

How building works, bibliography collections



Course material prepared by

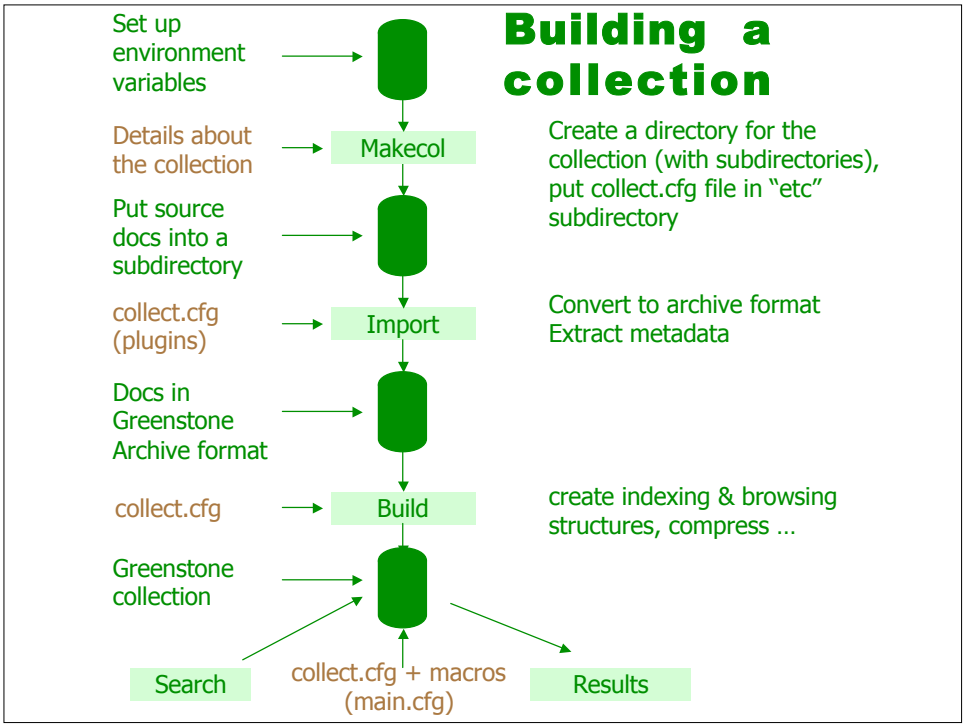
Greenstone Digital Library Project
University of Waikato, New Zealand

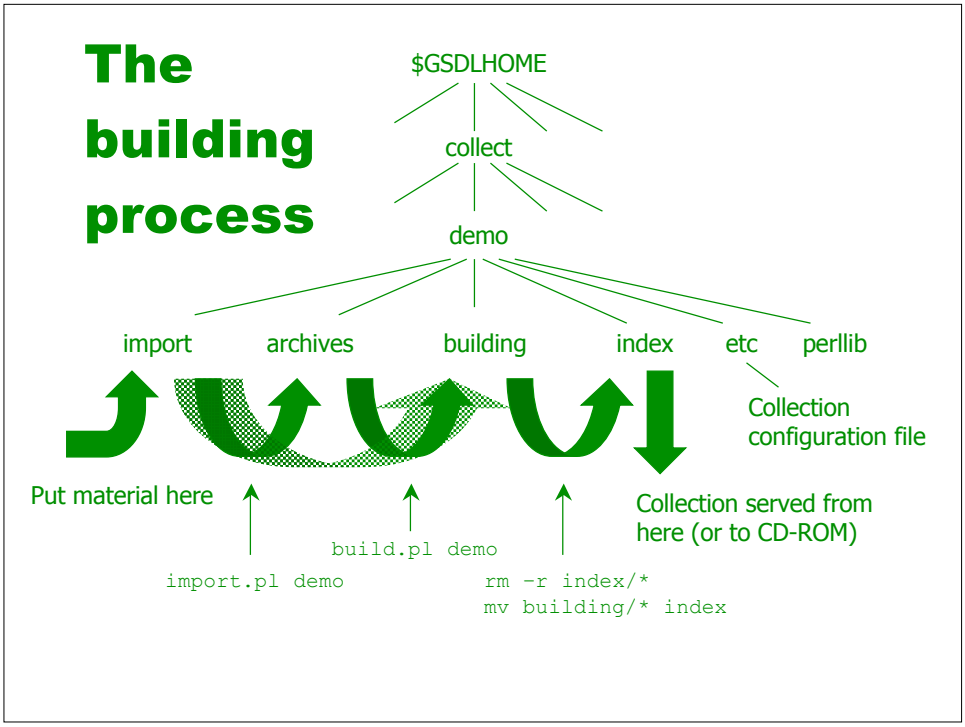
and National Centre for Science Information,
Indian Institute of Science, Bangalore



