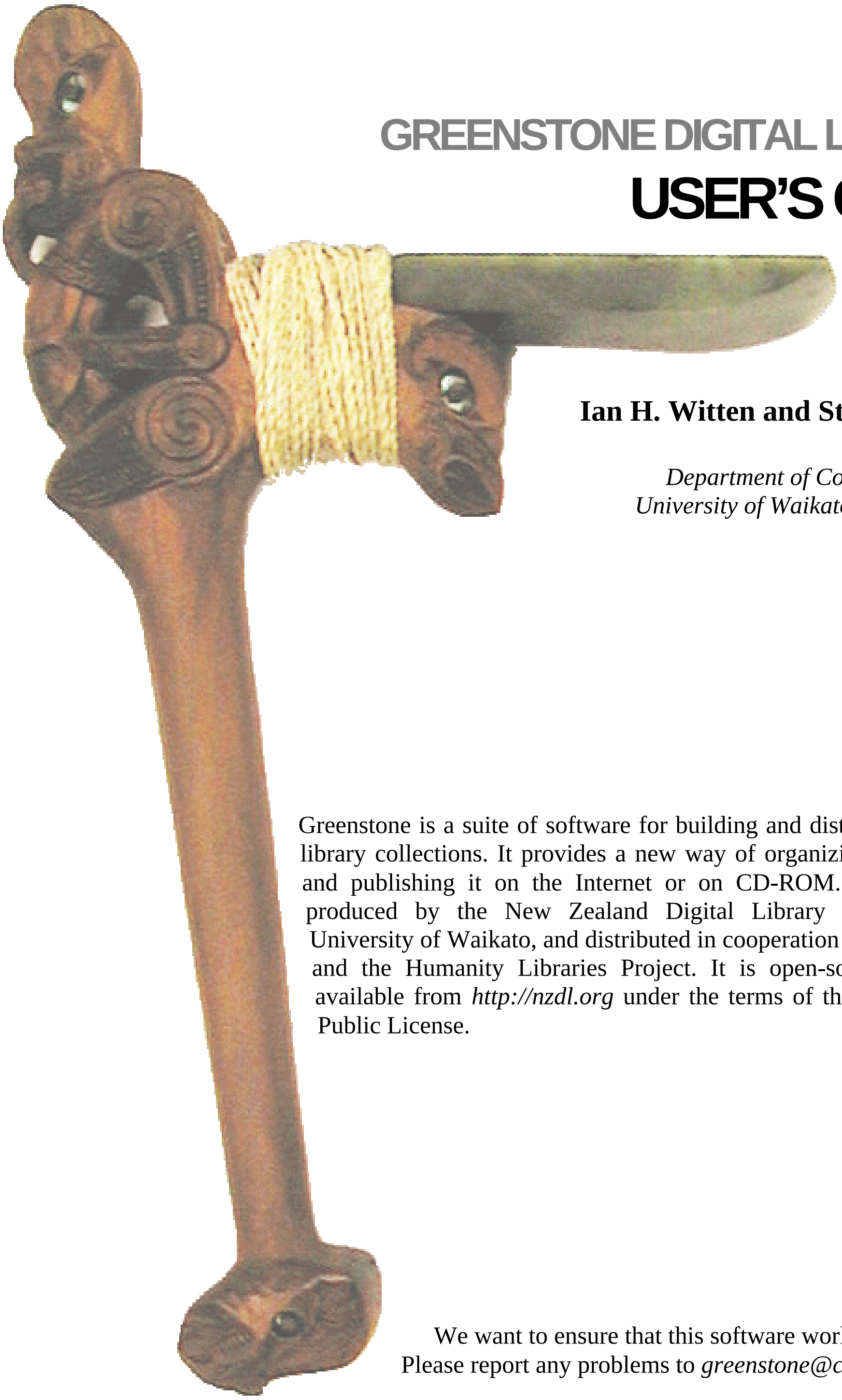


GREENSTONE DIGITAL LIBRARY

USER'S GUIDE

Ian H. Witten and Stefan Boddie

Department of Computer Science
University of Waikato, New Zealand

Greenstone is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato, and distributed in cooperation with UNESCO and the Humanity Libraries Project. It is open-source software, available from *http://nzdl.org* under the terms of the GNU General Public License.

We want to ensure that this software works well for you.
Please report any problems to *greenstone@cs.waikato.ac.nz*

Greenstone gsdl-2.31 **February 2001**

About this manual

This manual provides a comprehensive description of how to use the Greenstone software for accessing and building digital library collections.

Section 1 gives an overview of the capabilities of the software. Section 2 explains how to use Greenstone collections. The interface is self-explanatory—the best way to learn is by doing—and this section comprises the on-line help information for a typical collection. Section 3 explains how to build your own library collections using the Greenstone Collector, a set of Web pages that take you step by step through the process of building a collection. Section 4 introduces the administration facility that allows the system administrator to monitor what is going on and control who can build collections.

Appendices list the features of the Greenstone software, and give a glossary of terms used throughout the Greenstone documentation.

Companion documents

The complete set of Greenstone documents includes three volumes:

- Greenstone Digital Library Installer's Guide
- Greenstone Digital Library User's Guide *(this document)*
- Greenstone Digital Library Developer's Guide

Acknowledgements

The Greenstone software is a collaborative effort between many people. Rodger McNab and Stefan Boddie are the principal architects and implementors. Contributions have been made by David Bainbridge, George Buchanan, Hong Chen, Elke Duncker, Carl Gutwin, Geoff Holmes, John McPherson, Craig Nevill-Manning, Gordon Paynter, Bernhard Pfahringer, Todd Reed, Bill Rogers, and Stuart Yeates. Other members of the New Zealand Digital Library project provided advice and inspiration in the design of the system: Mark Apperley, Sally Jo Cunningham, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui and Lloyd Smith. We would also like to acknowledge all those who have contributed to the GNU-licensed packages included in this distribution: MG, GDBM, WGET, WVWARE, PDFTOHTML.

Contents

1 Overview of Greenstone

Greenstone is a comprehensive system for constructing and presenting collections of thousands or millions of documents, including text, images, audio and video.

1.1 Collections

A typical digital library built with Greenstone will contain many collections, individually organized—though they bear a strong family resemblance. Easily maintained, collections can be augmented and rebuilt automatically.

There are several ways to find information in most Greenstone collections. For example, you can *search for particular words* that appear in the text, or within a section of a document. You can *browse documents by title*: just click on a book to read it. You can *browse documents by subject*. Subjects are represented by bookshelves: just click on a bookshelf to look at the books. Where appropriate, documents come complete with a table of contents: you can click on a chapter or subsection to open it, expand the full table of contents, or expand the full document into your browser window (useful for printing). The New Zealand Digital Library website (*nzdl.org*) provides numerous example collections.

On the front page of each collection is a statement of its purpose and coverage, and an explanation of how the collection is organized. Most collections can be accessed by both *searching* and *browsing*. When searching, the Greenstone software looks through the entire text of all documents in the collection (this is called “full-text search”). In most collections the user can choose between indexes built from different parts of the documents. Some collections have an index of full documents, an index of paragraphs, and an index of titles, each of which can be searched for particular words or phrases. Using these you can find all documents that contain a particular set of words (the words may be scattered far and wide throughout the document), or all paragraphs that contain the set of words (which must all appear in the same paragraph), or all documents whose titles contain the words (the words must all appear in the document's title). There might be other indexes, perhaps an index of sections, and an index of section headings. Browsing involves lists that the user can examine: lists of authors, lists of titles, lists of dates, hierarchical classification structures, and so on. Different collections offer different browsing facilities.

1.2 Finding information

Greenstone constructs full-text indexes from the document text—that is, indexes that enable searching on any words in the full text of the document. Indexes can be searched for particular words, combinations of words, or phrases, and results are ordered according to how relevant they are to the query.

In most collections, descriptive data such as author, title, date, keywords, and so on, is associated with each document. This information is called *metadata*. Many document collections also contain full-text indexes of certain kinds of metadata. For example, many collections have a searchable index of document titles.

Users can browse interactively around lists, and hierarchical structures, that are generated from the metadata that is associated with each document in the collection. Metadata forms the raw material for browsing. It must be provided explicitly or be derivable automatically from the documents themselves. Different collections offer different searching and browsing facilities. Indexes for both are constructed during a "building" process, according to information in a collection configuration file.

Greenstone creates all searching and browsing structures automatically from the documents themselves: nothing is done manually. If new documents in the same format become available, they can be merged into the collection automatically. Indeed, for many collections this is done by processes that awake regularly, scout for new material, and rebuild the indexes—all without manual intervention.

1.3 Providing flexibility

Source documents come in a variety of formats, and are converted into a standard form for indexing by "plugins." Plugins distributed with Greenstone process plain text, HTML, WORD and PDF documents, and Usenet and E-mail messages. New ones can be written for different document types (to do this you need to study the *Greenstone Digital Library Developer's Guide*). To build browsing structures from metadata, an analogous scheme of "classifiers" is used. These create browsing indexes of various kinds: scrollable lists, alphabetic selectors, dates, and arbitrary hierarchies. Again, Greenstone programmers can create new browsing structures.

1.4 Multimedia and multilingual documents

Collections can contain text, pictures, audio and video. 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.

Unicode, which is a standard scheme for representing the character sets used in the world's languages, is used throughout Greenstone. This allows any language to be processed and displayed in a consistent manner. Collections have been built containing Arabic, Chinese, English, French, Mäori and Spanish. Multilingual collections embody automatic language recognition, and the interface is available in all the above languages (and more).

1.5 Usage

Collections are accessed over the Internet or published, in precisely the same form, on a self-installing Windows CD-ROM. Compression is used to compact the text and indexes. A Corba protocol supports distributed collections and graphical query interfaces.

