Building digital libraries using open source software

Learn how to build your own digital library with the Greenstone digital library software, an open-source system for managing collections

Ian H. Witten
Computer Science Department
Waikato University
New Zealand
www.nzdl.org
www.greenstone.org

"Greenstone turns a ragtag menagerie of documents in various formats into an easy-to-use collection that can run on a standalone laptop in a Ugandan village’s information center”

Agenda
- What is a digital library?
- What does Greenstone do?
- Example Greenstone collections
- Searching and browsing
- Metadata extraction plugins
- Documentation and help
- Greenstone Librarian Interface
  - A simple collection
  - An image collection, adding metadata
  - Exploring further Greenstone features
- Different interface languages
What is a “digital library”?

Collection of digital objects (text, video, audio) along with methods for access and retrieval, [user] and for selection, organization, and maintenance [lib]

- Traditional user/librarian distinction is blurred
- Computers make information active
- Kitchens for knowledge preparation
- WWW ≠ DL!—organization, selectivity
- Nice Web site ≠ DL!—import new documents easily

Data = smallest discernible difference (change of state)
Information = difference between data and your expectations
Knowledge = accumulation of your set of expectations
Wisdom = the value attached to knowledge

Example

Humanity Development Library
for sustainable development and basic human needs

- 160,000 pages
- 30,000 images
- 800 books
- 430 magazines
- 340 kg
- US$20,000
- CD-ROM
- US$6
- Win3.1x upward
- Stand-alone
- and intranet server
- Web browser user interface

Global Help Project, Antwerp (+ UN agencies)
What is the Greenstone software?

- Software suite for building, maintaining, and distributing digital library collections
- Comprehensive, open-source
- Developed by New Zealand Digital Library Project at the University of Waikato
- Distribution and promotion partners:
  - UNESCO
  - Human Info NGO, Belgium
  - NCSI, Bangalore; UCT, Cape Town; Dakar, Senegal; Almaty, Kazakhstan; ...
  - You!

UNESCO: Distributing Greenstone DL software

Sustainable development

“Give a man a fish, feed him for a day
Teach a man to fish, feed him for life”

Greenstone software on CD-ROM

- GNU licensed
- Fully documented ... in English/French/Spanish/Russian
- Language interfaces ... Arabic Chinese Czech ... Thai Turkish
- Unix/Windows/Mac OS-X
- Trivial to install
- GUI interface for gathering, enriching, building ...
- Serve collections on Web or write them to CD-ROM
- Document formats: HTML, Word, PDF, PS, plain text, e-mail
- Metadata formats: XML, DC, OAI, MARC, ...

download from http://greenstone.org
What we wanted

- “Collections” of digital material
- Individualized, depending on metadata etc
- Up to several Gb of text ...
- ... + associated images, movies, whatever
- Fully searchable
- Served on WWW, or published on CD-ROM
- Multi-platform (Unix + all Windows + Mac)
- Multi-format documents and metadata
- Multi-lingual: documents and interfaces
- Multimedia
- Metadata: standard and non-standard

Greenstone DL software

Access
- Accessible via any Web browser
- Server runs on Windows and Unix
- Collections can be published on CD-ROM

Searching/browsing
- Full-text and fielded search
- Flexible browsing facilities
- Metadata-based (Dublin Core)
- Collection-specific
- Hierarchical phrase browsing supported
- Creates all access structures automatically

Extensible
- Plugins — new document, metadata formats
- Classifiers — new metadata browsers

Multilingual
- Documents and interfaces
- Chinese, Arabic, Maori, Russian etc (+ European)
- Multimedia: video, audio collections exist
Some places where it’s used

- **Individual sites ...**
  - China: Tang Dynasty poetry
  - Germany: Digitale Bibliothek Information und Medien
  - India: Archives of Indian labour
  - India: collections in Hindu and Kannada
  - Russia: Mari El Republic government information
  - UK: Gresham College Archives, Kids’ Digital Library
  - US: Gutenberg, Aladin, NY Botanical Garden, ...

