Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 6. EXAMPLE OF DIGITAL LIBRARY SOFTWARE: GREENSTONE

LESSON 4. DOCUMENTED EXAMPLE GREENSTONE COLLECTIONS

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.

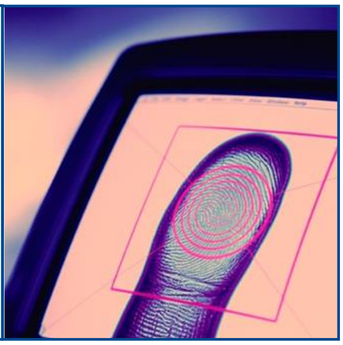
We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Learning Objective

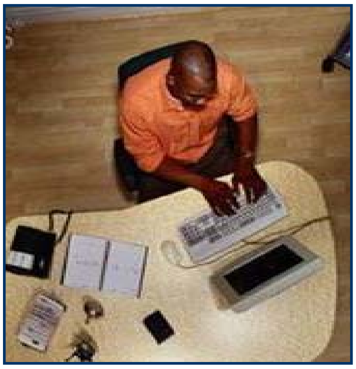
At the end of this lesson you will able to:

- identify the role of the **collection configuration file** in the Greenstone collection building process, and

- identify a source for learning more about the **customization of collections** built using Greenstone.



Introduction



In this lesson, we introduce you to some sample collections built using Greenstone.

A study of this documentation together with a study of these collections' access features will help you improve the customization of your own collections.

To better understand access features, let's have a look "behind the scenes": What happens during the collection building process?

Handling User Interface Features

Various features are available for the user interface. How does Greenstone manage them?

Collection Searching

To support various search features, Greenstone builds **indexes** for different components of source documents, including words from the full text as well as text associated with various metadata fields, at the time of collection building.

The required indexes for a collection are specified in a file named “configuration file”.

Document Browsing

Greenstone constructs browsing structures using **classifier** modules. These classifiers can be associated with specific metadata fields to build browsing by this field. Classifier information is to be specified in the collection configuration file.

Handling User Interface Features

Presentation of Search Results

These features are controlled using **format strings**. Format strings are specified in the collection configuration file, introduced by the keyword **format** followed by the name of the element to which the format applies. All format strings are interpreted at the time the pages are displayed.

Multilingual Support

The Greenstone interface is controlled by **macros** that are language specific and are stored in files with the **.dm** file extension. This means that, if you are interested in adding your language interface to Greenstone, you need to **create a new language specific macro** and tell Greenstone to use this macro when a user selects this language.

Greenstone uses UNICODE encoding standard internally to support multiple languages. UNICODE is an international standard for representing the charsets of all language. Through Unicode, Greenstone allows any language to be processed and displayed properly. Greenstone also supports other character sets that have their own precincts, converting these to UNICODE internally and then back to the specific charset for display.

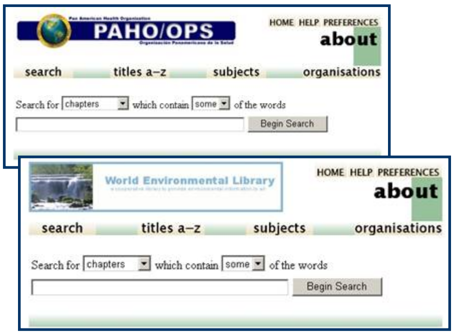
For searching using non-Latin scripts, the user needs to have **IME** (Input Mapping Editor) **installed** on their system and enable this to be accessed from the web browser.

Handling User Interface Features

You probably noticed that the search, browse and presentation features are mainly determined by the **collection configuration file** (**collect.cfg**).

Each Greenstone collection has its own collection configuration file, which acts as a collection model to serve specific types of source documents with specific indexing, browsing, and search features.

Greenstone uses the collect.cfg file during the collection building process, to construct search and browse structures and also during presentation of results.



