Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 6. EXAMPLE OF DIGITAL LIBRARY SOFTWARE: GREENSTONE

LESSON 1. GREENSTONE TUTORIAL

NOTE

Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Learning Objectives

At the end of this lesson you will be able to:

- recognize the key features of **Greenstone digital library software**;
- identify the **installation** requirements and options;
- identify which **interface features** can be provided to end users, and
- identify the three different **collection building approaches** provided by Greenstone.

Introduction

A very important aspect of the creation, organization and provision of access to digital library collections is **the software used for this purpose**.

Though it is possible today to create simple digital library collections using prevailing Web and database technologies, it would take a lot more effort to develop software that integrates functions such as metadata, full text search and retrieval, support for multiple document types and formats, multilingual content and access management.

Affordable software is thus a key requirement for the successful development and deployment of digital libraries.

A colleague from another department has talked to me about Greenstone, an open source suite of software for building digital library collections. Can you please collect some information about it?

This team is in charge of developing the digital library for their organization.

Greenstone overview

Greenstone is a freely available 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, and developed and distributed in cooperation with **UNESCO** and the **Human Info NGO**.

Greenstone is open-source software, issued under the terms of the GNU General Public License.

The aim of the software is to empower users, particularly in universities, libraries and other public service institutions, to build their own digital libraries.

Greenstone overview

We should focus more closely on Greenstone's key features. Are they consistent with the project we have in mind?

Digital library developers can benefit from the variety of features supported by Greenstone, such as multiplatform availability, the capability to provide access in different ways and manage different file formats, media and languages.

Greenstone allows powerful indexing, search and browse methods.

Developers can also choose among various interfaces to build collections and customize the end-user interface.

Let's have a look at these features in a more detailed way...

Greenstone overview

The following are the features supported by Greenstone:

Multiplatform availability

Greenstone is available for 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).

Access and Distribution

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.

Collection building

Supports a variety of **interfaces** for collection building.

Greenstone overview

The following are the features supported by Greenstone:

Powerful Indexing

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.

Powerful Search and Browse

Since Greenstone does full text and field based indexing, you are provided with a **variety of search options**.

File formats

Greenstone supports different file formats, such as HTML, PDF, DOC, RTF, E-mail, Plain text, PPT etc. These file formats are converted into a standard XML-based internal format for indexing using 'plugins'.

Greenstone overview

The following are the features supported by Greenstone:

Extensibility and Configurability

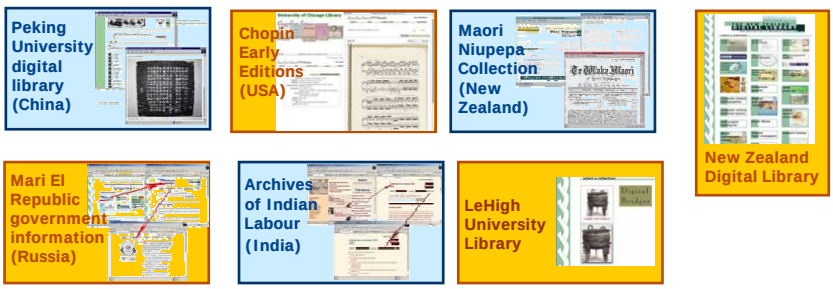
New plugins can be developed for file formats not supported by Greenstone. Greenstone allows you to configure a collection to customize the interface, indexing, browsing and presentation features according to your requirements.

Multimedia and Multilingual support

Greenstone lets you build collections of non-textual multimedia documents such as audio, video, and pictures accompanied by textual description or metadata to allow searching and browsing. Unicode, an encoding standard for representing a large number of language scripts, is used throughout Greenstone. This facilitates building, searching and browsing documents in any Unicode-compliant language.

Greenstone overview

In order to have a more precise idea of what type of collections can be built with Greenstone, the team examines some of the various **digital library collections that have been developed using Greenstone**. Such collections have been developed around the world, in several languages and in various domains, including historical, educational, cultural, and research.



