

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 developed and distributed in cooperation with UNESCO and the Human Info NGO. It is open-source software, available from *http://greenstone.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-2.50 **January 2004**

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 Librarian Interface. 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 four volumes:

- Greenstone Digital Library Installer's Guide
- Greenstone Digital Library User's Guide *(this document)*
- Greenstone Digital Library Developer's Guide
- Greenstone Digital Library: From Paper to Collection

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 inspiration for the design of the system: Mark Apperley, Sally Jo Cunningham, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui and Lloyd Smith. We also thank all who have contributed to the packages included in this distribution: MG, GDBM, WGET, WV, PDF2HTML, PERL.



Contents

About this manual ii

1 OVERVIEW OF GREENSTONE 1

1.1 Collections 1

1.2 Finding information 2

1.3 Document formats 2

1.4 Multimedia and multilingual documents 3

1.5 Distributing Greenstone 3

2 USING GREENSTONE COLLECTIONS 5

2.1 Using a Greenstone CD-ROM 5

2.2 Finding information 6
How to find information 7
How to read the documents 8
What the icons mean 9
How to search for particular words 9
Scope of queries 11
Advanced search features 11

2.3 Changing the preferences 13
Collection preferences 13
Language preferences 13
Presentation preferences 14
Search preferences 15

3 MAKING GREENSTONE COLLECTIONS 17

3.1 The librarian's interface 18
Getting started 18

Assembling the source material 20
Enriching the documents 22
Designing the collection 28
Building the collection 30
Previewing 31
Help 31

3.2 Librarian Interface user guide 31
Starting Off 31
Hunting For Your Files 33
Downloading Chosen Files 33
Collecting Files for Your Collection 35
Enriching the Collection with Metadata 37
Designing Your Collection's Appearance 43
Producing Your Collection 50
Previewing the Collection 51
Miscellaneous 52
Metadata Set and Profile Editing 53

3.3 Tagging document files 56

3.4 The Collector 58
Logging in 59
Dialog structure 60
Collection information 61
Source data 62
Configuring the collection 66
Building the collection 67
Viewing the collection 68
Working with existing collections 68
Document formats 70

4 ADMINISTRATION 72

5.1 Configuration files 75

5.2 Logs 75

5.3 User management 76

5.4 Technical information 77

APPENDIX A SOFTWARE FEATURES 78

APPENDIX B GLOSSARY OF TERMS 82



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 searching and browsing are constructed during a "building" process, according to information in a collection configuration file.

Greenstone creates all index structures automatically from the documents and suppporting files: 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 Document formats

Source documents come in a variety of formats, and are converted into a standard XML form for indexing by "plugins." Plugins distributed with Greenstone process plain text, HTML, WORD and PDF documents, and Usenet and Email 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 Distributing Greenstone

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, and text extraction from proprietary document formats. 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.



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 in conjunction with pre-compiled binaries. 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 that collection. To use it, simply 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, then the *Setup.exe* file inside it. The Greenstone *Setup* program will be entered, which guides

you through the setup procedure. Most people respond *yes* to all the questions.

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, "Development Library" 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

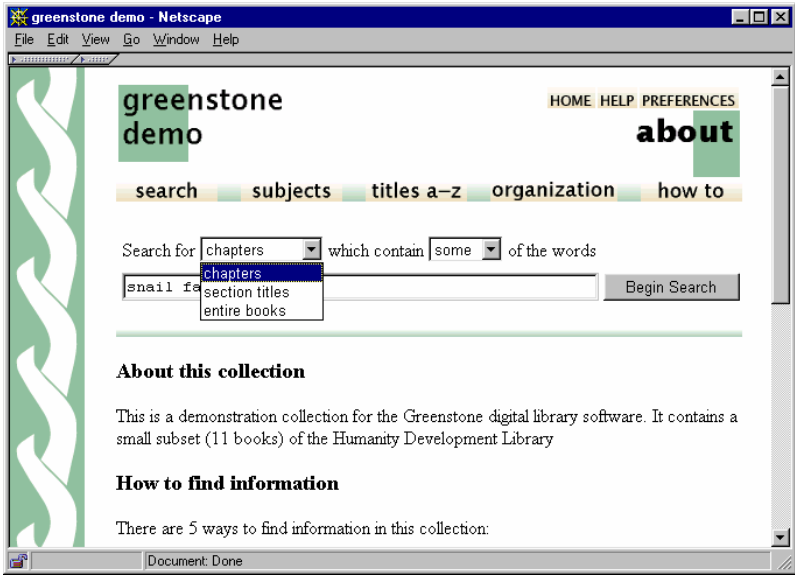
The easiest way to learn how to use a Greenstone collection is to try it out. Don't worry—you can't break anything. Click liberally: most images that appear on the screen are clickable. If you hold the mouse stationary over an image, most browsers will soon pop up a message that tells you what will happen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



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 last four 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 can 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 will 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 that can be further expanded by clicking on them. 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. All these buttons are visible in Figure 1.

How to read the documents

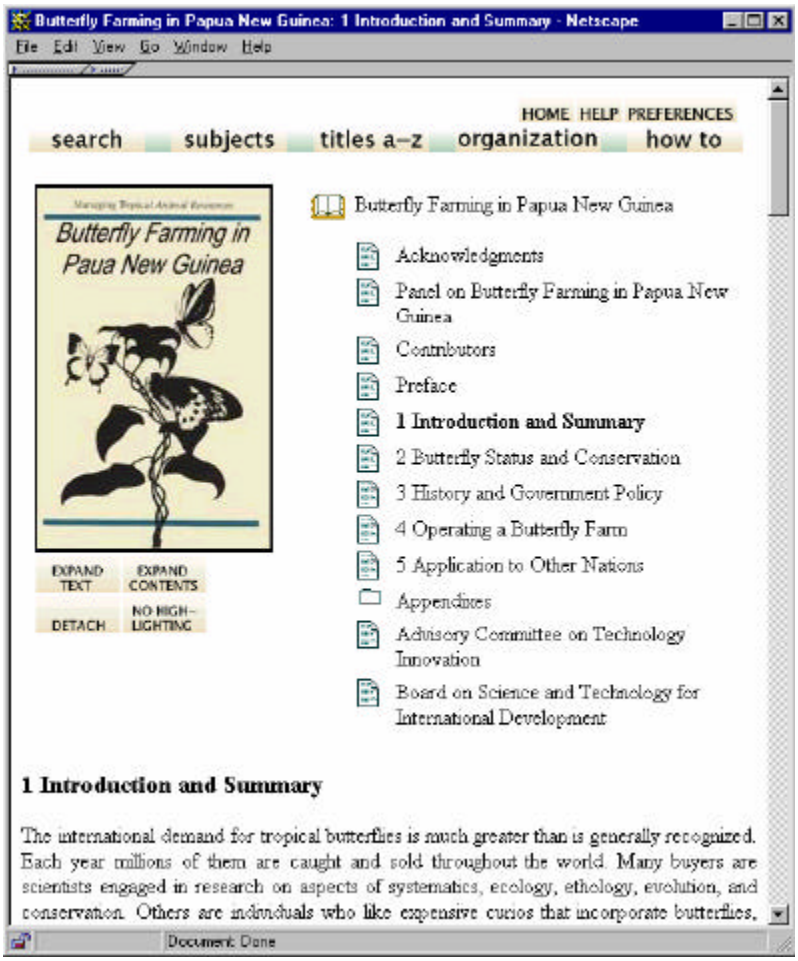
In the Demo collection, you can tell when you have arrived at an individual book because there is a photograph of its front cover (Figure 2). Beside the photograph is a table of contents: the entry in bold face marks where you are, in this case *Introduction and Summary*—Section 1 of the chosen book. 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 ("The international demand for tropical butterflies …" 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



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
	Open this page in a new window
	Expand table of contents
	Display all text
	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, 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?
```

Phrase matches are case-insensitive if *ignore case differences* is set on the Preferences page.

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 between search terms 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.

2.3 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. When you adjust your search preferences, you should press the *set preferences* button shown in Figure 3. After setting preferences, do not use your browser's "back" button—that would unset them! Instead, click one of the buttons on the access bar near the top of the page.

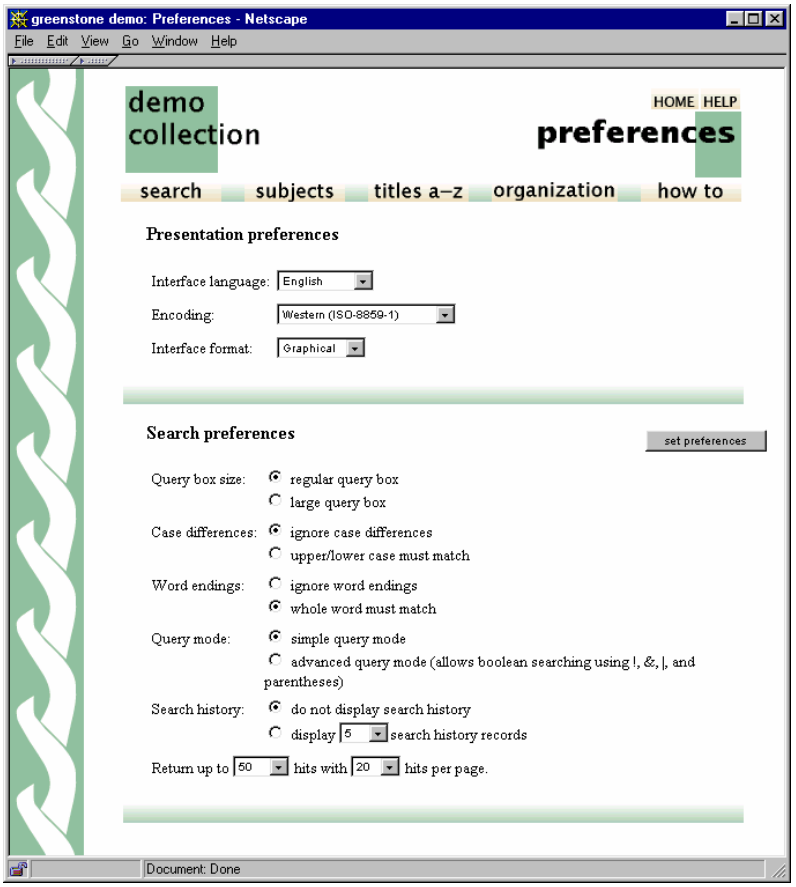
Collection preferences

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.

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

Figure 3
The Preferences page



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

Under *Search preferences* in Figure 3, the first pair of buttons allows you to get a large query box, so that you can easily do paragraph-sized searching. In Greenstone, it is surprisingly quick to search for large amounts of text. The next 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.

Using the next button pair you can switch to the "advanced" query mode described above, which allows you to specify more precise queries by combining terms using AND (&), OR (|), and NOT (!). 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, through the last entry in Figure 3.



3
Making Greenstone Collections

The simplest way to build new collections is to use Greenstone's "librarian" interface (GLI). This allows you to collect sets of documents, import or assign metadata, and build them into a Greenstone collection. It supports five basic activities, which can be interleaved but are nominally undertaken in this order:

1. Copy documents from the computer's file space, including existing collections, into the new collection. Any existing metadata remains "attached" to these documents. Documents may also be gathered from the web through a built-in mirroring facility.
2. Enrich the documents by adding further metadata to individual documents or groups of documents.
3. Design the collection by determining its appearance and the access facilities that it will support.
4. Build the collection using Greenstone.
5. Preview the newly created collection, which will have been installed on your Greenstone home page as one of the regular collections.

The librarian interface allows you to add what people call "external" metadata to documents, metadata that pertains to the document as a whole. But documents often need to be structured into sections and subsections, and "internal" metadata might be associated with each part. In Greenstone, source documents can be tagged with this information, and we explain this in Section 3.2.

Finally, an alternative way of building collections is provided by the Collector, which helps you creat new collections, modify or add to existing ones, or delete collections. It predates the librarian interface, and for most practical purposes the librarian interface should be used instead of the Collector. It is described in Section 3.3.

To harness the full power of Greenstone to build advanced collections, you will also need to read Chapter 2 of the *Developer's Guide*.

3.1 The librarian's interface

To convey the operation of Greenstone's librarian interface, we work through a simple example. Figures 4 to 15 are screen snapshots at various points during the interaction. This example uses documents in the Development Library Subset (DLS) collection, which is distributed with Greenstone. For expository purposes, the walkthrough takes the form of a single pass through the steps listed above. A more realistic pattern of use, however, is for users to switch back and forth through the various stages as the task proceeds.

Getting started

Launch the librarian interface under Windows by selecting *Greenstone Digital Library* from the *Programs* section of the *Start* menu and choosing *Librarian Interface*. If you are using Unix, instead type

