

CREATING DIGITAL LIBRARY COLLECTIONS WITH GREENSTONE

AUTHORS

Ian H. Witten
New Zealand Digital Library Project
Department of Computer Science
University of Waikato
New Zealand
Phone: +64 7 838-4246
Fax: +64 7 858-5095
ihw@cs.waikato.ac.nz

David Bainbridge
New Zealand Digital Library Project
Department of Computer Science
University of Waikato
New Zealand
Phone: +64 7 838-4246
Fax: +64 7 858-5095
davidb@cs.waikato.ac.nz

1. ABSTRACT

The Greenstone digital library software is a comprehensive system for building and distributing digital library collections. It provides a way of organizing information based on metadata and publishing it on the Internet. This paper introduces Greenstone and explains how librarians use it to create and customize digital library collections. Through an end-user interface, they add documents and metadata to collections, create new collections whose structure mirrors existing ones, and build collections and put them in place for users to view. More advanced users can design and customize new collection structures.

Keywords: digital libraries, [library systems,] collections management, user interfaces

2. INTRODUCTION

Digital libraries are organized, focused collections of information. They concentrate on a particular topic or theme—and good digital libraries will articulate the principles governing what is included. They are organized to make information accessible in particular, well-defined, ways—and good ones will include a description of how the information is organized [1].

The Greenstone digital library software is a comprehensive suite of software for building and distributing digital library collections [2]. It provides a new way of organizing information and publishing it on the Internet or on removable media (e.g. CD-ROM/DVD). It is widely used in a large number of countries: see www.greenstone.org for a representative selection of example sites.

Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato, and distributed as open source, multilingual software in cooperation with UNESCO and the Human Info NGO. The dissemination of educational, scientific and cultural information, and particularly its availability in developing countries, is central to UNESCO's goals, and appropriate, accessible technology such as Greenstone is seen as a vital tool in this context.

Aim and scope

Greenstone aims to enable users, particularly in universities, libraries, and other public service institutions throughout the world, to build their own digital library collections in the fields of education, science and culture. UNESCO hopes this will encourage the effective deployment of digital libraries to share information and, where appropriate, place it in the public domain. The key points that Greenstone makes it its core business to support include:

- Design and construction of collections
- Distributed on the web and/or removable media
- Customized structure depending on available metadata
- End-user collection-building interface for librarians
- Reader and librarian interfaces in many languages
- Multiplatform operation.

The Appendix summarizes some relevant facts about Greenstone, grouped under “Technical” and “User base”.

The liaison with UNESCO and Human Info has been a crucial factor in the development of Greenstone. Human Info began using Greenstone to produce collections in 1998, and provided extensive feedback on the reader’s interface. UNESCO wants to empower developing countries to build their own digital library collections—otherwise they risk becoming read-only societies in the information revolution. UNESCO selected Greenstone in 2000, and arranges user testing, helps with internationalization, and mounts courses. Internationalization is a central goal: today the Greenstone reader’s interface is available in 35 languages, and the librarian’s interface, including all documentation, is available in 4 (English, French, Spanish, Russian).

Software distribution and development

Greenstone is issued under the terms of the GNU General Public License. It originated in 1996 [3], and the current production version (Greenstone2) was designed about 7 years ago, although it is continually being extended. A complete redesign and reimplementaion, Greenstone3, has been described [4] and released, informed by experience with the current system and the problems and challenges faced by users, international collection developers, and practicing librarians. Greenstone3 allows documents to be dynamically added to collections; provides more flexible ways to dynamically configure the run-time system by adding new services; lowers the overhead incurred by collection developers when accessing this flexibility to organize and present their content; and modularizes the internal structure. The design is based on widely accepted standards that were unavailable when Greenstone2 was designed.

The production version, Greenstone2, is recommended for end-user librarians, while Greenstone3 is an emerging system currently intended for experimental use by computer scientists and information technologists. Greenstone3 is fully compatible with its predecessor, and can run old collections without any modifications whatsoever and make them indistinguishable from the original both visually and in terms of interaction. Librarian-level users can adopt Greenstone2 today, secure in the knowledge that though the software is developing their collections will still run in exactly the same form tomorrow. The remainder of this article focuses exclusively on Greenstone2.

Standards

Many popular document and metadata standards are incorporated into Greenstone. As we shall see, it can deal with documents in HTML, Word, PDF, PostScript, PowerPoint, and Excel formats (amongst others); images in TIFF, GIF, PNG, and JPEG formats (amongst others); and metadata in Dublin Core, MARC, CDS/ISIS, and ProCite formats (amongst others). It can deal with multimedia formats such as MP3, MIDI, and QuickTime. Greenstone’s catholic approach to document and metadata standards creates many demands for conversion facilities. For example, users can change metadata elements from one metadata scheme to another by making choices interactively as they drag documents from one collection to another, or in other circumstances they can either use a default mapping to convert, for example, MARC records to Dublin Core, or define their own crosswalk file.