Agenda

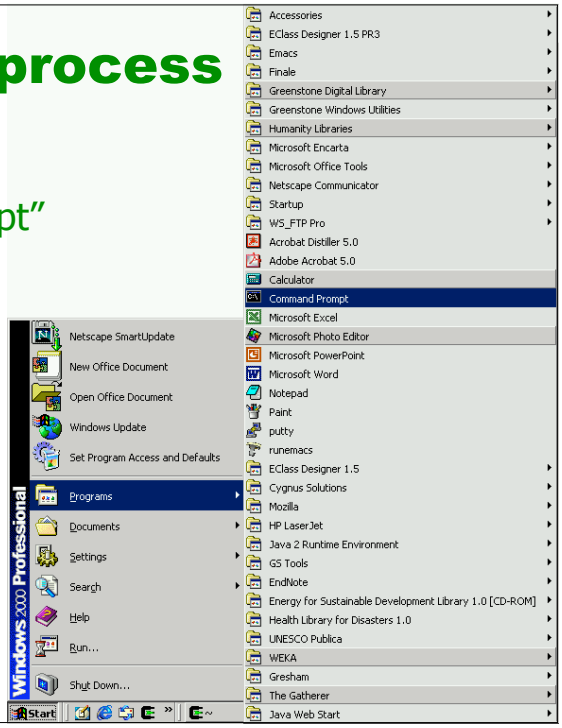
- Building a collection
- The dreaded black screen
- More on building
- Bibliography collections
- Authentication


Building a collection
Set up environment variables
Details about the collection
Makecol
Create a directory for the collection (with subdirectories), put collect.cfg file in "etc" subdirectory
Put source docs into a subdirectory
collect.cfg (plugins)
Import
Convert to archive format
Extract metadata
Docs in Greenstone Archive format
collect.cfg
Build
create indexing & browsing structures, compress ...
Greenstone collection
Search
collect.cfg + macros (main.cfg)
Results


The building process
$GSDLHOME
collect
demo
import
archives
building
index
etc
perllib
Collection configuration file
Put material here
Collection served from here (or to CD-ROM)
build.pl demo
import.pl demo
rm -r index/*
mv building/* index

