How to build a digital library 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
http://nzdl.org/

Agenda

Part 1: What Greenstone can do

Part 2: Building a collection
  Plugins, classifiers, format statements

Part 3: Running Greenstone
  ◆ installation
  ◆ collection-building
  ◆ administrative/maintenance pages
Part 1: What Greenstone can do

- What is a digital library?

- Greenstone software ... illustrated
  - collections
  - demo of Humanity Development Library
  - documents (+ pictures, voice, music)
  - multilingual (+ Maori, French, Arabic, Chinese)

- Finding documents: searching and browsing
  - full-text searching, fielded searching
  - metadata-based browsing
  - Lists, alphabetic lists, date lists
  - Hierarchical browsing structures

- Document types and formats
  - plugins and classifiers

- Configuring a collection

“Digital library” ... means different things to different people!

Collection of digital objects (text, video, audio) along with methods for access and retrieval, [user] and for selection, organisation, 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
Building DLs using open source software

May 2001

Ian H. Witten

Greenstone software

Collections
- "Library" = set of separate collections
- "Collection" = set of separate documents
- Multigigabyte collections

Documents
- Hierarchical document model
- Multimedia picture, voice, music, video collections
- Multi-language documents Unicode throughout
- Multi-language interfaces French, Chinese, Arabic ...

Access
- Web browser or CD-ROM
- Searching full-text and fielded, ranked or boolean
- Browsing hierarchical indexes created from metadata
- Metadata Dublin core + collection-specific extensions

Importing
- Plugins different document types and metadata specifications
- Classifiers create browsing indexes (collection editor decides)

Distributing
- Compression techniques throughout uses MG
- Distributed collections coming soon with Corba
- Open-source software free, extensible

Collections: on the Web
nzdl.org
(demo, not service)
Building DLs using open source software

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(!)/95/98/NT
- Stand-alone
- Intranet server
- Web browser user interface

Global Help Project, Antwerp (+ UN agencies)

May 2001
The United Nations says ...

- We are profoundly concerned at the deepening mal-distribution of access, resources and opportunities in the information and communication field ...
- A new type of poverty, "information poverty," looms ...
- Most developing countries ... are not sharing in the communications revolution ...
- The knowledge gap is widening

Statement on Universal Access to Basic Communication and Information Services, 1997
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
A book
Building DLs using open source software

Metadata specified at any level

A bibliography collection

French documents + French interface

UNESCO, Paris

May 2001
Building DLs using open source software

Ian H. Witten

Arabic documents + English interface

Chinese documents (pictures of text) + Chinese interface

May 2001
Building DLs using open source software

Chinese documents + Chinese interface

Acronym extraction plugin
Language identification plugin

Email plugin
Searching and browsing

- Searching
- Metadata-based browsing

Searching:
- multiple indexes
  (editor chooses)

Metadata-based browsing:
- Subject
- Title
- Publisher
- "HowTo"

Dublin Core:
- text
- metadata
Building DLs using open source software

A different collection means different indexes.

Ranked OR
Boolean AND
Full Boolean queries, plus other search preferences

Multilingual searching (Unicode)
Browsing: different "classifier" types

List classifier (Howto metadata)

AZList classifier (Title metadata)
Building DLs using open source software

DateList classifier (Date metadata)

Custom-made classifier (Title metadata)

Simple variant of AZList (2 lines of PERL)
Hierarchy classifier (Subject metadata)

Multilevel hierarchy

Information specified in auxiliary file

Multilevel hierarchy
Multilevel hierarchy

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
- ~ a dozen plugins (TEXT, HTML, EMAIL, WORD, RTF, PDF, PS …)

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)
How to build a digital library using open source software

Part 1: What Greenstone can do

Part 2: Building a collection
Plugins, classifiers, format statements

Part 3: Running Greenstone
installation
collection-building
administrative/maintenance pages

May 2001
Part 2: Building a collection

- What Greenstone does
  - Using the Collector
  - Altering the configuration
  - GML: Greenstone markup language
  - `collect.cfg` for the Demo collection
  - Plugins
  - Classifiers
    - List
    - AZList
    - DateList
    - HDLList
    - Hierarchy
  - Format strings
    - classifiers
    - search results
    - document text

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!**
- Gnu licensed
The pen is mightier than the sword! Building and distributing information 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.

Software "wizard" for building new collections

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 handiwork.
Building DLs using open source software

May 2001

Ian H. Witten

Title for collection: Women's History Excerpt

Contact email address: ianette@cs.waikato.ac.nz

About this collection: This collection is an excerpt of women's history primary source material and associated information from web sites around the world. The collection comprises documents.

