# Greenstone:
# Collection management for digital works

**Gordon W. Paynter, Ian H. Witten, David Bainbridge and Stefan Boddie**

The New Zealand Digital Library Project
The University of Waikato, Hamilton, New Zealand.

The Greenstone Digital Library Software is a comprehensive package for creating, maintaining, presenting and disseminating collections of digital resources (http://www.greenstone.org/, [10]). Greenstone collections offer effective full-text searching and metadata-based browsing facilities that are attractive and easy to use, and a user-friendly interface called *The Collector* makes it easy for people to assemble their own library collections from disparate source documents. To address the exceptionally broad demands of digital libraries, the system is public and extensible: it is distributed under the terms of the GNU General Public License and users are encouraged to contribute modifications and enhancements.

## The New Zealand Digital Library Project

The New Zealand Digital Library Project (NZDL) is a collective of researchers based in New Zealand with collaborators around the globe (http://www.nzdl.org/). The project originated in the Department of Computer Science at The University of Waikato and combines original research with the development of the Greenstone Digital Library Software.

The first NZDL collection was created in 1995 to demonstrate research into text compression and indexing [6]. It was based on several thousand computer science technical reports downloaded from over 300 public FTP sites, converted from their native PostScript into plain text, and indexed using the compression and indexing software (http://www.nzdl.org/cstr). Users could search for documents using any combination of words, and receive an ordered list of documents whose full text included those words, along with hyperlinks back to the original documents. The result was striking: it frequently drew attention to many extremely pertinent but previously unknown documents (such as obscure PhD theses), without the need to invest any effort in manual metadata production.

Several new collections followed. The Hamilton Public Library *Youth Oral History* Project is based on interviews with some of the city's oldest residents about life in Hamilton when they were young. This collection, held as cassette tapes, photographs and abstracts in the reference department of the city's central library, was difficult to access and search. In collaboration with the NZDL, the materials were digitized and turned into a Greenstone collection. Users can search for words and phrases, then read the transcriptions and listen to the original voices (http://www.nzdl.org/ohist).

The next major collection was national in scope: in 1988-1989 the Alexander Turnbull Library scoured New Zealand's libraries for the best surviving copies of any newspaper published for a Maori audience between 1842 and 1932, added cataloging data, and made the resulting "Maori Newspaper Collection" available to interested parties on microfiche in 1996. This is a very important cultural and historical resource, but is difficult to access (because the archive is physically located in Wellington, and because it is on microfiche). Project members Te Taka Keegan and Mark Apperley (whose great grandfather was an editor of one of the early papers) had the images converted from microfiche to TIFF image files in 1999, and organized Government funding, graduate students, and volunteers to convert the images to text and proofread them. The resulting *Niupepa* collection was officially launched to Maori Schools early this year and is available over the Internet (http://www.nzdl.org/niupepa) or on CD-ROM. Users can choose

between Maori and English-language interfaces, navigate the cataloging metadata, search the text of every paper, and view the results as text or as images of the original newspapers.

The multilingual nature of the software is exploited by a series of humanitarian collections based on traditional library materials, such as medical, agricultural and engineering textbooks, which have been distributed to developing nations on CD-ROM. Most of these collections were not created by the NZDL: they were created using the Greenstone software by organizations like the United Nations University [5], The Global Help Project [1], The Pan-American Health Organization [3], and The United Nations Educational, Scientific and Cultural Organization (UNESCO) [4]. (Several are mirrored at http://www.nzdl.org/ and http://www.humanitylibraries.net/). More humanitarian collections are in production: the Human Info NGO is compiling 20 million pages of humanitarian and development information on Greenstone CD-ROMs (http://humaninfo.org/), while UNESCO is preparing to distribute ten thousand CD-ROMs containing the Greenstone software, with interfaces and full documentation in three languages (English, French, Spanish) to developing nations.

## The collection management problem

The usual solution to creating a digital library of digital resources is the most obvious: just put them on the Web. But consider how much effort is involved in constructing a Web site for a digital library. To be effective it needs to be visually attractive and ergonomically easy to use, incorporate convenient and powerful searching capabilities, and offer rich and natural browsing facilities. Above all it must be easy to maintain and augment, which presents a significant challenge if any manual organization is involved.

The alternative is to automate these activities through software tools. But the broad scope of digital library requirements makes this a daunting prospect. Our motivating problems required facilities for

- importing documents in a variety of formats,
- importing documents in different languages and character sets,
- handling the same document in multiple times in different formats,
- importing multimedia documents,
- multilingual user interfaces,
- multilingual information retrieval,
- metadata standards compliance,
- metadata in non-standard formats, and
- multiple operating systems.

Notwithstanding intense research activity in the digital library field, comprehensive software systems for creating digital libraries are still not widely available. Yet the collections above, for all their variety, share a common structure: digital resources are imported, indexed, searched, and displayed. The goals of the NZDL are to meet these needs for users with a broad range of skills, presentation requirements, and source documents. The next section describes the Greenstone software, our solution to these problems.