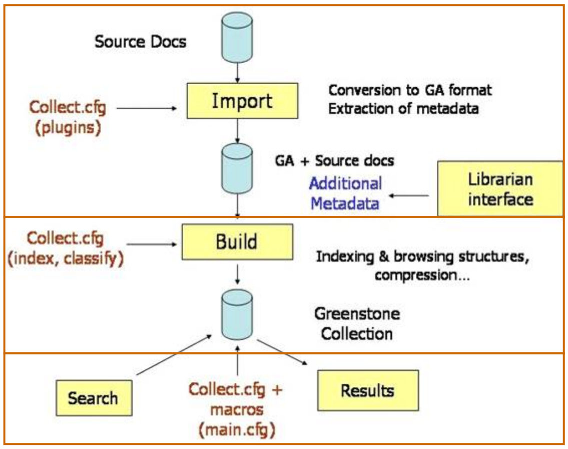
Search for chapters which contain some of the words

Search for chapters which contain some of the words

The collection model of a collection may be used for building different collections with similar configuration requirements.

Collection building process

Let's have a look at the following diagram which represents the collection building process:



Sets of source documents and the configuration file are imported into Greenstone for collection building. Greenstone imports these documents, extracts the full text and metadata and stores them in an internal XML-like format (called Greenstone Archive format - **GA format**). Greenstone uses relevant format-specific plugins mentioned in the collection configuration during this step.

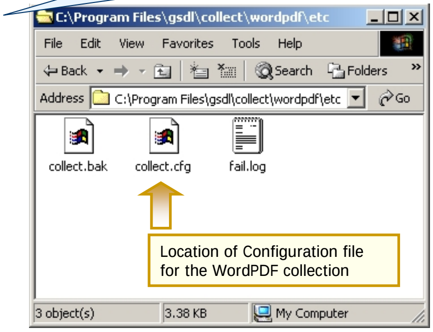
It then builds text search indexes and browsing structures based on the **indexes and classifiers** mentioned in the collection configuration file. The configuration file thus determines the searching and browsing structures built for the collection.

The presentation of search/browse results and the collection interface is determined by **format** strings in the configuration file and **macros** in **main.cfg**, the main configuration file of the Greenstone collection library.

Collection building process

This collection configuration file seems to play a very important role in the process!

But how is it physically formed??



Location of Configuration file for the WordPDF collection

The collection configuration is located in the **etc** directory of each collection directory (within the **gsdl/collect** folder).

It is a **text file** and can be **opened and read using any text editor** (but do not make changes without knowing the implications!!).

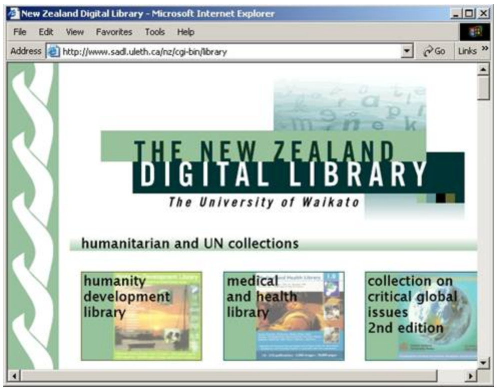
You may want to open the configuration file of the Greenstone demo collection and study its contents.

Collection building process

COMPONENTS OF CONFIGURATION FILE	
creator	E-mail address of the collection's creator
maintainer	E-mail address of the collection's maintainer
public	Whether collection is to be made public or not
beta	Whether collection is beta version or not
indexes	List of indexes to build (used in searching)
defaultindex	The default index
subcollection	Define a subcollection based on metadata
indexsubcollections	Specify which subcollections to index
defaultsubcollection	The default indexsubcollection
languages	List of languages to build indexes in
defaultlanguage	Default index language
collectionmeta	Defines collection-level metadata
plugin	Specify a plugin to use at build time (to extract text and convert it into GA format)
format	A format string - to control presentation of results
classify	Specify a classifier to use at build time – for building browse structures
searchtype	specify plain or form-based searching

Each line of the collection configuration file is essentially an “attribute, value” pair. Each attribute gives a piece of information about the collection that affects how it is supposed to look or how documents are to be processed.

