Part I

Background Information, Current Scenario, Major Issues & Data Standards

Introduction

1.1 Theses and dissertations are known to be the rich and unique source of information, often the only source for research work that does not find its way into various publication channels. Doctoral dissertations are manifestation of result of four to five years of intense work involving huge investment of resources, both mental and physical and infrastructure and other support from the universities. A thesis reflects quality of research work conducted by a student and the ability of an institution to lead and support original work of research in a given discipline. Research is characterized by originality, improvements and innovations. Scientific research, in particular, is cumulative in nature. The present research is built upon the past knowledge. The first and foremost step in research is, therefore "literature search" on the given topic. Scholarly communication system have evolved mechanisms to index past and current research publications through subject-specific bibliographic databases, such as MEDLARS in medical sciences, COMPENDEX in engineering and technology and INSPEC in physics, electrical and electronic engineering. Most of these bibliographic databases index research articles from national and international journals, conference proceedings and chapters in books. Most databases, baring one (Dissertation Abstracts International from Proquest) do not index theses and dissertations. Even the coverage of Dissertation Abstracts International is selective to those universities who have signed-up with Proquest. It does not contain bibliographic details of dissertations submitted to universities in India.

Why ETD?

1.2 The process of scrutiny, validation and approval of doctoral dissertations is confined to few experts (identified by the university on recommendation of theses supervisor). It is not open to the scientific community at large, and therefore, quality is sacrificed. The theses collection in most of the Indian libraries, are kept in closed access, making it difficult for other students to access them. It remains an un-tapped and under-utilized asset, leading to unnecessary duplication and repetition that, in effect, is the antitheses of research and wastage of huge resources, both human and financial. The UGC's Regulatory Framework aims at evolving a mechanism to improve the quality, accessibility and availability of Indian theses and to implement uniform standards for creating metadata of doctoral theses and a system for collecting and collating this standardized data.

1.3 Electronic version of theses provides broader exposure to research students through greater accessibility. It offers opportunities to use new forms of creative scholarship through use of interactive elements, multimedia, hyperlinks, etc. It provides opportunities to research students for professional development as they learn the basic skills of scholarly publishing in electronic format. It prepares them for their future career and lead to the most expressive rendering of their discoveries and ideas. In other words, ETDs are new genre of documents that are being continuously redefined with evolution of e-publishing technology. It is expected that better type of ETDs may develop with development in e-publishing technology and their adaptation. While the simplest form of

ETDs can be thought of as "electronic paper" where the underlying authoring goal is to produce a paper form with diagrams and images in colour. Richer ETD may include links, hyperlinks from table of contents, tables, figures and indexes that are targeted to locations in body of the document. More evolved ETDs may include multimedia contents following international standards, and files including data and interactive or dynamic forms that may be harder to transform into print media.

1.4 Implementation of ETD would lead to streamline of workflow and save time and labour as checking of submissions and cataloguing of ETD would be faster, moving and handling of paper copy is eliminated and delay in binding is removed.

Present Scenario in India

1.5 There are a number of agencies that are involved in collection, compilation and presentation of metadata of theses in India. Some of the major initiatives are as follows:

> *Association of Indian Universities (AIU) publishes list of theses awarded in various universities in its weekly publication called "University News". The AIU has also published a number of bibliographies on theses submitted to the Indian universities in various subject disciplines. However, coverage of University News and subject-specific bibliographies of theses published by the AIU is far from complete.*
>
> *The INFLIBNET and DELNET host databases of bibliographic records of Ph.D. theses submitted to various universities in India consisting of 1,40,000 and 3,953 bibliographic records respectively. Vidhyanidhi, a nation-wide effort on theses and dissertation, currently supported by the Ford Foundation and Microsoft India, hosts more than 500 theses in full-text and 85,000 bibliographic records of theses submitted to the universities in India. Vidyanidhi is a member of the Networked Digital Library of Theses and Dissertations (NDLTD), a global initiative with more than 174 members from different countries of the world. ETD@iisc is another initiative that hosts about 86 theses. It provides guidelines for submission, document conversion guidance, theses templates and sets the workflow for online theses submission.The National Social Science Documentation Centre (NASSDOC) a wing of the Indian Council of Social Science Research (ICSSR), hosts a library for the social scientists with Ph.D. theses in social sciences as its core collection. As a national institution, NASSDOC aims to build a truly representative collection of Ph.D. theses in each of the areas that make up the social sciences. The NASSDOC systematically acquires Ph.D. theses in social sciences submitted to the Indian universities. Currently, the NASSDOC has 4924 Ph.D. theses in social sciences in its collection.*