The New Zealand Digital Library (*nzdl.org*) provides many example collections, including historical documents, humanitarian and development information, technical reports and bibliographies, literary works, and magazines.

Being open source, Greenstone is readily extensible, and benefits from the inclusion of Gnu-licensed modules for full-text retrieval, database management, text extraction from proprietary document formats, and Z39.50 protocol support. Only through international cooperative efforts will digital library software become sufficiently comprehensive to meet the world's needs with the richness and flexibility that users deserve.

4 OVERVIEW OF GREENSTONE

2
Using Greenstone Collections

The Greenstone software is designed to be easy to use. Web-based and CD-ROM collections have interfaces that are identical. Installing the Greenstone software from CD-ROM on any Windows or Linux computer is very easy indeed; a standard installation setup program is used. A collection can be used locally on the computer where it is installed; also, if this computer is connected to a network, the software automatically and transparently allows all other computers on the network to access the same collection.

The next section describes how to install a Greenstone CD-ROM. Then we look at the searching and browsing facilities offered by a typical Greenstone collection, the "Demo" collection that is supplied with the Greenstone software. Other collections offer similar facilities; if you can use one, you can use them all. The following section explains how to customize the interface for your own requirements using the Preferences page.

2.1 Using a Greenstone CD-ROM

The Greenstone digital library software itself comes on a CD-ROM, and you or your system manager have probably installed it on your system, following the instructions in the *Greenstone Digital Library Installer's Guide*. If so, Greenstone is already installed on your computer and you should skip the rest of this section.

Some Greenstone collections come on a self-contained Greenstone CD-ROM that includes enough of the software to run just this one collection. To use it, just put it into the CD-ROM drive on any Windows PC. Most likely (if "autorun" is enabled on your PC), a window will appear inviting you to install the Greenstone software. If not, find the CD-ROM disk drive (on current Windows systems you can get this by clicking on the *My Computer* icon on the desktop) and double-click it, or the *Setup.exe* file inside it. In either case the Greenstone *Setup* program will be entered, which guides you through the setup procedure. Most people respond *yes* to all the questions except for the one which offers to install the Netscape browser; if you already have a browser you probably don’t need to install a new one.

When the installation procedure has finished, you’ll find the library in the *Programs* submenu of the Windows *Start* menu, under the name of the collection (for example, “Humanity Libraries” or “United Nations University”).

Once the software has been installed, the library will be entered automatically every time you re-insert the CD-ROM if autorun is enabled.

2.2 Finding information

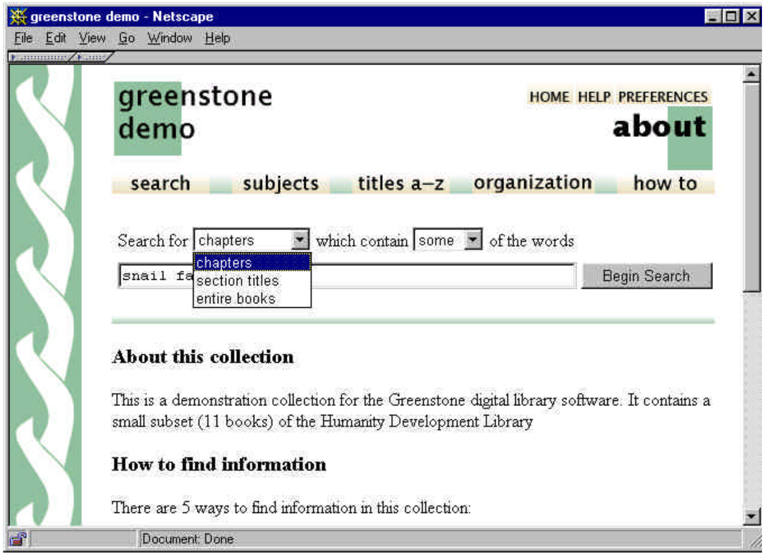
The easiest way to learn to use a Greenstone collection is to try it out. Don’t worry—you can’t break anything. Click liberally: most images that appear on the screen are clickable. If you hold the mouse stationary over an image, most browsers will soon pop up a message that tells you what will happen if you click.

Experiment! Choose common words like “the” and “and” to search for—that should evoke some responses, and nothing will break.

Greenstone digital library systems usually comprise several separate collections—for example, computer science technical reports, literary works, internet FAQs, magazines. There will be a home page for the digital library system which allows you to access any collection; in addition, each collection has its own “about” page that gives you information about how the collection is organized and the principles governing what is included in it. To get back to the “about” page at any time, just click on the “collection” icon that appears at the top left side of all searching and browsing pages.

Figure 1 shows a screenshot of the “Demo” collection supplied with the Greenstone software, which is a very small subset of the Humanity Development Library collection; we will use it as an example to describe the different ways of finding information. (If you can’t find the Demo collection, use the Humanity Development Library instead; it looks just the same.) First, almost all icons are clickable. Several icons appear at the top of almost every page; Table 1 shows you what they mean.

Figure 1
Using the Demo collection



About this collection

This is a demonstration collection for the Greenstone digital library software. It contains a small subset (11 books) of the Humanity Development Library

How to find information

There are 5 ways to find information in this collection:

The "*search … subjects … titles a-z … organization … how to*" bar underneath gives access to the searching and browsing facilities. The leftmost button is for searching, and the ones to the right of it—four, in this collection—evoke different browsing facilities. These may differ from one collection to another.

How to find information

Table 2 shows the five ways to find information in the Demo collection.

You can *search for particular words* that appear in the text from the "search" page. (This is just like the "about" page shown in Figure 1, except that it doesn't contain the *about this collection* text.) The search page can be reached from other pages by pressing the *search* button. You can *access publications by subject* by pressing the *subjects* button. This

Table 1 What the icons at the top of each page mean

greenstone demo	This takes you to the "about" page
HOME	This takes you to the Digital Library's home page, from which you ca select another collection
HELP	This provides help text similar to what you are reading now
PREFERENCES	This allows you to set some user interface and searching options that wi then be used henceforth

Table 2 What the icons on the search/browse bar mean

search	Search for particular words
subjects	Access publications by subject
titles a–z	Access publications by title
organization	Access publications by organization
how to	Access publications by “how to” listing

brings up a list of subjects, represented by bookshelves. You can *access publications by title* by pressing the *titles a-z* button. This brings up a list of books in alphabetic order. You can *access publications by organization* by pressing the *organization* button. This brings up a list of organizations. You can *access publications by how to listing* by pressing the *how to* button. This brings up a list of “how to” hints. You can see these buttons in Figure 1.

How to read the documents

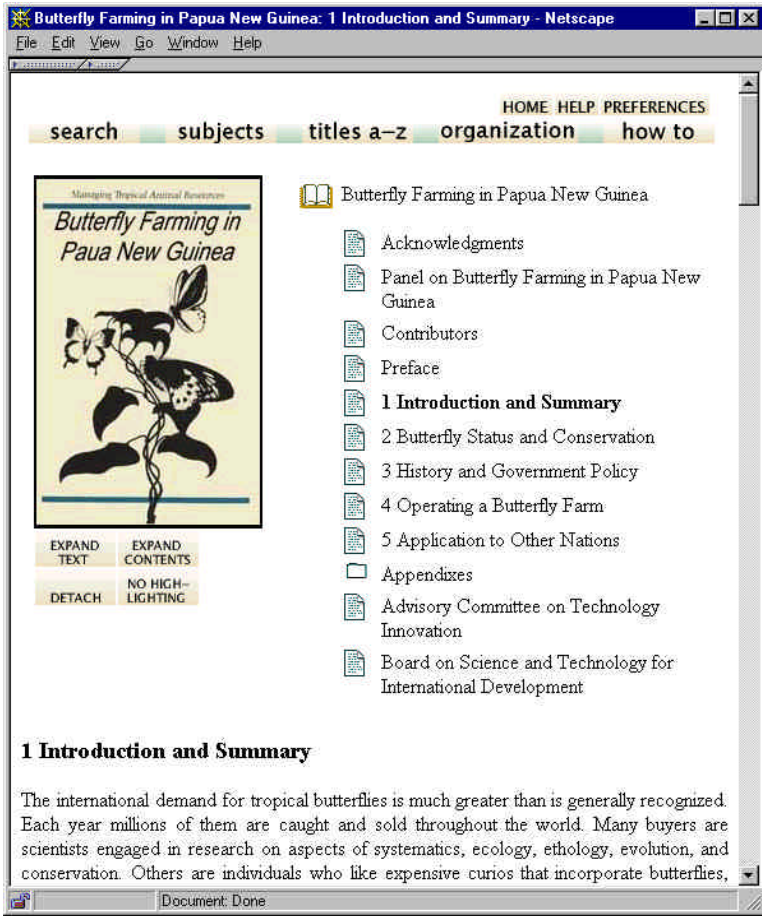
In the Demo collection, you can tell when you have arrived at an individual book because there is a photograph of its front cover (Figure 2). Beside the photograph is a table of contents; the entry in bold face (Section 1 in Figure 2) marks where you are. This table is expandable: click on the folders to open them or close them. Click on the open book at the top to close it.

Underneath is the text of the current section (*Introduction and Summary* in the example, beginning at the very bottom of the illustration). When you have read through it, there are arrows at the end to take you on to the next section or back to the previous one.

Below the photograph are four buttons. Click on *detach* to make a new browser window for this book. (This is useful if you want to compare books, or read two at once.) If you have reached this book through a search, the search terms will be highlighted: the *no highlighting* button turns this off. Click on *expand text* to expand out the whole text of the current section, or book. Click on *expand contents* to expand out the whole table of contents so that you can see the titles of all chapters and subsections.

