Annals of Library and Information Studies
Vol. 63, December 2016, pp. 274-288

Full-text ETD retrieval in library discovery system: designing a framework

Prosenjit Sarkar[a] and Parthasarathi Mukhopadhyay[b]

[a]Superintendent (Library Services), The University of Burdwan, Burdwan, West Bengal, India,
Email: prosenjit.toton@gmail.com

[b]Associate Professor, Department of Library and Information Science, University of Kalyani, West Bengal, India,
E-mail: psmukhopadhyay@gmail.com

Received: 12 August 2016; revised: 19 November 2016; accepted: 22 December 2016

This paper discusses designing an open source software based library discovery system for full-text ETD retrieval on the basis of a cataloguing framework developed by using available global standards and best practices in the domain of theses cataloguing. The purpose of this prototype framework is to provide a single-window search and retrieval system for end users for discovering ETD at metadata level and at full-text level. The prototype framework is based on three-layer architecture with Koha ILS as backend metadata provider, Apache-Tika as full-text extractor and VuFind as discovery system. A MARC-21 bibliographic format, especially designed to handle TDs, is working as data handler mechanism in Koha ILS and the harvester of VuFind is tuned to fetch bibliographic data related to ETD in marcxml format. The user interface of VuFind is also configured to support accessing ETDs from global-scale services like NDLTD, OATD, IndCat, ShodhGanga etc. apart from the local level ETD collection in order to provide an all-in-one search interface for users.

Keywords: Library discovery; ETD retrieval system; Full-text retrieval system; VuFind; ETD cataloguing

Introduction

India has a vast higher education system. As of July 31, 2016, India has 759 universities, 350 of which are state universities, 47 central universities, 123 deemed universities and 239 private universities (source:http://www.ugc.ac.in/oldpdf/alluniversity.pdf) ; and 15 IITs, 8 IIMs, 8 IIITs, 20 NITs and 4 NITTRs (source:http://www.aicte-india.org/einp.php). Every year these higher academic institutions are producing thousands of research outputs in the form of theses and dissertations (TDs) but most of these TDs are underutilized and lying obscure in the libraries of the TDs producing institutions. There are various methods available for increasing awareness of and distribution of TDs. These include building a digital repository of electronic theses and dissertations (ETDs), and making bibliographic information available in the library catalogue. Indian researchers are facing problems in not only accessing these TDs but also even in retrieving TDs available at the institutional level from the library OPAC. A few university libraries in India are cataloguing TDs and also opening up these valuable resources in the library OPAC, but all these services are restricted by two serious lacunae – i) non availability of books and TDs from a single-window search interface (particularly for library OPACs built on the top of SOUL, LibSys etc.); and ii) search is limited only to the metadata level and thereby not allowing users to search in the content part of TDs (all ILSs are presently limited by this inefficiency of retrieval including Virtua ILS and Koha). As a result the challenge lies in developing a framework for ETD retrieval where these resources will be available from the same user interface alongside traditional library materials like books, journals etc. and the framework will serve additional utility of searching TDs at full-text level.

In India, several attempts have been made by higher academic regulatory bodies to encourage developing ETD systems in Indian universities, IITs, IIMs and in this regard University Grants Commission (UGC) published two regulations; first is UGC Regulations, 2005, which is called *Submission of Metadata and Full-text of Doctoral Theses in Electronic Format* and the second one is UGC Regulations, 2009, named as *Minimum Standards and*

Procedure for the award of M.Phil/Ph.D. Degree and both regulations have been published in Gazette of India. ETDs are considered as valuable information sources, but these resources are still grossly underutilized because of the lack of discoverability[1]. Therefore, cataloguing of ETDs and subsequent inclusion of these resources in discovery services are important aspects for libraries and end users.

An OPAC or WebOPAC facilitates library patrons to access cataloguing data/metadata and provides direct access to a library's bibliographic database from anywhere at any time. But the search interfaces of these OPACs are not up to-date in comparison with the latest Web-based search tools. On the other hand the success of a library depends on the efficient discovery of library resources. Library resources and collections are useless without their seamless discovery to the library users. As the discovery process has moved online, library and information centers have started to work to solve the discovery problem and introduced easy and effective discovery processes for the users, especially new generation users. Influenced by the Google phenomenon, library users expect an easy, user friendly, Google-like search interface and facilities from the library OPAC/WebOPAC. They expect for a user friendly search interface for searching their desired documents irrespective of their forms and formats. In short, users need a search interface which is Google-like simple and OPAC-like elegant. Emergence of new tools changed the library catalogue search interfaces in that direction. These tools are used on the top of the existing integrated library system (ILS) and provide user friendly new catalogue search interface with an improved look and feel. Considering the end users' needs, library and information centers have been trying to implement new discovery services called Web-scale discovery service. Web-scale discovery service provides centralized access to current library resources including library catalogue, article databases, institutional repositories through a single-window search interface, minimizing the earlier provision of searching across decentralized silos.

Vaughan[2] defined "web-scale discovery service is a service capable of searching across a vast range of pre-harvested and indexed content quickly in a seamless manner. Contents are harvested from local library resources, remotely hosted digital libraries, and institutional repositories to create a vastly comprehensive centralized index of variety of objects well suited for quick search and retrieval. Content is enabled via harvesting facilitating access to their metadata or full-text content for indexing purposes". Hoeppner[3] described that Web-scale discovery is "a pre-harvested central index coupled with a richly featured discovery layer providing a single search across a library's local, open access, and subscription collections". The Central index of discovery service is possibly generated from different categories of contents, such as bibliographic and holdings information from a library's resource management system, metadata or full-text from institutional repositories and from open access repositories; and discovery layer is the user interface and search system for discovering, displaying, and interacting with the collection of contents in library systems.

Web-scale discovery service is now common in higher academic and research libraries. The goal of developing Web-scale discovery systems is to help library users in discovering library materials from a single search box and to make library research as intuitive as Google. Many proprietary and open source tools are available to provide Web-scale discovery service. Discovery systems provide discovery interfaces which are known as next generation OPACs and provide user friendly single-window search box/interface, facilitate discovery of a variety of objects from globally distributed resources and provide links to full-text from citation/bibliographic records in search results.

A NISO white paper by Breeding[4] described that Web-scale discovery systems have different facets. Discovery interfaces, termed as "next-generation catalogues", are the replacement to OPAC modules of integrated library systems (ILS). They provide an advanced end-user interface used by researchers to submit queries, receive results, and facilitate content selections. A discovery interface includes various 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 delivered via a Web browser, to perform tasks such as presentation of a 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 implementation and provide access to remote index platforms. Recently, Bento box style option, where retrieved results are clustered according to origin such as from catalogue, from repository etc. is becoming popular in discovery interfaces. Discovery systems facilitate access to a large, diverse information landscape of scholarly materials including ETDs — irrespective of formats of the materials and their locations. Library and Information professionals have welcomed the advances in discovery services for their clienteles and have been implementing new discovery services in their libraries. Many commercial and open source discovery interfaces are available. Major commercial interfaces are ExLibris Primo**,** SirsiDynix's Enterprise, Innovative Interfaces' Encore, VTLS' Visualizer, BiblioCore, ProQuest AquaBrowser (an illustrative list only); and open source discovery interfaces are Blacklight, Scriblio, eXtensible Catalog/XC and VuFind. Discovery interfaces added with integrated library systems provide many advantages such as –

(i) New, state of the art, search tool that provides easy resources discovery;

(ii) Facilitates searching of all library resources from a single search interface;

(iii) Provides easy access to full-text documents and supports indexing at the full-text level; and

(iv) One-stop resource portals incorporating meta-searching discovery tools, federated search options, browsing functionalities and query forwarding to external sources.