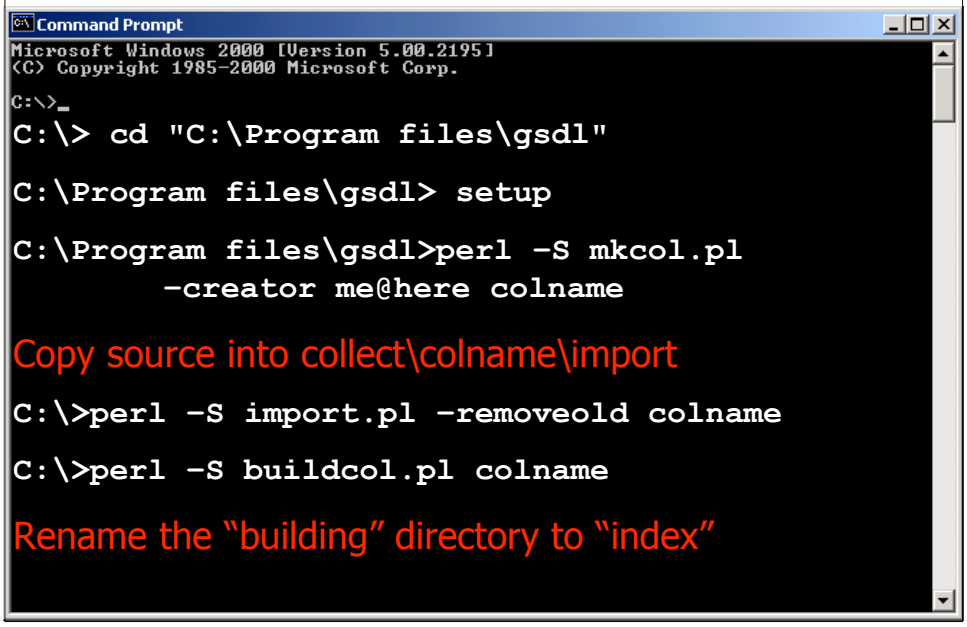
The building process

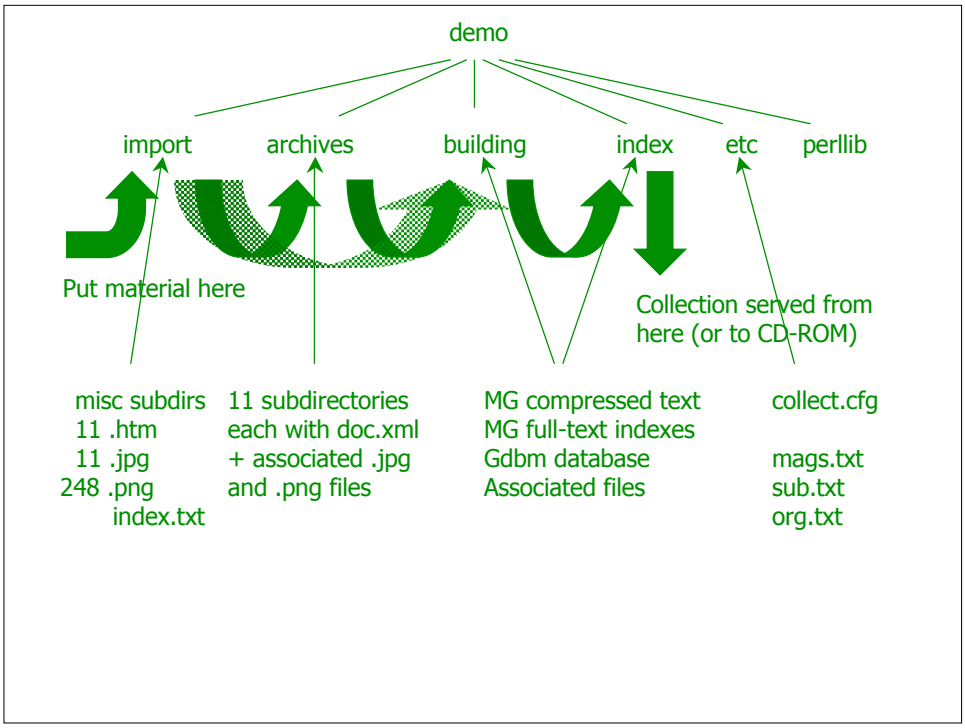
Get "Command Prompt" from Start menu

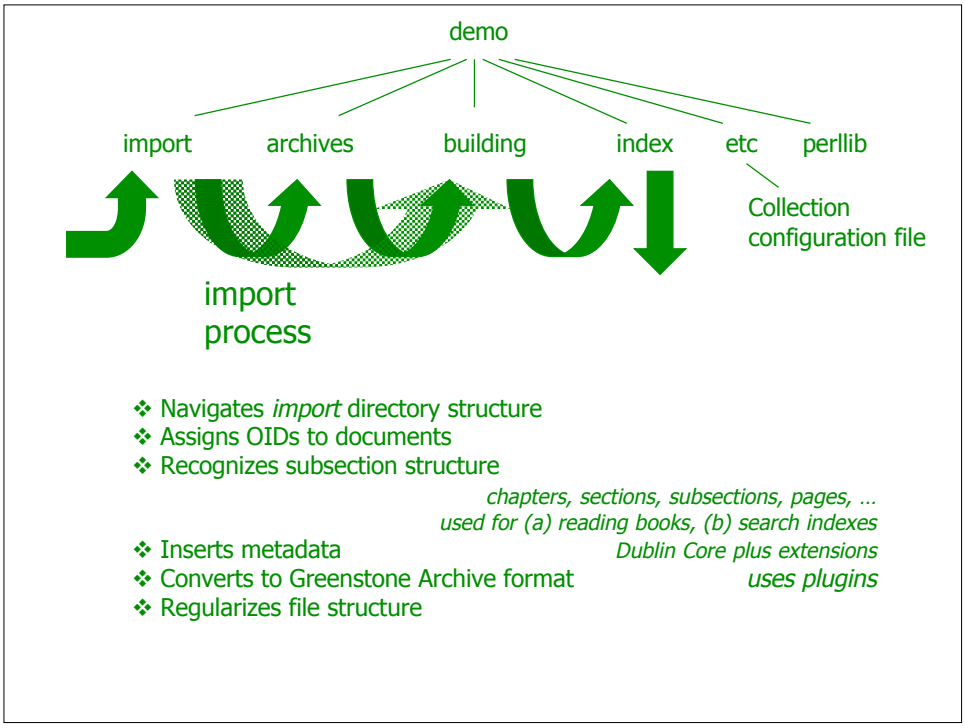




The building process







- Navigates *import* directory structure
- Assigns OIDs to documents
- Recognizes subsection structure

chapters, sections, subsections, pages, …
used for (a) reading books, (b) search indexes

