

# The Greenstone digital library software

**Ian H. Witten**

New Zealand Digital Library Project  
Department of Computer Science  
University of Waikato, New Zealand  
ihw@cs.waikato.ac.nz

Digital libraries are large, organized, focused collections of information. The Greenstone software is intended to help people design and build such collections quickly and easily. Collections may be large—some comprise Gbytes of text; others include many millions of short documents. Additionally, far larger volumes of information may be associated with a collection—typically audio, image, and video. Greenstone is international and multilingual: it is widely used in many different countries; interfaces and collections exist in many of the world’s languages; and it is being distributed by UNESCO as part of the “Information for All” programme. It is multiplatform: it runs on all Windows, Unix, and Macintosh OS/X systems. Users access collections over the web, or from self-contained, self-installing CD-ROMs.

A basic Greenstone collection of new material with a standard look and feel can be set up in just a few minutes (this operation is followed by the mechanical process of building the collection, which may take from a few moments for a tiny collection to several hours for a multi-Gbyte one; perhaps a day if it involves many different full-text indexes.) However, digital library collections may be customized in a wide variety of different ways, and some collections, particularly large ones, have their own idiosyncratic requirements. The design and debugging process for sophisticated collections can take days—longer if iterative redesign guided by usability testing is involved. Of course, as the number of collections grows and the variety of styles increases, it becomes more likely that some existing collection will match new requirements. With Greenstone, it is easy to reuse a collection design.

The facilities that a collection provides, and the user interface for searching and browsing, are highly customizable at many different levels. Users can easily specify what document formats will be included (e.g. HTML, Word, PDF, PostScript, PowerPoint, Excel); where already-available metadata (if any) comes from (e.g. XML files, OAI archives, Latex bibliographies); what searchable indexes will be provided (e.g. full text, perhaps differentiated by language, and certain metadata such as titles); and what browsing structures will be available (e.g. list of authors, titles, classification hierarchy).

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 of the library. It is the key to providing searching and browsing facilities. Greenstone incorporates a range of mechanisms at different levels to capitalize on this.

## Designing collections

With Greenstone, users design their collections individually—typically by taking an existing collection that closely matches their needs and adapting its structure as necessary. The resulting design is recorded in a short file called the “collection configuration file.” It specifies such things as the collection’s title, the creator’s email address, a description of the purpose and principles governing what is included, what input file types should be included in the collection, where the metadata comes from and what form it takes, and how the collection will look to the user. Most of the customization that non-programming users perform in Greenstone takes place in this file. It depends crucially on the availability of metadata, and the structures defined are only produced if appropriate metadata is provided.

**Searching** the full text of all documents in the collection is a basic facility, included by default in all collections. Collection designers can determine whether searching should be on a paragraph, section, and/or whole-document level (this affects the scope of matches to a given query). They can also ask for full-text indexes to be built on metadata items (e.g. titles, authors). They can split the collection into sub-collections that can each be searched individually, or use language metadata to restrict searches by language.

**Browsing** capabilities vary from collection to collection depending on the metadata available and the facilities that the collection designer wishes to provide. Greenstone includes predefined browsing structures based on certain kinds of metadata. Any textual metadata can be presented as an alphabetically sorted list, which can optionally be tabbed into alphabetic ranges that are chosen automatically to include a reasonable number of documents in each range. Date metadata can be presented in a list that allows selection by year and month. Metadata with hierarchical structure, such as library classifications, can be presented as a tree whose nodes open to reveal the data beneath. In this case the user must provide an auxiliary file giving labels for intermediate nodes of the hierarchy (e.g. subject headings corresponding to each classification number).

**Formatting** statements are used to control the presentation to the user of each “screen” that the system generates. They determine how target documents are displayed—whether they are preceded by title, for example, or indented. They control the search results page, where they determine what metadata is presented as a “snippet” that stands for matching documents, whether it should be preceded by an appropriate document icon, whether there should be a hyperlink and what is its target. In collections that include different versions of a document (e.g. Word and HTML extracted from it), links to both versions can be presented in the search results list so that users can determine which one to read. Format statements also apply to the browsing mechanisms mentioned above.

## Other features

Greenstone is multilingual: currently there are interfaces in Arabic, Chinese, Czech, Dutch, French, Galician, German, Hebrew, Indonesian, Italian, Japanese, Kazakh,

Maori, Portuguese, Russian, Spanish, Thai, Turkish and English. 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 text fragments, a task that has been performed many times by people with no expertise in Greenstone.

Installing Greenstone and building collections requires no more than ordinary computer literacy skills. However, those with greater knowledge can do more. Knowing HTML, one can hook Greenstone widgets like the full-text search mechanism or browsers into one's own pages. Knowing JavaScript, one can incorporate browsing mechanisms such as image maps; and using Perl one can add entirely new browsing facilities such as stroke-based or Pinyin browsing for Chinese. Greenstone is designed on the philosophy that simple things should be simple, while complex things should be possible.

Several advanced facilities are included that can be incorporated into a collection by a small addition to its configuration file. Here are three examples. One can add the ability for users to browse around a phrase hierarchy that has been extracted automatically from the full text of a document collection, which is designed to resemble a paper-based subject index or thesaurus. Acronyms and their definitions can be automatically extracted from the full text of documents and presented to the user along with their definitions and the places they occur. The language of each document can be automatically identified and used to provide separate full-text indexes for each language.

One challenge with a rich system like Greenstone is the need for good, up to date, documentation. In fact, from a user's point of view the chief bottleneck when customizing collections is documentation, not the facilities that are provided. Collection builders often need advice and assistance from others in order to learn how to tailor the software to meet ever-changing new requirements. There is a lively email discussion group for assistance with Greenstone; participants hail from over 40 different countries. There is extensive documentation in English, French, Spanish, and Russian, and the system is comprehensively described in a recent book entitled *How to build a digital library* (Witten and Bainbridge, Morgan Kaufmann, San Francisco, 2003). UNESCO has begun a series of regional training programs in the use of Greenstone, and it is included in the FAO's new Information Management resource kit.