## The Greenstone Digital Library Software

The Greenstone software operates under most variants of Unix (including Linux, FreeBSD and MacOS X) and all versions of Microsoft Windows (from Version 3 onwards). A standard Web server like Apache (or a commercial equivalent) can be used to make a library available over the Internet, though the *local library* version is a popular alternative for Windows users who do not have access to a Web server. The executable Linux and Windows programs, and the Greenstone source code, can be downloaded from http://www.greenstone.org/.

This section describes the software. First, we explain what Greenstone does for its users—and what it can do for you—though a survey of some of its major features (many more are omitted for brevity). We examine the software from the point of view of a "digital library patron", and also from the perspective of a "collection maintainer" creating a collection. We then describe how Greenstone itself has benefitted from other open-source projects that came before it; and again we have space to acknowledge only a few.

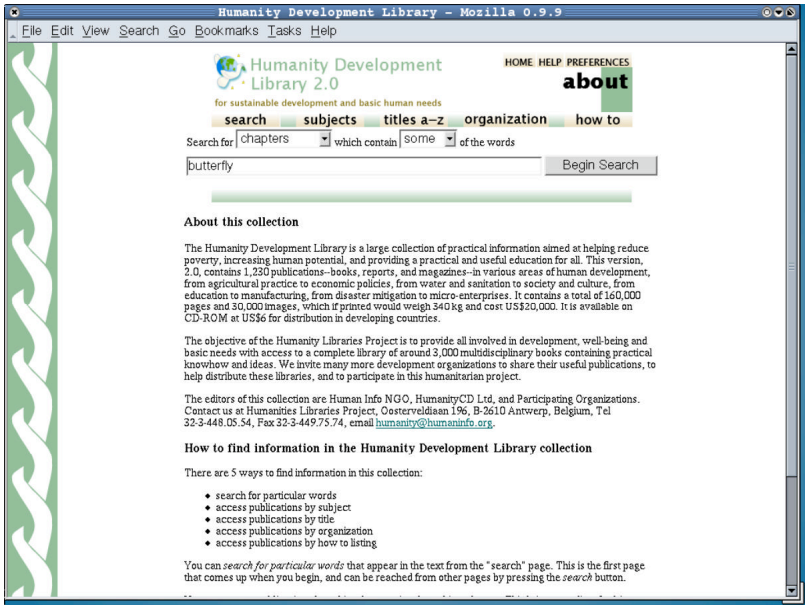


**Figure 1**: Searching the Humanity Development Library for *butterfly*.


**Figure 2**: Browsing the Humanity Development Library by *subject*.

## Searching and browsing collections

Greenstone digital libraries are arranged in *collections*. A collection comprises several (typically several thousand, or several million) documents, and a library may include any number of collections, each organized differently. Collections built with Greenstone offer effective, attractive searching and browsing facilities based on metadata and the full-text of electronic documents.

The most common use of Greenstone is to look for information in a collection. Most collections support both searching and browsing, although they differ depending on the collection design and the metadata available. Typically you can search for particular words that appear in the text, or within a section of a document, or within a title or section heading. A variety of interfaces exist for browsing collections by *title*, *subject*, *date*, or any other metadata chosen by the collection maintainer.

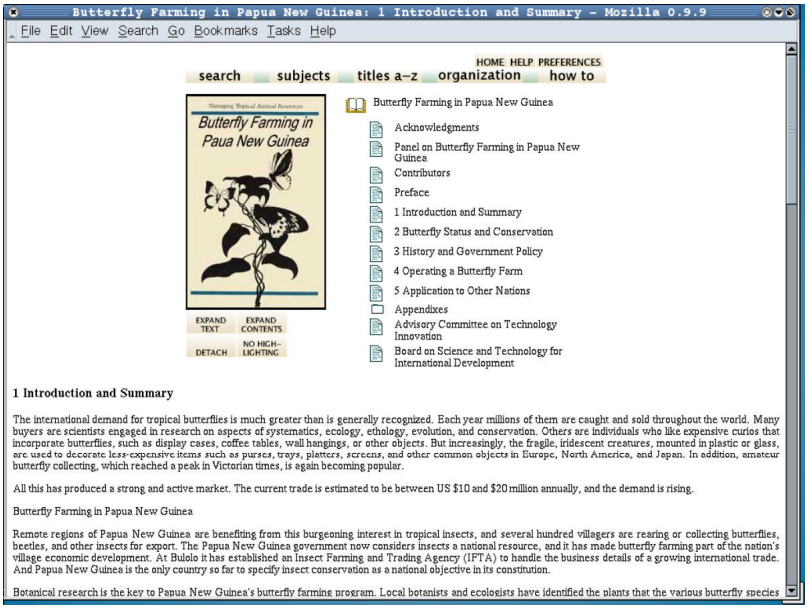


**Figure 3**: A document from the Humanity Development Library.

An example of searching is shown in Figure 1 where documents in the Global Help Project's *Humanity Development Library* (http://humanitylibraries.net/) are being searched for chapters matching the word *butterfly*. In Figure 2 the same collection is being browsed by subject: by clicking on the bookshelf icons the user has discovered an item under *Section 16, Animal Husbandry*. Pursuing an interest in butterfly farming, the user selects a book by clicking on its book icon (Figure 3).