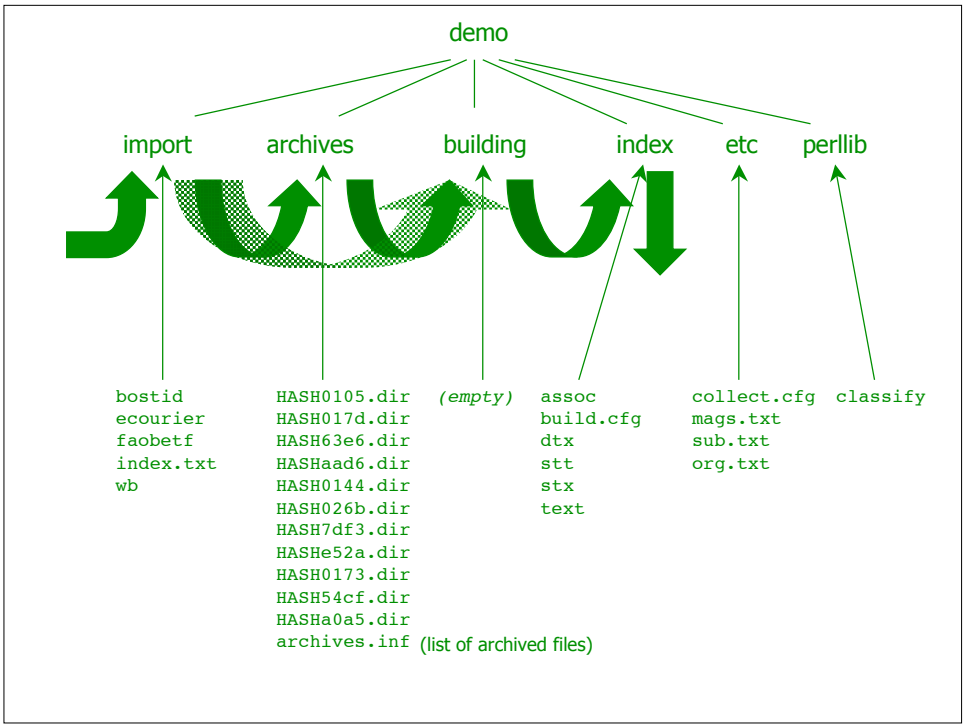
- Inserts metadata

Dublin Core plus extensions

- Converts to Greenstone Archive format

uses plugins

- Regularizes file structure



Contents of demo/index

build.cfg used by receptionist to determine indexes

```
builddate  951855434
indexmap   section:text->stx  section:Title->stt  document:text->dtx
numbytes   3029746
numdocs    11
```

associated files

```
assoc:
  HASH0141.dir
  HASH0169.dir
  HASH01a3.dir
  HASH01b4.dir
  HASH01ba.dir
  HASH01d6.dir
  HASH0f76.dir
  HASH863c.dir
  HASH8b94.dir
  HASHc5b3.dir
  HASHd803.dir
```

mg text

```
text:
  demo.bdb
  demo.t          text
  demo.td   dictionary
  demo.ti   text index
  demo.tsd      stats
```

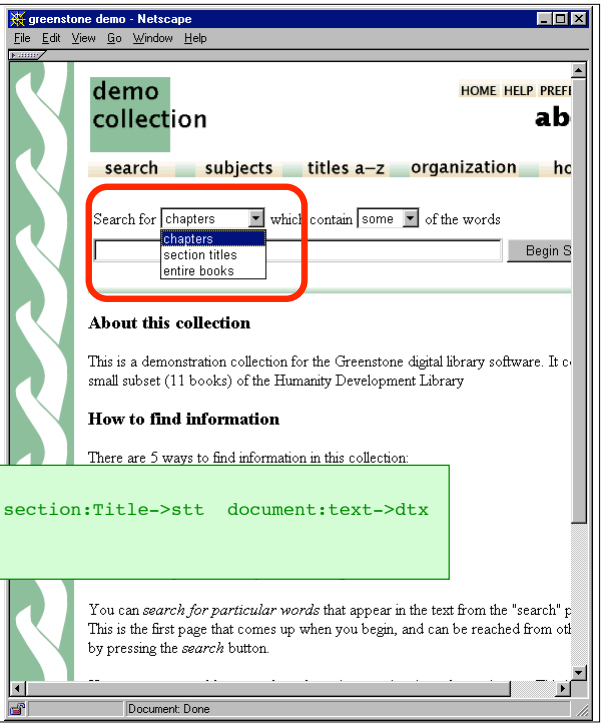
document database

mg indexes

```
stx:   stt:   dtx:
  demo.i      inverted file
  demo.tiw      doc weights
  demo.wa    approx weights
  demo.idb        term dict
  demo.ib1    stem indexes:
  demo.ib2      casefolded,
  demo.ib3    stemmed, both
```

build.cfg

```
builddate  951855434
indexmap   section:text->stx  section:Title->stt  document:text->dtx
numbytes   3029746
numdocs    11
```