```
cd ~/gsdl
cd gli
./gli.sh
```

where *~/gsdl* is the directory containing your Greenstone system. To begin, you must either open an existing collection or start a new one. Figure 4 shows the user in the process of starting a new collection. She has selected *New* from the file menu and begun to fill out general information about the collection—its title, the E-mail address of the person responsible for it, and a brief description of the content—in the popup window. The collection title is a short phrase used throughout the digital library to identify the collection's content: existing collections have names like *Food and Nutrition Library*, *World Environmental Library*, and so on. When you type the title, the system assigns a unique mnemonic identifier, the collection "name", for internal use (you can change it if you like). The Email address specifies the first point of contact for any problems encountered with the collection.

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 collection's initial page.

Figure 16. Starting a new collection

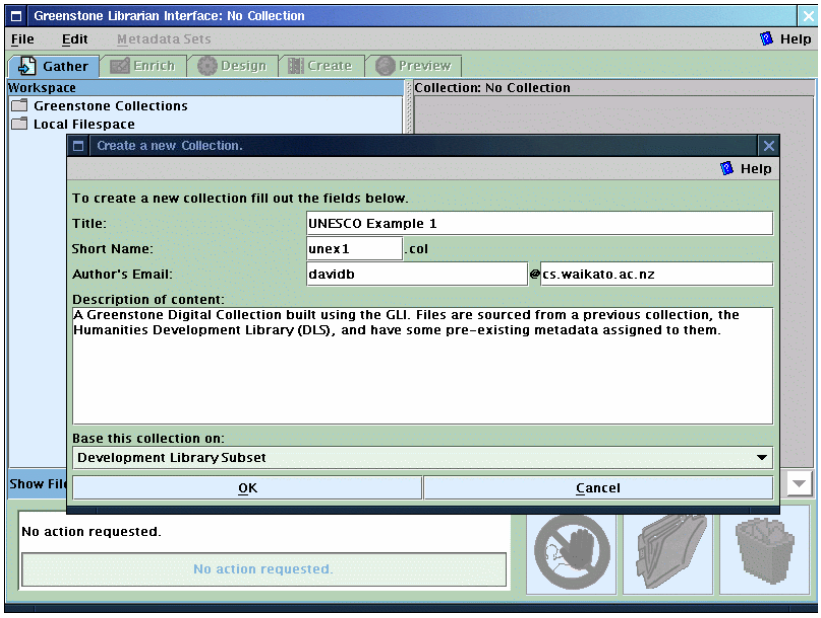
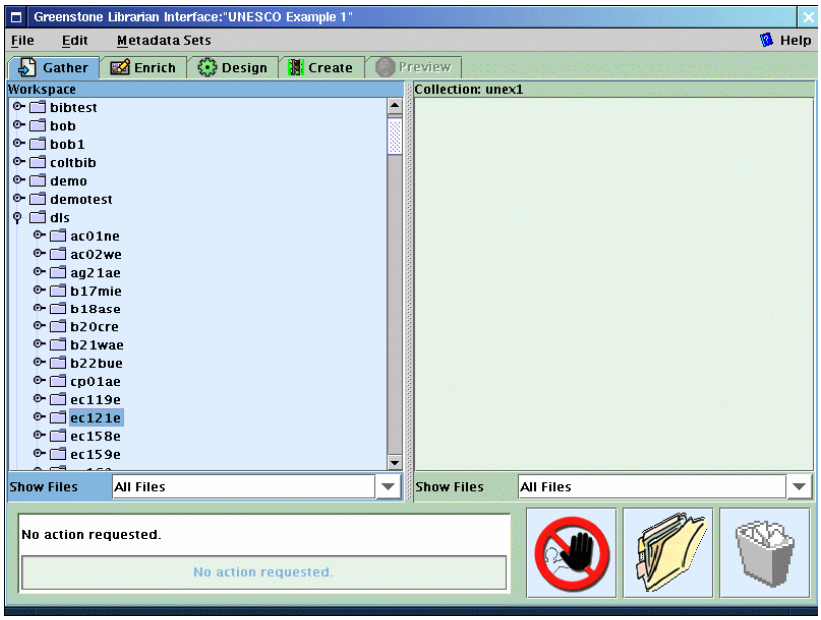


Figure 5. Exploring the local file space



At this point, the user decides whether to base the new collection on the same structure as an existing collection, or to build an entirely new kind of collection. In Figure 4 she has chosen to base it on the *Development Library Subset* collection. This implies that the "DLS" metadata set which is used in this collection will be used for the new collection. (In fact, this metadata set has been used to build several Greenstone collections that share a common structure and organization but with different content, including the *Development Library Subset* and *Demo* collections delivered as samples with Greenstone.)

The DLS metadata set contains these items:

- Title
- Subject
- Language
- Organization
- Keyword (i.e."Howto").

(There is, in addition, a metadata item called *AZList* which is used to determine which bucket of the alphabetic list contains the document's title, with values like "A-B" or "C-D-E". This is used to give precise control over the divisions in the list. For most other collections it is absent, and Greenstone assigns the buckets itself.)

If, instead, the user had chosen "New Collection" at this point, she would have been asked to select what metadata sets should be used in the new collection. Three standard sets are pre-supplied: Dublin Core, the DLS metadata set mentioned above, and a set that comprises metadata elements extracted automatically by Greenstone from the documents in the collection. The user can also create new metadata sets using a popup panel activated through the "metadata" menu.

Several different metadata sets can be associated with the same collection; the system keeps them distinct (so that, for example, documents can have both a Dublin Core *Title* and a DLS *Title*). The different sets are clearly distinguished in the interface. Behind the scenes, metadata sets are represented in XML.

Assembling the source material

After clicking the *OK* button on the "new collection" popup, the remaining parts of the interface, which were grayed out before, become active. The *Gather* panel, selected by the eponymous tab near the top of Figure 4, is displayed initially. This allows the user to explore the local file space and existing collections, gathering up selected documents for the new collection. The panel is divided into two sections, the left for

browsing existing structures and the right for the documents in the collection.

Operations available at this stage include:

- Navigating the existing file structure hierarchy, and the one being created, in the usual way.
- Dragging and dropping files into the new collection.
- Multiple selection of files.
- Dragging and dropping entire sub-hierarchies.
- Deleting documents from the nascent collection.
- Creating new sub-hierarchies within the collection.
- Filtering the files that are visible, in both the local file system and the collection, based on predetermined groups or on standard file matching terms.
- Invoking the appropriate program to display the contents of a selected file, by double-clicking it.

Care is taken to deal appropriately with name clashes when files of the same name in different parts of the computer's directory structure are copied into the same folder of the collection.

In Figure 5 the user is using the interactive file tree display to explore the local file system. At this stage, the collection on the right is empty; the user populates it by dragging and dropping files of interest from the left to the right panel. Such files are "copied" rather than "moved": so as not to disturb the original file system. The usual techniques for multiple selection, dragging and dropping, structuring the new collection by creating subdirectories ("folders"), and deleting files from it by moving them to a trashcan, are all available. Through an extra panel, which can optionally be activated, the user can browse external Web sites and selectively mirror them; the result of this copying appears as another top-level folder on the left-hand side of the *Gather* panel.

Existing collections are represented by a subdirectory on the left called "Greenstone Collections," which can be opened and explored like any other directory. However, the documents therein differ from ordinary files because they already have metadata attached, and this is preserved when they are moved into the new collection. Conflicts may arise because their metadata may have been assigned using a different metadata set from the one in use for the new collection, and the user must resolve these. In Figure 6 the user has selected some documents from an existing collection and dragged them into the new one. The popup window explains that the metadata element *Organization* cannot be automatically imported, and asks the user to either select a metadata set and press *Add* to add the

metadata element to that set,[1] or choose a metadata set, then an element, and press *Merge* to effectively rename the old metadata element to the new one by merging the two. Metadata in subsequent documents from the same collection will automatically be handled in the same way.

When large file sets are selected, dragged, and dropped into the new collection, the copying operation may take some time—particularly if metadata conversion is involved. To indicate progress, the interface shows which file is being copied and what percentage of files has been processed.

Special facilities are provided for dealing with large file sets. For example, the user can choose to filter the file tree to show only certain files, using a dropdown menu of file types displayed underneath the trees. In Figure 7, only the HTM and HTML files are being shown (and only these files will be copied by drag and drop).

Enriching the documents

The next phase in collection building is to enrich the documents by adding metadata. The *Enrich* tab brings up a new panel of information (Figure 8), which shows the document tree representing the collection on the left and on the right allows metadata to be added to individual documents, or groups of documents.

Documents that are copied during the first step come with any applicable metadata attached. If a document is part of a Greenstone collection, previously defined metadata is carried over to the new collection. Of course, this new collection may have a different metadata set, or perhaps just a subset of the defined metadata, and only metadata that pertains to the new collection's set is carried over. Resolution of such conflicts may require user intervention via a supplementary dialog (Figure 6). Any choices made are remembered for subsequent file copies.

The *Enrich* panel allows metadata values to be assigned to documents in the collection. For example, new values can be added to the set of existing values for an element. If the element's values have a hierarchical structure, the hierarchy can be extended in the same way.

---

[1] This option is disabled if an element of the same name already exists.

Figure 6. Importing existing metadata

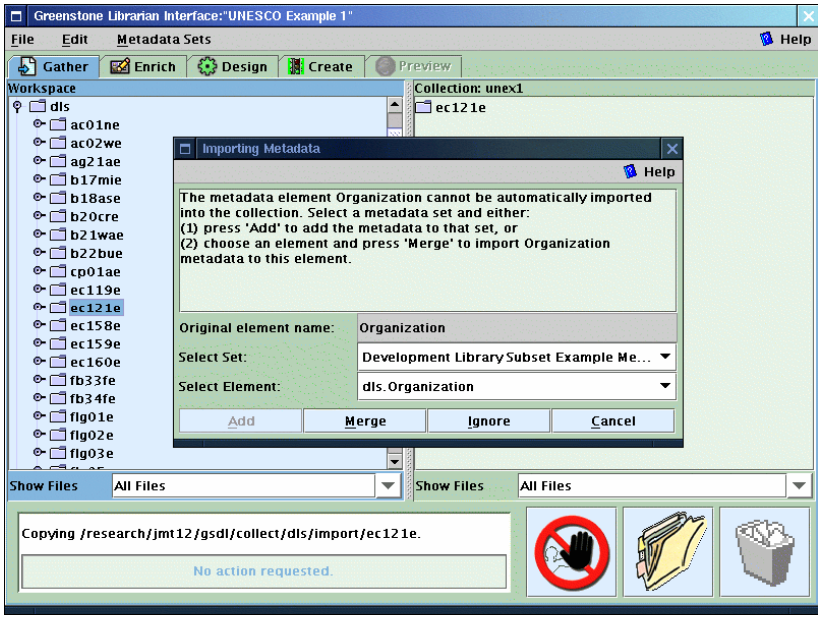


Figure 7. Filtering the file trees

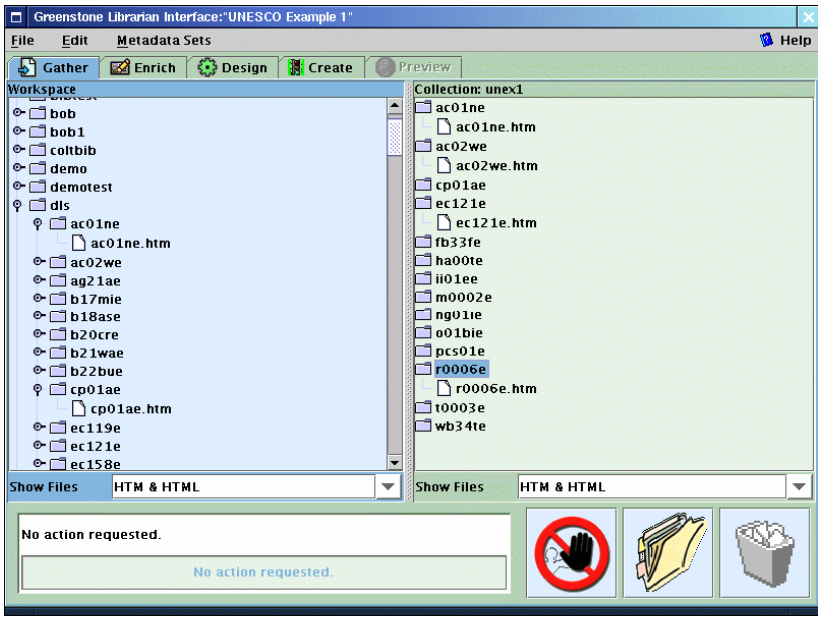


Figure 8. Assigning metadata using *Enrich* view

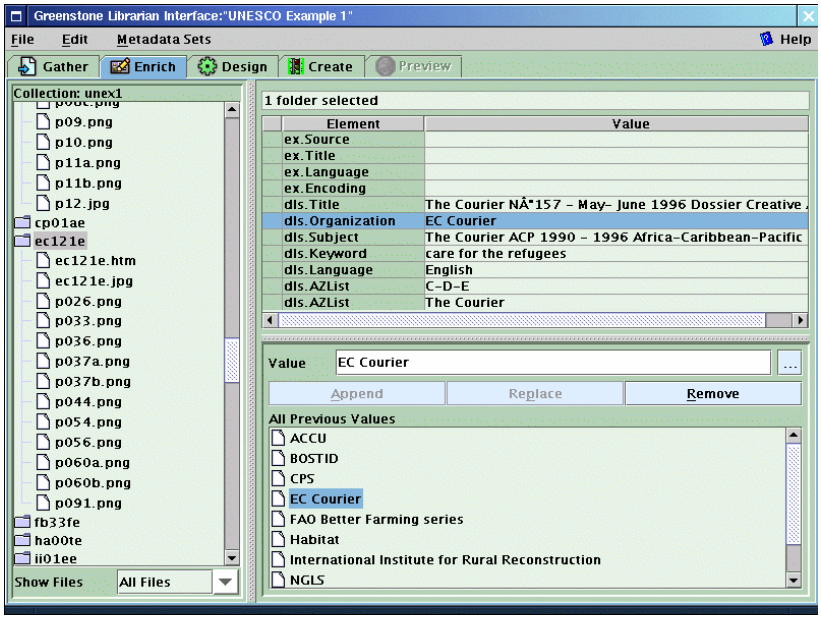
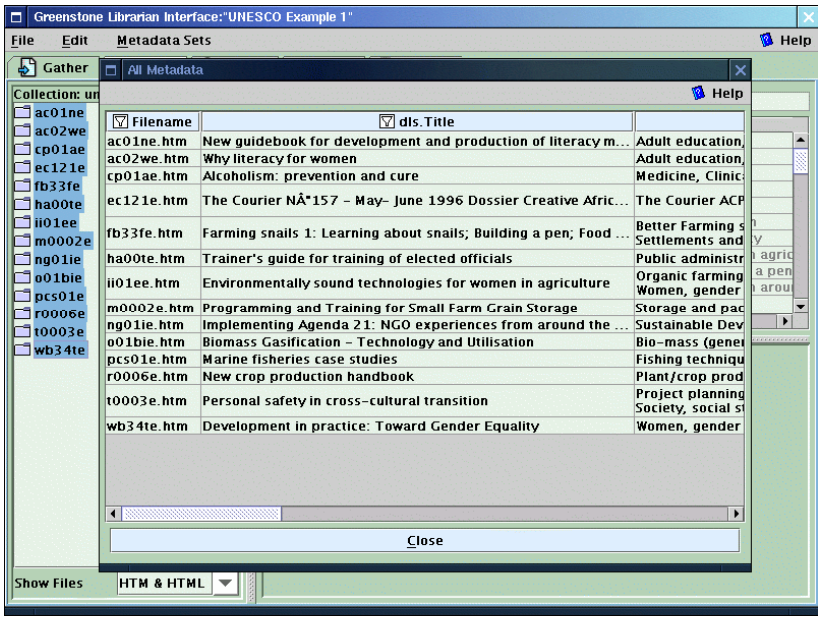


Figure 9. Viewing all metadata for selected files



Metadata values can also be assigned to folders, in just the same way. Documents in these folders for which this metadata is unspecified inherit the metadata values. However, they can subsequently be overridden by supplying different ones for the document itself.

Operations at this stage include:

- Assigning new and existing metadata values to documents.
- Assigning metadata to an individual document.
- Assigning metadata to a folder (this is inherited by all documents in the folder, including those in nested folders).
- Assigning hierarchical metadata, whose structure can be dynamically updated if required.
- Editing or updating assigned metadata.
- Reviewing the metadata assigned to a selection of files and directories.

For our walkthrough example, in Figure 8 the user has selected the folder *ec121e* and assigned "EC Courier" as its *Organization* metadata. The buttons for updating and removing metadata become active depending on what selections have been made.

During the enrichment phase, or indeed at any other time, the user can choose to view all the metadata that has been assigned to documents in the collection. This is done by selecting a set of documents and choosing *Assigned Metadata* from the metadata sets menu, which brings up a popup window like that in Figure 9 that shows the metadata in spreadsheet form. For large collections it is useful to be able to view the metadata associated with certain document types only, and if the user has specified a file filter as mentioned above, only the selected documents are shown in the metadata display.

The panel in Figure 10 allows the user to edit metadata sets. Here, the user is looking at the *Subject* element of the DLS set. The values of this element form a hierarchy, and the user is examining, and perhaps changing, the list of values assigned to it. The same panel also allows you to change the "profile" for mapping elements of one metadata set to another. This profile is created when importing documents from collections that have pre-assigned metadata.

Figure 10. Editing the metadata set

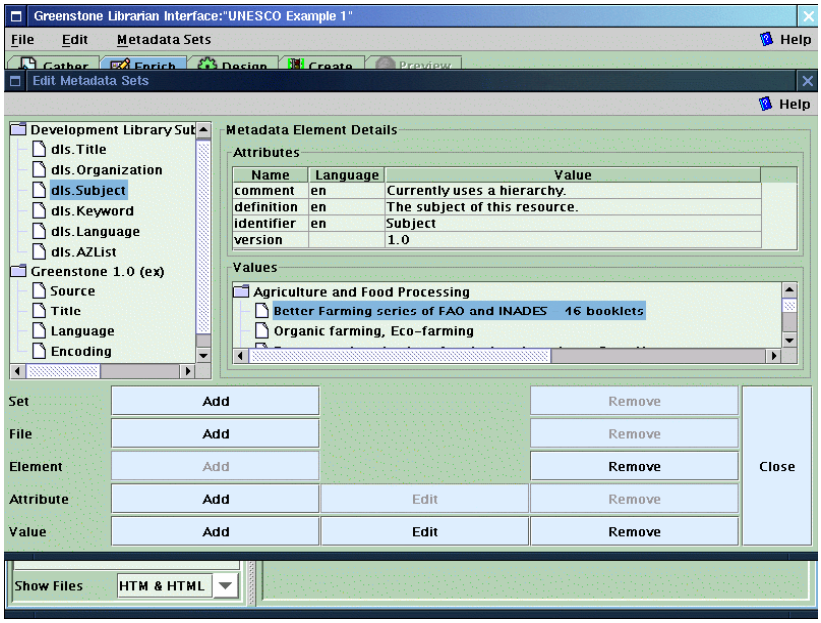


Figure 11. Designing the collection

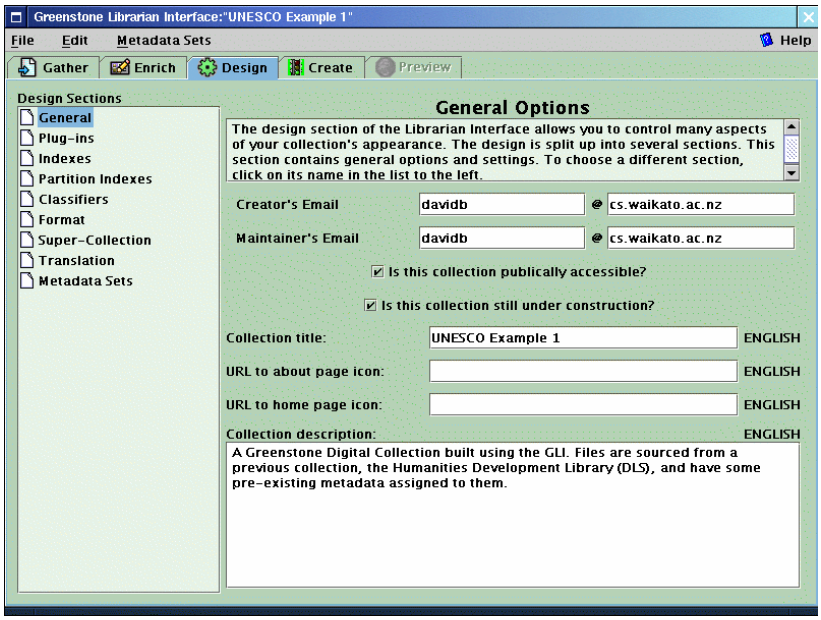


Figure 12. Specifying which plug-ins to use

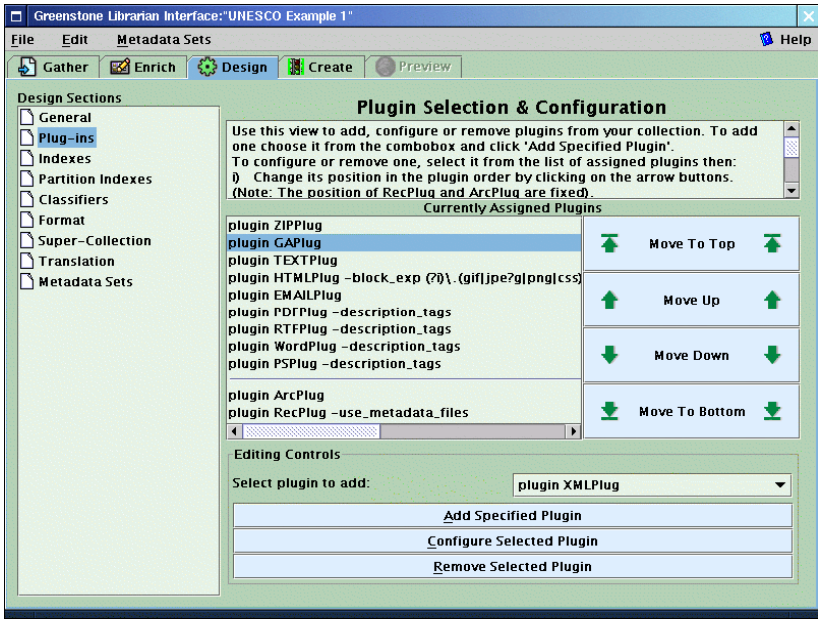
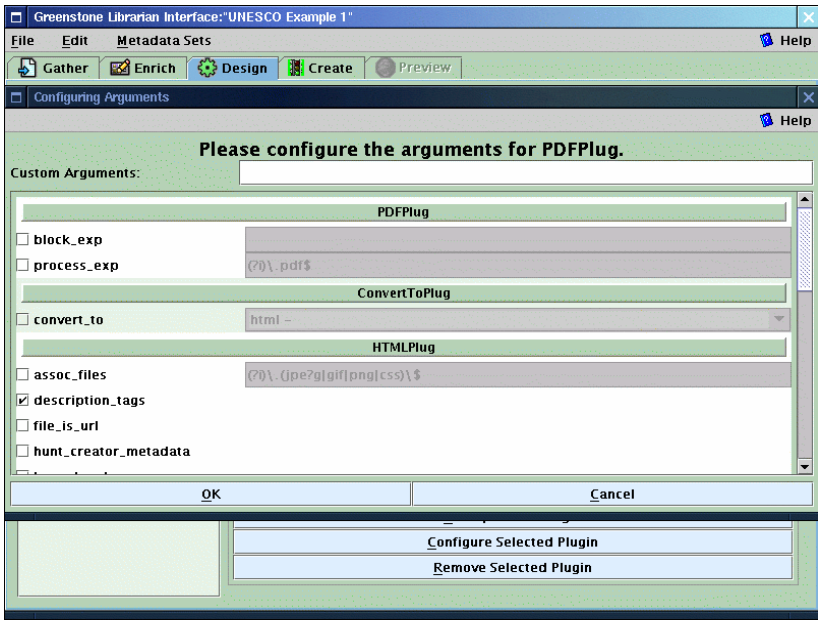


Figure 13. Configuring arguments to a plug-in



Designing the collection

The *Design* panel (Figures 12–13) allows one to specify the structure, organization, and presentation of the collection being created. As noted earlier, the result of this process is recorded in a "collection configuration file," which is Greenstone's way of expressing the facilities that a collection requires. This step involves a series of separate interaction screens, each dealing with one aspect of the collection design. In effect, it serves as a graphical equivalent to the usual process of editing the configuration file manually.

Operations include:

- Reviewing and editing collection-level metadata such as title, author and public availability of the collection.
- Defining what full-text indexes are to be built.
- Creating sub-collections and having indexes built for them.
- Adding or removing support for predefined interface languages.
- Constructing a list of plug-ins to be used, and their arguments.
- Presenting the list to the user for review and modification.
- Configuring individual plug-ins.
- Constructing a list of "classifiers," their arguments, assignment and configuration.
- Assigning formatting strings to various controls within the collection, thus altering its appearance.
- Reviewing the metadata sets, and their elements, used in the collection.

In Figure 11 the user has clicked the *Design* tab and is reviewing the general information about the collection, entered when the new collection was created. On the left are listed the various facets that the user can configure: Document Plug-ins, Search Types, Search Indexes, Partition Indexes, Cross-Collection Search, Browsing Classifiers, Format Features, Translate Text, Metadata Sets. Appearance and functionality varies between these. For example, clicking the *Plug-in* button brings up the screen shown in Figure 12, which allows you to add, remove or configure plug-ins, and change the order in which the plug-ins are applied to documents.

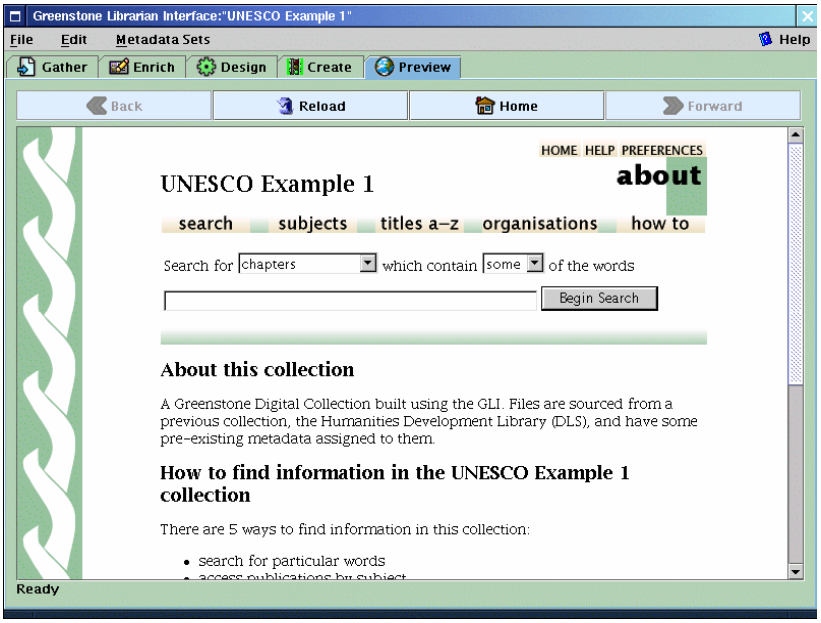
Plug-ins and classifiers have many different arguments or "options" that the user can supply. The dialog box in Figure 13 shows the user specifying arguments to some of the plug-ins. The grayed-out fields become active when the user adds the option by clicking the tick-box beside it. Because Greenstone is a continually growing open-source

system, the number of options tends to increase as developers add new facilities. To help cope with this, Greenstone has a “plug-in information”



Figure 14. Getting ready to create new collection

Figure 15. Previewing the newly built collection



utility program that lists the options available for each plug-in, and the librarian interface automatically invokes this to determine what options to show. This allows the interactive user interface to automatically keep pace with developments in the software.

Building the collection

The *Build* panel (Figure 14) is used to construct a collection based on the documents and assigned metadata. The brunt of this work is borne by the Greenstone code itself. The user controls this external process through a series of separate interaction screens, each dealing with the arguments provided to a certain stage of the creation process.

The user observes the building process though a window that shows not only the text output generated by Greenstone's importing and index-building scripts, but also progress bars that indicate the overall degree of completion of each script.

Figure 14 shows the *Create* view. On the left are shown groups of options that can be applied during the creation process: Import, Build, Message-Log. The user selects appropriate values for the options. This figure illustrates a popup "tool tip" that is available throughout the interface to explain the function of each argument. In this case the user is selecting the verbosity level for the collection creation process, and the tool tip

advises that this is a number between zero and three that controls how much information about the process is printed to the standard error stream—0 gives a little, 3 gives lots.

When satisfied with the arguments, the user clicks *Build Collection*. Greenstone continually prints text that indicates progress, and this is shown along with a more informative progress bar.

Previewing

The *Preview* panel (Figure 15) is used to view the collection that has been built. In practice, previewing often shows up deficiencies in the collection design, or in the individual metadata values, and the user frequently returns to earlier stages to correct these. This tab becomes active once the collection has been created. The newly created collection will also have been installed on your Greenstone home page as one of the regular collections.

Help

On-line help is always available, and is invoked using the *Help* item at the right of the main menu bar at the top of each of the Figures. This opens up a hierarchically structured file of help text, and account is taken of the user's current context to highlight the section that is appropriate to the present stage of the interaction. Furthermore, as noted above, whenever the mouse is held still over any interactive object a small window pops up to give a textual "tool tip," as illustrated near the bottom of Figure 14.

3.2 Librarian Interface user guide

Starting Off

This section covers how to create, save and load a collection.

Creating a New Collection

To create a new collection, open the "File" menu and choose "New". Several fields need to be filled out—but you can change their values later if you need to, in the design view.

"Title" is the text displayed at the top of your collection's home page. It can be any length.

"Short Name" is the collection's filename. It must be unique.

"Author's Email" should be a valid email address.

"Description of content" should describe, in as much detail as possible, what the collection is about. Use the [Enter] key to break it into paragraphs.

Now choose whether to view or hide the metadata that Greenstone extracts automatically from documents. (This option can be changed later—see Section 10.1—Preferences.)

Finally you must specify whether the new collection will have the same appearance and metadata sets as an existing collection, or whether to start a default "New Collection".

Click "OK" to create the collection. If you chose "New Collection" you are prompted for the metadata sets to use in it. You can choose more than one, and you can add others later.

Clicking "Cancel" returns you to the main screen immediately.