This research study investigates and addresses the topic of cataloguing of ETDs and their indexing at the full-text level; and tools and technologies available to libraries of higher academic institutions to make their resources discoverable and accessible by the scholarly communities that they serve. To facilitate the discoverability of ETDs and to provide Web-scale discovery service this research developed a prototype single-window search system named as ETD@BU. Function of the discovery layer in the prototype framework is schematically illustrated in Fig. 1**.**

ETD@BU framework is based on open standards and open source software and contains three layers – ILS layer (by Koha ILS), Discovery layer (by VuFind) and Full-text layer (by Apache Tika as full-text extractor). Koha is an open source ILS and VuFind is an open source discovery tool, which is configured for this prototype to sit on top of Koha ILS and to connect to Apache Tika for full-text indexing. VuFind also provides an end-user search interface to discover ETD alongside other library materials (see Fig. 2). This prototype uses MARC-21 Bibliographic Format designed and fine tuned to catalogue library resources including ETDs.

Literature Review

Before developing a framework for organizing, managing and facilitating discovery of ETDs at the metadata level as well as at the full-text level, a comprehensive literature review was carried-out to

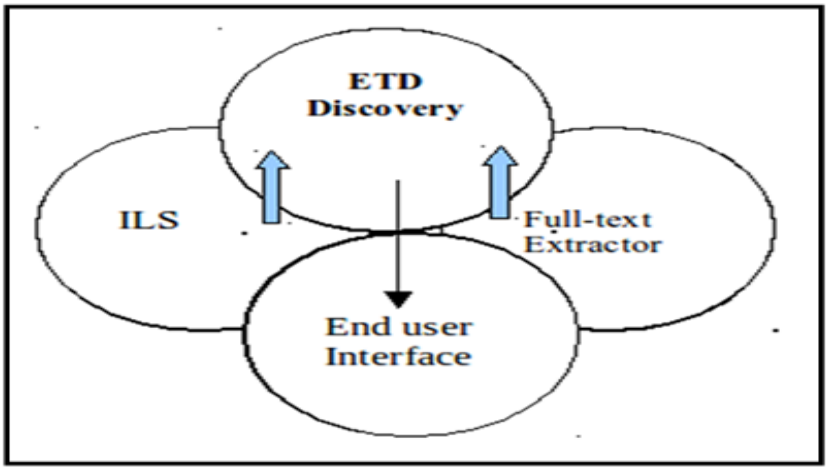

Fig. 1—Functions of ETD discovery system

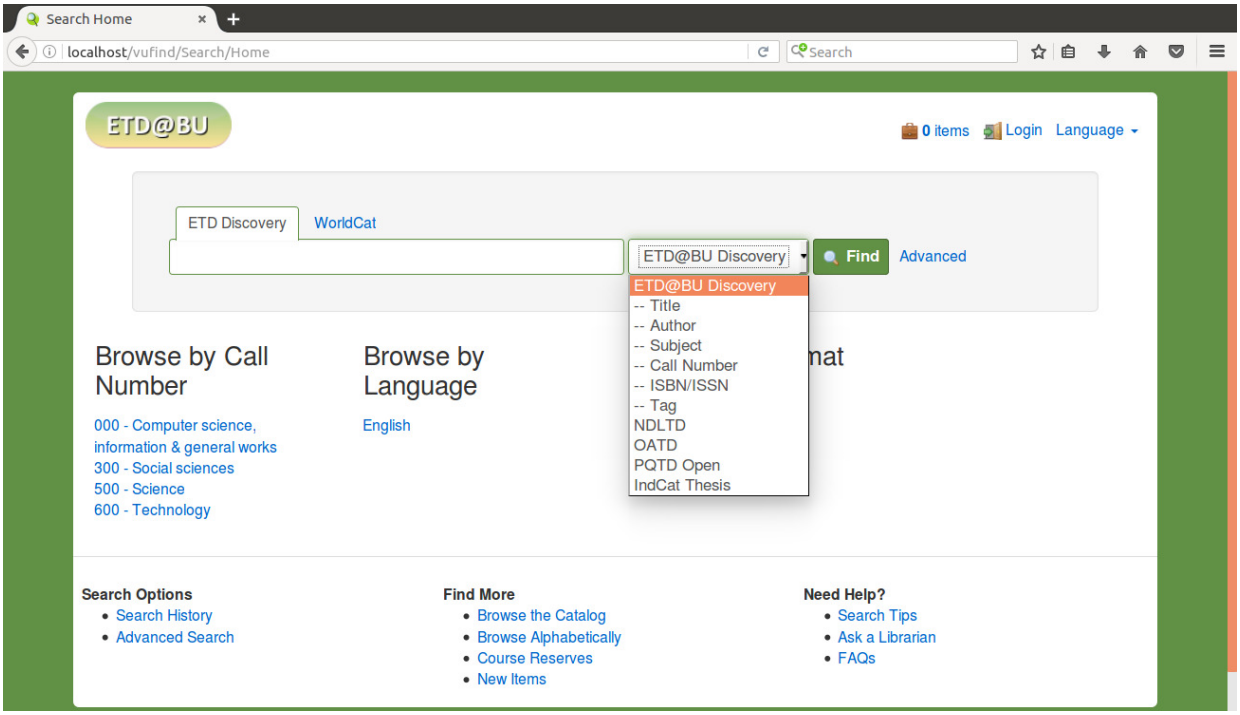

Fig. 2—ETD@BU discovery interface with VuFind on top and Koha as backend ILS.

obtain an idea of cataloguing standards and practices related to ETDs and their retrieval. This literature review is divided into two parts – the first one is cataloguing of ETDs and the second one is Web-scale discovery and its implementation in a higher academic library set up.

Cataloguing of ETDs

Hoover[5] reported that once ETDs are submitted in the repository, access to ETDs may be provided through the library's online public access catalogue (OPAC). Lubas[6], McCutcheon[7] and Smith II[8] advocated for the availability of ETDs bibliographical information via online public access catalogue (OPAC) to increase discoverability. Therefore, cataloguing of ETDs is very important. Wolverton Jr. et al[9] published a book in the year 2009 which is a comprehensive guide for cataloguing of ETDs. The book contains six chapters, describing different aspects of cataloguing of TDs/ETDs, and reporting several survey results of cataloguing of both TDs and ETDs. The authors reported an interesting observation that studies/articles on cataloguing of theses and dissertations are quite small in number in comparison to TDs produced in each year. A study by Frank and Rowe[10] focussed on a general overview of issues involved in providing access to ETDs via the cataloguing process. They raised a few relevant questions related to thesis retrieval such as – (i) how do the ETDs affect the cataloguing effort to describe them? (ii) how to integrate bibliographical information of ETDs into the library OPAC. It means that when library users search the OPAC for their required information they will find ETDs along with other types of documents. This approach will increase discoverability of ETDs. Another study by Hoover[5] reported that the most suitable way to find a specific ETD is an author or title search of an institution's OPAC. Keller[11] reported that subject analysis is the main problem in TDs cataloguing. Frank & Rowe[10] and Khurshid[12] opined that assigning and adding subject headings to TDs in the library catalogue is a time consuming activity and expensive in relation to required skilled manpower. Hoover & Wolverton[13] conducted a survey and its result reveals that the cataloguing of print TDs and ETDs is more or less similar and identified that the only difference is format. McMillan[14] presented various e-document cataloguing tools including AACR2R. The author opined that AACR2R provides the guidelines for elements to include in the MARC record for an ETD cataloguing. Another study by McMillan[15] stated goals of cataloguing ETDs and described different