In some collections, the documents do not have this kind of hierarchical structure. In this case, no table of contents is displayed when you get to an individual document—just the document text. In some cases, the document is split into pages, and you can read sequentially or jump about

Figure 2
A book in the Demo collection


Butterfly Farming in Papua New Guinea: 1 Introduction and Summary - Netscape
File Edit View Go Window Help
HOME HELP PREFERENCES
search subjects titles a–z organization how to
Managing Tropical Animal Resources
Butterfly Farming in Paua New Guinea
EXPAND TEXT
EXPAND CONTENTS
DETACH
NO HIGH-LIGHTING
Document: Done

Butterfly Farming in Papua New Guinea

- Acknowledgments
- Panel on Butterfly Farming in Papua New Guinea
- Contributors
- Preface
- **1 Introduction and Summary**
- 2 Butterfly Status and Conservation
- 3 History and Government Policy
- 4 Operating a Butterfly Farm
- 5 Application to Other Nations
- Appendixes
- Advisory Committee on Technology Innovation
- Board on Science and Technology for International Development

1 Introduction and Summary

The international demand for tropical butterflies is much greater than is generally recognized. Each year millions of them are caught and sold throughout the world. Many buyers are scientists engaged in research on aspects of systematics, ecology, ethology, evolution, and conservation. Others are individuals who like expensive curios that incorporate butterflies,

from one page to another.

What the icons mean

When you are browsing around the collection, you will encounter the items shown in Table 3.

How to search for particular words

From the search page, follow these simple steps to make a query:

- Specify what units you want to search: in the Demo collection you can search section titles or the full text of the books.

Table 3 Icons that you will encounter when browsing

📙	Click on a book icon to read the corresponding book
📚	Click on a bookshelf icon to look at books on that subject
📄	View this document
📁	Open this folder and view contents
📖	Click on this icon to close the book
📂	Click on this icon to close the folder
▶	Click on the arrow to go on to the next section ...
◀	... or back to the previous section
DETACH	Open this page in a new window
EXPAND CONTENTS	Expand table of contents
EXPAND TEXT	Display all text
HIGH-LIGHTING	Highlight search terms

- Say whether you want to search for all or just some of the words
- Type in the words you want to search for into the query box
- Click the *Begin Search* button

When you make a query, the titles of up to twenty matching documents will be shown. There is a button at the end to take you on to the next twenty. From there you will find buttons to take you on to the third twenty or back to the first twenty, and so on. However, for efficiency reasons a maximum of 100 is imposed on the number of documents returned. You can change these numbers by clicking the *preferences* button at the top of the page.

Click the title of any document, or the little icon beside it, to open it. The icon may show a book, or a folder, or a page: it will be a book icon if you are searching books; otherwise if you are searching sections it will be a folder or page icon depending on whether or not the section found has subsections.

SEARCH TERMS

Whatever you type into the query box is interpreted as a list of words called “search terms.” Each search term contains nothing but alphabetic characters and digits. Terms are separated by white space. If any other characters such as punctuation appear, they serve to separate terms just as though they were spaces. And then they are ignored. You can't search for words that include punctuation.

For example, the query

```
Agro-forestry in the Pacific Islands: Systems for
Sustainability (1993)
```

will be treated the same as

```
Agro forestry in the Pacific Islands Systems for
Sustainability 1993
```

QUERY TYPE

There are two different kinds of query.

- Queries for all the words. These look for documents (or chapters, or titles) that contain all the words you have specified. Documents that satisfy the query are displayed.
- Queries for some of the words. Just list some terms that are likely to appear in the documents you are looking for. Documents are displayed in order of how closely they match the query. When determining the degree of match,
  - the more search terms a document contains, the closer it matches;
  - rare terms are more important than common ones;
  - short documents match better than long ones.

Use as many search terms as you like—a whole sentence, or even a whole paragraph. If you specify only one term, it doesn't much matter whether you use an *all* or a *some* query, except that in the second case the results will be sorted by the search term's frequency of occurrence.

Scope of queries

In most collections you can choose different indexes to search. For example, there might be author or title indexes. Or there might be chapter or paragraph indexes. Generally, the full matching document is returned regardless of which index you search.

If documents are books, they will be opened at the appropriate place.

Advanced search features

While the above is enough to meet most searching needs, some more advanced search features are provided. These are activated from the Preferences page, which is reached by clicking the *preferences* button at the top of the page—see section 2.2 below. After changing your preferences, do not click your browser's *Back* button—that would undo the changes. Instead, click any of the buttons on the search/browse bar.

CASE SENSITIVITY AND STEMMING

When you specify search terms, you can choose whether upper and lower case must match between the query and the document: this is called "case sensitivity." You can also choose whether to ignore word endings or not: this is called "stemming."

Under *Search options* on the Preferences page you will see a pair of buttons labeled *ignore case differences* and *upper/lower case must match*; these control the case sensitivity of your queries. Below is a pair of buttons labeled *ignore word endings* and *whole word must match*: these control stemming.

For example, if the buttons *ignore case differences* and *ignore word endings* are selected, the query

```
African building
```

will be treated the same as

```
africa builds
```

because the uppercase letter in "African" will be transformed to lowercase, and the suffixes "n" and "ing" will be removed from "African" and "building" respectively (also, "s" would be removed from "builds").

Generally case differences and word endings should be ignored unless you are querying for particular names or acronyms.

PHRASE SEARCHING

If your query includes a phrase in quotation marks (" and "), only documents containing that phrase, exactly as typed, will be returned.

If you want to use phrase searching, you need to learn a little about how it works. Phrases are processed by a post-retrieval scan. First the query is issued in the normal way—all the words in the phrase are included as search terms—and then the documents returned are scanned to eliminate those in which that phrase does not appear.

During the post-retrieval scan, phrases are checked just as they are, including any punctuation. For example, the query

```
what’s a “post-retrieval scan?”
```

will first retrieve all documents that match all of the words

```
what s a post retrieval scan
```

and then the documents returned will be checked for the phrase

```
post-retrieval scan?
```

ADVANCED QUERY MODE

In *advanced query mode*, which can be selected on the Preferences page, the queries for *all* of the words, described above, are actually Boolean queries. They consist of a list of terms joined by logical operators & (and), | (or), and ! (not). Absent operators are interpreted as & (and): thus a query without any operators returns documents that match *all* the terms.

If the words AND, OR, and NOT appear in your query they are treated as ordinary search terms, not operators. For operators you must use &, |, and !. In addition, parentheses can be used for grouping.

USING SEARCH HISTORY

When you switch on the “search history” feature on the Preferences page you will be shown your last few searches, along with a summary of how many results they generated. Click the button beside one of the previous searches to copy the text into the search box. This makes it easy to repeat slightly modified versions of previous queries.

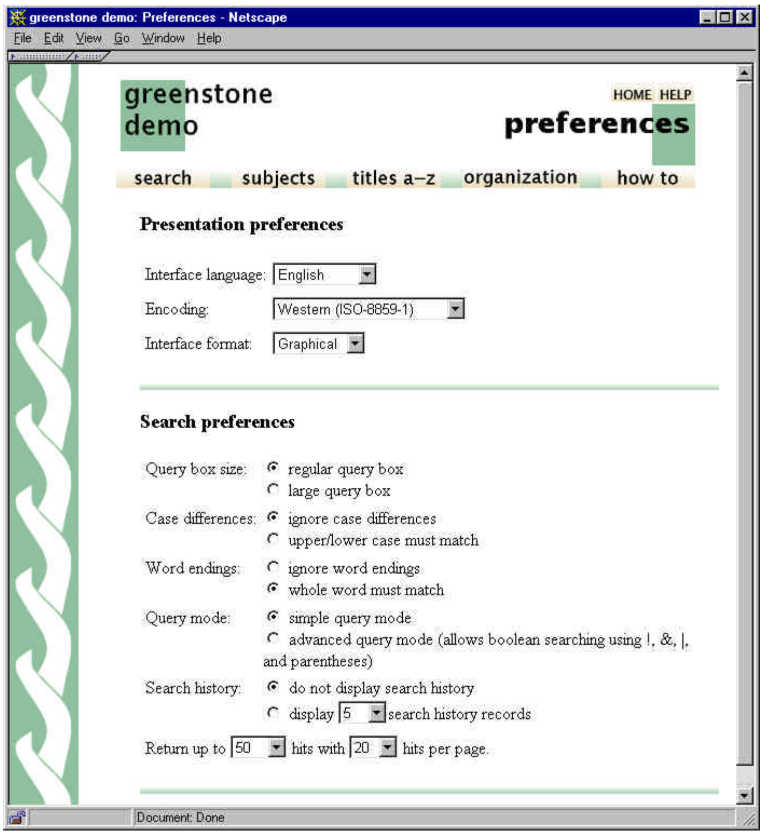
2.2 Changing the preferences

When you click the *preferences* button at the top of the page you will be able to change some features of the interface to suit your own requirements. The preferences depend on the collection; an example is shown in Figure 3.

Collection preferences

Some collections comprise several subcollections, which can be searched independently or together, as one unit. If so, you can select which subcollections to include in your searches on the Preferences page.

Figure 3
The Preferences page



Presentation preferences

Interface language: English
Encoding: Western (ISO-8859-1)
Interface format: Graphical

Search preferences

Query box size: ● regular query box ○ large query box
Case differences: ● ignore case differences ○ upper/lower case must match
Word endings: ○ ignore word endings ● whole word must match
Query mode: ● simple query mode ○ advanced query mode (allows boolean searching using !, &, |, and parentheses)
Search history: ● do not display search history ○ display 5 search history records
Return up to 50 hits with 20 hits per page.

Language preferences

Each collection has a default presentation language, but you can switch to a different language if you like. You can also alter the encoding scheme used by Greenstone for output to the browser—the software chooses sensible defaults, but with some browsers better visual results can be used by switching to a different encoding scheme. All collections allow you to switch from the standard graphical interface format to a textual one. This is particularly useful for visually impaired users who use large screen fonts or speech synthesizers for output.