Example Collections


New Zealand Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.sadl.uleth.ca/nz/cgi-bin/library
Go
Links
THE NEW ZEALAND
DIGITAL LIBRARY
The University of Waikato
humanitarian and UN collections
humanity development library
medical and health library
collection on critical global issues 2nd edition

The New Zealand Digital Library Website

As we have seen, the collection configuration file allows you to **customize the "look and feel"** of your collection and the way in which the documents are processed and presented.

To facilitate understanding of how this customization can be achieved, the New Zealand Digital Library team has prepared a few sample collections and documentation about how these collections are configured.

Let's have a look at them…

Example Collections

The documented example collections include the following:

Development Library Subset (DLS)

Development Library Subset: The DLS collection has the same structure as the Greenstone demo collection. It presents very interesting features: browsing by 'subject', 'organization' and 'how to' fields; searching on chapters and section titles; variety of formats for showing browse and search results and hierarchical browsing of documents. The collection configuration is quite complex. Chapter and section title level searching and hierarchical browsing of documents is made possible by full text tags embedded in the HTML source documents.

MSWord and PDF demonstration

Microsoft Word and PDF demonstration: This collection contains a few documents in PDF, Microsoft Word, RTF, and Postscript formats, demonstrating the ability to build collections from documents in different formats. The interface supports browsing by document title and full text searching. You can view the documents in HTML or native format. The collection configuration file is very simple.

Greenstone Archives collection (email)

Greenstone Archives: This is a collection of email messages from the Greenstone mailing list archives. It includes messages from the beginning of the mailing list in April 2000 up until fairly recently. This collection uses the Email plug-in, which parses files in email formats. The collection configuration file is very simple. The interface supports browse by 'subject', 'from' and 'dates' fields of e-mail messages and search on message text, header, from and subject fields.

Example Collections

simple image collection

Simple image collection

Simple image collection: This is a very basic image collection containing no text and no explicit metadata. The configuration file is probably the simplest!

bibliography collection

Bibliography collection (with fielded searching)

Bibliography Collection: This has about 4,000 bibliography (BibTeX) entries made from the Computational Learning Theory (COLT) Bibliography. This collection incorporates a form-based search interface that allows fielded searching. It also supports browsing by ‘dates’ and ‘phrases’. Collection configuration is fairly complex.

OAI example

OAI demonstration collection

OAI Example: This collection demonstrates Greenstone's ability to retrieve (‘harvest’) metadata from an OAI-compliant (Open Archives Initiative interoperability protocol). This collection consists of metadata retrieved from a collection of photographs taken at the inaugural Joint Conference on Digital Libraries.

Example Collections

MARC example

MARC record collection

MARC Example: This is a simple collection containing some MARC records from the Library of Congress catalog (http://catalog.loc.gov/). MARC records that include Beowulf in their title have been selected for inclusion in this collection. You can browse the records by author, title and subject and view records in MARC format.

bibliography supplement

Bibliography supplement

Bibliography Supplement: This tiny collection of 10 bibliography entries illustrates the "super collection" facility which searches several collections together, seamlessly. It operates together with the Bibliography collection, and its configuration file is almost the same. Look under Preferences Settings for selecting the "super collection" feature.

CDS/ISIS example

CDS/ISIS collection

CDS/ISIS Example: This collection is built from a CDS/ISIS database of about 150 bibliography entries. It uses the ISIS Plug plug-in, which reads the standard ISIS .mst and .fdt files and converts them to Greenstone metadata. The plug-in also handles the subfield data extraction. The collection incorporates a form-based search interface that allows fielded searching.

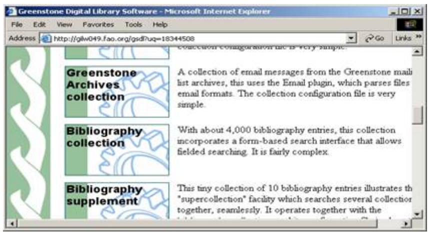
authentication and formatting demo

Authentication, and customizing menus

Authentication and formatting demo: This has the same material as the original Greenstone demo collection, but shows two independent features: non-standard document formatting, and controlled access to the documents via user authentication.

Example Collections

You can access these collections in different ways...

