



Building Digital Library Collections with **Greenstone**

Ian H. Witten

New Zealand Digital Library Project
Computer Science Department
University of Waikato
New Zealand
nzdl.org
2007

Introduction

This tutorial is a practical, hands-on, laboratory-style workshop in which attendees build their own digital library collections using the Greenstone digital library software, a comprehensive, open-source system for constructing, presenting, and maintaining information collections. Collections will be built from HTML documents; Word, PDF and PostScript documents; images in various formats; MP3 and MIDI audio; MARC records; and more. For each collection, various different full-text search indexes and metadata-based browsers will be created. Interoperability will be demonstrated with MARC, METS, MODS, DSpace, and Fedora.

About the Greenstone Digital Library Software

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.”

The Greenstone software runs under Unix, Windows and Mac (OS/X), and is issued as source code under the GNU public license. Attendees will learn enough to install the software, set up a digital library system, build their own collections, and customize them. Those with programming skills should be able to extend and tailor the system extensively. Moreover, all attendees will be equipped with extensive course material that is freely redistributable.

Workshop content

The course is centered upon Greenstone’s “librarian” interface. This facility allows users to gather together sets of documents, import or assign metadata, build them into a Greenstone collection, and serve it from their web site. It supports seven basic activities: opening an existing collection or defining a new one; copying documents into it, with metadata attached (if any); mirroring documents from the Web if required; enriching the documents by adding further metadata to individual documents or groups; designing the collection by determining its appearance and the access facilities it will support; building it using Greenstone; and previewing the newly created collection from the Greenstone home page.

The interface explicitly supports four levels of user: Library Assistants, who can add documents and metadata to collections, and create new ones whose structure mirrors that of existing collections; Librarians, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions); Library Systems Specialists, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl scripts); and Experts, who can perform all functions.

Collections built with Greenstone automatically include effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. They are easily maintainable and can be rebuilt entirely automatically. Searching is full-text, and different indexes can be constructed (including metadata indexes). Browsing utilizes hierarchical structures that are created automatically from metadata associated with the source documents. Collections can include text, pictures, audio, and video. The interface to collections can be extensively customized. Documents can be in any language: the interface has been translated into about thirty languages.

Although primarily designed for Web access, collections can be made available, in precisely the same form, on CD-ROM or DVD. The system is extensible: software “plug-ins” accommodate different document and metadata types.

Intended audience

The tutorial is designed for those who want to build their own digital library but do not want to write their own software. It is intended for librarians and other information workers who are interested in building their own digital collections.

The Greenstone Librarian Interface is designed for end users. No programming ability is required. Attendees should be familiar with HTML and the Web, and be aware of representation standards such as Unicode and Dublin Core.

Topics

Overview

- What does Greenstone do?—Examples
- Platforms and installation
- Documentation and help

Building collections

- Building collections from HTML, Word/PDF, Images
- Creating Greenstone CD-ROM
- Adding and using metadata
- Browsing classifiers, search indexes
- Multimedia/Scanned Image collections

Customizing

- Under the hood: collection configuration file
- Customizing with macros
- Personalizing your home page
- Different interface languages
- Examples of what others have done

Reaching out

- Bibliographic (MARC) data: metadata crosswalks
- OAI: serving and ingesting
- Generating METS; using MODS
- Interoperating with DSpace
- Searching Fedora repository using Web Services

Concluding discussion

Instructor

Dr. Ian H. Witten (ihw@cs.waikato.ac.nz)

Workshop material

1. Workbook containing
 - Introduction
 - Greenstone Digital Library Software factsheet
 - PowerPoint slides for all workshop sessions
 - Laboratory exercises
 - How to format the output of your collection
2. Tutorial CD-ROM containing
 - Greenstone software
 - Documented example collections
 - Greenstone Wiki collection
 - All language interfaces (about 40)
 - Java runtime environment
 - Ghostscript
 - ImageMagick graphics package
 - Full documentation (4 manuals)
 - Installer that installs all of the above
 - Tutorial exercises
 - Sample files for tutorial exercises

Sample files for tutorial exercises

Small set of HTML files (*simple_html*)
Word and PDF documents (*Word_and_PDF*)
Difficult PDF documents (*difficult_documents*)
Image files (*images*)
Large set of HTML files (*tudor*)
MARC records (*marc*)
Multimedia collection (*beatles/advbeat_large*)
Small multimedia collection (*beatles/advbeat_small*)
Scanned image collection (*niupepa*)
Files exported from OAI (*oai*)
Files exported from DSpace (*dspace*)
CDS/ISIS files (*isis*)
Files for interface customization exercise (*custom*)

Further documentation on the Tutorial CD-ROM

Greenstone Digital Library Installation Guide
Greenstone Digital Library User's Guide
Greenstone Digital Library Developer's Guide
Greenstone Digital Library: From Paper to Collection