Presentation preferences

Depending on the collection, there may be other options you can set that control the presentation. Collections of Web pages allow you to suppress the Greenstone navigation bar at the top of each document page, so that once you have done a search you land at the exact Web page that matches without any Greenstone header. To do another search you will have to use your browser's "back" button. These collections also allow you to suppress Greenstone's warning message when you click a link that takes you out of the digital library collection and on to the Web itself. And in some Web collections you can control whether the links on the "Search Results" page take you straight to the actual URL in question, rather than to the digital library's copy of the page.

Search preferences

Two pairs of buttons control the kind of text matching in the searches that you make. The first set (labeled "case differences") controls whether upper and lower case must match. The second ("word endings") controls whether to ignore word endings or not. It is possible to get a large query box, so that you can easily do paragraph-sized searching. It is surprisingly quick to search for large amounts of text.

You can switch to an "advanced" query mode which allows you to combine terms using AND (&), OR (|), and NOT (!). This allows you to specify more precise queries. You can turn the search history feature, described above, on and off. Finally, you can control the number of hits returned, and the number presented on each screenful.

3 The Collector

The Collector is a facility that helps you create new collections, modify or add to existing ones, or delete collections. To do this you will be guided through a sequence of Web pages which request the information that is needed. The sequence is self-explanatory: this section takes you through it.

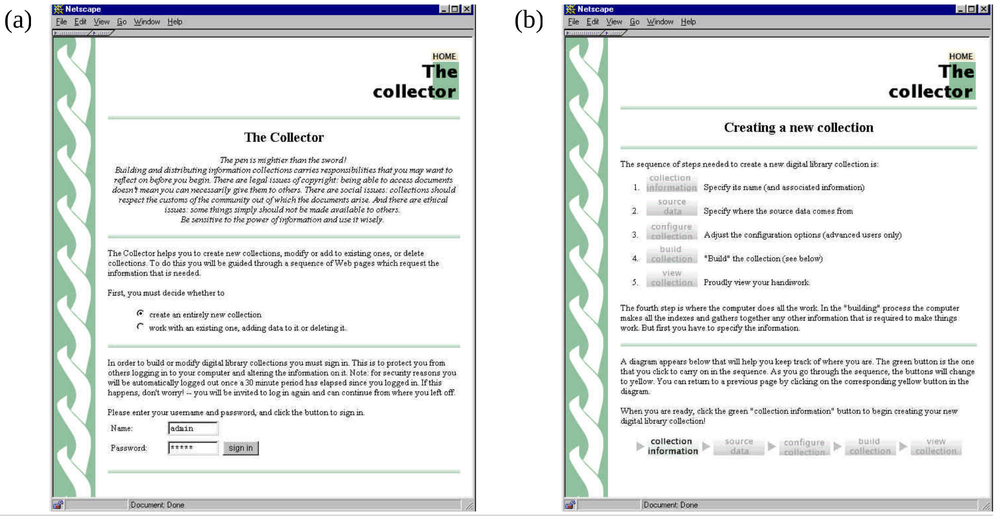
But first, building and distributing information collections carries responsibilities that you should reflect on before you begin. There are legal issues of copyright: being able to access documents doesn't mean you can necessarily give them to others. There are social issues: collections should respect the customs of the community out of which the documents arise. And there are ethical issues: some things simply should not be made available to others. The pen is mightier than the sword!—be sensitive to the power of information and use it wisely.

On the default Greenstone setup, you can access the Collector by clicking the appropriate link on the front page. Alternatively, the person who set up your Greenstone system may have created a special URL for the Collector. If not, you can access it by entering the special URL

http://localhost/gsdl /cgi-bin/library?a=collector&p=intro

into your browser. If the Greenstone software is running on another computer, you should substitute that computer's domain name (e.g. *www.yourcomputer.com*) for *localhost*.

In Greenstone, the structure of a particular collection is determined when the collection is set up. This includes such things as the format of the source documents, how they should be displayed on the screen, the source of metadata, what browsing facilities should be provided, what full-text search indexes should be provided, and how the search results should be displayed. Once the collection is in place, it is easy to add new documents to it—so long as they have the same format as the existing documents,



The Collector

The pen is mightier than the sword!
Building and distributing information collections carries responsibilities that you may want to reflect on before you begin. There are legal issues of copyright: being able to access documents doesn't mean you can necessarily give them to others. There are social issues: collections should respect the customs of the community out of which the documents arise. And there are ethical issues: some things simply should not be made available to others.
Be sensitive to the power of information and use it wisely.

The Collector helps you to create new collections, modify or add to existing ones, or delete collections. To do this you will be guided through a sequence of Web pages which request the information that is needed.

First, you must decide whether to

● create an entirely new collection
○ work with an existing one, adding data to it or deleting it.

In order to build or modify digital library collections you must sign in. This is to protect you from others logging in to your computer and altering the information on it. Note: for security reasons you will be automatically logged out once a 30 minute period has elapsed since you logged in. If this happens, don't worry! -- you will be invited to log in again and can continue from where you left off.

Please enter your username and password, and click the button to sign in.

Name: admin
Password: *****

Creating a new collection

The sequence of steps needed to create a new digital library collection is:

1. collection information Specify its name (and associated information)
2. source data Specify where the source data comes from
3. configure collection Adjust the configuration options (advanced users only)
4. build collection "Build" the collection (see below)
5. view collection Proudly view your handiwork.

The fourth step is where the computer does all the work. In the "building" process the computer makes all the indexes and gathers together any other information that is required to make things work. But first you have to specify the information.

A diagram appears below that will help you keep track of where you are. The green button is the one that you click to carry on in the sequence. As you go through the sequence, the buttons will change to yellow. You can return to a previous page by clicking on the corresponding yellow button in the diagram.

When you are ready, click the green "collection information" button to begin creating your new digital library collection!

Figure 4 Using the Collector to build a new collection (continued on next pages)

and the same metadata is provided, in exactly the same way.

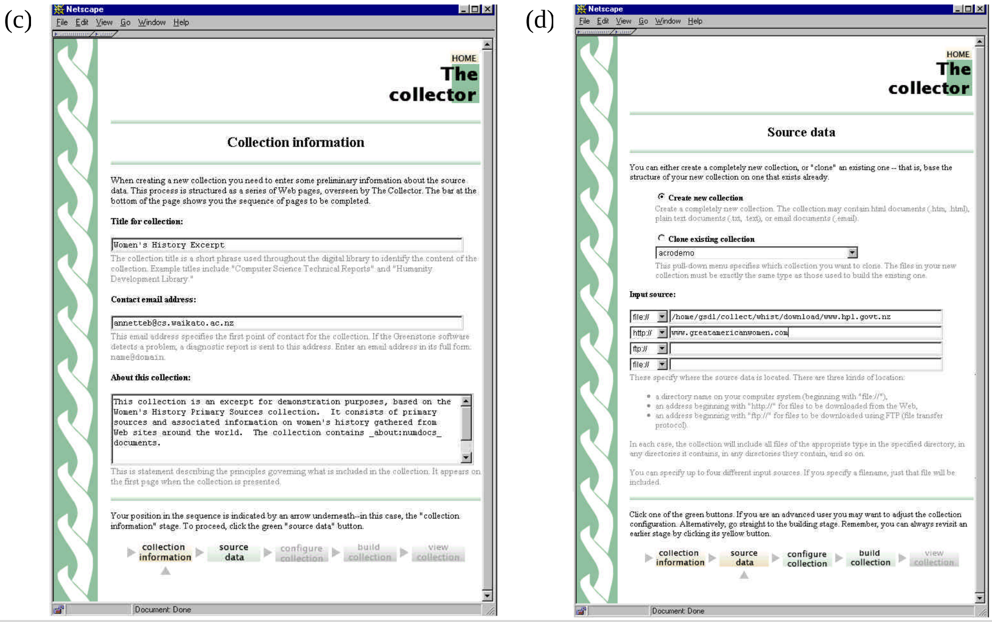
The Collector has the following basic functions:

1. create a new collection with the same structure as an existing one;
2. create a new collection with a different structure from existing ones;
3. add new material to an existing collection;
4. modify the structure of an existing collection;
5. delete a collection; and
6. write an existing collection to a self-contained, self-installing CD-ROM.

Figure 4 shows the Collector being used to create a new collection, in this case from a set of HTML files stored locally. You must first decide whether to work with an existing collection or build a new one. The former case covers options 1 and 2 above; the latter covers options 3–6. In Figure 4a, the user opts to create a new collection.

3.1 Logging in

Either way it is necessary to log in before proceeding. Note that in general, people use their Web browser to access the collection-building facility on a remote computer, and build the collection on that server. Of course, we cannot allow arbitrary people to build collections (for reasons of propriety if nothing else), so Greenstone contains a security system which forces people who want to build collections to log in first. This



(c)

Collection information

When creating a new collection you need to enter some preliminary information about the source data. This process is structured as a series of Web pages, overseen by The Collector. The bar at the bottom of the page shows you the sequence of pages to be completed.

Title for collection:

`Women's History Excerpt`

The collection title is a short phrase used throughout the digital library to identify the content of the collection. Example titles include "Computer Science Technical Reports" and "Humanity Development Library."

Contact email address:

`annetteb@cs.waikato.ac.nz`

This email address specifies the first point of contact for the collection. If the Greenstone software detects a problem, a diagnostic report is sent to this address. Enter an email address in its full form: name@domain.

About this collection:

```
This collection is an excerpt for demonstration purposes, based on the
Women's History Primary Sources collection.  It consists of primary
sources and associated information on women's history gathered from
Web sites around the world.  The collection contains _about:numdocs_
documents.
```

This is statement describing the principles governing what is included in the collection. It appears on the first page when the collection is presented.

