Cataloguing Theses and Dissertations: Designing an Integrated Processing and Retrieval System

Prosenjit Sarkar, Superintendent (Library Services)
Central Library, The University of Burdwan, Burdwan-713104,
Email: prosenjit.toton@yahoo.co.in,
prosenjit.toton@gmail.com

Dr. Parthasarathi Mukhopadhyay, Sr. Lecturer
Department of Library and Information Science
The University of Burdwan, Burdwan – 713 104
Email: psmukhopadhyay@gmail.com,
psm_bu@india.com

Abstract

Theses and dissertations in print and electronic format is a valuable academic resources, this resources must be available to scholars. Library catalogue serve as interface and resource discovery tool to find this valuable academic documents. In view of the importance of organization and access to these resources, this paper explores cataloguing of theses and dissertations and its retrieval along with other types of documents in an integrated processing environment. The software framework that provides integrated processing and retrieval environment is based on open source software and open standard. Designing of the framework takes into consideration shortcomings of the maximum used library automation software implemented in India.

Keywords: ETD cataloguing, ETD-MS and MARC-21 mapping, Theses retrieval, ETD management in library automation.

INTRODUCTION: Theses and Dissertations (TDs) in print format and Electronic Theses and Dissertations (ETDs) in digital format represent academic heritage of an institution. These TDs and ETDs constitute a major vessel in scholarly communication [1]. They signify and describe most current research topics. Since long time libraries have preserved and circulated these valuable academic resources. These

Preprint version of the following paper

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

nascent knowledge objects must be available to library users. Library catalogue (or OPAC) provides an interface to find and to navigate resources of a library and thereby serves as a resource discovery tool for the library collections. Advances in Information and Communication Technologies (ICTs) and its application in libraries have changed whole scenario of library's housekeeping jobs including cataloguing. Card catalogue is replaced by automated computer based Machine Readable Catalogue and accessible from anywhere at any time.

An OPAC is the key to library's holdings as because it helps library users to search for library materials either in print format or in digital/electronic format. The online catalogue made available over the Web is called Web OPAC. The concept of Web OPAC is of recent origin. It helps library users to access cataloguing data/metadata and providing with direct access to a library's bibliographic database from anywhere at any time. This research paper starts with investigation of the state of theses/dissertations management in Web OPAC vis-a-vis library catalogue. Web OPACs of elite Indian Institutes were evaluated against carefully crafted criteria. It has been observed that majority of the institutions in India are using LibSys or SOUL as their library management systems (LMSs) and presently no Web OPAC is providing integrated search facilities of Theses and Dissertations (along with other bibliographic materials) through their LMSs. It means that user has to select bibliographic database (i.e. book or theses) before conducting search and a search query can't be forwarded to all the available bibliographic databases at a given time. Therefore, retrieval of documents against a search statement is confined to specific material types (see Fig. 1

Preprint version of the following paper

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

and Fig. 6). Again, the attributes and/or data elements that are displayed for each thesis/dissertation retrieved against a query vary greatly from software to software possibly due to lack of ISBD – like display standard.

Now, the question comes that why popular library management software are not providing integrated search environment to retrieve documents irrespective of documentary forms (i.e. books, articles, theses etc.). The reason is simple – these software are not providing integrated processing environment for different types of documents. Generally, library management software organize books/monographs, journals, electronic resources etc. by following standard content designator schemes like MARC-21 family of standards, CCF or UNIMARC, and theses and dissertations are treated on the basis of proprietary data structure (varies from software to software).

In view of the foregoing, a prototype system has been developed by using KOHA LMS for providing integrated processing and retrieval facility for all types of documents including theses and dissertations. The system maps MARC-21 bibliographic format with ETD-MS (a metadata standard for cataloguing of Electronic Theses and Dissertations) and the mapping is used for designing a thesis-specific data entry framework. This framework supports integrated processing of theses and dissertations in a library setup.

LITERATURE REVIEW: Before designing Integrated Processing and Retrieval System (IP&RS), an extensive literature review was done to ascertain knowledge about the cataloguing of Theses and Dissertations in both traditional print

Preprint version of the following paper

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

format and electronic/digital format; and also to get some ideas about IP&RS in libraries attached with Indian higher education institutes.