Emerging digital library standards are also supported. Greenstone incorporates web mirroring software, so that whole sites can be downloaded using the HTTP protocol, to a prespecified depth, and ingested into a collection. Metadata (and documents, if appropriately referenced) from an Open Archives Initiative (OAI) server can easily be ingested too, and any Greenstone collection can be served over the OAI protocol for metadata harvesting (OAI-PMH). Greenstone collections can be exported into the METS metadata encoding and transmission standard, and METS collections can be imported into Greenstone. (The particular form that Greenstone uses has been submitted to the METS Board as a proposed METS Profile.) An option has recently been added that allows end users—typically librarians, not computer specialists—to export a collection from Greenstone and import it into DSpace, and vice versa [5].

3. THE GREENSTONE LIBRARIAN INTERFACE

Greenstone users employ the “Librarian” interface to create and maintain digital library collections. This is intended to help librarians (and others who compile electronic anthologies) expedite the construction and organization of digital information collections. Only a few minutes of the user’s time are needed to set up a collection based on a standard design and initiate the building process, assuming that documents and metadata are already available in electronic form.

More than a few minutes may be required to actually build the full-text indexes and browsing structures that comprise the collection, and compress the text. Some collections contain Gbytes of text; millions of documents. Additionally, even larger volumes of information may be associated with a collection—typically audio, image, and video, with textual metadata. Once initiated, the mechanical process of collection-building may take from a few moments for a small collection to several hours for a multi-Gbyte one that involves many full-text indexes. The Librarian interface monitors all this and provides visual feedback over progress.

Naturally, customized collections that have their own idiosyncratic requirements—as most substantial collections do—take longer to set up, and the design and debugging process can take days, weeks if iterative usability testing is involved. The Greenstone designers wholeheartedly endorse Alan Kay’s maxim that “simple things should be simple, complex things should be possible” [6].

The facilities that Greenstone provides, and the user interface through which library readers access them, are highly customizable at many different levels. Even librarians who need to produce new collections in just a few minutes can dictate what document formats (e.g. HTML, Word, PDF, PostScript, PowerPoint, Excel) or image formats (e.g. TIFF, GIF, PNG, JPEG) will be included, what forms of metadata (e.g. MARC records, OAI archives, ProCite, BibTex or Refer files, CDS/ISIS databases) are available, what searchable indexes will be provided (e.g. full text, perhaps partitioned by language or other features, and selected metadata such as titles or abstracts), and what browsing structures will be constructed (e.g. list of authors, titles, dates, classification hierarchy). Advanced users can control the presentation of items on the screen, personalizing each and every page that Greenstone serves up. All these facilities can be controlled through the Librarian interface.

There are many additional features of Greenstone that lie outside the Librarian interface. Users can translate the interface into different natural languages. If they know HTML they can hook into Greenstone widgets like the full-text search mechanism or browsers from their own Web pages. If they know JavaScript they can incorporate browsing mechanisms such as image maps, and using Perl they can add entirely new browsing facilities, such as stroke-based or Pinyin-based browsing for Chinese. Some new requirements are best met by altering the Greenstone “receptionist” program, written in C++, to add new facilities at runtime.

The Greenstone Librarian interface is targeted at four different levels of user.

Assistant Librarians gain access to the basic features of the Librarian interface: adding documents and metadata to existing collections, creating new collections whose structure mirrors existing ones, and rebuilding collections to reflect changes.

Librarians, the regular or default users of the Librarian interface, perform all the Assistant Librarian tasks above, and can also design new collections—adding, for example, new document types, new full-text indexes, and new metadata browsing features. They typically design a new collection by identifying an existing one that closely matches their needs and adapting its structure as necessary.

Library Systems Specialists can perform all the functions of Librarians, and in addition customize collections in more complex ways, such as those that involve defining and using regular expressions—for example, partitioning collections based on filename or directory structure.

Expert users are those who are experienced with Greenstone and are familiar with running Perl scripts and examining their output. These users can access all features of the Greenstone Librarian interface.

4. THE ROLE AND STRUCTURE OF METADATA

A digital library’s organization is reflected in the interface it presents to users. Much of the organization rests on metadata—structured information about the resources (typically documents) that the library contains. Metadata is the stuff in the traditional card catalogs of bricks-and-mortar libraries (whether computerized or not). It is “structured” in that it can be meaningfully manipulated without necessarily understanding its content. For example, given a collection of source documents, bibliographic information about each document would be