Your position in the sequence is indicated by an arrow underneath--in this case, the "collection information" stage. To proceed, click the green "source data" button.

(d)

Source data

You can either create a completely new collection, or "clone" an existing one -- that is, base the structure of your new collection on one that exists already.

● **Create new collection**
Create a completely new collection. The collection may contain html documents (.htm, .html), plain text documents (.txt, .text), or email documents (.email).

○ **Clone existing collection**
acrodemo
This pull-down menu specifies which collection you want to clone. The files in your new collection must be exactly the same type as those used to build the existing one.

Input source:

file://	/home/gsdl/collect/whist/download/www.hpl.govt.nz
http://	www.greatamericanwomen.com
ftp://	
file://	

These specify where the source data is located. There are three kinds of location:

- a directory name on your computer system (beginning with "file://"),
- an address beginning with "http://" for files to be downloaded from the Web,
- an address beginning with "ftp://" for files to be downloaded using FTP (file transfer protocol).

In each case, the collection will include all files of the appropriate type in the specified directory, in any directories it contains, in any directories they contain, and so on.

You can specify up to four different input sources. If you specify a filename, just that file will be included.

Click one of the green buttons. If you are an advanced user you may want to adjust the collection configuration. Alternatively, go straight to the building stage. Remember, you can always revisit an earlier stage by clicking its yellow button.

Figure 4 (Continued)

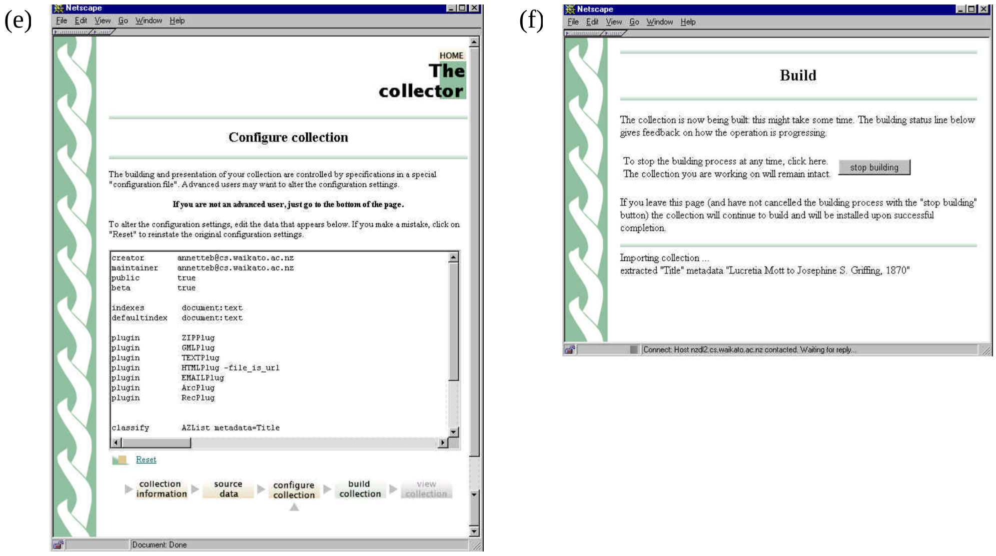
allows a central system to offer a service to those wishing to build information collections and use that server to make them available to others. Alternatively, if you are running Greenstone on your own computer you can build collections locally, but it is still necessary to log in because other people who use the Greenstone system on your computer should not be allowed to build collections without prior permission.

3.2 Dialog structure

Upon completion of login, the page in Figure 4b appears. This shows the sequence of steps that are involved in collection building. They are:

1. Collection information
2. Source data
3. Configuring the collection
4. Building the collection
5. Viewing the collection.

The first step is to specify the collection’s name and associated information. The second is to say where the source data is to come from. The third is to adjust the configuration options, which requires considerable understanding of what is going on—it is really for advanced users only. The fourth step is where all the (computer’s) work is done.



Configure collection

The building and presentation of your collection are controlled by specifications in a special "configuration file". Advanced users may want to alter the configuration settings.

If you are not an advanced user, just go to the bottom of the page.

To alter the configuration settings, edit the data that appears below. If you make a mistake, click on "Reset" to reinstate the original configuration settings.

```
creator       annetteb@cs.waikato.ac.nz
maintainer    annetteb@cs.waikato.ac.nz
public        true
beta          true

indexes        document:text
defaultindex   document:text

plugin         ZIPPlug
plugin         GMLPlug
plugin         TEXTPlug
plugin         HTMLPlug -file_is_url
plugin         EMAILPlug
plugin         ArcPlug
plugin         RecPlug


classify       AZList metadata=Title
```

Build

The collection is now being built: this might take some time. The building status line below gives feedback on how the operation is progressing.

To stop the building process at any time, click here.
The collection you are working on will remain intact.

If you leave this page (and have not cancelled the building process with the "stop building" button) the collection will continue to build and will be installed upon successful completion.

Importing collection ...
extracted "Title" metadata "Lucretia Mott to Josephine S. Griffing, 1870"

Connect: Host nzdl2.cs.waikato.ac.nz contacted. Waiting for reply...

Figure 4 (Continued)

During the "building" process the system makes all the indexes and gathers together any other information that is required to make the collection operate. The fifth step is to check out the collection that has been created.

These five steps are displayed as a linear sequence of gray buttons at the bottom of the screen in Figure 4b, and at the bottom of all other pages generated by the Collector. This display helps users keep track of where they are in the process. The button that should be clicked to continue the sequence is shown in green (*collection information* in Figure 4b). The gray buttons (all the others, in Figure 4b) are inactive. The buttons change to yellow as you proceed through the sequence, and the user can return to an earlier step by clicking the corresponding yellow button in the diagram. This display is modeled after the "wizards" that are widely used in commercial software to guide users through the steps involved in installing new software.

3.3 Collection information

The next step in the sequence, collection information, is shown in Figure 4c. When creating a new collection, it is necessary to enter some information about it:

- title,
- contact E-mail address, and

- brief description.

The collection title is a short phrase used through the digital library to identify the content of the collection. Example titles include *Food and Nutrition Library*, *World Environmental Library*, *Humanity Development Library*, and so on. The E-mail address specifies the first point of contact for any problems encountered with the collection. If the Greenstone software detects a problem, a diagnostic report is sent to this address. Finally, the brief description is a statement describing the principles that govern what is included in the collection. It appears under the heading *About this collection* on the first page when the collection is presented.

The user's current position in the collection-building sequence is indicated by an arrow that appears in the display at the bottom of each screen—in this case, as Figure 4c shows, the collection information stage. The user proceeds to Figure 4d by clicking the green source data button.

3.4 Source data

Figure 4d is the point where the user specifies the source text that comprises the collection. You can either create a completely new collection, or "clone" an existing one—that is, base the structure of your new collection on one that exists already. Creating a totally novel collection with a completely different structure from existing ones is a major undertaking, and is not what the interactive Collector interface is designed for. The most effective way to create a new collection is to base its structure on an existing one, that is, to clone it.

If you clone an existing collection, you need to specify (on a pull-down menu) which collection you want to clone. Note that some collections use non-standard input file formats, while others use metadata specified in auxiliary files. If your new input lacks this information, some browsing facilities may not work properly.

For this reason we do not recommend cloning the Demo collection, or any similary collections (e.g. Humanity Development Library, Food and Nutrition Library) unless you are sure that your new files are in the same format and contain the same metadata specification file.

The alternative to cloning an existing collection is to create a completely new one. A bland collection configuration file is provided that accepts a wide range of different document types and generates a searchable index of the full text and an alphabetic title browser. More information about the different document formats that can be accommodated is given in Section 3.9 below.

Boxes are provided to indicate where the source documents are located: up to four separate input sources can be specified. There are three kinds of specification:

- a directory name on the Greenstone server system (beginning with “file://”)
- an address beginning with “http://” for files to be downloaded from the Web
- an address beginning with “ftp://” for files to be downloaded using FTP.

In each case of “file://” or “ftp://” the collection will include all files in the specified directory, any directories it contains, any files and directories *they* contain, and so on. If instead of a directory a filename is specified, that file alone will be included. For “http://” the collection will mirror the specified Web site.

In this case (Figure 4c) the new collection will contain documents taken from a local file system as well as a remote Web site, which will be mirrored during the building process.

3.5 Configuring the collection

Figure 4e shows the next stage. The construction and presentation of all collections is controlled by specifications in a special collection configuration file (see below). Advanced users may use this page to alter the configuration settings. Most, however, will proceed directly to the final stage.

In this case the user has made a small modification to the default configuration file by including the file_is_url flag with the HTML plugin. This flag causes URL metadata to be inserted in each document, based on the filename convention that is adopted by the mirroring package. This metadata is used in the collection to allow readers to refer to the original source material, rather than to a local copy.

3.6 Building the collection

Figure 4f shows the “building” stage. Up until now, the responses to the dialog have merely been recorded in a temporary file. The building stage is where the action takes place.

During building, indexes for both browsing and searching are constructed according to instructions in the collection configuration file. The building process takes some time: minutes to hours, depending on the size of the collection and the speed of your computer. Some very large collections take a day or more to build.

When you reach this stage in the interaction, a status line at the bottom of the web page gives feedback on how the operation is progressing, updated every five seconds. The message visible in Figure 4f indicates that when the snapshot was taken, Title metadata was being extracted from an input file.

Warnings are written if input files or URLs are requested that do not exist, or exist but there is no plugin that can process them, or the plugin cannot find an associated file, such as an image file embedded in a HTML document. The intention is that you will monitor progress by keeping this window open in your browser. If any errors cause the process to terminate, they are recorded in this status area.

