



Introducing the Greenstone Digital Library Software

Ian H. Witten
 Computer Science Department
 Waikato University
 New Zealand
<http://greenstone.org>
<http://nzdl.org>

Tutorial exercises

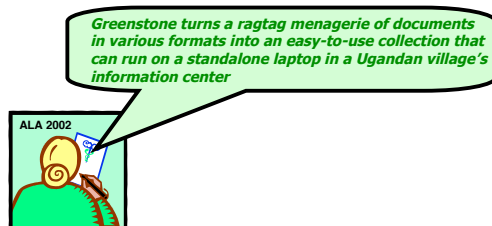
Installing Greenstone (local library version) (#1)
 Small collection of HTML files (#2)
 Simple image collection (#3)
 Collection of Word and PDF files (#4)
 The Tudor collection of HTML files (#5, 6, 7, 8)
 The Beatles multimedia collection (#9, 10)
 The Niupepa newspaper collection (#11, 12)
 Customizing the appearance of collections (#13)
 A bibliographic collection (#14)
 Open Archives Initiative (OAI) collection (#15, 16)
 Using METS as Greenstone's internal representation (#17)
 Greenstone—DSpace conversion (#18, 19)



Agenda

- Overview**
 - What does Greenstone do?
 - Greenstone facts; standards
 - Reader's Interface: examples of collections
- Librarian interface**
 - Build a collection in 30 sec (Hobbits)
 - Add metadata, classifiers, search indexes (Images)
 - Build a multimedia collection (Beatles)
- Going further**
 - ContentDM, DSpace lookalikes
 - Remote GLI
 - Working with big collections (for masochists only)
 - Under the hood: collection configuration file
- Customizing the look**
 - Different interface languages
 - Customizing with macros
 - Collection-specific macros, styles, stylesheets
 - Personalizing your home page
 - Examples of what others have done
- Reaching out**
 - Serving and acquiring OAI
 - DSpace and METS
 - Greenstone3

What we wanted



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 removable media
- ❖ Run anywhere, on any computer
- ❖ Fully internationalized
- ❖ Non-exclusive: documents and metadata in any format
- ❖ Non-prescriptive: standard and non-standard metadata

Tutorial exercise #1 Installing Greenstone (local library version)

Self-installing CD-ROM
 Greenstone
 ImageMagick: **yes**
 Ghostscript: **yes**
 Copy *sample_files* to desktop
 Try the reader's interface
 Librarian interface
 Navigate to **Home Folder\Desktop\sample_files**
Right-click

What we got: Greenstone

- Access**
- ❖ Accessible via any Web browser
 - ❖ Server runs on anything (all Windows + Unix + Mac)
 - ❖ Collections can be published on CD-ROM/DVD
 - ❖ Trivial to install
 - ❖ GUI interface for building and publishing collections
- Searching/ browsing**
- ❖ Collection-specific
 - ❖ Full-text and fielded search
 - ❖ Flexible browsing facilities
 - ❖ Metadata-based (Dublin Core recommended)
 - ❖ Creates all access structures automatically
- Extensible**
- ❖ Plugins — new document, metadata formats
 - ❖ Classifiers — new metadata browsers
- Multi-***
- ❖ Multilingual: Documents *and* interfaces
 - ❖ Multimedia: image, video, audio collections exist
 - ❖ Multiformat: Documents *and* metadata

UNESCO: Distributing Greenstone DL software

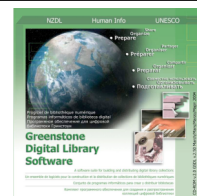
Sustainable development

*"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>



Greenstone facts

- Distribution**
- ❖ Open source: Gnu GPL
 - ❖ Distributed via SourceForge since: Nov 2000
 - ❖ Average downloads: 5000/month since Nov 2000
 - ❖ Humanitarian CD-ROMs produced: 30
 - ❖ Distribution for each one: 5000/year
- International**
- ❖ Languages for interface: 50
 - ❖ Languages for full software + manual: 10
 - ❖ Countries represented on email lists: 30
 - ❖ UNESCO training courses in: 10
- UN Agencies**
- ❖ UNESCO, Paris ("Information for All")
 - ❖ FAO, Rome (Info Management Resources)
 - ❖ UNU, Japan (CD-ROM collections of UNU)
- Technical centers**
- ❖ University of Waikato, New Zealand
 - ❖ Indian Institute of Sciences, Bangalore
 - ❖ University College, London
 - ❖ University of Cape Town, South Africa
 - ❖ University of Lethbridge, Canada



Sample collections (see greenstone.org)

International

Argentina Human Rights Commission
Tasmania State Library
Peking University Digital Library
Gresham College, London
Oxford Digital Library, University of Oxford
University of Applied Sciences, Stuttgart
Association of Indian Labour Historians, Delhi
Indian Institute of Management, Kozhikode
Indian Institute of Science, Bangalore
Vimercate Public Library, Milan, Italy
Netherlands Institute for Scientific Information Services
Philippine Government Information Network
Mari El Republic, Russia
Slavonski Brod Public Library, Slovenia
Vietnam National University
Welsh Books Council

Argentina
Australia
China
England
Germany
India
India
Italy
Netherlands
Philippines
Russia
Slovenia
Vietnam
Wales

U.S.

Auburn University, Alabama
Detroit Public Library
Hawaiian Electronic Library
ibiblio project, University of North Carolina
Illinois Wesleyan University
Lehigh University, Pennsylvania
New York Botanical Garden
University of California at Riverside
University of Chicago Library
University of Illinois
Texas A&M University
Washington Research Library Consortium

Standards

- Metadata**
- ❖ Can use any metadata set, Dublin Core supplied
 - ❖ Plugins for
- | | |
|----------|--------|
| XML | Refer |
| MARC | OAI |
| CDS/ISIS | METS |
| ProCite | DSPACE |
| BibTex | |
- METS can be used as Greenstone's internal representation
- Serving**
- ❖ Web
 - ❖ Can publish Greenstone collections on CD-ROM
 - ❖ Can publish Greenstone collections on OAI
 - ❖ Export collections to METS
 - ❖ Export collections to DSpace (ready for DSpace's batch import program)
- Documents**
- ❖ Plugins for
- | | | |
|------------|-------------|------------------------------|
| PDF | ZIP | Images (GIF, JPEG, TIFF ...) |
| PostScript | Excel | MP3 |
| Word, RTF | PPT | Ogg Vorbis |
| HTML | Email | MediaWiki |
| Plain text | Source code | UnknownPlug |
| Latex | RealMedia | (e.g. for audio, MPEG, Midi) |

