Greenstone: Open Source DL Software

Ian H. Witten, David Bainbridge and Stefan Boddie

Computer Science Dept, Waikato University, New Zealand
E-mail: {ihw, davidb, sjboddie}@cs.waikato.ac.nz

Greenstone is a comprehensive system for constructing and presenting collections of thousands or millions of documents, including text, images, audio and video. Libraries contain many collections, individually organized—though they bear a strong family resemblance. Easily maintained, collections can be augmented and rebuilt automatically.

Finding text information in the library

Greenstone constructs full-text indexes from the document text, and from metadata elements such as title and author. Indexes can be searched for particular words, Boolean combinations, or phrases, and results are ranked by relevance or sorted by a metadata element.

Browsing involves hierarchical lists that the user can examine interactively. Metadata (based round the Dublin Core) is the raw material for browsing, and must be provided explicitly or be derivable automatically from the source documents. Different collections offer different searching and browsing facilities. Indexes for both are constructed during a "building" process, according to information in a collection configuration file.

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

Plugins provide flexibility

Source documents come in a variety of formats, and are converted into a standard form for indexing by "plugins." Plugins distributed with Greenstone process HTML, WORD and PDF documents, Usenet and E-mail messages; new ones can be written for different document types. To build browsing structures from metadata, an analogous scheme of "classifiers" is used. These create browsing indexes of various kinds: scrollable lists, alphabetic selectors, dates, and arbitrary hierarchies.

Multimedia and multilingual documents

Collections can contain text, pictures, audio and video. Currently non-textual material is either linked into the textual documents or accompanied by textual descriptions (such as figure captions) to allow full-text searching and browsing. The architecture, however, permits implementation of plugins and classifiers for non-textual data.

Unicode is used throughout, allowing any language to be processed and displayed in a consistent manner. Collections have been built containing Arabic, Chinese, English, French, Mäori and Spanish. Multilingual collections embody automatic language recognition, and the interface is available in all the above languages (and more).

Usage

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

The New Zealand Digital Library (nzdl.org) provides many example collections, including historical documents, humanitarian and development information, technical reports and bibliographies, literary works, and magazines. Other examples appear in a companion article[1] and sidebar.[2]

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

---

[1] Witten *et al.*, "The promise of digital libraries in developing countries."

[2] Apperley *et al.*, "Niupepa: An historical newspaper collection."