Input sources:

- File:// C:\My Documents\misc notes\whist.html
- File:// www.greatamericanwomen.com
- File://
Building DLs using open source software

May 2001

Ian H. Witten

Status updated every 5 secs

Importing collection...
extracted 'Title' metadata "Lucretia Mott to Josephine S. Griffing, 1870"
Collection configuration file

creator  annetteb@cs.waikato.ac.nz
maintainer annetteb@cs.waikato.ac.nz
public  true
beta  true
indexes  document:text
defaultIndex  document:text
plugin  ZIPPlug
plugin  GMLPlug
plugin  TEXTPlug
plugin  HTMLPlug –file_is_url
plugin  EMAILPlug
plugin  ArcPlug
plugin  RecPlug
classify  AZList metadata=Title
collectionmeta collectionname "Women’s History Excerpt"
collectionmeta collectionextra "This collection is an excerpt for demonstration purposes, based on the Women’s ... ... contains _about:numdocs_ documents"
collectionmeta .document:text "documents"

Alter configuration

- Add full-text index of titles ...
- ... or authors
- Add alphabetic author browser
- Include Word documents
- Include PDF documents
- Separate index for each language
- Extract acronyms and add list
- Extract keyphrases and add browser (coming real soon)
- Extract phrase hierarchy and add browser
- Alter the format of any of the above
- Alter the format of any of the above
- Restrict collection’s interface langs
- Change default interface language

<table>
<thead>
<tr>
<th>expresses</th>
<th>document:Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>indexes</td>
<td>document:Title</td>
</tr>
<tr>
<td>classify</td>
<td>AZList metadata=Title</td>
</tr>
<tr>
<td>plugin</td>
<td>ZIPPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>GMLPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>TEXTPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>HTMLPlug –file_is_url</td>
</tr>
<tr>
<td>plugin</td>
<td>EMAILPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>ArcPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>RecPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>WordPlug</td>
</tr>
<tr>
<td>plugin</td>
<td>PDFPlug</td>
</tr>
<tr>
<td>languages</td>
<td>en fr es</td>
</tr>
<tr>
<td>plugin</td>
<td>PDFPlug –extract_acronyms</td>
</tr>
<tr>
<td>classify</td>
<td>phind</td>
</tr>
<tr>
<td>format</td>
<td>PreferenceLangs en</td>
</tr>
<tr>
<td>CGI ARG shortname=1 argdefault =fr</td>
<td></td>
</tr>
</tbody>
</table>
## Dublin Core metadata

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
<td>A name given to the resource</td>
</tr>
<tr>
<td>Creator</td>
<td>Creator</td>
<td>An entity primarily responsible for making the content of the resource</td>
</tr>
<tr>
<td>Subject and keywords</td>
<td>Subject</td>
<td>The topic of the content of the resource</td>
</tr>
<tr>
<td>Description</td>
<td>Description</td>
<td>An account of the content of the resource</td>
</tr>
<tr>
<td>Publisher</td>
<td>Publisher</td>
<td>An entity responsible for making the resource available</td>
</tr>
<tr>
<td>Contributor</td>
<td>Contributor</td>
<td>An entity responsible for making contributions to the content of the resource</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>A date associated with an event in the life cycle of the resource</td>
</tr>
<tr>
<td>Resource type</td>
<td>Type</td>
<td>The nature or genre of the content of the resource</td>
</tr>
<tr>
<td>Format</td>
<td>Format</td>
<td>The physical or digital manifestation of the resource</td>
</tr>
<tr>
<td>Resource identifier</td>
<td>Identifier</td>
<td>An unambiguous reference to the resource within a given context: this is the object identifier or OID</td>
</tr>
<tr>
<td>Source</td>
<td>Source</td>
<td>A Reference to a resource from which the present resource is derived</td>
</tr>
<tr>
<td>Language</td>
<td>Language</td>
<td>A language of the intellectual content of the resource</td>
</tr>
<tr>
<td>Relation</td>
<td>Relation</td>
<td>A reference to a related resource</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
<td>The extent or scope of the content of the resource</td>
</tr>
<tr>
<td>Rights management</td>
<td>Rights</td>
<td>Information about rights held in and over the resource</td>
</tr>
</tbody>
</table>

## GML: Greenstone markup language

```xml
<gsdlsection>
<metadata>
  <gsdlsourcefilename> uu02fe.txt </gsdlsourcefilename>
  <gsdldoctype> indexed_doc </gsdldoctype>
  <Identifier> HASHa723e7e164df07c833bfc4 </Identifier>
  <Title> Freshwater Resources in Arid Lands </Title>
  <gsdlassocfile> cover.jpg:image/jpeg </gsdlassocfile>
  <gsdlassocfile> p21.jpg:image/jpeg </gsdlassocfile>
  <gsdlassocfile> p22.jpg:image/jpeg </gsdlassocfile>
</metadata>

This is the text of the document

</gsdlsection>
```
Building DLs using open source software