Greenstone: Platforms

- ❖ Operating system:
 - Windows (any version)
 - Linux (any version)
 - Unix (most versions, e.g. Solaris)
 - Mac OS X
- ❖ Restrictions:
 - No longer runs under Windows 3.1/3.11
 - For Librarian interface (GLI), need Java—which is no longer supported on Windows 95

Local library vs Web library

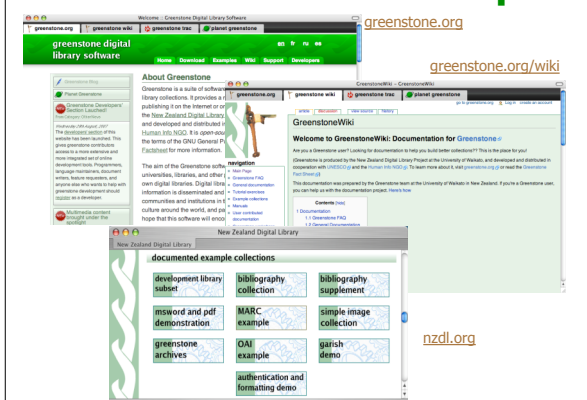
- ❖ **Local library: stand-alone**
 - Serves collections on a standalone PC ...
 - And on others on the same network
 - Includes built-in Web server
- ❖ **Web library: uses external web server**
 - Apache, Microsoft PWS/IIS
 - Need to configure web server (geeky job?)
- ❖ **Windows: Both local library and web library**
 - Normally use local library (else must set up server)
 - Web library works with Microsoft PWS, IIS
- ❖ **Unix, Mac OS/10: Web library only**
 - Use Apache (or other web server)
 - Linux binaries supplied
 - Tested on SUN Solaris, Mac OS/10
 - Need GDBM (standard on Linux)
- ❖ **Greenstone developed on Linux**

Documentation and help

Manuals on the CD-ROM (gsdl/docs)

- **From Paper To Collection** (30pp)
 - Scanners and scanning, OCR, 3 examples – from 1,000 to 100,000 pages, Creating an electronic collection
- **Installer's Guide** (36pp)
 - Versions of Greenstone, installation procedure, Greenstone collections, setting up the web server, configuring your site, personalizing your installation
- **User's Guide** (90pp)
 - Overview of Greenstone, using Greenstone collections, the collector, administration, software features, glossary of terms
- **Developer's Guide** (113pp)
 - Understanding the collection building process, getting the most out of your collections, the Greenstone runtime systems, configuring your Greenstone site

Documentation and help



Example collections

development library subset

The DLS collection has the same structure as the Greenstone demo collection. It's fairly complex, and if you're just starting out you might prefer to look at some other collections first (e.g. [msword and pdf demonstration](#), or the [greenstone archives](#), or the [simple image collection](#)).

msword and pdf demonstration

This contains a few documents in PDF, MSWord, RTF, and Postscript formats, demonstrating the ability to build collections from documents in different formats. The collection configuration file is very simple.

greenstone archives

A collection of email messages from the Greenstone mailing list archives, this uses the Email plugin, which parses files in email formats. The collection configuration file is very simple.

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).

Example 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.

simple image collection

This very basic image collection contains no text and no explicit metadata -- which makes it rather unrealistic. The configuration file is about as simple as you can get.

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.

garish demo

This collection also contains the same material as the Greenstone demo. Its appearance has been altered to show how the pages generated can be set out differently. It relies on a non-standard macro file that is supplied with Greenstone.

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 .jdt files and converts them to Greenstone metadata.

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\$1
- ❖ Win3.1x upward
- ❖ Stand-alone
- ❖ and intranet server
- ❖ Web browser user interface

Global Help Project, Antwerp (+ UN agencies)



Agenda

Overview

- What does Greenstone do?
- Greenstone facts; standards
- Reader's Interface: examples of collections

Librarian interface

- Build a collection in 30 sec (Hobbits)
- Add metadata, classifiers, search indexes (Images)
- Build a multimedia collection (Beatles)

Going further

- ContentDM, DSpace lookalikes
- Remote GLI
- Working with big collections (for masochists only)
- Under the hood: collection configuration file

Customizing the look

- Different interface languages
- Customizing with macros
- Collection-specific macros, styles, stylesheets
- Personalizing your home page
- Examples of what others have done

Reaching out

- Serving and acquiring OAI
- DSpace and METS
- Greenstone3



Overview

- What does Greenstone do?
- Greenstone facts; standards
- Reader's Interface: examples of collections

Librarian interface

- Build a collection in 30 sec (Hobbits)
- Add metadata, classifiers, search indexes (Images)
- Build a multimedia collection (Beatles)