MARC fields for cataloguing. The author mentioned the need for institutional policy for assigning subject descriptor and recommended author supplied keyword as the substitute of subject descriptor and recommended 653 MARC tag for this purpose, 520 for abstract and 856 to describe the electronic location and access information. Frank and Rowe[10] described the example of the University of South Florida for cataloguing paper formatted TDs and ETDs and commented that both items provide single cataloguing records indicating both formats. A study by Genoni & Cowan[16] investigates providing access to bibliographical information of Australian theses, highlights the problems with regard to both accessibility and quality of those bibliographic information. McMillan[15] opined that many libraries programatically derived the cataloguing records directly from ETD systems to the library OPAC. Some authors such as Koulouris & Anagnostopoulos[17] and Surratt & Hill[18] reported about semi-automated methods for cataloguing of ETDs; it is a time saving cost effective process of cataloguing ETD. Bevan[19], El-Sherbini & Klim[20] and Lubas[6] discussed various aspects of cataloguing ETDs, such as staff training, current cataloguing practices, the impact of metadata on library cataloguing practice, the role of librarians related to selecting metadata standards for cataloguing of ETDs, decisions about assigning subject descriptors, author supplied keywords etc. McCutcheon[7] reported various levels of treatment options for cataloguing ETDs, such as, from the creation of basic bibliographic information of ETDs to information with full subject descriptor and name authority control. Sarkar & Mukhopadhyay[21] explores cataloguing of TDs/ETDs and its retrieval along with other type materials in an integrated processing environment and developed a software framework by using open source software and open standards for this purpose.

Discovery services

Due to the information explosion, a huge amount of printed documents and online information resources are now available, the library does not facilitate centralised access to its different information silos, nor does it provide a user-friendly search and retrieval system for scholarly community whose expectations are influenced by popular search engine Google. Searching across library resources is a difficult task, requiring high alertness and steep learning curve. To improve discoverability of resources and their retrieval, currently universities and research libraries are introducing various new discovery services in their libraries. New discovery services include next generation catalogues, federated search, and Web-scale discovery, in addition to their traditional integrated library systems. Han[22] reported how new discovery services use the cataloguing records and the problems that libraries faced in bibliographic control to work with new discovery services. Sarkar & Mukhopadhyay[23] reported how metadata can be used to organize and facilitate discovery of ETDs. Web-scale discovery service is an important issue in the current library and information science domain. Higher academic libraries and information centers are implementing Web-scale discovery services to replace traditional OPACs/WebOPACs to increase discoverability of a wide range of objects in different formats from a single-box search interface. Copenhaver & Koclanes[24] reported that Web-scale discovery services have become the new standards for higher academic and research libraries within the past decade. A study conducted by Lundrigan et al[25] evaluated user satisfaction with the discovery service Summon, results described a high level of satisfaction overall, although this was heavily influenced by the quality of search results over ease of use. Study provides insight into the information-seeking behavior and search preferences of a user when a discovery layer is implemented in a research library. Vaughan[2] reported why Web-scale discovery service should be implemented in higher academic and research libraries. In their study, Balaji Babu & Krishnamurthy[26] analyse the paradigm shift of library automation to resource discovery by exploring the applications of resource discovery and reported the current position of India on adapting resource discovery applications. A study by Goodsett[27] compared three discovery services such as EBSCO Discovery Service, Ex Libris' Primo and Serials Solutions' Summon – on the basis of different parameters and described their strengths and weaknesses. Breeding[4] provides an overview of the current resource discovery environment. Hoeppner[3] described Web-scale discovery concepts and terminology, shares findings from his interviews with major *Web-scale discovery* service providers, and provides a template checklist, which librarians can use during their own exploration of these systems.

Several authors discussed about the use and implementation of VuFind resource discovery tool in the libraries. Han[22] described VuFind implementation in the University of Illinois Library and reported faceted navigation options provided by VuFind are more extensive than is provided by other library systems. Houser[28] in his article reported that VuFind is designed and developed to sit on top of an integrated library system (ILS) and replace the OPAC interface and provides a single discovery interface for other local resources such as a digital repository or local databases. Denton & Coysh[29] presented the results of implementation of a discovery layer as a next-generation catalogue, based on usability testing and an online survey in an academic library. Authors expected that the outcome will assist VuFind implementers and other next-generation catalogues to improve their own systems. Said authors also reported that incomplete, inconsistent and inaccurate metadata is the reason of providing less than ideal results of faceted navigation display. Another study by Ho et al[30] presents a case study of what is involved in implementing the VuFind discovery tool and reports on VuFind hardware and software architectures and some initial testing results and evaluation. Selection and implementation of a discovery tool in the library is a complex process. Chickering & Yang[31] evaluated and compared eleven proprietary, and three open source discovery tools; these findings will help librarians to adopt a discovery tool in their libraries. A research report by Katz et al[32] investigated different ways in which implementation of the open source discovery tool VuFind provides information to its users, showing how these mechanisms can be combined with authority data to enhance discoverability of library resources.

Objectives of the study

The three-level objectives of this study related to design of ETD@BU are:

- To develop ETD cataloguing policy, standards and framework by using global recommendations, open standards and open source software;
- To provide an easy and effective discoverability for scholarly objects like ETDs (alongside other library materials) from a single-window search service; and
- To develop mechanisms for indexing and retrieving ETD at full-text level with a facility to query-forwarding mechanisms for external sources.

Methodology

The library OPAC/WebOPAC acts as an information gateway, making possible the integration of variety of library resources within single access and search tool. The methodology for this research work conceptually can be divided into two areas:

Part 1 – Developing a cataloguing policy and standard for ETDs following global standards and best practices; identifying an open source ILS for cataloguing and integrated retrieval of ETDs; and identifying an open source Web-scale discovery tool.

Part 2 - Designing a software framework for ETD cataloguing and their integrated retrieval via OPAC/WebOPAC and implementation of Web-scale discovery system and services for retrieval of ETDs at metadata level and at full-text level with a user friendly single-window search interface.

Developing a cataloguing policy and standard for ETDs and selecting open source ILS and discovery tool

This research developed a cataloguing policy framework for ETDs following global standards and best practices. And for this purpose this research study developed a set of rules and software framework for ETD cataloguing. This study consulted AACR2R (Anglo American Cataloguing Rules 2nd revision) cataloguing tool and various guidelines prepared by different organisations such as Library of Congress, OCLC and Australian National Bibliographic Database for the cataloguing of TDs in both printed and digital formats. An ETD is an electronic resource and according to AACR2R an ETD may be considered either as manuscript or as monograph. For a long time TDs have been considered as unpublished documents due to their limited and restricted access and distribution. Emergence of ETDs has changed the distribution and access scenario of TDs. So there is a valid ground for the consideration of ETDs as published documents that can be treated as monographs. Therefore, ETDs may be catalogued as monographs. But the problem arises when TDs in both printed and digital formats are accepted by universities and higher academic

institutions for the requirement of any degree. Is it is necessary to create multiple records for each format of the ETDs? Or is a single bibliographic record, noting the multiple formats, enough? How to manage bibliographic records for different formats of a TD in a catalogue database? How to manage records for different formats of TD resources in a catalogue database? In India, many universities and higher academic institutions accept TDs in both formats. To solve this problem, this research work studied and compared three guidelines for cataloguing of ETDs which are discussed in the following section.