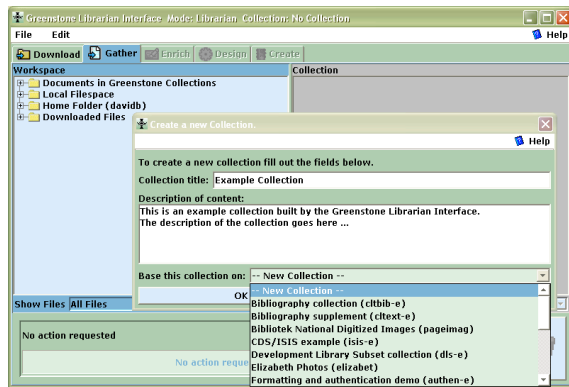


Figure 1. Starting a new collection

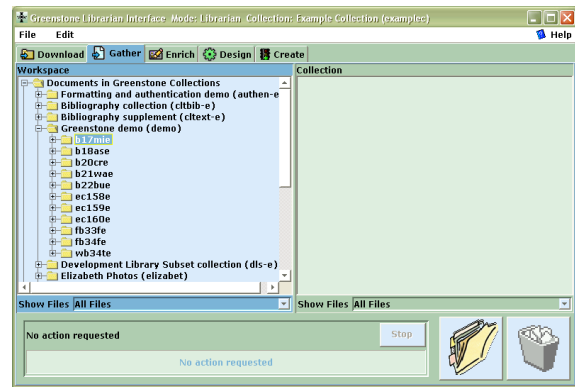


Figure 2. Exploring the local file space

metadata for the collection. The structure is made plain, in terms of which pieces of text represent author names, which are titles, and so on. The notion of “metadata” is not absolute but relative: it is only really meaningful in a context that makes clear what the data itself is [7]. For example, given a collection of bibliographic information, metadata might comprise information about each bibliographic item, such as who compiled it and when.

The use of metadata as the raw material of organization is really the defining characteristic of digital libraries: it is what distinguishes them from other collections of online information. It is metadata that allows new material to be sited within a library and hooked into existing structures in such a way that it immediately enjoys first-class status as a member of the library. Adding new material to ordinary online information collections requires manually linking it in with existing material, but the only manual work needed when adding new items to a digital library is to determine metadata values for each one. If a standard metadata scheme is used, even that may be unnecessary: the information may already be available from another source.

In Greenstone, one or more metadata sets are associated with each collection. There are a few pre-prepared sets, of which Dublin Core is one. Modifications to existing sets and new ones can be defined using an auxiliary Greenstone application called GEMS (Greenstone Editor for Metadata Sets). One important set is the *extracted* metadata set, which contains information extracted automatically from the documents themselves (e.g. HTML *Title* tags, *meta* tags, or built-in Word author and title metadata). This is always present behind the scenes, though it may be hidden from the user. The system keeps metadata sets distinct using namespaces. For example, documents can have both a Dublin Core *Title* (*dc.Title*) and an extracted *Title* (*ex.Title*); they do not necessarily have the same value. Behind the scenes, metadata in documents, and metadata sets themselves, are represented in XML.

In order to expedite manual assignment of metadata, the Librarian interface allows metadata to be associated with document folders as well as with individual documents. This means that users can take advantage of existing document groupings to add shared metadata in one operation. Within the interface users can organize the document hierarchy by dragging items around and creating new sub-hierarchies, which may expedite joint metadata assignment. Metadata values assigned to a folder remain with that folder and are inherited by all files nested within it. If the user subsequently selects a file and changes an inherited metadata value, a warning appears that doing so will override the inherited value. (Of course, these warnings can be turned off: for experienced users they soon become annoying.)

Metadata in Greenstone can be a simple text string (e.g. title, author, publisher). Or it can be hierarchically structured, as with hierarchical classification values, in which case new values can be placed in the classification tree. In addition, it is multivalued: each element can have more than one value. This is used, for example, for multiple authors. The Librarian interface allows existing metadata values to be reused where appropriate, encouraging consistency in metadata assignment by eliminating the need to retype duplicate values.

5. WORKING WITH THE LIBRARIAN INTERFACE

Within the Librarian interface, users collect sets of documents, import or assign metadata, and build them into a Greenstone collection. It is an interactive platform-independent Java application that runs on the same computer that operates the Greenstone digital library server.¹ It is closely coupled to the server, and tightly integrated with

¹ There is also an applet version that allows collections to be constructed remotely.

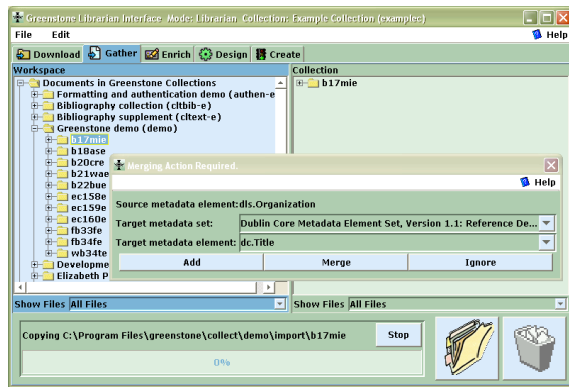


Figure 3. Importing existing metadata

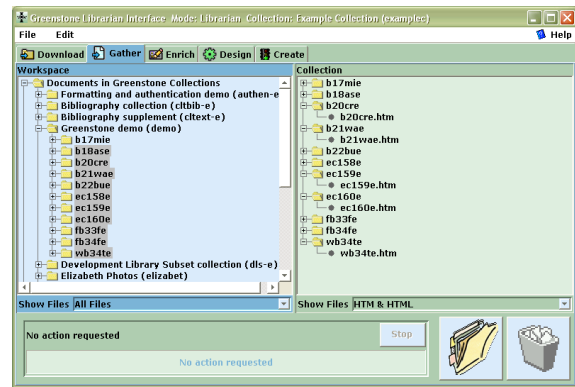


Figure 4. Filtering the file trees

Greenstone’s collection design and creation process. It incorporates various open-source packages for such tasks as file browsing, HTML rendering, web mirroring, and efficient table sorting.