Click on each collection for more information

Greenstone overview

Here are some features the team has found about **digital library collections** built using Greenstone:

They predominantly consist of **textual objects** (called 'documents'), in various forms: structured and unstructured text documents, page images, and text documents with accompanying images... They may comprise a very large number of documents.

Greenstone collections may also consist of **audio, photographic pictures, video and other digital objects**.

Each collection provides a **uniform interface** through which all documents in it can be accessed.

Metadata are automatically extracted from document objects and/or assigned explicitly during collection building.

A Greenstone digital library may include **many different collections**, each **organized differently**, although there is a strong family resemblance in how collections are presented.

Greenstone overview

A variety of collections can be built using Greenstone! This is clear from the features we have identified. I think we should summarize them in a short report...

Can you help the team to list the features supported by Greenstone?

List the features in the text box.

Then, click on View answer to see the complete list.

View answer

Greenstone installation

This software seems to be appropriate to our needs! How can we obtain it? And what do we need to install it?

The latest version of Greenstone software can be downloaded freely from http://www.greenstone.org, together with extensive documentation on how to use it.

The software and the documentation are also included on this CD-ROM: you can find them in the Resources section.

Greenstone's installation is guided, but you may want to have more detailed instructions.

Where can I find step-by-step instructions for installing Greenstone?

Detailed information and step-by-step instructions for installing Greenstone on Windows and Linux are available in the Greenstone Digital Library Installer's Guide (2.50). You can find these documents in the Greenstone Digital Library Documentation section (*Resources-software&tools-Greenstone*).

Greenstone installation

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:

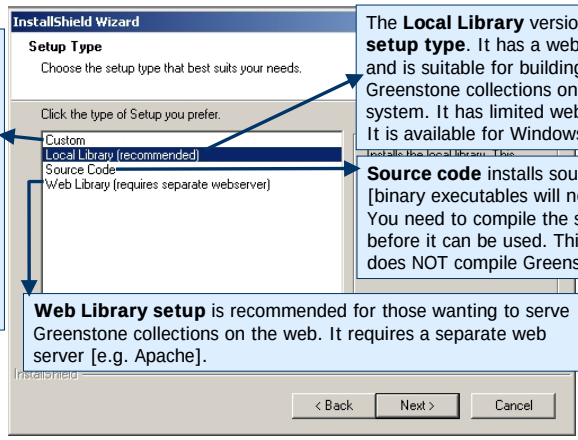
Java Run-time Environment (JRE) version 1.4 or above (Install JRE before installing GSDL) – JRE is required for GLI
[Not required for default Windows installation] Web Server (Apache Recommended)
PERL - gets installed automatically
C++ compiler, if you wish to compile the source code (Visual Studio or GCC)
A Web Browser (Internet Explorer 4 and above or Netscape 4.5 and above)

Now let's focus on a few critical points concerning the installation...

Greenstone installation

When you install Greenstone, you are asked to select one from the four available types of setup:

Custom setup allows you to install any or all of the features provided by the other three setup types. It also allows you to install additional features not included in the other three-setup types.



The **Local Library** version is the **default setup type**. It has a web server built-in and is suitable for building and viewing Greenstone collections on a stand-alone system. It has limited web serving abilities. It is available for Windows platform only.

Source code installs source code only [binary executables will not be installed]. You need to compile the source code before it can be used. This installation does NOT compile Greenstone for you.

Web Library setup is recommended for those wanting to serve Greenstone collections on the web. It requires a separate web server [e.g. Apache].

