Exploring Open Source solutions in the management of ETD processes

Chetan S Sonawane

Librarian, KMC College, Khopoli, Tal – Khalapur, Dist – Raigad, Maharashtra, India

Abstract

ETD which refers to the Electronic version of Theses and Dissertation is the outcome of researcher’s scholarly work at the Master and Doctoral degree level respectively. With the advancement in ICT’s (Information and Communication Technologies) almost all kinds of research material is available online. ETDs are not exception to this; digital repositories of developed and developing countries are mushrooming, which in turn will foster research. Open Source Software solutions has played a great role in providing wider access to research work by enabling Open Access. Various Open Source Software’s (OSS) are available out of this, which one to choose is the real bottleneck. Some of the remarkable solutions are DSpace, EPrints, Fedora, ETD-db etc. Most of the available ETDs are metadata driven and facilitates the search based on metadata entered. Some new Open Source solutions are being made available in recent years which include Islandora, ABCD ETD, Samvera repository (formerly Hydra) etc. The big challenge the ETDs are facing is of metadata enhancement. The Network Digital Library of Thesis and Dissertation (NDLTD) has formed an international consortium and is an organization dedicated to promotion, adoption, creation, use, dissemination, and preservation of ETDs. NDLTD adopted an interoperable ETD metadata standard (i.e. ETD-MS) in 2001 which was based on Dublin Core. There are several limitations of the metadata fields that are essential in the management process of ETDs in the widely used Open Source Software’s like DSpace, EPrints, Fedora, ETD-db and Islandora. Additional tools and platforms have emerged to support the various component of submission and distribution process. There is a need to explore adding more solutions to current genre of Open Source ETDs like discovery solutions, artificial intelligence, recommender system and any other relevant solution so as to make enable the wider dissemination of research work.

Keywords: ETD, Open Source Software’s, Metadata, Digital Repository, NDLTD

1. Introduction

ETD forms an important educational discourse in the learning process leading to the attainment of the Master and Doctoral degree. With ICT evolution, Thesis and Dissertations which were earlier available in only print form and confined accessible only to few, now became accessible to the research scholars throughout the world. The electronic form is robust and it has helped in providing wider access to all kinds of research materials. ETD initiatives have come a long way since 1987 when the first meeting to discuss the concept of a worldwide digital library of graduate student research took place\(^1\). Virginia Tech first began exploring ETDs in the early 1990’s and began accepting ETDs from 1996. They founded the Network Digital Library of
Theses and Dissertations (NDLTD), a membership organization that engages in many activities to promote the growth of ETDs. Since from its foundation, it has grown enormously with over 170 members. Online repositories of ETDs are accessible from anywhere at any time provided that the hosting Institution/University is having sufficient infrastructure. Upon automation of the submission process, ETDs are quickly available after submission. Born-digital documents provide the full text for indexing, enhancing the searching possibilities. While the practice of managing the Thesis and Dissertations (TDs) in both forms i.e. Print and Electronic still continues in many academic institutions of the World. The retrospective conversion of old TDs still continues by scanning the pages of paper TDs already submitted in the Institutions and Universities and is a daunting task with several limitations imposed like they require much more storage space, they do not easily support full text searching, they cannot be flexibly manipulated and they don’t motivate the researcher’s/author’s to learn about electronic publishing (i.e. to prepare them for electronic submission of paper’s proposals or other works now commonly required).

The advancement in ICT has changed the way of organizing, processing and disseminating information to end users. The advocacy of Open Access (OA) has made it easy to access any information in the digital environment. The educational institution have also started to deposit their content in the digital form called “Institutional repository (IR)” and made it available to users. Apart from journal article and Eprint, Electronic Theses and Dissertations (ETDs) are the most frequent document type found and consulted for use in the Digital Era. It is noteworthy to note that making ETDs freely available to users, has clear benefit to student authors, since readability of their researches is enhanced and they become much more visible in the research community. What’s more important is open access to ETDs is of help to universities, since it increases the awareness of their research activities around the globe. During the last decade (2005-2015), the growth of ETDs globally has been increased exponentially among universities and other organizations. Moreover, there is an increase in both number of ETDs initiatives and ETDs collections. Open access and open source technologies helped ETDs and other research material repositories to grow rapidly. The Open source repositories are metadata driven and facilitate the search based on their own or through third party search engines like discovery system.