You can stop the building process at any time by clicking on the *stop building* button in Figure 4f. If you leave the web page (and have not cancelled the building process with the *stop building* button), the building operation will continue, and the new collection will be installed when the operation completes.

Potential problems

If you are operating under Windows and using the Web Library version of Greenstone, you may experience problems with the Collector. The particular symptoms depend on a combination of the version of Windows you are using, and the kind (and even the version) of web server that you have. Sometimes the feedback status line does **not** get updated, although the building process may be proceeding and may terminate correctly. Sometimes the building process terminates with a failure message from Greenstone.

If you do experience problems with the Collector, you can still build collections from the command line. Read the first few pages of the Developer’s Guide for a detailed walk-through of how to do this.

3.7 Viewing the collection

When the collection is built and installed, the sequence of buttons visible at the bottom of Figures 1a–e appears at the bottom of Figure 4f, with the View collection button active. This takes the user directly to the newly built collection.

Finally, there is a facility for E-mail to be sent to the collection’s contact

E-mail address, and to the system's administrator, whenever a collection is created (or modified.) This allows those responsible to check when changes occur, and monitor what is happening on the system. The facility is disabled by default but can be enabled by editing the *main.cfg* configuration file.

3.8 Working with existing collections

When you enter the Collector you have to specify whether you want to create an entirely new collection or work with an existing one, adding data to it or deleting it. By creating all searching and browsing structures automatically from the documents themselves Greenstone makes it easy to add new information to existing collections. Because no links are inserted by hand, when new documents in the same format become available they can be merged into the collection automatically.

To work with an existing collection, you first select the collection from a list that is provided. Some collections are "write protected" and cannot be altered: these ones don't appear in the selection list. With the collection, you can

- Add more data and rebuild the collection
- Edit the collection configuration file
- Delete the collection entirely.

Add new data

The files that you specify will be added to the collection. Make sure that you do not re-specify files that are already in the collection—otherwise two copies will be included. Files are identified by their full pathname, Web pages by their absolute Web address. You specify directories and files just as you do when building a new collection.

If you add data to a collection and for some reason the building process fails, the old version of the collection remains unchanged.

Edit configuration file

Advanced users can edit the collection configuration file, just as they can when a new collection is built. Part 5 below explains the configuration settings.

Delete the collection

You will be asked to confirm whether you really want to delete the collections. Once deleted, Greenstone can not bring the collection back!

3.9 Document formats

When building collections, Greenstone processes each different format of source document by seeking a "plugin" that can deal with that particular format. Plugins are specified in the collection configuration file. Greenstone uses the filename to determine document formats—for example, *foo.txt* is processed as a text file, *foo.html* as HTML, and *foo.doc* as a Word file.

Here is a summary of the plugins that are available for widely-used document formats. More detail about these plugins, and additional plugins for less commonly-used formats, can be found in the *Greenstone Digital Library Developer's Guide.*

TEXTPlug (**.txt*, **.text*)

This interprets a plain text file as a simple document. It adds *title* metadata based on the first line of the file.

HTMLPlug (**.htm*, **.html*; also *.shtml*, *.shm*, *.asp*, *.php*, *.cgi*)

This processes HTML files. It extracts *title* metadata based on the <title> tag; other metadata can be extracted too. There are many options available with this plugin.

WORDPlug (**.doc*)

This imports Microsoft Word documents. There are many different variants on the Word format—and even Microsoft programs frequently make conversion errors. Greenstone uses the program *wvWare* to convert Word files to HTML. However, *wvWare* does not cope with documents in the Rich Text Format (RTF), and for these the program *RHTC* is used. Sometimes these programs fail—for example, *wvWare* does not cope with some older Word formats—and if so, WORDPlug resorts to a simple extraction algorithm that finds all ASCII text strings in the input file.

A collection created using the default collection configuration file, used when creating brand new collections, will show the HTML equivalent of the file when the user clicks the *document* icon. In some cases the HTML version will be poorly formatted—but the text that has been extracted from the document is still useful for indexing and searching. Consequently, we recommend altering the format strings in the collection configuration file to give the user access to the original Word file, instead of the HTML version. Whether the user will see this file depends on how his or her browser is set up. Usually, when viewing on a Windows computer, Word files will be displayed correctly.

PDFPlug (**.pdf*)

This imports documents in PDF Adobe's Portable Document Format. Like WORDPlug, it uses an independent program, in this case *pdftohtml*, to convert PDF files to HTML.

As with WORDPlug, by default collections will display the HTML equivalent of the file when the user clicks the *document* icon; however, the format strings in the collection configuration file can be adjusted to give the user access to the original PDF file instead, and we recommend that you do this.

The *pdftohtml* program fails on some PDF files. What happens is that the conversion process takes an exceptionally long time, and often an error message relating to the conversion process appears on the screen. If this occurs, the only solution that we can offer is to remove the offending document from the collection.

PSPlug (**.ps*)

This imports documents in PostScript. It relies on a standard Linux program, called *ps2html*, being already installed on your computer. This is available on most Linux installations, but not on Windows. Thus PSPlug will not work on Windows computers.

EMAILPlug (**.email*)

This imports files containing E-mail, and deals with common E-mail formats such as are used by the Netscape, Eudora, and Unix mail readers. Each source document is examined to see if it contains an E-mail, or several E-mails joined together in one file, and if so its contents are processed. The plugin extracts *Subject*, *To*, *From*, and *Date* metadata. However, this plugin does not yet handle MIME-encoded E-mails properly—although legible, they often look rather strange.

ZIPPlug (*.gz*, *.z*, *.tgz*, *.taz*, *.bz*, *.zip*, *.tar*)

This plugin handles the following compressed and/or archived input formats: gzip (*.gz*, *.z*, *.tgz*, *.taz*), bzip (*.bz*), zip (*.zip .jar*), and tar (*.tar*). It relies on the programs *gunzip*, *bunzip*, *unzip*, and *tar*, which are standard Linux utilities. ZIPPlug is disabled on Windows computers.

4
Administration

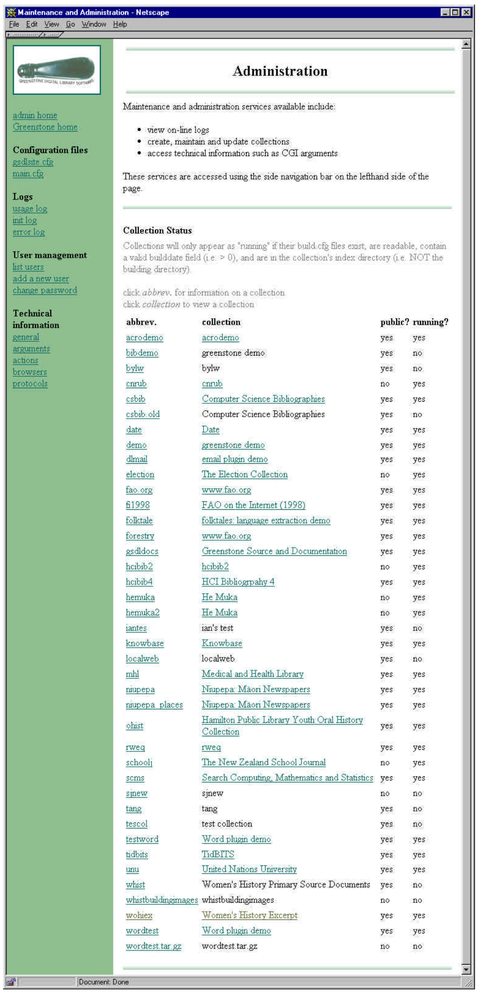
An “administrative” facility is included with every Greenstone installation. On the default Greenstone setup, you can access this facility by clicking the appropriate link on the front page. Otherwise, you can access it by typing into your browser the special URL

http://localhost/gsdl /cgi-bin/library?a=status&p=frameset

The entry page, shown in Figure 5a, gives information about each of the collections offered by the system. Note that *all* collections are included—for there may be “private” ones that do not appear on the Greenstone home page. With each is given its short name, full name, whether it is publicly displayed, and whether or not it is running. Clicking a particular collection’s abbreviation (the first column of links in Figure 5a) brings up information about that collection, gathered from its collection configuration file and from other internal structures created for that collection. If the collection is both public and running, clicking the collection’s full name (the second link) takes you to the collection itself.

The collection we built in Sections 3.2–3.6 has been named *wohiex*, for *Women’s History Excerpt*, and is (barely) visible near the bottom of Figure 5a. Figure 5b shows the information that is displayed when this link is clicked. The first section gives some information from the configuration file, and the size of the collection (about 1000 documents, about a million words, over 6 Mb). The next sections contain internal information related to the communication protocol through which collections are accessed. For example, the filter options for “QueryFilter” show the options and possible values that can be used when querying the collection.

The administrative facility also presents configuration information about the installation and allows it to be modified. It facilitates examination of the error logs that record internal errors, and the user logs that record usage. It enables a specified user (or users) to authorize others to build



Administration

Maintenance and administration services available include:

- view on-line logs
- create, maintain and update collections
- access technical information such as CGI arguments

These services are accessed using the side navigation bar on the lefthand side of the page.