1.6 In spite of a number of sporadic efforts mentioned above, India neither has a comprehensive and authentic source of information nor a mechanism to obtain information on all Ph.D. theses submitted to the universities in India. The situation calls

for a regulatory framework essentially to create a National Database of Theses and Dissertations for India as well as to initiate the process of electronic submission of theses and dissertations in universities.

International Scenario

Networked Digital Library of Theses and Dissertation (NDLTD)

1.7 The movement on electronic theses and dissertation is led world-wide by the Networked Digital Library of Theses and Dissertation (NDLTD) (http://www.ndltd.org/) initiative, taken-up by the Virginia Tech University, USA. It is an open federation consisting of currently174 member universities and research institutions from all over the world. NDLTD's activities are focused on universities, libraries, faculties and research students in order to support authoring, indexing, archiving, dissemination and retrieval of electronic theses and dissertations worldwide.

1.8 NDLTD's vision is to increase the availability of student research documents for scholars and to preserve it electronically and to empower students to convey a richer message through the use of multimedia and hypermedia technologies. NDLTD encourages and supports universities to unlock their information resources and to advance digital library technology by sharing of experiences, tools, technology and knowledge.

UNESCO's Guide to ETD

1.9 The UNESCO has launched a project for development of an international framework for creation of electronic theses and dissertations (ETDs). The project aims to contribute to enhancing the production, access and archiving of scientific information by using the possibilities of new technologies. The specific objectives of the UNESCO project are:

- Establishing and disseminating guidelines, workflow models and best practices;
- Establishing a model training programme for project managers responsible for ETD programmes;
- Carrying out training courses and pilot projects

1.10 With the objectives mentioned above, the UNESCO has developed a "Guide to Electronic Theses and Dissertations (ETDs)", a resource targeted to the research students who are writing theses or dissertations, for faculty who want to mentor ETD authors, for research administrators who want to initiate ETD programmes, and for IT administrators at universities. The Guide promotes the sharing of knowledge locked up in universities, and the collaboration of universities worldwide is designed specifically for academic researchers and their mentors.

1.11 Some of the major ETD initiatives taken-up the world over are given below:

Australian Digital Theses Programme (ADT) (http://adt.caul.edu.au/): ADT Programme is a collaborative initiative involving university libraries in Australia. The initiative is voluntary, any Australian university can join the programme. ADT software, available free to members in the programme, was designed to be transportable and flexible which can be installed at each member institution with minimum modification.

Cyber-theses (http://www.cybertheses.org/)
The Cybertheses started as a cooperative project amongst selected universities in France with an aim to publish and distribute electronic theses on the web. Cybertheses is now open to all institutions of higher learning in France. It allows theses to be indexed online using common metadata model. The Cybertheses database contain the metadata of theses from participating institutions. It provides an efficient indexing system and rapid searching, even while significantly increasing the visibility and the distribution of the theses.

Dissertation Online
In German, "Dissertation Online" project, funded by the German Research Foundation, is a multi-site, multidisciplinary project that involves several educational and research institutions in Germany**.** The project aims at bringing scholarly publications, such as dissertations, diploma and master theses online. The project is highly successful and works closely with the NDLTD. A Bureau of Coordination has been established in the German National Library (DDB) to coordinate all developments of "DissOnline.de".

