UNIT 8  CREATING DIGITAL LIBRARIES USING GSDL

Structure

8.0  Objectives
8.1  Introduction
8.2  Technical Features
8.3  Installation of GSDL on Windows
8.4  Greenstone Interfaces
8.5  Collection Building In Greenstone
8.6  Summary
8.7  Answers to Self Check Exercises
8.8  Keywords
8.9  References and Further Reading

8.0  OBJECTIVES

After going through this Unit, you will be able to:

• explain the technical features of Greenstone Digital Library (GSDL) Software;
• install GSDL on your system; and
• build a digital collection for the web as well as CD-ROM for your library.

8.1  INTRODUCTION

Greenstone is an open-source, multilingual software, issued under the terms of the GNU General Public License for building and distributing digital library collections. 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. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM in the form of a fully-searchable, metadata-driven digital library.

Greenstone has been produced by the New Zealand Digital Library Project at the University of Waikato, and is now being further developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium.

The exact user base for Greenstone is unknown. However, since it is being distributed on SourceForge, since November 2000, it has been found that the average downloads per month since then is around 4500.

The advantages of GSDL are:

• It is based on FOSS platform and has active community supporting it.
• It is Multi-platform application and can run on various operating system platforms, including Windows (any version), Linux, Sun Solaris, and Mac OSX. It is available in both binary (executable) and source code form for the Windows (all versions), Linux, and Mac OS X operating systems and in source code form for other operating systems (Unix).
• A Greenstone Collection can be served on the World Wide Web or it can be exported to a CD-ROM and accessed from the CD-ROM or local hard disc without the need for Internet connectivity.

• Greenstone can build indexes from full text documents and also metadata associated with these documents. It supports creation of indexes for various metadata fields, either automatically extracted or manually assigned.

• It uses Perl-scripting, MG(PP) or Lucene for indexing, Apache (or built-in webservice), XML, which are proven technologies.

• Greenstone lets you build collections of multimedia documents such as audio, video, and pictures accompanied by textual description or metadata to allow searching and browsing.

• UNICODE compliant facilitating building, searching and browsing documents in any Unicode-compliant language.

• Separate modules are available for different uses:
  – JAVA-based interface for management
  – Web-browser based access to collections
  – CLI client : remote collection building

• Multi-metadata (with editor)
• Practical GLI interface for editing/managing GSDL
• Plug-ins for most document formats also available as well as for crosswalks for ISIS, Dspace, e-mails, MARC, MARCXML.

The Unit has been adapted from the Greenstone official documentation and the IMARK tutorial developed by FAO. Both the documents are available under the terms of either the GNU General Public License (http://www.gnu.org/licenses/gpl.html) and the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), for distribution and modification. The documents used are listed in the References and Further Readings section for further reference and you may refer them for further details.

8.2 TECHNICAL FEATURES

Multiplatform user friendly application

Greenstone runs on all versions of Windows, Unix/Linux, and Mac OS-X. The process of installation is quite simple. The default Windows installation does not require any configuration. End users routinely install Greenstone on their personal laptops or workstations. The Institutional users, however, generally run it on their main web server, where it interoperates with standard web server software i.e. Apache.

Interoperability

It is highly interoperable, based on contemporary standards. Greenstone can harvest documents over OAI-PMH and include them in a collection. Greenstone can ingest documents in METS (Metadata Encoding and Transmission Standard) form. This facilitates export and import of any collection to and from DSpace through DSpace batch import program.
Digitisation and Digital Libraries – DSpace and GSDL

Interfaces

Greenstone has two separate interactive interfaces, the Reader interface and the Librarian interface. End users access the digital library through the Reader interface, which operates within a web browser. The Librarian interface is a Java-based graphical user interface (also available as an applet) that makes it easy to gather material for a collection (downloading it from the web where necessary), enrich it by adding metadata, design the searching and browsing facilities that the collection will offer the user, and build and serve the collection.

Metadata formats

Users define metadata interactively within the Librarian interface. Unlike DSpace Greenstone allows several sets of metadata, including locally produced ones to be merged. The metadata sets are predefined:

- Dublin Core (qualified and unqualified)
- RFC 1807
- NZGLS (New Zealand Government Locator Service)
- AGLS (Australian Government Locator Service)

All metadata are stored in XML-format with the documents. Metadata can also be extracted from XML-statements within the documents. It can be assigned easily through the GSDL Librarian interface using Greenstone’s Metadata Set Editor. “Plug-ins” are used to ingest externally-prepared metadata in different forms, and plug-ins exist for: XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace and METS.

