ISBN 978-1-84626-xxx-x
Proceedings of 2009 International Conference on Computer Science and Information Technology
Singapore, Oct 9-11,2009, pp. 232-236

Enhancing Full text Search Capability in Library Automation Package: A Case Study with Koha and Greenstone Digital Library Software

Anuradha, K.T. [1] and Sivakaminathan, R.[2]

[1] Technical Officer, National Centre for Science Information, Indian Institute of Science, Bangalore 560 012

[2] Project Trainee, National Centre for Science Information, Indian Institute of Science, Bangalore 560012

Abstract. There are many library automation packages available as open source software. Though they provide advanced features of searching and retrieving of bibliographic records, none of them facilitate full text indexing and searching. Most of the available open source digital library software facilitate indexing and searching of full text documents in different formats. An effort is made here to enable full text search feature in a widely used open source library automation package viz., Koha, by integrating it with an open source digital library software viz., Greenstone Digital Library Software (GSDL), by making use of Search and Retrieval by URL (SRU) feature available in both Koha and GSDL. The implementation of this can be accessed at: http://dharmaganja.ncsi.iisc.ernet.in:8082

Keywords: Library Automation Package, Digital Library, Full text search, SRU

Introduction

Libraries have been looking forward for the better technologies even before the onset of the computers. The introduction of the typewriter into libraries was a revolutionary concept in late 1800's. Later stages of modernization witnessed the introduction of unit record equipment, the move of offline computerization, use of online systems. By the mid-60, computers were being used for the production of machine readable catalogue records (MARC) by the Library of Congress (LOC). Between 1965 and 1968, LOC began the MARC I project, followed quickly by MARC II. MARC was designed as way of "tagging" bibliographic records using 3-digit numbers to identify fields. In 1974, the MARC II format became the basis of a standard incorporated by National Information Standards Organization (NISO). This was a significant development because the standards created meant that a bibliographic record could be read and transferred by the computer between different library systems (Nelson, 1991).

Library automation software, integrating all the activities and routines of the library is essential software for the libraries and is referred to as Integrated Library Automation Package (ILAP). An ILAP means an enterprise resource planning system for a library, used to track items owned, orders made, bills paid, and patrons who have borrowed the items. In other words it is one where all the library activities such as acquisitions, cataloguing, circulation, serials, and the Online Public Access Catalogue (OPAC) are automated. There are many ILAP available in the market that meets the needs as well as budgets. However, with the open source software movement catching up, a few open source library automation package (LAP) are also available, which are comparable with any commercial LAP. To name a few: Koha, Evergreen, OPAL, PhpMyBibli, OpenBook, OpenBiblio. Among these Koha is the first open source library automation software and is widely used (Anuradha, 2009).

A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system and the software used to enable this functionality is a Digital Library Software (DLS). More detailed working definition of digital library is available at: http://www.diglib.org/about/dldefinition.htm There are many open source DLS available. To name a few: Greenstone Digital Library Software (GSDL), Dspace, Fedora.

1.1 Need for the study:

Most of the open source library automation software are either in developing stage or do not have full functional modules. Of them, a few satisfy key functional requirements of ILAP and support essential modules like acquisition, cataloguing, circulation, serials control and OPAC but still full text search and retrieval feature is not available in any ILAP. This feature is available in most of the DLS. Hence an attempt is made here to integrate DLS into ILAP. Doing so will result in an ILAP with full text indexing, searching and retrieving option. Also, DLS can index documents of large size and in different formats such as PDF, DOC, RTF, PPT. A few DLS stores documents in compressed form there by saving hard disk space.

1.2 Objective:

The main objective is to facilitate full text indexing and searching in an ILAP by integrating it with a DLS. An effort is made to add full text indexing and searching feature in Koha Version 3.0.2 by integrating it with GSDL Version 2.80.

Methods

Among open source ILAP, Koha is widely used. GSDL is a quite popular DLS. Both the software are compatible with many library standards such as SRU/W feature, Z39.50 feature, MARC record import. To facilitate full text searching in Koha, Search and Retrieval through URL (SRU) functionality available in both Koha and GSDL is used. Hence a SRU request is sent to GSDL from Koha for fulltext searching.