(A) Libraries Australia prepared guidelines for cataloguing theses for the *Australian National Bibliographic Database* (http://www.nla.gov.-au/librariesaustralia/files/2015/05/ Theses.pdf). In this guideline three separate rules for cataloguing TDs are given:

(i) General guidelines for cataloguing TDs;

(ii) Cataloguing ETDs; and

(iii) TDs presented/submitted in multiple formats.

The third part of the standard defined that many TDs are now produced in both formats, i.e. electronic and print. In an ideal situation each format should be described as a separate record, adding a MARC 530 (additional physical form) note and optionally a MARC 776 (linking entry) to indicate the existence of the alternative form. This is consistent with the practice outlined in the *Guidelines for cataloguing remote access electronic resources* which is discussed in the Anglo American Cataloguing Rules, 2nd edition, 2005 revision (AACR2) rule 9.0A1. This guideline advised that when a T/D is produced in multiple formats (e.g. both print and electronic formats), or the original is reproduced in a varied format, that both formats of a T/D can be catalogued as one record and recommended some MARC fields such as 007 for Physical Description, 245 $h for General Material Designation, 533 for Reproduction Note, 534 for Original Version Note etc. and also provided an example of a print thesis reproduced in electronic format.

(B) *Special Cataloging Guidelines by OCLC* (https://www.oclc.org/bibformats/en/

specialcataloging.html) described that most TDs are unpublished items made available through copies but ETDs that are remotely accessible via the Web should be treated as published items and be catalogued as original electronic publications, as explained in AACR2R 9.4B2. According to AACR2R 9.4B2 rule, all remote access electronic resources should be considered as published documents. This guideline does not provide any more information about cataloguing ETDs and for more information suggests consulting Cataloging Electronic Resources: OCLC-MARC Coding Guidelines *(*https://www.oclc.-org/support/services/worldcat/documentation/cataloging/electronicresources.en.html). This research consulted the said document, but the document does not provide any specific information regarding ETDs cataloguing.

(C) Another set of guidelines named as *Rules and Tools for Cataloging Internet Resources: Instructor Manual* (http://www.loc.gov/-catworkshop/courses/ cataloginginternet/ pdf/-ceig1_IM-FINAL.pdf) prepared by Steven J. Miller for the Library of Congress and other organisations does not provide any specific information regarding cataloguing ETDs.

Cataloguing TDs presented/submitted in multiple formats is the ideal guidelines for cataloguing TDs. After studying and consulting the above three guidelines, this research selected the standard for cataloguing ETDs used for *Australian National Bibliographic Database* i.e. TDs presented/submitted in multiple formats prepared. This is a cost effective and time saving method for cataloguing of ETDs from the managerial point of view. Moreover it reduces the chance of duplicate entries (for different formats of the same TD) in catalogue database and thereby gets rid of data redundancy and naturally ensures a better retrieval experience for end users. The MARC-21 data model has been used by the *Australian National Bibliographic Database* for the TDs (both printed format and electronic) cataloguing. MARC-21 is an open standard. This study already recommended using open standards in the system under design, so MARC-21 has been used as data model for cataloguing of ETDs. In addition to MARC-21 data elements used by *Australian National Bibliographic Database* for cataloguing of TDs in both formats, this research added some other MARC-21 data elements for cataloguing TDs in both formats such as MARC 520 for T/D abstract, 653 for uncontrolled index term, 655 for genre and 720 for the name and role of T/D supervisor/guide etc.

The software framework for cataloguing and integrated processing and retrieval of TDs along with other types of library materials uses Koha as it is more robust in providing integrated search and retrieval facilities in comparison to SOUL and LibSys ILS software (the commercial ILSs available in India)[21]. The screenshot (Fig. 3) demonstrates that Koha retrieved three documents from the host database against search query 'web'. Out of these three retrieved documents, the first two documents are books and the third one is ETD. It is clear from Fig. 3 that Koha provides integrated search and retrieval of all types of library materials from its OPAC interface.

Many proprietary and open source Web-scale discovery tools are available in the domain of library discovery. It has already been mentioned that the software framework for this research is designed and developed by using open source software. Examples of some open source discovery tools are Blacklight, Scriblio, eXtensible Catalog/XC and VuFind. To identify a suitable Web-scale discovery tool for this research, some pre-defined selection criteria have been taken into consideration. Criteria include user base, download statistics, continuous revision or upgrading and active user forum. On the basis of these criteria, this research identified and selected the VuFind open source discovery tool for the software framework. The filled-in MARC format for ETD (Fig. 4) as adopted in view of the discussion made in previous section is as follows:

Part 2 - Designing the framework

Design and development of the framework for ETD retrieval at metadata level and indexing at full-text level in library discovery system requires judicious strategies and planning. Obviously, one software can-not perform all the operations related to different activities. Therefore, the system requires creating a platform where different application software can be integrated in a seamless and harmonized manner to perform specific jobs. It is already reported that the framework is designed and developed by using open source software. The methodology related to developing the framework may be divided into the following logical groups:

i) Development of the basic cluster of the software framework;

ii) Selection, installation and configuration of ILS and design of ETD data-entry standard therein;

iii) Selection, installation and configuration of Discovery tool; and

iv) Installation and linking of Full-text extractor in the software framework.