Greenstone installation

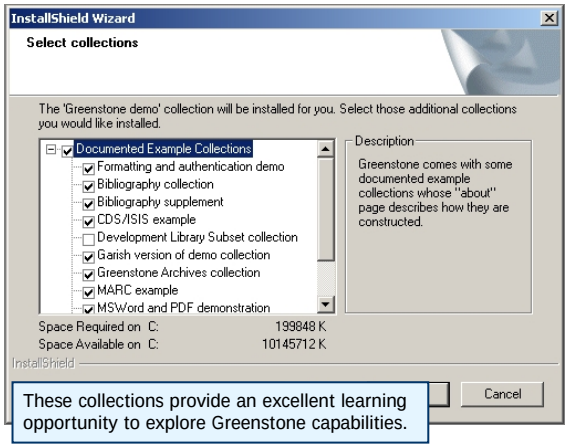
Some suggestions for choosing your setup

For **Windows and Linux**, installing from binaries (not source code) is recommended, unless if you want to work at source code level which requires compilation.
Note that collection building is not supported in **Windows 3.x**.
Since **Linux** is the primary operating system under which Greenstone has been developed, it is more robust and efficient on this platform. You may like to opt for Linux platform if you have the required support for installation and maintenance.

For Windows, if you do not feel comfortable configuring a suitable web server (e.g. Apache or Internet Information Server), we recommend that you use the **Local Library setup** (this setup is not available in other platforms). This setup is easy to install and comes with an inbuilt web server. We would recommend it over the Web library **unless**:
a) you expect a large number of simultaneous users, or
b) you are already using a web server to serve other stuff, or
c) you need to offer 24/7 service that will automatically start after rebooting etc.
To serve large number of simultaneous users and provide 24/7 services on the Internet, you should opt for Web Library setup (all platforms).

Greenstone installation

During installation, a demonstration Greenstone collection ('**Greenstone demo**') is automatically installed.



These collections provide an excellent learning opportunity to explore Greenstone capabilities.

The Greenstone CD-ROM also comes with several additional **pre-built example collections**, demonstrating the capability of Greenstone to build a variety of collections with different types of source documents and collection configuration features.

These example collections (except the Development Library Subset DLS collection which is quite large) **are installed automatically** under Local Library and Web Library setup options. You can also selectively install these pre-built collections by opting for 'custom' setup (during installation).

Greenstone installation

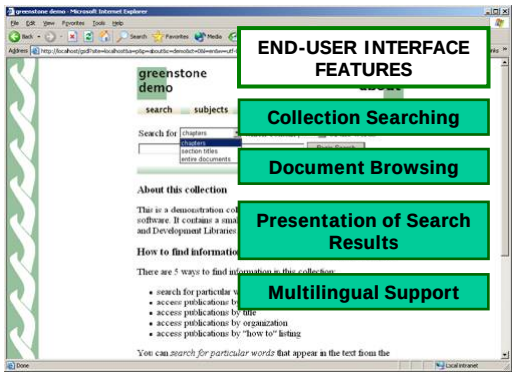
A key feature supported by Greenstone is **exporting digital library collections to CD-ROM**. The exported collection can be accessed directly from the CD-ROM without requiring any installation or can be installed on the hard disc.

This feature, however, is available for Windows platform only.

Utilization of this functionality **requires the CD-ROM export module**. This module is automatically installed when you install Greenstone ('Export CD-ROM' option in the File menu).

Greenstone User Interface

An end-user accesses a Greenstone digital library collection through its **user interface**. Before building your collection, it's very important to understand how Greenstone supports various features in the 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**.

Let's have a look at these features...

Greenstone User Interface

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".

Collection Searching

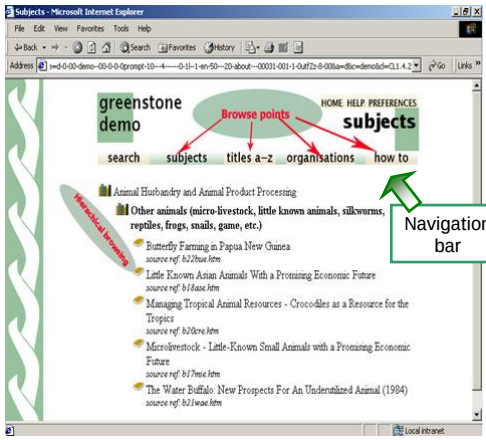
PLAIN SEARCH

- **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.

