

How collection building works



Course material prepared by

Greenstone Digital Library Project
University of Waikato, New Zealand

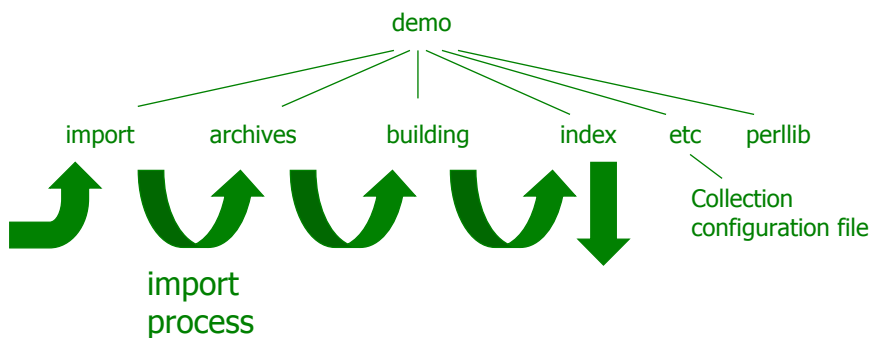
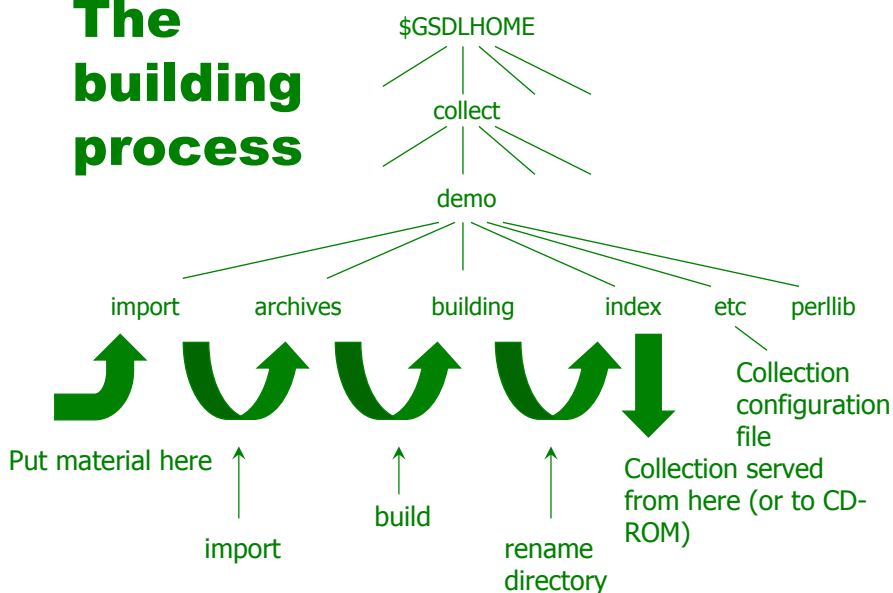
and National Centre for Science Information,
Indian Institute of Science, Bangalore

