Greenstone Digital Library Management System – a functional review based on selected criteria

Ramesh Parichi and Faizul Nisha

Introduction

Open Source Digital Library Software is an application where source code and rights are provided under a GNU General Public License. The users can download and use, and modify and redistribute customized versions. Dspace, Eprints, FEDORA, Greenstone, etc, are some examples of open source Digital Library Management Systems (DLMS). DLMS helps in structuring and managing digital libraries; it provides all the essential and advanced functionalities such as searching and browsing, acquisitions and metadata bases over the Internet. This paper provides an overview of the usage and utility of Greenstone3 based on selected criteria to help library professionals in learning about various features available in it. The present work is an outcome of our experience of working with Greenstone3. It embodies our reflections as well as assessment.

Greenstone digital library – introduction

Greenstone Digital Library (GSDL) is a suite of software developed by the New Zealand Digital Library Project at the University of Waikato in 1995. The distribution and promotion partners for Greenstone are: UNESCO; Human Info NGO, Belgium; NCSI, Bangalore; UCT, Cape Town; Dakar, Senegal; Almaty, Kazakhstan; and an individual.(Tramboo *et al.*, 2012).

The Greenstone website was launched in 1998, and later in 2000 Greenstone collaborated with UNESCO. Greenstone has released several versions of the software and a Java-based Greenstone librarian Interface (GLI) was developed in 2003. Information about user groups and mailing lists are posted on the website. It is distributed on SourceForge, and the present Greenstone version is 3.06.

Functional review and critical analysis of Greenstone

The Greenstone Digital Library Management Software is functionally reviewed and analysed based on selected broad criteria, viz., user friendliness, administration, customization, international metadata standards and interoperability, search and browse capabilities, security aspects and statistical functions.

User interface

Greenstone software is available for all operating systems: Windows, Unix/Linux and Mac OS-X. Windows installation is the default, and the configuration steps and time taken is much less compared to DSpace. Greenstone can be installed on desktops, laptops and servers, and be made available online via standard web server software (e.g. Apache). Additional softwares like ImageMagick, Ghostscript and Tomcat web server are installed along with the core installation. Not only is it easy to install, the software is also easy to maintain.

Greenstone is available in 59 international languages. Greenstone uses Unicode thus enabling automatic language recognition of multilingual collections. Hence, any language supported by the browser can be processed, converted dynamically and displayed in a consistent manner (Wikipedia, 2015).

Greenstone provides three modes for acquisitions: command line, web interface and GLI. In the earlier version of Greenstone, GLI worked in four modes: Library Assistant, Librarian, Library System Specialist and Expert. The modes are designed according to the level of expertise of the user. Beginners can choose the Librarian mode which is the default mode. The higher the mode, Greenstone provides more options for building a digital library. In Greenstone version 3, there are only three modes – Library Assistant, Librarian and Expert. The Librarian interface is available in English, French, Spanish and Russian.

Greenstone accepts almost all kinds of document formats. The documents in the collections can be technical documents, newspaper articles, educational journals, oral history, visual art, videos, MIDI pop music collections, ethnic folksongs, etc.

Plugins are available for all kinds of documents such as PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, email (various formats), source code, multimedia (including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio and a generic plugin that can be configured for other audio formats such as MPEG, MIDI, etc. Suitable plugins can be installed if the document format is not included. Plugins can also be downloaded and added from the Internet.

Administration

An "administrative" function in Greenstone enables specified users to authorize new users to build collections, protect documents, so that they can only be accessed by registered users with a password, examine the composition of all collections and so on. Greenstone also maintains user logs. The administrator will have the privileges of adding users and deleting or creating groups. At present, the groups are Administrator and Cobuilder.

End-users can build collections in three ways. Using the GLI, they can import existing collections, the web interface depositor and through Command Line Interface using Perl Command line build scripts. The GLI has a user-friendly graphical interface developed using Java. The collections are stored in several collect directories. Greenstone2 only had a single collect directory and a single run time Common Gateway Interface program. The collection is a batch mode process and not interactive.

There is a provision to upload documents via web, MediaWiki, OAI, Z39.50 and SRU. The documents can also be uploaded from the local system folders into the collection using drag and drop method in the Gather section of the Librarian Interface. The configuration details of the collection structure are stored in the file "collection.cfg". The file structure in Greenstone is hierarchical.