Document formats

Plug-ins are also used to ingest documents. For textual documents, there are plug-ins for: PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code. For multimedia documents, there are plug-ins for: Images (any format, including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio, and a generic plug-in that can be configured for audio formats, MPEG, MIDI, etc.

Languages

One of Greenstone’s unique strengths is its multilingual nature. The reader’s interface is available in the following languages: Arabic, Armenian, Bengali, Catalan, Croatian, Czech, Chinese (both simplified and traditional), Dutch, English, Farsi, Finnish, French, Galician, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Kyrgyz, Latvian, Maori, Mongolian, Portuguese (BR and PT versions), Russian, Serbian, Spanish, Thai, Turkish, Ukrainian, Vietnamese

The Librarian interface and the full Greenstone documentation (which is extensive) is in: English, French, Spanish, and Russian.

In GSDL the server (library.exe) uses PERL-scripts to create web-pages and forms to deal with the library of documents and its indexes. The documents are stored in their native format as such (PDF, DOC, HTML, XML etc.) which are converted (‘imported’) as XML in a collection with their text-only content. ‘Plug-ins’ for each type of content extract words from the documents and pass them
Creating Digital Libraries
Using GSDL

Metadata are also stored in XML. A web-interface allows searching, browsing results and opening full-text documents either in original or converted format.

There are three indexers available in GSDL:

- MG (‘Managing Gigabytes’) : at section level (=field), Boolean or ranked
- MGPP : word level indexing (field, phrase + proximity) with Boolean+ranking
- Lucene (from the Apache SF) : field+proximity indexing but either on whole document or section, Boolean+ranking plus : single-character wildcards and range-searching; allows incremental collection building (not possible with MG(PP))

Unlike DSpace, GSDL allows several sets of metadata, including locally produced ones, even merged. Dublin Core (v.1.1) is provided together with RFC 1807, Development Library Subset, as well as LOM required for indexing learning objects. All metadata are stored in XML-format with the documents and can also be extracted from XML-statements within the documents. Metadata can be assigned easily through the GSDL Librarian interface. One limitation is that since GSDL does not use a DB for handling its XML-data, this imposes real limitations on speed.

Self Check Exercise

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.

1) Enumerate technical features of GSDL.

...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................
...........................................................................................................................................................................................

8.3 INSTALLATION OF GSDL ON WINDOWS

Before installing the software, be sure you have all the hardware and software requirements!

Hardware and software requirements

Storage requirements:
- 50MB for a binary installation
- 155MB for compiling Greenstone from source code
- 200MB for optional Greenstone demonstration collections
- 5MB for documentation
- 24MB for Greenstone’s “CD exporting” function
Software:

- Java Run-time Environment (JRE) version 1.4 or above (Install JRE before installing GSDL) - JRE is required for GLI
- [Not required for default Windows installation] Web Server (Apache Recommended)
- PERL - gets installed automatically
- C++ compiler, if you wish to compile the source code (Visual Studio or GCC)
- A Web Browser

There are different options for getting the GSDL software:

1) UNESCO CD-ROM (version 2.70) or FAO IMARK CD-ROM, (but this is an earlier version 2.51) which contain the Greenstone software, plus documented example collections, four language interfaces (English French Spanish Russian), the Export to CD-ROM package, the ImageMagick graphics package, the Java runtime environment, and an installer that installs all of these.

2) IITE Digital Libraries in Education CD-ROM, or a Greenstone workshop CD-ROM. This CD-ROMs contains the tutorial exercises and a set of sample files to be used for these exercises apart from the requisite software listed above.

3) Download directly from http://www.greenstone.org that contains the latest version of Greenstone.

You will need Java to run Greenstone. You might already have it installed on your system otherwise, download it from http://java.sun.com. To work with image collections, you need ImageMagick (from http://www.imagemagick.org).

Most Greenstone CD-ROMs have AutoPlay feature and start the installation process as soon as they are inserted into the drive. If installation does not begin by itself, locate the file setup.exe and double click it to start the installation process.