Agenda



- ❖ Building a collection
- ❖ The dreaded black screen
- ❖ More on building

The building process

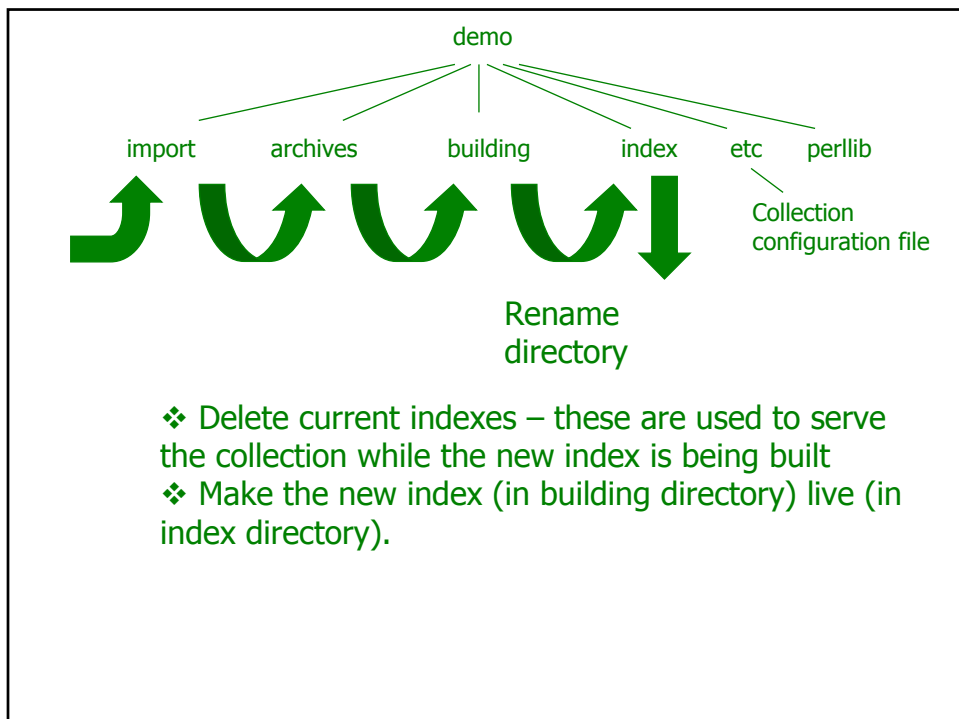
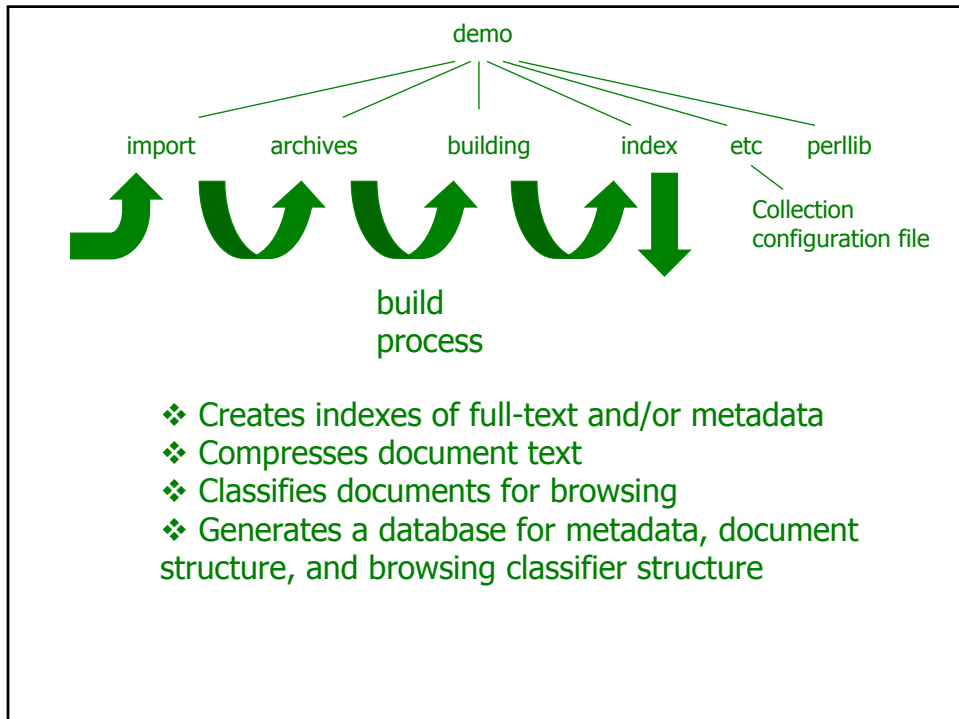