The Librarian interface supports six basic activities, which can be interleaved but are normally undertaken in this order.

1. If required, download any documents from the Web that need to be included in the collection. This optional step is only relevant when the collection will contain material that is sourced from the Web.
2. Bring documents into a collection—whether to populate a new collection or update an existing one. Metadata files may also be brought in. Users browse the the computer’s file space to find documents to include, and drag and drop them into place. Any documents imported from existing Greenstone collections come with existing metadata attached.
3. Enrich the documents by adding metadata to them manually. Documents can be grouped into folders (it is easy to retain existing folder structures when dragging documents in under step 1), and any metadata assigned to folders is inherited by all documents nested within them.
4. Design the collection by determining its appearance and the access facilities that it will support: full-text search indexes, browsing structures, the format of items on the pages that Greenstone generates, etc. This design facility is not available at the Assistant Librarian user level.
5. Build the collection using Greenstone. This work is done by the computer; users are presented with a progress bar. This is the point where Expert users might examine the output of Perl scripts, which are presented in a scrolling window, to determine if anything is going wrong.
6. Pass the newly-created collection to the Greenstone digital library server for previewing. The collection is automatically installed as one of those in the user’s personal digital library, and a web page is opened showing the collection’s home page.

To convey the operation of the Librarian interface we work through a small example. Figures 1 to 12 are screen snapshots at various points during the interaction. This example uses documents in the Humanity Development Library Subset 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.

Assembling source material

To commence, users either open an existing collection or begin a new one. Novice users (“Assistant Librarians”) generally work with existing collections, adding documents and/or metadata. However, they can begin a new collection by copying the structure of an existing one, effectively creating an empty shell exactly like an existing collection, and adding documents and metadata to it. Collection design involves more advanced skills.

Figure 1 shows the process of starting a new collection. Having selected *New* from the file menu, the user fills out general information about the collection—its name and a brief description of the content—in the popup window shown. The name is a short phrase used to identify the collection throughout the digital library: existing collections have names like *Food and Nutrition Library*, *World Environmental Library*, and so on. The description is a statement about the principles that govern what is included in the collection, and appears under the heading *About this collection* on the collection’s home page.

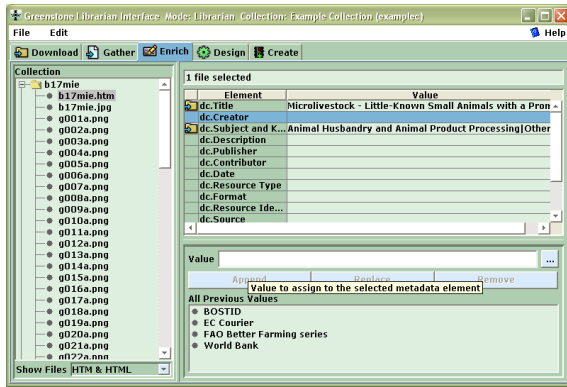


Figure 5. Assigning metadata using the *Enrich* view

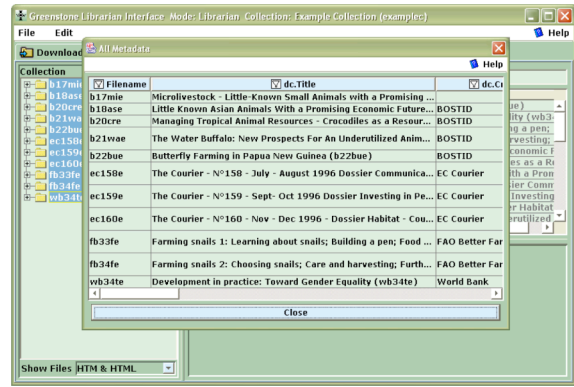


Figure 6. Viewing all metadata assigned to selected files

At this point, the user decides whether to base the new collection on an existing one, selecting from the menu pulled down in Figure 1, or design a new one. In this example we will design a new collection, and now it is necessary to select one or more metadata sets for it. We choose Dublin Core from a popup menu (not shown in the Figure). Now 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 Figures 1–4, is active initially. It allows the user to explore the local file space and existing collections, gathering selected documents into the new collection. The panel is divided into two sections, the left for browsing existing file structures and the right for organizing the documents in the collection.

Users navigate the existing file structure hierarchy in the usual way. They can select files or directories, drag them into the collection on the right, and drop them there. Entire file hierarchies can be dragged and dropped, and files can be multiply selected in the usual way. Users can navigate around the collection on the right too, adjusting the file hierarchy by dragging items around, creating new sub-hierarchies, and deleting files if necessary. Another possible source of documents is the web itself, and the *Download* panel can be used for this. This panel has many options: mirroring depth, automatically download embedded objects like images, only mirror from the same site, etc. URLs are entered into the panel (typically cut and pasted from a browser), and the system maintains a queue of items to download. The actual download operation is accomplished by a widely-used open-source mirroring utility. The resulting files appear as another top-level folder, called *downloads*, on the left-hand side of the *Gather* panel.