If you download Greenstone over the web then just double-click installer.

If Greenstone is already installed on your system then completely remove the old version before installing a new one. You need not remove any pre-packaged collections that you may have installed for this.

The following steps need to be carried out to install Greenstone:

1) Install the Java 2 Runtime Environment (latest version).
2) After installing J2RE, go for GSDL folder choose setup gsdl 2.70.
3) Choose setup Language. English (US) is the default. We choose English
4) Welcome to the InstallShield Wizard for the Greenstone Digital Library Software. Click <Next>
5) License Agreement. Accept the agreement and then click <Next>
6) Choose location to install Greenstone. Leave at the default and click <Next>
7) Setup Type. Leave at the default (Local Library) and click <Next>
8) (For older installers you must now select collections. Leave at the default, Documented Example Collections, and click <Next>.)

9) Set admin password. Choose a suitable password and click <Next> (If your computer will not be serving collections online, the password doesn’t matter)

10) Click <Install> to complete the installation

11) Files are copied across and Installation is complete.

If you are installing from a CD-ROM, the installer will offer to install ImageMagick, and Java, if necessary.

To invoke the Greenstone Reader’s interface, go to the Greenstone Digital Library Software item under Programs on the Windows Start menu and select Greenstone Digital Library. To invoke the Librarian interface, go to the same item and select Greenstone Librarian Interface.

**Installing ImageMagick on a Windows system**

Once Greenstone has been installed, ensure that ImageMagick is installed on your system, if you wish to build any image collections. If you are installing from a Greenstone CD-ROM, you will be asked whether you want to install ImageMagick: say **Yes**. If you are not, you will need to download ImageMagick (from http://www.imagemagick.org). To install this program you must have Windows “Administrator” privileges.

The remaining steps are straightforward, and, as before, it is recommend that you use the default settings. Here is what you need to do for installing ImageMagick:

1) “This will install ImageMagick 5.5.7 Q8. Do you wish to continue?” **Yes**
2) “Welcome to the ImageMagick Setup Wizard” Click <Next>
3) “Information: Please read the following …” Click <Next>
4) “Select Destination Directory …” Leave at default and click <Next>
5) “Select Start Menu Folder …” Leave at default and click <Next>
6) “Select Additional Tasks …” Leave at default and click <Next>
7) “Ready to Install”. Click <Install>
8) Files are copied across
9) “You have now installed …” Click <Next>
10) “Setup has finished …”. Deselect “View index.html” and click <Finish>.

### 8.4 GREENSTONE INTERFACES

GSDL comprises two interfaces, the Librarians Interface and the Website which serves as the user interface.

The “librarian’s interface” in GSDL is for creation, management and updating collections. It is programmed in JAVA highly based on creation of the necessary commands.

The website is served by internal www-server or Apache. Webpages are created by Perl and Java Servlets which is customisable via CSS and text-files.
A) Librarian’s Interface

A JAVA-PERL applet (gliserver.pl) provides an interactive graphical interface for the Greenstone Librarian Interface with the following main functions:

1) **Gathering** - documents into a Selecting files from ‘local file space’ or Local Network or downloading using protocols viz. WWW, OAI (Open Archives Initiative), Z39.50, SRW (Search and Retrieve Web service), MediaWiki.

![Fig. 8.1: Librarian’s Interface- Collection Building](image)

2) **Enriching** - cataloguing with metadata, i.e. assign values to metadata-fields -Dublin Core and/or others or local sets. Metadata editor allows creating/changing sets and assigning values- automatic inheriting for lower levels, multiple values, picklists or hierarchical at level1|level2|level3

![Fig. 8.2: Librarian’s Interface- Metadata Input](image)
3) **Design** – this involves selection of plugins (e.g. GA, TEXT, PPT, Word, PDF, RTF, e-mail, XLS, Fox, DB, as well as ISIS, DSpace, MARC, ProCite…), defining Search index, Partitioning of sub-collections and setting Browsing classifiers, hierarchical or A-Z.

4) ‘plug-ins’ (filters), Indexing the documents and providing preview facility for direct access to webpage with search-interface produced by GLI is done at this stage. Once build is successful then the collection needs to be linked to previewing.
Digitisation and Digital Libraries – DSpace and GSDL

Self Check Exercise

Note: i) Write your answers in the space given below.
    ii) Check your answers with the answers given at the end of this Unit.

