OPEN SOURCE SOFTWARE AND BUILDING DIGITAL LIBRARY USING GSDL SOFTWARE

Nabajyoti Das

Abstract

The paper presents and analyses the development of open source software and the building of digital library with reference to the GSDL software. It describes digital library and its importance in details. It also discusses with illustrations how to build a digital library by using the open source software i.e. 'Greenstone Digital Library'. Searching and browsing full text information is also described taking example from New Zealand Digital Library. It identifies the strength and limitation of GSDL. It is predicted that GSDL is becoming popular digital library software because of its flexibility and low cost/no cost of ownership. The author conclude that because of its cost effectiveness and flexibility GSDL can be a powerful tool in bridging the gap of digital divide in India.

Keywords: Digital Library, GSDL, Open Source Software

1. Introduction

Now-a-days Information Technology (IT) becomes an indispensable concern of developing countries like India. One of the vital components of IT is software. Presently the developing countries are fully confronted the copyright and illegal copying of software, which widely affect on the prospect of very high and recurrent software cost. All commercial software companies distribute their software in compiled form. We know that once software has been complied into a computer readable form, it is practically impossible to understand the internal functioning of the software, and it can not be modified. By doing so, the software companies gain monopoly on improving their software by adding features or fixing bugs and this is how the software becomes expensive. By following this practice the software companies have gained a monopolistic market and have also drive rivals and technological predation one upon others out of business.

These strategies adopted by commercial software companies, has given rise to an unhealthy dependence on proprietary software, huge expenditure on licensing fee, growth of gray market in pirated software, troublesome environment in local software industries and most importantly discouraged innovation in the software industry at global level. In this background a development, which is attracting the interest, is the freedom of research and development offered by Open Source Software (OSS). In OSS the source code (human readable set of instructions, which makes a software) is distributed along with the executable form (the computer readable set of instruction, which makes a software, also known as complied form of a computer software) (Suman and Bhardwaj; 2003: 9). There is a hope that this Open Source Software will solve such issues.

2. What is Open Source Software?

"Open Source Software" (OSS) is a marketing name for Free Software, coined in Feb 1998 as an attempt to overcome the confusion over the word "free" in the English language. Open Source refers to the fact that the source code of the software is open to and for the world to take, to modify and to reuse. Open Source Software, as used in this article, refers to software distributed in source form which can be freely modified and redistributed, i.e. Open Source Software is freely modifiable and redistributable software. More precisely, it refers to four kinds of freedom, for the users of the software :

- The freedom to run the program, for any purpose,
- The freedom to study how the program works, and adapt it to users' needs. Access to the source code is a precondition for this.
- The freedom to redistribute copies so one can help another.
- The freedom to improve the program, and release the improvements to the public, so that the whole community can get the benefits. Access to the source code is a precondition for this.

A program is Open Source Software if users have all of these freedoms. Thus, users should be free to redistribute copies, either with or without modifications, either gratis or charging a fee for distribution, to anyone anywhere. Being free to do these things means that the developer does not have to ask or pay for permission. The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

Therefore, another group has been using the term Free Software instead of Open Source Software. They consider the Open Source Software is something close to Free Software, but not identical. They prefer the term 'Free Software' because, once we have heard that it refers to freedom rather than price, it calls to mind freedom. The word 'open' never refers to freedom. However, this author have no intention to discuss the differences of these two terms. OSS is copyrighted and distributed with General Public License (GPL) terms to design to ensure that the source code will always be available.

Some Open Source Software are mentioned bellow :

Operating Systems -	Linux (or GNU/Linux), FreeBSD/OpenBSD, NetBSD, GNU/Hurd, etc.
Windowing Systems -	The X Window System, XFree86, etc.
Desktop Environments -	GNOME, KDE, GNUStep, XFce, etc.
Languages -	GNU C/C++ , Perl , Python , Tcl, etc.
Web Browsers -	Mozilla (Netscape 6), etc.