Going further

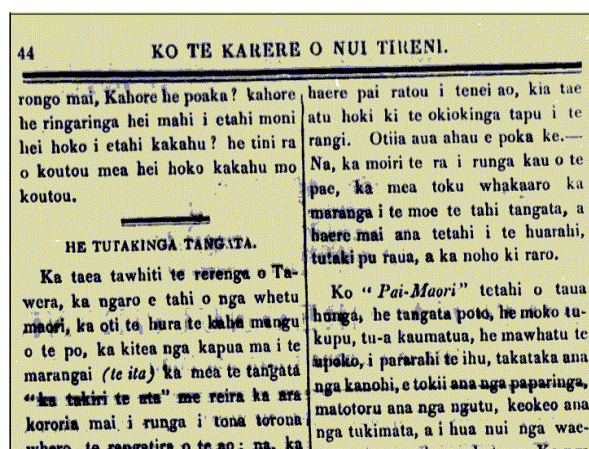
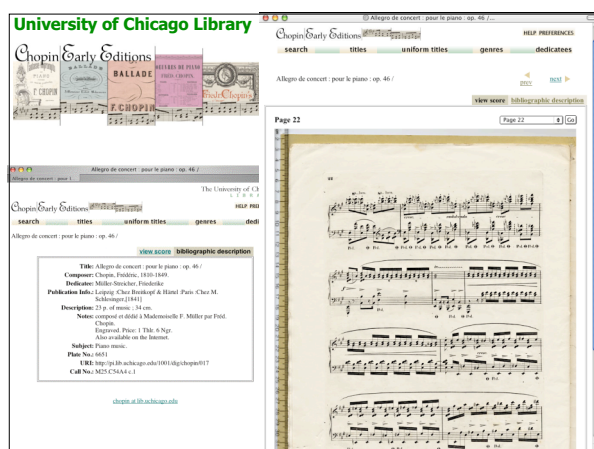
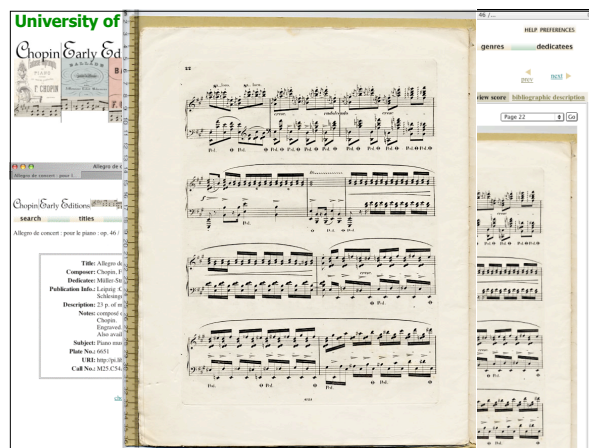
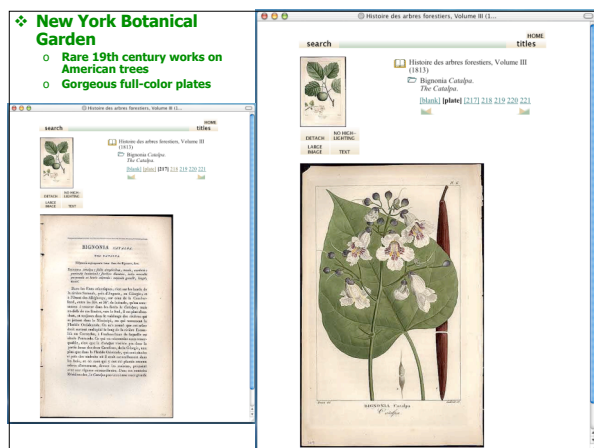
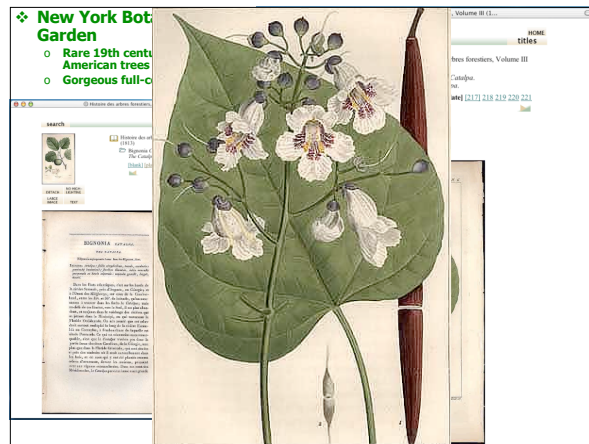
- ContentDM, DSpace lookalikes
- Remote GLI
- Working with big collections (for masochists only)
- Under the hood: collection configuration file

Customizing the look

- Different interface languages
- Customizing with macros
- Collection-specific macros, styles, stylesheets
- Personalizing your home page
- Examples of what others have done

Reaching out

- Serving and acquiring OAI
- DSpace and METS
- Greenstone3



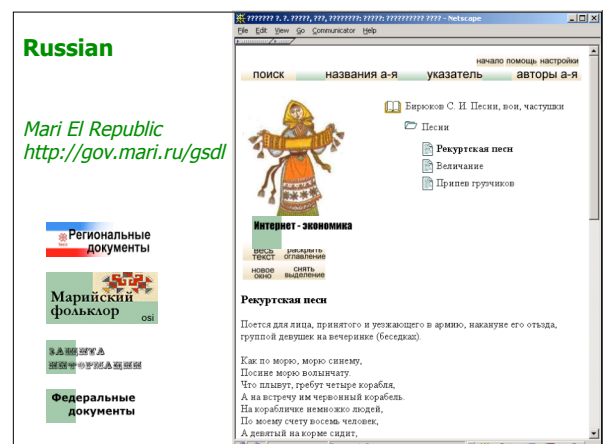


Chinese documents (pictures of text) + Chinese interface

Peking University Library



Classic Chinese literature



Agenda



Overview
What does Greenstone do?
Greenstone facts; standards
Reader's Interface: examples of collections

Librarian interface
Build a collection in 30 sec (Hobbits)
Add metadata, classifiers, search indexes (Images)
Build a multimedia collection (Beatles)

Going further
ContentDM, DSpace lookalikes
Remote GLI
Working with big collections (for masochists only)
Under the hood: collection configuration file

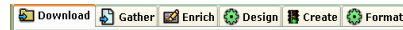
Customizing the look
Different interface languages
Customizing with macros
Collection-specific macros, styles, stylesheets
Personalizing your home page
Examples of what others have done

Reaching out
Serving and acquiring OAI
DSpace and METS
Greenstone3

The Greenstone Librarian Interface (GLI)

- ❖ Building collections
- ❖ Interactive Java program
- ❖ Runs on anything
- ❖ Build a collection on the computer you are on
- ❖ ... plus new applet version
- ❖ Includes metadata editor

Caveat: cannot deal with such huge metadata collections as Greenstone can

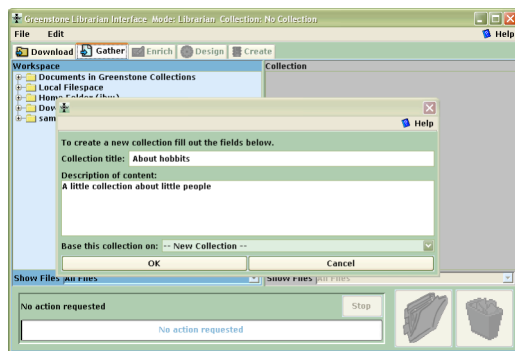


(Tutorial exercise: small collection of HTML files)