An Information Retrieval (IR) System contains different types of databases. Library catalogue is one kind of database. Chowdhury [2] mentioned about different factors that are to be considered before developing an IR system. He opined that developing of database is the first step of designing an IR system. Frank and Rowe [3] reported after the completion of thesis or dissertation and degree awarded, how do people know about the TDs? How libraries should store and catalogue these intellectual resources? He mentioned cataloguers have been using MARC format and Dublin Core metadata format for cataloguing theses and dissertations. He also pointed out; (i) need to increase awareness of ETD management, (ii) need to integrate bibliographical information of TDs and ETDs into the library OPAC. It means that when library users search OPAC for their required information they will find TDs along with other types of documents. This approach will help user to retrieve all types of documents available on his/her topic of search.

It is observed that researchers in India, who are working in the domain of ETD management, have not mentioned technical factors related with cataloguing of Theses and Dissertations and ETDs. The need, advantages and mechanisms, related with the development of integrated processing environment is another neglected area in the domain of ETD management. Most of the published literatures are case studies of different ETD repositories. Vijaykumar [4] mentioned that in India 8000-10,000 doctoral degrees are awarded every year. INFLIBNET has already developed an

Preprint version of the following paper

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

online database of Indian theses (IndCat – http://indcat.inflibnet.ac.in/indcat/). Urs [5] advocated that India with its enormous system of higher education is a reservoir of extensive doctoral research. No exact statistics is available for activities related with doctoral research. There is no mechanism to deposit, catalogue and archive Indian Ph.D. theses. She mentioned that annually 25,000 to 30,000 Ph.D. theses are produced in India. Vijaykumar et al. [6] reported that 39% University Librarians in India want to provide access to ETDs through Library LAN, 29% suggested access over campus Intranet and 32% are suggesting for Global access through Internet.

University Grants Commission in its regulation, which is called *UGC (Submission of Metadata and Full-text of Doctoral Theses in Electronic Format) Regulations, 2005* [7], mentioned about cataloguing of TDs and advocated about mandatory submission of TDs in electronic format and giving efforts to achieve bibliographic control of TDs in India.

Electronic Theses and Dissertations: a Sourcebook for Educator, Students and Librarians, edited by Edward Fox et al. is a valuable book for ETD researchers. This is published in 2004 and contributors of this book are eminent experts in the ETD domain [8]. The authors of this book prescribed a scheme of ETD management on the basis of NDLTD.

Another book entitled *Electronic Theses and Dissertations: Developing Standards and Changing Practices for Libraries and Universities* authored by Robert E. Wolverton Jr. et al. published in 2009 [9], containing six chapters, is a guide for cataloguing TDs and ETDs. This book enumerated several survey results about

different aspects of cataloguing TDs and ETDs. Wolverton reported that literature on cataloguing and access of Theses and Dissertations is less in comparison with other document types. He advised to include extra MARC fields (specific to ETDs) for efficient organization and seamless access to ETDs. .

SCOPE: This research paper is related with three levels of scope: -

i) To understand how Web OPACs of different Library Management Systems are functioning and providing searching and retrieval facilities for TDs along with other types of documents;

ii) To identify the problems related with the integrated processing and retrieval of TDs and ETDs; and

iii) To design software framework for the identified problems by using open source software and open standards.

At present, different Library Management Systems packages are being used in Indian Libraries, attached with the Higher Academic and Research Institutions. However, for the present research study two LMSs have been studied extensively against carefully crafted criteria, namely LibSys and SOUL, because these two software are maximum used in India. Already 2003 institutes installed SOUL (till 18th March 2010) [10]; more than 1000 Libraries are using LibSys [11].

The abovementioned two LMSs have different modules. Only cataloguing module of these LMSs (including Web OPAC) have been studied for this research work.

Preprint version of the following paper

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

OBJECTIVES: The objectives of the present study are delineated in the title of the topic. Search and retrieval of TDs along with other type documents in different Web OPACs using different Library Management Systems are discussed here and a prototype system has been designed to reduce the problems of integrated access. The specific objectives are stated below: -

i) Mapping of MARC-21 format with ETD-MS;

ii) Designing standardized cataloguing framework for TDs and ETDs;

iii) Setting a set of cataloguing rules for TDs in both print and electronic format;

iv) Designing integrated processing and retrieval system; and

v) Exploring the suitability of open source software and open standards in system design.