Greenstone User Interface



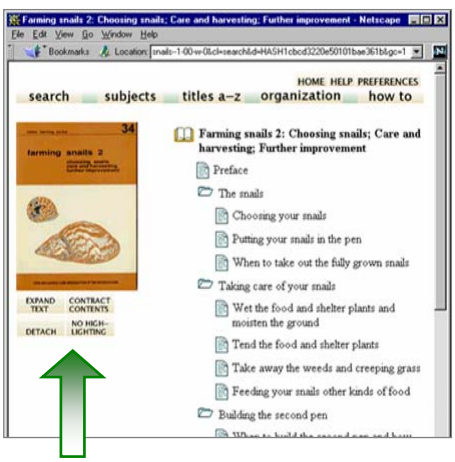
Document Browsing

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.

Greenstone User Interface



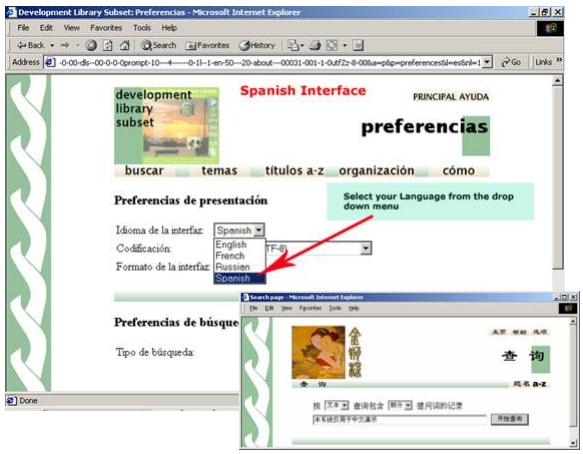
Presentation of Search Results

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.

Greenstone User Interface



Multilingual Support

Through the preferences setting, the user can **change the language of the Greenstone interface**.

Greenstone can also support indexing and searching of **document collections** in **non-Latin scripts**.

Greenstone User Interface

So, by selecting “Spanish” will I have documents in Spanish?

How would you answer this question?

- Yes
- No

Select the answer of your choice

Building Greenstone collections

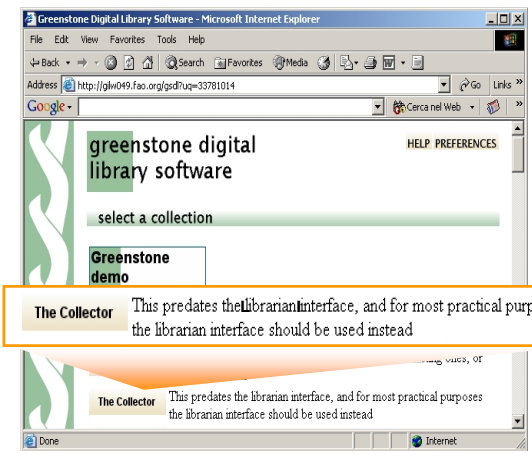
When you build a collection using Greenstone, you can choose among **three collection building approaches**:

- Collector,
- Command line mode, and
- Librarian Interface.

While the Collector and Librarian Interfaces are easy to use, the Command line mode requires understanding and correct usage of the commands and associated parameters.
The Greenstone Librarian Interface (GLI) provides the most advanced and at the same time a very user friendly approach to collection building and metadata management.

In the next few screens we will look at each of these approaches, by focusing on how to use them and on available features.

Building Greenstone collections



Approach 1: Collector

The Collector approach provides a web-based collection building wizard. You can invoke it by selecting the option **Collector** from the Greenstone home page.

Though collection building is easy, it does not support assignment and management of metadata.