Other important ETD Projects

- MI (University Microfilms International) (http://wwwlib.umi.com/)
- National Library of Canada (http://www.nic-bnc.ca/index-e.html)
- TUG Electronic Theses Project (http://www.lib.lutwaterloo.ca/TUG/ETD)
- JETD Project (http://www.fics.utoronto.ca/etd)
- Digitale Dissertationen of Humboldt University at Berlin: (http://www.edoc.hu-berlin.de)
- University Lumiere Lyon 2 Digital Theses Project (http://www.univ-lycon2.fr/search.html)

Major Issues

Plagiarism

1.12 The risk of plagiarism is one of the important concerns that most students and faculty have. Although plagiarism cannot be ruled out even in print environment, it, however, cannot be denied that availability of documents in electronic format makes it easier for authors to copy. However, the risks of exposure of plagiarism is much larger in a scenario where theses are available in electronic format publicly given the fact that

most scholars and researchers still work in fields where a fairly small group of workers have detailed knowledge of their work. Moreover, the technology that has made ETD possible, also provides mechanism to detect plagiarized passages in electronic documents. Several software packages have now been developed that detect plagiarism. The software examines document files submitted for detection of plagiarism. It extracts the text portions from these documents and looks through them for matching words in phrases of a specified minimum length. When it finds matching files that share enough words in a number of phrases, a report is generated which contain the document text with the matching phrases underlined. Widespread use of such packages would increase risks of detection and, therefore, plagiarism.

1.13 Moreover, since ETDs as read more often then printed theses, there is a strong psychological pressure to discourage plagiarism. While on one hand, students would be more careful about consequences of plagiarism once detected, the associated faculty are likely be more diligent than with paper works at the time of checking the validity and quality of results reported. In short, detection by machines or other users, and threat of severe penalties are likely to discourage students from considering plagiarism with regards to ETD.

Intellectual Property Right (IPR) and Copyright

1.14 The owner of copyright of a book or any other written document belongs to its creator or author, irrespective of media used for its presentations, i.e. paper or electronic. The author of an electronic thesis or dissertation is its copyright holder and thus owns the intellectual property contained in it. It is for author to decide how their works will be reproduced, modified, distributed, performed in public or displayed in public. However, an author may use another author's work with certain restrictions know as "fair use". The owner of ETD, i.e. a research student must agree in writing to host his / her thesis on the web with or without restrictions. Such declaration / undertaking is taken from the student at the time of submission of thesis on Student Approval Form (Annexure I).

1.15 Such a declaration gives a university or institution non-exclusive rights to archive and host an electronic theses on the institute ETD repository. It does not take away the right of an author to use intellectual contents of his / her theses for writing papers, books or taking patents, etc.

Metadata

1.16 Metadata is the term used to describe data about data. The primary function of metadata is to facilitate information access, search and retrieval. To achieve this goal, the metadata provides information known about the document, such as its title, creator (author), publisher, and date of publication, etc. in order to facilitate access, search and retrieval of document. It usually includes information about the intellectual content of the document (i.e. subject keywords or descriptors), digital representation data, and security or rights management information.

1.17 Besides providing access to intellectual contents of a document, a function analogous to bibliographic records, digital objects also require metadata about applications and formats used for creating a digital object. Such metadata is required to provide long-term access to a digital resource. In short, the following three types of metadata are associated with the digital objects:

- Descriptive Metadata: Include content or bibliographic description consisting of keywords and subject descriptors.
- Administrative or technical Metadata: Incorporates details on original source, date of creation, version of digital object, file format used, compression technology used, object relationship, etc. Administrative data may reside within or outside the digital object and is required for long-term collection management to ensure longevity of digital collection.
- Structural Metadata: Elements within digital objects that facilitate navigation, e.g. table of contents, index at issue level or volume level, page turning in an electronic book, etc.

1.18 Meta data support efficient and effective organization, access and retrieval of information contents in a digital library. Meta data is used in effective designing of browsing and search interfaces of a digital library. With attributes of a digital objects defined in the metadata, it is a simple task to organize digital objects into predefined categories specified in search / browsing interfaces.

Metadata Schemes

1.19 Institutions dealing with electronic theses and dissertations have either developed their own standards or adapted existing metadata standards. These metadata standards attempt to describe the author, the work, and the context in which the work was produced in a way that will be useful to the researcher as well as the librarians and / or technical staff maintaining the work in its electronic form.