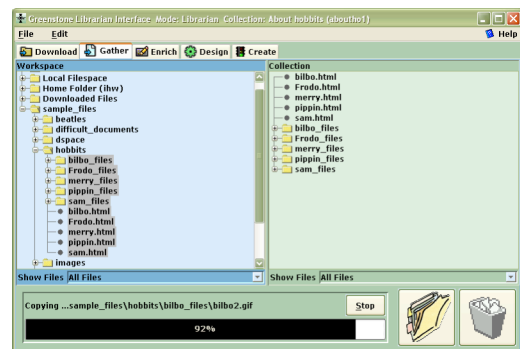
Invoke GLI: build a small collection of HTML files

- ❖ Gather
- ❖ Create
- ❖ Look at extracted metadata
- ❖ Set up shortcut in the Librarian interface

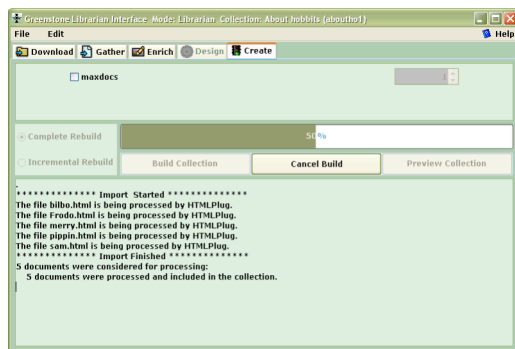
Create a new collection



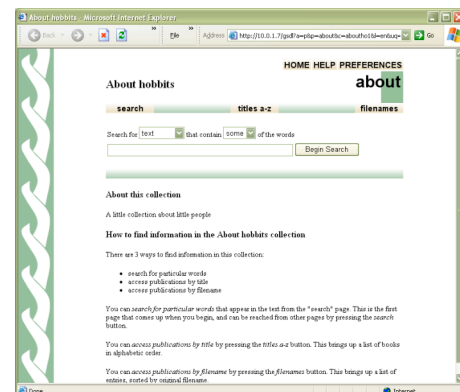
Gather: Gather the files together



Create: Build the collection



Preview: admire the result



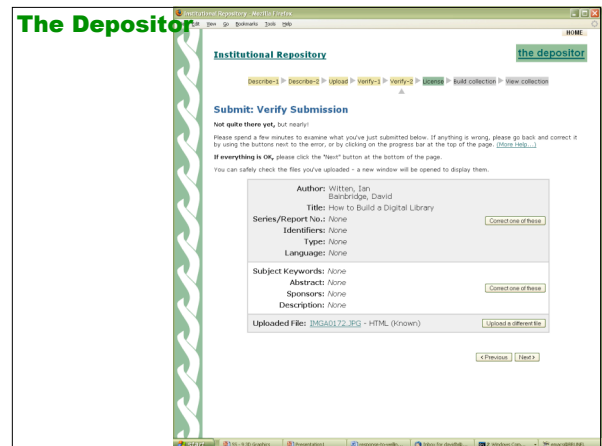
Serving and acquiring OAI
DSpace and METS
Greenstone3

2007

The Depositor



The Depositor



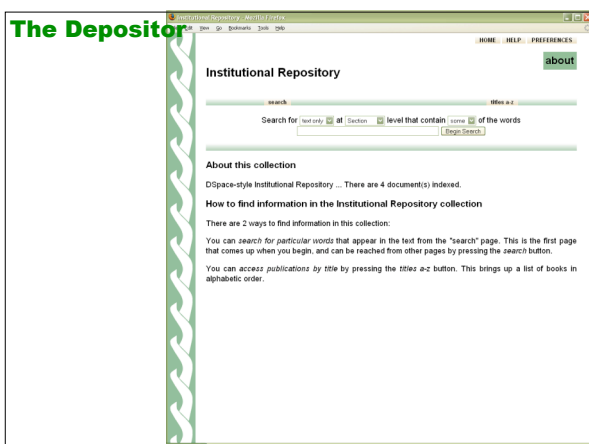
The Depositor



The Depositor



The Depositor



Remote GLI/Applet version

❖ Collection on remote Greenstone server

❖ Edit it locally

▪ gather

▪ enrich

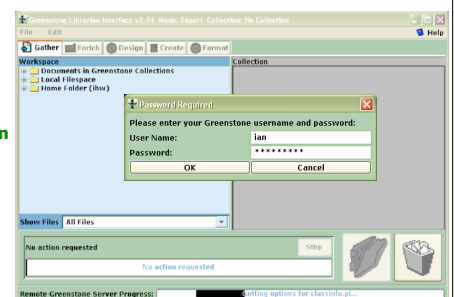
▪ design

▪ create

❖ Authentication

▪ uses admin facility

❖ Locking



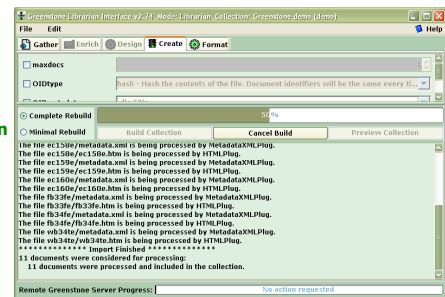
Remote GDI/Applet version

- ❖
- Collection on remote Greenstone server**

- ❖ **Edit it locally**
 - gather
 - enrich
 - design
 - create

- ❖ **Authentication**
 - uses admin facility

- #### ❖ Locking



Building a collection

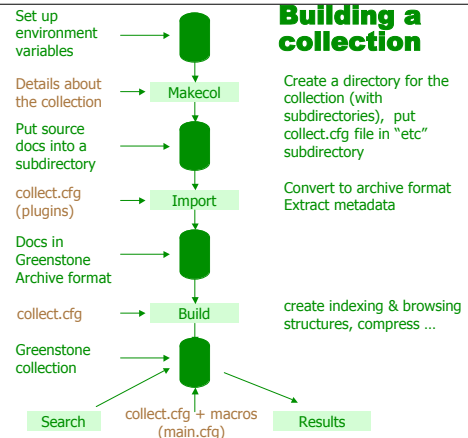
Create a directory for the collection (with subdirectories), put collect.cfg file in "etc" subdirectory

Convert to archive format
Extract metadata

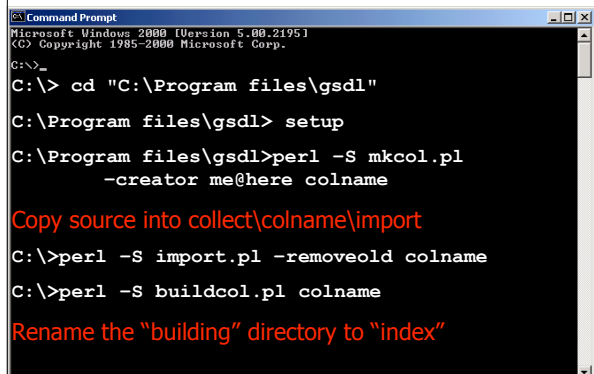
- create indexing & browsing structures, compress ...