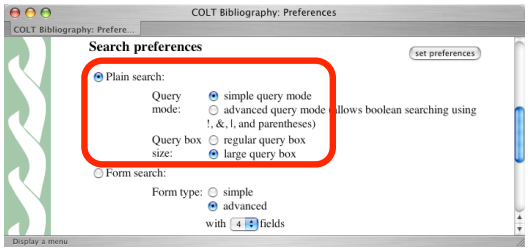




Bibliography collections

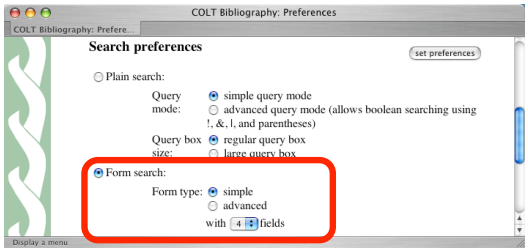
bibliography collection	With about 4,000 bibliography entries, this collection incorporates a form-based search interface that allows fielded searching. It is fairly complex.
bibliography supplement	This tiny collection of 10 bibliography entries illustrates the "supercollection" facility which searches several collections together, seamlessly. It operates together with the Bibliography collection, and its configuration file is almost the same.
MARC example	Based on some MARC records from the Library of Congress, this is a simple collection (and does not allow form-based searching).
OAI example	Using the Open Archive Protocol and the Import-From feature, this retrieves metadata from an archive and builds a collection from the records. In this case they are images, so both the OAI and Image plugins are used.
CDS/ISIS example	This collection is built from a CDS/ISIS database of about 150 bibliography entries. It uses the ISISPlug plugin, which reads the standard ISIS .mst and .fdt files and converts them to Greenstone metadata.