Buttons, like menus, have one character underlined. To "click" the button, press [ALT] and the underlined character at the same time.

Saving the Collection

Save your work regularly by opening the "File" menu and choosing "Save". Saving a collection is not the same as making it ready for use in Greenstone (see Section 8—Producing Your Collection).

The Librarian Interface protects your work by saving it whenever you exit the program or load another collection.

Saved collections are written to a file named for the collection and with file extension ".col", located in a folder of the same name within your Greenstone installation's "collect" folder.

Opening an Existing Collection

To open an existing collection, choose "Open" from the "File" menu to get the Open Collection prompt. A list of your Greenstone collections appears. Select one to see its description, and click "Open" to load it. If you seek a collection that resides outside Greenstone's "collect" folder, click "Browse" for a file system browsing dialog.

In case more than one Greenstone Librarian Interface program is running

concurrently, the relevant directories are "locked" to prevent interference. On opening a collection, a small temporary lock file is created in its folder. Before opening a collection, the Librarian Interface checks to ensure that no lock file already exists. You can tell whether a collection is locked by the colour of its icon: green for a normal collection, red for a locked one. However, when the Librarian Interface is exited prematurely the lock file is sometimes left in place. When you open such a collection, the Librarian asks if you want to "steal" control of it. Never steal a collection that someone else is currently working on.

Hunting For Your Files

The Librarian Interface can run in different configurations. This section only applies when the "hunt" and/or "mirror" views are enabled. If these tabs do not appear, advanced users can enable them by editing the "config.xml" file in the Librarian Interface installation folder (or, on a multiuser system, in your home directory's ".gli" folder) to set the values of "workflow.browse" and "workflow.mirror" to "true".

When using web resources, the Librarian Interface operates in two stages. This section describes the first, Hunting, where you browse the Internet for files of interest. The next section describes the second stage, Mirroring, where these files can be downloaded.

The Hunt view

The Hunt view is accessed by clicking on its tab. Most of the screen shows a web page. Hyperlinks work in the usual way. Underneath the page is a status bar that shows you what the browser is doing.

You type URL addresses into the address field above the page.

To the left and right of the address field are web browser buttons for Back, Reload, Home, Go, Stop and Forward.

Downloading Chosen Files

To enable web mirroring see Section 3.0—Hunting For Your Files. To download web pages you need to install the mirroring tool wget (version v1.8 recommended) and make it accessible from the Librarian Interface "install" folder.

Once web resources, the second stage is to download (or "mirror") the files you need. This section explains the Librarian Interface's mirroring process.

The Mirror view

Access the "Mirror" view by clicking its tab. The top half of the screen shows the downloading controls. The bottom half is initially empty, but will show a list of pending downloading jobs if there are some.

Files are downloaded into a folder in the workspace tree called "Public WebCache" (only present when mirroring is enabled), and can be used in all collections built with the Librarian Interface. When a collection is open a second folder, "Private WebCache", appears, which only that collection can access files from. Files in both these areas are named by their full web URL. A new folder is created for each host, followed by others for each part of the path. This ensures that each file is distinct.

Use the first of the download configuration controls, "Source URL", to enter the URL of a target resource. If you have come from the "Hunt" view, this field points to the last page visited. Use the "Download Depth" control to determine how many hyperlinks deep to go when downloading: 0 means follow no hyperlinks and grabs just the target resource. The depth limit is ignored when downloading media other than html pages. Use the "Destination Folder" radio button control to choose whether the files are cached into the shared public folder or the private collection-specific folder (if available). Next, there are four checkbox controls which can be set to turn on the specified feature for a specific download. A fifth checkbox control "Automatically remove failed downloads..." does not pertain to a specific download, but instead clears the list of any failed download information, and prevents any future failures from appearing in the list. The final control is a "Download" button, which adds to the download list a new job corresponding to the configuration settings.

The download list has an entry for each unfinished download. For each entry, a central region with several lines of text and a progress bar is flanked by two buttons. The text gives details of the task, and updates as the task progresses, while the bar measures progress. The "play" triangle on the left button starts the current task and then changes to a double vertical bar, which pauses the current download. The "stop" square on the right button removes the current download task from the list. Download tasks are removed from the list when they finish successfully (ones that fail are controlled by the "Automatically remove failed downloads..." checkbox).

Section 10.1—Preferences describes how to establish an Internet connection via a proxy. If authentication is needed, the proxy server prompts for identification and password. The Librarian Interface does not store passwords between sessions.

Collecting Files for Your Collection

Once you have a new collection you need to get some files into it. These may come from your ordinary file space, or from other Greenstone collections. Some may already have attached metadata. This section describes how to import files.

The Gather View

The Librarian Interface starts with the Gather view. To return to this view later, click the "Gather" tab directly below the menu bar.

The two large areas titled "Workspace" and "Collection" are used to move files into your collection. They contain "file trees", graphical structures that represent files and folders.

Select an item in the tree by clicking it. (There are other ways; see below.) Double-click a folder, or single-click the switch symbol beside it, to expand (or collapse) its contents. Double-click a file to open it using its associated application program (see Section 10.2—File Associations).

The Workspace file tree shows the sources of data available to the Librarian Interface—the local file system (including disk and CD-ROM drives), the contents of existing Greenstone collections, and the public and private download caches if Web mirroring is enabled. You can copy and view these files but you cannot move, delete, or edit them. Navigate this space to find the files you want to include in the collection.

The Collection file tree represents the contents of the collection so far. Initially, it is empty.

You can resize the spaces by mousing over the grey bar that separates the trees (the shape of the pointer changes) and dragging.

Beneath is the Status Area, which describes the state of the Librarian Interface: how many items are selected and what action is requested. It reports on the progress of actions that involve files, which can take some time to complete. The "Stop" button stops any action that is currently in progress.

Two large buttons occupy the lower right corner of the screen. "New Folder", with a picture of a folder, creates new folders (see Section 5.2—Creating folders). "Delete", with a garbage can, removes files. Clicking the Delete button will remove any selected files from the Collection file tree. Alternatively, files can be deleted by dragging them onto the Delete

button.

Creating Folders

Use folders in the Collection file tree to group files together and make them easier to find. Folders can be placed inside folders. There is virtually no limit to how many folders you can have or how deeply they can be nested.

To create a new folder, optionally select an existing folder in the Collection Tree and click the New Folder button. The new folder appears within the selected one, or at the top level if none is selected. You are prompted for the folder's name (default "New Folder").

Folders can also be created by right-clicking over a folder, choosing "New Folder" and proceeding as above.

Adding Files

Files can be copied into the collection by dragging and dropping. The mouse pointer becomes a ghost of the selected item (or, if more than one is selected, the number of them). Drop the selection into the Collection Tree to copy the files there (if the source was the Workspace Tree) or move them around within the collection (if the source was the Collection Tree).

When copying multiple files, they are all placed in the target folder at the same level, irrespective of the folder structure they occupied originally. When you copy a second file with the same name into the same folder, you are asked whether to overwrite the first one. Respond "No" and the file will not be copied, but the others will be. To cancel all remaining copy actions, click the "stop" button.

Only the "highest" items in a selection are moved. A folder is higher than its children. You cannot select files within a folder and also the folder itself.

When you add a file, the Librarian Interface searches through the source folders for auxiliary files containing metadata previously assigned to the added file and, if it finds one, begins to import this metadata. As the operation proceeds, you may be prompted (perhaps several times) for extra information to match the imported metadata to the metadata sets in your collection. This process involves many different prompts, described in the section 6.8—Importing Previously Assigned Metadata. For a more detailed explanation of associating metadata with files read Chapter 2 of

the Greenstone Developer's Guide—Getting the most out of your documents.

Removing Files

There are several methods for removing files and folders. You must first indicate what items to remove by selecting one or more files and folders as described in section 5.1—The Gather View.

Once files have been selected, click the "delete" button to remove them, or press the [Delete] key on your keyboard, or drag them from the collection to the delete button and drop them there.

Filtering the Tree

The "Show Files" pull-down menu underneath each tree shows a list of predefined filters, such as "Images". Choosing this temporarily hides all other files in the tree. To restore the tree, change the filter back to "All Files". These operations do not alter the collection, nor do they affect the folders in the tree.

You can specify a custom filter by typing in a pattern to match files against. Use standard file system abbreviations such as "*.*" or "*.doc" ("*" matches any characters).

Enriching the Collection with Metadata

Having gathered several files into the collection, now enrich them with additional information called "metadata". This section explains how metadata is created, edited, assigned and retrieved, and how to use external metadata sources (also see Chapter 2 of the Greenstone Developer's Guide—Getting the most out of your documents).

The Enrich View

Use the Enrich view to assign metadata to the documents in the collection. Metadata is data about data—typically title, author, creation date, and so on. Each metadata item has two parts: "element" tells what kind of item it is (such as author), and "value" gives the value of that metadata element (such as the author's name).

On the left of the "Enrich" view is the Collection Tree. To the right is the Metadata Table, which shows metadata for any selected files or folders in the Collection Tree. Columns are named in grey at the top, and can be resized by dragging the separating line. Clicking any row transfers its

details to the MetaEdit Controls below. If several files are selected, black text indicates that the value is common to all of the selected files, while grey text indicates that it is not. Black values may be updated or removed, while grey ones can be removed from those that have it, or appended to the others.

A folder icon may appear beside some metadata entries. This indicates that the values are inherited from a parent (or ancestor) folder. Inherited metadata cannot be edited or removed, only appended to or overwritten. Click on the folder icon to go immediately to the folder where the metadata is assigned.

The MetaEdit Controls at the lower right appear only when a file is selected from the tree and a row is selected from the table. Use them to update, append, and remove the metadata value. The value field is for entering or editing the metadata value. Beside it is a button labelled "..." which, when clicked, opens a larger editing box. In the buttons below, "Append" assigns the value as new metadata and adds it to any existing values for the selected element, "Replace" overwrites the selected existing value with the new one, and "Remove" clears the selected value. Underneath, labelled "All Previous Values", is the "Value Tree".

The Value Tree expands and collapses. Usually it is a list that shows all values entered previously for the selected element. Clicking an entry automatically places it into the value field. Conversely, typing in the text field selects the Value Tree entry that starts with the characters you have typed. Pressing [Enter] auto-completes the typing with the selected value.

Metadata values can be organised into a hierarchy. This is shown in the Value Tree using folders for internal levels. Hierarchical values can be entered using the character "\" to separate the levels. For example, "Cards\Red\Diamonds\Seven" might be used in a hierarchy that represents a pack of playing cards. This enables values to be grouped together. Groups can also be assigned as metadata to files.

Greenstone extracts metadata automatically from documents into a metadata set whose elements are prefixed by "ex.". This has no value tree and cannot be edited, so the edit controls are hidden if such an entry is selected. The "..." button still serves to expand the value, but the text cannot be edited.

Selecting Metadata Sets

Sets of predefined metadata elements are known as "metadata sets". An example is the Dublin Core metadata set. When you add a metadata set to

your collection, its elements become available for selection. You can have more than one set; to prevent name clashes a short identifier that identifies the metadata set is pre-pended to the element name. For instance the Dublin Core element Creator becomes "dc.Creator". Metadata sets are stored in the Librarian Interface's metadata folder and have the suffix ".mds".

To add a metadata set, choose "Metadata Sets" from the menu bar and select the "Import Set" action. A list appears that shows the sets stored in the Librarian Interface's metadata folder. Choose one and open it, or click "Browse" to locate metadata set files stored elsewhere. If the metadata elements have associated value trees, you will be asked whether to import all values associated with the elements in the set, just those values that make up the structure of hierarchy-based metadata, or no values at all.

To install a newer version of a metadata set, simply add it as above. The Librarian Interface merges the sets, but does not alter values you have entered. You may be asked how to merge certain elements. For example if the current set and the one you are importing share a common element (which is likely if you are installing a new version of a set), you are shown as much information about the existing and new elements as possible, and asked how to proceed. Options include merging the elements, renaming the new one, replacing the old element entirely, or skipping this element. When merging two elements you are confronted with the same options, but this time at the "attributes" (rather than "elements") level. You can cancel the import operation at any time.

The ".mds" files are expressed in XML format. You can edit an existing metadata set or create a new one with an ordinary text editor. If you are starting a new file, copy the Document Type Definition and be sure to follow it, otherwise the Librarian Interface will be unable to load the metadata set. Use of an XML validator or validating editor is recommended.

To export a metadata set, or part of one (e.g. its assigned value hierarchy, or all its values), return to the Librarian Interface and choose "Export Set" from the "Metadata Set" menu. You will be asked to select appropriate export options, and a file to export into.

Appending New Metadata

We now add a metadata item—both element and value—to a file. First select the file from the Collection file tree on the left. The action causes any metadata previously assigned to this file to appear in the table at the right.

Next select the metadata element you want to add by clicking its row in the table.