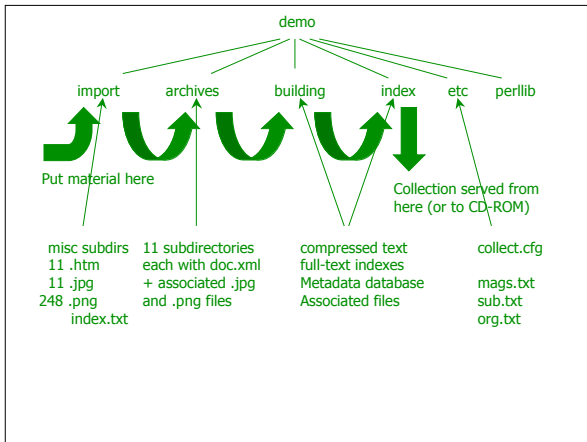
Results

Serving and acquiring OAI
DSpace and METS
Greenstone3



The building process





How big?

PAPERSPAST

today

- 50,000 searchable newspaper issues
- 250,000 pages
- 3M articles
- 9 GB of raw text
- 27 GB of metadata
- 50M unique terms (uncorrected OCR)
- 900,000 non-searchable pages
- All pages searchable
- All pages searchable
- DLConsulting Ltd
- uses Lucene
- Input: METS/ALTO XML files
- National Library of Singapore
- Mar 2008
- 600,000 searchable newspaper issues
- mid 2009
- 2M pages
- Will continue to grow

Papers Past Home Introduction Search

Search News

mid 2008

mid 2009

Find out more

NATIONAL LIBRARY OF NEW ZEALAND

Agenda

- Overview**
 - What does Greenstone do?
 - Greenstone facts; standards
 - Reader's Interface: examples of collections
- Librarian interface**
 - Build a collection in 30 sec (Hobbits)
 - Add metadata, classifiers, search indexes (Images)
 - Build a multimedia collection (Beatles)
- Going further**
 - ContentDM, DSpace lookalikes
 - Remote GLI
 - Working with big collections (for masochists only)
 - Under the hood: collection configuration file
- Customizing the look**
 - Different interface languages
 - Customizing with macros
 - Collection-specific macros, styles, stylesheets
 - Personalizing your home page
 - Examples of what others have done
- Reaching out**
 - Serving and acquiring OAI
 - DSpace and METS
 - Greenstone3

Under the hood: 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        ArchPlug
plugin        RecPlug

classify      Hierarchy -hfile sub.txt -metadata Subject -sort Title
classify      SOLRlist -metadata Title
classify      Hierarchy -hfile org.txt -metadata Organization -sort Title
classify      List -metadata Howto

format        SearchWList "<td valign=top>[link][icon][link]</td>
<td>[if]{{parent(All': '):Title},{parent(All': '):Title}: }
[link][Title][link]</td>"
format        CLAVList "<tr>[link][Howto][link]"
format        DocumentImages true
format        DocumentText "<h3>[Title]</h3>\n\n[Text]"

