Building DLs Using Open Source Software

Ian H. Witten

New Zealand Digital Library Project Department of Computer Science University of Waikato, New Zealand


This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
Browsing around a digital library

Greenstone: Open source system for creating and delivering digital library collections

Ian H. Witten
New Zealand Digital Library Project
Computer Science Department
Waikato University
New Zealand
http://greenstone.org

Agenda

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- (Using the Collector)
- The power of open source

May 2003
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)
- Multi-format documents
- Multi-lingual: documents and interfaces
- Multimedia
- Metadata: standard and non-standard

Collections: on the Web

nzdl.org

(demo, not service)
Building DLs using open source software

Greenstone collections: on CD-ROM

- UNESCO
- Global Help Project
- United Nations University
- World Health Organization
- Pan American Health Organization

May 2003
Building DLs using open source software

Kataayi Multipurpose Cooperative
Rural Uganda
(20 km from Masaka)

Example

Humanity Development Library
for sustainable development and basic human needs

- 160,000 pages
- 30,000 images
- 1230 books
- 340 kg
- US$20,000
- CD-ROM
- US$6
- Win3.1x(!)/95/98/NT
- Stand-alone and intranet server
- Web browser user interface

Global Help Project, Antwerp (+ UN agencies)

May 2003
Building DLs using open source software

May 2003

Ian H. Witten

**Agenda**

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- Using the Collector
- The power of open source

**Collection of pictures**
(pictures of text)

*Alexander Turnbull Library, NZ*
Voice
(and pictures)
Hamilton Public Library

Music

May 2003
Building DLs using open source software

May 2003

Ian H. Witten

Chinese documents (pictures of text) + Chinese interface

Peking University Library

Chinese (Chinese & English interfaces)

Classic Chinese literature
Building DLs using open source software

Ian H. Witten

May 2003
Building DLs using open source software

May 2003

Ian H. Witten

Spanish

PAHO, WHO

Turkish

Koleksiyon nasıl okunur

May 2003
Building DLs using open source software

May 2003

Russian collection from Mari El Republic

http://gov.mari.ru/gsdI

Agenda

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- Using the Collector
- The power of open source
Building DLs using open source software

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

May 2003
Building DLs using open source software

Multiple search indexes

Collection-dependent

May 2003

Ian H. Witten
Multilingual searching

- Search for documents which contain the words

- Results

- Search page - Microsoft Internet Explorer

- Language extraction demo - Microsoft Internet Explorer

May 2003
Building DLs using open source software

Browsing using classifiers

AZList classifier (Title metadata)

DateList classifier (Date metadata)

May 2003
Building DLs using open source software

Hierarchy classifier (Subject metadata)

May 2003
Building DLs using open source software

Metadata extraction plugins

Acronym extraction plugin

Language identification plugin

May 2003
Building DLs using open source software

Email plugin

Phrase hierarchy extraction + thesaurus browsing

May 2003
Building DLs using open source software

May 2003

Ian H. Witten

Agenda

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- Using the Collector
- The power of open source

Collection configuration file

- creator sjboddie@cs.waikato.ac.nz
- maintainer sjboddie@cs.waikato.ac.nz
- public true
- beta true
- indexes section:text section:Title document:text
- defaultindex section:text
- plugin GAPlug
- plugin ArcPlug
- plugin RecPlug
- classify Hierarchy hfile=sub.txt metadata=Subject sort=Title
- classify HDList metadata=Title
- classify Hierarchy hfile=org.txt metadata=Organization sort=Title
- classify list metadata=Books
- format SearchVList "<td valign=top>[link][icon][/link]</td>
  - "<td>{If}{[parent(All': '):Title], [parent(All': '):Title]}:[Title]}:
  - "[link][Title][/link]"
- format CL4VList "<br>[link][Howto][/link]"
- format DocumentImages true
- format DocumentText "<h3>[Title]";
  "<p>[Text]"

collectionmeta collectionname "greenstone demo"
collectionmeta "This is a demonstration collection for the Greenstone digital library software. It contains a small subset (11 books) of the Humanity Development Library"
collectionmeta icon collectionname "/gsdl/collection/demo/images/demo.gif"
collectionmeta icon collectionname "/gsdl/collection/demo/images/demo.gif"
collectionmeta .section:Title "section titles"
collectionmeta .document:text "entire books"
collectionmeta .section:text "chapters"
Building DLs using open-source software

**Alter configuration**

- Add full-text index of titles
  - indexes document:Title
- ... or authors
  - classify AZList – metadata Creator
- Add alphabetic author browser
  - plugin WordPlug
- Include Word documents
  - plugin PDFPlug
- Include PDF documents
  - languages en fr es
- Separate index for each language
  - plugin PDFPlug – extract_acronyms
- Extract acronyms and add list
  - plugin OAIPlug
- Import OAI metadata
  - classify phind
- Extract phrase hierarchy and browser
- Alter the format of any of the...
  - format PreferenceLangs en|fr|es
  - format ...
  - cgiarg shortname=1 argdefault =fr

**Agenda**

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- Using the Collector
  - The power of open source
Building and distributing collections carries responsibilities ...
legal ... social ... ethical ...
Be aware of the power of information and use it wisely

First, you must decide whether to
- create an entirely new collection
- work with an existing one, adding data to it or deleting it.

Second, you must decide whether:
- build a completely new collection
- work with an existing one, adding data to it or deleting it.

Name: admin
Password: 12345
Sign in

Specify its name (and associated information)
Specify where the source data comes from
Adjust the configuration options (advanced users only)
"Build" the collection
Proudly view your handwork
Building DLs using open source software

Status updated every 5 secs

May 2003
Building DLs using open source software

May 2003

Ian H. Witten

Agenda

- Context
- Documents and interfaces
  - Different document types
  - ... and interface languages
- Searching and browsing
  - Different search indexes
  - ... and browsing functionality
- Collection configuration
- Using the Collector
- The power of open source
Building DLs using open source software

The power of open source: Greenstone uses ...

- Ghostscript: Interpreter for Adobe Postscript documents (Postscript plugin)
- Kea: Keyphrase extraction program (to generate metadata)
- pdftohtml: Converter for PDF documents (PDF plugin)
- rftohtml: Converter for RTF documents (RTF plugin)
- TextCat: Detects languages and document encodings
- wvWare: Converter for Word documents (Word plugin)
- Xlhtml: Converter for Excel/Powerpoint documents (plugins)
- XML::Parser: Parses XML documents, used to read and write Greenstone’s internal XML document format

and ...

- MG: Creates compressed full-text indexes and performs searches
- GDBM: Database used for metadata etc
- wget: Downloading pages from the Web when creating collections
- YAZ: Client and server implementation of Z39.50
- Stemmer: English language stemmer
- GCC: C/C++ compiler
- CVS: Version control system
- Perl: Used for plugins etc
- Apache: Web server used by many Greenstone installations

May 2003
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

Distributed
- CORBA protocol allows remote access
- Z39.50 server/client for backwards compatibility

What you see — you can get!
- Open-source software: free, extensible

Download Greenstone from

http://greenstone.org

- Linux-born
- Runs on Linux, Solaris, Mac OS/X; portable to other Unix versions
- Runs on every version of Windows: 95, 98, NT, 2000, Me, XP
  —even 3.1/3.11!