Development of ETD Repository at IITK Library using DSpace

R. Mishra, S. K. Vijaianand, Noufal P. P. and Gaurav Shukla

P. K. Kelkar Library,
Indian Institute of Technology Kanpur, India.
{rmishra, vanand, noufal, gshukla }@iitk.ac.in

Abstract. Establishing an ETD repository at any organization having large volume of these is really a challenging task. As the Indian Institute of Technology Kanpur had started its research programs from 1963 onwards, we have over 9000 of M.Tech and Ph. D theses in different areas of Science, Engineering, and Humanities & Social Sciences. We have completed the digitization of all submitted theses. Our ETD powered by DSpace with more advanced features like cross-collection search, linking to referred IITK theses, user friendly design and customized work flow for different modes of submission of theses to ETD etc.. In this paper, we clearly explain and share the experiences we have gained from the digitization to accessing ETD, and the challenges we faced, enhancements we incorporated with special emphasis on technical developments, and the lessons we have learnt during the various stages of development of our project for ETD.

Keywords: ETD (Electronic Theses and Dissertations), Institutional Repositories, Open Archives Initiatives (OAI), Digitisation

1 Introduction

Nowadays Institutional Repositories (IRs) constitute an integral part of present day digital libraries permitting global access to the scholarly publications. Establishing an IR has become the buzzword among library fraternity. IR gives exposure to researchers and their thought content to the wider community and affords an opportunity for future research enhancement and long term preservation of information. Electronic Theses and Dissertations (ETD) are one of the major components of any

Institutional Repository. IITK had started its research programs from 1963 onwards and we have over 9000 of M.Tech and Ph. D theses in different disciplines of Science, Engineering, and Humanities & Social Sciences. In early 2005, IITK Library has initiated its Digitization Program with a multi-

ARD Prasad & Devika P. Madalli (Eds.): ICSD-2007, pp. 249-259, 2007

pronged approach and vision. Initially our emphasis has been for establishment of ETD as a part of our Digitization Initiatives.

2 Digitization of IITK Theses and Dissertations

Digitization of theses has proved to be a most stupendous task; nonetheless a challenging one! Scanning of 900,000 pages with graphs, images, charts and programming codes with pale typed papers dating back to 1963 was really a challenging and painstaking job. We have availed the scanning facility available at Indian Institute of Information Technology Allahabad (IIITA), a Govt. of India, Mega Scanning Centre, Million Book Project (MBP) of Universal Digital Library (UDL). IITK signed an MoU with IIITA on 15 July, 2005 for scanning the existing theses. The first phase of scanning theses was successfully completed well within twelve months time. Now we have developed a digitization facility in the library which is well equipped with Minolta PS7000 scanner and required accessories. Following is the schematic description of the various processes involved in the digitization job:

2.1 Specifications

After careful consideration, the following decisions with regard to scanning specifications were undertaken:

(i) *TIFF* for archival and *PDF* for presentation purpose.
(ii) *Resolution of Image:* 600 dpi TIFF and PTIFF (Processed). PTIFF has been subjected again for scan fixing to resize it to a uniform shape and size.
(iii) *Compression:* CCITT 4 Fax
(iv) *Conversion* to PDF format using Adobe Acrobat: PDF is the de facto standard for the secure and reliable distribution and exchange of electronic documents and forms around the world.

2.2 Quality Control

Quality Checking is no doubt one of the essential processes in digitization to ensure quality output and to get the most reliable and consistent data. As mentioned earlier, the state of most of theses has led to the discrepancies that crept in during scanning; and hence we took a decision for thorough and proper Quality checking. The salient features of quality checking were: completeness, contrast, sharpness, skew, resolution, bit depth level of compression and data conversion, missing, duplicate and misplaced pages, file naming etc. Finally, having put in tremendous efforts and time, we completed the quality checking and rectified the problematic items as late as April, 2006.