Documents are typically presented as Web pages generated from the source documents by Greenstone. In Figure 3 the front cover of the book is displayed as a graphic on the left, and an automatically constructed table of contents appears at the start of the document. The current focus, *Introduction and Summary*, is shown in bold in the table of contents with its text starting further down the page. Most Greenstone collections present documents as automatically-generated Web pages. This allows documents in different source formats to be presented in a consistent manner, and lets users view the entire collection with a standard Web browser—no special viewing applications are required. Of course, the collection maintainer may choose to present the original source document instead of (or in addition to) the HTML version.

All the icons in the screenshots of Figures 1-3 are clickable. Those icons at the top of the page return to the library home page, provide help text, and allow you to set user interface and searching preferences. The navigation bar underneath gives access to the searching and browsing facilities, which differ from one collection to another.

## Internationalization

The international Unicode character set is used throughout Greenstone so that documents in any language and character encoding can be imported. (In fact, Greenstone can automatically detect the language and encoding of most documents.) Collections of documents in Arabic, Chinese, Cyrillic, English, French, Spanish, German, and Maori are publicly available. The NZDL Web site (http://www.nzdl.org/) hosts many of these collections, and the Greenstone Web site (http://www.greenstone.org/) links to further examples.

It makes little sense to have a collection whose content is in Chinese or Russian, but whose supporting text—instructions, navigation buttons, labels, images, help text, and so on—are in English. Consequently, the entire Greenstone interface has been translated into a range of languages, and the interface language can be changed by the user as they browse from the *Preferences* page. This is possible because the language-dependent images and Web pages are generated from a set of "macro files" that

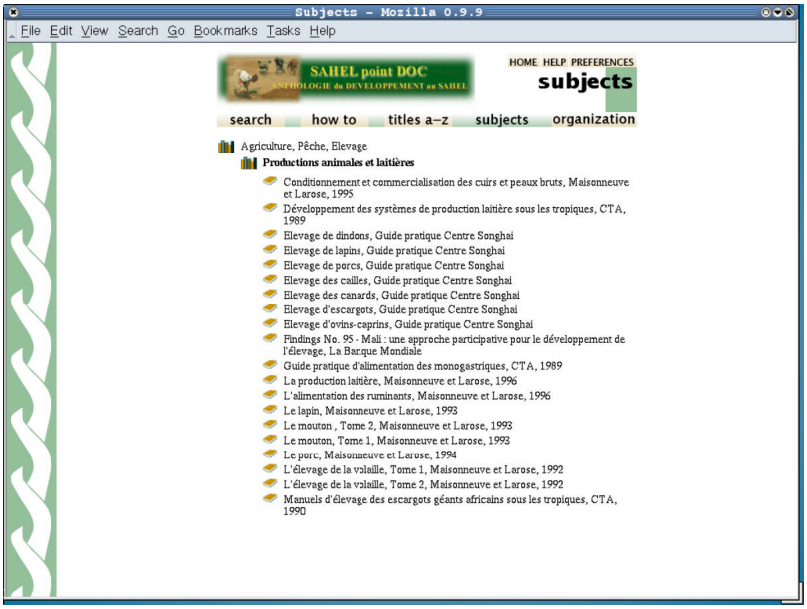


**Figure 4**: A Web page viewed through the French interface.

have been translated by Greenstone users in other parts of the world and contributed back to the project. (The same flexibility provides text-only versions of the interface to accommodate visually impaired users.) Figure 4 shows an example a Russian collection using the Cyrillic interface.

Finally, to aid collection editors in other countries, UNESCO and the Human Info NGO are translating the Greenstone manual and Web site into French and Spanish.

## Creating collections

The overall purpose of Greenstone is to take a collection of electronic documents and, under the editorial control of the *collection maintainer*, turn it into an organized, attractive, accessible digital library. Collections comprise many documents: thousands, tens of thousands, or even millions. In the best case, they are well-organized, consistently formatted, and annotated with descriptive metadata and internal links. It is far more common, however, that the source documents benefit from no such editorial attention—instead, a potential collection is a jumble of documents in different formats with no documented relationships and little metadata. In either case, Greenstone excels at drawing the documents into a cohesive whole.

Library collections are more than just documents, however, and Greenstone cannot function without editorial oversight. For example, Figure 1 shows the statement of purpose and coverage (provided by the maintainer) that must accompany each collection, along with an explanation of how this particular collection is organized (generated from the maintainer's specification).

The maintainer's specifications are stored in the *collection configuration file*, and include the collection name and purpose, the formats of the source documents and their metadata, what browsing facilities should be provided, what full-text search indexes should be provided, and how search results and documents should be displayed. Greenstone uses this file and the source documents to construct the collection. Once the collection is in place, it is easy to add new documents—so long as they have the same format as the existing documents, and the same metadata is provided.

