BUILDING DIGITAL LIBRARY COLLECTIONS WITH GREENSTONE

Ian H. Witten

OVERVIEW

This hands-on workshop will demonstrate how to build a variety of different kinds of digital library collections with 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. We will discuss support for Dublin Core, OAI, and METS, and show how collections can be exported to and from DSpace.

Attendees will receive this open source software on a CD-ROM that we will provide, along with various sample files, so that they can take it home and install it on their own machines.

SUBJECT MATTER

The workshop 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.

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.

LENGTH

2 ½ days.

SCHEDULE

Greenstone: Overview of features, capabilities, applications
Platforms, installation, configuration; accessing example collections
Librarian interface: Building collections, adding and using metadata
Librarian interface: indexes, partitioned indexes, browsing classifiers
Librarian interface: customizing collections, advanced features
Creating Greenstone CD-ROMs
Standards: Dublin Core, OAI, METS; interoperation with DSpace
Multilingual support; interface languages
Concluding discussion

INTENDED AUDIENCE

The workshop 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 students, researchers, and practitioners, in any area of IR, who are interested in the details of building digital libraries.

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.

MATERIALS

Participants will receive:
1. Tutorial CD-ROM containing
   Greenstone software
   Full documentation (4 manuals)
   Sample collections

2. Participants will receive an extensive (350 pp) handout that includes
   PowerPoint slides for the tutorial
   Laboratory exercises plus screen snapshots
   Greenstone Digital Library User's Guide

3. Further documentation is available on the Workshop 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