Type the value into the value field. Do not use the character "\", as it is used for constructing hierarchies. When finished, click "Append" to add the new value as metadata for the chosen file. The value immediately appears in the Metadata table.

You can also add metadata to a folder, or to several multiply selected files at once. It is added to all files within the folder or selection, and to child folders. Keep in mind that if you assign metadata to a folder, any new files in it automatically inherit the folder's values.

When you add metadata to multiple files at once, you will be prompted for confirmation for any files that already have a value for that metadata. You are shown the name of the file in question, the element's title, previously-assigned values for this element, and the new value. The buttons offer different options: "Append" appends the metadata to the file without altering any existing values; "Append All" adds the new value to all other files too, without requiring individual confirmation; "Skip File" proceeds to the next file; "Cancel" undoes any changes and cancels the action.

If you choose metadata that occurs in some of the selected files and click "Append", it is added to the other files in the selection too.

You can add structure to metadata values by using paths as described in Section 6.1—The Enrich View. Correct any mistakes in creating hierarchies by using the metadata set editor explained in Section 11.1—Editing Metadata Sets.

Adding Previously Defined Metadata

To add metadata that has an existing value, first select the file, then select the required value from the value tree, expanding hierarchy folders as necessary. The value of the selected entry automatically appears in the Value text field (alternatively, use the value tree's auto-select and auto-complete features). Click "Append" to add the metadata to the selected file.

The process of adding metadata with already-existing values to folders or multiple files is just the same.

Updating Metadata

To update the value of a piece of metadata, first choose the file to which that value applies, and then the metadata element whose value you want to change. Your selection appears in the metadata edit controls. Edit the value field and click “Replace” to alter the metadata.

The process is the same when updating a folder with child folders or multiple files, except that when you click “Replace” you are asked what to do with the other files. The buttons offer different options: “Replace” replaces any previous value with the new one; “Replace All” adds the new value in the same way to all other files; “Skip File” skips the current file and proceeds to the next; “Cancel” undoes any changes and cancels the action. You can only update metadata that is common to all files selected. For a folder, this means that all its contents must share the same metadata.

The value tree shows all previous values, not just those currently assigned. Thus the value you have replaced will remain in the value tree.

Removing Metadata

You remove metadata the same way as you update it. First select a file from the file tree, then use the metadata table to select the metadata. If the metadata has a value assigned, the “Remove” button in the MetaEdit Controls becomes active. Click it to remove the metadata from the specified file. Other files remain unchanged, and the value remains in the Value Tree.

When you remove metadata from a folder, or from several files, you are presented with various options: removing the metadata from this file, removing it from this and all other files, and skipping this file. You can cancel the operation at any time. If you choose metadata that is not common to all the selected files and dick “Remove”, the metadata is removed from those files that have it; all others are unaffected.

Reviewing Assigned Metadata

Sometimes you need to see the metadata assigned to many or all files at once—for instance, to determine how many files are left to work on, or to get some idea of the spread of dates.

Select the files you wish to examine, and from the “Metadata Set” menu choose “Assigned Metadata...”. A window called “All Metadata”, dominated by a large table with many columns, appears. The first column

shows file names; the rows show all metadata values assigned to those files.

Drawing the table can take some time if many files are selected. You can continue to use the Librarian Interface while the "All Metadata" window is open.

Click "Close" to hide the window.

You can also view the "All Metadata" table by selecting the files you wish to examine, right-clicking, and choosing "Assigned Metadata...". If a folder has been selected, all its child files are included in the table.

When it gets too large, you can filter the "All Metadata" table by applying filters to the columns. As new filters are added, only those rows that match them remain visible. To set, modify or clear a filter, click on the "funnel" icon at the top of a column. You are prompted for information about the filter. Once a filter is set, the column header changes colour.

The prompt has a "Simple" and an "Advanced" tab. The Simple version filters columns so that they only show rows that contain a certain metadata value ("*" matches all values). You can select metadata values from the pull-down list. The Advanced version allows different matching operations: must start with, does not contain, alphabetically less than and is equal to. The value to be matched can be edited to be any string (including "*"), and you can choose whether the matching should be case insensitive. Finally, you can specify a second matching condition that you can use to specify a range of values (by selecting AND) or alternative values (by selecting OR). Below this area is a box that allows you to change the sort order (ascending or descending). Once you have finished, click "Set Filter" to apply the new filter to the column. Click "Clear Filter" to remove a current filter. Note that the filter details are retained even when the filter is cleared.

For example, to sort the "All Metadata" table, choose a column, select the default filter setting (a Simple filter on "*"), and choose ascending or descending ordering.

Importing Previously Assigned Metadata

If metadata in a form recognized by the Librarian Interface has been previously assigned to a file—for example, when you choose documents from an existing Greenstone collection—it is imported automatically when you add the file. To do this, the metadata must be mapped to the metadata sets available in the collection.

The Librarian Interface prompts for the necessary information. The prompt gives brief instructions and then shows the name of the metadata element that is being imported, just as it appears in the source file. This field cannot be edited or changed. Next you choose what metadata set the new element should map to, and then the appropriate metadata element in that set. The system automatically selects the closest match, in terms of set and element, for the new metadata.

Having checked the mapping, you can choose "Add" to add the new metadata element to the chosen metadata set. (This is only enabled if there is no element of the same name within the chosen set.) "Merge" maps the new element to the one chosen by the user. Finally, "Ignore" does not import any metadata with this element name.

Once you have specified how to import a certain piece of metadata, the mapping information is retained for the collection's lifetime. To correct any mistakes during importing, use the metadata set editor described in Section 11—Editing Metadata Sets.

For details on the metadata.xml files which Greenstone uses to store the metadata, see Chapter 2 of the Greenstone Developer's Guide—Getting the most out of your documents.

Designing Your Collection's Appearance

Once your files are marked up with metadata, you next decide how it should appear to users as a Greenstone collection. What kind of information is searchable? What ways are provided to browse through the documents? What languages are supported? Where do the buttons appear on the page? These things can be customized; this section describes how to do it.

The Design View

With the Librarian Interface, you can configure how the collection appears to the user. The configuration options are divided into different sections, each associated with a particular stage of navigating or presenting information.

On the left is a list of different views, and on the right are the controls associated with the current one. To change to a different view, click its name in the list.

To understand the stages and terms involved in designing a collection, first read Chapters 1 & 2 of the Greenstone Developer's Guide.

General Settings

This section explains how to review and alter the general settings associated with your collection. First, under "Design Sections", click "General".

Here the values provided during collection creation can be modified.

At the top of the page is an instruction box, which appears for each of the different sections. It contains a brief list of instructions to remind you what functionality is available.

First are the contact emails of the collection's creator and maintainer. Then come two checkboxes for whether the collection should be publicly accessible, and whether it is still under construction. The following field allows you to change the collection title. The next one specifies (in the form of a URL) the icon to show at the top left of the collection's "About" page, and the next is the icon used in the Greenstone library page to link to the collection. Finally comes the "Collection Description" text area as described in section 2.1—Creating A New Collection.

Document Plugins

This section describes how to configure the document plugins the collection uses. It explains how you specify what plugins to use, what parameters to pass to them, and in what order they occur. Under "Design Sections", click "Document Plugins".

To add a plugin, select it using the "Select plugin to add" pull-down list near the bottom and then click "Add Plugin". A window appears entitled "Configuring Arguments"; it is described later. Once you have configured the new plugin, it is added to the end of the "Currently Assigned Plugins" list. Note that a plugin may only occur once in the list.

To remove a plugin, select it in the list and click "Remove Plugin".

Plugins are configured by providing arguments. To alter them, select the plugin from the list and click "Configure Plugin" (or double-click the plugin). A "Configuring Arguments" dialog appears with three parts: a text field for entering custom arguments, an area containing controls for specifying arguments, and two buttons at the bottom.

There are different kinds of controls. Some are checkboxes, and clicking one adds the appropriate option to the plugin. Others are text strings, with a checkbox and a text field. Click the box to enable the argument, then

type appropriate text (regular expression, file path etc) in the box. Others are pull-down menus from which you can select from a given set of values. Still others allow multiple selections from a list. To add a value, select it and click "Add"; to remove it, select it and click "Remove". To learn what an argument does, let the mouse hover over its name for a moment and a description will appear.

When you have changed the configuration, click "OK" to commit the changes and close the dialog, or "Cancel" to close the dialog without changing any plugin arguments.

The plugins in the list are executed in order, and the ordering is sometimes important. Two plugins, ArcPlug and RecPlug, are vital to the collection building process, and are fixed in place at the end of the list (with a separator line). To change the ordering of the other ones, select the plugin you want to move and click "Move To Top", "Move Up", "Move Down", or "Move To Bottom".

Search Types

This section explains how to modify a new design feature in Greenstone, Search Types, which allow fielded searching. Under "Design Sections", click "Search Types".

When you enter the Search Types view, first check "Enable Advanced Searches", which activates the other controls. This migrates the collection to Greenstone 2.4 format, which supports fielded searching, and means that (a) the index design is different (explained in section 7.5—Search Indexes), (b) there are more text fragments to translate (see Section 7.10—Translation), and (c) the collection will not be usable under older Greenstone installations. If you later uncheck this field, most of your collection will be migrated back to Greenstone 2.39. However the Librarian Interface cannot convert the new index specifications into older ones, so you will have to re-enter them manually.

To add a search type, select it from the "Search Types" list and click "Add Search Type". Each type can only appear in the list once.

To remove a search type, select it from the "Currently Assigned Search Types" list and click "Remove Search Type". The list must contain at least one search type.

To change to order of a search type, select it from the list and click "Move Up" or "Move Down". The first one will be the default.

Search Indexes

Indexes specify what parts of the collection are searchable. This section explains how to add and remove indexes, and set a default index. Under "Design Sections", click "Search Indexes".

To add an index, type a name for it into the "Index Name" field. Select which of the possible information sources to index by clicking the checkboxes beside them. The list shows all the assigned metadata elements, as well the full text. Having selected the data sources, choose the granularity of the index, using the "At the level" menu. Once these details are complete, "Add Index" becomes active (unless there is an existing index with the same settings). Click it to add the new index.

To remove an index, select it from the list of assigned indexes and click "Remove Index".

The default index, the one used on the collection's search page, is tagged with "[Default Index]" in the "Assigned Indexes" list. To set it, select an index from the list and click "Set Default". To reset it, click "Clear Default".

If advanced searching is enabled (via the Search Types view), the index controls are different. Each index is based on just one data source. There is a new pseudo-data source "allfields" which provides searching across all specified indexes at once. Levels are not assigned to a specific index, but apply across all indexes: thus indexes and levels are added separately. Indexes are removed in the same way as above, but the default index can no longer be set—it is simply the first index assigned.

To create indexes on all sources, click the "Add All" button. The name of each index will default to the source name. To change the name, select an index, change its details, and click "Replace Index".

Partition Indexes

Indexes are built on particular text or metadata sources. The search space can be further controlled by partitioning the index, either by language or by a predetermined filter. This section describes how to do this. Under "Design Sections", click "Partition Indexes".

The "Partition Indexes" view has three tabs; "Define Filters", "Assign Partitions" and "Assign Languages". To learn more about partitions read about subcollections and subindexes in Chapter 2 of the Greenstone Developer's Guide.

Define Filters

Filters allows you to group together into a subcollection all documents in an index for which a metadata value matches a given pattern.

To create a filter, click the "Define Filters" tab and enter a name for the new filter into the "Name the subcollection filter" field. Next choose a document attribute to match against, either a metadata element or the name of the file in question. Enter a regular expression to use during the matching. You can toggle between "Including" documents that match the filter, or "Excluding" them. Finally, you can specify any of the standard PERL regular expression flags to use when matching (e.g. "i" for case-insensitive matching). Finally, click "Add Filter" to add the filter to the "Defined Subcollection Filters" list.

To remove a filter, select it from the list and click "Remove Filter".

To alter a filter, select it from the list, change any of the values that appear in the editing controls and click "Replace Filter" to commit the changes.

Assign Partitions

Having defined a subcollection filter, use the "Assign Partitions" tab to build indexes for it (or for a group of filters). Select the desired filter (or filters) from the "Defined Subcollection Filters" list, enter a name for your partition in the "Partition Name" field, and click "Add Partition".

To remove a partition, select it from the list and click "Remove Partition".

To make a partition the default one, select it from the list and click "Set Default".

To clear the default partition, click "Clear Default".

Assign Languages

This section details how to restrict search indexes to particular languages. You do this by generating a partition using the "Assign Languages" tab of the "Partition Indexes" view.

To add a new language to partition by, use the "Assign Languages" tab to build an index for it. Select the desired language from the "Language to add" pull-down list and click "Add Language".

To remove a language, select it from the "Language Selection" list and click "Remove Language".