An electronic thesis and dissertation (ETD) system provide all necessary management services throughout the lifecycle of a student’s thesis or dissertation, from initial submittal to the thesis office, through the iterative review and approval process, to the final publication in a digital repository. ETD is primarily sought to ease the submission, storage, dissemination and the cost of handling theses and dissertations. This is the primary reason for ETD development, especially to developed countries. In the developing countries however, there are a few isolated initiatives, which are, if not in its infancy stage, rather slow to meet stakeholder requirements. Nowadays Institutions/Universities work in collaborations to have better access to ETDs. Aimed towards facilitating the research conduct and enhancing the productivity of scholars, ETD perspective
have expanded towards interoperability, open access and global visibility\textsuperscript{7}. The availability of various types of software solutions for managing IRs has hampered selection of appropriate solution and has rather become a time-consuming process for institutions. The process of software selection for sustainable system is multi-dimensional. Selection of an ETD software solution for an IR that maintains a diverse range of scholar documents (report, articles, journals etc) is not just based on its ability to meet the ETD-specific requirement, but the other characteristics related to the interoperability of system\textsuperscript{8}. There are more than 15 OSS available for managing IR. While NDLTD has formed an international consortium, ETD repositories operate through a consortium model established for this purpose in different countries either through state level collaboration or through an establishment of an apex body looking after higher education of a particular country.

With the development of the digital collections, new descriptive metadata standards began to appear – these include Metadata Object Description Schema (MODS), Dublin Core (DC), Metadata Encoding and Transmission Standard (METS) and Learning Object Metadata (LOM)\textsuperscript{9}. Today the metadata enhancement is one of the big challenges in the context of ETDs\textsuperscript{8}. Dublin Core (DC) metadata standard which was created in 1995 in Dublin, Ohio became very popular with the development of the web based digital collection, but it does not describe some essential pieces of information that pertains to dissertations. That is why the ETD-MS was created\textsuperscript{10}. NDLTD assumes an international leadership role in the ETD initiatives. NDLTD was the developer of the ETD Metadata Standard (i.e. ETD-MS) in 2001 which was based on DC standard. ETD-MS setup a guideline for cataloguing ETDs because the standard metadata set is tailored to capture such information as committee members (advisors), degree names, and degree level that are part of ETDs. NDLTD also encode that standard for cross walking with the MARC-21 standard and XML schema as well\textsuperscript{11}.

There are at least 5 open-source software solutions for ETD submission management. They include TAPIR, Valet, OpenETD, Vireo and Weller ETD Project (formerly Jarrow)\textsuperscript{12}. ETDs distribution platform includes IRs of Universities / Institutions, Open ETD portals like DART Europe, NDLTD and a commercial player PQDT Global – ProQuest Dissertation and Theses offers a powerful distribution platform and compliments Open Access dissemination\textsuperscript{13}. There has been growing proliferation of ETD distribution platforms on the Internet and still difficulties are faced by the students in searching for and retrieving these works of scholarship dispersed across multiple sites online. Growing trend can be seen towards dual submission of ETDs to both the University repository and to the ProQuest system. Both system offer functional ETD management and Internet publishing system with some degree of support with submission, public distribution, collection management, discovery, access control, archiving and technical/user support\textsuperscript{14}. Though the users sometimes face difficulties with regard to efficient retrieval of useful resources and hence there is a need to explore considering more add-on solutions.
The success of a library depends on the efficient discovery of library resources. Influenced by the Google phenomenon, library users expect an easy, user-friendly, Google-like search interface and facilities from the library. Discovery services provide some kind of solution for this. Many proprietary and Open Source Web-Scale discovery tools are available in the domain of library discovery, considering the scope of this paper open source software solutions will be explored for applicability. Among the emerging technologies Artificial Intelligence (AI) heads the list, and is poised to amplify the utility and reach of library services. AI can personalize the library experience for patrons, connecting them more efficiently to resources that best align with their goal. The relevance of AI for the retrieval of most useful document will be reviewed by means of an Open Source Artificial Intelligence Markup Language (AIML) as a set of eXtensible Markup Language (XML), which is able to relate and represent expression in natural language.

Another open source solution proposed for the Open Source ETD is a Recommender system. Recommender systems are software tools which provide suggestion for items that a user may be interested in. The most suitable Recommender System identified for ETD is Apache Mahout.

2. Selection of Open Source Solution for ETD Management

The management of Electronic Theses and Dissertations (ETDs) raises a number of issues concerning the process of ETD creation, ingestion, access, archiving and preservation. Digital repository system providers generally provide a platform for ETD management, including function for ETD submission, ingestion, dissemination and retrieval. There are a number of Digital Library software’s available as “Open Source” as well as “Proprietary”. The Open Source Software (OSS) helps the libraries especially in eliminating vendor lock-in and reducing the cost considerably with greater flexibility. These software are generally available free of cost with capability to capture, store, index, preserve and redistribute all scholarly research materials in digital formats. Libraries in the present era are facing the problem of insufficient budget for experimentation and implementation of technology and also procurement of resources. To manage the information resources, there is a necessity of software and tools. This situation necessitates the libraries to stand up alone with the latest technology. Most of the software’s being used in the libraries are proprietary’s which means there is an involvement of cost without the actual source code availability. On the contrary, OSS’s are just opposite which revolves around sharing and collaboration. By nature, OSS is free for anyone to copy, modify and redistribute which increases the possibility of being used as per the requirements.