In Figure 2 the interactive file tree display is being used to explore the local file system. At this stage the collection on the right is empty; the user populates it by dragging files of interest from the left-hand panel and dropping them into the right-hand one. Such files are copied rather than moved, so as not to disturb the original file system. 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, which the Librarian interface preserves when it moves them into the new collection. Conflicts may arise because their metadata may have been assigned according to a different metadata set from the one attached to the new collection, and the Librarian interface helps the user resolve these.

In Figure 3 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 new element to that set, or choose a metadata set and element, and press *Merge* to effectively rename the old metadata element to the new one by merging the two. Metadata in subsequent documents will be imported in the same way automatically.

When large file sets are selected, dragged, and dropped into the collection, the copying operation may take some time—particularly if metadata must be converted too. The Librarian interface indicates progress by showing which file is being copied and what percentage of files has been processed. The implementation is multi-threaded: users can proceed to another stage while copying is still in progress.

Special mechanisms are needed for dealing with large file sets. For example, the user can filter the file tree to show only certain files, using a dropdown menu of file types displayed underneath the trees. In the right-hand

panel of Figure 4, only HTM and HTML files are being shown (and only these files will be copied by drag and

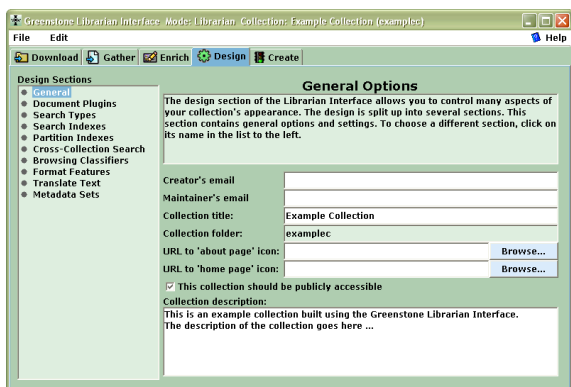


Figure 7. Designing the collection

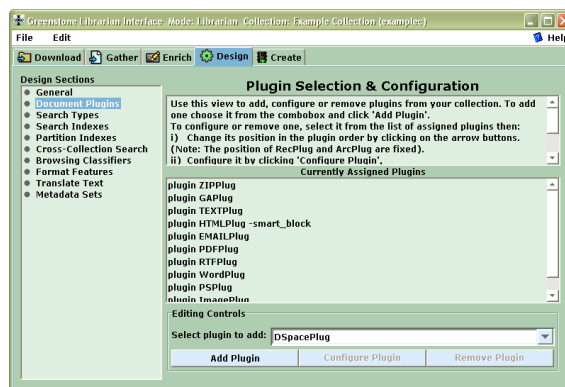


Figure 8. Specifying which plug-ins to use

drop). In fact, the left-hand panel is showing the same part of the file space without filtering, and you can see the additional .png and .jpg files that are present there.

Adding metadata to documents

The next phase of collection-building is to enrich the documents by adding metadata. This is where the Librarian users spend most of their time: enhancing the collection by selecting individual documents and manually adding metadata. We have already discussed two features of the Librarian interface that help with this task:

- Documents that are copied during the first step come with any applicable metadata attached.
- Whenever possible, metadata is extracted automatically from documents.

The Librarian implements two further features that expedite manual metadata assignment:

- Metadata values can be assigned to several documents at once, either by virtue of them being in a folder, or through multiple selection.
- Previously-assigned metadata values are kept around and made easy to reuse.

The *Enrich* tab brings up the panel of information shown in Figure 5. On the left is the document tree representing the collection, while on the right metadata can be added to individual documents, or groups of documents. Users often want to see the document they are assigning metadata to, and if they double-click a document in the pane on the left it is opened by the appropriate viewing program. Here the user has selected a document and typed “new creator” as its *dc.Creator* metadata. The buttons for appending, replacing and removing metadata become active depending on what selections have been made. Values previously assigned to *Creator* metadata are shown in the pane labeled “All previous values.”

At any time the user can view all the metadata that has been assigned to the collection. The popup window in Figure 6 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.

Designing a collection

All except “Assistant Librarian” users of the Librarian interface have the ability to design new collections, which involves specifying the structure, organization, and presentation of the collection being created. 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 in a machine-readable form.

Collection design has many aspects. Users might review and edit collection-level metadata such as title, author and public availability of the collection. They might define what full-text indexes are to be built. They might create sub-collections and have indexes built for them. They might add or remove support for predefined interface languages. They will need to decide what document formats will be included. In Greenstone, document types are processed by modules called “plug-ins,” and each plug-in may need to be configured by specifying appropriate arguments. The collection designer will need to specify what browsing structures will be