collectionmeta collectionname "greenstone demo"
collectionmeta collectionextra "This is a demonstration collection for the
Greenstone digital library software.\nIt contains a small
subset (11 books) of the Humanity Development Library"
collectionmeta iconcollectionmall "/gsdl/collect/demo/images/demo.gif"
collectionmeta iconcollection "/gsdl/collect/demo/images/demo.gif"
collectionmeta section:Title "section titles"
collectionmeta document:text "entire books"
collectionmeta section:text "chapters"
  
```

- name, icon, etc
- description
- email of creator
- search indexes
- plugins
- classifiers

how to format

- documents
- query results
- classifiers

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
- Import OAI metadata
- Extract phrase hierarchy and add browser
- Alter the format of any of the above
- Restrict collection's interface langs
- Change default interface language

indexes	document:Title
indexes	document:Creator
classify	AZList -metadata Creator
plugin	WordPlug
plugin	PDFPlug
languages	en fr es
plugin	PDFPlug -extract_acronyms
plugin	OAIPlug
classify	Phind
format ...	
format	PreferenceLangs en fr es
cgiarg	shortname=1 argdefault =fr

Agenda

- Overview**
 - What does Greenstone do?
 - Greenstone facts; standards
 - Reader's Interface: examples of collections
- Librarian interface**
 - Build a collection in 30 sec (Hobbits)
 - Add metadata, classifiers, search indexes (Images)
 - Build a multimedia collection (Beatles)
- Going further**
 - ContentDM, DSpace lookalikes
 - Remote GLI
 - Working with big collections (for masochists only)
 - Under the hood: collection configuration file
- Customizing the look**
 - Different interface languages
 - Customizing with macros
 - Collection-specific macros, styles, stylesheets
 - Personalizing your home page
 - Examples of what others have done
- Reaching out**
 - Serving and acquiring OAI
 - DSpace and METS
 - Greenstone3

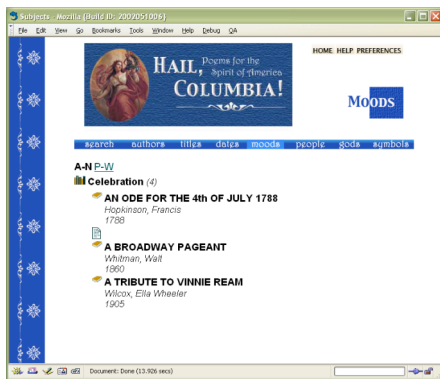
Illinois student project



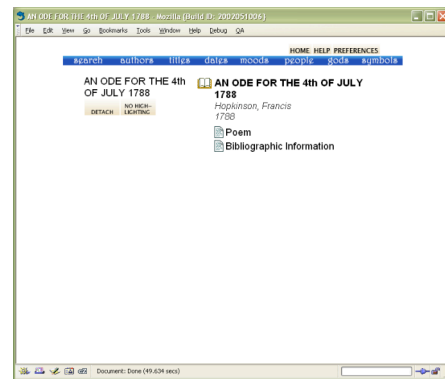
Illinois student project



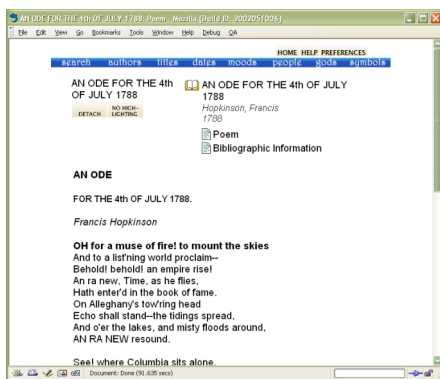
Illinois student project



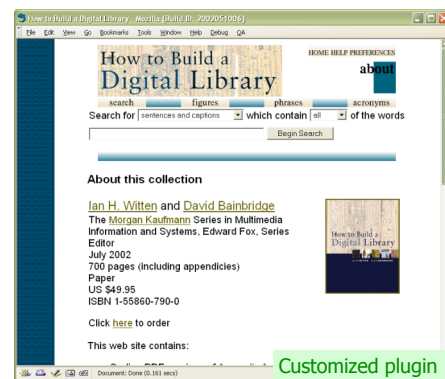
Illinois student project



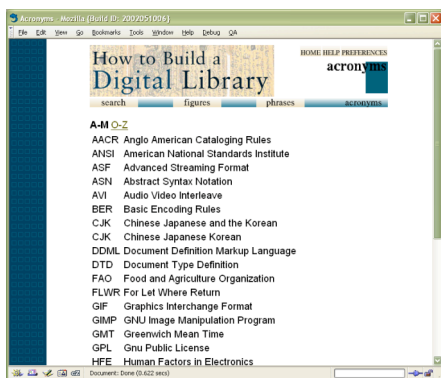