Server-type software - Samba, Apache, PhP, Zope, MySQL, PostgreSQL, etc.
Office Suites - Open Office, KOffice, etc.
Productivity Applications - ABIWord, GNU Image Manipulation Program, Dspace, GSDL, KOHA, etc.
General Utilities - GNU Utilities, etc.

3. Digital Libraries

In the present day context due to the wide use of information technology and digital/electronic storage media, it becomes a challenge to the library professionals how to acquire, organize, store and retrive various information available in digital form. This has initiated the concept of creation of 'digital library'. Digital library has a number of machine readable study materials as well as other publications such as text, images, sound, videos, and any combination of text, images, sound, videos etc. in digital form and facilitates remote access to several databases. The basic concept behind a digital library is to exploit the facilities of information technology with a mission of sharing resources available globally for providing nascent information to the users' community at right time. Hence, a typical digital library has a media server connected to high speed networks, and we also call it 'Virtual Library'. Unlike a conventional library where users are provided with physical materials from many sources, a digital library is a group of attributed repositories that users see as single repository in a digital form. The functioning of a digital library is controlled by machines with minimum human interventions.

3.1. Why Digital Library

There are a number of reasons for creating digital library. In a library, even in an automated library searching information in printed books in response to some queries will be time consuming and sometimes impossible also. Because, information are available in the books are not structured information. But, in digital resources searching information on a particular query is possible if those are organized with the help of metadata sets. In a digital library the delivery of the materials is different from removing of a book from shelf and checking out. This is because the book in digitised form can be copied to a user's computer for reading, but the book still remains in the source computer (server). It can again be loaned in the name of another user. Again sharing and circulation of printed books in a wide area is also time consuming, but, it is very easy to share the digital resources through network, even multiple user can use digital resources at a time available in a single source. Therefore, it becomes essential to acquire, organize, store and disseminate information available in digital form.

Here we have to consider two possibilities: (1) the materials oroginally available in digital form, and (2) the materials in digitised form. In the second one there is an involvement of the creation of

digital information from conventional stage, which generally is a two stage process. The first stage is digitization. This is obviously the conversion of physical medium, say a printed book, into digital representation. It has no effect on the information content of the original material. The second stage is the computerization process to make the computer extract information from the digitised image by using Optical Character Recognition (OCR) software. This stage allows the information from original book or document to be made available to the computer, and make possible to index the text for retrieval and is also able to reformat the text for different forms of output such as compressing, changing the font size & type and graphical manipulation etc. However, once digitised, the problems are not over, searching, retrieving and delivery may be problematic in real life.

4. Greenstone Digital Library (Gsdl)

To build up a wel organized digital library we may opt for open source software, preferably **'Greenstone Digital Library' (GSDL).** Greenstone Digital Library Software is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato headed by Prof. Ian H. Witten. GSDL is supported by the UNESCO and the Human Info NGO Belgium for spreading the benefits of this software to developing countries. It is ***open-source***, multilingual software, issued both source and binary code under the terms of the GNU General Public License (GPL). Greenstone is powerful and flexible software which is of great potential interest to libraries and information centers and other public and private institutions in South Asia. It provides a new way of organizing information and making it available over the Internet. ***Collections*** of information comprise large numbers of ***documents*** (typically several thousand to several million), and a uniform interface is provided to them. The structure of a collection is determined by a configuration file. The collections range from educational journals to oral history, from newspaper articles to technical documents, from visual art to videos, from MIDI pop music collections to ethnic folksongs, etc.

Making information available using Greenstone is far more than just "putting it on the Web." The information becomes searchable, browsable, and maintainable. Each collection, prior to presentation, undergoes a "building" process that, once established, is fully automatic. This creates all the structures used for access at run-time. Searching utilizes various indexes of text and/or metadata, while browsing utilizes metadata such as title and author. When new material appears, it is incorporated into the collection by rebuilding. Greenstone supports the metadata harvesting of collection(s) in 38 different languages including Bengali, Hindi and Kannada.