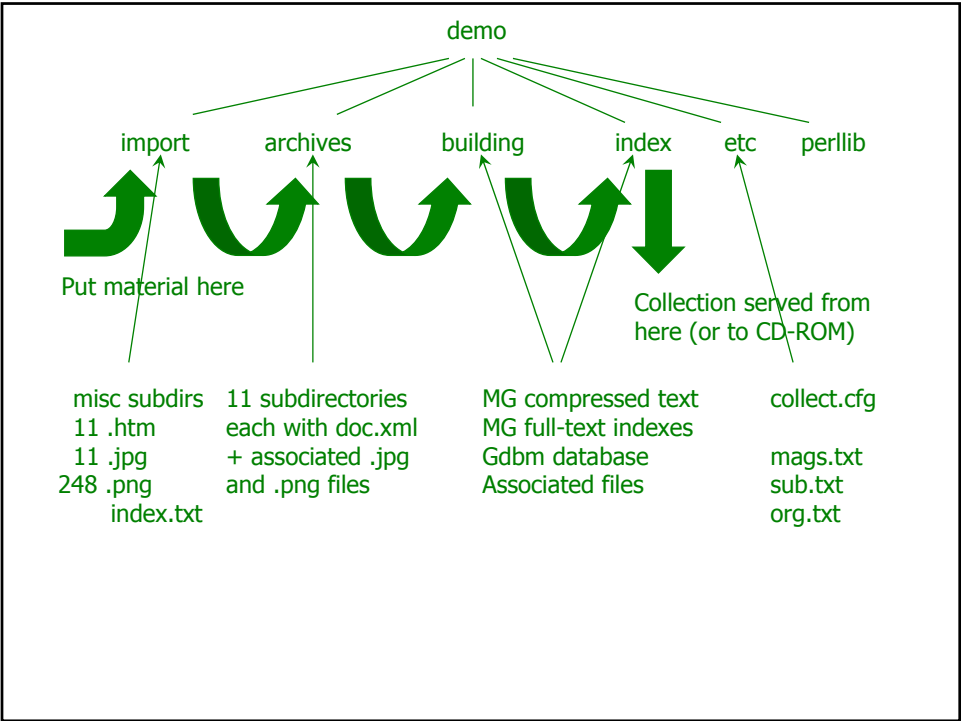
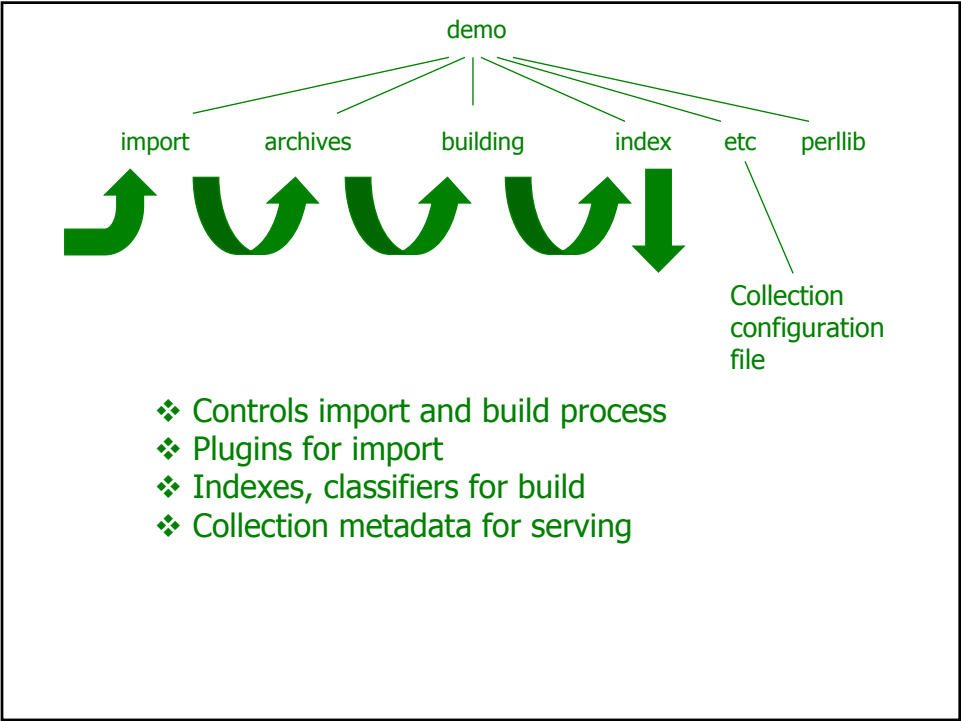
- ❖ Navigates *import* directory structure
- ❖ Assigns OIDs to documents
- ❖ Recognizes subsection structure