In GLI, documents can be downloaded using web, mediawiki, OAI, Z39.50 and SRU (Search/Retrieve via URL). The documents are first imported into the extensible markup language (XML)-compliant Greenstone Archive Format, then searchable indexes and the collection information database including the hierarchical structure that support browsing is built, before making them online.

The Librarian interface is developed using Java and it has following sections. They are:

- *Gather*: The files and documents which are to be uploaded into the collection are selected in this module. Greenstone provides a provision to upload documents from the system directories, directly from the web, specified URLs and also a Wiki. The system is user friendly, so users can use drag and drop methods to upload documents.
- *Enrich*: In this section, the user enriches the documents by adding metadata to each uploaded document. All metadata elements need not be added.
- *Design*: The user can choose from the plugins available according to the document format. Greenstone has provided the facilities to index, classify and design the searching and browsing facilities that the collection will offer to the user
- *Create*: This section deals with building the collection. An option for a minimal build or a full build is available. The collection is then stored on a server along with a preview option. The collection can also be built using command line interfaces using Perl command building scripts and the tomcat is restarted to see the new collections. This method is convenient if the users have knowledge of Perl language.
- *Format*: This section deals with developing the appearance of the digital library. Experience with HTML programming and other languages is helpful.
- *Convert*: The collections web programming build can be either published on an external drive or provided online. The section provides options to convert the digital library into an offline library. The import of already built collections in other versions is possible. The collection folder is copied into the site and conversion scripts have to be run. This conversion script will create a new collection Config.xml and buildConfig.xml using the old collect.cfg and build.cfg files.

New documents can be uploaded into existing collections via the web interface, Greenstone Depositor. This web interface was withdrawn from Greenstone 2.4x and the GLI was enhanced and strengthened. The incremental collection building in Greenstone2 is achieved through the web interface depositor. However, the Apache Web server should be enabled. The metadata is added to the uploaded document, and the collection can be rebuilt using the create panel or command line interface.

The two main components in Greenstone are "receptionists" and "collection servers". The user interacts and inputs through the receptionist, the request is then dispatched to the appropriate collection servers after analysing for the user interface, and the processed information is returned back to the receptionist for providing it to the user. The receptionist reads in the interfaceConfig.xml file and loads up all the different action classes. Other actions may be loaded on the fly as needed. Collection server is a specific type of service cluster that act as an abstract mechanism that handles the content of the collection, while receptionists are responsible for the user interface. The communication between the receptionist and a collection server is through a defined protocol which depends on the configuration of the computer where the DLMS is running.

The create panel has a build collection button which initiates the collection building process. A progress bar indicates the time completed and remaining. After successful collection, building the preview collection is clicked to launch a web browser showing the home page of the collection.

This build process can also schedule tasks and the frequency of build, add update or delete. Errors can be corrected in collection building.

The collections are published both online and on a self-installing Windows external drive. Compression is used to compact the text and indexes. A Corba protocol supports distributed collections and graphical query interfaces. The collections can be exported to other formats such as METS, XML, Dspace, etc., and also to an external drive.

In Greenstone, security is through username and password. Presently, only two groups the administrator and collection builder provide security support. Authenticated users are allowed to add, modify and delete the document collection. The other security features such as LDAP, Shibboleth and IP authentication that are available in DSpace are not available in Greenstone. Logs of user activity and errors in administration and maintenance page can record all queries made to every Greenstone collection.

Customization

Greenstone allows new document and metadata formats to be accommodated by writing plugins for them in Perl. If the uploading document is in the same proprietary format, it

can be merged into the collection automatically without writing plugins. It has been observed that sometimes the plugins, particularly HTMLPlug, PDFPlug, failed for some documents. In Greenstone, the documents which have similar characteristics are grouped into collections and are uploaded using one uniform interface. The user has the provision to create a user-defined collection structure or use the existing collection structure.

Most of the interfaces are defined by Extensible Stylesheet Language Transformation (XSLT) files which are stored in the interfaces/interface-name/transform. The changes made to these files will reflect immediately. The modified files can be copied into the transform directory of sites and collections folder so that this interface file is used in preferences to the other interface XSLT files when using the collection. A completely new interface can also be designed by keeping the old one intact. The new interface directory needs the name of the directory and the images, transform directories along with interfaceConfig.xml files. The language of the interface can be modified using the preferences page. The customization can also be collection specific.