The first stage of the collection building process is to *import* the source documents. In order to accommodate different kinds of source documents, software modules called *plugins* are used to translate the various input formats into a canonical XML format. Plugins already exist for many document formats: plain text; HTML; Email messages; BibTex and Refer bibliographies; Microsoft's RTF, Word, Excel and Powerpoint; Adobe PostScript and Portable Document Format (PDF); Text Encoding Initiative

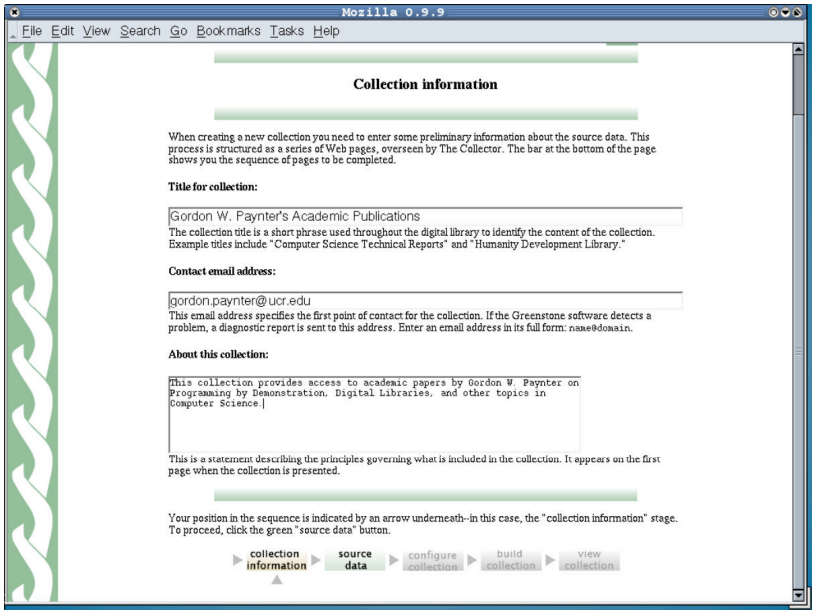


**Figure 5**: Specifying a new collection's name and purpose with The Collector.

(TEI) Lite; GIF, JPEG and other images; and many more. There are also "recursive" plugins, like RecPlug, which reads all the files in a directory and passes them on to the other plugins; and ZIPPlug, which recognizes and decompresses compressed files. The architecture permits plugins for non-textual data, though these are usually linked to textual documents or accompanied by textual descriptions (such as photo captions) to allow searching and browsing.

Once a collection has been *imported*, it must be *built*. Building is the process whereby the search and browse facilities are created from the imported documents. Greenstone provides *maintainability* by creating all searching and browsing structures automatically from the metadata accompanying the imported documents. No links are inserted by hand. This means that when new documents are added to the collection, they are easily incorporated into the search and browse structures by rebuilding the collection. Indeed, for some collections this is done by programs that wake up regularly, scout for new material, import it, and rebuild the indexes—all without human intervention.

## The Collector

*The Collector* is an interface that makes it easy for people to create, update and maintain their own library collections [9]. Users can easily build new collections from material on the Web or from their local files (or both), and collections can be updated and new ones brought on-line at any time. Collections may be built and served locally from the user's own Web server, or (given appropriate permissions) remotely on a shared digital library host.

The Collector has facilities for creating new collections (optionally reusing the structure of existing ones), modifying the structure of a collection, adding new material to a collection, deleting a collection, and writing a collection to a self-contained, self-installing Windows CD-ROM. We will illustrate The Collector by showing how it can be used to create a new collection.

Before beginning, the collection editor must log in. In general, users access the collection-building facility remotely, and build the collection on a Greenstone server. Of course, we cannot allow arbitrary people to build collections (for reasons of propriety if nothing else), so a built-in security system forces people to log in first. After logging in, a new page appears that shows the sequence of steps involved in collection building:

1. Collection information
2. Source data

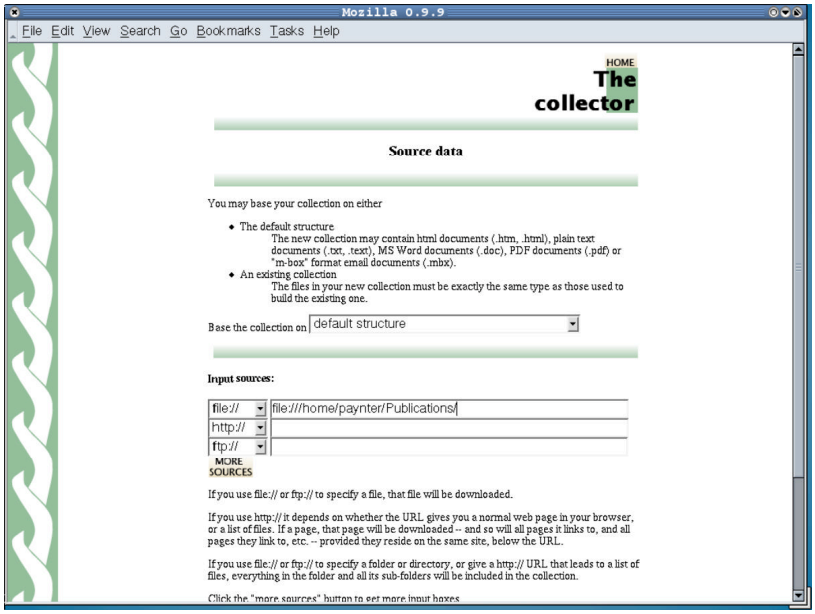


**Figure 6**: Specifying the structure and source documents for a new collection.

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 screenshot in Figure 5 gives the flavour of the interaction. The second step is to say where the source data is to come from. The third is to adjust the configuration options, which requires some understanding of the system but can be skipped by inexperienced users (Greenstone provides a default) and edited later. The fourth step is where all the (computer's) work is done, and incorporates the building and importing steps described above. The final step is to check out the collection that has been created.