METHODOLOGY: Web OPAC, as an information gateway, makes possible the integration of many of the library's resources within a single access tool. The methodology for this research work can be divided into following conceptual areas:-

Part 1 - Identification of problem;

Part 2 - Solution of the problem. This part also can be divided into following areas: -

i) Identification of suitable open source LMSs;

ii) Mapping of MARC-21 format with ETD-MS;

iii) Framework design;

iv) Data entry standardization; and

Preprint version of the following paper

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

v) Integrated searching and retrieval.

Part 1 - Identification of problem: Web OPAC of elite Indian educational institutes was checked against some carefully crafted criteria from the month of January to March 2010. One such criteria is that OPAC database should have TDs/ ETDs metadata and they can be accessed through LMSs. Most of the Indian Higher Academic Libraries (with Web based OPACs) are using either LIbSys or SOUL as their LMS.

After carefully searching Web OPACs of different Libraries; Jayakar Library, Pune University [12] and Central Library, Calcutta University [13] were taken into consideration for the present study as because both the OPACs have TD metadata along with other documents in their OPAC database. Pune University is using LibSys as their LMSs and SOUL is being used by Calcutta University and result found that both the LMSs are not providing integrated search facilities of TDs along with other types of documents. Screenshots of search results of both the LMSs are given below.

Fig. 1 Web OPAC home page of Jayakar Library, Pune University.

Preprint version of the following paper

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

From the above screenshot it is clear that catalogue databases are organized according to different types of documents e.g. books, theses, manuscripts etc. and there is no option to search a topic across different material types.

Fig. 2 search query in book database.

In fig. 2 it is shown that with the search queries, Web OPAC users have to select one database. Here book database have been selected.

Fig. 3 search results of book database.

Preprint version of the following paper

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

Fig. 4 search query in theses database.

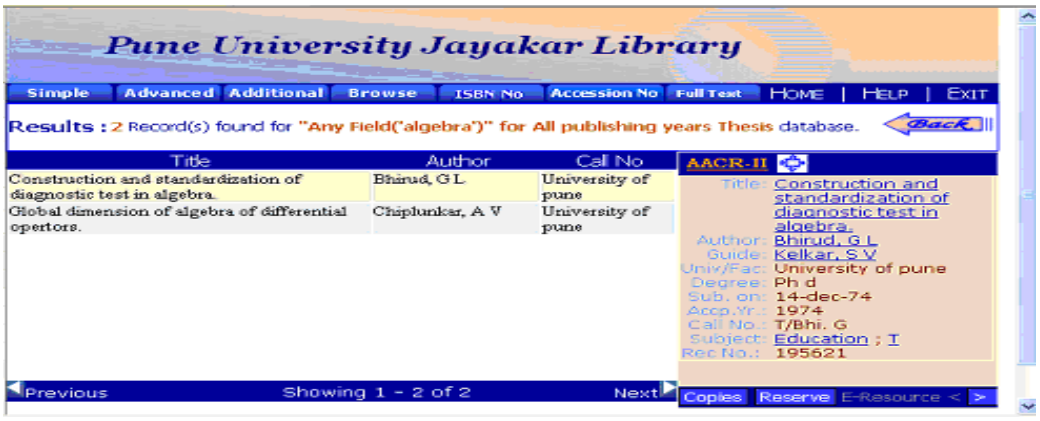

Fig. 5 search result of theses database.

From the above screenshots (from Fig. 1 to 5) it is clear that integrated searching of TDs along with other types of documents is not possible in LIbSys implementation in Pune University library. All types of documents by

Preprint version of the following paper

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

a particular author or on a particular subject can't be retrieved against search query.

Let's take the case of the Central Library, Calcutta University.