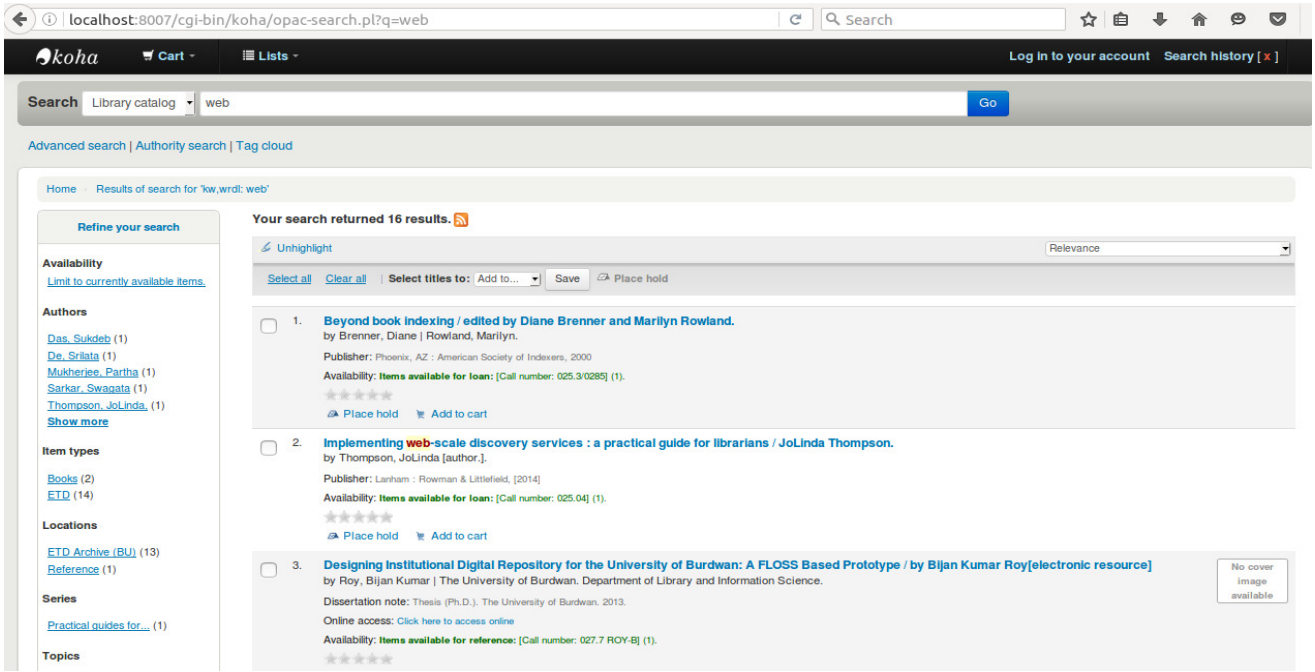

Fig. 3—Result retrieved from Koha OPAC search

000	05449nam a22003737a 4500
003	lfl
005	20160904151504.0
007	cr aa aaaaa
007	ta
008	160904b xxulllll llll 00l 0 eng d
040	_aThe University of Burdwan _beng _cCentral Library _dCentral Library
082	_221 _a027.7 _bROY-D
100	_aRoy, Bijan Kumar _91
245	_aDesigning Institutional Digital Repository for the University of Burdwan: A FLOSS Based Prototype _b _c/ by Bijan Kumar Roy _h[electronic resource]
500	_aTitle from screen page; viewed 5th January 2016.
500	_aSubmitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the Department of Library and Information science.
502	_aThesis (Ph.D.). _cThe University of Burdwan. _d2013.
504	_a.
516	_aElectronic publication; full text available in PDF format.
520	_aInstitutional Digital Repositories (IDRs) have become a hot topic and have been an emerging research field in Library and Information Science because of rising costs, flat budgets, and restricted access to information, as well as rapid changes in technology, scholarly practice, and patron expectations etc. There is increasing awareness that universities and research institutions lose valuable digital and print material due to difficulties in accessing them and lack of good preservation practices ….. _b
530	_aAlso available in print format.
534	_exxxiii, 451p. ; 30 cm. _pOriginal.
538	_aMode of access: World Wide Web.
650	_aInstitutional repository _95
650	_aInstitutional digital repository _9115
650	_aDigital library _93
653	_aIR
655	_aElectronic Thesis or Dissertation _9163
710	_aThe University of Burdwan. _bDepartment of Library and Information Science. _9166
720	_aBiswas, Subal Chandra _eSupervisor
720	_aMukhopadhyay, Parthasarathi _eSupervisor
856	_uhttp://localhost/ftt/bijanthesis.pdf _qAdobePDF (.pdf)

Fig. 4—MARC view of an ETD catalogued in Koha

Development of the software framework:

The software framework for designing the ETD discovery system ETD@BU is divided into many sub clusters which are described below:

a) Basic cluster

This framework like many other systems is based on Linux-Apache-MySQL-PERL/PHP (LAMP) architecture. Ubuntu 16.04 LTS version has been used as operating system (OS); Apache as web server; MySQL is used as backend database management system and PERL/PHP as open source programming environment. Programmes written in either PERL or PHP can run smoothly in AMP framework. This cluster also includes Default Java (OpenJDK version 1.8) and Apache-Solr. Apache-Solr is a comprehensive open source text retrieval engine and the major features of the tool are – support for an array of search operators, faceted navigation, auto suggestion feature, search statistics and so on.

b) ILS cluster

This research selected Koha (version 16.05.03, the latest stable release as of August 31, 2016) open source ILS for organization and management of ETDs. Koha is customisable and has an active user forum with huge user base. Koha can run on LAMP architecture though it can run also in other OS. Moreover, Koha can be configured to act as OAI data provider and can transfer metadata (in oai_dc and marcxml formats) to service providers against OAI verbs. For example, Koha can serve the OAI request http://localhost:8007/cgi-bin/koha/oai.pl?verb=ListIdentifiers&metadataPrefix=marcxml as follows:

```
<OAI-
PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmln
s:xsi="http://www.w3.org/2001/XMLSchema-
instance"xsi:schemaLocation="http://www.openarchives.or
g/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-
PMH.xsd">
<responseDate>2016-09-15T18:01:59Z</responseDate>
<request verb="ListIdentifiers" metadataPrefix="marcxml"
>http://localhost:8007/cgi-bin/koha/oai.pl</request>

<ListIdentifiers>

<header>
<identifier>KOHA-LFL:1</identifier>
<datestamp>2016-06-02T22:32:48Z</datestamp>
</header>

<header>
<identifier>KOHA-LFL:2</identifier>
<datestamp>2016-01-21T03:27:02Z</datestamp>
</header>

<header>
<identifier>KOHA-LFL:3</identifier>
<datestamp>2016-01-21T03:47:14Z</datestamp>
</header>

<header>
<identifier>KOHA-LFL:4</identifier>
<datestamp>2016-01-21T04:02:58Z</datestamp>
</header>

<header>
<identifier>KOHA-LFL:5</identifier>
<datestamp>2016-01-21T13:51:42Z</datestamp>
</header>

<header>
<identifier>KOHA-LFL:6</identifier>
<datestamp>2016-04-25T22:42:21Z</datestamp>
</header>

<resumptionToken cursor="49">marcxml/49/1970-01-
01/2016-09-15/</resumptionToken>

</ListIdentifiers>

</OAI-PMH>
```

This unique feature of Koha makes it an ideal candidate to handle the ILS cluster of the software framework.

c) Discovery cluster

On the basis of some pre-defined criteria such as user base, download statistics, continuous revisions and active user forum, this research selected VuFind software (version 3.0.3) as discovery cluster. VuFind is a comprehensive open source discovery tool and presently in use in different global scale organizations for a variety of purposes. It 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 a variety of formats. This feature ensures that VuFind can retrieve data (in oai_dc or marcxml formats) from the ILS cluster seamlessly and then can throw data to Apache-Solr using the SolrMarc indexing tool for efficient indexing.

d) Full-text cluster

This research used Apache Tika as an extractor for this cluster. The developer of the project in its introduction says "The Apache Tika toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF)". More importantly, the project homepage also mentioned that "all of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more" (source: https://tika.apache.org/). The above-mentioned promises and its compatibility with the discovery tool VuFind made Tika an automatic choice for the full-text layer of the software framework in comparison with other full-text extractors. The mechanism works in three levels – i) first we need to store full-text theses in web document root and the cataloguing record in MARC-21 bibliographic format includes hyperlink to related full-text file (in pdf format) in tag 856 subfield $u; ii) configuring VuFind to use Apache-Tika as extractor; and iii) fine tuning of import mechanisms of VuFind to index the full-text (here pdf files) files for end user retrieval.

The search interface of VuFind discovery tool (named as ETD@BU) is primarily acting as user interface for search and retrieval. It supports simple search as well as advanced search interface with many sophisticated search techniques and filtering options. The Screenshot in Fig. 5 described that ETD@BU is providing integrated searching of different types of library resources like books, ETDs etc.