3 Setting up of IT Infrastructure for ETD @IITK

IT Infrastructure for ETD, is one of the core component for this proper planning and right expertise is required especially when you consider user demand, Push /Pull of Technology and ever-growing and emerging global standards. Major areas where our DL team was actively involved is given below. Brainstorming sessions were conducted right from proposal preparation down to ETD launching :

- Selection of hardware and software
- Configuration and customization
- Selection of Metadata Standard
- Designing standard workflow pattern
- Content Management
- Metadata extraction and injection
- Uploading: Mode of Submissions

3.1 Selection of Hardware and Software

3.1.1 Hardware

At the initial stage, we have used a test bed server for installation and customization. Later on we procured a high-end server.

As part of the exercise to establish a digitization facility in our library, we have procured a MINOLTA PS7000 scanner with requisite accessories for further digitization and development work.

3.1.2 Software

After a detailed study and close observation of available Content Management and Digital Library Management Systems, we decided to choose *DSpace*, an Open Source Software from MIT and HP due to its features like granularity, adherence to standards, multi-format support, customizable interface, OAI-PMH compliant, support with fully qualified DCMI, remote submission, authorization and reviewing, community/sub-community based collection architecture, import and export features, Persistent Identifiers Handle System, Open URL Support, Lucene search engine and generation of statistical reports, etc.

3.2 Configuration and customization

Having gone through the due process of systems study and analysis, we successfully installed and configured DSpace (Version: 1.2.2) on Linux platform on 19 July 2005 for our test bed. Later, we have acquired a dedicated high-end server for hosting our ETD for the benefit of academic community and have installed and configured latest version of DSpace (Version.1.4) with all prerequisites.

3.3 Designing standard workflow pattern

Designing user friendly, systematic and simple workflow is one of the essential requirements of any ETD system to be successful. If you choose any commercial system, for its management, the vendor is responsible for configuration and customization as per your needs. But in the case open software like DSpace, you have to thoroughly study and become familiar with it to get the best out of it. We have successfully customized and restructured the entire workflow pattern as per our needs.

3.4 Selection of metadata standard

We have chosen DCMI as metadata standards due to its usability, flexibility, repeatable elements, qualifying nature and wide popularity.

3.5 Content management

After the quality checking of the data we have gone through the following set of processes for content management to make digitized versions more meaningful and secure:

3.5.1 Scanning & insertion of signed certificate page

During the current academic year, 2005-06, the students have been asked to submit direct to the library a CD containing the thesis and abstract. As part of the content handling, once the certificate page is scanned and inserted into thesis, the item becomes more meaningful and authentic.

3.5.2 Extracting abstract

For digitized theses, we have utilized our existing database for extracting abstract, like other metadata values. This searchable abstract provides depth retrieval at the abstract level in addition to the keywords and subject heading. For current year, as mentioned earlier, we have collected abstract pdf from the scholars, which is searchable.

We have used a tool PDFBOX Library available at http://www.pdfbox.org/ for extracting text from PDF files. It comes with an ExtractText utility, which converts PDF files to text files.

3.5.3 Embedded watermarking

Considering the possibilities of potential infringement of the copyright for all submitted theses, we decided to embed IITK Logo in each page of thesis as a water mark. We have developed a script for automatic embedding and batch processing.

We are using a tool iText Library freely available at http://www.lowagie.com/iText/. It comes with a utility to embed watermark in Pdf text file. But in our case, we have manipulated it to embed in PDF images.

4 Metadata extraction and injection

As mentioned earlier in case of abstract extraction, we have used the bibliographic database of theses powered by iitKLAS, an Oracle supported Library Package developed in-house. We have developed a Perl Script for Extracting Metadata from the existing library database and injected it in our new system, powered by DSpace.

A Java program written by us, connects to our current oracle database and extracts all the metadata for each item separated by a specifier line by line. This program creates a file **"MetaDataDetails"** and writes to it.

5 Uploading: mode of submissions

We have formulated three precise strategies for submission and uploading of theses to ETD at the various phases of the program. As mentioned earlier, we had envisaged three phases of theses submission; namely the Batch processing, the CD mode and the proposed Online submission, as detailed in Fig. 1.

5.1 Submission mode 1: batch processing

Firstly, we have customized the work flow for submission of digitized theses as batch process. The metadata was extracted from our existing theses database by using a Java program developed by our team. After extraction we have injected those to ETD by using customized export/import utility of DSpace.