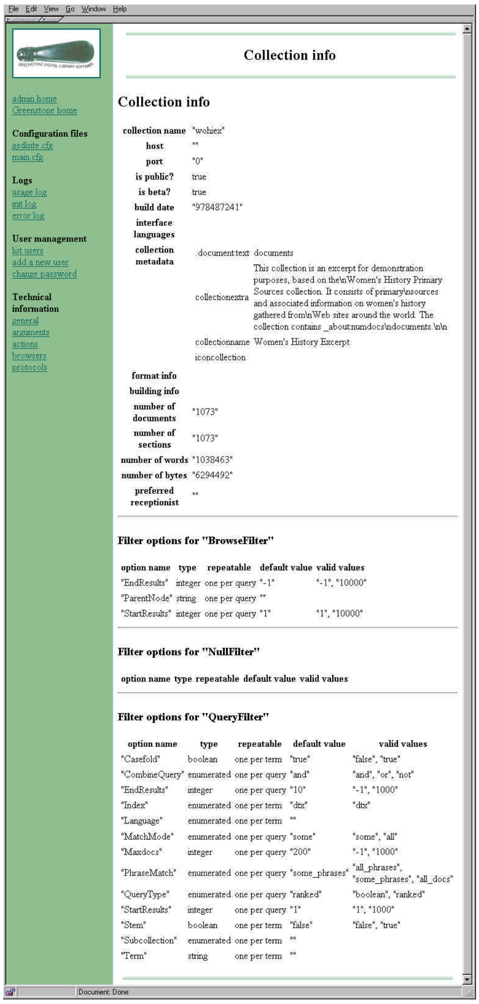
Collection Status

Collections will only appear as "running" if their build.cfg files exist, are readable, contain a valid builddate field (i.e. > 0), and are in the collection's index directory (i.e. NOT the building directory).

click *abbrev.* for information on a collection
click *collection* to view a collection

abbrev.	collection	public?	running?
acrodemo	acrodemo	yes	yes
bibdemo	greenstone demo	yes	no
bylw	bylw	yes	no
cnrub	cnrub	no	yes
csbib	Computer Science Bibliographies	yes	yes
csbib.old	Computer Science Bibliographies	yes	no
date	Date	yes	yes
demo	greenstone demo	yes	yes
dlmail	email plugin demo	yes	yes
election	The Election Collection	no	yes
fao.org	www.fao.org	yes	yes
fi1998	FAO on the Internet (1998)	yes	yes
folktale	folktales: language extraction demo	yes	yes
forestry	www.fao.org	yes	yes
gsdldocs	Greenstone Source and Documentation	yes	yes
hcibib2	hcibib2	no	yes
hcibib4	HCI Bibliogrpahy 4	yes	yes
hemuka	He Muka	no	yes
hemuka2	He Muka	no	yes
iantes	ian's test	yes	no
knowbase	Knowbase	yes	yes
localweb	localweb	yes	no
mhl	Medical and Health Library	yes	yes
niupepa	Niupepa: Māori Newspapers	yes	yes
niupepa_places	Niupepa: Māori Newspapers	yes	yes
ohist	Hamilton Public Library Youth Oral History Collection	yes	yes
rweq	rweq	yes	yes
schoolj	The New Zealand School Journal	no	yes
scms	Search Computing, Mathematics and Statistics	yes	yes
sjnew	sjnew	no	no
tang	tang	yes	no
tescol	test collection	yes	no
testword	Word plugin demo	yes	yes
tidbits	TidBITS	yes	yes
unu	United Nations University	yes	yes
whist	Women's History Primary Source Documents	yes	no
whistbuildingimages	whistbuildingimages	no	no
wohiex	Women's History Excerpt	yes	yes
wordtest	Word plugin demo	yes	yes
wordtest.tar.gz	wordtest.tar.gz	no	no

(a)



Collection info

Collection info

collection name	"wohiex"	
host	""	
port	"0"	
is public?	true	
is beta?	true	
build date	"978487241"	
interface languages		
collection metadata	.document:text	documents
	collectionextra	This collection is an excerpt for demonstration purposes, based on the\nWomen's History Primary Sources collection. It consists of primary\nsources and associated information on women's history gathered from\nWeb sites around the world. The collection contains _about:numdocs\ndocuments.\n\n
	collectionname	Women's History Excerpt
	iconcollection	
format info		
building info		
number of documents	"1073"	
number of sections	"1073"	
number of words	"1038463"	
number of bytes	"6294492"	
preferred receptionist	""	

Filter options for "BrowseFilter"

option name	type	repeatable	default value	valid values
"EndResults"	integer	one per query	"-1"	"-1", "10000"
"ParentNode"	string	one per query	""	
"StartResults"	integer	one per query	"1"	"1", "10000"

Filter options for "NullFilter"

option name	type	repeatable	default value	valid values

Filter options for "QueryFilter"

option name	type	repeatable	default value	valid values
"Casefold"	boolean	one per term	"true"	"false", "true"
"CombineQuery"	enumerated	one per query	"and"	"and", "or", "not"
"EndResults"	integer	one per query	"10"	"-1", "1000"
"Index"	enumerated	one per term	"dtx"	"dtx"
"Language"	enumerated	one per term	""	
"MatchMode"	enumerated	one per query	"some"	"some", "all"
"Maxdocs"	integer	one per query	"200"	"-1", "1000"
"PhraseMatch"	enumerated	one per query	"some_phrases"	"all_phrases", "some_phrases", "all_docs"
"QueryType"	enumerated	one per query	"ranked"	"boolean", "ranked"
"StartResults"	integer	one per query	"1"	"1", "1000"
"Stem"	boolean	one per term	"false"	"false", "true"
"Subcollection"	enumerated	one per term	""	
"Term"	string	one per term	""	

(b)

Figure 5 The Greenstone Administration facility

collections and add new material to existing ones. All these facilities are accessed interactively from the menu items at the left-hand side of Figure 5a.

4.1 Configuration files

There are two configuration files that control Greenstone's operation, the site configuration file *gsdlsite.cfg* and the main configuration file *main.cfg*.

The *gsdlsite.cfg* file is used to configure the Greenstone software for the site where it is installed. It is designed for keeping configuration options that are particular to a given site. Examples include the name of the directory where the Greenstone software is kept, the HTTP address of the Greenstone system, and whether the *fastcgi* facility is being used. The entries in this file are described in the *Greenstone Digital Library Installation Guide.*

The *main.cfg* file contains information that is common to the interface of all collections served by a Greenstone site. It includes the E-mail address of the system maintainer, whether the status and collector pages are enabled, whether logs of user activity are kept, and whether Internet "cookies" are used to identify users.

4.2 Logs

Three kinds of logs can be examined: usage logs, error logs and initialization logs. The last two are only really of interest to people maintaining the software.

All user activity—every page that each user visits—can be recorded by the Greenstone software, though no personal names are included in the logs. Logging, disabled by default, is enabled by including the lines

```
logcgiargs true
usecookies true
```

in the main system configuration file. Both options are false by default, so that no logging is done unless they are set. It is the *logcgiargs* line that actually turns logging on and off. However, without *usecookies* the user's identity will not be recorded.

Each line in the user log pertains to a page visit. Each line in the user log records a page visited—even the pages generated to inspect the log files! It contains (a) the IP address of the user's computer, (b) a timestamp in square brackets, (c) the CGI arguments in parentheses, and (d) the name of the user's browser (Netscape is called "Mozilla"). Here is a sample line, split and annotated for ease of reading:

```
      /fast-cgi-bin/niupepalibrary
(a)   its-www1.massey.ac.nz
(b)   [Thu Dec 07 23:47:00 NZDT 2000]
(c)   (a=p, b=0, bcp=, beu=, c=niupepa, cc=, ccp=0, ccs=0, cl=, cm=,
      cq2=, d=, e=, er=, f=0, fc=1, gc=0, gg=text, gt=0, h=, h2=, hl=1,
      hp=, il=l, j=, j2=, k=1, ky=, l=en, m=50, n=, n2=, o=20, p=home,
      pw=, q=, q2=, r=1, s=0, sp=frameset, t=1, ua=, uan=, ug=,
      uma=listusers, umc=, umnpw1=, umnpw2=, umpw=, umug=, umun=, umus=,
      un=, us=invalid, v=0, w=w, x=0, z=130.123.128.4-950647871)
(d) “Mozilla/4.08 [en] (Win95; I ;Nav)”
```

The last CGI argument, “z”, is an identification code or “cookie” generated by the user’s browser: it comprises the user’s IP number followed by the timestamp when they first accessed the digital library.

The log file is placed in the *etc\usage.txt* directory in the Greenstone file structure (see the *Greenstone Digital Library Developer’s Guide*). When logging is enabled, every action by every user is logged. However, only the last 100 entries in the log file are displayed by the *usage log* link in Figure 5a.

4.3 User management

Greenstone incorporates an authentication scheme which can be used to control access to certain facilities. At the moment this is only used to restrict the people who are allowed to enter the Collector and certain administration functions. If, for a particular collection, it were necessary to authenticate users before returning information to to them, this is possible too—for example, documents could be protected on an individual basis so that they can only be accessed by registered users on presentation of a password. However, no current collections use this facility). Authentication is done by requesting a user name and password, as illustrated in Figure 4a.

From the administration page users can be listed, new ones added, and old ones deleted. The ability to do this is of course also protected: only users who have administrative privileges can add new users. It is also possible for each user to belong to different “groups”. At present, the only extant groups are “administrator” and “colbuilder”. Members of the first group can add and remove users, and change their groups. Members of the second can access the facilities described above to build new collections and alter (and delete) existing ones.

When Greenstone is installed, there is one user called *admin* who belongs to both groups. The password for this user is set during the installation process. This user can create new names and passwords for users who belong just to the *colbuilder* group, which is the recommended way of giving other users the ability to build collections. User information is recorded in two databases that are placed in the Greenstone file structure (see the *Greenstone Digital Library Developer’s Guide*).

Currently, the system limits certain activities to people belonging to certain groups—namely, *administrator* and *colbuilder*. It would be easy to use these same security features to control access to other parts of the library, and to create public and private versions of collections in which certain parts were only available to certain users.

4.4 Technical information

The links under the *Technical information* heading show further information on the installation. The *general* link gives access to technical information, including the directories where things are stored. The *protocols* menu item gives information about each of the collections offered by the system. The user interface code (called the "receptionist") uses *actions* to communicate the wishes of the user. These actions correspond to the CGI argument labeled *a*. For example, if *a=status* the receptionist invokes the *status* action (which displays the status page). A menu item gives access to lists of all actions supported by the system, and another leads to the arguments that these actions take.