To set the default language, select it from the list and click "Set Default".

To clear the default language, click "Clear Default".

Cross-Collection Searching

Greenstone can search across several different collections as though they were one. This is done by creating a "super-collection" that comprises the individual collections. Under "Design Sections", click "Cross-Collection Search".

The Cross-Collection Search view shows a checklist of available collections. The current collection is ticked and cannot be deselected. To add another collection to be searched in parallel, click it in the list (click again to remove it). If only one collection is selected, there is no cross-collection searching.

For further details, see Chapter 1 of the Greenstone Developer's Guide.

Classifiers

This section explains how to assign "classifiers", which are used for browsing, to the collection. Under "Design Sections", click "Browsing Classifiers".

To add a classifier, select it using the "Select classifier to add" pull-down list near the bottom and then click "Add Specified Classifier". A window appears entitled "Configuring Arguments"; instructions for this dialog are just the same as for plugins (see Section 7.3—Plugins). Once you have configured the new classifier, it is added to the end of the "Currently Assigned Classifiers" list.

To remove a classifier, select it from the list and click "Remove Selected Classifier".

To change the arguments a classifier, select it from the list and click "Configure Selected Classifier" (or double-click on the classifier in the list).

The ordering of classifiers in the collection's navigation bar is reflected in their order here. To change it, select the classifier you want to move and click "Move To Top", "Move Up", "Move Down", or "Move To Bottom".

For further information on classifiers read Chapter 2, Greenstone

Developer's Guide—Getting the most out of your documents.

The CustomAZList classifier is a special classifier that builds an alphabetical selection list ("AZList") and allows you to specify the letter ranges. This classifier has its own configuration dialogue. When a metadata element is selected, the "Ranges" tree automatically becomes populated with appropriate values. Expand or collapse the tree as desired. Select any two values and click "Merge" to specify a range, or select a previously merged value and click "Split" to restore the values contained within. When satisfied with the ranges, click "OK" to begin processing the documents in the collection. You can "Cancel" the dialog without making any changes to the collection.

Format Features

Format commands control the structure and appearance of the collection. They affect such things as where buttons appear when a document is shown, and what links are displayed by the DateList classifier. Format commands are not easy to develop, and you should read Chapter 2 of the Greenstone Developer's Guide. This section discusses the format settings, and how the Librarian Interface gives access to them. Under "Design Sections", click "Format Features".

You can apply a format command to anything in the "Choose Feature" pull-down list, which includes each classifier and a predefined list of features. When you select a feature, there are two types of control. Some features are simply enabled or disabled, and this is controlled by a checkbox. Others require a format string to be specified. For these there is a pull-down list ("Affected Component") for selecting which part of the feature the string applies to (if necessary), a text area ("HTML Format String") for entering the string, and a selection of predefined "Variables". To insert a variable into the current position in the format string, select it from the pull-down list and click "Insert".

You can specify a default format for a particular component by selecting the blank feature. This format is then applied to all applicable features unless otherwise specified.

To add a new format command, fill out the information as explained above and click "Add Format". The new format command appears in the list of "Currently Assigned Format Commands". Only one format command can be assigned to each feature/component combination.

To remove a format command, select it from the list and click "Remove Format".

To change a format command, select it from the list, modify the settings, and click “Replace Format”.

For more information about variables and the feature components, read Chapter 2 of the Greenstone Developer’s Guide.

If the “Allow Extended Options” checkbox is ticked, some advanced formatting options are enabled. The list of features that can be formatted is changed slightly, and more variables are available to be used in the format command, providing greater control over the page layout.

Translate Text

This section describes the translation view, where you can define language-specific text fragments for parts of the collection’s interface. Under “Design Sections”, click “Translate Text”.

First choose the an entry from the “Features” list. The language-specific strings associated with this feature appear below. Use the “Language of translation” pull-down list to select the target language, and type the translated text into the text area, referring to the “Initial Text Fragment” if necessary. Click “Add Translation” when finished.

To remove an existing translation, select it in the “Assigned Translations” table and click “Remove Translation”.

To edit a translation, select it, edit it in the “Translated Text” text area, and click “Replace Translation”.

Metadata Sets

This section explains the metadata set review panel. Under “Design Sections”, click “Metadata Sets”.

This view is used to review the metadata sets that the collection uses, and the elements that are available within each set. Choose from the list of “Available Metadata Sets” in order to see details of their elements. This view is read-only.

Producing Your Collection

Having collected the documents for the collection, annotated them with metadata, and designed how the collection will appear, you can now produce the collection using Greenstone. This section explains how.

The Create View

This section explains the Create view used to produce a collection.

Import and Build Settings

This section explains how to access the various import and build settings. For more information of importing and building read Chapter 1 of the Greenstone Developer's Guide—Understanding the collection-building process.

The first two entries in the list on the left are "Import" and "Build", which give settings that apply to the import and build scripts respectively.

Controlling the various settings is done in a similar way to the "Configuring Arguments" window described under Section 7.3—Plugins. Some fields require numeric arguments, and you can either type these in or use the up and down arrows to increase or decrease the current value (in some cases, the interface restricts the range you can enter). Others are enabled by clicking a checkbox (click again to disable).

Message Log

The third item on the left is "Message Log". This shows the output that Greenstone generated when it built the collection before. Select the log you want by clicking on the desired date in the "Log History" list.

The Progress View

When you start to build a collection, the view changes immediately. The controls described in the first section of this section are replaced by two progress bars and a text area. The bars indicate progress through the import phase, then the build phase. The text area shows the Message Log mentioned in the previous section.

Previewing the Collection

This section explains how to use the "Preview" view to inspect the collection you have produced.

The Preview View

Once you have built a collection the "Preview" tab on the main screen becomes enabled. This allows you to inspect the new collection. It shows a simplified Web browser with initial page (and home page) set to the

new collection's "About" page. You can navigate the collection using standard hyperlink clicks.

Miscellaneous

This section describes features of the Librarian Interface that are not associated with any particular view.

Preferences

There are three "General" options. If "View Extracted Metadata" is checked, the various controls dealing with metadata always show all metadata that has been extracted automatically from documents. Deselecting it hides this metadata (although it is still available during collection design, and within the final Greenstone collection).

If "Show File Size" is checked, the file size is shown next to each file in the Workspace and Collection file trees in the Gather and Enrich views.

The third "General" option is a pull-down list of the various languages that the Librarian Interface can be presented in. These correspond to the dictionaries located in the "classes" folder of the Librarian Interface's directory. If you change the dictionary by choosing one from the list, you must restart the Librarian Interface in order to load the new language strings from the dictionary.

The Librarian Interface can support different workflows by determining which of the various view tabs are visible. Use the "Workflow" tab to customise what views are available by checking the boxes next to the views that you want to be available. Alternatively, use the pull-down list at the bottom to select predetermined configurations. Closing the preferences dialog establishes these workflow settings. These settings are stored with the collection, not in the Librarian Interface configuration file.

The "Connection" tab lets you alter the path to the locally-running Greenstone library server, which is used when Previewing collections. It also lets you set proxy information for connecting to the Internet (e.g. when Browsing or Mirroring your files; see 3.0 and 4.0 for details). Check the box to enable proxy connection and supply details of the proxy host address and port number. The proxy connection is established when you close the Preferences dialog.

During the course of a session the Librarian Interface may give warning messages which inform you of possibly unforeseen consequences of an action. You can disable the messages by checking the "Do not show this

warning again" box. You can re-enable warning messages using the "Warnings" tab. Check the box next to warning messages you want to see again.

File Associations

The Librarian Interface uses particular application programs to open particular file types. This section explains how to assign and edit these file associations.

To alter file associations open the "File" menu and click "File Associations...".

To add an association, select the target file extension from the pull-down list, or type in a new extension (do not include the "."). Next either type command that launches the desired application in the appropriate field, or choose the application from the "Browse" dialog. "%1" can be used in the launch command to insert the name of the file being opened. Once these are filled out, "Add Association" is enabled and can be clicked to add the association.

To edit an association, select an existing file extension. Any existing associated command is shown in the launch command field. Edit it, and then click "Replace Association".

To remove an association, select an existing file extension and click "Remove Association". (The file extension remains in the "For Files Ending" pull-down list.)

File associations are stored in the Librarian Interface's main folder, in a file called "associations.xml".

Metadata Set and Profile Editing

This section explains how to edit metadata sets used by the Librarian Interface. This is the only way to remove a value from the "Previous Values" tree. Although you can use the Enrich view to remove a certain value from a record, the value remains in the value tree. To remove it (or any part of the metadata set, including its elements), use the metadata set editor.

The same tool is used to alter the instructions that map metadata from files imported into the collection to existing metadata sets. These are called "importing profiles".

To edit a metadata set or importing profile, choose “Metadata Sets” from the menu bar and select the “Edit Set” action.

Editing Metadata Sets

On the left of the “Edit Metadata Sets” dialog is a list showing what metadata sets and profiles can be edited. Click one of these and its details will appear on the right in one or more tables. Beneath are buttons for adding, editing or removing the various parts; alongside them is the “Close” button. Many buttons are greyed out initially, and are activated by selections in the tables. Now we describe how to edit sets and elements.

To define a new metadata set, beside “Set” click “Add”, fill out the information requested, and click “OK”. “Namespace” is a short identifier for the new set (e.g. “dc” for Dublin Core; “dls” for the Development Library Subset).

To remove a metadata set, select it in the list on the left and click “Remove”. A confirmation prompt will appear; confirming it permanently removes the set and all associated metadata.

Some information is associated with each metadata set, such as its creator and creation date. We call these “attributes” of the metadata set, and you can alter them. Beside “Attribute”, click “Add” to add an attribute to the selected metadata set, fill in the requested information—name, language and values—and click “OK”. Each metadata set is considered unique, so for a new metadata set the pull-down list for the name and value are initially empty. Beside “Attribute”, “Edit” becomes active when an attribute is selected in the table and leads to the same dialog as “Add” (except that the current value is already filled out). Beside “Attribute”, “Remove” becomes active when the attribute is selected; when clicked the attribute is removed.

Double clicking on a set in the list on the left will display a list of metadata elements in that set. You can add an element, remove it, and remove values from it. To add an element, beside “Element” click “Add” and specify the new element’s name.

To remove a metadata element, select it and beside “Element” click “Remove”. This permanently removes the element and all metadata associated with it.

Just as information is associated with each metadata set, information can also be associated with each metadata element—metadata about

metadata! Again we call these "attributes"; Examples are a language-specific name for the element, its definition, or perhaps a general comment.

You edit the attributes of an element in the same way that you edit the attributes of a metadata set, explained above. In this case the pull-down lists in the add and edit prompts may contain values from the same attribute of other elements within the set.

You can also alter the "value tree" for an element, which contains all the values that have been assigned to it. You can "Add" a value whenever an element is selected. Choose a parent folder (if any), enter the value and click "OK" to put the new value in the tree. You can "Edit" a value that you have selected in the value tree; click "OK" to commit the changes. Note that changing the parent subject will cause the value to be moved to that subject. You can "Remove" a value that you have selected in the tree—but note that this does not remove all metadata referring to this value, and if the value is still in use it will be restored the next time you save.

Once you have finished changing the metadata set, click "Close".

Editing Metadata Import Profiles

Double-click the "Importing Profiles" item to see a list of importing profiles for importing from other collections into this one. Each profile is named after the collection to which it applies—that is, the collection that documents are coming from. You can add a profile by clicking "Add" beside "Profile" and specifying the name of the collection that it should apply to. You can remove a profile by selecting it and click "Remove" beside "Profile".

When a profile is selected, its mapping table appears. Each line gives a correspondence between a metadata element in the collection that the metadata comes from, and a metadata element in the collection being constructed. You can edit this table. To add a new mapping, select a source profile, then click "Add" beside "Attribute". The standard attribute dialog box appears, except that the language field is disabled and the "Values" pull-down list contains all the elements currently available in the collection. Mappings can be edited and removed as described above.

Once you have finished changing the metadata importing profile, click "Close".

3.3 Tagging document files

Source documents often need to be structured into sections and subsections, and this information needs to be communicated to Greenstone so that it can preserve the hierarchical structure. Also, metadata - typically the title - might be associated with each section and subsection.

The source documents from an OCR process are typically a set of word processor files, including images. If these are represented as Microsoft Word files, they can be input into Greenstone using the Word plugin. Alternatively, they can be converted to HTML and input using the HTML plugin.

In either case, the hierarchical structure of a document may be indicated by inserting tags in the text as follows:

```
<!--
<Section>
  <Description>
    <Metadata name="Title">Realizing human rights for poor
       people: Strategies for achieving the international
       development targets</Metadata>
  </Description>
-->
(text of section goes here)
<!--
</Section>
-->
```

The <!-- ... --> markers are used because they indicate comments in HTML; thus these section tags will not affect document formatting. You must include these markers around your section tags, even if the document you are working with is not HTML (e.g. if it's a Microsoft Word file).

In the Description part (between the <Description> and </Description> tags) other kinds of metadata can be specified, but this is not done for the style of collections we are describing here.

It is important to remember that you are creating a hierarchical table of contents when you insert section tags into your document. This means that sections can be nested within other sections. In fact, all sections must be nested within a single enclosing section that encompasses the entire document.

The following example demonstrates a document with two chapters, the

second of which contains two subsections. For real examples of sourcedocuments tagged in this way, look at the source documents for the Demo or DLS collections.

```
<!--
<Section>
  <Description>
    <Metadata name="Title">My Document</Metadata>
  </Description>
  <Section>
    <Description>
      <Metadata name="Title">Chapter 1</Metadata>
    </Description>
-->
(text of chapter 1 goes here)
<!--
  </Section>
  <Section>
    <Description>
      <Metadata name="Title">Chapter 2</Metadata>
    </Description>
    <Section>
      <Description>
        <Metadata name="Title">Subsection 1</Metadata>
      </Description>
-->
(text of sub-section 1 goes here)
<!--
    </Section>
    <Section>
      <Description>
        <Metadata name="Title">Subsection 2</Metadata>
      </Description>
-->
(text of sub-section 2 goes here)
<!--
    </Section>
  </Section>
</Section>
-->
```

Note that metadata assigned from within a section tag in a source document takes precedence over that assigned to the document as a whole. This means that you should not explicitly specify Title metadata for the top-level section within a source document unless you want it to override the title you gave it when specifying metadata. In the above example, unless you want to override the document's existing title you should omit the line that reads:

```
<Metadata name="Title">My Document</Metadata>
```

3.4 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. As an alternative to using the Collector, you can also build collections from the command line—the first few pages of the Developer's Guide give a detailed walk-through of how to do this. The Collector predates the librarian interface described in Section 3.1, and for most practical purposes the librarian interface should be used instead of the Collector.

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.

To access the Collector, click the appropriate link on the digital library home page.

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

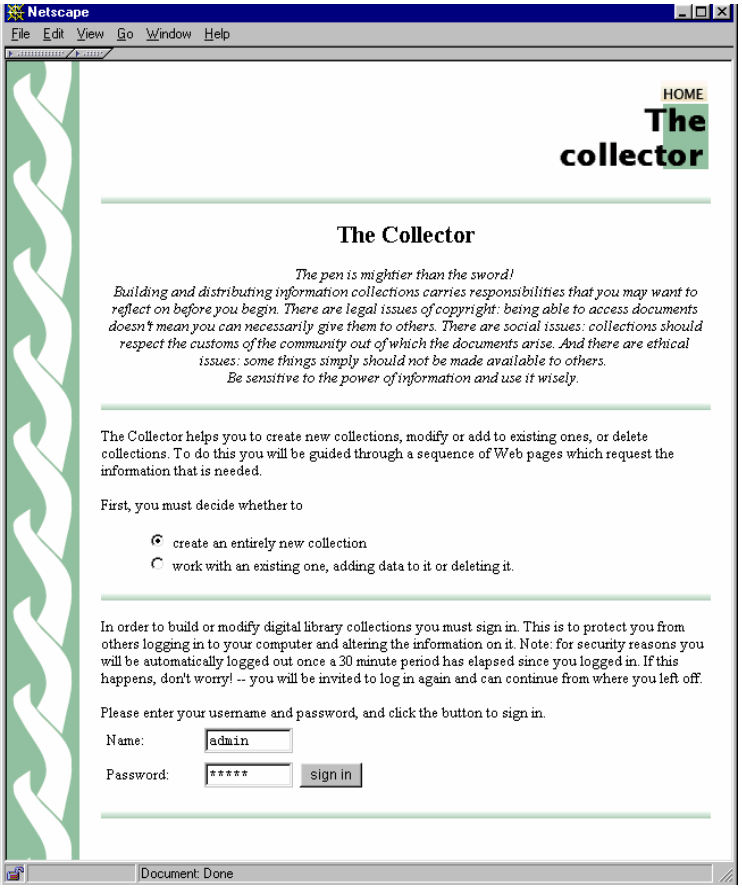
The Collector has the following basic functions:

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 16 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

Figure 16 Using the Collector to build a new collection (continued on next pages)

(a)

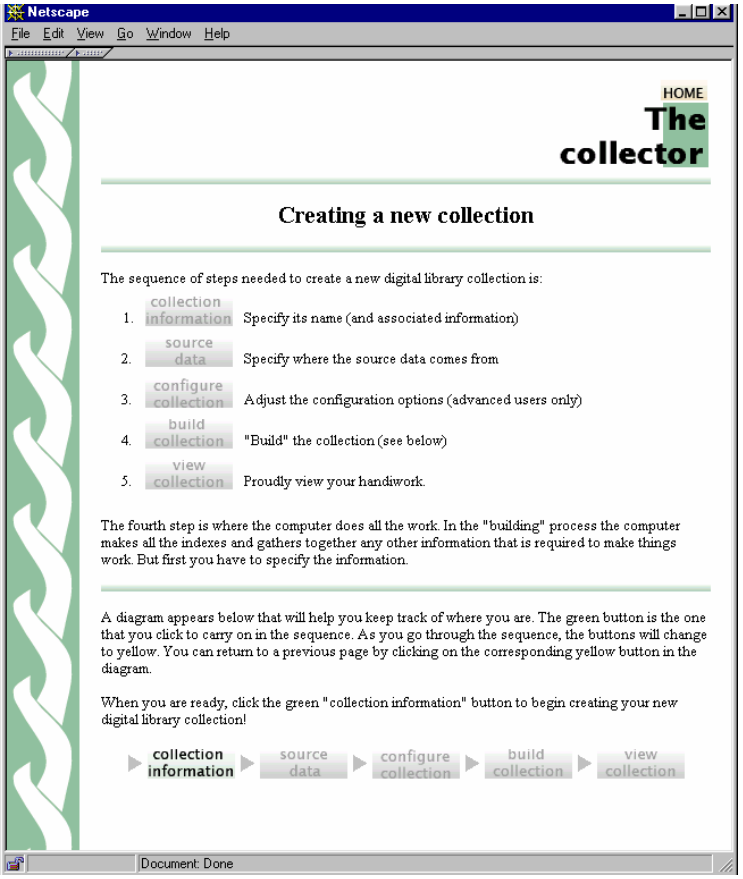


whether to work with an existing collection or build a new one. The former case covers options 1 and 2 above; the latter covers options 3–6. In Figure 16a, the user opts to create a new collection.

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 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

Figure 16 (Continued) (b)



in because other people who use the Greenstone system on your computer should not be allowed to build collections without prior permission.

Dialog structure

Upon completion of login, the page in Figure 16b 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, a step that becomes more

useful as you gain experience with Greenstone. The fourth step is where all the (computer's) work is done. 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 view 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 16b, 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 16b). The gray buttons (all the others, in Figure 16b) 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.

Collection information

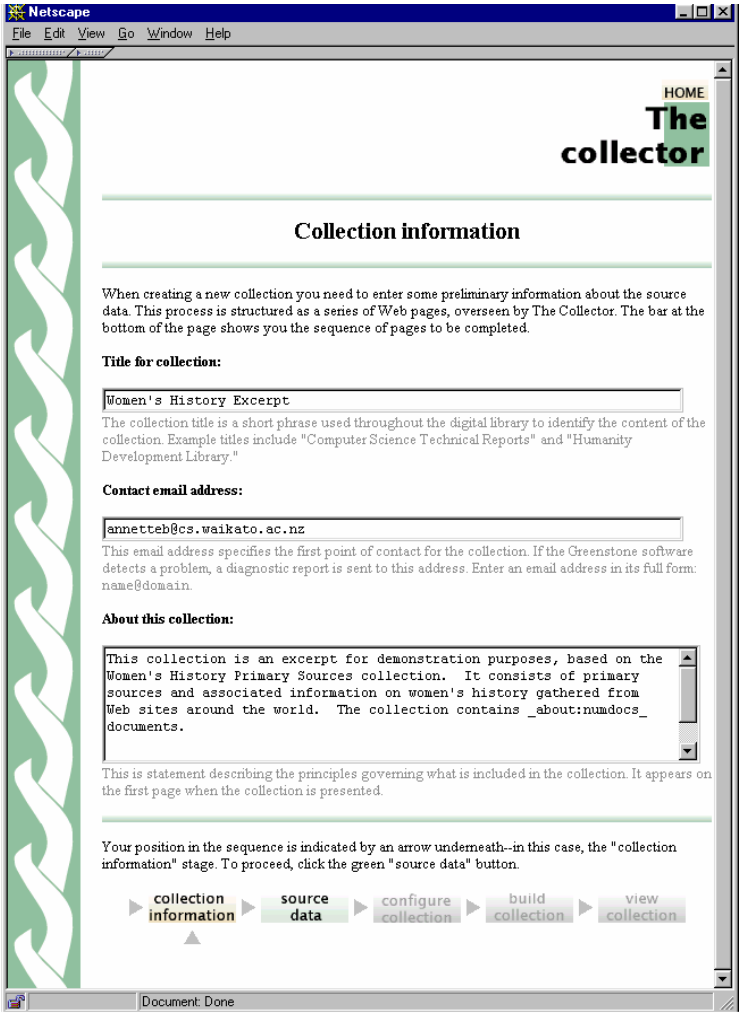
The next step in the sequence, collection information, is shown in Figure 16c. When creating a new collection, it is necessary to enter some information about it:

- 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*, *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 may be 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 16c shows, the collection information stage. The user proceeds to Figure 16d by clicking the green source data button.

Figure 16 (Continued) (c)

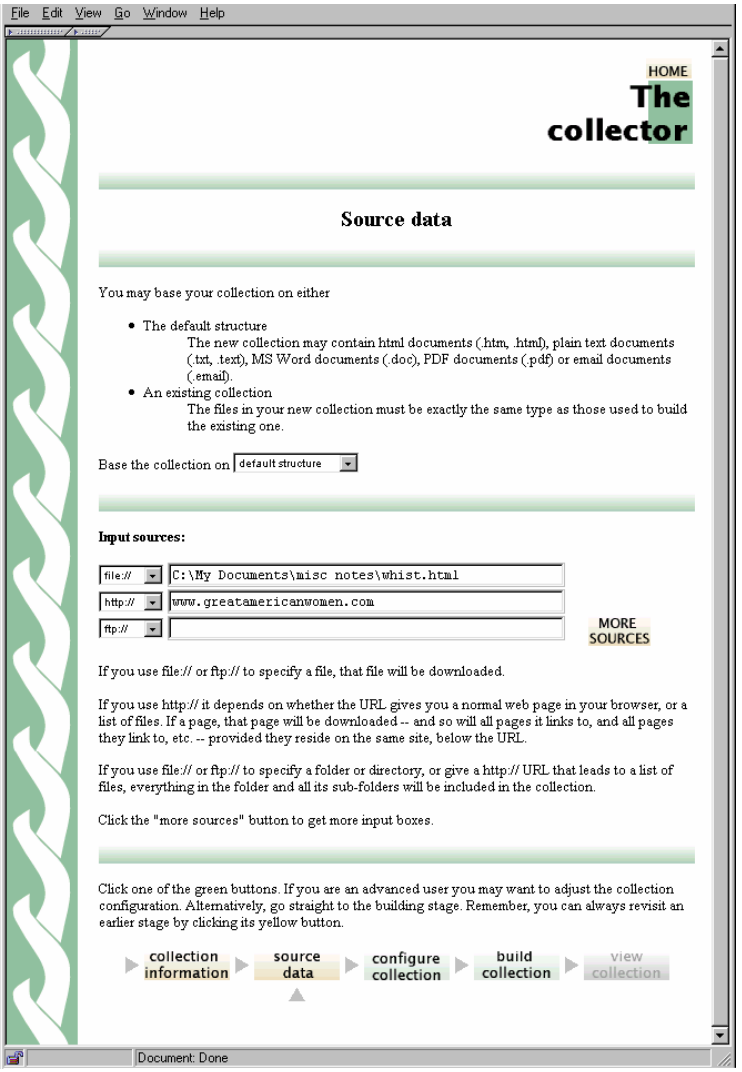


Source data

Figure 16d is the point where the user specifies the source text that comprises the collection. You may either base your collection on a default structure that is provided, or on the structure of an existing collection.

If you opt for the default structure, the new collection may contain HTML documents (files ending in *.htm, .html*), or plain text documents (files ending in *.txt, .text*), Microsoft Word documents (files ending in *.doc*), PDF documents (files ending in *.pdf*) or E-mail documents (files ending in *.email*). More information about the different document formats that

Figure 16 (Continued) (d)



can be accommodated is given in the section on "Document formats" below.

If you base your new collection on an existing one, the files in the new collection must be exactly the same type as those used to build the existing one. 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 example, if you clone the Demo collection you may find that the *subjects*, *organization*, and *how to* buttons don't work.

Boxes are provided to indicate where the source documents are located: up to three separate input sources can be specified in Figure 16d. If you need more, just click the button marked "more sources."

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 anonymous FTP.

If you use *file://* or *ftp://* to specify a file, that file will be downloaded.

If you use *http://* it depends on whether the URL gives you a normal web page in your browser, or a list of files. If a page, that page will be downloaded—and so will all pages it links to, and all pages they link to, etc.—provided they reside on the same site, below the URL.

If you use *file://* or *ftp://* to specify a folder or directory, or give a *http://* URL that leads to a list of files, everything in the folder and all its subfolders will be included in the collection.

You can specify sources of more than one type.

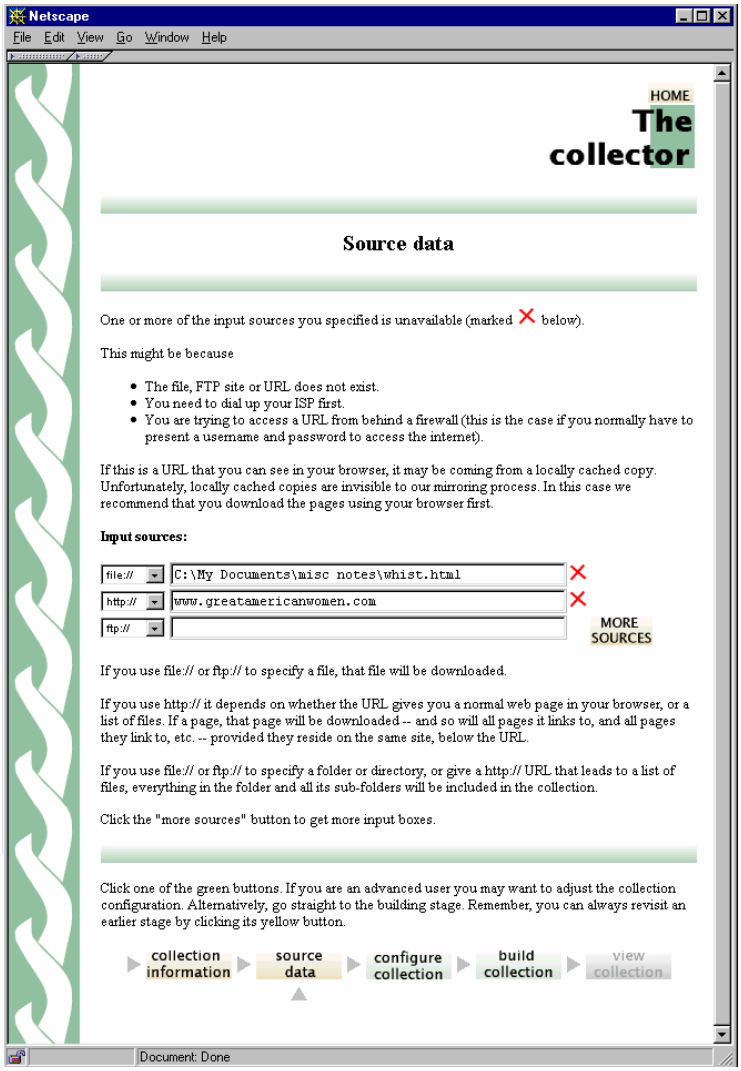
In this case (Figure 16d) the new collection will contain documents taken from a local file system as well as a remote web site, which will be mirrored during the building process.

When you click the *configure collection* button to proceed to the next stage of building, the Collector checks that all the sources of input you specified can be reached. This might take a few seconds, or even a few minutes if you have specified several sources. If one or more of the input sources you specified is unavailable, you will be presented with a page like that in Figure 16e, where the unavailable sources are marked (both of them in this case).

Sources might be unavailable because

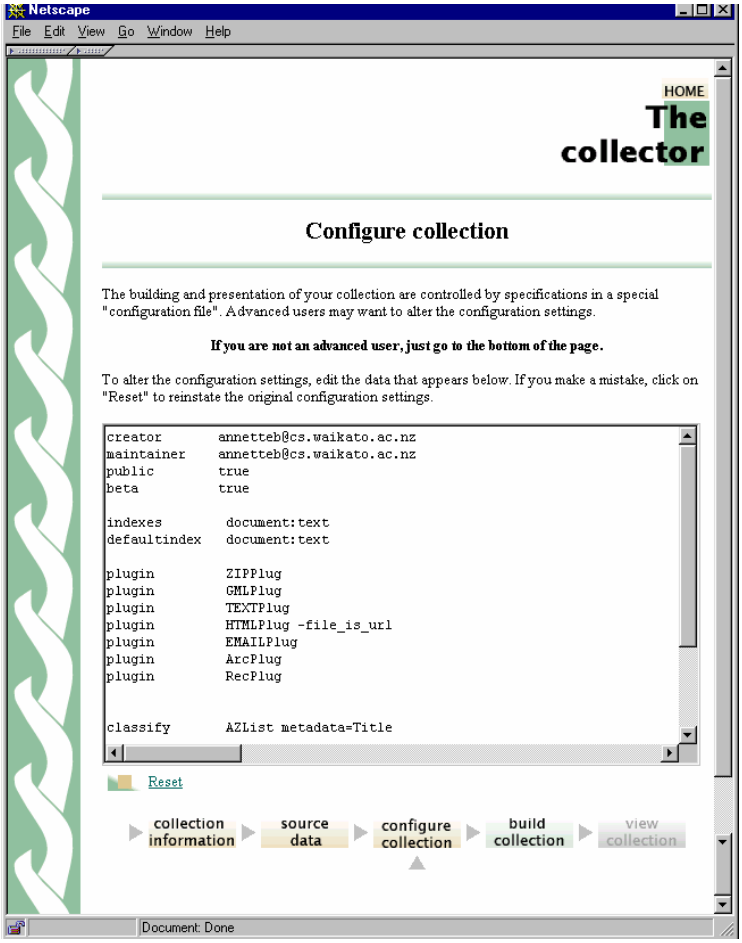
- the file, FTP site or URL does not exist;
- you need to dial up your ISP first;
- you are trying to access a URL from behind a firewall.

Figure 16 (Continued) (e)



The last case is potentially the most mysterious. It occurs if you normally have to present a username and password to access the Internet Sometimes it happens that you can see the page from your Web browser if you enter the URL, but the Collector claims that it is unavailable. The explanation is that the page in your browser may be coming from a locally cached copy. Unfortunately, locally cached copies are invisible to the Collector. In this case we recommend that you download the pages using your browser first.

Figure 16 (Continued) (f)

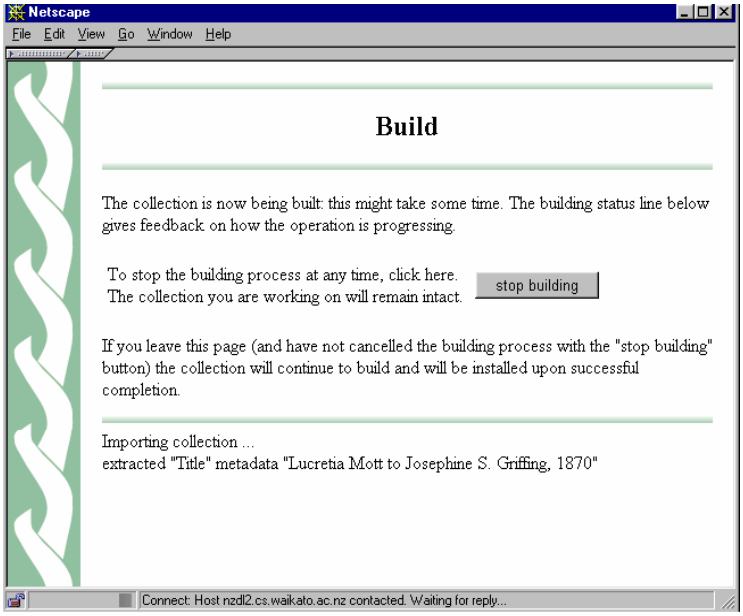


Configuring the collection

Figure 16f 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. Indeed, in Figure 16d both the *configure collection* and the *build collection* buttons are displayed in green, signifying that step 3 can be bypassed completely.

In our example 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

Figure 16 (Continued) (g)



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.

Building the collection

Figure 16g 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 16f 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 16g. 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.

Viewing the collection