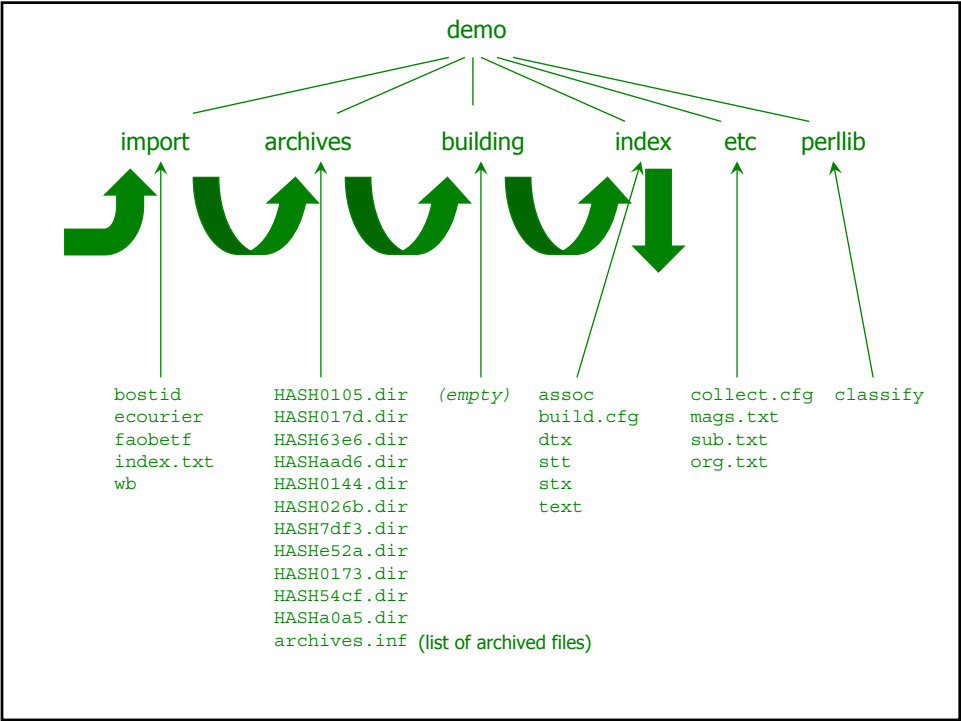
*chapters, sections, subsections, pages, ...
used for (a) reading books, (b) search indexes*

- ❖ Inserts metadata
- ❖ Converts to Greenstone Archive format
- ❖ Regularizes file structure

*Dublin Core plus extensions
uses plugins*







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.ldb
demo.t text
demo.td dictionary
demo.ti text index
demo.tsd stats
```

document database

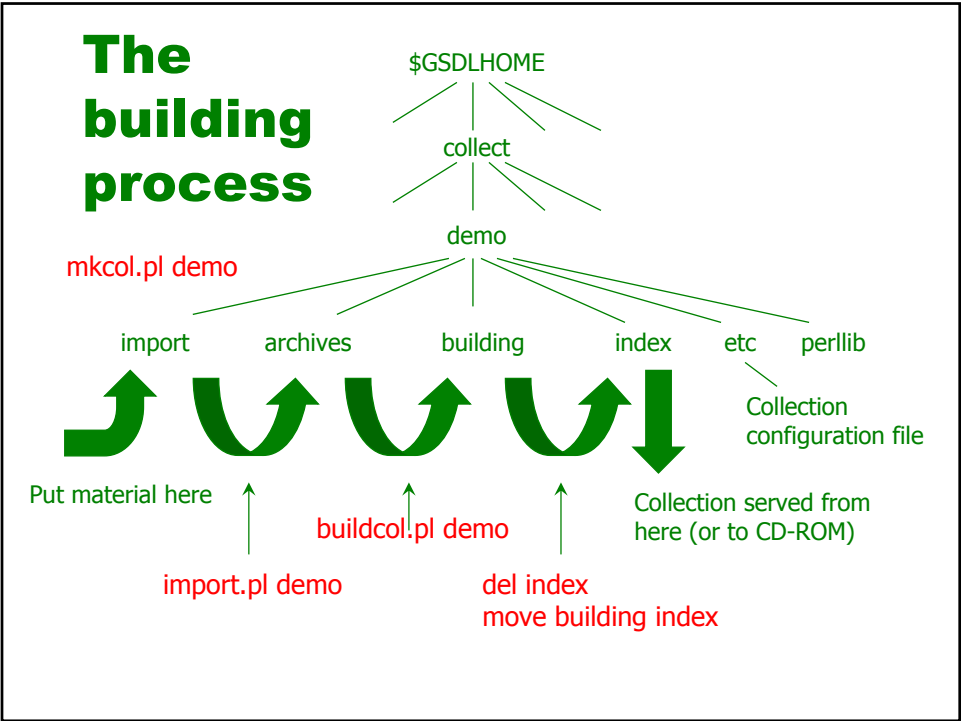
mg indexes