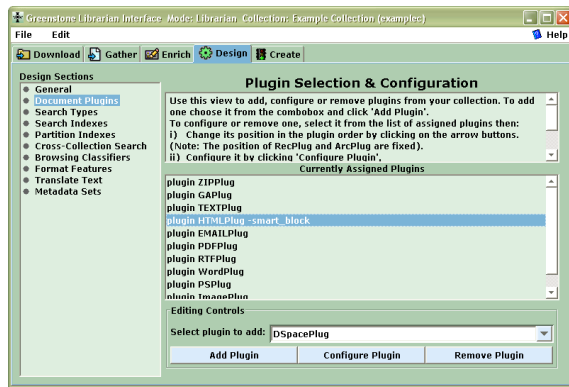


Figure 9. Configuring the arguments to a plug-in

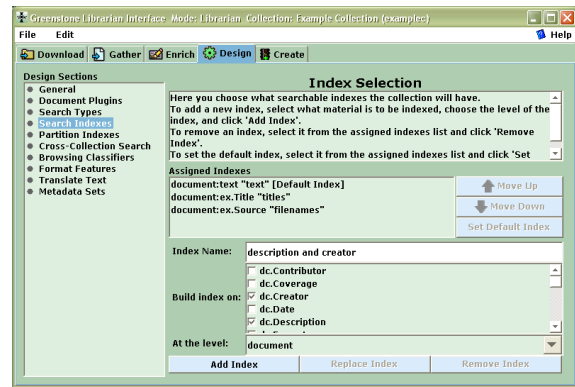


Figure 10. Adding a full-text-search index

constructed—in Greenstone, these are built by modules called “classifiers,” which also have various arguments. It will also be necessary to specify the formatting of various items in the collection’s user interface. Sensible, generally-applicable defaults are supplied for all these features.

Users accomplish this process with the *Design* panel illustrated in Figures 7–10. It has 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 process of editing the raw collection configuration file manually. In Figure 7 the user has clicked the *Design* tab and is reviewing general information about the collection, which was 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, and Metadata Sets. For example, clicking the *Document Plug-in* button brings up the screen shown in Figure 8, which allows you to add, remove or configure plug-ins, and change the order in which the plug-ins are applied to documents.

Both plug-ins and classifiers have many different arguments or “options” that the user can supply. The dialog box in Figure 9 shows the user specifying arguments to a plug-in. The grayed-out fields become active when the user adds the option by clicking the preceding tick-box. Because Greenstone is a continually growing open-source system, the number of options tends to grow as developers add new facilities. To help cope with this, Greenstone has a “plug-in information” 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.

In Figure 10 the user is adding a new full-text-search index to the collection, in this case based on both *dc.Creator* and *dc.Description* metadata. In Figure 11 she is adding a “cross-collection search” capability so that other collections are searched whenever this one is.

Building the collection

The next step is to construct the collection formed by the documents and assigned metadata. The brunt of this work is borne by the Greenstone code itself. The user observes the building process through a window that shows the text output generated by Greenstone’s importing and index-building scripts—filtered for brevity in all but the Expert user level—along with a progress bar that indicates the overall degree of completion. Figure 12 shows the Expert (i.e. most detailed) version of the *Create* view through which users control collection building. On the left are groups of options that can be applied during the creation process: Import, Build, and Message Log. The user selects values for the options if necessary, and clicks *Build Collection*.

6. BEYOND THE LIBRARIAN INTERFACE

Most of the customization that non-programming users perform in Greenstone takes place in the collection configuration file, which the Librarian interface creates. It depends crucially on the availability of metadata, and the structures defined are only produced if appropriate metadata is provided. However, Greenstone has more advanced customization features. Our philosophy is to target the most common features and make them accessible to librarian-level users without particular training in computer science. But users who are prepared to dig deeper can accomplish more.

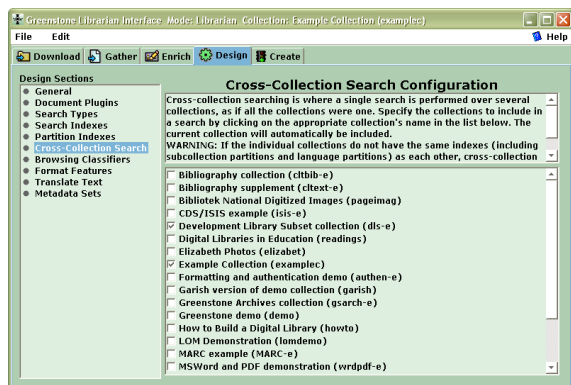


Figure 11. Adding a cross-collection search facility

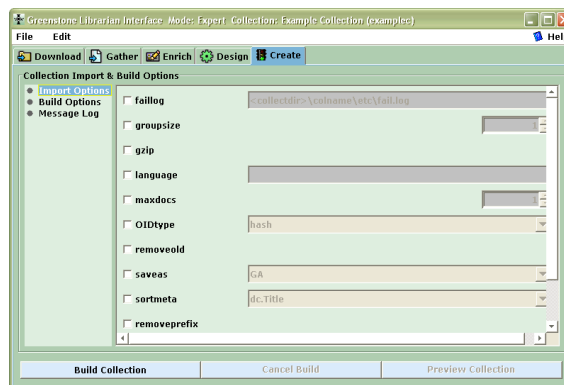


Figure 12. Getting ready to create the new collection

Macros

Greenstone incorporates a macro facility, expressed as an extension of HTML. It includes the ability to define macros and perform textual substitution. Currently, for example, there are interfaces in over thirty-five languages, from Arabic to Turkish, Bosnian to Ukrainian, Chinese to Vietnamese. To accommodate these variants, and to allow the language interfaces to be updated when new facilities are added, all web pages are passed through a macro expansion phrase before being displayed. This means that a new language can be added by providing a new set of language-specific macros, a task that has been performed many times by people with no expertise in Greenstone.