1.20 There are quite a few metadata schemes. Some of these schemes are applicable to documents received in a library, others have broader scope. Some of the important metadata schemes are as follows:

- Machine Readable Catalogue (MARC)
  Most traditional library systems exchange and store records using MARC format. MARC format has approximately 1,000 fields, several with repeatable sub-field. The use of this format allows very detailed description of the items. The MARC records for theses is not very robust, it often requires manipulation of fields to accommodate variations that are specific to theses as a type of document. Moreover, there are several fields in MARC record that are not applicable to theses and dissertations. Although, the MARC records have to be used for cataloguing theses and dissertations in a library, it is not designed for the ETDs.

- Dublin Core Metadata Elements Sets (DCMES)
  Dublin Core is a set of 15 attributes divided into three groups, i.e. content, intellectual property and instanciation. Associate to Dublin core are Dublin Core qualifiers that enhance the identification of items. Annexure III provides qualified Dublin Core Metadata elements for an ETD.

  Most of the institutional repositories use Unqualified Dublin Core (www.dublincore.org) metadata to ensure interoperability. Since OAI is based on the exchange of metadata, getting the metadata right is fundamentally important for a repository. The OAI compliant software automatically produce the necessary Dublin Core metadata for harvesting by service providers.

- ETD-MS
  The ETD-MS standard is developed by the Networked Digital Library for Theses and Dissertations (NDLTD) which is used for submission of electronic theses and dissertations at Virginia Tech University. ETD-MS incorporates basic Dublin Core elements with the addition of several elements that further describe parts of the ETD process. The ETD-MS is the only metadata standard supported by the software. Annexure IV provides guidelines for information contents for each element of ETD-MS.

Standards for Metadata Harvesting

Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

1.21 The Open Archives Initiative was originally proposed to enhance access to e-print / pre-print archives. Gradually, the scope of the initiative has broadened to cover any kind of digital content including images and videos. The OAI-PMH is a protocol devised to make machine-readable metadata widely available for use. The development of the OAI-PMH protocol took root in a meeting that was convened in late 1999 at Santa Fe, New Mexico to address problems of the e-print world. As disciplinary e-print servers became more common, it was difficult to support searching across multiple repositories. Repositories needed greater capabilities to automatically identify and access papers that had been deposited in other repositories. Need was felt to build a framework to bring about a kind of integration of these e-print/pre-print archives to solve these problems. The major work was to define an interface to permit e-print servers to expose their metadata for the papers it held, so that search services or other similar repositories could then harvest its metadata. These archives would then act as a federation of repositories by giving a single search platform for multiple collections.

Metadata standards and OAI-PMH

1.22 For the purpose of interoperability, the OAI Protocol for Metadata Harvesting specifies unqualified Dublin Core, encoded in XML, a mandatory metadata schema as the lowest common denominator. Almost any metadata scheme can be "downgraded" into unqualified Dublin Core. However, each server is also free to offer metadata in one or more schemas, and a harvester can request that metadata in any format in addition to the unqualified Dublin Core.

The OAI-PMH Framework

1.23 There are two classes of participants in the OAI-PMH framework:

- Data Providers: Data Providers, or repositories, administer systems that support the OAI-PMH as a means for exposing their metadata. All universities would serve as data providers once they have set-up their OAI-complaint ETD repositories.

- Service Providers: Service Providers, or harvesters, use metadata harvested via the OAI-PMH as a basis for building value-added services, such as building subject gateways, email alerts, etc. The central agency designated to maintain National Theses Database would work as service provider for ETD in India.