2) What functions are available in the Librarian’s Interface?

...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................
...................................................................................................................

B) Greenstone User Interface

Although the user interface of different Greenstone collections may appear remarkably similar, each one can provide varying search, browse and display features, depending on access requirements, nature of documents comprising the collection and metadata associated with these documents. As a digital library developer you can define the desired end-user interface features for your collection at the designing stage.

Collection Searching

Greenstone supports different ways of searching collections. They can be grouped in two main categories: “plain search” (through Google-like single search box) and “form-based search”.

- **Plain search:**
  
  **Simple** - Users can search for words or phrases in the full text of the document or limit the search to a specific index (e.g. document title or author) by selecting the available index from the drop-down box.

  **Advanced** - Boolean queries.

- **Form-based search**
  
  **Simple** - Users can search for words or phrases across different fields.

  **Advanced** - Users can search for words or phrases across different fields, with support for Boolean query combination, case folding and stemming.

Document Browsing

Greenstone supports browsing of documents in a collection by specific metadata fields.

Available browse elements for a collection are shown on the navigation bar in the collection home page. Hierarchical browsing of classification-like structures (e.g. a subject classification) with different levels is possible.
Presentation of Search Results

The web pages the users see when using Greenstone are not pre-stored but are generated “on the fly” as they are needed. This includes the way the browse and search results appear and individual documents are presented. After obtaining a document (selected from results of browse/search), a user can:

- view complete content or contract it (in a full-text tagged document);
- highlight matching search terms or not; and
- detach the document for viewing in a different window.
Digitisation and Digital Libraries – DSpace and GSDL

Greenstone supports multilingual interface. Through the preferences setting, the user can change the language of the Greenstone interface. It can also support indexing and searching of document collections in non-Latin scripts.

8.5 COLLECTION BUILDING IN GREENSTONE

You will need some source files like those in the sample_files\Word_and_PDF folder to work on the collection building.

1) Start a new collection called reports, fill out appropriate fields for it, and choose Dublin Core as the metadata set.

2) Copy the 12 files from sample_files ‘! Word_and_PDF ‘! Documents into the collection. You can select multiple files by clicking on the first one and shift-clicking on the last one, and drag them all across together. (This is the normal technique of multiple selection.)

3) Switch to the Create panel, and build and preview the collection.

4) Again, this collection contains no manually assigned metadata. All the information that appears—title and filename—is extracted automatically from the documents themselves. Because of this the quality of some of the title metadata is suspect.

5) Back in the Librarian Interface, click the Enrich tab to view the automatically extracted metadata. You will need to scroll down to see the extracted metadata, which begins with “ex.”. The PostScript documents (cluster.ps and langmodl.ps do not have extracted titles: what appears in the titles a-z list is just the first few characters of the document).

6) Manually adding metadata to documents in a collection

   In the Enrich panel, manually add Dublin Core dc.Title metadata to one of these documents. Select word03.doc and double-click to open it. Copy the title of this document (“Greenstone: A comprehensive open-source digital library software system”) and return to the Librarian Interface. Scroll up or down in the metadata table until you can see dc.Title. Click in the value box, paste in the metadata and press Enter.

7) Now add dc.Creator information for the same document. You can add more than one value for the same field: when you press Enter in a metadata value field, a new empty field of the same type will be generated.

8) Close the document when you have finished copying metadata from it. External programs opened when viewing documents must be closed before building the collection, otherwise errors can occur.

9) Next add title and creator metadata for a few of the other documents.

If you build and preview your collection at this point, you will find that nothing has changed. You need to alter the collection design to use the new Dublin Core metadata instead of the original extracted metadata.
10) **Collection design; branding a collection with an image**

Change to the **Design** panel, which is split into several sections. The first section **General** appears. This allows you to modify the values you provided when defining the collection, if desired. You can also brand the collection using a suitable image.

11) Click on the *Browse...* button associated with **URL to about page icon**, and browse to the image `sample_files` '! Word_and_PDF' '! wrpdf.gif' on your computer. When you select this image, Greenstone automatically generates an appropriate URL for the image. **Preview** the collection.