The digital library functionality is hooked into the user interface through “dynamic macros” whose expansions are determined by the system (in terms of other macros). For example, the search widget is generated by a dynamic macro. Users can incorporate this widget into their own Web pages, provided they go through the macro expansion phase. A total of about twenty dynamic macros provides access to Greenstone’s full user interface functionality. Users who work with Greenstone can capitalize on the macro system to radically alter the style of the pages generated, and some have produced attractive new designs for the Greenstone user interface [8].

Altering the run-time system

The part of Greenstone that serves collections to users is called the “receptionist,” and one sometimes has to resort to changing this program to achieve a desired level of customization. This rarely involves large changes, but creates software management difficulties in dealing with different parallel versions.

Our system development strategy is to accept the inevitability of occasionally having to build a special-purpose collection-dependent receptionist to achieve some desired features, and to note what is required with a view to incorporating it as an option within the standard Greenstone code.

7. CONCLUSIONS

A general-purpose digital library system like Greenstone must cater for a wide range of users. We have targeted the Librarian interface at four different user levels: *assistant librarians*, who can add to existing collections and create new ones with the same structure; *librarians*, who can, in addition, design new collections; *library systems specialists*, who can customize collections in more complex ways; and *expert users*, who can deal with every aspect of the system.

A digital library may be customized in a wide variety of different ways, and virtually every collection has its own idiosyncratic requirements. Although a basic Greenstone collection of new material with a standard look and feel can be set up in just a few minutes, most users want more personalization. As the number of collections grows and the variety of styles increases, it becomes more likely that some existing collection will match new requirements.

It is always difficult to produce good, up to date, documentation for a richly functional software system. In fact, from a user’s point of view the chief bottleneck in customization is documentation, not the facilities that are

provided. Consequently collection builders need access to advice and assistance from others, in order to continue to learn how to tailor the software to meet ever-changing requirements. There is a lively Greenstone email discussion group; participants hail from around 70 countries.

Digital libraries have the advantage over other interactive systems that their user interfaces are universally based on metadata. Metadata is the glue that allows new documents to be added and immediately become first-class citizens. It is also the key to user interface customization, and Greenstone incorporates a range of mechanisms at different levels to capitalize on this.

8. ACKNOWLEDGEMENTS

The Greenstone Digital Library software has grown out of the stimulating research environment of the New Zealand Digital Library project, and we gratefully acknowledge the profound influence of all project members. In particular, John Thompson pioneered the Librarian interface described in this paper.

9. REFERENCES

- [1] Lesk, M. (2005) *Understanding digital libraries*. Morgan Kaufmann, San Francisco.
- [2] Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann, San Francisco, CA.
- [3] Witten, I. H., Cunningham, S.J. and Apperley, M. (1996) "The New Zealand Digital Library Project." *D-Lib Magazine* 2(11) (www.dlib.org/dlib/november96/newzealand/11witten.html)
- [4] Bainbridge, D., Don, K.J., Buchanan, G.R., Witten, I.H., Jones, S., Jones, M. and Barr, S.I. (2004) "Dynamic digital library construction and configuration." *Proc European Digital Library Conference*, Bath, England.
- [5] Witten, I.H., Bainbridge, D., Tansley, R., Huang, C.Y. and Don, K.J. (2005) "StoneD: A bridge between Greenstone and DSpace." Submitted to *Joint Conference on Digital Libraries*, Denver.
- [6] Davidson, C. (1993) "The man who made computers personal." *New Scientist*, No. 1978, pp. 32–35; June.
- [7] Lagoze, C. and Payette, S. (2000) "Metadata: Principles, practices and challenges." In *Moving theory into practice: digital imaging for libraries and archives*, edited by A.R. Kenney and O.Y. Rieger. Research Libraries Group, Mountain View, CA, pp. 84–100.
- [8] Zhang, A. (2003) "Customizing the Greenstone User Interface." Washington Research Library Consortium; August; <http://www.wrlc.org/dcpc/UserInterface/interface.htm>.

10. APPENDIX: GREENSTONE DIGITAL LIBRARY SOFTWARE FACTSHEET

Greenstone is a suite of software for building and distributing digital library collections. It is not a digital library but a tool for building digital libraries. It provides a new way of organizing information and publishing it on the Internet in the form of a fully-searchable, metadata-driven digital library. It has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium. It is open-source, multilingual software, issued under the terms of the GNU General Public License. Its developers received the 2004 IFIP Namur award for “contributions to the awareness of social implications of information technology, and the need for an holistic approach in the use of information technology that takes account of social implications.”

<http://www.greenstone.org>

Technical

Platforms. Greenstone runs on all versions of Windows, and Unix, and Mac OS-X. It is very easy to install. For the default Windows installation absolutely no configuration is necessary, and end users routinely install Greenstone on their personal laptops or workstations. Institutional users run it on their main web server, where it interoperates with standard web server software (e.g. Apache).