Figure 5 shows a typical page. The window has been scrolled down to show that the five steps are displayed as a linear sequence of buttons at the bottom of each page to help users keep track of where they are in the process. The current step is indicted by a triangle beneath its label, and the button that should be clicked to continue the sequence is shown in green; other buttons are grayed out because they are inactive, or coloured yellow if they are already completed (and can be revisited by clicking on them).

The first step is shown in Figure 5. The collection title is a short phrase used throughout the digital library to identify the content of the collection. The email address specifies the first point of contact for any problems encountered with the collection. If the Greenstone software detects a problem, a diagnostic report is sent to this address. Finally, a brief description is required, this will appear under the heading *About this collection* on the collection's home page (e.g. Figure 1).

In step 2, the user specifies the source of the structure and documents that comprise the collection. In top half of Figure 6, a user can elect to create a completely new collection based on the default configuration, or to "clone" an existing configuration file. Creating a totally novel collection can be a major undertaking, and the most effective way to create a new collection is often to take the structure of an existing one and add new documents.

In the lower section of Figure 6, the user is asked to nominate a set of "Input Sources": these are locations where The Collector will look for source documents. In this case, the user has chosen a folder on their hard disk. Any number of input sources can be specified, and there are three kinds of specification:

- a directory name on the Greenstone server (beginning with "file://")

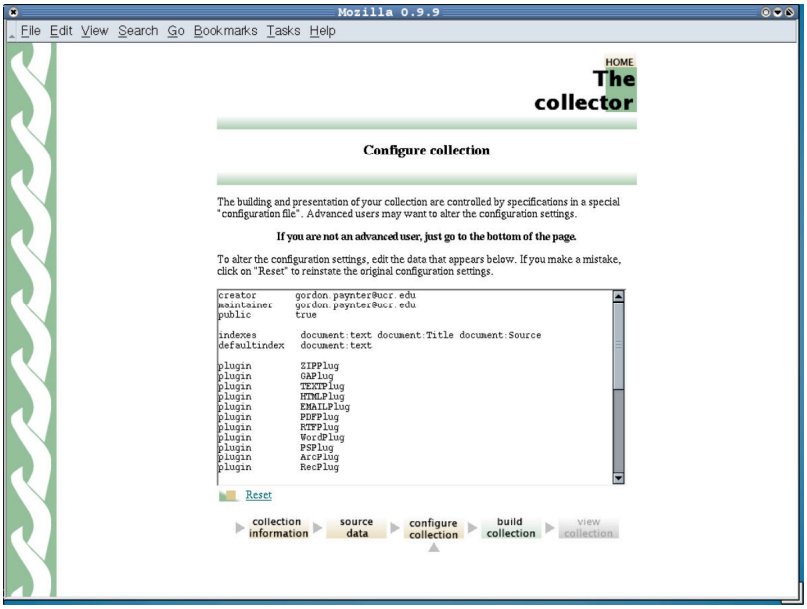


**Figure 7**: Editing the collection configuration file.

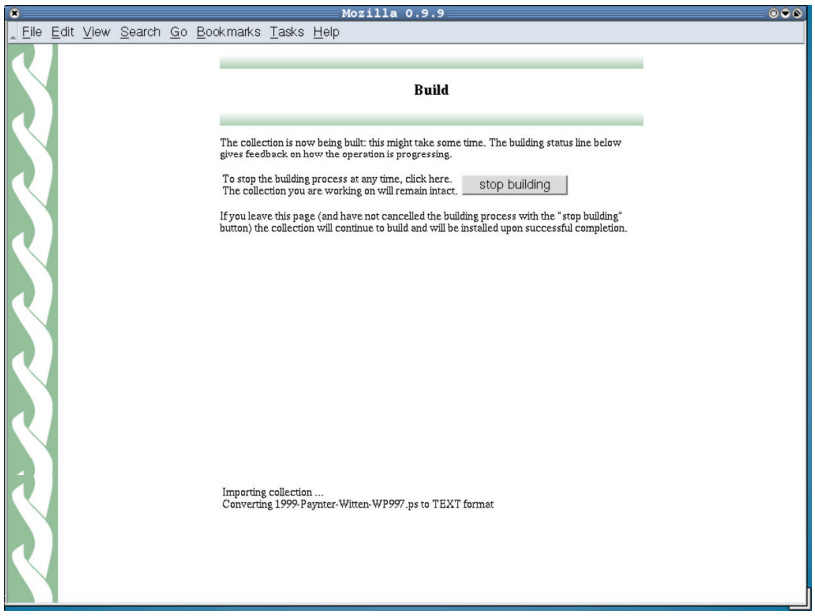


**Figure 8**: Building a collection in the Collector.

- an address for files to be downloaded from the Web (beginning with “http://”)
- an address for files to be downloaded using FTP (beginning with “ftp://”)

Step 3 is to outline the construction and presentation of the collection by editing the collection configuration file. Advanced users may edit the file though the interface in Figure 7; others will use the default settings and proceed directly to the next stage. Figure 7 shows the default settings, which suffice for this collection; they specify three search indexes, a broad selection of plugins, and two browsing facilities (not visible).