```
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
```

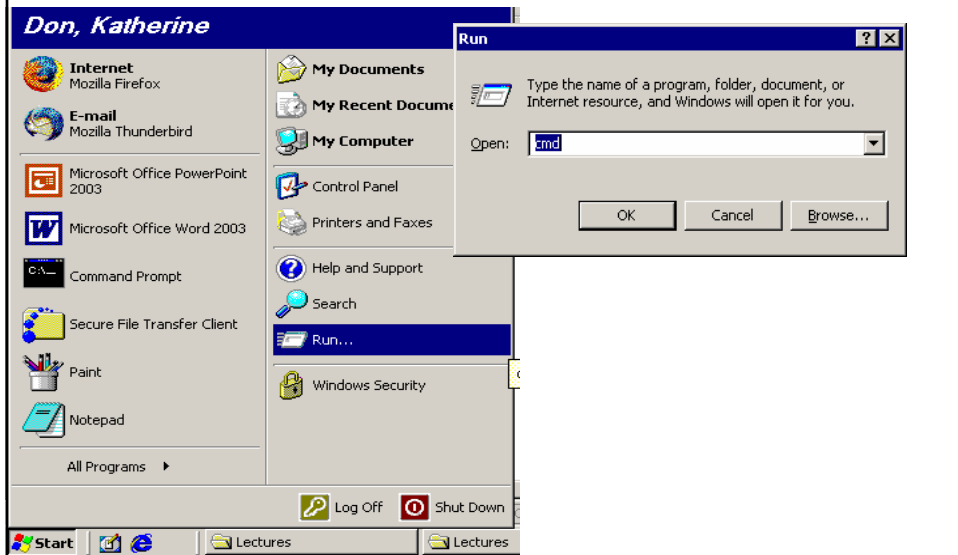


Agenda

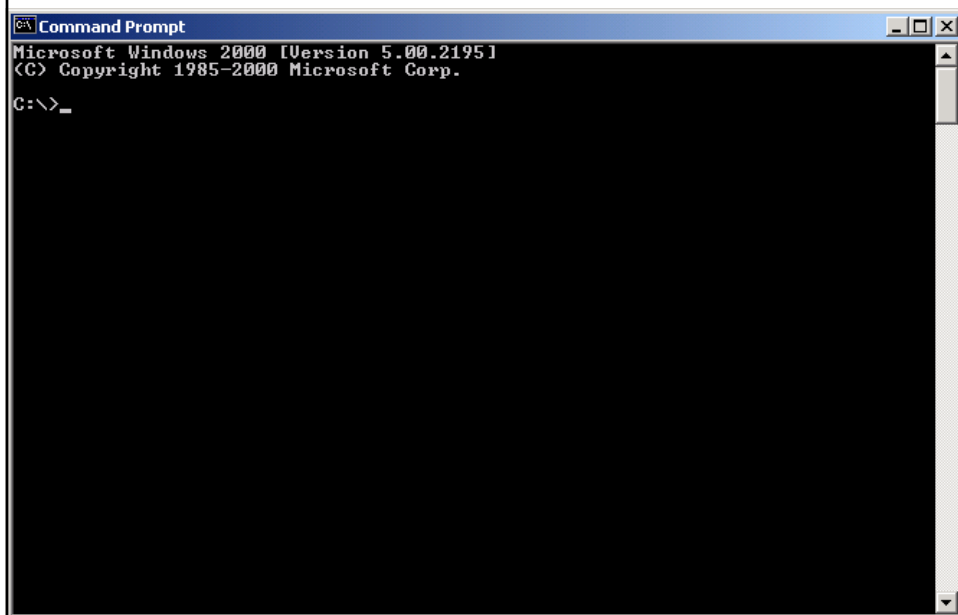
- ❖ Building a collection
- ➔ ❖ The dreaded black screen
- ❖ More on building



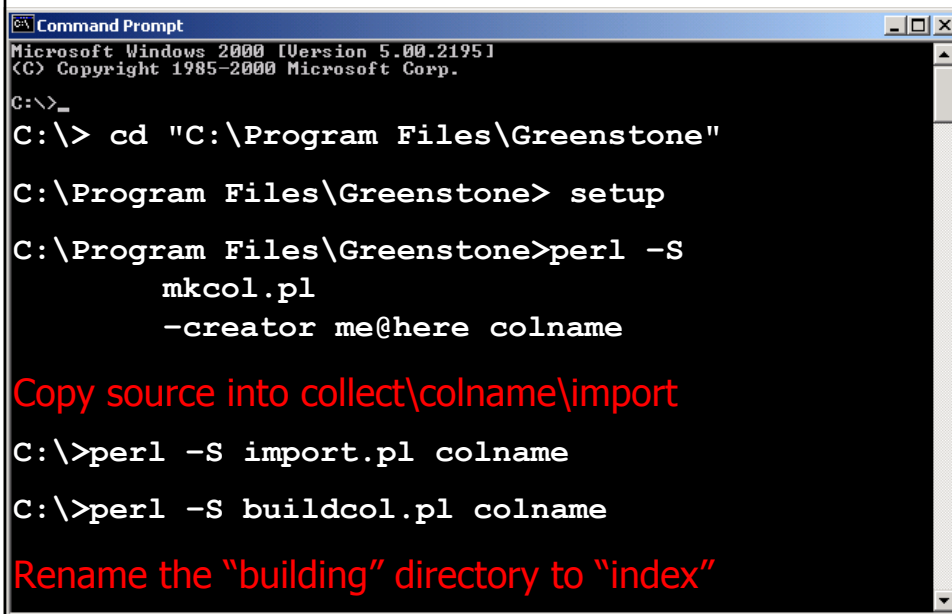
Start a command prompt



Command Prompt



The building process