1.1. Overview of Koha

In 1999 when the Horowhenua Library Trust (HLT) in New Zealand, was looking for a Y2K compliant replacement for their library system, Katipo Communications proposed a new system, using open source tools to be released under the GPL. Koha (the Maori word for 'gift' or 'donation') went live at HLT in January 2000, and was the world's first open source ILAP and is distributed under GNU GPL license.

Latest version is Koha-3.0.2 (Linux platform only) and Koha 2.9.x (for Windows and other platforms) (http://koha.org). It runs on different platform like Linux, MacOSx, FreeBSD, Solaris, and Windows. Originally developed on the Linux OS, is written in Perl, uses Apache web server, has better support for multi-RDBMS like MySQL, PostgreSQL. OPAC interface is in CSS with XHTML. It supports all major library standards such as MARC record import/export, Z39.50 and SRU/W feature. Koha-3.x supports Zebra fulltext search engine as backend, in addition to MySQL/PostgreSQL. Records are stored internally in an SGML-like format and can be retrieved in MARCXML, Dublin Core, MODS, RSS, Atom, RDF-DC, SRW-DC, OAI-DC, and Endnote; and the OPAC can be used by citation tools such as Zotero. Koha's default installation supports running Zebra which is configured to support SRU queries on bibliographic and authority data. Zebra itself is capable of detecting Z39.50 or HTTP and responding with SRU if the incoming request is HTTP.

1.2. Overview of GSDL:

GSDL (http://www.greenstone.org), a suite of software for building, publishing and distributing digital library collections, either on the Internet or on CD-ROM. It is compatible with many library standards such as SRU/W feature, Z39.50 feature, MARC record import. These features of Greenstone make it a very good selection for integrating it with library automation package for full text indexing and searching. It is

produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO.

1.3. Full text search with SRU

SRU/W (Search/Retrieve through URL or Web service) is a web-services based protocol for querying the databases and returning the search results. It uses the Common Query Language (CQL) as the format for submitting the queries. Although CQL is a formal language for representing queries to information retrieval systems, it has been designed to be human readable and writable. It allows both simple and very complex and powerful queries. Search results from SRU/W are in XML format (Retrieved online on June 9th, 2009 http://www.loc.gov/standards/sru/index.html).

Implementation Details

To implement full text search and retrieval feature in Koha V3.0.2, the software is downloaded from www.koha.org and installed on Ubuntu 8.04. Koha 3.0.2 available for Linux platform contains four components viz., OPAC, Intranet, Daemons and Database. OPAC is used search and locate library holdings. There are two components in the OPAC package. The Main-OPAC contains all the Perl scripts and the HTML-OPAC contains all the HTML templates for all the HTML pages. In OPAC, search can be done in different ways: advance search, search by subject and basic search. Koha allows field based searching by making use of field tags such as Title, Author, ISBN. INTRANET is the back office and the front desk side of the system. There are 9 components in the INTRANET package. The Main-INTRANET component contains all the main Perl scripts to handle the navigation, login, logout and provide connection to the other components. The HTML-INTRANET component contains all the HTML templates for all pages for the INTRANET side of the system. Daemons contain all the scripts for all the daemons in the system. There is currently only one daemon -- the Z39.50 Daemon component -- in the Daemons package. This component provides the connectivity to Z39.50 servers for querying of library material using the Z39.50 protocol. Database components make use of MySQL database, which has 116 tables. One of the tables is 'biblioitems', which contains all the metadata information and passes to Zebra search engine in MARC format for indexing.

GSDL v2.80 downloaded from www.greenstone.org is also installed. After indexing for each collection (a set of documents form a collection) 9 folders are created viz., archives, building, index, etc, import, perllib, images, macros, tmp. The documents to be indexed are kept in import. The collect.cfg file in etc folder determines the look and feel of the collection.

A few freely available e-books are downloaded and are catalogued in Koha, giving all the required bibliographical details. The same e-books are indexed in GSDL. In Koha OPAC full text search option is added and they are searched through SRU feature available both in Koha and GSDL. In the following paragraphs full text indexing and full text searching modules are explained in detail.