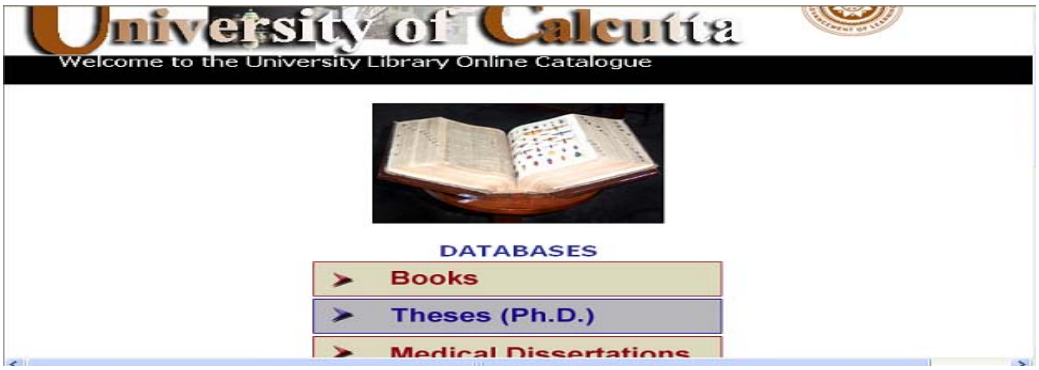

Fig. 6 Web OPAC home page of Central Library, Calcutta University.

From the above snapshot (Fig.6) it is observed that catalogue databases are organized according to different types of documents. Calcutta University is using SOUL as their LMS.

From the above six screenshots (Fig. 1 to Fig. 6) we came into conclusion that no Web OPACs of the above two LMSs are providing integrated searching facilities of Theses and Dissertations along with other types of library materials.

Part 2 – Solution of the problem towards IP&RS:

To solve this problem, a prototype system has been developed by using KOHA LMS for providing integrated processing and retrieval facility for all types of documents including theses and dissertations. KOHA allows creation and customization bibliographic frameworks for different document

types on the basis of MARC-21 bibliographic format. This research selected ETD-MS as a global data standard for ETD. The data elements of ETD-MS are then mapped with semantically related Tags/subfields of MARC-21 bibliographic data format. This crosswalk of ETD-MS with MARC-21 bibliographic format act as a base for integrated processing.

The crosswalk, as mentioned above, may be represented as below:-

Mapping of MARC-21 bibliographic formats with ETD-MS [14]

♦ dc.title	245a (Title proper)
♦ dc.title.alternative	246 (Alternative title)
♦ dc.title.translated	242 (Translated title)
♦ dc.creator	100a (Author)
♦ dc.subject	650a (Topical heading)
♦ dc.subject	653a (Uncontrolled index term)
♦ dc.description.abstract	520a (Abstract)
♦ dc.description.note	504 (Bibliographic note)
♦ dc.publisher	260a+b (Place &TD producing Institution Name)
♦ dc.cotributor	720a (Name of the guide/supervisor)
♦ dc.cotributor role	720e (Designation e.g. guide, supervisor etc.)
♦ dc.date	008 (Submission date) (character position 7-10)
♦ dc.type	655 (Document type)
♦ dc.identifier	856u (Electronic location & access information)
♦ dc.language	008 (Language) (character position 35-37)

Preprint version of the following paper

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

- thesis.degree.name 502a (Name of the degree)
- thesis.degree.level 502a (Doctoral/Masters)
- theses.degree.discipline 710b (Name of the subject)
- theses.degree.granter 502a (Name of the Institution)

This crosswalk has been utilized as a base for designing an ETD/TD-specific bibliographic framework for entering data items. The framework has also been standardized by generating pick up list supports for Leader, Control fields and Number and Code fields.

The screenshots as given below through Fig. 7 to Fig. 9 shows the thesis-specific framework and data entry activities in developing catalogue database on ETD/TD.