In Greenstone3, customization of presentation of collections is based on XSLT and other agents that govern the definite functions of Digital library. XSLT templates are used to format all the parts of the pages. The GSF format language contains format elements which help in formatting the presentation. The user can display a different icon for each document in a list depending on which organization it came from using gsf-switch element. These templates modified can be collection specific and have to be put at prescribed standard places. Each collection has a set of service text strings which are used to display information in the collection. These strings are found in .properties files in the WEB-INF/classes file of the home folder. The text can be changed and replaced with another informative text strings.

In Greenstone2, the appearance of all pages is governed by macro files and these macro files are stored in Greenstone/macros and images and CSS in Greenstone/web/style. The customization of these macros enables various presentations of pages into different languages; print variables into the page text (e.g. number of search hits). Macro files are stored in a gsdl/macros folder, where each file defines one or more "packages" (A "package" is a group of macros) loaded on startup (note difference between Local and Web Library) listed in etc/main.cfg**.** The macro files have .dm extension, the macros start with underscores and end with underscores and a macro value is stored in curly brackets. The important macros are base.dm, style.dm, English.dm, English2.dm, home.dm and query.dm. These files control the display style of each page, text in the pages and description and appearance of the home page.

Collection-specific macros are stored in gsdl/collect/mycol/macros/extra.dm or include argument [c=collectionname] for each macro. Personalization of home pages is also possible. Presentation of search/browse results and collection interface is determined by "format" strings and "macros". Interface customization can be done using Front Page Design, Header for the Digital Library, Collection Icon, Cover Images, etc. A Corba protocol allows agents (e.g. in Java) to use all the facilities associated with document collections. Finally, the source code, in C++ and Perl, is available and accessible for modification.

ImageMagick is a robust collection of tools and libraries to read, write and manipulate an image in many image formats including TIFF, JPEG, PNG, PDF, PhotoCD and GIF. With ImageMagick, one can create images dynamically, making it suitable for Web applications. Ghostscript is a suite of software based on an interpreter for Adobe Systems Postscript and Portable Document Format page description languages. It is used in conversions. Both these are available by default during installation and are required to build collections and for advanced conversions of Postscript documents.

The page display of the resultant digital library, including the display page that appears on clicking the browsing classifiers on making a search, is governed by the features provided below the format tab. The format has four sections:

(1) general, in which all the settings and information about the collection and web page can be entered;
(2) the Search pane allows one to change the search indexes, level and partitions;
(3) format makes provisions to change the appearance of the digital library by selecting, adding and modifying available format features using XML; and
(4) the Translate Text option is used when language specific strings are to be added.

International metadata standards and interoperability

In Greenstone, the metadata provided explicitly or derived automatically from various sources is used as the raw material for browsing indexes. For most electronic documents, Dublin Core is used; however, Greenstone also offers a provision to use other metadata schemas. Users can define metadata interactively within the Librarian interface. These metadata sets are predefined and they are Dublin Core, RFC 1,807 (a format for bibliographic records), NZGLS (New Zealand Government Locator Service) and AGLS (Australian Government Locator Service). However, in Greenstone3, the assigned metadata sets are Dublin Core, Greenstone metadata set (gs) and Extracted Greenstone metadata 1.1 (ex).

New metadata sets can be defined using Greenstone's Metadata Set Editor. Plugins are used to ingest externally prepared metadata in different forms, and plugins exist for: XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace and METS. Metadata database file types such as MARC, OAI, CDS/ISIS, BibTex, Refer and Procite can be imported into Greenstone, but their metadata cannot be immediately viewed or edited in the Librarian Interface. Exploding a metadata database file splits into individual records with viewable and editable metadata. This process is irreversible.

The explodable files will have a green icon in the collection tree.

Greenstone is highly interoperable using contemporary standards. It incorporates a server that can serve any collection over the Open Archives Protocol for Metadata Harvesting (OAI-PMH). Any collection can be exported to (Metadata Encoding and Transmission standard) METS. Greenstone can import documents in METS form (The Library of Congress, 2015).