32 ADMINISTRATION

Appendix A
Software features

Accessible via Web browser

Collections are accessed through a standard Web browser (Netscape or Internet Explorer) and combine easy-to-use browsing with powerful search facilities.

Full-text and fielded search

The user can search the full text of the documents, or choose between indexes built from different parts of the documents. For example, some collections have an index of full documents, an index of sections, an index of titles, and an index of authors, each of which can be searched for particular words or phrases. Results can be ranked by relevance or sorted by a metadata element.

Flexible browsing facilities

The user can browse lists of authors, lists of titles, lists of dates, classification structures, and so on. Different collections may offer different browsing facilities and even within a collection, a broad variety of browsing interfaces are available. Browsing and searching interfaces are constructed during the building process, according to collection configuration information.

Creates access structures automatically

The Greenstone software creates information collections that are very easy to maintain. 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. This means that if new documents in the same format become available, they can be merged into the collection automatically. Indeed, for many collections this is done by processes that wake up regularly, scout for new material, and rebuild the indexes—all without manual intervention.

Makes use of available metadata

Metadata, which is descriptive information such as author, title, date, keywords, and so on, may be associated with each document, or with individual sections within documents. Metadata is used as the raw material for browsing indexes. It must be either provided explicitly or derivable automatically from the source documents. The Dublin Core metadata scheme is used for most electronic documents, however, provision is made for other schemes.

Plugins extend the system's capabilities

In order to accommodate different kinds of source documents, the software is organized in such a way that "plugins" can be written for new document types. Plugins currently exist for plain text, HTML, Word, PDF, PostScript, E-mail, some proprietary formats, and for recursively traversing directory structures containing such documents. A collection may have source documents in different forms. In order to build browsing indexes from metadata, an analogous scheme of "classifiers" is used: classifiers create browsing indexes of various kinds based on metadata.

Designed for multi-gigabyte collections

Collections can contain millions of documents, making the Greenstone system suitable for collections up to several gigabytes.

Documents can be in any language

Unicode is used throughout the software, allowing any language to be processed in a consistent manner. To date, collections have been built containing French, Spanish, Maori, Chinese, Arabic and English. On-the-fly conversion is used to convert from Unicode to an alphabet supported by the user's Web browser.

User interface available in multiple languages

The interface can be presented in multiple languages. Currently, the interface is available in Arabic, Chinese, Dutch, English, French, German, Maori, Portuguese, and Spanish. New languages can be added easily.

Collections can contain text, pictures, audio, and video

Greenstone collections can contain text, pictures, audio and even video clips. Most non-textual material is either linked in to the textual documents or accompanied by textual descriptions (such as figure captions) to allow full-text searching and browsing. However, the architecture permits implementation of plugins and classifiers even for non-textual data.

Uses advanced compression techniques

Compression techniques are used to reduce the size of the indexes and text. Reducing the size of the indexes via compression has the added advantage of increasing the speed of text retrieval.

Administrative function provided

An "administrative" function enables specified users to authorize new users to build collections, protect documents so that they can only be accessed by registered users on presentation of a password, examine the composition of all collections, and so on. Logs of user activity can record all queries made to every Greenstone collection.

New collections appear dynamically

Collections can be updated and new ones brought on-line at any time, without bringing the system down; the process responsible for the user interface will notice (through periodic polling) when new collections appear and add them to the list presented to the user.

Collections can be published on the Internet or on CD-ROM

The software can be used to serve collections over the World-Wide Web. Greenstone collections can be made available, in precisely the same form, on CD-ROM. The user interface is through a standard Web browser (Netscape is provided on each disk), and the interaction is identical to accessing the collection on the Web—except that response times are more predictable. The CD-ROMs run under all versions of the Windows operating system.

Collections can be distributed amongst different computers

A flexible process structure allows different collections to be served by different computers, yet be presented to the user in the same way, on the same Web page, as part of the same digital library.

Operates on both Windows and Unix

Greenstone runs under both Windows (3.1/3.11, 95/98, NT) and Unix (Linux and SunOS). Any of these systems can be used as a webserver. New collections cannot be built on low-end Windows systems (3.1/3.11).

What you get with Greenstone

The Greenstone Digital Library is open-source software, available from the New Zealand Digital Library (*nzdl.org*) under the terms of the GNU General Public License. The software includes everything described above: Web serving, CD-ROM creation, collection building, multi-lingual capability, plugins and classifiers for a variety of different source document types. It includes an autoinstall feature to allow easy installation on both Windows and Unix. In the spirit of open-source software, users are encouraged to contribute modifications and enhancements.

Appendix B
Glossary of terms

Term	Meaning
autoconf	Unix program used to configure the Greenstone software installation package to suit your system
Autorun	Windows feature that starts a program automatically whenever a CD-ROM is inserted
Boolean query	Query to an information retrieval system that may contain AND, OR, NOT
Browsing	Accessing a collection by scanning an organized list of metadata values associated with the documents (such as author, title, date, keywords)
buildcol.pl	Program used to build collections
Building	Process of creating the indexing and browsing structures that are used to access a collection
C++	Programming language in which the majority of the Greenstone software is written
Casefolding	Making uppercase and lowercase words look the same, for searching purposes
CGI	Common Gateway Interface, a scheme that allows users to activate programs on the host computer by clicking on Web pages
CGI script	Code associated with a button, menu, or link on a Web page that specifies what the host computer is to do when it is clicked
cgi-bin	Directory in which CGI scripts are stored
Classifier	Greenstone code module that examines document metadata to form an index for browsing
Collection	Set of documents that are brought together under a uniform searching and browsing interface
Collection configuration file	File that specifies how a collection is to be imported and built, what indexes and language interfaces are to be provided, etc
Collection server	Program responsible for providing access to a collection when it is being used

Configuration file See collection configuration file, main configuration file, site configuration file

CVS Concurrent Versioning System, a scheme for maintaining source code used throughout Greenstone

db2txt Tool for viewing a GDBM database as text (see GDBM)

Demo collection A subset of the Humanities Development Library, distributed with the Greenstone software and used for illustration in this tutorial

Digital library Collection of digital objects (text, audio, video), along with methods for access and retrieval, and for selection, organization, and maintenance

Document Basic unit from which digital library collections are constructed; it may include text, graphics, sound, video, etc.

Dublin core A standard way of describing metadata

Fast CGI Facility that allows CGI scripts to remain continuously active so that they do not have to be restarted from scratch every time they are invoked

Filter program That part of a collection server that implements querying and browsing operations

Format string A string that specifies how documents and other listings are to be displayed

GB-encoding Standard way of encoding the Chinese language

GDBM GNU DataBase Manager, a program used within the Greenstone software to store metadata for each document

GIMP GNU Image-Manipulation Program used (on Unix) to create icons in Greenstone

GML Greenstone Markup Language, a file format used for storing documents internally

GNU license Software license that permits users to copy and distribute computer programs freely, and modify them—so long as all modifications are made publicly available

Greenstone The name of this digital library software

GSDL Abbreviation for Greenstone Digital Library

%GSDLHOME% Operating system variable that represents the top-level directory in which all Greenstone programs and collections are stored (*$GSDLHOME* on Unix systems)

%GSDLOS% Operating system variable that represents the operating system currently being used (*$GSDLOS* on Unix systems)

hashfile Program used at import or build time to generate the OID of each document

HDL Humanity Development Library, a Greenstone collection of humanitarian information for developing countries

HTML	HyperText Markup Language, the language in which Web documents are written
import.pl	Program used to import documents
Importing	Process of bringing collections of documents into the Greenstone system
Index	Information structure that is used for searching or browsing a collection
InstallShield	Windows program, used by Greenstone CD-ROMs, that allows a system to be installed from a CD-ROM
Main configuration file	File that contains specifications common to all collections served by this site
Metadata	Descriptive data such as author, title, date, keywords, and so on, that is associated with a document (or document collection)
MG	Managing Gigabytes, a program used by the Greenstone system for full-text indexing, that incorporates compression techniques (see Witten, I.H., Moffat, A. and Bell, T. *Managing Gigabytes: compressing and indexing documents and images*, Morgan Kaufmann, second edition, 1999)
mgbuild	MG program for building a compressed full-text index
mgquery	MG program for querying a compressed full-text index
mkcol.pl	Program that creates and initializes the directory structure for a new collection
New Zealand Digital Library	Research project in the Computer Science Department at the University of Waikato, New Zealand, that created the Greenstone software (*nzdl.org*)
OID	Object Identifier, a unique identification code associated with a document
Perl	Programming language used for many of the text-processing operations that occur during the building process
Ping	Message sent to a system to determine whether it is running or not
Plugin	Code module for handling documents of different formats, used during the importing and building processes
Protocol	Set of conventions by which a receptionist communicates with a collection server
Ranked query	Natural-language query to an information retrieval system, for which the documents that match the query are sorted in order of relevance
Receptionist	Program that organizes the Greenstone user interface
RTF	Rich Text Format, a standard format for interchange of text documents
Searching	Accessing a collection through a full-text search of its contents (or parts of contents, such as section titles)
Server	See Collection server, Web server
setup.bat, setup.sh, setup.csh	Command used to set up your environment to recognize the Greenstone software
Site configuration	File that contains specifications used to configure the Greenstone

file	software for the site on which it is installed
Stemming	Stripping endings off a query term to make it more general
STL	Standard template library, a widely-available library of C++ code
txt2db	Program used at build time to create the GDBM database
Unicode	Standard scheme for representing the character sets used in the world’s languages
UNU	The United Nations University; also used to refer to a Greenstone collection created for that organization
Web server	Standard program that computers use to make information accessible over the World Wide Web