Step 4 is where the computer imports and builds the new collection according to the specification. The building stage is potentially very time-consuming. Small collections take a minute or so but large ones (extremely large) can take a day or more. The Web is not a supportive environment for lengthy activity of this kind. Figure 8 shows the build in action: progress is displayed in a status area at the bottom of the

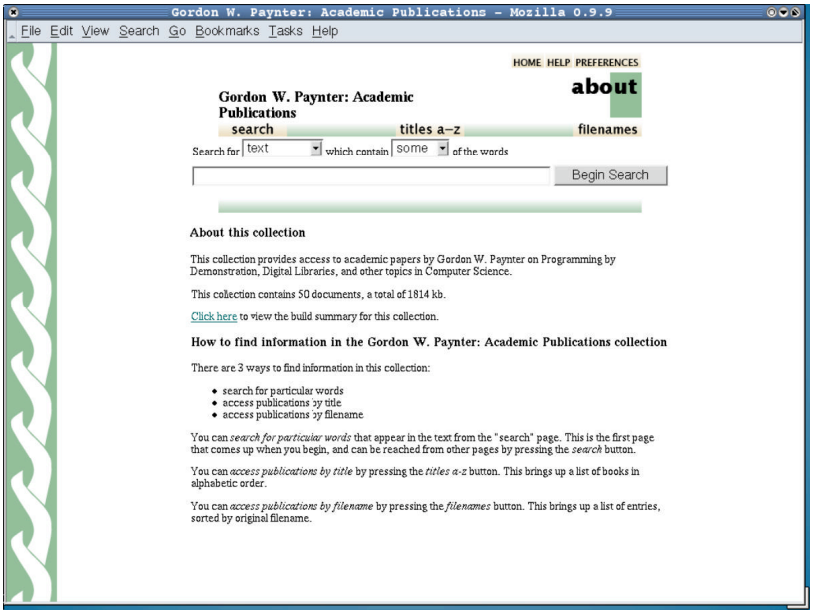


**Figure 9**: The new collection's homepage.

building screen, updated every five seconds. The intention is that the user will monitor progress by keeping this window open in their browser, but there is no reliable way to detect if users leave the building page, and no way to let them return. If the window is closed The Collector will continue building the collection.

Afterwards the contents of the building process is moved to the area for active collections and the final "View collection" button becomes active. By updating the active Web site only at the last moment, we ensure that if an earlier version of the collection already exists, it remains available to users right up until the new one is ready. Persistent document identifiers ensure the changeover is almost always invisible to users. Finally, email is sent to the collection's contact email address and the digital library server administrator notifying them that the collection has been created.

Figure 9 shows the new collection's main page. It has links to two browsing interfaces, one based on document titles and one based on filenames. If the user wants to change these, or any other features of the collection, they can do so by re-entering The Collector and electing to "work with an existing collection".

The Collector was provoked by our work on digital libraries in developing countries, and in particular by the observation that effective human development blossoms from empowerment rather than gifting. Disseminating information originating in the developed world, as most of the above-mentioned collections do, is very useful for developing countries. But a more effective strategy for sustained long-term development is to disseminate the capability to create information collections rather than the collections themselves [7]. This allows developing countries to participate actively in our information society rather than observing it from outside, and will help ensure that intellectual property remains where it belongs—in the hands of those who produce it.

### Distributing a library on CD-ROM

Almost all digital libraries are "located" on the Web, where any user with Internet access can use them. However, it is often desirable to have a collection immediately to hand, and in some cases the Web is not accessible—in developing countries that lack the necessary infrastructure, for example—or the available bandwidth is inadequate. For these reasons, Greenstone digital libraries, including both the collection and a Web server to host them, can be published on CD-ROM.

**Environment: free operating system and development software**

| Program | License | Principal authors | Description |
|---|---|---|---|
| The GNU Compiler Chain (gcc) | GPL | Free Software Foundation | The C and C++ compiler used by most Greenstone developers. http://gcc.gnu.org/ |
| Concurrent Versions System (CVS) | GPL | Brian Berliner<br>Jeff Polk<br>David D. Zuhn<br>Jim Kingdon | The version control system used to organize the Greenstone source code. http://cvshome.org/ |
| Cygwin | GPL, X11 | Cygnus Software<br>Red Hat Software | A library for making GNU software work on Windows computers. http://cygwin.com/ |
| Perl | Artistic License | Larry Wall | The programming Language used in plugins, import scripts, and build scripts http://perl.com/ |
| Apache | Apache Software License | The Apache Project | A Web server used by many Greenstone installations. http://apache.org/ |