Supported functions

How to use it

More details

Click on each button to read the description

Building Greenstone collections

Approach 1: Collector

Supported functions

The Collector approach supports the following functions:

- Create a new collection
- Add new material to an existing one
- Modify the structure of an existing one
- Delete a collection
- Write an existing collection to a CD-ROM

How to use it

Collection building using the Collector involves the following steps:

- Specify collection information - its name and associated info
- Specify source data - where the source data comes from
- Configure collection - Adjust the configuration options (advanced use)
- Build the collection
- Access the collection!

More details

More details on using the Collector for building collections is given in Greenstone digital library 2.50 – User's guide (section 3.4). You can find this document in the Greenstone Digital Library Documentation section (*Resources-software&tools-Greenstone*).

Building Greenstone collections

Approach 2: Command line mode

(MS-DOS mode for Win 95/98, 'cmd' mode for Win NT/2000/XP, shell level for Linux)

This approach requires the use of a set of Perl scripts and thus requires understanding and correct usage of the commands and associated parameters. Assignment and management of metadata is very difficult.

Supported functions

How to use it

More details

Click on each button to read the description

Building Greenstone collections

Approach 2: Command line mode

Supported functions

Command line mode supports simple and advanced collection building.

How to use it

Collection building using the Collector involves the following steps:
- Set up Greenstone environment variables
- Use **mkcol** script to create new collection
- Move source documents to the **import** folder of the new collection
- Adjust the configuration options by editing the **collect.cfg** file
- Import the collection using **import** script
- Build the collection using the **build** script

More details

More details of using the Command line mode for building collections is given in Greenstone digital library 2.50 – Developer's guide (section 1, pages 1-25). You can find this document in the **Greenstone Digital Library** Documentation section.

Building Greenstone collections

Hmmm...we aren't expert developers... can we find a simpler way?? I would like to build collections without external support...

Yes, we can use GLI! It is very user friendly and it is the most advanced approach as well!

Approach 3: GUI-based Librarian Interface (GLI)

With version 2.40 onwards, Greenstone now includes the **Librarian Interface (GLI)**, a Java-based interface for building digital library collections. Installation of GLI is integrated with installation of Greenstone.

This provides a very user-friendly approach to building collections. We strongly recommend that, for most practical purposes, you use GLI for collection building.

Supported functions

How to use it

More details

Click on each button to read the description

Building Greenstone collections

Approach 3: GUI-based Librarian Interface (GLI)

Supported functions

GLI is the most advanced approach to collection building and also metadata management. The GLI has excellent metadata management support and also supports definition of custom metadata fields, if required.

How to use it

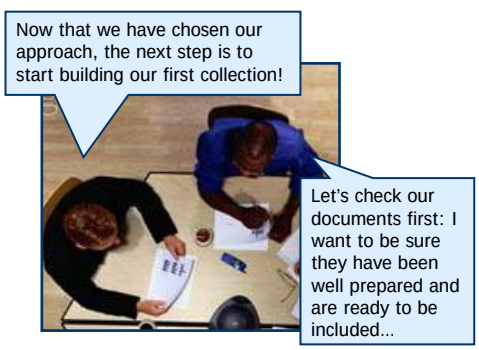
Building collections using GLI involves following steps:

- Gathering – gather the source documents that will comprise the collection and associate appropriate metadata element set
- Enriching – assign metadata for each source document
- Designing – specify collection configuration in terms of indexes, classifiers, display formats, etc.
- Creating – build the collection
- Previewing – preview the newly built collection

More details

A practical example of collection building using GLI is discussed in lesson 6.2 More details of the GLI approach to collection building is given in Greenstone digital library 2.50 – User's guide (section 3.2, pages 31-53). You can find this document in the **Greenstone Digital Library** Documentation section.

Building Greenstone collections



Before building a Greenstone digital library collection, **source documents should be prepared carefully** to obtain the desired access (search and browse) and display features.

The care and attention paid during the preparation of the source documents will reflect on the quality of the digital library collection offered to the users.

Particularly, **assigning explicit metadata** to each source document during collection building offers the most powerful approach to support various browse and field-based searching features. Metadata information is also useful in producing a suitable presentation for search results.