4.1. Required Software

To create the digital collection(s) using GSDL some other associated software are required. The GSDL can be downloaded (free) from www.greenstone.org and www.sourceforge.net. Greenstone

CD-ROMs have also been published by the United Nations and other humanitarian agencies for distribution in developing countries. Following are the associated software :

i) Java Runtime Environment (JRE) version 1.4 or above (Free download from http://Java.sun.com/j2se/downloads.html)
ii) Image Magick Software (Free download from www.imagemagick.org)
iii) Web Browser Software (Internet Explorer, Netscape, etc.) (download from www.msn.com; www.netscape.com).
iv) Web Server Software – Apache, PWS/IIS.

4.2. Platforms

Both the source code and binaries of GSDL are available for Windows (95, 98, 2000, XP) and Linux (Red Hat and other Clones). It is also available for Mac and Sun Solaris, but here the source code has to be compiled.

4.3. Source Code

Source code is available in GCC and Perl for Linux and VC++ and Perl for Windows.

4.4. GSDL Installation

After downloading the above software from the respective websites, those have to install. Regarding installation of GSDL it is very essential to install the Java Runtime Environment (JRE) before installation of GSDL. GSDL has four types of installation setup –

Local Library – It is default setup. It has web server built-in and is suitable for building and viewing the Greenstone collections in a stand alone system. It is available for Windows platform only.

Web Library – It is recommended for those who wanting to serve Greenstone collections on the web. It requires a separate web server like Apache and Microsoft PWS/IIS.

Source Code – Only the source code will be installed and binary executables will not be installed.

Custom – This setup allows installing any or all of the features provided by the above three setup types.

4.5. GSDL Interface

GSDL has two separate interactive interfaces- User Interface and Librarian Interface. End users access the digital library collections through the User Interface, which operates within a web server.

The Librarian Interface is a Java-based graphical User Interface that makes it easy to gather material for collection, enrich it by adding metadata, design the searching and browsing facilities and build and serve the collection to the end users.

4.6. Metadata Formats

GSDL has four predefined metadata sets, such as Dublin core (DC), RFC 1807, New Zealand Government Locator Service (NZGLS), and Australian Government Locator Service (AGLS). New metadata sets can also be defined using Greenstone's Metadata Sets Editor (MSE).

4.7. Interoperability

GSDL can harvest documents over OAI-PMH (Open Archives Protocol for metadata Harvesting) and include them in a collection. Any collection can be exported to METS and the Greenstone can ingest documents in METS form. Any collection can be exported to Dspace ready for Dspace's batch import program, and any Dspace collection can be imported into Greenstone.

4.8. Plug-Ins

Relevant plug-ins has to use to ingest externally prepared metadata in different forms. In Greenstone built-in plug-ins exist for- XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, Dspace, METS. Plug-ins is also used to ingest documents in different formats. For textual documents, the plug-ins are: PDF, Postscript, Word, RTF, HTML, Latex, ZIP, Excel, PPT, Email, TEXT, Index, Open Document, Book, etc. For multimedia, the plug-ins are – For Images - GIF, TIFF, JPG, JIF, PPT, W3Img, ; For Audio - Video - DAT, MP3 , AVI, WAVE, MPEG (1, 2, 3, 4), MIDI, Real Media, etc.

4.9. Collection Building

For building a digital collection the library professionals have do work with Greenstone Librarian Interface (GLI). The GLI is a Java based interface for building digital library collections and this provides very user-friendly approach. The librarian interface can be run in one of four modes: Librarian, Assistant Librarian, Library System Specialist and Expert users. Modes control the level of detail within the interface, and can be changed through 'Preferences' in the 'File' menu. The GLI supports six basic activities, i.e. **Download, Gather, Enrich, Design, Create and Build & Preview** the collection, which may be considered as modules of GLI. Among these six activities the later five are indispensable for making Greenstone collection.