Functions of ETD@BU

ETD@BU is a prototype Web-scale discovery system. It may be considered as a next generation catalogue for providing a user friendly search interface with impressive look and feel that facilitates easy discovery and retrieval of ETDs along with other library resources at metadata level and full-text level. ETD@BU provides faceted navigation and facilitates access to various related external sources other than the host ILS, such as PQDT, OATD, NDLTD, IndCat databases, ShodhGanga etc. and thereby ensures discovery and retrieval of ETDs from different sources from a single-window search system. This is a special feature in VuFind through which any number of external datasets can be integrated with the discovery layer (see Fig. 7 and Fig. 8). ETD@BU is also capable to provide full-text indexing by using the mechanisms as described in previous section. With additional tools and technologies, ETD@BU can harvest data from other sources and index them for searching and browsing, together with the library's own bibliographical records to increase the performance of the system. Like many other VuFind

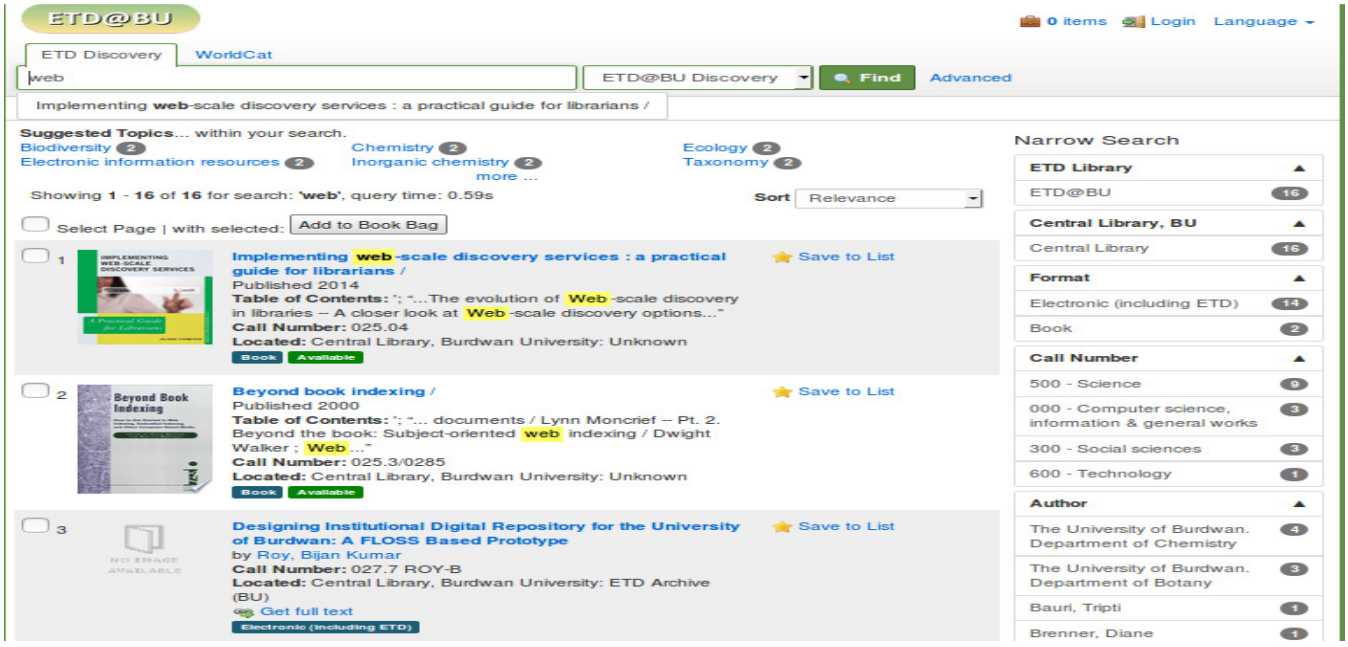

Fig. 5—Integrated searching from ETD@BU

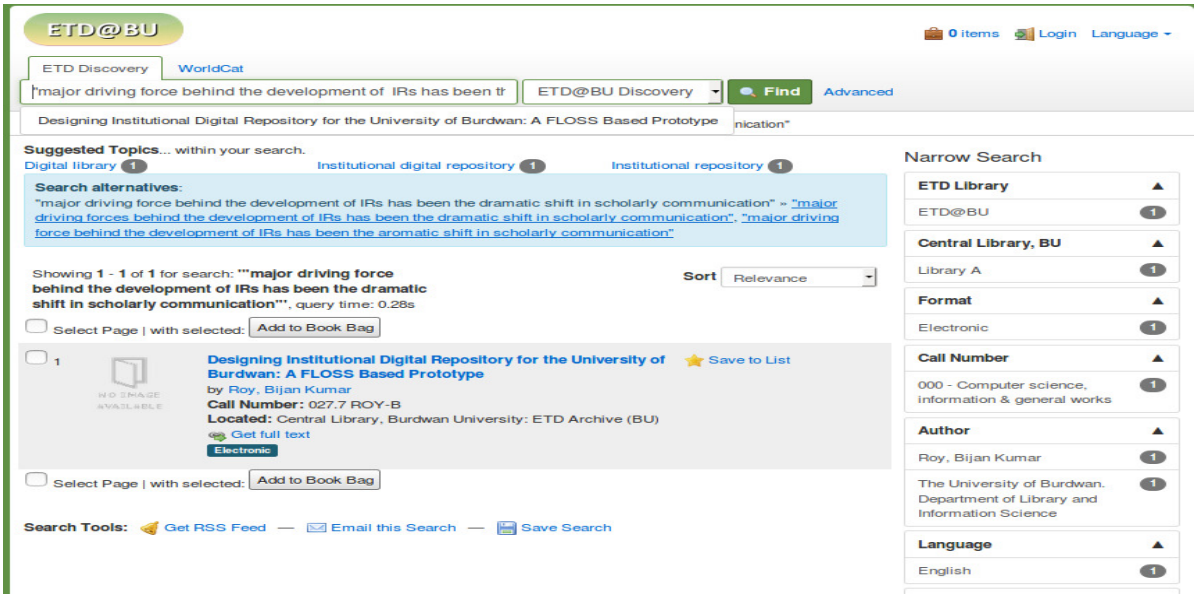

Fig. 6—ETD@BU providing full-text indexing

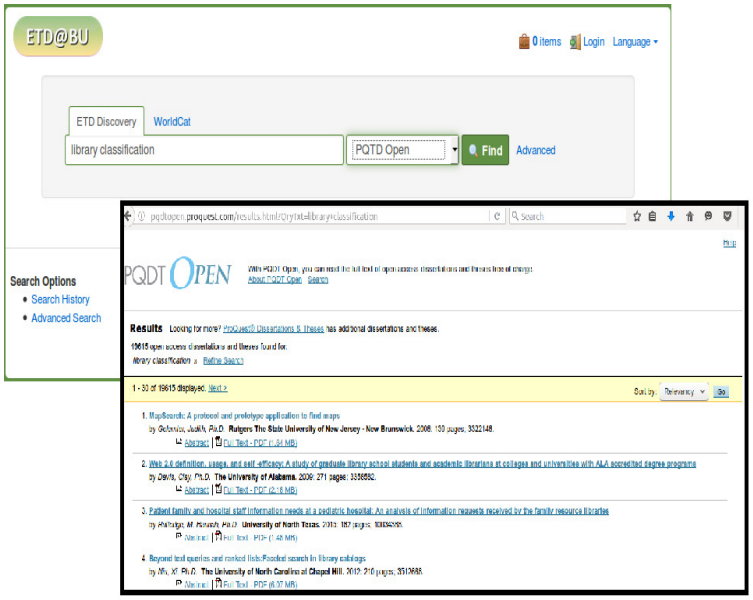

Fig. 7—Searching PQDT OPEN with search term from ETD@BU Search Interface

implementations, ETD@BU can narrow search facilities by faceted navigation and advanced search features of VuFind (powered by Apache-Solr) give end users different filtering options like range search, Boolean search, relational search and positional search. The facilities of browsing of resources by call number, author, format, era and region, etc. help users in structured navigation. This prototype also experimented with an array of Bengali script based full-text ETD resources and the system performance is quite encouraging in handling Indic scripts based resources. It may be reported that the Department of Library and Information Science, Kalyani University, West Bengal, India has developed Bengali script based VuFind interface as the first ever Indic script interface of VuFind and it is now included in the official release of VuFind.