Plain search



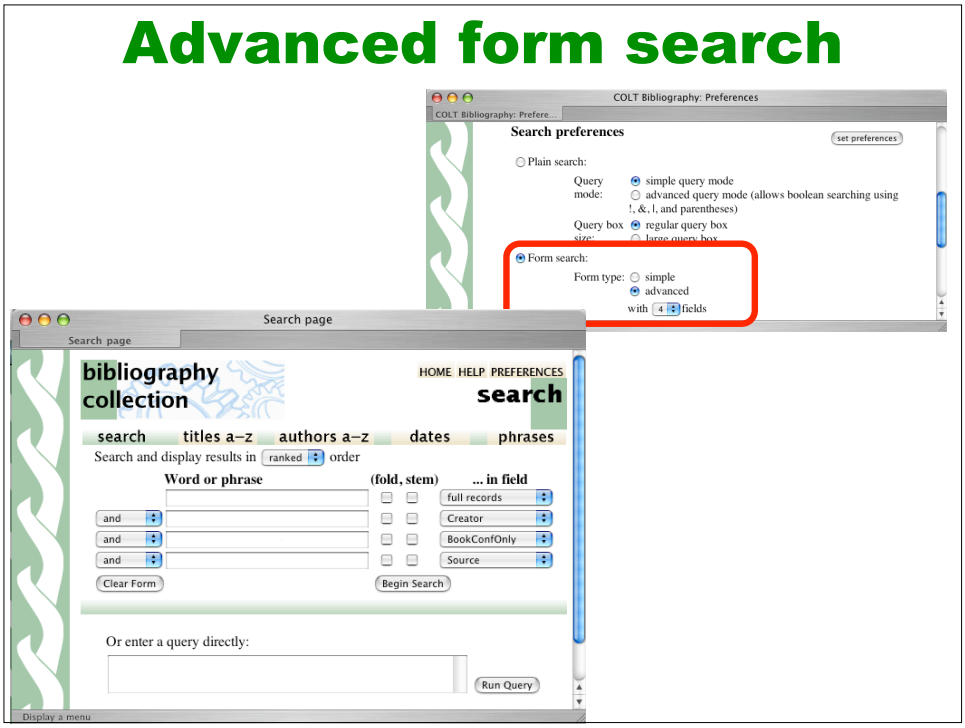


Simple form search

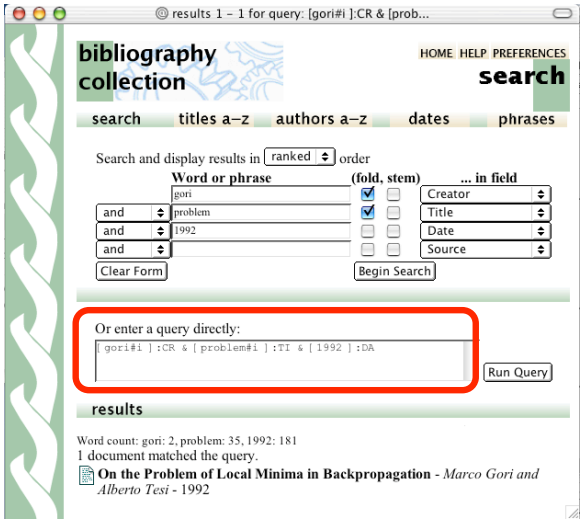




Advanced form search

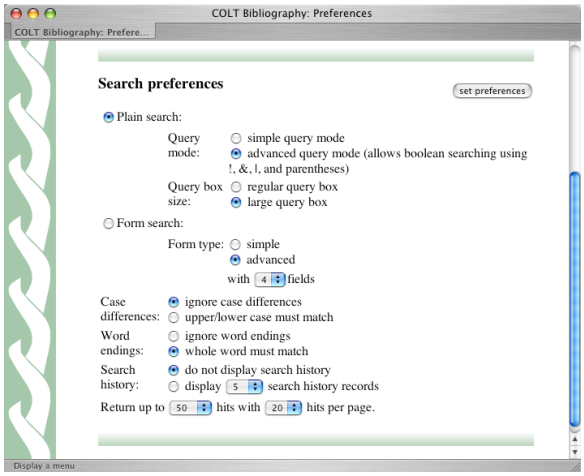


Advanced form search



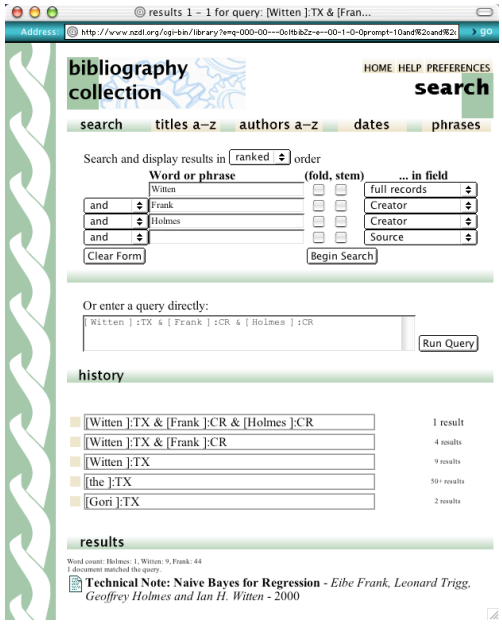
- *Fielded searching*
  - [terms]:Field
  - Specifies searching within a particular field
- *Term modifiers*
  - #icus /x
  - Stemming, casefolding and weighting
- *Fields*
  - TI title
  - CR creator
  - SU subject
  - TX text
  - ...

Advanced plain search



- *Regular vs large query box*
- *Boolean operators*
  - & AND
  - | OR
  - ! NOT
- *Phrase searching ("...")*
- *Case folding*
- *Stemming (word endings*
- *Search history*
- *Number of returned results*

Search history



(works in plain search too)

Bibliography supplement

bibliography collection

With about 4,000 bibliography entries, this collection incorporates a form-based search interface that allows fielded searching. It is fairly complex.

bibliography supplement

This tiny collection of 10 bibliography entries illustrates the "supercollection" facility which searches several collections together, seamlessly. It operates together with the Bibliography collection, and its configuration file is almost the same.

Bibliography supplement

Same as bibliography collection except …

http://127.0.0.1/gsdl/collect/coltextra-e/etc/collect.cfg

```
creator    ihw@cs.waikato.ac.nz
public true

collectionmeta collectionname      "COLT Bibliography supplement"
collectionmeta iconcollectionsmall "_httpcollection_/images/coltextra-e.gif"
collectionmeta iconcollection      "_httpcollection_/images/coltextra.gif"
collectionmeta collectionextra 'This collection, which contains
    _about:numdocs_ BibTeX entries, is a small adjunct to the sample
    bibliography collection. It contains a few Computational Learning
    Theory (COLT) papers from the year 2003.'

supercollection coltextra-e coltbib-e

searchtype    form plain
indexes       text metadata
collectionmeta .text "full records"

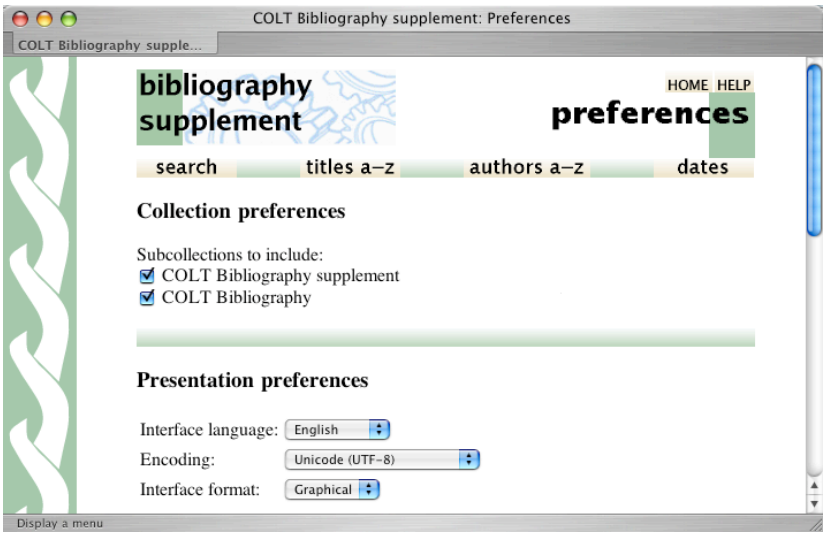
plugin ZIPPlug
plugin BibTexPlug
plugin GAPlug
plugin ArcPlug
plugin RecPlug

classify   AZList -metadata Title
classify   AZCompactList -metadata Author -buttonname Creator
classify   DateList

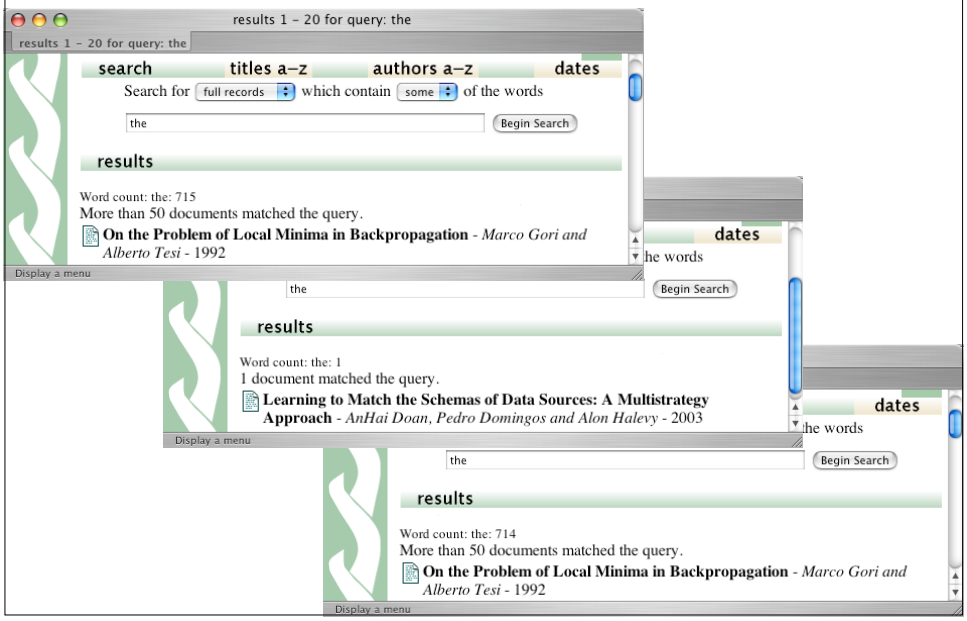
format VList "<td valign=top>[link][icon][/link]</td>
```