Illinois student project



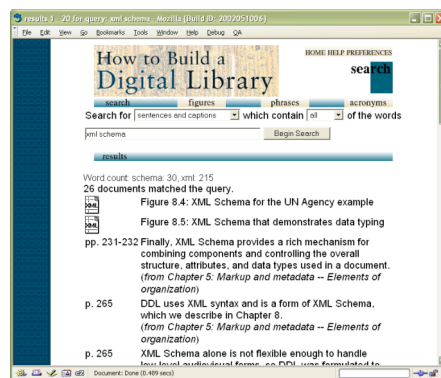
"How to build" web site



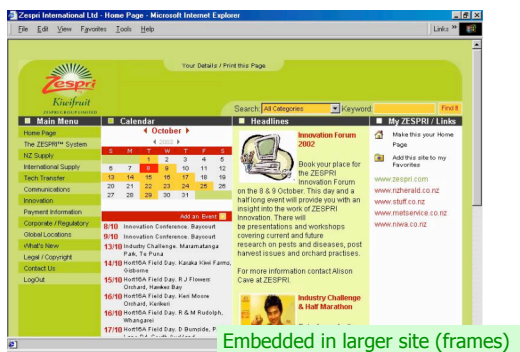
“How to build” web site



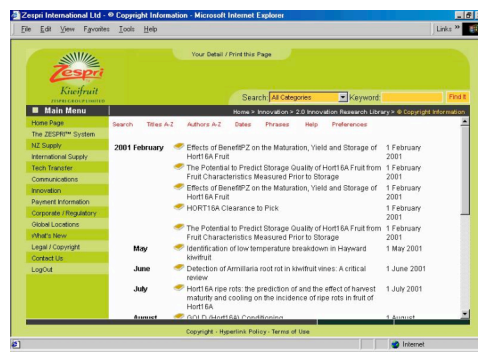
“How to build” web site



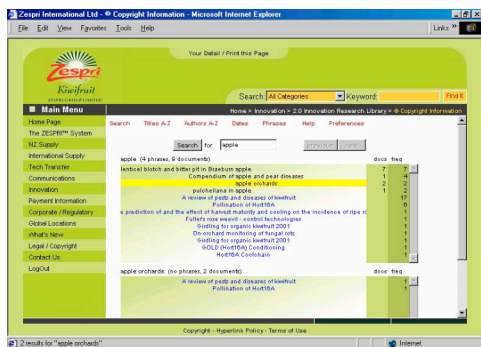
Zespri innovation library



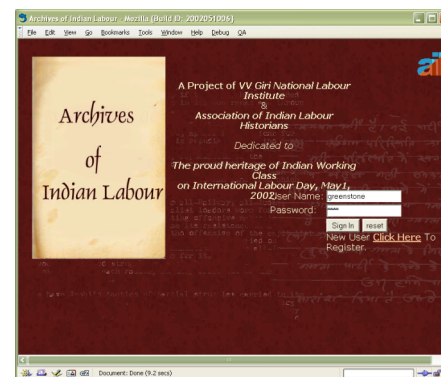
Zespri innovation library



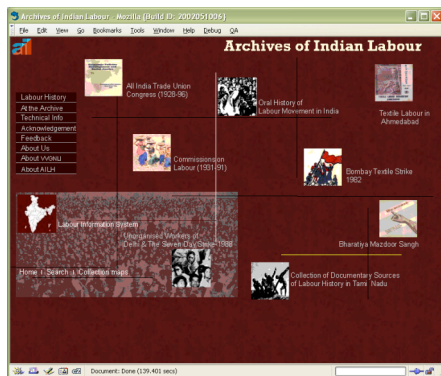
Zespri innovation library



Archives of Indian labour



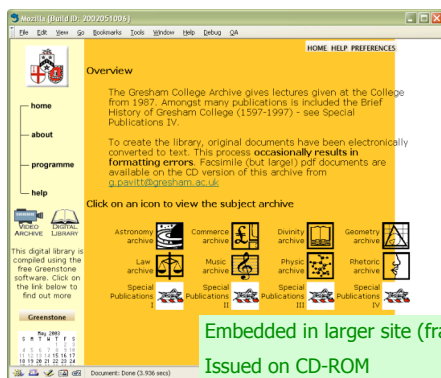
Archives of Indian labour



Archives of Indian labour

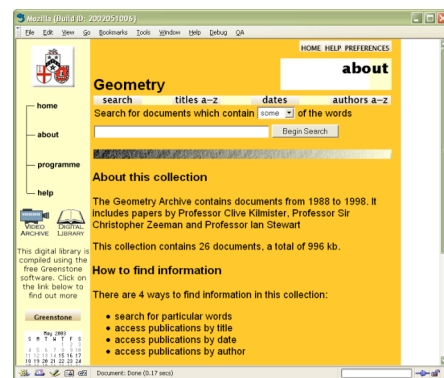


Gresham College archive

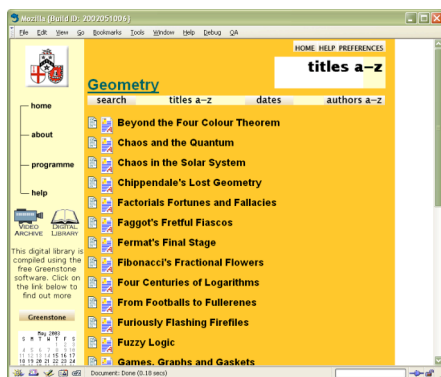


Embedded in larger site (frames)
Issued on CD-ROM

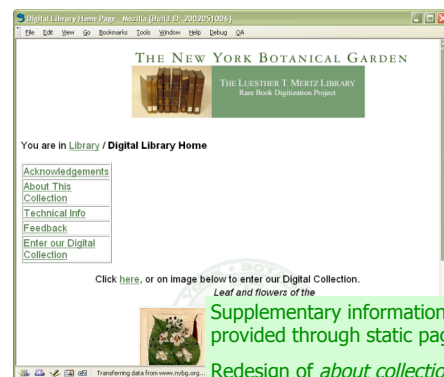
Gresham College archive



Gresham College archive

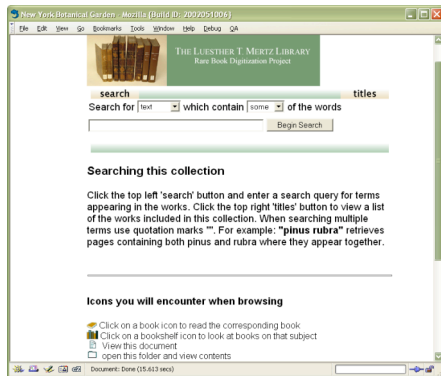


The NY Botanical Gardens

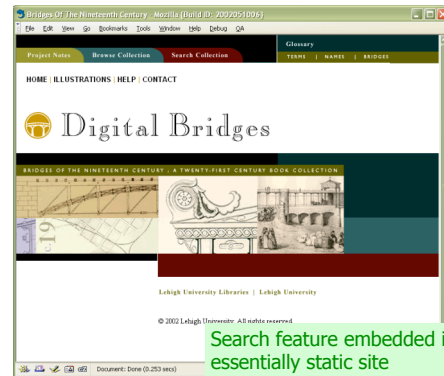


Supplementary information
provided through static pages
Redesign of *about* collection page

The NY Botanical Gardens

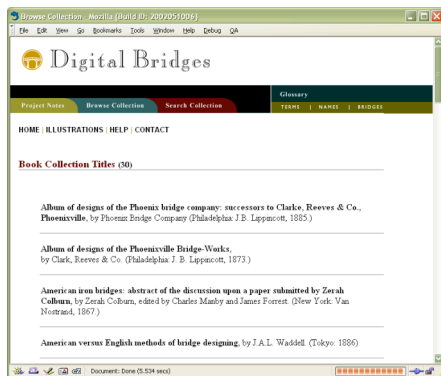


Lehigh University

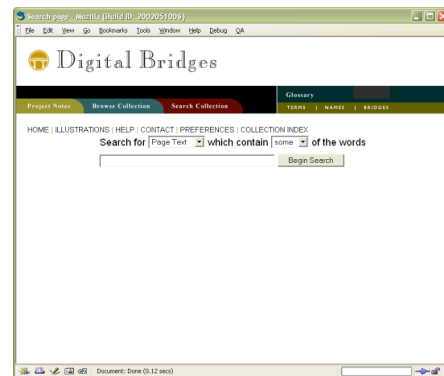


Search feature embedded into essentially static site

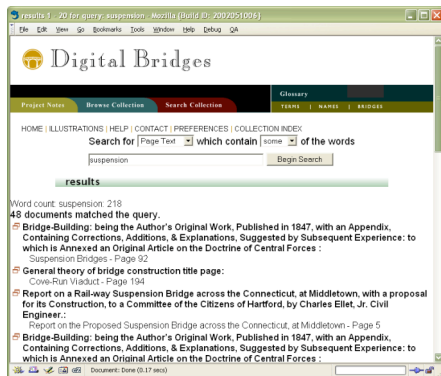
Lehigh University



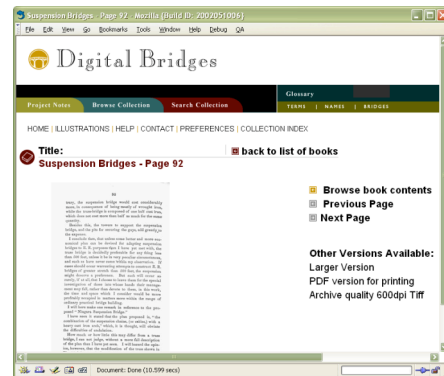
Lehigh University



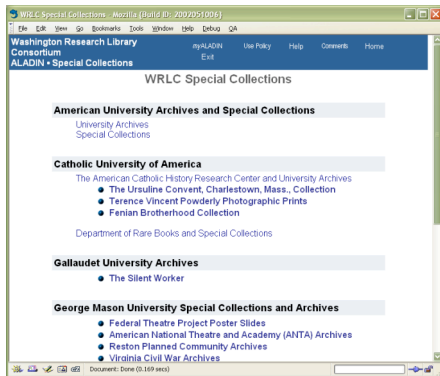
Lehigh University



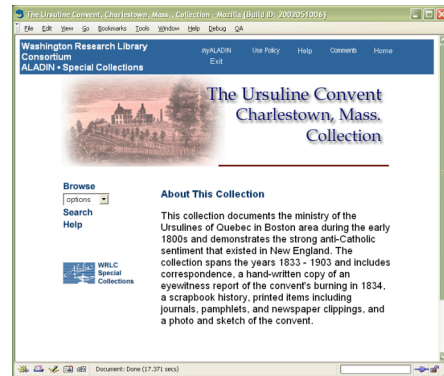
Lehigh University



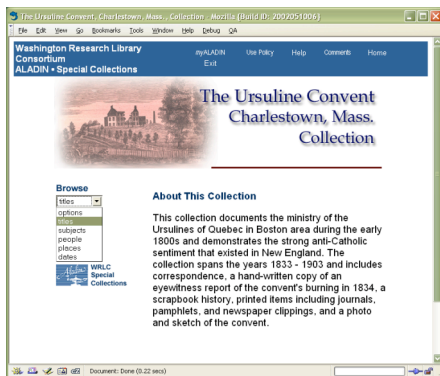
Aladin digital library system



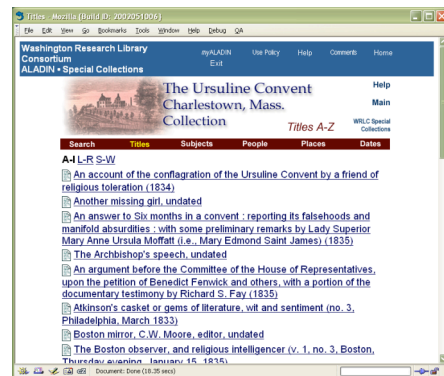
Aladin digital library system



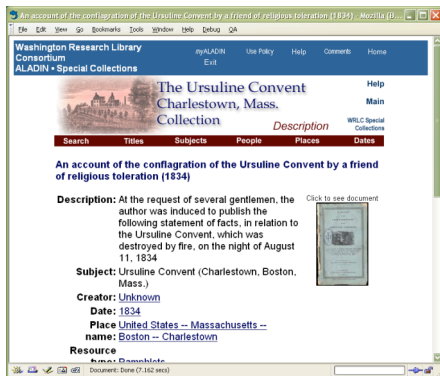
Aladin digital library system



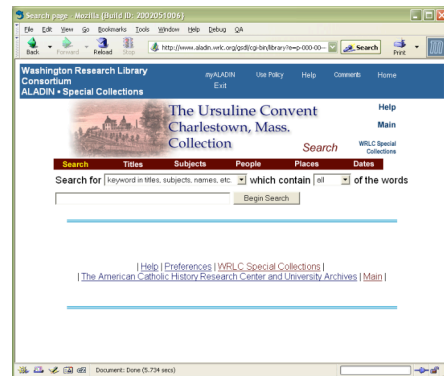
Aladin digital library system



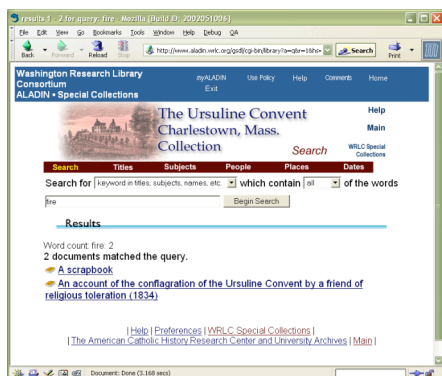
Aladin digital library system



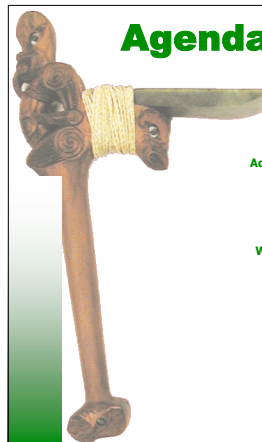
Aladin digital library system



Aladin digital library system



Agenda

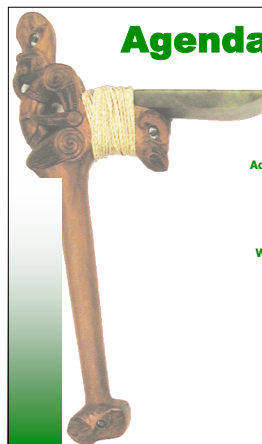


- Overview**
 - What does Greenstone do?
 - Greenstone facts; standards
 - Reader's Interface: examples of collections
- Librarian interface**
 - Build a collection in 30 sec (Hobbits)
 - Add metadata, classifiers, search indexes (Images)
 - Build a multimedia collection (Beatles)