At present there are a number of OSS solutions promoted to manage repositories system, and most of them can also handle ETDs. Due to the different nature and complexities of these packages, there are limited guidelines, which can be used by academic institutions to choose the most appropriate application for their ETD management system implementation. Nowadays there is a broad consensus on the vital importance of openness and dissemination of scientific information resources over the web. It makes sense when following the ethos of Open Source movement, to simultaneously endorse Open Access (OA) movement. Both serve the same objectives with open source concentrating, as the name suggest, on access to source code for the
software and OA on the access to content. The irony of choosing a closed source, or proprietary package to achieve an OA objective can hardly be overlooked, although special programming support may be required to implement OSS\textsuperscript{19}. In order to secure the stability of an open-source based repository system, prior to the software selection, a manager needs to be ensured about the sufficiency of experience of the related staff in terms of programming and database management skills for the selected platform. Furthermore, to achieve the best performance, the selected software should be compatible with institution’s infrastructure and computer environment (such as operating system platform, server type, security policies, etc). In order to select the most compatible software, each solution should be checked from the platform-related aspect such as: Compatible operating systems, Compatible databases, development languages and hardware\textsuperscript{8}. One of the major challenges of OSS generally is its frequent updates and upgrading the old system is another big challenge, especially if it is managed by in-house technical man powers\textsuperscript{18}.

3. Available Open Source Solutions for ETD Management

Libraries have a range of choices of open source repository software for ETD management. These includes ABCD ETD, Archimede, ARNO, DAITSS, Dienst, DSpace, EPrints, ETD-db, Fedora, Greenstone, i-Tors, Invenio, IR-Plus, Islandora, MyCoRe, Samvera etc\textsuperscript{18}. While choosing to build a repository based on open-source software offers many opportunities for development and customization, it also comes with challenges. Aside from the time and technology cost required to get the repository from day one of development to a full production instance, there are also important ongoing workflow considerations. With a hosted repository platform, the library pays for customer support as part of the annual maintenance fees. With open source, there are online communities of developers using the same platform who can offer advice, but bug squashing may definitely be more challenging\textsuperscript{20}. The most widely used repositories are DSpace, EPrints, Fedora and ETD-db, the newly available open source solutions which have caught attention are ABCD ETD repository, Samvera repository (formerly Hydra) and Islandora repository. In the following sections review of the most widely used open source repository software viz. DSpace, EPrints, Fedora and ETD-db and the relatively new and emerging software’s like ABCD ETD repository, Samvera repository and Islandora repository is presented.

3.1 DSpace

DSpace is an open source digital repository system that capture, store, indexes, preserve, and redistribute an organization’s research output in digital form. As institutional repository software DSpace is making its mark, with an increasing number of institutions around the globe installing, evaluating, it for managing their digital assets. DSpace provides a long-term solution for physical storage and management of digital items in a secure, professionally managed repository. It is adaptable to different community needs with different workflows, different levels of access control mechanism and content moderation procedure\textsuperscript{21}. The first public version of DSpace was released in November 2002, as a joint effort between developers from MIT (Massachusetts
Institute of Technology) and HP (Hewlett Packard) Labs. Following the first user group meeting on March 2004, a group of interested institutions formed the DSpace Federation, which determined the governance of future software development by adopting the Apache Foundation’s community development model as well as establishing the DSpace committer group. In July 2007 as the DSpace user community grew larger, HP and MIT jointly formed the DSpace foundation, a not-for-profit organization that provided leadership and support.

DSpace is the first open source digital repository system to tackle the complex problem of how to accommodate the differing submission workflows needed for a multidisciplinary system. DSpace has a comprehensive and customizable Dublin Core metadata collection for describing items individually. The authorization processes from the registration to submission and archival makes DSpace unique from other software. It has a strong and flexible administrative and security features like email/password based authentication, email notifications in different workflow steps and persistent identifiers (handle) for each item archived. DSpace provides both simple and advanced search and browse features. It supports full text search and thumbnail display of images in search results. DSpace is compliant to OAI-PMH (Open Archive Initiative - Protocol for Metadata Harvesting). DSpace is written using Java programming language with a built in repository based either on PostgreSQL or ORACLE as backend database with two web interfaces JSPUI and XMLUI. DSpace offers an interface that can be extensively customized using XML, and a standard organizational scheme (Communities are made up of collections are made up of sub-collections are made up of objects). It also supports OAI-PMH v2.0 to export METS. The latest version of DSpace support faceted search and browsing functionality under Apache Solr. The DSpace code is currently licensed under BSD open source license.

3.2 EPrints

EPrints is general digital repository management software that was developed by the University of Southampton in 2002 and is released under GPL license. It is open source software suitable for creating digital repositories compliant with OAI-PMH with Perl based plug-in architecture for importing and exporting data and converting objects (for search engine indexing). Its developer goal was creating a flexible platform for building high quality and scalable repositories. EPrints is a Perl based IR solution, which is easy to get up and running, and provides an excellent discovery and submission interface. However additional configuration beyond the standard configuration is more challenging. The configuration files are written in Perl script. There is good documentation provided, but good knowledge of Perl is required to administer the system. Eprints also uses a custom built repository, but support the standard simple Web-service offering Repository deposit or SWORD interchange protocol, meaning that digital object can be extracted and shared.