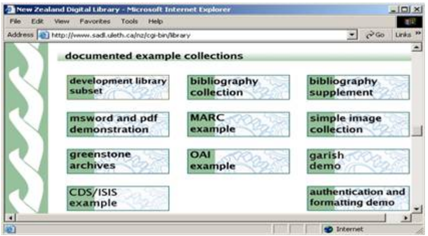

Greenstone Digital Library Software - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://gilw049.fao.org/gsdl?uq=18344508
Go
Links
Greenstone Archives collection
Bibliography collection
Bibliography supplement

A collection of email messages from the Greenstone maili list archives, this uses the Email plugin, which parses files email formats. The collection configuration file is very simple.

With about 4,000 bibliography entries, this collection incorporates a form-based search interface that allows fielded searching. It is fairly complex.

This tiny collection of 10 bibliography entries illustrates th "supercollection" facility which searches several collectior together, seamlessly. It operates together with the

When you install Greenstone in either **Local Library** or **Web Library** mode, by default Greenstone will install all the example collections.


New Zealand Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.sadl.uleth.ca/nz/cgi-bin/library
Go
Links
documented example collections
development library subset
bibliography collection
bibliography supplement
msword and pdf demonstration
MARC example
simple image collection
greenstone archives
OAI example
garish demo
CDS/ISIS example
authentication and formatting demo
Internet

If you want to access the example collections online from the **New Zealand Digital Library website**, you can find them under **Documented example collections**.

● **Example collections on this CD-Rom**

○ http://www.nzdl.org

By Installing Greenstone from this IMARK CD, you will find all the example collections, except the “MS Word and PDF demonstration" collection, **in the Greenstone home page**. You can access a specific collection by clicking on the collection icon.
Please note that the **source documents have not been included** in this IMARK CD: if you want to look at these documents, please go to the **New Zealand Digital Library website**.

Summary

- Greenstone uses a collection-specific configuration file (**collect.cfg**), a text file which determines the search, browse and display features of the collection.
- The collection configuration file allows you to **customize the “look and feel”** of your collection and the way in which the documents are processed and presented.
- The New Zealand Digital Library team has prepared a few sample collections and documentation about how these collections are configured, to facilitate understanding of how this customization can be achieved.
- The sample collection can be accessed from:
  - New Zealand Digital Library Website
  - Greenstone home page (after the installation)
  - This IMARK CD-ROM (Resources section)

Exercises

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

Good luck!

Exercise 1

What happens “behind the scenes”?

To provide the following features...	Greenstone uses...
COLLECTION SEARCHING	
DOCUMENT BROWSING	
PRESENTATION OF SEARCH RESULTS	
MULTILINGUAL SUPPORT	