If you are on the web, you can easily make your own Greenstone-style icon by going to and following the instructions there.

http://www.greenstone.org/make-images.html

**Document plugins**

12) Now look at the **Document Plugins** section, by clicking on this in the list to the left. Here you can add, configure or remove plugins to be used in the collection. There is no need to remove any plugins, but it will speed up processing a little. In this case we have only Word, PDF, RTF, and PostScript documents, and can remove the ZIPPlug, TEXTPlug, HTMLPlug, EMAILPlug, ImagePlug, ISISPlug and NULPlug plugins. To delete a plugin, select it and click *<Remove Plugin>*. GAPlug is required for any type of source collection and should not be removed.

13) **Search types and fielded searching**

Go to the **Search Types** section. This specifies what kind of search interface and what search indexes will be provided for the collection. Let’s add a form search option. Click *<Enable Advanced Searches>*; this allows form searching to be added to the collection.

14) To include “form search” as well as the default “plain search”, pull down the **Search Types** menu and select **form**; then click *<Add Search Type>*.

Plain search will be the default search type as it is first in the list.

**Search indexes**

15) The next step in the **Design** panel is **Search Indexes**. These specify what parts of the collection are searchable (e.g. searching by title and author). Delete the `ex.Title` and `ex.Source` indexes, which are not particularly useful, by selecting them one at a time and clicking *<Remove Index>*. Only the `text` index remains.

16) Now add a Title index based on `dc.Title` by providing an **Index Name** (e.g. “Document Title”) and selecting `dc.Title` from the **Index Source** box. Then click *<Add Index>*.

17) You can add indexes based on any metadata. Add an index called “Authors” based on `dc.Creator` metadata.

*The next two sections are Partition Indexes and Cross-Collection Search. In this exercise, we will not make any changes to these.*
18) The Browsing Classifiers section adds “classifiers,” which provide the collection with browsing functions. Go to this section and observe that Greenstone has provided two classifiers, AZLists based on ex. Title and ex. Source metadata. Remove both of these by selecting them in turn and clicking <Remove Classifier>.

19) Now we add an AZList classifier for dc.Title metadata. Select AZList from the Select classifier to add drop-down list and click <Add Classifier>.

20) A popup window Configuring Arguments appears. Select dc.Title from the metadata drop-down list and click <OK>.

21) Now add an AZCompactList classifier. Click <Add Classifier> and configure it to use dc.Creator metadata, with button name “Creator”. Click <OK>.

The last three sections are Format Features, Translate Text and Metadata Sets. In this exercise, we will not make any changes to these.

22) Switch to the Create panel, and build and preview the collection.

23) Check that all the facilities work properly. There should be three full-text indexes, called text, Document Title, and Authors. In the titles a-z list should appear all the documents to which you have assigned dc.Title metadata (and only those documents). In the authors a-z list should appear one bookshelf for each author you have assigned as dc.Creator, and clicking on that bookshelf should take you to all the documents they authored.

In the similar fashion you can build up collection for other types of file formats. For details visit the tutorial site of Greenstone.

8.6 SUMMARY

Greenstone is a freely available open source software for building and distributing digital library collections through Internet or. Multiplatform availability, the capability of providing access in different ways and managing different file formats, media and languages are some of the major advantages of Greenstone. The Librarian Interface provides the most advanced and at the same time a very user friendly approach to collection building and also metadata management.

In this Unit we discussed the technical features of Greenstone, installation process and building a digital library.

8.7 ANSWERS TO SELF CHECK EXERCISES

1) Technical features of GSDL are:
   • Multiplatform user friendly application
   • Interoperability
   • Independent librarian and user interfaces
   • Supports variety of Metadata formats
   • Supports variety of Document formats
   • Supports multiple Languages
2) Following functions are available in the Librarian’s Interface:

- Creation of New Collection
- Selection Metadata
- Gathering
- Enrich
- Design
- Create

8.8 KEYWORDS

Lucene : Open source search engine.
Perl : A script programming language that is similar in syntax to the C language and that includes a number of popular UNIX facilities.
UNICODE : An international encoding standard for use with different languages and scripts, by which each letter, digit, or symbol is assigned a unique numeric value that applies across different platforms and programs.
XML : Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable.

8.9 REFERENCES AND FURTHER READING