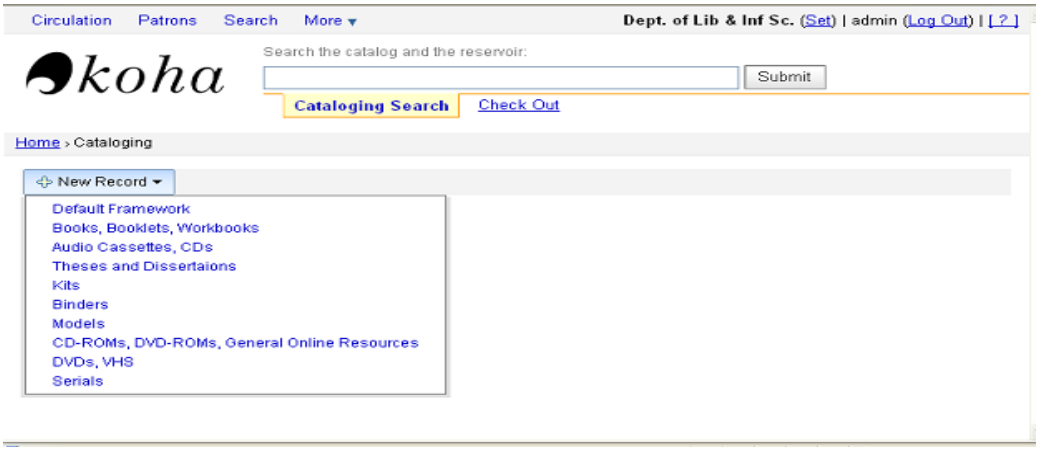

Fig. 7 Here user can select TD format from a list of different frameworks.

Preprint version of the following paper

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

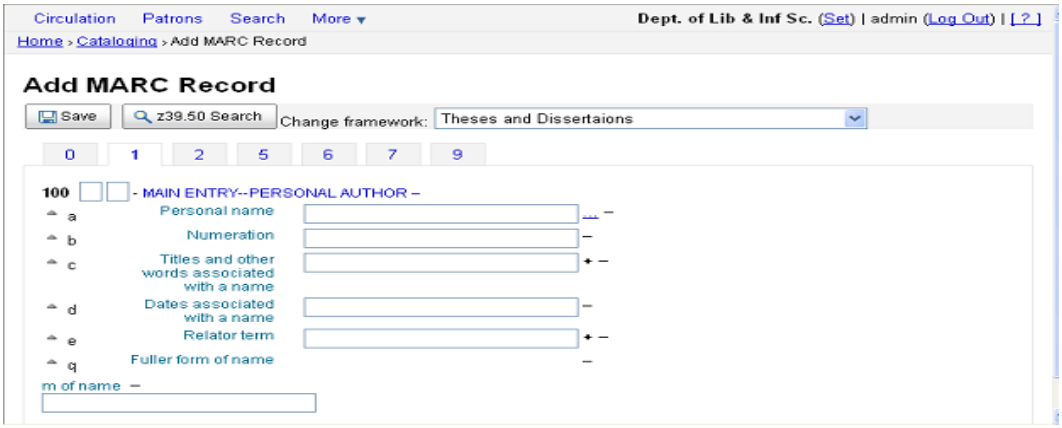

Fig. 8 this screenshot shows 100 tag and subfields of the framework.

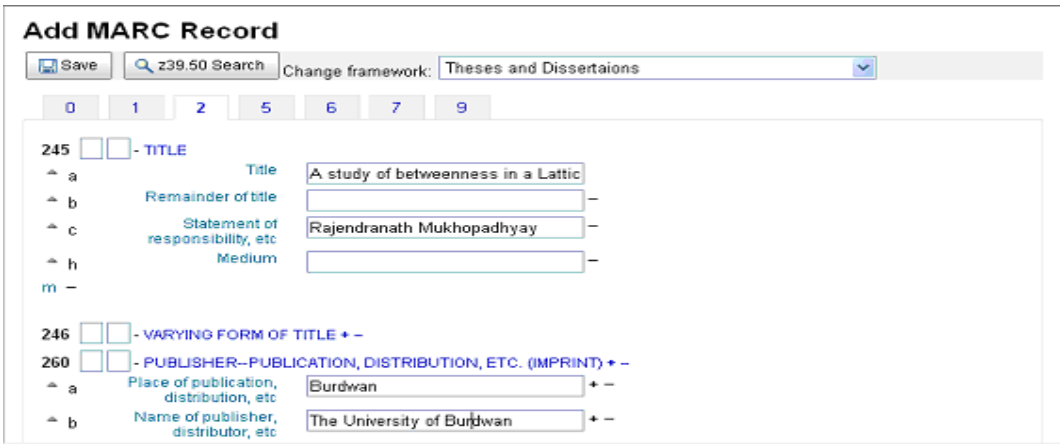

Fig. 9 this is a filled-in subfields for the tags 245 and 260.

The integrated processing ensures that data items from different document-specific frameworks are collected in a single database (Catalogue database of KOHA) and single framework (MARC-21 bibliographic framework). This feature of KOHA helps in retrieving different document types against a search query.