3.3 Fedora
Fedora (Flexible Extensible Digital Object Repository Architecture) is a robust, modular, open source repository system for the management and dissemination of digital content for access and preservation. It can be used for complex and very large digital collection such as historic and cultural material as well as scientific data. It is capable of archiving different kinds of content and its metadata with various storage options. The information can be access via web API and web based administrator GUI. It also supports and provides RDF search, OAI-PMH, GSearch (Fulltext) and rebuilds utilities with multiple and customer driven front-ends. Fedora is distributed under the terms of the Apache 2.0 open source license. Fedora is written in Java, tested under Windows, Linux and Mac OS. Fedora modular architecture is built on the principle that interoperability and extensibility is best achieved by providing a limited set of stable, standard-based repository services and common patterns for integrating with other best-practice systems and applications. These services are provided via RESTful APIs in accordance with modern web standards. Fedora provides a foundation upon which many types of repository frameworks can be built, including the popular Samvera and Islandora projects. In May 2009 collaboration on related projects and growing synergies between the DSpace Foundation and Fedora commons organization led to the joining of the two organizations to pursue their common mission in a not-for-a profit called DuraSpace. Currently the DSpace software and Fedora software is stewarded by DuraSpace.

3.4 ETD-db

The ETD-db is open source software written using Perl script with MySQL database developed by VirginiaTech. This software provide a standard interface for web users, researchers and library professionals to manage the various fields and metadata related to collection of ETD. The software is currently maintained by Paul Mather, and is available free of charge to members of the NDLTD. The software is developed by Anthony Atkins and is endorsed by NDLTD. Currently this is the most widely used E-theses package in use, in part due to the support it has from the NDLTD. Despite this, there is currently little directional development, with some institution choosing to install the “out-of-the-box” version, while others make their own changes to the system, which are not made available to the general community. Perl which is native to most of the Linux and UNIX install, and MySQL is also very common. In addition to the standard Perl installation, it is also necessary to install additional ‘Perl modules’ which enhances the functionality of language, but it requires the efforts of the reasonably experienced system administrators to do the prerequisite installation. There is no formal software license applied to the module.

4. Innovative solutions for ETD Management

There is no single best system for ETD management overall. Rather, it is up to decision makers at each institution to choose an approach that best fits their university’s values, goals, and needs. There is a need for a single portal for ETDs that allow for search and discovery of these unique works of scholarship wherever the full text resides. Therefore further investigation into possible
solutions for such an ETD portal would be a boon not only to universities and ETD authors, but to the diverse researchers, students, professionals, and interested citizenry who could benefit from easier access to this growing corpus of knowledge\textsuperscript{14}. The most widely used system; Virginia Tech’s ETD-db, was developed in December 1999 and is showing its age. Unfortunately ETD-db has not had a major release since 2004\textsuperscript{12}. Other systems are being used to manage ETDs, most notably DSpace, but DSpace’s workflow is based on a pre-print self archiving model, not a thesis model. Its inflexible workflows are not tailored for ETD submission. The open source solutions developed for submission namely TAPIR as well as open source solutions developed like ETD-db lacks key features that are necessary for managing future ETD collections. The time is ripe for the development of the next-generation ETD system that is capable of managing the entire ETD workflow\textsuperscript{2}. Some new solutions have emerged which has drawn considerable attention in recent years, which includes Islandora, Samvera (formerly Hydra) and ABCD ETD solution which is reviewed in following sections:

4.1 Islandora

Islandora is a new entrant in the institutional repository software market\textsuperscript{12}. Islandora is an open source software framework that focuses on collaborative management, and discovery of digital objects, and provides a comprehensive and interoperable starting point for the implementation of a graphical interface to a Fedora repository\textsuperscript{25}. Islandora ecosystem is comprised of three major components: Drupal, Islandora and Fedora. Drupal, as the front end of the system, offers an array of collaborative tools and applications for presentation. Fedora, at the back-end of the system, provides a data store within which digital assets are managed. Islandora, the glue which holds the system together, facilitates the communication and messaging services between the other parts of the ecosystem. The core design principle is the separation of data from presentation in order to leverage the strength of individual component and make extensive use of other OSS tools. The most commonly integrated components are Djatoka, an open source JPEG image viewer, and ffmpeg, a video conversion utility that allows for playback of different video file types. To communicate with Fedora, Islandora uses the Representational State Transfer (REST) architecture to retrieve content from the database in fedora. REST is a commonly used methodology for computer networking to send or request data between systems. To manipulate and prepare the data for presentation in Drupal, Islandora makes use of Drupal Hook system. Hooks are code snippets providing functionality for content models that contain information about data objects\textsuperscript{26}. Islandora is developed by the University of Prince Edward Island in Canada and supported by an independent company, Discovery Garden. As Islandora is built on top of Drupal, all of the necessary content management system (CMS) and application framework component are included. As an example, any existing Drupal theme will automatically work with Islandora because it uses standard Drupal components. Furthermore, by leveraging such a well-established and sophisticated CMS like Drupal, Islandora makes it much easier to organize repository collection for public consumption in ways that may be much more challenging for DSpace or EPrints.
Fedora is a well-established open source repository for organizing digital objects. As the DuraSpace foundation has stewarded both DSpace and Fedora, some software designed for DSpace also works with Islandora. For example, DuraCloud, an offsite backup and preservation module for digital object can be used with Islandora. Furthermore, Discovery Garden has released a toolkit written in PHP name “tuque” which allow developers to ingest objects into a Fedora repository independently of Islandora. This indicates that any module built with tuque could potentially export into a DSpace Institutional Repository, or conceivably an EPrint one as well, if tuque library was expanded^12.