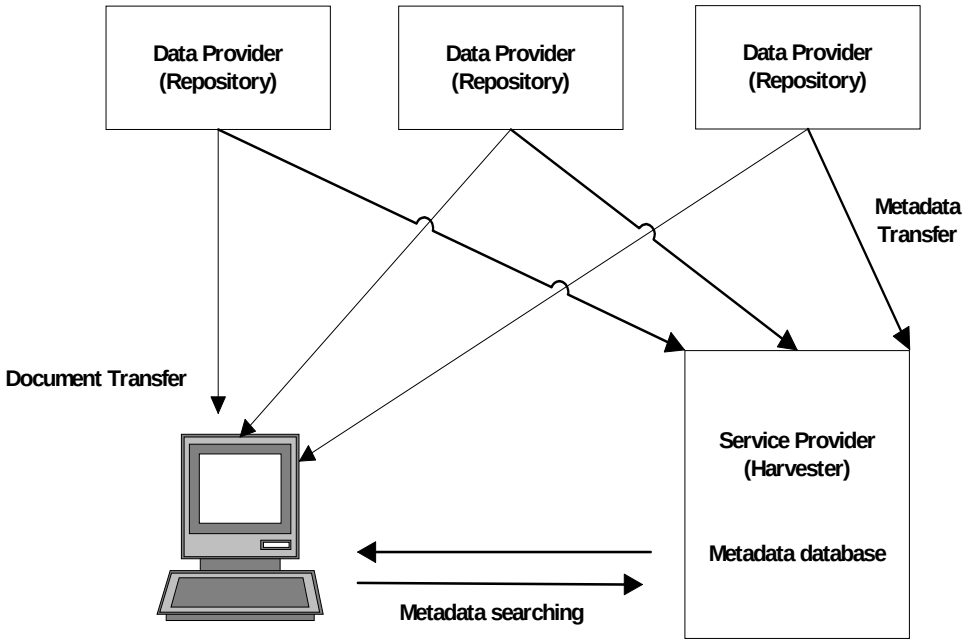

Figure 1: The OAI-PMH Architecture

1.24 The metadata stored in the data providers' database is transferred in bulk to the metadata database of the service providers. The transfer of metadata is done in a series of requests and responses between the data provider and the service provider/harvester. The

OAI-PMH Protocol depends upon the HTTP-transaction framework for communication between a harvester and a repository. Requests may be made using either the HTTP GET or POST methods. All successful replies are encoded in XML, and all exception and flow-control replies are indicated by HTTP status codes.

Standards for Data Format for Theses

1.25 The preparation of an ETD involves making of the electronic copy of the thesis / dissertation. There are many file formats available for text, image, sound and video. Basically, file formats can be proprietary or non-proprietary i.e. open file formats. MS-Word is an example of proprietary file format and OpenOffice.org is an example of non-proprietary file format.

1.26 Open standard file formats are strongly recommended to upload documents in the ETD. Using open standard file formats is very advantageous in the long run. The reason is proprietary standards require proprietary software which may not be accessible to everyone. Proprietary standards may not be backward compliant i.e. file formats created using an older version of the software is usually not readable by the newer versions. In the archival perspective it is feared that if the proprietary software become obsolete and go out of use there is no way one can retrieve the documents that have been written using these software. Being proprietary standards, nobody can even attempt to write programs to read these older versions of the file formats. Point in case is the WordStar document format of Microsoft. The software has become obsolete and is not available. And the current versions of MS-Office cannot read these files. Whereas, for open standards anyone can write a program to retrieve the document even if the software becomes obsolete. In cases where use of proprietary software is unavoidable, it should be urged to include the required software with the ETD file so that the user does not require to purchase additional software. Some of the open standard document formats are described below:

Open Text File Formats

Hypertext Mark-up Language (HTML)

1.27 HTML is the language with which Web pages are designed. This standard has been defined by the World Wide Web Consortium (W3C). HTML allows web documents to be created with ease. The primary objective of using HTML is to build a web page that communicates readily and effectively to make the document on the web most compelling to access and read. HTML is a plain text file and any text editor as simple as Notepad can be used to create HTML documents.

eXtensible Markup Language (XML)

1.28 XML provides a structured representation of data that can be implemented broadly and is easy to deploy. XML is a subset of SGML (Standard Generalized Markup Language), modified and optimized for delivery over the Web. This standard has been

defined by the World Wide Web Consortium (W3C). XML can be used to format and transfer data in an easy and consistent way. XML is more flexible, because one can define his/her own tags/elements. Hence it is possible to tailor the XML documents for different needs, and makes it possible to use XML to represent all kind of data for different purposes. XML is also a plain text format.