```
Command Prompt
Microsoft Windows 2000 [Version 5.00.2195]
(C) Copyright 1985-2000 Microsoft Corp.

G:\>_
C:\> cd "C:\Program Files\Greenstone"

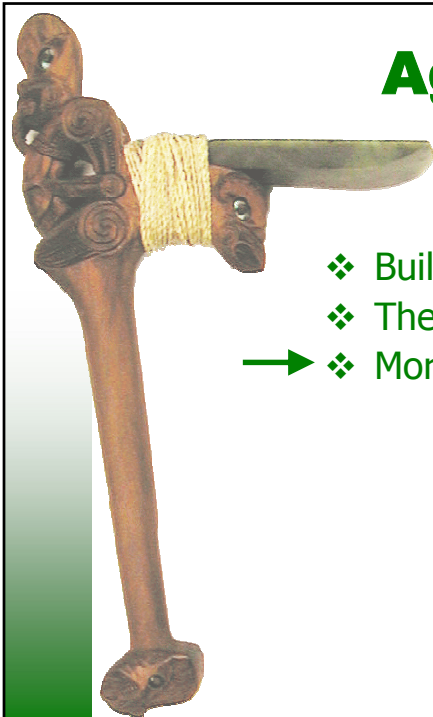
C:\Program Files\Greenstone> setup

C:\Program Files\Greenstone>perl -S
    mkcol.pl
    -creator me@here colname

Copy source into collect\colname\import
C:\>perl -S import.pl colname
C:\>perl -S buildcol.pl colname