To build the digital collection the GLI has to select from the program section of the start menu or desktop. The opening window of GSDL - GLI will appear (**Fig. 1**). Here the option 'File' and then

'New' have to select, then a popup window will appear to give the name and the description of the content of the collection (**Fig. 2**).

Then one metadata element set has to select out of the built-in four metadata set. The activity **'Gather'** is for gathering the source documents that will comprise the collection (**Fig. 3**). For this the target files/folders have to drag from the workspace (Left side) and drop it in the Collection (Right Side).

'Enrich' is to enter the data of the source document(s) in the metadata fields to assign metadata for each source document (**Fig. 4**). Metadata have to prepare for all the source documents in the collection.

Fig-1: Starting Window of GLI -1

Fig. 2: Starting Window of GLI - 2

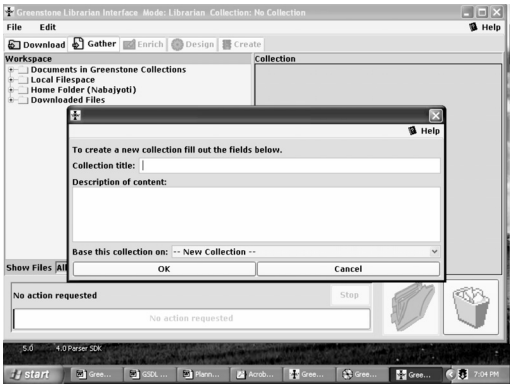
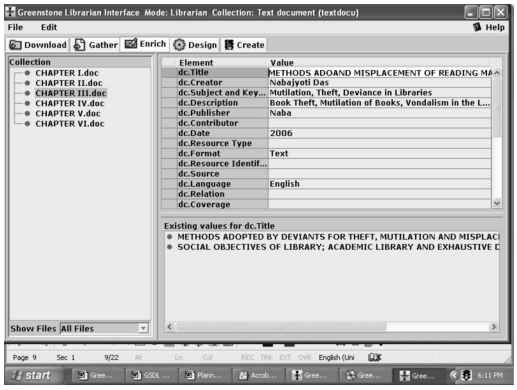


Fig. 3: Gathering the Collection

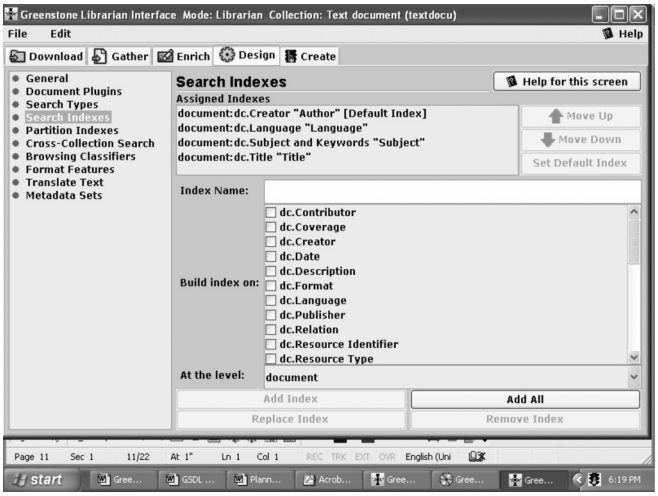
Fig. 4: Window for Enriching the Collection



'Design' is for specifying collection configuration in terms of indexes, classifiers, display formats, document plugins, etc **(Fig. 5)**. Automatic extraction of simple metadata such as Title, Date,

etc. is possible. Explicit metadata has to be extracted via 'Classifiers' e.g. Subject, Author, Organization, etc.