The collections of the Dspace and Greenstone can be exported and imported into each other's collections. DSpace imports the Greenstone collection through batch import program and a DSpace collection can be imported into Greenstone. Both DSpace and Greenstone have the concept of a single document being present in alternative formats or versions, flat file structure and they follow Dublin Core wherein the fields are repeatable; it is possible to use multiple occurrences of Identifier metadata to point to these alternatives. Both DSpace and Greenstone uses METS standards, and the file structures being different, they cannot be logically mapped. StoneD is a joint distribution, and a bridged version of Dspace and Greenstone has been developed to solve the above problem (Witten *et al.*, 2005).

Search and browse capabilities

The search and browse features can be designed using the elements available in the Collection Design panel of GLI: Document plugins, Create Search Indexes, Browsing Classifiers and the Cross collection search panel.

Document plugins are used to convert the document into the Greenstone archive format. The Search indexes determine whether to confine the search to paragraph, chapter or the entire text of the document. Through Create Search Indexes, new indexes are added and default indexes are removed. The default text index should not be removed. The arrangement of these indexes indicates the appearance.

The Browsing Classifiers help browse the collection by a metadata element. New classifiers can be added, removed and arranged in an order. Use of A-Z CompactList Classifier brings bookshelf icons under the browsing classifier. It groups together the documents that appear multiple times with the same metadata and does not differ with A-Z list in any other manner. The user can design and implement user-defined explicit metadata browsing structures by writing "classifiers". The metadata can be either Hierarchical or List type. The Collection configuration file "collect.cfg" determines content conversion, extraction and building of indexes and browsing structures, indexes, classifiers and plugins. Finally, the cross-collection search panel shows a checklist of available collections where collections are selected in which the search should be done. Greenstone supports full-text tagging for hierarchical document browsing, automatic text extraction and indexing.

Greenstone allows multi-language support for browse and search interface. The search and browse of the collection can be Metadata or field-based. The advanced levels of search techniques such as Boolean, Ranked, provisions for search history, search terms and highlighting is available in Greenstone. The strings in the obtained search results can also be formatted using the configuration file.

The Word search is provided as Greenstone constructs full-text indexes from the document text which means that these indexes enable searching of any word in the full-text of the document.

Further, indexes can also be searched for particular words, combinations of words or phrases.

The search can also be performed using metadata elements such as author, title, date, keywords and so on. Many document collections also contain full-text indexes of certain kinds of metadata. The results are displayed according to the relevancy to the query performed and the desired document can be opened with a mouse click from the result list (University of Waikato, 2007).

The search can also be performed by subject that is represented by bookshelves that come complete with a table of contents and users can choose a chapter or subsection to open it, expand the full table of contents or expand the full document into a browser window. The non-textual material is either linked into the textual documents or accompanied by textual descriptions such as figure captions to allow full-text searching and browsing. The browsing facilities vary between different collections and even within a collection. Browsing and searching interfaces are constructed during the building process, according to collection configuration information.

As with Dspace, non-textual material is either linked to the documents or accompanied by descriptions (such as figure captions) that allow full-text searching and browsing through a web browser. All searching and browsing structures are built directly from the documents themselves. No links are inserted by hand, but existing links in originals are maintained. The scouting for new material and rebuilding the indexes is done all without manual intervention.

Unique features

The popularity of Greenstone can be observed in the statistics mentioned in Table I.

The major strengths of Greenstone can be expressed in the words of scalability, flexibility, configurability and extensibility. All the configuration settings for content presentation, design of layout, extraction, of metadata, indexing, field-based searching and browsing, data compression techniques make Greenstone user friendly. As there is only batch mode; no interactive content management is possible. There is no check method for identification of duplicate items. Greenstone on Windows is less robust than with other operating systems such as Linux. It was reported that during collection building process of some documents, the system gets hanged up and the user has to turn off the browsers and the interfaces, and no alternate method is available to restore the system. Adding metadata to the content can be complex, especially for beginners (GSDL, 2015).

The following Table II provides a summary of the features of Greenstone DLMS.