Interoperability. Greenstone is highly interoperable using contemporary standards, It incorporates a server that can serve any collection over the Open Archives Protocol for Metadata Harvesting (OAI-PMH), and Greenstone can harvest documents over OAI-PMH and include them in a collection. Any collection can be exported to METS (in the Greenstone METS Profile, approved by the METS Editorial Board and published at <http://www.loc.gov/standards/mets/mets-profiles.html>), and Greenstone can ingest documents in METS form. Any collection can be exported to DSpace ready for DSpace’s batch import program, and any DSpace collection can be imported into Greenstone.

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

Metadata formats. Users define metadata interactively within the Librarian interface. These metadata sets are predefined:

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

New metadata sets can be defined using Greenstone’s Metadata Set Editor. “Plug-ins” are used to ingest externally-prepared metadata in different forms, and plug-ins exist for

XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace, METS

Document formats. Plug-ins are also used to ingest documents. For textual documents, there are plug-ins for

PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code

For multimedia documents, there are plug-ins for

Images (any format, including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio, and a generic plug-in that can be configured for audio formats, MPEG, MIDI, etc.

User base

Distribution. As with all open source projects, the user base for Greenstone is unknown. It is distributed on SourceForge, a leading distribution centre for open source software.

Distributed via SourceForge since:	Nov 2000
Average downloads since then:	4500/month
Currently running at:	4500/month
Proportion of downloads that are documentation:	60%
Proportion of downloads that are software:	40%
Of these, 80% are Windows binaries	
15% are Linux binaries	
5% are source	

Examples. Examples of public Greenstone collections (see <http://www.greenstone.org> for URLs) can be found at:

Association of Indian Labour Historians, Delhi
Auburn University, Alabama
California University at Riverside
Chicago University Library
Detroit Public Library
Gresham College, London
Hawaiian Electronic Library
Illinois Wesleyan University
Indian Institute of Management
Kyrgyz Republic National Library
LeHigh University, Pennsylvania
Mari El Republic, Russia
National Centre for Science Information, Bangalore, India
Netherlands Institute for Scientific Information Services
New York Botanical Garden
Peking University Digital Library
Philippine Research Education and Government Information Network
Secretary of Human Rights of Argentina
Slavonski Brod Public Library, Slovenia
State Library of Tasmania
Stuttgart University of Applied Sciences
Texas A&M University Center for the Study of Digital Libraries
University of Illinois
University of North Carolina ibiblio project
Vietnam National University
Vimercate Public Library, Milan, Italy
Washington Research Library Consortium
Welsh Books Council

UN agencies with an interest in Greenstone include

UNESCO, Paris
Sponsors distribution of the Greenstone software as part of its *Information for All* programme

Food and Agriculture Organization (FAO), Rome
Their *Information Management Resource Kit* uses Greenstone as the (only) example of digital library software in the *Digitization and digital libraries* self-instructional module

Institute for Information Technology in Education (IITE), Moscow
Have commissioned an extensive course on *Digital libraries in education* that uses Greenstone for all the practical work

United Nations University (UNU), Japan
Two CD-ROM collections of UNU material have been produced

Humanitarian collections. Greenstone is used by Human Info NGO in Belgium to produce collections of humanitarian information and distribute them on CD-ROM widely throughout the developing world. (For more information, contact Michel Loots <mloots@humaninfo.org>)

Number of humanitarian collections:	approx 35–40
Annual distribution of each one:	approx 5,000 copies

Languages. One of Greenstone's unique strengths is its multilingual nature. The reader's interface is available in the following languages:

Arabic, Armenian, Bengali, Catalan, Croatian, Czech, Chinese (both simplified and traditional), Dutch, English, Farsi, Finnish, French, Galician, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Kyrgyz, Latvian, Maori, Mongolian, Portuguese (BR and PT versions), Russian, Serbian, Spanish, Thai, Turkish, Ukrainian, Vietnamese

The Librarian interface and the full Greenstone documentation (which is extensive) is in:

English, French, Spanish, and Russian.

Training is a bottleneck for widespread adoption of any digital library software. Many international training courses have been run.

UNESCO
has sponsored training courses in Bangalore (2002 and 2003), Almaty (2003), Senegal (2004), Suva (2004)

Self-study courses
FAO and UNESCO IITE have produced training material on Greenstone

Digital Library conferences
There have been Greenstone tutorials (on several occasions) at all major digital library conferences: JCDL, ECDL, ICADL, ICDL

Librarian conferences
There have been Greenstone workshops and presentations at LITA, DLF, ALA Annual Conference

Payson Institute, Tulane University
has run courses that use Greenstone collections as a resource in locations in Africa (Burkina Faso, Cameroon, Cote d'Ivoire, Democratic Republic of Congo, Ghana, Rwanda, Senegal, Sierra Leone, Togo) and Latin America (Argentina, Bolivia, Colombia, Ecuador, Guatemala)

Others
There have been several Greenstone courses in India (e.g. Khozikode, Poona), some in Canada (Vancouver, Calgary, Edmonton), one in Cuba (Havana).

E-mail support

Number of people on Greenstone email lists:	600
Number of countries represented:	70
Number of messages (excluding spam):	150/month