Fig. 5: Indexing in the Design Panel

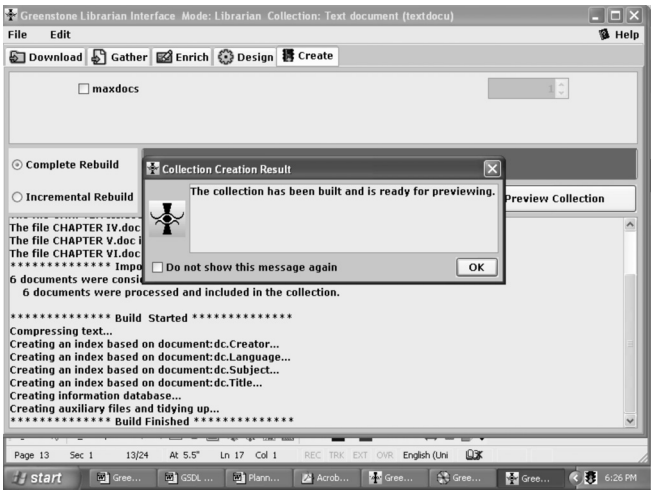


'Create' is for initiating the building process and **'Build Collection'** is to build the collection in the GSDL system in compressed form of the source documents and coordinating all the activities done in the previous modules **(Fig 6).**

'Preview collection' is a link to the 'User Interface', by which one can view the current collection.

'Download' is for downloading files and websites from the Internet.

Fig. 6: Create and Building Collection



4.10. Collection Search And Browse :

Collection in GSDL can be searched and browse by using the 'User Interface' i.e Greenstone Digital Library. The library is browsed by web browser (e.g. Internet Explorer). Fig. 7 shows the Home page of the New Zealand Digital Library.

Fig: 7 : The New Zealand Digital Library Home Page

For searching 'Full Text Search, Metadata (Field) Search, Boolean Search' can be applied and full text/ full audio – video can be browse accordingly. In the following we have entered to the World Environmental Library of New Zealand Digital Library.

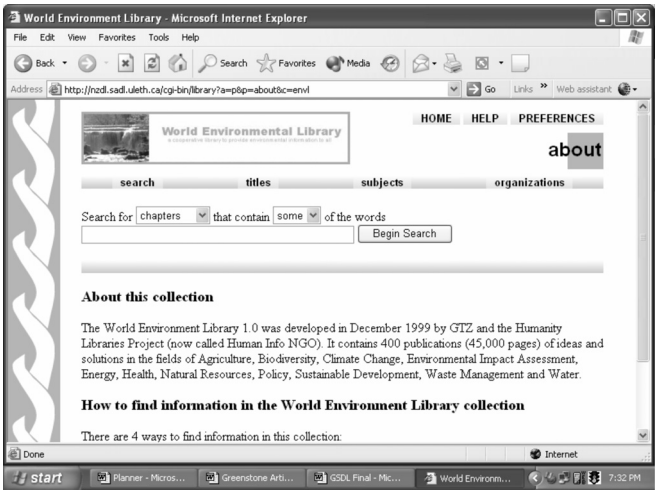


Fig. 8: Searching the World Environmental Library (WEL) collection.

Source: http://nzdl.sadl.uleth.ca/cgi-bin/library?a=p&p=about&c=envl

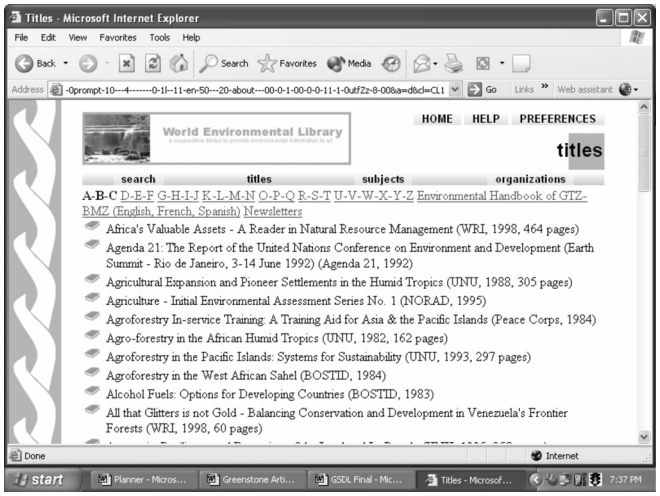


Figure 9: Browsing the WEL collection by all Title.