4.2 Samvera (formerly Hydra)

Samvera is an open source, multi-institutional project that gives institution a framework to build and deploy robust and durable digital repositories (the “body”) supporting multiple “heads”: feature-rich, digital asset management application with tailored workflows. Samvera provides both a technical solution and software as well as a vibrant community infrastructure, giving like-minded institution the ability to collaborate and realize the benefit of pooled development. Samvera as a repository solution provide a versatile and feature-rich environment for end-users and repository administers alike. As a technical framework, it provides a toolkit of reusable components that can be combined and configured in different arrays to meet a diversity of content management needs. As a community framework, Samvera provides like-minded institutions with the mechanism to combine their individual development efforts, resources and priorities into a collective solution with breadth and depth that exceed the capacity of any single institution to create maintain or enhance on its own. Samvera software is free and open source and available under Apache 2 license^27.

A Samvera repository solution provides functionality for the full range of CRUD services (create, read, update and delete). To do this it employs a number of Ruby on Rails based components (or “gems”) in conjunction with three other “best of breed” open source software products: Fedora, Solr and Blacklight. The four major components on which Samvera is based works in a following manner: Fedora repository software provides a robust, durable repository layer for persisting and managing digital objects, Solr indexes provides fast access to information about an institution’s resources, Blacklight, a Ruby on Rail plugin that sits above Solr and provides faceted searching, browsing and tailored views on objects, Samvera gems the Ruby on Rails components that integrate the building block to form a complete, flexible and extensible digital repository solution^28.

4.3 ABCD repository solution

ABCD software is fully based on ISIS technology, some of the technical characteristics which has made ABCD repository software suitable for ETD repository application are Freedom of database structure, ETD database from the existing standard formats, ETD database using
CEPAL database for full text access, Full text indexing, The use of non Latin script, Integrated end user search etc. The characteristics are discussed in short below:

*Freedom of database structure*

ISIS records carry their individual structural description as a “header” within themselves, unlike in a relational “table” – based databases in which all records in the same table share the same structure by necessity, and therefore each record can have its own, different structure, making it a “document” rather than a record. In fact for most record-related operation in ISIS, there is no need to formally describe the structure even, so one could consider ISIS as “scheme-less” record. As a result of this, ISIS accepts any structure – definition tool, and so does ABCD.

ABCD offers flexibility to create full text database of ETD using existing MARC 21 format, CEPAL, Dublin Core or users can create new databases from scratch using ISIS field definition tables and its powerful formatting language. This is rather unique feature of this software. Not much other software provides such flexibility for database management. It supports all kinds of standards and non-standards.

*ETD database from the existing standard formats*

ABCD repository solution permits to use MARC-21 (or UNIMARC) cataloguing formats, LILACS, AGRIS, CEPAL. Other standard available are Dublin Core, METS, Z39.50, ISAD/G and – from version 2 on – Unicode.

*ETD database using CEPAL database for full text access*

With CEPAL database once can offer full text access with ETD database. In the object field where upload of files/chapters, content pages, full text PDFs, cover pages, images of theses and dissertations can be uploaded by means of a field provided in the database. The image and PDF files are stored in management-defined folders. ABCD version 2 file-upload dialog allows storing the files in hierarchy of folders, starting from a “root” point defined by a system manager.

*Full-text indexing*

ISIS had capabilities to extract individual words, except for non meaningful words defined in a “stopwords”-list from any field of record in the database. As ABCD is based on ISIS for data-storage and – retrieval on the one hand, and PHP for the web-interface creation on the other hand, some nice PHP tools like FckEditor offers full HTML editor which can be embedded into a cataloguing form to create full documents.