supercollection coltextra-e coltbib-e

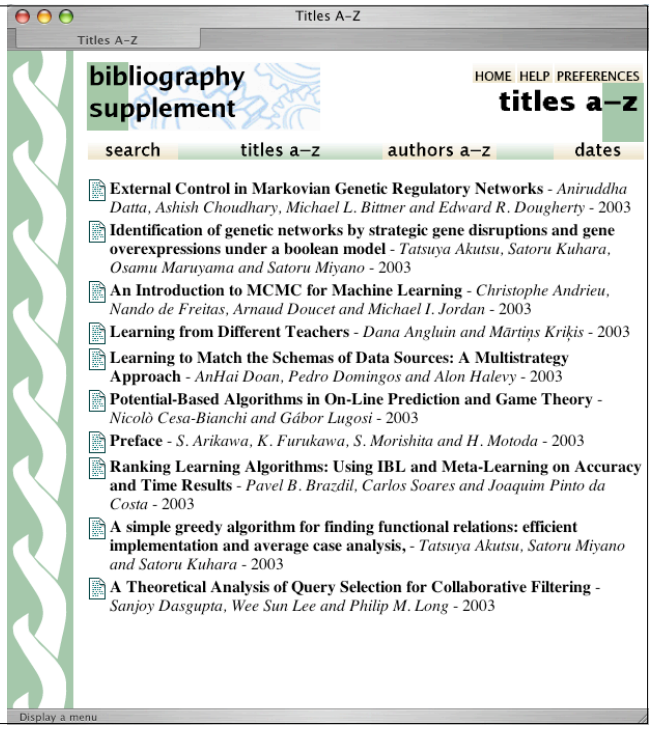
Bibliography supplement



Cross-collection searching ...



... but not browsing



Bibliography collections

MARC example

Based on some MARC records from the Library of Congress, this is a simple collection (and does not allow form-based searching).

Map MARC field numbers to Greenstone metadata:

gsdl/etc/marctodc.txt

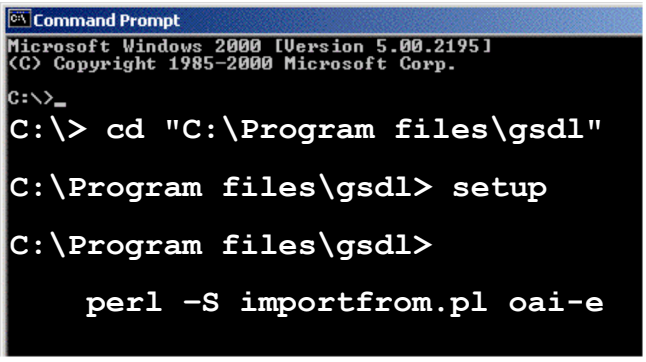
720 --> Creator	Uncontrolled name
100 --> Creator	Personal name
110 --> Creator	Corporate name
111 --> Creator	Meeting name
520 --> Description	Summary, note
856 --> URL	Electronic location
260 --> Publisher	Publication, etc
787 --> Relation	Nonspecific reln note
540 --> Rights	Reproduction notes
024 --> MarcIdentifier	Identifier
786 --> MarcSource	Source
546 --> MarcLanguage	Language
650 --> Subject	Subject
653 --> Subject	Index term
245 --> Title	Title statement
655 --> Type	Index term - genre