- **Larger projects**
  - New Zealand Digital Library Project
  - Human Info NGO, Belgium: Humanitarian collections
  - Singapore/Australia: Global Library Services Network
  - UNESCO distributes Greenstone Digital Library Software

- **New York Botanical Garden**
  - Rare 19th century works on American trees
  - Gorgeous full-color plates
Chinese documents
(pictures of text)
+ Chinese interface

Peking University Library

Chinese
(Chinese & English interfaces)

Classic Chinese literature
Hindi/Kannada

Collections: on the Web

nzdl.org
(demo, not service)
Searching and browsing

- Searching
- Metadata-based browsing

Subject | Title | Publisher | “HowTo”
---|---|---|---
Dublin Core | ad hoc
What are documents?

- Hierarchical document model
  
  Sections, subsections, ..., paragraphs

- Metadata
  
  Dublin core, for searching and browsing

- Multimedia
  
  Picture, Voice, Music

- Multi-language documents
  
  Maori, French, Arabic, Chinese, ...

- Multi-language interfaces
  
  French, Chinese, ...

Hierarchical document model

- Metadata specified at any level
  
  Title metadata
Searching and browsing

- Searching
- Metadata-based browsing

Subject  Title  Publisher  “HowTo”

Dublin Core  ad hoc

Multiple search indexes

text  metadata
Collection-dependent

Multilingual searching
Ranked OR
Boolean AND

Full Boolean queries, plus other search preferences
Form search

Form search
Advanced form search
Browsing using classifiers

AZList classifier (Title metadata)

DateList classifier (Date metadata)
List classifier (Howto metadata)

Hierarchy classifier (Subject metadata)
Custom classifier (Title metadata)

Simple variant of AZList (a few lines of PERL)
Metadata extraction plugins

Acronym extraction plugin

Language identification plugin
Email plugin

Phrase hierarchy extraction
Different document types and formats: plugins and classifiers

Plugins
- format-specific parsing of source documents (and metadata specs)
- pipeline: files are passed to each plugin in turn
- Document plugins: TEXT, HTML, EMAIL, WORD, RTF, PDF, PS, PPT, EXCEL
- Metadata plugins: XML, OAI, MARC, Latex, Refer

Classifiers
- Create browsing indexes
- VLists, HLists, and DateLists
- Hierarchical structure of lists
- List
  - one-level: single VList
  - two-level: HList and Vlist
  (e.g. SectionList, AZList, DateList)

Getting information

- Available at: www.greenstone.org
  - Software
  - Demo collections
  - FAQ
  - Examples
  - Documented example collections

- Documentation:
  - Installer’s Guide
  - User’s Guide
  - Developer’s Guide
  - From Paper to Collection

- Mailing lists:
  - Greenstone User’s List
  - Greenstone Developer’s List
Documentation and Help

- greenstone.org
  - Examples
  - support
    (+ join mailing list)
  - Configuration files for nzdl.org collections
- FAQ
  - General info section
- nzdl.org
  - Documented example collections

Initial view of the Librarian Interface
We will work in “Librarian” mode

Starting a new collection
Filling out information about the collection

Selecting a metadata set ...
... use Dublin Core

... in different languages, e.g. Marathi
The Gather panel

Copying some files into the collection
Inspecting the files in the right-hand panel

The Create panel
Partway through the building process

Collection has been built
The collection’s initial page

titles a-z page
Bilbo Baggins character page

Initial view of the Enrich panel
Selecting `bilbo.html` and scrolling down to view extracted metadata

Building an image collection

Base new collection on existing ‘image-e’ collection
Copy the sample photo images

Building the photo image collection
Preview the new photo image collection

Browse the new photo image collection
View metadata extracted from Ascent.jpg

Adding a metadata set to the collection
Choose dublin.mds (Dublin Core)

Entering *dc.Title* field for *Ascent.jpg*
Customizing the collection

The design panel

The Design panel’s Format Features zone
Changing the format statement to take advantage of the new metadata

The resulting change in the collection
Configuring arguments for ImagePlug
(set thumbnailsize to 50)

Thumbnail preview now only 50 pixels wide
Add metadata to describe Ascent.jpg

Modifying the format statement to use the new dc.Description metadata
PREVIEWING THE EFFECT OF THE MODIFIED FORMAT STATEMENT

ADDING A NEW CLASSIFIER FOR BROWSING

DESCRIPTION METADATA
Setting options for the classifier
(some of these options are unavailable in Librarian mode)

New ‘descriptions’ browse feature
Adding searching index based on dc.Description

Using searchable index built on dc.Description
Search descriptions for mention of ‘autumn’

Controlling the building process

Using maxdocs to limit number of documents built into collection
Collection now builds much more quickly ...