Preparation of source collection and metadata

You should consider the following while preparing input documents.

If you wish to support full text searching or if you expect Greenstone to automatically extract metadata (e.g. document title)...

This will not be possible if source documents are in image formats (e.g. image PDF and JPEG). Ensure that the input documents are in a format from which Greenstone can extract text (e.g. Text PDF, Word Doc or RTF, and HTML).

If you expect Greenstone to correctly extract the document 'Title' metadata automatically...

It will be useful to assign correct document properties, particularly for Microsoft Word documents. The first readable text line in the PDF is likely to be used as a document title – ensure that this is appropriate. Similarly, ensure that HTML documents have meaningful document titles. However, if you are explicitly assigning the document title as a metadata field and using this for indexing, these precautions are not necessary.

...

Preparation of source collection and metadata

You should consider the following while preparing input documents.

If you find that some Word and PDF documents are not getting converted and indexed properly by Greenstone...

We suggest that you find an alternative way of converting these to HTML and using them for collection building. Greenstone uses third party software for converting Word and PDF documents and since there are so many versions of these documents, conversion may sometimes fail.

If you want to support section level searching or hierarchical browsing of large documents...

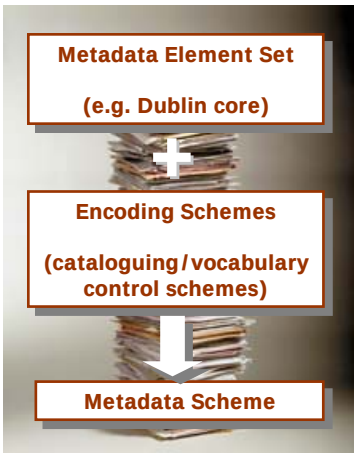
You should incorporate appropriate full text tags in the text document. You will find more details of full text tagging in the following documents:

Tagging Document Files, Section 3.3, in Greenstone Digital Library User's Guide 2.50;
Librarian Interface: adding and using metadata (gsdl-4-GLI2.pdf) in Greenstone training workshop material.

You can find these documents in the Greenstone Digital Library Documentation section (*Resources-software&tools-Greenstone*).

...

Preparation of source collection and metadata



A key decision to be made during the preparation of the input collection is **what metadata scheme** is to be used.

Adoption of Dublin Core should be considered seriously since Greenstone has internal support for this standard.

Decisions regarding encoding schemes (cataloguing rules, vocabulary control schemes) to be used and formats for rendering of content of each field are also important.

These decisions lie outside the digital library software, but form a **very important part of the collection building process**. The metadata element set and the cataloguing/vocabulary control schemes together comprise the "metadata scheme" (or "metadata specification") for the collection and should be adhered to during collection building.

Summary

Greenstone is a freely available suite of software for building and distributing digital library collections on the Internet or on CD-ROM.

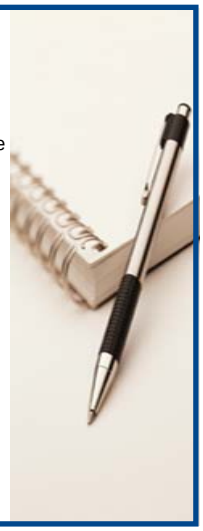
Digital library developers can benefit from the **variety of features supported** by Greenstone, such as multiplatform availability, the capability of providing access in different ways and managing different file formats, media and languages. Greenstone also allows powerful indexing, search and browse methods.

When you install Greenstone, you can select the **type of setup** that best suits your needs. Available types are: custom, source code, local library and web library setup.

You can choose among three **collection building approaches**: Collector, Command line mode, and Librarian Interface (GLI).

The **Librarian Interface** provides the most advanced and at the same time a very user friendly approach to collection building and also metadata management.

Before building collections using Greenstone, pay attention to the **preparation process**, including the adoption of a suitable **metadata scheme** for the collection being developed.