*The use of non-latinscripts*

ABCD as a web based solution can use non-latin script (e.g. Amharic, Chinese, Greek etc) as part of web-browser capabilities.
Integrated end user search

Any of the created databases can be added to the central part of the ABCD “Site”, which provides links (as a portal) relevant to the library’s user and a meta-search option to search over a list of databases. The result are subsequently presented in a result-page with links to the hits for each individual database, which leads to the OPAC display (with options like reservation or online ordering) of that database.

5. Open Source ETD Submissions solutions

There are at least 5 open-source software solutions for ETD submission management. They include TAPIR, Valet, OpenETD, Vireo and Weller Project (formerly Jarrow). There have been many attempts to create ETD systems, but most of them failed to gain momentum. Some submission tools were developed as add-on solution to existing repository solutions. They’re developed as a plug-in to the existing repository solution so as to integrate it into the repository solutions. There were some years when no absolute activity took place with regard to development of open source ETD submission tools, but in recent years tools like OpenETD and Vireo were developed which provided some kind of solutions, but there is still no software that can handle the entirety of the submission process.

5.1 TAPIR

Theses Alive Plug-in for Institutional Repositories or TAPIR which was developed by University of Edinburgh in 2003 was developed as a software package based on DSpace. It is available under a 3 clause “modified BSD license”. The software focuses heavily on collaboration tools designed to assist students and faculty with the editing and revision of theses. Today, collaboration tools for editing documents are widely available from Word’s track changes to Adobe Reader’s tool for PDF document markup. As such TAPIR’s innovative focus is less relevant.

TAPIR was developed as an add-on solution to DSpace which was written in Java; however it was created long time back in 2003 and then after it did not provided new versions of the software compatible with new versions of DSpace. It was created for DSpace 1.2. Parts of the DSpace code were modified with TAPIR to deliver some of its core functionality. Currently DSpace 6.0 version is available and the TAPIR is unsuitable for it, this only makes TAPIR an unlikely candidate for adoption.

5.2 Valet

VTLS (Visionary Technology in Library Solutions) inc. which developed an open source ETD submission module named Valet and publicly released it in 2005 is available under Mozilla Public License 1.1. Valet is a Perl based module but coupled with a Fedora digital repository. The product is best suited when deployed in conjunction with VTLS proprietary institutional
repository, VITAL. There has been no update since its initial release in 2005, there is also very little documentation available which resulted in making the installation and use difficult.

ARROW (Australian Research Repositories Online to the World) Project which was released during the summer of 2008, led by Monash University in Australia forked Valet when they created a Java version called Squire, which is also open source. Squire release does not provide any kind of documentation nor got it updated for the last nine years which makes it useless.

5.3 OpenETD

In 2010 Rutgers University developed OpenETD software as an add-on solution as well as standalone submission system and is available under GNU General Public License 3.0. This system is repository agnostic. The software exports the theses and dissertations in METS/XML compliant files for import into the desired repository solution. This software does not impose any kind of restriction regarding the choice of a particular repository solution. OpenETD is based on Linux, MySQL, PHP this ETD solution utilizes better web application architecture than its predecessor. It provides features like email notification, automatic margin and page number detection, along with simple use management facility, this system is quite compelling. OpenETD is based on post defense process. The software also offers very limited ability to customize forms or the logic around a submission process.

5.4 Vireo

Vireo is a Java based DSpace plugin developed by Texas Digital Libraries, and is available under a 3 part “modified BSD license”. Vireo is an online system for the submission and management of ETDs. Vireo was released in 2011 and is under active development. The software was developed to integrate into DSpace, for ease of use and ability to make simple system customization has led to an active user community that is beginning to grow outside of Texas area. Vireo offers many of its features as like that of OpenETD does, such as automatic email notification and templates. Although Vireo handles more of the process than any other system, beginning immediately after defense and handling corrections in addition to basic ingesting, it does not handle any pre-defense procedures, such as scheduling or initial committee approval.

5.5 Weller ETD Project

Weller ETD Project (formerly known as Jarrow) software was designed to integrate with Islandora, an institutional repository solution built on top of Fedora and Drupal. The Weller ETD Project strives to assist scholars and institutions in the publication and dissemination of their work. This ETD module is designed to serve two complementary goals: to collect information about the theses from people involved and to manage the process preparing, approving, submitting, accepting and publishing the theses, dissertation or projects. A sub goal is to manage the organization of the defense, as it is so tightly coupled to the submission process.
The workflow management system is broken into three components, which work together to allow for wildly different workflow to be configured. The three components are Data Model, the Form Editor and the Workflow Manager. Each component is dependent on the component before it. The Weller ETD Project began in early 2012 at University of North British Columbia (UNBC) as a joint initiative of the Library and the office of the Dean of Graduate Programs. The Project was last modified in 2013, and is actively maintained and is under active development.

6. Metadata issues in ETD management processes

NDLTD developed an ETD specific item-level metadata schema ETD-MS, based on Dublin Core, is a primarily descriptive schema; it does not include element for life cycle management event or structural relationships. Most of the existing metadata standards fail to provide for the unique properties of ETDs. Most of the metadata standards have fail to provide the level of specificity required to sufficiently describe the item. ETD-MS published by NDLTD provides low barrier to entry and broad application, which focuses on repository interoperability.

Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility. To describe digital resources accurately, metadata creators try to follow, as closely as possible, the thinking of the creator/author and also to anticipate what users might want to discover and how they’ll retrieve the information. As noted by many researchers, one of the key issues for information retrieval and all other content-based text management application is document indexing. The generation of accurate indexing term is fundamental to the discovery, use, and reuse of digital resources. Sometimes in addition to metadata created by professionals, incorporating author generated keywords can expand descriptive metadata. Some libraries (like Virginia Tech which were earlier ETD adopters) include only author-created metadata. But viewing the growing interdisciplinary nature of subjects and importance of specific topics, the author supplied keywords may not be sufficient.

Most of the digital repository management software (e.g. DSpace, Eprints, Fedora, and Islandora) supports simple Dublin Core and Qualified Dublin Core by default. “Simple Dublin Core” is DC metadata that uses no qualifiers. It applies only 15 elements without any qualifiers. On the other hand, “Qualified Dublin Core” uses additional qualifiers to increase specificity or precision of the metadata. For example, a “Date” is a DC element which may be specified to identify a particular kind of date (date of last modification, date of publication etc). The Dublin Core Metadata Element Set (DCMES) or Dublin Core consist of small set of resource description categories which is notably different from many of the other metadata schemas due to its ease of use and interoperability. As DCMES consist of 15 basic elements only it has several limitations. These fields, being generic for any kind of digital resource, do not capture any specific information about specialized content such as maps, images, video, objects, learning materials, etds etc. A generic metadata schema is not sufficient enough to describe different types of resources with all relevant elements.
The diversity of disciplines and formats of content in repositories makes it difficult to use a blanket metadata format and allow only high-level controlled vocabularies to be used across the repositories. Repositories should be allowed to select an appropriate metadata schema and appropriate controlled vocabularies at the community or collection level, instead of continuing with existing metadata and descriptive practices into a single Dublin Core (DC) schema. Standards schema such as MODS and Visual Resource Association VRA Core are widely supported and provide more granular description than simple DC. Integration of these metadata standards into repository environment would enable batch ingestion of a wider variety of materials without having to write additional transformational programs. This will help each repository community to choose a metadata standard that better fits its need and content.

Along with providing more options for descriptive metadata, repository should support preservation, structural and right management. DSpace is most widely used repository solution, but it currently provides a limited amount of preservation metadata in the form of filetype identification and a checksum for each bitstream submitted. There is also provenance data captured for each event such as submission, approval and edited metadata. All of this data is created and captured by DSpace itself, but there is strong need to make use of additional user-supplied metadata. Preservation metadata enhances the library’s ability to manage activities related to digital item’s format, authenticity, and stability over time. Add-on solutions for ETDs

Some of the add-on solutions which can be proved useful for ETDs and should be considered for adoption include Discovery System, Artificial Intelligence (AI) and Recommender System. Discovery Service or Web-Scale Discovery Service is now common in higher academic and research libraries. The goal of developing Discovery Service is to help library users in discovering local ETDs and Global ETDs from a single search box and make library research as intuitive as Google. Many open source tools are available which includes Blacklight, VuFind, eXtensible Catalogue (XC) and Franklin. The underutilization of ETDs pose a serious challenge for research and development, the lack of discoverability is the reason behind it. To overcome this challenge the applicability of discovery service will be explored in following section. AI applications to ETD repository will be reviewed. The Recommender System which generate meaningful recommendation to a collection of users for items or product that might interest them is based on algorithms and is widely used by e-commerce websites. The applicability of Recommender System for Open Source repositories will be reviewed in the following section.

7. Discovery Solution

A Discovery interface includes features such as relevancy-based search results, faceted navigation, and other features consistent with web-based resources. Discovery interfaces have multiple functionalities such as end-user interface, usually deliverable via a Web browser, to perform task such as presentation of search box for end-user queries, presents advanced query
options and presentation of search results listed either in a brief form or in full-record displays; interoperability with a link resolver to present links to full-text from citation records in search results; local search and retrieval, usually through an integrated indexing; ability to interactively communicate with the library’s ILS (Integrated Library System) implementation and provide access to remote index platforms. VuFind which is widely used, widely downloaded, regularly updated, and which is having active user forum is a good discovery solution for adoption. VuFind is based on AMP architecture and uses Apache-Solr as its default text retrieval engine. It supports OAI/PMH version 2.0 as its standard for harvesting and includes support for MARC and XML format. Apache Tika could help in refining the search result, its toolkits which extracts and extend metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, which makes Tika useful for search engine indexing, content analysis, translation, and much more. VuFind system offers supports for loading records from specific repository system like DSpace, EPrints and Greenstone Digital Library (GSDL). In Samvera repository Blacklight discovery tool is integrated. The globally available ETD repositories can also be integrated with local repository solution for wider dissemination of research by means of discovery tools.