The Screenshot (Fig. 6) illustrates the full-text indexing capability of ETD@BU discovery system.

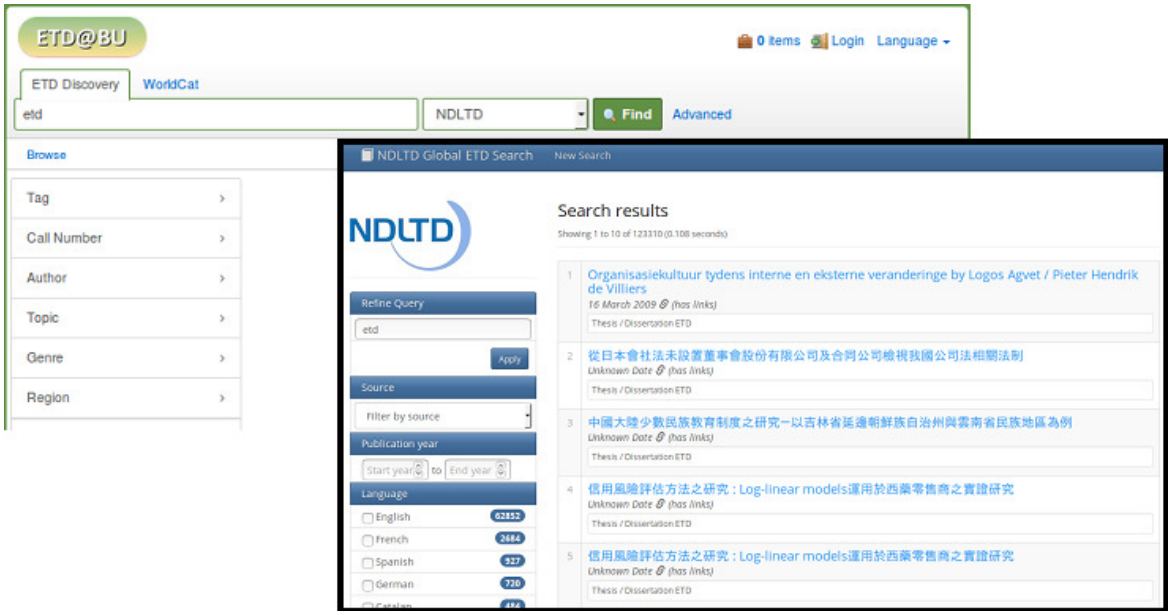

Fig. 8—Searching NDLTD with search term from ETD@BU Discovery Interface and result retrieved

Here a phrase is selected for searching and result retrieved the document where selected phrase can be found in contents of the thesis. ETD@BU discovery system provides federated search facilities and is capable to search remotely hosted open access ETD repositories and facilitates federated searching and browsing. The screenshot in Fig.7 illustrates that PQDT OPEN repository searching from ETD@BU search interface with a search query and result retrieved.

In similar way it is possible to search other remote open access ETD databases from ETD@BU Web-scale ETD discovery system. The search interface includes global-scale ETD services like NDLTD, OATD, ShodhGanga etc. For example, the Fig. 8 depicts ETD@BU providing searching and browsing facility for NDLTD.

The different screenshots given here describe the various functionalities of the ETD@BU discovery system. The universities and higher academic libraries in India can implement this type of improved discovery system and service in addition to or in place of the library OPAC to provide better search services for end users and thereby can increase discoverability of scholarly materials like ETDs. Apart from these user friendly retrieval features, the ETD@BU interface provides an array of value-added features like statistics generation (through Piwik) and chat service (through Mibew) to help determining usage statistics and to support librarian-user communication in real time respectively. It is worth noting here that Piwik support in VuFind is inbuilt and can be configured easily through settings in *config.ini* file but Mibew is an out-of-the-box solution which needs to be integrated with user interface by customizing the theme layout of VuFind. The Fig. 9 demonstrates the chat service (named as Ask the Librarian) through the integration of Mibew (an open source live support application) in the end user interface of the ETD@BU.

Conclusion

ETD@BU is a localised ETD discovery service that can facilitate searching of ETDs at the metadata level and at the full-text level from the local library collection as well as from remote open access repositories (which contain ETDs) through query forwarding mechanisms. Such a facility ensures the much required one window access to all possible resources. The locally available resources can be configured to be searched at the full-text level. Koha (ILS) OPAC provides only integrated searching of different types of materials at metadata level and link to access full-text documents of the host institution. Koha, as ILS, does not provide full-text indexing facility. Discovery system ETD@BU (powered by VuFind) provides ETD metadata and links to access full-text documents and supports indexing at the

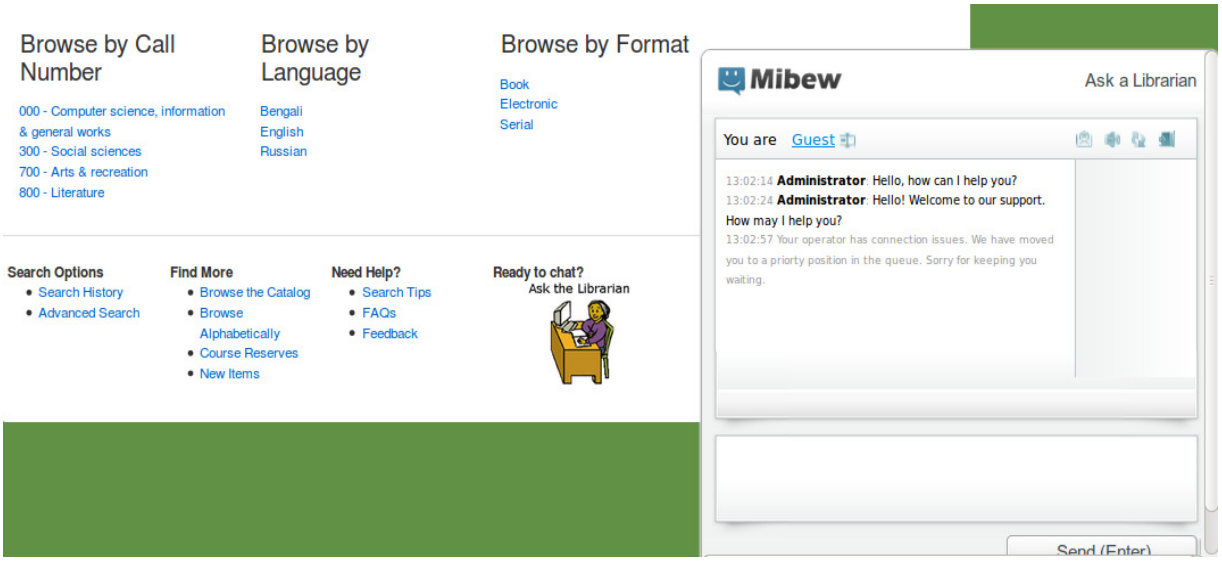

Fig. 9—Live chat support in ETD@BU Discovery Interface

full-text level which are illustrated in different screenshots. ETD@BU also provides integrated searching of different types of materials including ETDs. The beauty of ETD@BU is that it can narrow the searching by different facets such as call number, author, topic, genre and region. Implementation of Web-scale discovery services in universities and higher academic libraries will increase the standard of the organisations and attracts their clienteles by providing better quality services in discovering scholarly objects including ETDs from the next generation OPAC or discovery service like ETD@BU.