The perl script written reads the metadata from file **"***MetaDataDetails***"** and creates a directory structure for each collection with a "*dublin_core.xml***"** files containing all the metadata with corresponding DC Elements. This script also stores the files to be uploaded after being watermarked.

5.2 Submission mode 2: CD

Secondly, theses submitted during the first semester of the current year 2005-06 containing full text and an abstract as PDF file were uploaded by our team using customized work flow after the completion of essential content processing.

5.3 Submission mode 3: online (proposed)

Finally, for Future transactions, we have proposed the Online submission by scholars themselves to ETD. For this, we have meticulously customized the entire work flow process in terms of Registration, Submission, Reviewing, Validation etc. Once the approval of the authorities is there, Online theses submission to the library will start.

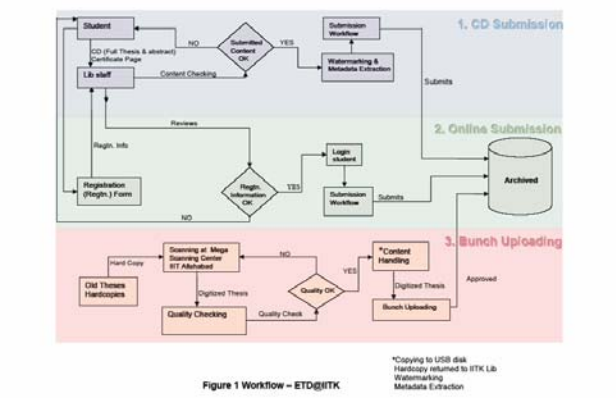

Figure 1. Uploading: Mode of Submissions

6 Additional options incorporated in our ETD

We have entirely customized our ETD to incorporate the advanced features. These features are very user friendly. We are discussing below technical and manipulative features which have been designed to make the ETD more meaningful.

6.1 Customization to incorporate additional 'browse' option

DSpace permits by default the following access options: (a) by Communities and Collections, (b) by Authors, (c) by Titles, (d) by Issue Date and (e) by New Collection and Recent Submission and (f) by Subjects.

In addition to the above noted default 'browse' options, we have customized DSpace to incorporate browsing facility by Supervisor/s and Citations.

The main files which are associated for browsing are "*Browse.java*", "*BrowseServlet.java*". By carefully studying it one can also add the other browsing facilities in the same manner as was done for browsing Authors and Subjects.

The files we created for browsing by Supervisors are "`ItemsBySupervisorServlet.java`**",** "`supervisors.jsp`**"**, "`items_by_supervisor.jsp`". Additional tables and sequences are also created `ItemsBySupervisor`, `CommunityItemBySupervisor`**,** `CollectioItemBySupervisor` and `itemsbysupervisor_seq`.

The additional servlet and files created for browsing by Citations/References are "`ItemsByReferenceServlet.java`", "`references.jsp`" and "`items_by_reference.jsp`"**.** In this regard we have created the following

table's `itemsByReference`, `CommunityItemByReference`, `CollectioItemByReference`, and sequence `itemsbysupervisor_seq`.

6.2 Customization to incorporate additional 'Search' option

DSpace offers by default the following search features: (1) Search all DSpace, (2) Bounded Search within a specified Community's Collection, (3) Simple search and (4) Advanced search.

6.2.1 Cross collection Search

Search involving more than one discipline, known as 'cross collection search', viz. Chemistry and Chemical Engineering; Materials Science and Physics; Metallurgy & Materials Science; Lasers and Biomedicine, etc. have been incorporated as additional features under the 'Advanced search' option.

Our collections are uniquely defined by the combination of department and degree type. e.g. M.Tech Thesis @ CSE , Ph.D Thesis @ AE, etc. where M.Tech or Ph.D is the degree type and CSE or AE is the department. The same idea is implemented on cross collection search. For each item two DC elements `description.department` and `description.degree-type` uniquely defines a collection for cross collection searching. We have done some changes in the following files "`advanced.jsp`" as a display file and "`QueryArgs.java`" as a core java which takes care of structuring the query.