1.29 The advantages of using XML-based applications are that they can be implemented and used irrespective of the device and platform being used. XML is device independent and platform independent. The data encoded in XML can be accessed irrespective of the device i.e. on a wireless handset, a palmtop, a laptop, an airport kiosk, a projector in a conference room, or a desktop PC or operating system, i.e. Windows, Unix, Linux, Sun Solaris, etc.

Portable Document Format (PDF)

1.30 Invented by Adobe Systems, Adobe Portable Document Format (PDF) is a publicly available specification used by various standards bodies around the world for electronic document distribution and exchange. As an open file format specification, PDF is available to anyone who wants to develop tools to create, view, or manage PDF documents.

1.31 The most popular and preferred format is PDF (Portable Document Format). Adobe Acrobat's Portable Document Format (.pdf) is recommended, since it retains all format codes and graphic images, appearing as the original paper document and also because it is easily portable. In addition .pdf files can be indexed and searched by keywords. Apart from being an open standard, it maintains the integrity of the document. It can be converted to PostScript format, which can be used for electronic delivery and printed directly. Both Microsoft Word and WordPerfect files can be easily converted to .pdf files. Training and assistance in the conversion process to .pdf should be provided to the students.

TeX

1.32 TeX is a typesetting program designed for high-quality composition of material that contains a lot of mathematical and technical expression. (*http://www.tug.org/tex-ptr-faq*) It has been adopted by many authors and publishers who generate technical books and papers. It was created by Professor Donald Knuth of Stanford University, originally for preparation of his book series "The Art of Computer Programming". TeX has been made freely available by Knuth in a generic form.

1.33 TeX implementations are governed by the principle that the same input should produce the same output, modulo font availability and output device resolution. All implementations of TeX must pass a "trip test" that assures adherence to these guidelines.

1.34 TeX has been tailored for and installed on almost every platform (computer + operating system), and is available as freeware, shareware and commercial

implementations. The TeX program is usually accompanied by other software to form a complete and usable system.

LaTeX (Lamport TeX)

1.35 LaTeX is a document preparation system for high-quality typesetting *(http://www.latex-project.org)*. It is used for technical or scientific documents, but it can be used for almost any form of publishing. LaTeX is based on Donald E. Knuth's TeX typesetting language. LaTeX was first developed in 1985 by Leslie Lamport, and is now being maintained and developed by the LaTeX3 Project. LaTeX is available for free at *http://www.latex-project.org/ftp.html*.

Open Image File Formats

Portable Network Graphics (PNG)

1.36 PNG (pronounced 'ping'), the Portable Network Graphics file format, is an open raster image format. It is supported by the W3C and IETF and is expected to be released as ISO/IEC International Standard 15948. The latest version is PNG 1.2. It was developed in 1995 as a replacement for the GIF (GIF89a - Graphics Interchange Format) and a possible replacement for the TIFF (Tagged Image File Format). It is still not widely used and it has taken some time for Web browsers and image application software to support it. Now, PNG files have reasonable support among the leading browsers and can be created and manipulated within many image applications.

Joint Photographic Experts Group (JPEG)

1.37 The JPEG compression and its corresponding file format were developed in the late 1980s by independent members of the Joint Photographic Experts Group (JPEG). JPEG properly refers to the compression, with the file format officially termed JFIF (JPEG File Interchange Format). However, the format has become commonly known as JPEG and is usually given .jpg and .jpeg extensions. It is the most popular image format for the Web.

Poor Standard of Theses

1.38 Several researchers, research supervisors and education administrators would be reluctant to join ETD movement knowing the fact that research work being conducted in their institutions are of poor quality. We should take this as an opportunity to improve our research work. Considering the fact that ETDs are read more often then printed theses, there will a strong psychological pressure on to the research students as well as on the research supervisor to improve the quality of their research work. While on one hand, students would be more careful about the quality of their work, the associated faculty are likely be more diligent than with paper works at the time of checking the validity and quality of results reported. Moreover, qualitative research work available on the Internet from other universities would give an idea to researchers as to how they can improve their work.