- Going further**
 - ContentDM, DSpace lookalikes
 - Remote GLI
 - Working with big collections (for masochists only)
 - Under the hood: collection configuration file
- Customizing the look**
 - Different interface languages
 - Customizing with macros
 - Collection-specific macros, styles, stylesheets
 - Personalizing your home page
 - Examples of what others have done
- Reaching out**
 - Serving and acquiring OAI
 - DSpace and METS
 - Greenstone3

The Greenstone documentation wiki

❖ go to greenstone.org

Agenda



- Overview**
 - What does Greenstone do?
 - Greenstone facts; standards
 - Reader's Interface: examples of collections
- Librarian interface**
 - Build a collection in 30 sec (Hobbits)
 - Add metadata, classifiers, search indexes (Images)
 - Build a multimedia collection (Beatles)
- Going further**
 - ContentDM, DSpace lookalikes
 - Remote GLI
 - Working with big collections (for masochists only)
 - Under the hood: collection configuration file
- Customizing the look**
 - Different interface languages
 - Customizing with macros
 - Collection-specific macros, styles, stylesheets
 - Personalizing your home page
 - Examples of what others have done
- Reaching out**
 - Serving and acquiring OAI
 - DSpace and METS
 - Greenstone3

Greenstone 3

Greenstone 2 was designed about 7 years ago. Since then, new software and library standards have emerged ...

- ❖ **Easier to modify interface: generates data (XML) rather than presentation (HTML)**
 - XML can be transformed using XSLT, rather than HTML via macro files
- ❖ **More dynamic: more can be done at runtime**
 - e.g. sort search results on the fly, generate new classifiers at runtime
- ❖ **METS foundation: multiple hierarchies possible**
 - e.g. chapter-section-subsection hierarchy can coexist with page header-contents-footer hierarchy
- ❖ **Distributed modules, using SOAP**
 - Interface can be on different computer to collection, can talk to several different Greenstone sites
- ❖ **Java basis: platform independent**
 - Greenstone2 operates on different platforms, but at a high cost to implementors
- ❖ **Better software technology**
 - easier for people (e.g. our students!) to modify the software to make it do radically different things