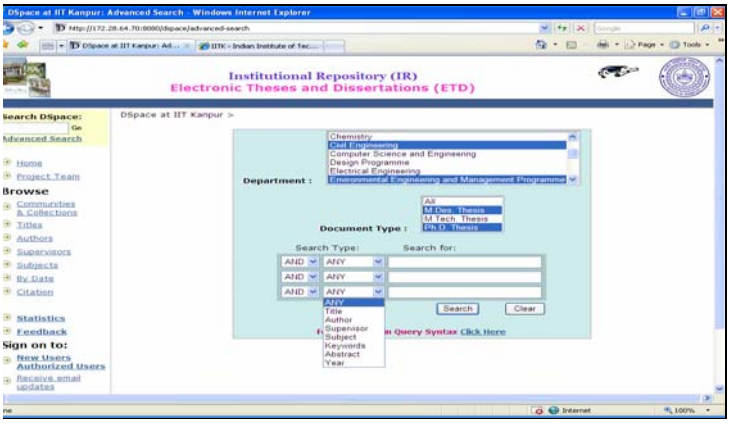

Figure 2. Cross Collection Search

6.3 Supervisor, Author, Subject and Reference count

The Strength/Count of each Browsing element is one of the interesting options. We have manipulated the file "`Browse.java`" This option will give you a picture about the number of items guided by a person, item created

by an author, items related to a particular subject and the no. of citations for a particular item.

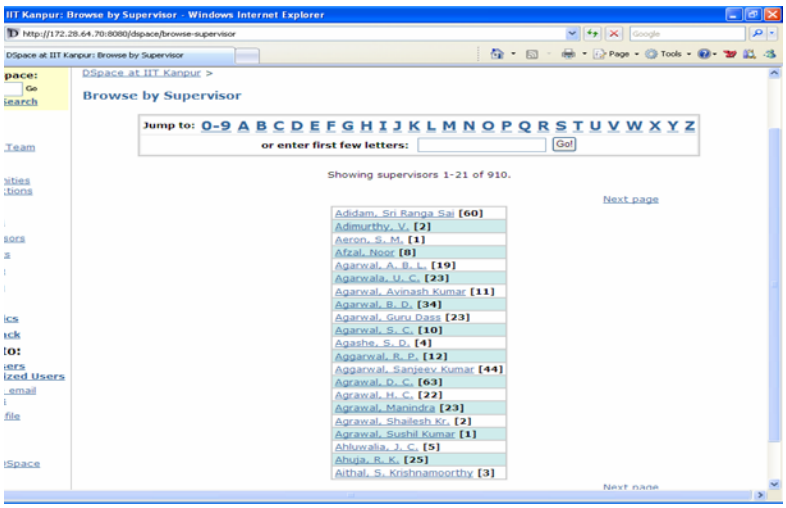

Figure 3. Supervisor Strength/Count

6.4 Keyword and Reference Linking

'Subject Keyword' is a common user's approach to find literature in a given subject area. An access has been provided to the various 'Subject Keywords' incorporated in each thesis, through a hyperlink.

A link to 'cited thesis/theses' from the existing IITK theses collection, as mentioned by author in 'References' showing citation impact of our theses by the IITK scholars, has been provided as an additional feature. Further, it is proposed to send an e-mail alert to the author/s and supervisors in case their theses are cited by any new scholars at IITK.

The java core file we changed "`ItemTag.java`" which takes care of rendering dc elements of Items for this purpose.