Exercises

The following five exercises will help you test your understanding of the concepts covered in the lesson and provide you with feedback.

Good luck!

Exercise 1

For which of the following operating platforms is Greenstone software available in binary (executable) form?

- ☐ Windows
- ☐ Unix
- ☐ Linux
- ☐ Macintosh

Please select the options of your choice (2 or more) and press “Check Answer”

Exercise 2

Greenstone uses a standard character encoding format for storing document content in order to support multiple languages.

What is this encoding standard?

Write the correct word in the box. Then click on the “Check Answer” button.

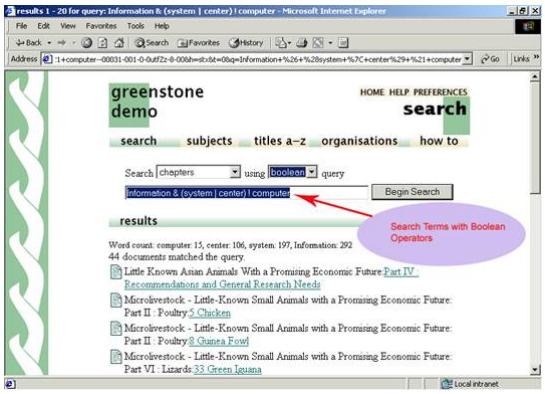
Exercise 3

When installing Greenstone in the Windows operating system, the default setup type is...

Write the correct word in the box

Exercise 4

Look at the Greenstone end-user interface below. What kind of collection searching does it provide?



- Plain search - simple
- Plain search - advanced
- Form-based search – simple
- Form-based search - advanced

Select the answer of your choice

Exercise 5

Which one of the following collection building approaches supports metadata management?

- Collector
- Librarian Interface (GLI)
- Command line mode

Please click on the answer of your choice

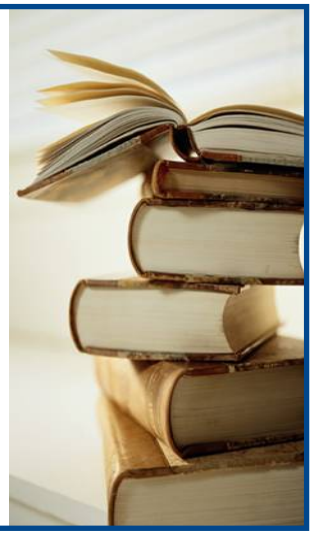
Resources on this CD

Greenstone software (v.2.51)

Greenstone is a suite of software for building and distributing digital library collections. This distribution includes everything you need to run Greenstone (including a pre-built demonstration collection) and to build new Greenstone collections. The Greenstone Digital Library Software (an open source product licensed by the University of Waikato) is being provided by UNESCO as a contribution to IMARK.

From Paper to Collection

This document explains how to create CD-ROM collections from paper documents. It describes in full detail the procedures and economics involved in the scanning and optical character recognition (OCR) processes, so that you end up with text in the right format to apply the Greenstone software (also provided on this CD). This document is being provided by UNESCO as a contribution to IMARK.



If you want to know more...

Online Resources:

Greenstone - Configuration files of demo collections in New Zealand Digital Library project: (http://www.greenstone.org/cgi-bin/library?a=p&p=colcfg)

Ian H. Witten. Examples of practical digital libraries: Collections built internationally using Greenstone. D-Lib Magazine, March 2003. (http://dlib.org/dlib/march03/witten/03witten.html)

Greenstone training workshop material. Greenstone Digital Library Project and NCSI, IISc. 2003: (http://www.greenstone.org/)

Customizing the Greenstone User Interface. An illustrated guide to customizing the Greenstone user interface. Written by Allison Zhang of the Washington Research Library Consortium (http://www.wrlc.org/dcpc/UserInterface/interface.htm)

Additional Reading:

Ian H. Witten and David Bainbridge 2003. How to build a digital library. Morgan Kaufman Publishers