**Greenstone components: free software in the main library program**

| Program | License | Principal authors | Description |
|---|---|---|---|
| Managing Gigabytes (mg) | GPL | Tim C. Bell,<br>Ian Witten<br>Alistair Moffat<br>Justin Zobel<br>Stuart Inglis<br>Craig Nevill-Manning<br>Neil Sharman<br>Tim Shimmin. | A program that creates compressed full-text indexes and performs searches using them. http://www.nzdl.org/html/mg.html [6] |
| GNU Database Manager (gdbm) | GPL | phil@cs.wwu.edu | The database used to store document text and metadata. http://www.gnu.org/software/gdbm/ |
| wget | GPL | Hrvoje Niksic | A program for downloading pages from Web sites, used to create collections from URLs. http://www.gnu.org/software/wget/ |
| YAZ | BSD-like | Index Data ApS. | Client and server implementations of the Z39.50 protocol. http://www.indexdata.dk/yaz/ |
| stemmer | GPL | Linh Huynh | An English-language stemmer. http://www.sourceforge.net/projects/stemmers/ |

**Table 1**: Free software directly incorporated into Greenstone (continued overleaf).

Greenstone CD-ROMs operate on a standalone PC under Windows (from Version 3 onwards) and the interaction is identical to accessing the collection on the Web—except that response is more reliable and predictable. If the PC is connected to a network (intranet or Internet), the Web server provided on each CD makes exactly the same information available to others through their standard Web browser. The requirement that Greenstone operate on early Windows systems is one that plagues the software design, but is crucial for many users in developing countries seeking access to humanitarian aid collections. Compression ensures that the greatest possible volume of information can be packed on to a CD-ROM.

**Plugins: free software used to import source documents**

| Program | License | Principal authors | Description |
|---|---|---|---|
| GhostScript | GPL | Peter Deutsch | Interpreter for Adobe Postscript documents, used by Postscript plugin.<br>http://www.gnu.org/software/ghostscript/ |
| Kea | GPL | Eibe Frank<br>Gordon Paynter | Automatic keyphrase extraction program used to generate metadata.<br>http://www.nzdl.org/Kea/ |
| pdftohtml | GPL | Gueorgui Ovtcharov<br>Rainer Dorsch | Converter for Adobe PDF documents, used by the PDF plugin.<br>http://www.ra.informatik.uni-stuttgart.de/~gosho/pdftohtml/ |
| rtftohtml | Public Domain | Chris Hector | Converter for RTF documents, used by the RTF plugin. |
| TextCat | GPL | Gertjan van Noord | A tool for automatically detecting languages and document encodings.<br>http://odur.let.rug.nl/~vannoord/TextCat/ |
| wvWare | GPL | Caolan McNamara<br>Dom Lachowicz<br>Martin Vermeer | Converter for Microsoft Word documents, used by Word plugin.<br>http://www.wvware.com/ |
| xlhtml | GPL | Steve Grubb | Converter for Microsoft Excel and Powerpoint documents, used by the Excel and Powerpoint plugins.<br>http://sourceforge.net/xlhtml/ |
| XML::Parser Perl module | Artistic License | Larry Wall<br>Clark Cooper | A module that parses XML documents, used to read and write Greenstone's internal XML document format.<br>http://www.netheaven.com/~coopercc/xmlparser/ |

**Table 1** (continued).

Digital libraries is an application area where free software licenses are inherently superior to their commercial counterparts. An aid organization distributing 50,000 copies of a medical library to towns and schools in a developing country is seldom in a position to pay licensing fees for so many copies of the software. And even if commercial software was provided at no cost, copyright law forbids the recipients from making further copies for other needy schools and towns. If the library is distributed under a license like the GNU GPL, however, each recipient is free to make additional copies—or new, derived works—for those who need them. This scenario illustrates the difference between the two types of free software: even when provided at no cost, traditional commercial software places limits on its users' freedoms. Such software is said to be "free" as in "free beer", but not "free" as in "free speech".

## Open source projects used by Greenstone

Greenstone is free software (in both the "free beer" and "free speech" senses) distributed under the GNU General Public License. The preceding sections have surveyed some (but not all) of the software's many features; it is a large and complex program—far too extensive, you may suspect, to have been created from scratch by one small group of programmers. In fact, this suspicion is well-founded. Greenstone has, in its turn, benefitted from the contributions made by uncountable programmers to other open-source projects, contributions that Greenstone has incorporated, altered, and extended in building the present system.

Table 1 shows a selection of the programs used directly by Greenstone. These projects range from the very large—like the GNU Compiler Chain and the Perl programming language—to small, specialized utilities like the stemmer and the TextCat language identifier.

This list is not, and can never be, exhaustive. For one thing, only a few projects are listed, and many crucial efforts are omitted (the Linux Kernel for example!). Further, only a few of the authors are credited: in reality, many of these projects (like Greenstone itself) have dozens or hundreds of contributors. And like Greenstone, many of these projects in turn incorporate smaller, earlier components; work that goes uncredited here, yet is essential to Greenstone's operation. (James Clark's "Expat" XML library, for example, is included by at least two of the programs in Table 1.) Indeed, it is

almost unfair to single out these few projects for attention, when so many others must go unacknowledged.