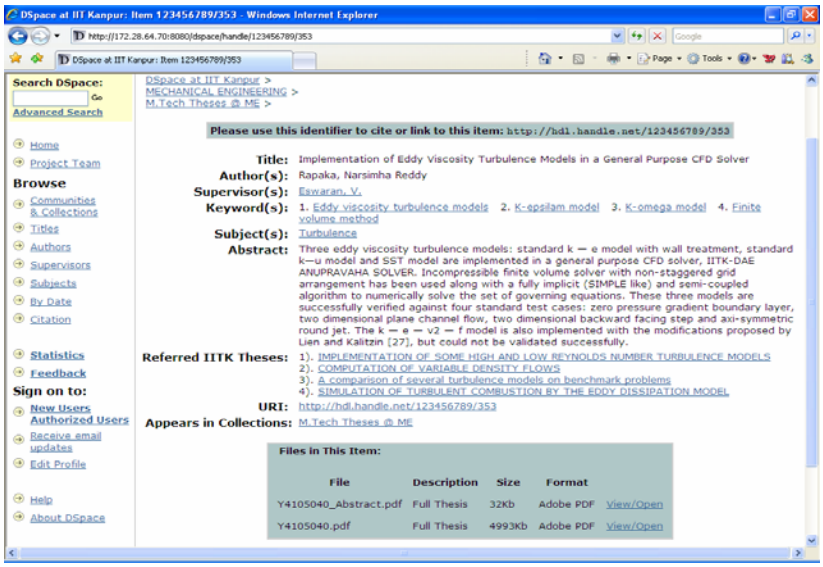

Figure 4. Keyword and Reference Linking

6.5 Redesigning 'Registration Form'

Normally, in DSpace, the registration form contains the following details: First Name, Last Name and Contact No. We redesigned this by providing additional options, i.e. Roll No., Degree and Department. These data elements are very essential for DSpace administrator to assign a scholar to a particular E-Group and for user to submit his/her item to collection without any problem.

File associated with redesigning registration form are "`Eperson.java`" the table eperson was altered and the above-mentioned additional fields have been added.

6.6 Redesigning 'Feedback Form'

Feedback is one of the processes in which part of the output of a system is returned to its input in order to regulate its future output. We redesigned the default Feedback Form, incorporating additional personal details of scholar, i.e. email-id, designation, degree, department, PF No./Roll No., and a set of features related to the facilities, and/or description parameters concerning content/metadata and workflow, these have been used to solicit ratings on defined parameters from the scholars submitting theses to the library.

Earlier this feedback used to be sent through email to the administrator but now we have created a specific table in the database as Feedback to store the values. Analysis of theses values is helpful for assessment of the ETD.

7 Conclusion

To sum up, digitization of theses with varied complex nature of pages and establishing an ETD system using an Open Source software has really been a challenging and interesting assignment for us. Experiences and lessons we have learnt from project initiation, quality checking of scanned document, content management and launching have been very rich and informative. In depth customization of DSpace to incorporate the foregoing advanced features like workflow for different modes of submission, various browsing and searching options and keyword and reference linking etc have been widely appreciated by our user community. We record with a sense of pride that once our ETD is put on the web, it would be the largest OAI compliant ETD repository in India and one of the top ten in the world.

Acknowledgments

The authors would like to place on record their sincere gratitude to Prof. S. G. Dhande, Director, IIT Kanpur, and to Prof. T.V. Prabhakar and Prof. Harish Karnick, Dept. of Computer Science & Engineering, for their initiative, encouragement and support in our endeavour to digitize IITK theses collection. We also express our thanks to Prof. M. D. Tiwari, Director, IIIT Allahabad and Mrs. Ratna Sanyal, Co-ordinator, Mega Scanning Centre, IIITA for providing scanning of over 9,000 theses at their institute.

References

[1] Adobe. http://www.adobe.com/
[2] Digital Library of India. http://dli.iiit.ac.in/
[3] DigiTool. http://www.exlibrisgroup.com/digitool.htm
[4] DSpace Federations. http://www.dspace.org/
[5] Dublin Core Metadata Initiative (DCMI). http://www.dublincore.org/
[6] Eprints. http://www.eprints.org/
[7] Gail, McMillan. (2003). Electronic Theses and Dissertations. Encyclopedia of Library and Information science. (pp1034-1040) New York: Marcel Dekker.
[8] Mishra, R., Vijaianand S. K., Noufal P. P., Rajesh Kumar andShukla, Gaurav., (2006). Digitisation Initiatives to Destress Library Collection: A case study of ETD at P. K. Kelkar Library, IIT Kanpur. To be published by International Conference on Digital Libraries, New Delhi, Dec. 5-8
[9] Open Archives Initiatives. http://www.openarchives.org/
[10] PDFbox. http://www.pdfbox.org/

[11] Registry of Open Access Repositories (ROAR). http://archives.eprints.org/index.php

[12] Rowlands, Ian., Nicholas, David. (2005) Scholarly communication in the digital environment. Aslib proceedings 57 (6) .pp 481-497

[13] UNESCO Guide to Electronic Theses and Dissertations. http://www.etdguide.org