Preprint version of the following paper

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

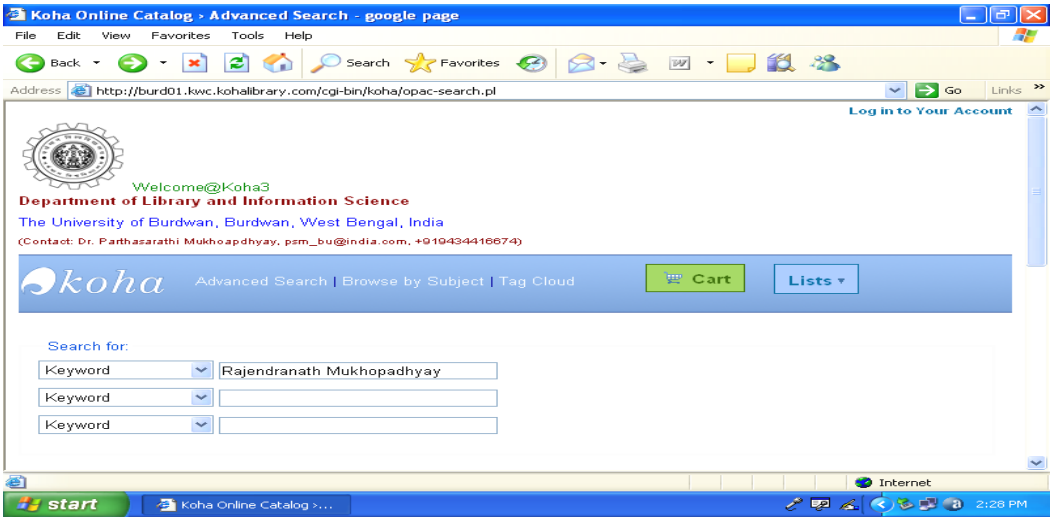

Fig. 10 keyword entered in the advanced search box of KOHA Web OPAC.

In the above screenshot it is seen that a keyword was entered in the advanced search box and results of this search queries is shown in the next screenshot.

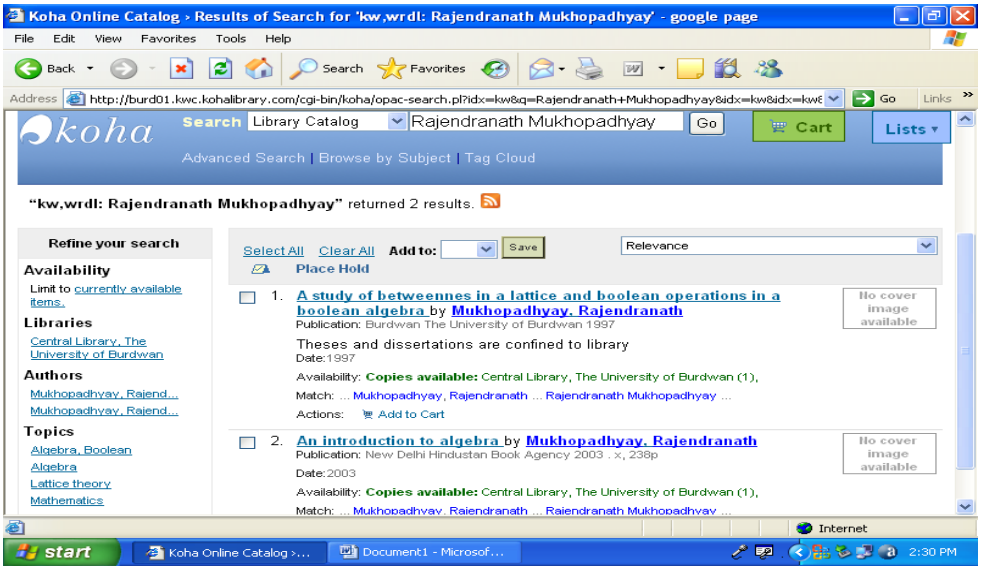

Fig. 11 search results found from KOHA LMS.

From the Fig. 11 it is seen that 2 results found for searched keyword 'Rajendranath Mukhopadhyay'. First one is thesis and second one is book.