References

1. Eaton J L, Enhancing Graduate Education Through electronic Theses and Dissertations. In Fox E A, Feizabadi S, Moxley J M, Weisser C R, (Eds), *Electronic Theses and Dissertations: a sourcebook for educators, students and librarians* (New York: Marcel Dekker, 2004), p.1-7.
2. Vaughan J, Dispatches from the Field: Web-Scale Discovery: Rapidly evolving tools more important than ever, *American Libraries,* 42(1/2) (2011) 32.
3. Hoeppner A, *The Ins and Outs of Evaluating Web-Scale Discovery Services.* Available at: http://www.infotoday.-com/cilmag/apr12/Hoeppner-Web-Scale-Discovery-Servic-es.shtml (Accessed on 12 May 2016).
4. Breeding M, *The Future of Library Resources: A white paper commissioned by the NISO Discovery to Delivery (D2D) Topic Committee.* (Baltimore: NISO, 2015). Available at: http://www.niso.org/apps/group_public/download.php/14487 /future_library_resource_discovery.pdf (Accessed on 12 May 2016).
5. Hoover L, Agriculture and food related theses and dissertations available on the web, *Journal of Agricultural & Food Information,* 7 (2-3) (2006) 87-108.
6. Lubas R L, Defining best practices in Electronic thesis and dissertation metadata, *Journal of Library Metadata*, 9 (3-4) (2009) 252-263.
7. McCutcheon S, Basic, fuller, fullest: treatment options for electronic theses and dissertations, *Library Collections, Acquisitions, & Technical Services,* 35 (2011) 64-68.
8. Smith II P L, Where IR you?: Using "open access" to extend the reach and richness of faculty research within a university. *OCLC Systems & Services: International digital library perspectives*. 24(3) (2008) 174-184.
9. Wolverton R E, Hoover L, Hall S & Fowler R, *Electronic theses and dissertations: developing standards and changing practices for libraries and universities.* (London; Routledge) 2009.
10. Frank I & Rowe W C, Indexing and accessing electronic theses and dissertations: some concern for users. In Fox E A, Feizabadi S, Moxley J M, Weisser C R, (Eds), *Electronic Theses and Dissertations: a sourcebook for educators, students and librarians* (New York: Marcel Dekker, 2004), p. 343-353.
11. Keller B, Subject content through title: a masters theses matching study at Indiana State University. *Cataloging & Classification Quarterly,* 15 (3) (1992) 69-80.
12. Khurshid Z, Inprovisations in cataloging of theses and dissertations. *Cataloging & Classification Quarterly,* 20 (2) (1995) 51-59.
13. Hoover L & Wolverton R E, Cataloging and treatment of theses, dissertations, and ETDs, *Technical Services Quarterly,* 20 (4) (2003) 3-57.
14. McMillan G, Electronic theses and dissertations: merging perspectives. *Cataloging & Classification Quarterly,* 22 (3/4) (1996) 105-125.

15. McMillan G, Implementing ETD services in the library. In. Fox E A, Feizabadi S, Moxley J M, Weisser C R, (Eds), *Electronic Theses and Dissertations: a sourcebook for educators, students and librarians* (New York: Marcel Dekker, 2004), p. 319-329.

16. Genoni P, & Cowan R, Bibliographic control of Australian higher degree theses: The Future Role of the Australian Digital Theses Program. *Australian Academic & Research Libraries,* 34 (2) (2003) 81-91.

17. Koulouris A & Anagnostopoulos A, Theses e-submission tool at the National Technical University of Athens, *OCLC Systems & Services: International Digital Library Perspectives,* 36 (2) (2010) 123-132.

18. Surratt B E & Hill D, ETD2MARC: a semiautomated workflow for cataloging electronic theses and dissertations. *Library Collections, Acquisitions, & Technical Services.* 28 (2004) 205–223.

19. Bevan S J, Electronic thesis development at Cranfield University, *Program: Electronic Library and Information Systems,* 39 (2) (2005) 100-111.

20. El-Sherbini M & Klim G, Metadata and cataloging practices. *The Electronic Library,* 22(3) (2004) 238-248.

21. Sarkar P & Mukhopadhyay P, Cataloguing theses and dissertations: designing an integrated processing and retrieval system, *SRELS Journal of Information Management,* 48 (4) (2011) 377-388.

22. Han M, New discovery services and library bibliographic control, *Library Trends*, 61(1) (2012) 162-172.

23. Sarkar P & Mukhopadhyay P, Metadata and discovery of electronic theses and dissertations, In Tiwari R, Mukhopadhyay P S, Ramesha B, Mahesh G, Singh A (Eds), *Trends and Development in Library and Information Science* (New Delhi: Zenith Publications, 2012), p. 315-323.

24. Copenhaver K & Koclanes A, Impact of web-scale discovery on reference inquiry, *Reference Services Review,* 44 (3) (2016) 266-281.

25. Lundrigan C, Manuel K &Yan M, "Pretty Rad": explorations in user satisfaction with a discovery layer at Ryerson University, *College & Research Libraries,* 76(1) (2015) 43-62.

26. Balaji Babu P & Krishnamurthy P, Library automation to resource discovery: a review of emerging challenges, *The Electronic Library,* 31 (4) (2013) 433-451.

27. Goodsett M, Discovery search tools: a comparative study, *Reference Reviews,* 28 (6) (2014) 2-8.

28. Houser J, The VuFind implementation at Villanova University, *Library Hi Tech,* 27 (1) (2009) 93-105.

29. Denton W & Coysh S J, Usability testing of VuFind at an academic library. *Library Hi Tech.* 29 (2) (2011) 301-319.

30. Ho B, Kelley K & Garrison S, Implementing VuFind as an alternative to Voyager's WebVoyage interface: One library's experience, *Library Hi Tech,* 27 (1) (2009) 82-92.

31. Chickering F W & Yang S Q, Evaluation and comparison of discovery tools: an update, *Information Technology And Libraries,* 33 (2) (2014) p. 5-30. Available at: http://ejournals.bc.edu/ojs/index.php/ital/article/view/3471 (Accessed on 12 May 2016).

32. Katz D, LeVan R & Ziso Y, Using authority data in VuFind. *Code4Lib Journal*. 14 (2011). Available at: http://journal.code4lib.org/articles/5354 (Accessed on 10 July 2016).