Rename the "building" directory to "index"
```

Agenda



- ❖ Building a collection
- ❖ The dreaded black screen
- ❖ More on building

Specifying metadata: XML metadata file

```
<?xml version="1.0" ?>
<!DOCTYPE GreenstoneDirectoryMetadata SYSTEM
"http://greenstone.org/dtd/GreenstoneDirectoryMetadata/1.0/GreenstoneD
irectoryMetadata.dtd">
<DirectoryMetadata>
  <FileSet>
    <FileName>nugget.*</FileName>
    <Description>
      <Metadata name="Title">Nugget Point, The Catlins</Metadata>
      <Metadata name="Place" mode="accumulate">Nugget Point</Metadata>
    </Description>
  </FileSet>
  <FileSet>
    <FileName>nugget-point-1.jpg</FileName>
    <Description>
      <Metadata name="Title">Nugget Point Lighthouse</Metadata>
      <Metadata name="Subject">Lighthouse</Metadata>
    </Description>
  </FileSet>
</DirectoryMetadata>
```

XML metadata format

Document type definition (DTD)

```
<!DOCTYPE GreenstoneDirectoryMetadata [
  <!ELEMENT DirectoryMetadata (FileSet*)>
  <!ELEMENT FileSet (FileName+,Description)>
  <!ELEMENT FileName (#PCDATA)>
  <!ELEMENT Description (Metadata*)>
  <!ELEMENT Metadata (#PCDATA)>
  <ATTLIST Metadata name CDATA #REQUIRED>
  <ATTLIST Metadata mode (accumulate|override) "override">
]>
```

Greenstone Archive Format: Example document

```
<?xml version="1.0" ?>
<!DOCTYPE GreenstoneArchive SYSTEM
"http://greenstone.org/dtd/GreenstoneArchive/1.0/GreenstoneArchive.dtd">
<Section>
  <Description>
    <Metadata name="gsdlsourcefilename">ec158e.txt</Metadata>
    <Metadata name="Title">Freshwater Resources in Arid Lands</Metadata>
    <Metadata name="Identifier">HASH0158f56086efffe592636058</Metadata>
    <Metadata name="gsdlassocfile">cover.jpg:image/jpeg</Metadata>
    <Metadata name="gsdlassocfile">p07a.png:image/png</Metadata>
  </Description>
  <Section>
    <Description>
      <Metadata name="Title">Preface</Metadata>
    </Description>
    <Content>
      This is the text of the preface
    </Content>
  </Section>
  <Section>
    <Description>
      <Metadata name="Title">First and only chapter</Metadata>
    </Description>
    <Section>
      <Description>
        <Metadata name="Title">Part 1</Metadata>
      </Description>
      <Content>
        This is the first part of the first and only chapter
      </Content>
    </Section>
  </Section>
</Section>
```

Greenstone archive format

Document type definition (DTD)

```
<!DOCTYPE GreenstoneArchive [
  <!ELEMENT Section (Description,Content,Section*)>
  <!ELEMENT Description (Metadata*)>
  <!ELEMENT Content (#PCDATA)>
  <!ELEMENT Metadata (#PCDATA)>
  <ATTLIST Metadata name CDATA #REQUIRED>
]>
```

demo/
index/
demo.ldb

Document database

```
[42]
<section>HASH863cfd85c90056aeb66bc3.7.1
-----
[HASH863cfd85c90056aeb66bc3.7.1]
<doctype>doc
<hastxt>1
<Title>National park restoration in Chad: luxury or necessity ?
<docnum>42
-----
[HASH863cfd85c90056aeb66bc3.8]
<doctype>doc
<hastxt>0
<Title>Developing World
<childtype>VList
<contains>".1;".2
<docnum>43
-----
[CL1]
<doctype>classify
<hastxt>0
<childtype>VList
<Title>Subject
<numleafdocs>17
<thistype>Invisible
<contains>".1;".2;".3;".4;".5;".6
-----
[CL1.2]
<doctype>classify
<hastxt>0
<childtype>VList
<Title>Communication, Information and Documentation
<numleafdocs>1
<contains>".1
<mdoffset>
```