3.1. Full text indexing

The main objective of this module is to catalogue fulltext documents in Koha and index it in GSDL for carrying out full text search. In Koha cataloguing module, the URL of the full text document is specified under the tag 856, which is repeatable field, there by multiple URL for the same document can be given. After filling up the required cataloguing details, the record is saved. After saving the catalogue information, a unique document number is assigned by Koha for each catalogued document. This document number along with other required metadata details and full text document location obtained through the catalogue form is passed on to GSDL for carrying out full text indexing. This is enabled by modifying additems.tmpl in Koha. A PHP script is invoked to carry out indexing in GSDL through command line collection building option.

3.2. Full text search feature

The main objective of this module is to enable the full text search in Koha and display the results of GSDL in Koha OPAC. For this purpose, four different perl scripts are written, viz., fulltextsearch.pl,

fulltextsearch1.pl, fulltextsearch.tmpl, fulltextsearch1.tmpl. In fulltextsearch.pl query term is obtained from the user and passed to GSDL through SRU technique. The URL which is passed is split into 4 parts with question mark as the delimiter. The four parts are GSDL location in the system, collection name, required query and Do option. Following example explains this.

Query string passed throguh SRU from Koha to GSDL:

First part (GSDL location) : http://vidya-mapak.ncsi.iisc.ernet.in/cgi-bin/library?

Second part (GSDL collection name) : e=p-01000-00---off-0koha--00-1--0-10-0---0---0prompt-10---4-------0-1l--11-en-50---20-about---00-3-1-00-0-0-11-1-0utfZz-8-00 koha

Third part (Query term and other search details): fqf=TX&t=0&q=software

Fourth part (DO option) : perform

In Koha Explain pragma can be configured via *the koha-conf.xml* file and can be retrieved by a search interface using a query such as:
http:*<serverhost>:9999/biblios?version=1.1&operation=explain Queries for results can be run using syntax such as the following:*
http:<serverhost>:9999/biblios?version=1.1&operation=searchRetrieve&query=harry&startRecord=1&maximumRecords=20&recordSchema=mods
Koha SRU link looks like: http://<IP>:Port/cgi-bin/koha/opac-search.pl?idx=ti&q=Query

The GSDL retrieved results are displayed in another window. To get back to Koha search window, a link is given to the document in Koha from the GSDL retrieved results window. Thus by indexing the bibliographical details in Koha and full text in GSDL, full text searching in Koha is enabled by sending a SRU request to GSDL. By linking the documents in GSDL to that of Koha, after viewing the full text documents in GSDL, user can get back to Koha OPAC page. The implementation of this can be accessed at: http://dharmaganja.ncsi.iisc.ernet.in:8082

Acknowledgement

This work is carried out as part of a project entitled “Enhancing Knowledge Innovation Culture of Libraries through Union Catalogues” (http://dharmaganja.ncsi.iisc.ernet.in), carried out at National Centre for Science Information, Indian Institute of Science, Bangalore, India, funded by International Development Research Centre, Canada (http://www.idrc.ca).

References

Anuradha, K.T. (2009), “Koha: An Overview”, Presented in NCSI-IDRC Workshop on Integrated Library Automation Packages: Basics of KOHA and NewGenLib. Retrieved July 2nd 2009, Website: http://dharmaganja.ncsi.iisc.ernet.in/workshopnew/index.html

Digital Library Federation (1998), "A Working definition of digital library", Retrieved July 3rd, 2009, Website: http://ww.diglib.org/about/dldefinition.htm

Greenstone Digital Library Software (2005), “About Greenstone“, Retrieved April 16th, 2009, Website: http://www.greenstone.org

Koha (1999), ”Koha: Features”, Retrieved July 2nd , 2009, Website: http://www.koha.org

Library of Congress (2007), “SRU: Search/Retrieval via URL”, Retrieved July 3rd, 2009, Website: http://www.loc.gov/standards/sru/simple.html

Nelson, N. M. (Ed.). (1991), “Library technology 1970-1990: Shaping the library of the future”, Research contributions from the 1990 Computers in Libraries Conference (In Supplements to Computers in Libraries, No. 25.). Westport, CT: Meckler.