GML: Greenstone markup language

This is the text of the preface

This is the first part of the conclusions

Rest of the document
Building DLs using open source software

GML: Greenstone markup language

This is the second part of the conclusions

Rest of the document

collect.cfg for the demo collection

May 2001
Building DLs using open source software

May 2001

Ian H. Witten

indexes section:text
section:Title
document:text
defaultindex section:text

collectionmeta .section:Title "section titles"
collectionmeta .document:text "entire books"
collectionmeta .section:text "chapters"

collect.cfg for the demo collection

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 GMLPlug
plugin HBPlug
plugin ArcPlug
plugin IndexPlug
plugin RecPlug

classify Hierarchy hfile=sub.txt metadata=Subject sort=Title
classify AZList metadata=Title

classify Hierarchy hfile=org.txt metadata=Organization sort=Title

classify List metadata=Howto

format SearchVList "<td valign=top>[link][icon][/link]</td>
    <td>{If}{[parent(All': '):Title], [parent(All': '):Title]:}
        [link][Title][/link]</td>">

format CL4VList "<br>[link][Howto][/link]"

format DocumentImages true
format DocumentText "<h3>[Title]</h3>

<text>

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

Plugins

Used by collection-building software to accomplish format-specific parsing of source documents

Plugin pipeline: files are passed to each plugin in turn until one is found that can process it

- GMLPlug processes .gml files generated during import
- HBPlug processes HTML marked up for UN collections
- ArcPlug processes .gml filelist in archives.inf
- IndexPlug assigns metadata from index.txt file
- RecPlug recurses through a directory structure

Also TEXTPlug, HTMLPlug, EMAILPlug, WORDPlug, RTFPlug, PDFPlug, PSPlug, FoxPlug, PrePlug, GBPlug, TCCPlug...
Building DLs using open source software

import/index.txt
Used by IndexPlug
to add metadata to .gml files

Filename   Subject metadata
key line   Organization metadata
   Howto metadata   Magazine metadata

key:

<table>
<thead>
<tr>
<th>key</th>
<th>Subject</th>
<th>Organization</th>
<th>Howto</th>
<th>Magazine</th>
</tr>
</thead>
<tbody>
<tr>
<td>bostid/b32bue</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;start a butterfly farm&quot;</td>
<td></td>
</tr>
<tr>
<td>fachref/fb31fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>fachref/fb34fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce little-known Asian farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals and micro-livestock in your farm&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec158e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;achieve gender equality&quot;</td>
<td>Magazine=The Courier</td>
</tr>
<tr>
<td>eecourier/ec159e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec160e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>wb/wb34te</td>
<td>6.4</td>
<td>wb</td>
<td>&quot;achieve gender equality&quot;</td>
<td></td>
</tr>
</tbody>
</table>

import/index.txt
Used by IndexPlug
to add metadata to .gml files

Filename   Subject metadata
key line   Organization metadata
   Howto metadata   Magazine metadata

key:

<table>
<thead>
<tr>
<th>key</th>
<th>Subject</th>
<th>Organization</th>
<th>Howto</th>
<th>Magazine</th>
</tr>
</thead>
<tbody>
<tr>
<td>bostid/b32bue</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;start a butterfly farm&quot;</td>
<td></td>
</tr>
<tr>
<td>fachref/fb31fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>fachref/fb34fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce little-known Asian farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals and micro-livestock in your farm&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec158e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;achieve gender equality&quot;</td>
<td>Magazine=The Courier</td>
</tr>
<tr>
<td>eecourier/ec159e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec160e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>wb/wb34te</td>
<td>6.4</td>
<td>wb</td>
<td>&quot;achieve gender equality&quot;</td>
<td></td>
</tr>
</tbody>
</table>

import/index.txt
Used by IndexPlug
to add metadata to .gml files

Filename   Subject metadata
key line   Organization metadata
   Howto metadata   Magazine metadata

key:

<table>
<thead>
<tr>
<th>key</th>
<th>Subject</th>
<th>Organization</th>
<th>Howto</th>
<th>Magazine</th>
</tr>
</thead>
<tbody>
<tr>
<td>bostid/b32bue</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;start a butterfly farm&quot;</td>
<td></td>
</tr>
<tr>
<td>fachref/fb31fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>fachref/fb34fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce little-known Asian farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals and micro-livestock in your farm&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec158e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;achieve gender equality&quot;</td>
<td>Magazine=The Courier</td>
</tr>
<tr>
<td>eecourier/ec159e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec160e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>wb/wb34te</td>
<td>6.4</td>
<td>wb</td>
<td>&quot;achieve gender equality&quot;</td>
<td></td>
</tr>
</tbody>
</table>

import/index.txt
Used by IndexPlug
to add metadata to .gml files

Filename   Subject metadata
key line   Organization metadata
   Howto metadata   Magazine metadata

key:

<table>
<thead>
<tr>
<th>key</th>
<th>Subject</th>
<th>Organization</th>
<th>Howto</th>
<th>Magazine</th>
</tr>
</thead>
<tbody>
<tr>
<td>bostid/b32bue</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;start a butterfly farm&quot;</td>
<td></td>
</tr>
<tr>
<td>fachref/fb31fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>fachref/fb34fe</td>
<td>14.12</td>
<td>fachfe</td>
<td>&quot;farm snails&quot;</td>
<td>Subject=16.11</td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce little-known Asian farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b3baxe</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce farm animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals and micro-livestock in your farm&quot;</td>
<td></td>
</tr>
<tr>
<td>bostid/b17mie</td>
<td>16.11</td>
<td>bostid</td>
<td>&quot;introduce small animals with a promising future&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec158e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;achieve gender equality&quot;</td>
<td>Magazine=The Courier</td>
</tr>
<tr>
<td>eecourier/ec159e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>eecourier/ec160e</td>
<td>23.15</td>
<td>ecc</td>
<td>&quot;Magazine=The Courier&quot;</td>
<td></td>
</tr>
<tr>
<td>wb/wb34te</td>
<td>6.4</td>
<td>wb</td>
<td>&quot;achieve gender equality&quot;</td>
<td></td>
</tr>
</tbody>
</table>
The List classifier creates a linear list of the specified metadata values.

The AZList classifier creates a list with A-Z tabs.

The DateList classifier creates a list of dates.

(recall: Howto metadata was specified in the index.txt file)
AZList classifier (Title) for the demo collection

classify AZList metadata>Title
Building DLs using open source software

DateList classifier

```
classify DateList metadata=Date
```

HDLList classifier (Title)

```
classify HDLList metadata=Title
```

May 2001
Building DLs using open source software

The HDLLList classifier

HDLLList is AZList + extra code to read etc/mags.txt and special-case magazines (2 lines)

search  subjects  titles a-z  organization  how to

classify HDLLList metadata=Title

e tc/mags . txt

identifier (matches Magazine metadata value)
position in browsing hierarchy
title for hierarchy browser

CERES 1
"Food and Nutrition Bulletin" 2 "Food and Nutrition Bulletin"
"The Courier" 3 "The Courier"
"BGKE Bulletin" 4 "BGKE Bulletin"
"Boiling Point" 5 "Boiling Point"
"Developing Ideas" 6 "Developing Ideas"
"GATE Magazine" 7 "GATE Magazine"
"Go Between" 8 "Go Between"
"Basin - News" 9 "Basin - News"

etc/org . txt

identifier (matches Organization metadata value)
position in the hierarchy that the browser implements
title to be displayed in hierarchy browser

accu 1 ACCU
ag21 2 "Agenda 21"
bgz 3 "BASIN - GTZ - SKAT"
bostid 4 BOSTID
cps 5 CPS
cpaz 6 "CTA Spore"
cf 7 "Commonwealth Foundation"
csa 8 "Computer Science Unplugged"
dox 9 "Development Consultancy Services"
dcc 10 "EC Courier"
dcpps 11 "EC DG8"
dgpp 12 "EC DG8"
decho 13 "Educational Concerns for Hunger Organization"
dac 14 DAC
dacsfa 15 "DAC Better Farming series"
ceres 16 "DAC Ceres"
ff 17 "Food First"
classify Hierarchy hfile=org.txt
metadata=Organization sort=Title
Building DLs using open source software

Format strings

(document text)

(classesifiers (separate HList or VList specs)

(search results)

(sensible defaults throughout)

(components)

(document text)

(classifiers)

(search results)

(html)

[link] [text] [link]

[link] [text]

[link] [text]

[link] [text]

(if statement)

List classifier (Howto)

<link> [Howto] [/link]

format CL4VList "<br>
[link] [Howto] [/link]"