Bibliography collections

OAI example

Using the Open Archive Protocol and the Import-From feature, this retrieves metadata from an archive and builds a collection from the records. In this case they are images, so both the OAI and Image plugins are used.

```
acquire OAI -src rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI1.1/jcdlpix.pl
            -getdoc
```



–getdoc
download docs from URLs in OAI Identifier field

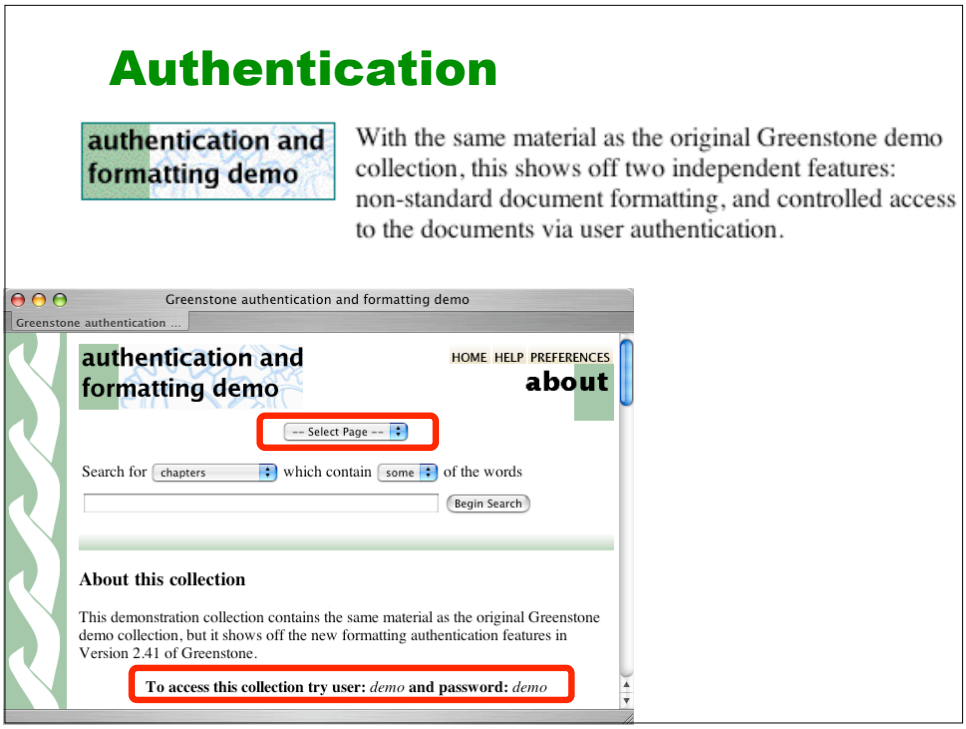
Bibliography collections

CDS/ISIS example

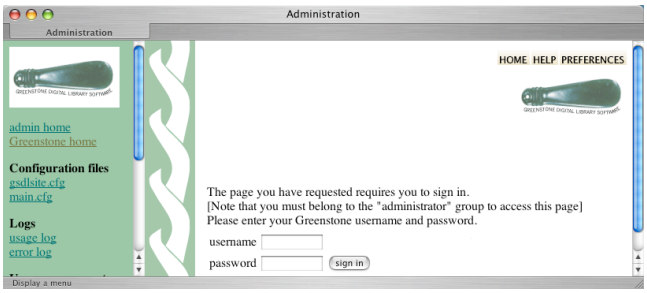
This collection is built from a CDS/ISIS database of about 150 bibliography entries. It uses the ISISPlug plugin, which reads the standard ISIS .mst and .fdt files and converts them to Greenstone metadata.

ISISPlug demo


Agenda
Building a collection
The dreaded black screen
More on building
Bibliography collections
Authentication


Authentication
authentication and formatting demo
With the same material as the original Greenstone demo collection, this shows off two independent features: non-standard document formatting, and controlled access to the documents via user authentication.
Greenstone authentication and formatting demo
Greenstone authentication ...
authentication and formatting demo
HOME HELP PREFERENCES
about
-- Select Page --
Search for chapters which contain some of the words
Begin Search
About this collection
This demonstration collection contains the same material as the original Greenstone demo collection, but it shows off the new formatting authentication features in Version 2.41 of Greenstone.
To access this collection try user: demo and password: demo

Admin pages




```
# Authentication directives

authenticate            document
public_documents        HASH30aa188f4d8ddaef558fc9

auth_group              allowed
```

Individual documents, or whole collection?

If individual docs, either list public or private documents

Which user groups can see the documents?

Authentication demo

Documentation and Help

- *Developer's Guide* (developer.pdf)
- Example collections

Questions?

Comments?

Discussion?

Feedback form!