Table I.
Facts about Greenstone

Distributed via SourceForge since:	11/2000
Average downloads per month since then:	4500
Currently running at:	4500
Proportion of downloads that are documentation:	33%
Proportion of downloads that are software:	67%
Proportion Windows/Linux/Mac/Source	82%/10%/3/5%
Number of people on Greenstone email lists:	750
Number of countries represented:	82
Number of messages per month (excluding spam):	165

Notes: The popularity of Greenstone can be observed in the statistics mentioned in the factsheet of the Greenstone website

Conclusion

The Greenstone DLMS is an easy to use and customizable architecture to build and manage online digital libraries. The choice of the software always depends on type/format of collection material, method of distribution, technical requirements such as the operating system and time frame for setting up a Digital Library. Librarians are most familiar and comfortable with Windows OS and GSDL, and its unique features make it attractive for libraries to build their institutional repositories, digital

Table II.
Features of Greenstone digital library management system

Feature	Greenstone
Year of development	1997
Type, license and pricing	Free and open source software. Distributed through SourceForge.com, GNU and distributed in cooperation with UNESCO
Latest release	Greenstone2.86 and 3.06 for Windows, Linux and Mac
Associated software	Apache, PERL, GNU C++ compiler, JAVA and GNU Database manager
Platforms	Linux, Unix, Windows and Mac
Databases	Its own
Programming language	C++, Perl and Java
Resource identifier	OAI identifier
Updating process	Free. Download the new version. Help and installation procedure is available
OAI-PMH	Yes
Supported item types for storage and rendition	Can store and manage all types of content
Metadata formats	Dublin core, qualified DC, METS, NZGLS (New Zealand Government Locator Service) and AGLS (Australian Government Locator Service); and RFC 1807
User interface functions	End-user deposition and multilingual support Command Line Interface, Reader interface and the Librarian interface Plugins-new document and metadata formats; classifiers-new metadata browsers; and multilingual: Documents and interfaces Chinese, Arabic, Maori, Russian, etc. (+European); multimedia: video, audio collections exist
Thumbnail preview	Images, audio, video
Browse and search functions	Field specific and Boolean logic Browsing can be done on any field Full-text and fielded search Flexible browsing facilities Metadata-based (Dublin core) Collection specific Hierarchical phrase browsing supported Creates all access structures automatically
Interoperability	Z39.50, OAI-MHP and METS
Syndication	–
Authentication and authorisation	User groups
Statistical reporting	Count of full records
Web server	Apache/IIS
Services and training	Training is provided

collections and then disseminate these to their users. The functional review on the features, usage and utility emphasize that Greenstone has both merits and problems. The software has been reviewed on selected criteria, and it is suggested to use Greenstone DLMS system as the most suitable and convenient for installation, uploading of documents and usage for small- and medium-sized organizations. The customization and entering metadata descriptions using GLI appear to be simple. The users find Greenstone having a minute edge over other DLMS. The present study can be used by libraries to compare Greenstone with other DLMS, learn the features and usage of Greenstone and set up Digital Libraries.

REFERENCES

Greenstone Digital Library Software (GSDL) (2015), available at: www.greenstone.org/factsheet (accessed 8 April 2015).

The Library of Congress (2015), "METS", available at: www.loc.gov/standards/mets/mets-profiles.html (accessed 8 April 2015).

Tramboo, S., Humma, S.M.S. and Gul, S. (2012), "A study on the open source digital library software's: special reference to DSpace, EPrints and Greenstone", *International Journal of Computer Applications*, Vol. 59 No. 16, available at: http://arxiv.org/ftp/arxiv/papers/1212/1212.4935.pdf (accessed 8 April 2015).

WIKIPEDIA (2015), "Greenstone software", available at: http://en.wikipedia.org/wiki/Greenstone_(software) (accessed 8 April 2015).

Witten, I.H., Bainbridge, D., Tansley, R., Huang, C.Y. and Don, K.J. (2005), "Stone D: a bridge between Greenstone and DSpace", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/witten/09witten.html (accessed 8 April 2015).

University of Waikato (2007), "Overview of Greenstone", *Greenstone Digital Library Users Guide*, Chapter 1, available at: www.greenstone.org/manuals/gsdl2/en/html/Chapter_overview_of_greenstone.htm (accessed 8 April 2015).

Faizul Nisha (*faizi123.drdo@gmail.com*) is based at DESIDOC, Delhi, India.