When the collection is built and installed, the sequence of buttons visible at the bottom of Figures 16b–f appears at the bottom of Figure 16g, 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 (see the *Greenstone Digital Library Developer's Guide*, Section 4).

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
- Export the collection to CD-ROM.

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.

Delete the collection

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

Export the collection

You can export the collection in a form that allows it to be written to a self-contained, self-installing Greenstone CD-ROM for Windows. Because commercial software that creates self-installing CD-ROMs is expensive, this facility includes a homegrown installer module.

When you export the collection, the dialogue informs you of the directory name in which the result has been placed. The entire contents of the directory should be written on to CD-ROM using a standard CD-writing utility.

The immense variety of different possible Windows configurations has made it difficult for us to test and debug the Greenstone installer under all possible conditions. Although the installer produces CD-ROMs that operate on most Windows systems, it is still under development. If you experience problems and you possess a commercial installation package (e.g. InstallShield), you can use it to create CD-ROMs from the information that Greenstone provides. The above-mentioned export directory contains four files that relate to the installation process, and three subdirectories that contain the complete collection and software. Remove the four files and use InstallShield to make a CD-ROM image that installs these directories and creates a shortcut to the program *gsdl\server.exe*.

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 generally 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)*

TEXTPlug 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)*

HTMLPlug processes HTML files. It extracts *title* metadata based on the <title> tag; other metadata expressed using HTML's metatag syntax can be extracted too. There are many options available with this plugin, documented in the *Greenstone Digital Library Developer's Guide*.

WORDPlug (.doc)*

WORDPlug imports Microsoft Word documents. There are many different variants on the Word format—and even Microsoft programs frequently make conversion errors. Greenstone uses independent programs to convert Word files to HTML. For some older Word formats the system resorts to a simple extraction algorithm that finds all text strings in the input file.

PDFPlug (.pdf)*

PDFPlug 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. Again, just replace the *<link>* ... *</link>* tags by *<srclink>* ... *</srclink>* ones.

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. Also, PDFPlug cannot handle encrypted PDF files.

PSPlug (*.ps)

PSPlug imports documents in PostScript. It works best if a standard Linux program, called *ps2ascii*, is already installed on your computer. This is available on most Linux installations, but not on Windows. If this program is not available, PSPlug resorts to a simple text extraction algorithm.

EMAILPlug (*.email)

EMAILPlug 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)

ZIPPlug 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. To access this facility, click the appropriate link on the front page.

The entry page, shown in Figure 17, 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 17) 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.

A collection named *wohiex*, for *Women's History Excerpt*, is visible near the bottom of Figure 17. Figure 18 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 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 17.

Figure 17 Greenstone Administration facility

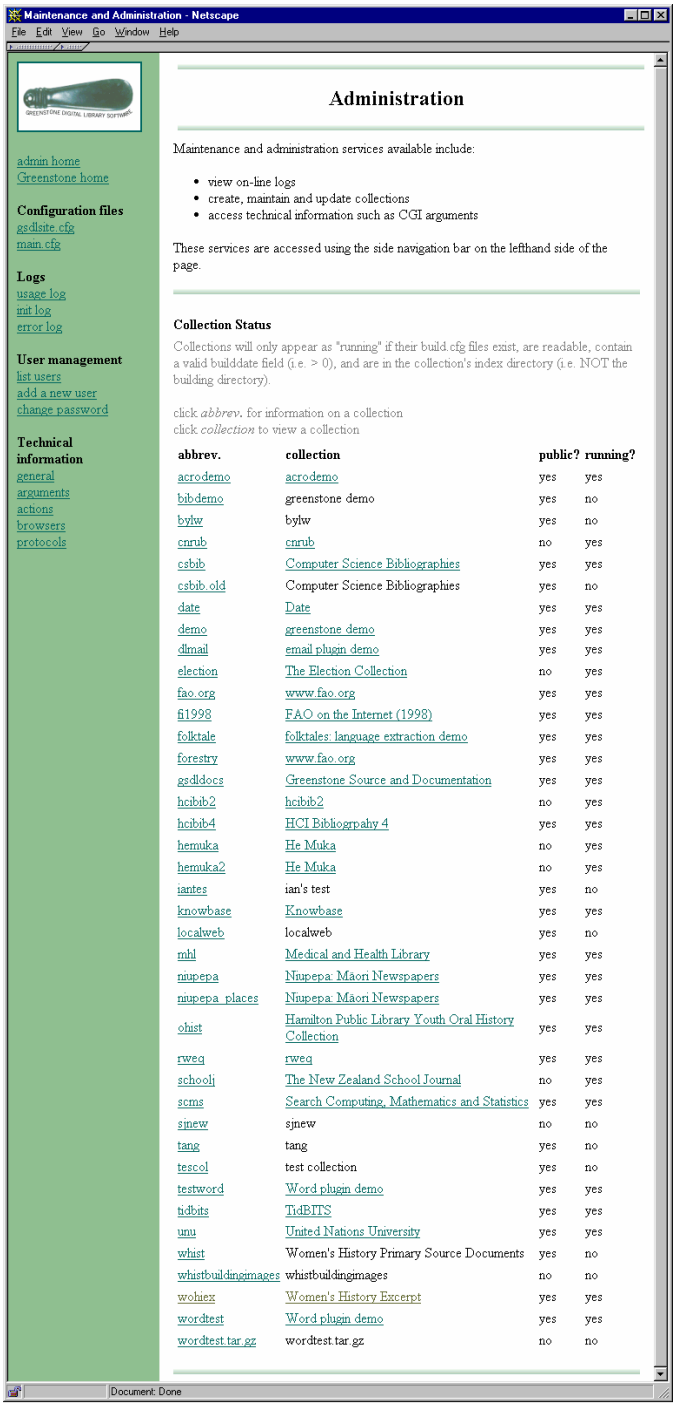
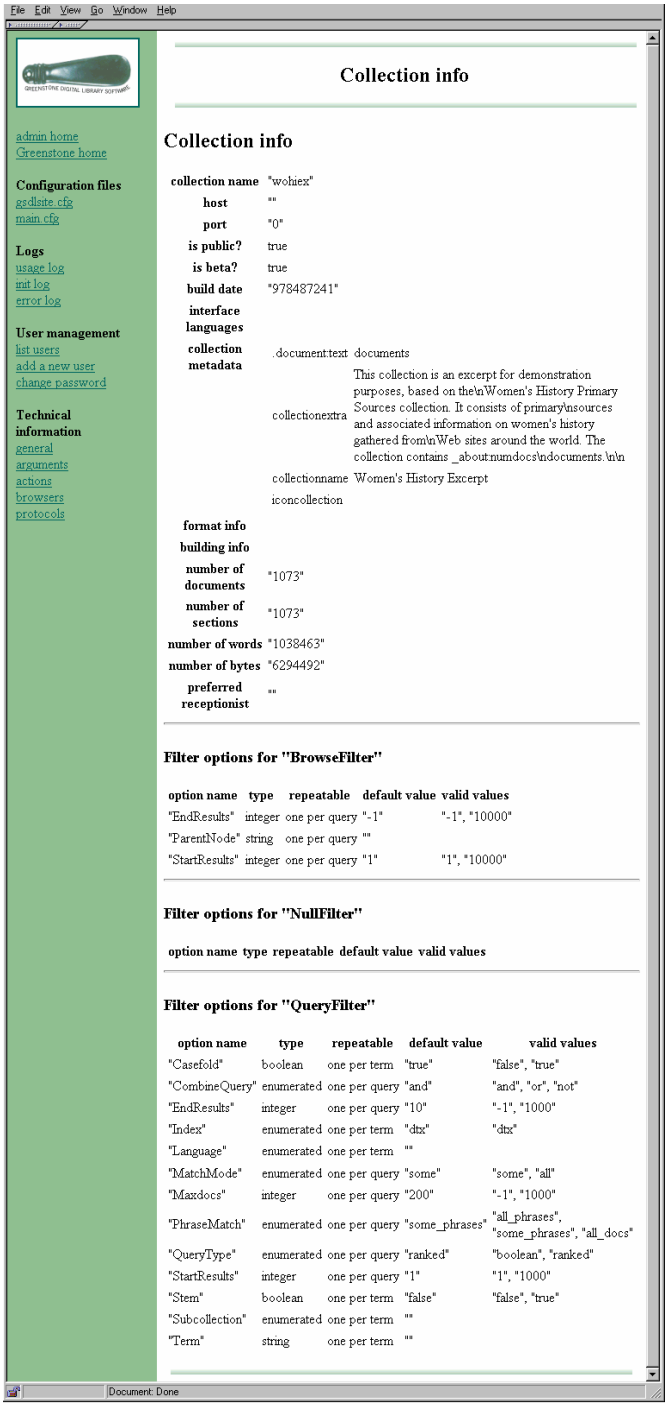


Figure 18 Information about the *Women's History Excerpt* collection



5.1 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.

5.2 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. By activating *usecookies* a unique identification code is assigned to each user, which enables individual user's interactions to be traced through the log file.

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 *usage.txt* is placed in the *etc* 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 17.

5.3 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 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 16a.

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*).

5.4 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, for each possible protocol type, information about each of the collections supported by that protocol.

Finally, 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.



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 some 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 and compressed archives 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 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/Me, NT/2000) and Unix (Linux and SunOS). Any of these systems can be used as a webserver. Collections cannot be built on low-end Windows systems (3.1/3.11), but pre-built collections can be transferred to them.

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	Greenstone 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	Greenstone 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
DL	Development Library, A Greenstone collection of humanitarian information for developing countries
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 Greenstone collection server that implements querying and browsing operations
Format string	A string that specifies how documents and other listings are to be displayed in Greenstone
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, an XML-compliant 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	Greenstone program used at import or build time to generate the OID of each document

HTML	HyperText Markup Language, the language in which web documents are written
import.pl	Greenstone 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	Greenstone program that creates and initializes the directory structure for a new collection
New Zealand Digital Library Project	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 Greenstone 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 and Web server
setup.bat, setup.sh, *setup.csh*	Script used to set up your environment to recognize the Greenstone software

Site configuration file	File that contains specifications used to configure the Greenstone 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 developed by Silicon Graphics
txt2db	Greenstone 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
XML	A standard format for structured documents and data on the web (the Greenstone Markup Language is an XML-compliant format)