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=Howto

format DocumentImages true

format DocumentText "<h3>[Title]</h3>

<p>[Text]"
Building DLs using open source software

May 2001

Format search results

- \[\text{format SearchXML: } \ast<td align=top>[\text{link}] [icon] [/link]\ast\] <td>[If]{[parent(All': '):Title], [parent(All': '):Title]: } [link] [Title] [/link]</td>\]

- [link][icon][/link]
- [parent(All': '): Title]
  - hierarchy enclosing parents
  - separated by ":"
  - title metadata
- [link][Title][/link]
  - use empty string if no parent

Format document text

- \[\text{format DocumentImages true}\]
- \[\text{format DocumentText } "<h3>[Title]\</h3>\n\n<p>[Text]\"}\]

Building a pen

97. Build a pen for your *snails* as soon as you can after you have burn the plant material. This is to keep other snails, mice, rats or snakes from getting into your pen.

98. A snail pen is a simple fenced-in area and you can build a fence using:
  - corrugated sheet metal or plastic sheets
  - woven plant material
  - woven plant material with chicken wire
### How to build a digital library using open source software

**Part 1: What Greenstone can do**

**Part 2: Building a collection**
- Plugins, classifiers, format statements

**Part 3: Running Greenstone**
- Installation
- Collection-building
- Administrative/maintenance pages
Admin support

❖ Add new user with ... privileges
  – E.g. admin, collection-building
❖ Check what collections are available
  – Including “private” ones not on home page
❖ Check summary info about a collection
  – build date
  – collection metadata
  – interface language preferences
  – number of docs/sections/words/bytes
❖ Logging
  – switch on or off
  – check user logs—every page access is logged
  – check system logs—errors are logged
❖ Notify whenever a new collection is built
❖ Browse technical info about the installation

DL admin ≠ computer systems admin
Installing Greenstone

- PERL 5 for collection-building
- Web server (e.g. Apache) fastCGI optional
- GCC for Unix compilation
- Visual C++ for Windows compilation
- GDBM (Gnu database manager) included for Windows
- MG, crypt included

Downloading collections

<table>
<thead>
<tr>
<th>collection</th>
<th>abbrev</th>
<th>built size (Mb)</th>
<th>download size (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic demonstration collection</td>
<td>arabic</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Bibliothèque pour le développement</td>
<td>tulane</td>
<td>492</td>
<td>340</td>
</tr>
<tr>
<td>Durable et les Besoins Essentiels</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese demonstration collection</td>
<td>chinese</td>
<td>1</td>
<td>470</td>
</tr>
<tr>
<td>Collection on critical global issues</td>
<td>ccgi</td>
<td>160</td>
<td>102</td>
</tr>
<tr>
<td>Computer science bibliography</td>
<td>csbib</td>
<td>866</td>
<td>112</td>
</tr>
<tr>
<td>Computer science technical reports</td>
<td>csstr</td>
<td>2010</td>
<td>1800</td>
</tr>
<tr>
<td>Food and Nutrition Library</td>
<td>fsl</td>
<td>155</td>
<td>98</td>
</tr>
<tr>
<td>HCI bibliography</td>
<td>hcbib</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>Humanity Development Library</td>
<td>hdl</td>
<td>199</td>
<td>387</td>
</tr>
<tr>
<td>Indigenous Peoples</td>
<td>ipc</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Medical and Health Library</td>
<td>mhl</td>
<td>142</td>
<td>73</td>
</tr>
<tr>
<td>Maori newspapers</td>
<td>niupepa</td>
<td>670</td>
<td>659</td>
</tr>
<tr>
<td>Oral history</td>
<td>ohist</td>
<td>430</td>
<td>421</td>
</tr>
<tr>
<td>Project Gutenberg</td>
<td>gutenberg</td>
<td>510</td>
<td>427</td>
</tr>
<tr>
<td>Sahel point doc</td>
<td>unesco</td>
<td>113</td>
<td>78</td>
</tr>
<tr>
<td>The computists weekly</td>
<td>tcc</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>Tidbits magazine</td>
<td>tidbits</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>United Nations University collection</td>
<td>unu</td>
<td>97</td>
<td>71</td>
</tr>
<tr>
<td>Virtual Disaster Library</td>
<td>paho</td>
<td>110</td>
<td>73</td>
</tr>
<tr>
<td>World Environment Library</td>
<td>envl</td>
<td>309</td>
<td>220</td>
</tr>
<tr>
<td>Women’s history</td>
<td>whist</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6363</td>
<td>5366</td>
</tr>
</tbody>
</table>
Demo

- Delete and re-install the software
- Look around the directory structure
- Use *The Collector* to build a collection from `C:\Perl\html\lib\CGI`
- Alter the Demo collection’s config file change the format of the *howto* list
- Browse Maintenance/Admin pages