Preprint version of the following paper

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

KOHA retrieved two documents from the database searched for above mentioned keyword. One document is book and another document is thesis.

From the above screenshot of search results it is clear that KOHA LMSs providing integrated search facilities of Theses and Dissertations along with other types documents like Books etc.

CONCLUSIONS: Generally Web OPAC is designed to tell the library users about the collections of a particular library. All Web OPACs allow users to search library's collections from any remote place at anytime, through their LMSs though their search and retrieval facilities may differ. In this research topic it is found that KOHA is more robust in providing search and retrieval facilities in comparison to SOUL and LibSys. KOHA is providing integrated search and retrieval facilities of Theses and Dissertations along with other library materials. Limitations of this research is that some other LMSs viz. Alice for Windows, New-Gen Lib, Troodon are being used by some Institutions. Due to non-accessibility of their Web OPAC, these three LMSs are beyond the scope of this research work.

References:

1. Wolverton (Robert E) et al. Electronic theses and dissertations: developing standards and changing practices for libraries and universities. London: Routledge, 2009, p 22.

Preprint version of the following paper

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

2. Chowdhury (G G). Introduction to modern information retrieval. London: Library Association Publishing, 1999, p 17-19.

3. Frank (Ilene); Rowe (Walter C). Indexing and accessing electronic theses and dissertations: some concerns for users. In Edward A Fox … et al. (Eds.), *Electronic theses and dissertations: a sourcebook for educators, students, and librarians.* New York: Marcel Dekker, 2004, p343-353.

4. VijayKumar (J K); Murthy (T A V). Need of a digital library for Indian theses and dissertations: a model on par with the ETD initiatives at international level. Retrieved August 24, 2007, from http://eprints.rclis.org/archive/00005655/

5. Urs (Shalini R). Vidyanidhi- the evolving Indian digital library of electronic theses initiative. Retrieved August 12, 2008, from http://edoc.huberlin.de/conferences/etd2003/urs-shalini/PDF/urs.pdf

6. VijayKumar (J K) et al. Introducing ETDs in universities: an Indian perspective. Retrieved August 12, 2008, from http://eprints.rclis.org/5670/3/vijaykumar_jk-paper.pdf

7. University Grants Commission (2005). UGC (Submission of Metadata and Full-text of Doctoral Theses in Electronic Format) Regulations, 2005. Retrieved June 25, 2007, from www.ugc.ac.in/new_initiatives/etd_hb.pdf

Preprint version of the following paper

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

8. Fox (Edward A) et al. (Eds.) Electronic theses and dissertations: a sourcebook for educators, students, and librarians. New York: Marcel Dekker, 2004.

9. Wolverton, Op. cit.,

10. http://www.inflibnet.ac.in/soul/ (Retrieved February 17, 2010).

11. http://www.libsys.co.in/ (Retrieved February 17, 2010).

12. http://lib.unipune.ernet.in:8080/opac/html (Retrieved February 17, 2010).

13. http://www.caluniv.ac.in/opac/index.html (Retrieved February 17, 2010).

14. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html (Retrieved August 12, 2008).

15. Mukhopadhyay, P. (2004a). Measuring Web Impact Factors: A webometric study based on the analysis of hyperlinks. *Proceedings of the IASLIC XXI National Seminar on Information Support for Rural Development* (pp. 411–425).

16. Mukhopadhyay, P. (2004b). Organization and dissemination of digital objects through web and CDROM: a framework for Indian libraries. *Proceedings of the International Conference on Digital Libraries,* (p. 470–8).

17. Mukhopadhyay, P. (2005). Use of FRBR as model of bibliographic description in online environment. *Vidyasagar University Journal of*

Library and Information Science, 51–69.

18. Mukhopadhyay, P. (2006a). Five laws and ten commandments: The open road of library automation in India. *Proceedings of the IASLIC 22nd National Seminar on Open Source Movement – Asian Perspective, at IIT, Roorke, 2006* (pp. 27–36).
19. Mukhopadhyay, P. (2006b). VidyaOnline: Design and Development of a FOSS based Virtual Learning Environment on Library and Information Science at Vidyasagar University, West Bengal. *Proceedings of the Conference on ICT for Facilitating Digital Learning Environment* (p. Paper–A).

Preprint version of the following paper

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