The titles are arranged alphabetically. We may get the full text of the required title by clicking the title and then the respective steps such as content, chapter, etc. Similarly, we can search or browse documents by subject, organization or by advance search using the 'design search' option. The collection can be distributed through Web and CD-ROM.

4.11. Observation

In GSDL some remarkable strengths and some limitation have been observed. The **Strengths** are - Configurability: Content extraction for indexing, presentation layout, metadata for browsing and field-based searching (though searching by little is difficult). Extensibility: Plug-ins for content extraction, Unicode for Multilanguage support, source code availability, Fulltext search on variety of document formats, XML, Unicode, Dublin Core support, Data compression, CD-ROM publishing, The Z39.50 protocol is supported for accessing external servers and for presenting Greenstone collections to external clients etc. The **Limitations** are - Interactive content updating and management not possible, No duplicate identification, Metadata handling appears to be little complex, Linux version seems to be more robust than Windows, Hangs while processing some documents during collection building – no way to gracefully handle.

5. Conclusion

Growth of OSS concept and GSDL can be viewed as an opportunity for the library & information professionals to come out from under the yoke of the proprietary platform and high software license fees. Because of its cost effectiveness and flexibility, GSDL can be a powerful tool in bridging the gap of digital divide in India. The aim of the Greenstone software is to empower users, particularly

in universities, libraries, and other public service institutions, to build their own digital libraries. Digital libraries are radically reforming how information is disseminated and acquired in UNESCO's partner communities and institutions in the fields of education, science and culture around the world, and particularly in developing countries. It is hoped that this software will encourage the effective deployment of digital libraries to share information and place it in the public domain. The support effort in India is coordinated by the CDDL at Indian Institute of Management, Kozhikode (greenstonesupport@iimk.ac.in) in collaboration with the **Greenstone development team** in order to ensure effective promotion and support for Greenstone in South Asia.

References

1. Arora, Jagadish (2006). Building Digital Libraries: an overview. DESIDOC Bulletin of Information Technology, 21 (6), 2006, 3-24.
2. Bainbridge, David, McKay, Dana and Witten, Ian H. (2004). Greenstone digital library developer's guide, Newzealand, Dept. of Computer Science, University of Waikato.
3. Baker, Thoma. The open source movement. Available at http://www.server.tiac.or.th/ tiacweb/ Baker/Section1_6.html. Accessed in Oct. 2007.
4. Information resource management using IT: a training programme, November 6 – 11, 2006, Centre for Media & Rural Documentation, National Institute of Rural Development, Hyderabad.
5. New Zealand Digital Library project (http://www.nzdl.org).
6. Rajashekar, T.B. (2002). Greenstone Digital Library Software (GSDL): overview. Available at http://scigate.ncsi.iisc.ernet.in/raja/opendl/gsdl-overview.pdf. Accessed on Nov. 1, 2007.
7. Suman, Yogesh and Bharadwaj (2003). Open source software and growth of Linux: the Indian prospective . DSEIDOC Bulletin of Information Technology, 23 (6). 9 -16.
8. Witten, Ian H. , Bainbridge, David and Boddie, Stefan J. (2001). Open source digital library software. D – Lib Magazine, 7 (10), Oct. 2001. Available at http://www.dlib.org/ dlib/october01/ witten/10witten.html
9. Witten, I.H., Bainbridge, D. and Boddie, S. J. (2001) Power to the people: end-user building of digital library collections. Proc Joint Conference on Digital Libraries, Roanoke, 94-103. Available at http://www.acm.org/pubs/articles/proceedings/dl/ 379437/p94-witten/p94-witten.pdf

Mr. Nabajyoti Das is presently working as Librarian cum Archive Officer in Jyoti Chitraban Film & Television Institute. He has received his MLISc degree from Gauhati University. He is also working as lecturer (part time) in the Department of Library & Information Science, Gauhati University.