## Conclusion

The Greenstone Digital Library Software has been successful in terms of both technical accomplishment and user adoption. For our researchers, it provides a consistent work environment and a constant stream of new problems, new data and new opportunities to experiment—recent work has focussed on automatic classification, automatic metadata creation, musical retrieval, and phrase-based browsing, among other topics. For our users, both institutional and individual, it is a stable, usable, documented system for distributing collections that can import existing data and metadata, and that exploits—but is not limited to—the reach of the Web.

Greenstone 2.0 was released under the GNU General Public License in September 1999. It was a major rewrite of the earlier code, and has been upgraded many times since: most recently, in April 2002 as Version 2.39. This version was a landmark release: it is being adopted by UNESCO, who are collaborating with the Human Info NGO to translate the manual into French and Spanish. UNESCO has also provided financial support for Greenstone's development, as have the New Zealand Government and The University of Waikato. Meanwhile, back at the New Zealand Digital Library Project, we are working on future releases of Version 2 and looking ahead to Version 3.

Greenstone's status as free software has been pivotal to its success. It has given us access to a large and constantly improving pool of software to reuse, without the expense of licensing fees or royalties. And by making the software and source code freely available, we have gained a base of users who have downloaded the software, used it, redistributed it, and sent us comments, bug reports and advice. A smaller but still significant group of users has downloaded the source code itself, fixed bugs, extended it to suit their own circumstances, and contributed their work back to the project.

Our experience with free software has not been completely without problems. In some cases, the programs we have incorporated have been of poor quality; however, the developers are usually receptive and responsive to feedback, and even when they are not, it is often easier to extend their work than to develop completely new programs. Interoperability with closed-source environments has also been problematic: we have been unable to find a good, free Windows installation program, for example, and users have reported incompatabilities when compiling the source code on Sun's proprietary Unix operating system. On the balance, however, these are minor problems.

We have encountered some resistance to the free software model—starting with ourselves. The initial decision to release Greenstone under the GPL was a difficult one because we felt we would be losing control of the work. In practice, all we have lost is the ability to choose who else should use the work. We have also encountered resistance to the *use* of our software: some prefer to pay for software rather than accept it for free. There are a number of possible reasons for this preference: they perceive the quality is lower, or the risk greater, or their legal recourse less assured. In fact, commercial vendors tend to be less responsive, and disclaim all liability for their software. But the perception that anything "free" is worthless persists: in one extreme case, an organization turned to Greenstone after spending hundreds of thousands of pounds on a commercial alternative that did not work, but refused to accept it without offering payment (which we were glad to receive).

Greenstone is evolving still. Only through an international cooperative effort will digital library software become sufficiently comprehensive to meet the world's needs. Currently, Greenstone is used at sites in Canada, Germany, New Zealand, Romania, Russia, the UK, the US, and elsewhere; collections range from newspaper articles to technical documents, from educational journals to oral history, from visual art to folksongs. The software has been used to build collections in many different languages, and to create CD-ROMs that have been published by the United Nations and other humanitarian agencies in Belgium, France, Japan, and the US for distribution in developing countries. Greenstone is mature, documented free software, supported by its creators and its large (and growing) community of users.

## Further Information

The Greenstone software and manuals (the Installer's Guide, the User's Guide [8], and the Developer's Guide) are available from the Web site (http://www.greenstone.org/). The site also contains the Greenstone FAQ, instructions for joining the Greenstone mailing lists, and infomration on how to prepare collections from paper source documents [2]. A forthcoming book, *How to build a digital library*, by two of the authors of the present article, describes the challenges of creating digital libraries in a more general way, and also describes Greenstone in more detail [10].

## References


1. Humanity Libraries (1998) *Humanity Development Library*. CD-ROM produced by the Global Help Project, Antwerp, Belgium.

2. Loots, M., Camarzan, D. and Witten I. H. (2001) *From Paper to Collection*. New Zealand Digital Library Project, New Zealand.

3. PAHO (1999) *Virtual Disaster Library*. CD-ROM produced by the Pan-American Health Organization, Washington DC, USA.

4. UNESCO (1999) *SAHEL point DOC: Anthologie du développement au Sahel*. CD-ROM produced by UNESCO, Paris, France.

5. UNU (1998) *Collection on critical global issues*. CD-ROM produced by the United Nations University Press, Tokyo, Japan.

6. Witten, I.H., Moffat, A. and Bell, T. (1999) *Managing Gigabytes: compressing and indexing documents and images* (second edition). Morgan Kaufmann, USA.

7. Witten, I.H., Loots, M., Trujillo, M.F. and Bainbridge, D. (2000) *The promise of digital libraries in developing countries.* Communications of the ACM, 55 (5) 82-85, May.

8. Witten, I.H., McNab, R.J., Boddie, S.J. and Bainbridge, D. (2001) *Greenstone: User's Guide*. New Zealand Digital Library Project, New Zealand.

9. Witten, I.H., Bainbridge, D. and Boddie, S.J. (2000) *Power to the people: End-user building of digital library collections.* Proc. Joint Conference on Digital Libraries, 94-103, Roanoke, VA, June.

10. Witten, I.H. and Bainbridge, D. (2002) *How to build a digital library*. Morgan Kaufmann, USA.