... but it only includes 3 documents
Exploring further Greenstone features

To create a new collection fill out the fields below:

Collection title: TudorX

Description of content:
Experimentation with Phind and partitioned collections based on sample HTML web site of Tudor England.

Base this collection on: -- New Collection --

Building basic collection based on large_html sample files
The basic collection

Adding extra features

Adding a Phind classifier
The refined browsing classifiers

“phrases” button appears in navigation bar
Interacting with Phind

Partitioning search indexes

Defining three filters for partitioning
(this facility is unavailable in Librarian mode)
Assigning partition names

Collection with partitioned index
Search relatives partition for ‘mary’

Search result for relatives partition
Adding hierarchical metadata

Metadata applies to all documents in folder
Choosing the Hierarchy classifier

Filling out *hfile* and *buttonname* fields
Browsing the manually assigned hierarchy
Starting a new collection based on MARC records

It’s OK to have no initial metadata sets
Sample MARC dataset copied into collection

Cull extraneous plugins and add MARCPlug
Default look of MARC-based collection

Default titles a-z for MARC collection
Document display in MARC format

Finding George Martin
Setting **AZCompactList** options for **Subject** metadata

(some of these options are unavailable in Librarian mode)

**subjects** button now appears in navigation bar
Top level of the subject hierarchy

Opening ‘Orchestral Music, Arranged’ in subject hierarchy
Extending format statement to include number of entries with bookshelves

Enhanced format statement for subject classifier
Switching on ‘form’ search type to support fielded queries

Querying for text including ‘peter best’ and title including ‘silver’
Brand collection with an image

Copying logo to Beatles collection

Adding image to the collection
Collection home page with image

Exporting to CD-ROM
Choosing tudorX as the collection to export

Different interface languages

“About” page:

English
Macro file

package Globals
    _header_ { The New Zealand Digital Library Project }
    _content_ { Oops. If you are reading this then an error has occurred in the runtime system. }
    _footer_ { Powered by &lt;a href="www.greenstone.org">Greenstone</a&gt;.

package query
    _content_ { If _cgiargqb_eq "large", _largequerybox__, _normalquerybox__ ... }

# ... the macro descriptions for _largequerybox__, _normalquerybox__ ... and other nested macros are omitted
# ... the macro descriptions for _largequerybox__, _normalquerybox__ ... and other nested macros are omitted

# ... the macro descriptions for _largequerybox__, _normalquerybox__ ... and other nested macros are omitted

# ... the macro descriptions for _largequerybox__, _normalquerybox__ ... and other nested macros are omitted

... and so on

... and so on

- Language specific text fragments actually gathered up in one file: english.dm, french.dm etc.
- Language specific fragments further divided: english.dm, english2.dm, french.dm, french2.dm
Maintenance issues

- Over 600 linguistic fragments per language in user interface
- No one person knows all languages
- No overlap between translators and developers
- Software developers and translators geographically dispersed
- Translations must keep pace with software developments

Translator’s interface
New phrases, and superseded ones

Generating Icons

- Use image editor
- Translate the content of the tab or icon
- Retain the same file names
- Place all tabs and icons into single folder and copy the folder to gsdl/images

<table>
<thead>
<tr>
<th>English</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>titles a–z</td>
<td>शीर्षक</td>
</tr>
<tr>
<td>ttitlgr.gif</td>
<td>ttitlgr.gif</td>
</tr>
<tr>
<td>titles a–z</td>
<td>शीर्षक</td>
</tr>
<tr>
<td>ttitlof.gif</td>
<td>ttitlof.gif</td>
</tr>
<tr>
<td>titles a–z</td>
<td>शीर्षक</td>
</tr>
<tr>
<td>ttitlon.gif</td>
<td>ttitlon.gif</td>
</tr>
</tbody>
</table>