7.2 Artificial Intelligence (AI)

Artificial Intelligence (AI) which is still in its nascent stage and the time to adoption horizon expected is next four to five years. AI encompasses the 5 broad general area of research which includes (1) Expert System (2) Natural Language Processing (3) Fuzzy Logic (4) Robotics and (5) Neural networks. The AIML language which is a set of XML element capable of representing and linking natural language expressions, allows the creation of engine capable of maintaining simple dialogue. AI agents such as chatterbots and location-based services are shifting the focus of librarian, while also optimizing search engine result to increase student access. Some findings indicate that AI agent can alleviate the burden on librarian answering technical questions and free up time to increase focus on other duties, such as teaching and improving skills. AI agent can also assist students using large databases by locating relevant resources based on personalized data. For example, an agent can take into consideration students past course experiences to locate article that align with their knowledge level. The AIML language which is an open source can help in enabling the use in research for improvement in creation of robotic engines able to recognize information written in natural languages. AIML is an XML compliant language which makes it possible to customize an Alicebot or creating one from scratch from beginning, which can be created for DSpace using XML. Development of metadata parser software of digital repositories for DSpace platform which will convert the input from XML to AIML is necessary. There are 20 or so more additional tags often found in AIML files and it is possible to create your own so-called “custom predicate”.

The Center University Euripides of Marilia – UNIVEM, called Univem Aberto has created metadata parser software XML2AIML. The DSpace platform enables the digital repository web server to receive request from external media, off the server. With request, we are able to have
access to all digital files that are submitted to the repository. The request made to the web server is parsed by the repository and all files that fit the information passed by DC will an XML response that will be sent to the XML2AIML software. XML2AIML software can do search request to the server of the repositories, the metadata already created for the standard DSpace platform will be used. AIML language has several limitations in creating robotic conversational engine, limiting the number of questions these robotic engines can be asked.

7.3 Recommender System

The goal of Recommender System is to generate meaningful recommendations to a collection of users for items or products that might interest them. Some of the examples of users of Recommender System are Amazon and Netflix. Suggestions for books on Amazon, or movies on Netflix, are real world example of the operation of industry-strength recommender system. The design of such recommendation engine depends on the domain and the particular characteristic available. Recommender systems are increasingly used on Web to help users find material relevant to their interest. Recommender systems have been proved useful by commercial establishments to boost sale, there have been less recommenders used in research. From a user point of view, recommenders are either personalized, i.e. recommendation are based on the knowledge of the user’s preference or past activity, or non-personalized, recommendation are the same for all users. There are two important classes of recommendation: collaborative filtering (CF) and content-based filtering (CBF). CF makes use of past interaction data to recommend new known item to a users based on the assumption of similarity (e.g. to other users in the case of CF). These technologies have proven extremely effective in supporting exploratory browsing, but for effective functioning it depends on the data generated by users. CF can not recommend any items no user has referred and is known as cold-item.

Another approach to build recommender system for repositories is to make use of CBF. This method attempts to recommend items using the assumption of content similarity, such as cosine, based on attribute (feature) of each item. Hybrid method combines both CF and CBF. The integration of the recommender in EPrints repositories is especially easy as it is provided on EPrint Bazzar Plugin (http://bazaar.eprints.org/466/) and is based on CBF. For DSpace Apache Mahout is an ideal recommender system based on CF. Apache Mahout is an open source suite of machine learning libraries designed to be scalable and robust. For DSpace a CF based recommender System that will rely on users rating is most suitable. In order to implement recommender system in DSpace there are a number of steps that will need to be completed which includes: (1) Creating a table in DSpace database in order to store user’s recommendation (2) Creating a way to manipulate this table with basic operations CRUD (3) Create a recommender system with the help of Mahout (4) Adopting a method for adding or updating a rating on DSpace item and provide a list of recommended item on item display page by modifying User Interface (UI). Although there is an alternative of CBF, CF recommenders are known for their better results in most cases. Hybrid approach can be applied by combing the results produced by CF and CBF and then after merging its result to produce a final list.
8. Conclusion

Although Libraries invest huge amounts in building their collections and platforms to host various resources including ETDs, sometimes theses resources get under-utilized due to lack of powerful search tools and repository solutions inability in retrieving useful resources. Some of the solutions developed by Institutions / Universities to support ETD management have not updated themselves on a regular basis and thus became irrelevant for adoption. Whereas the add-on solutions mentioned in this paper are created by commercial establishments and are regularly updating and providing innovative approaches and hence should be considered for integration into ETD repository.
References


21. P Jobish, etd@IISc: A DSpace-based ETD-MS and OAI compliant Theses Repository Service of Indian Institute of Science, [http://docs.ndlted.org/dspace/bitstream/2340/229/1/049Mini.pdf](http://docs.ndlted.org/dspace/bitstream/2340/229/1/049Mini.pdf)