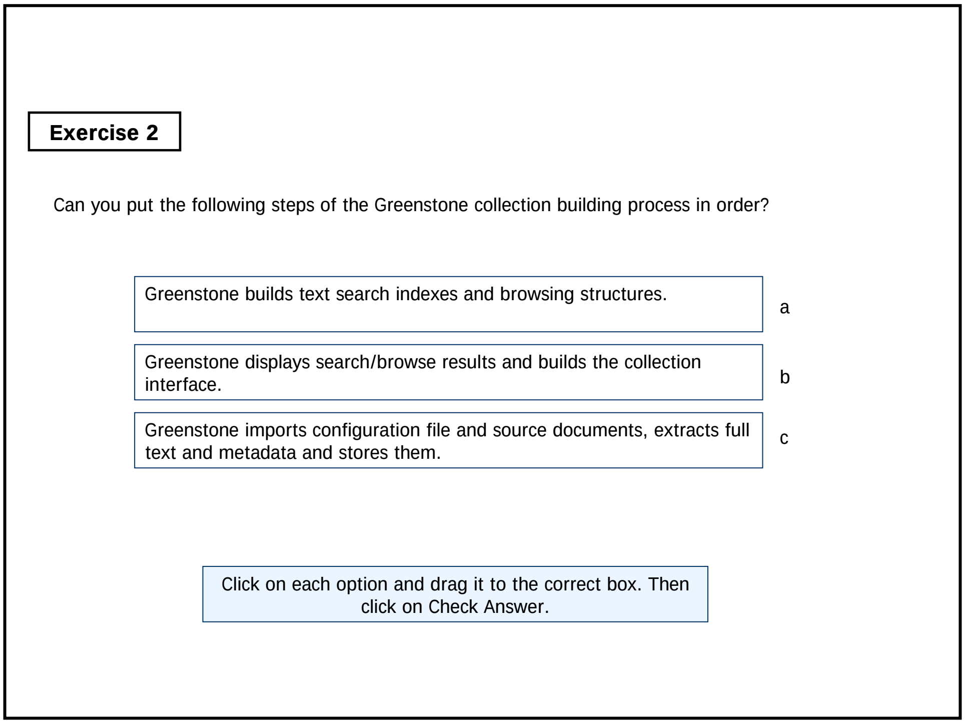
macros stored in files .dm

format strings

indexes

classifiers

Click on each option and drag it to the correct box. Then click on Check Answer.



Exercise 2

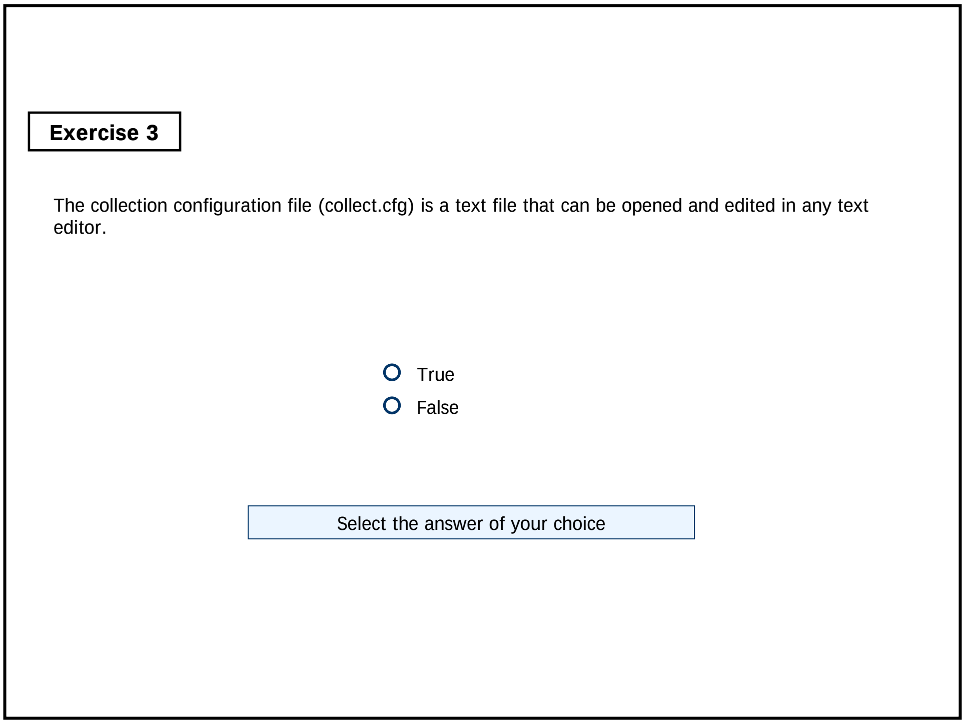
Can you put the following steps of the Greenstone collection building process in order?

Greenstone builds text search indexes and browsing structures. a

Greenstone displays search/browse results and builds the collection interface. b

Greenstone imports configuration file and source documents, extracts full text and metadata and stores them. c

Click on each option and drag it to the correct box. Then click on Check Answer.



Exercise 3

The collection configuration file (collect.cfg) is a text file that can be opened and edited in any text editor.

○ True

○ False

Select the answer of your choice

If you want to know more…

Online resources:

Ian H. Witten et al. Inside Greenstone collections. Department of Computer Science, University of Waikato, New Zealand. June 2003. http://www.cs.waikato.ac.nz/~ihw/greenstone/inside.pdf

New Zealand Digital Library website: http://www.nzdl.org

Home page customization:

Customizing your collection. In: Greenstone training workshop material. Greenstone Digital Library Project and NCSI, IISc. 2003 (document gsdl-6-Home-page-etc.pdf).
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