost of the project [2]. In the study of [4], the development of Greenstone Digital Library, was based on open source software. This was sought as a solution to expensive proprietary software. The Focus on Access to Institutional Resources (FAIR) Programme was made possible with funding from Joint Information Systems Committee (JISC) in collaboration with the Higher and Further Education Community [16].

Aside from the cost, training was cited by [14] as one of the best ETD practices. It includes workshops and documents that are used to educate students on the submission process. Additionally, volunteer efforts were exerted by library staff, graduate deans and IT personnel. This collaboration among stakeholders could uphold the objectives set [1]. Additionally, training could alleviate the nurse’s lack of awareness and understanding of the value of TnD as found in the study of [6].

4) Recent trend

Knowledge discovery was conceptualized by Ref. [30]. He was concerned with the lengthy ETDs, that is, more than hundred pages that researchers would scan through; and the different languages that ETDs are written. First, he investigated the digital library approaches towards discovering knowledge easily. The result served as foundation for his system proposal. The system is based on natural language processing; it summarizes document and creates a concept map based on the relevance of topics. As of this writing however, evaluation of studies on knowledge discovery and how it facilitates the ETD goal is yet minimally explored.

From the studies presented, IETD and CRIS were developed following a more organized step compared to other initiatives. While focusing on the metadata is very significant, there are other factors that may contribute to the development of ETD that will better suit the needs of the researchers. The following emerged as causes to the failure of some projects: 1) unforeseen possibility of collaboration and interoperability; 2) lack of participation of stakeholders; 3) lack of training; 4) lack of policy; and 5) complexity of software.

Another observation is that, none of the studies reviewed reflect adoption to development methodology. Such methodology is an essential software engineering concept to ensure the development of quality software. Likewise, prior to the project implementation, the authors have opted to analyze existing ETDs, to benchmark recent ETD developments. However, few of the projects were evaluated after it had gone live. This process is essential, considering that a software system does not take three to five years before software problems are encountered. By doing this, the software is not only maintained but likewise, innovations may also be integrated to better meet the needs of the stakeholders.

IV. Conclusion

In this paper, the author sought to determine the goals for developing ETD, the challenges that hinder these goals and the approaches adopted in mitigating these challenges. In conclusion thereof, the findings are contributory in shaping development paradigms for ETDs, especially to developing countries whose resources are limited and studies are also scarce. Factors that affect the development of ETD in the Philippine context may likewise be studied; such as plagiarism and intellectual property concerns. In other countries, the collaboration of institutions is already realized hence the centralization of ETDs. Another study henceforth may be conducted to explore the feasibility of collaborations in the Philippine setting.

In terms of the development of ETD application, findings showed that none of the studies reviewed have presented specific methodology adopted. Hence, a comparative study may be conducted to show if the application of software engineering concept promotes quality ETD software. The methodology will be sought not only to provide a robust foundation for development but also lay mechanisms that are essential in sustaining the goals for developing ETD.

REFERENCES


Terry Ann C. Cayabyab resides in San Carlos City, Pangasinan, Philippines. Her undergraduate is B.S. in computer science from Lorma Colleges, San Fernando, La Union. Degrees in the graduate studies are masters in mathematics and masters in information technology, graduated in the years 2008 and 2010, from Don Mariano Marcos Memorial State University-South La Union Campus, Agoo, La Union and Colegio de Dagupan, Dagupan City consecutively. Currently she is writing her dissertation for the Doctorate in Information Technology program at De La Salle University, Manila, Philippines. She is a faculty member of the School of Information Technology Education at Colegio de Dagupan located at Dagupan City Philippines. Research interest on electronic theses and dissertations and have presented “Library Productivity Tool Using ETD” in Phuket, Thailand last October 2013.