

Advanced document types



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



- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ❖ PowerPoint documents
- ❖ CDS/ISIS

Print documents

- ❖ Need to be converted to an electronic form – scanning produces a set of images
- ❖ To add each page as an individual image, process using ImagePlug
- ❖ To group them into a single document, process using PagedImgPlug
 - Requires an 'item' file which lists all the pages and gives additional metadata

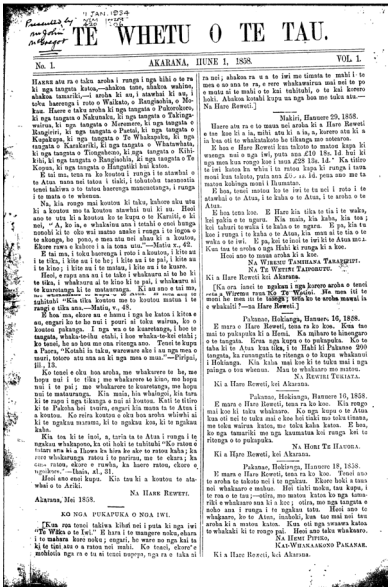
Print documents

- ❖ Can add metadata to the images to enable searching.
- ❖ If full text searching is desired, use OCR (Optical Character Recognition) to generate an electronic version of the text
- ❖ Alternatively, if the documents are small and few, manually type the text into a file.
- ❖ Text files can be included with the images in the item file.

Sample document

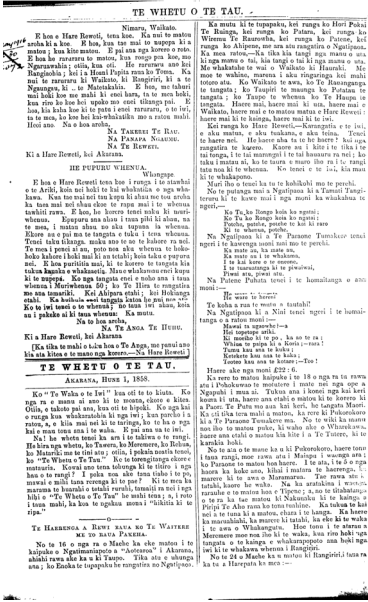
- ❖ 4 newspaper page images and their scanned text
- ❖ 10_1_1.item
- ❖ images/10_1_1_1.gif, 10_1_1_2.gif, 10_1_1_3.gif, 10_1_1_4.gif
- ❖ text/10_1_1_1.txt, 10_1_1_2.txt, 10_1_1_3.txt, 10_1_1_4.txt

10_1_1_1



TE WHETU O TE TAU.
No. 1.
AKARANA, HUNE 1, 1858.
VOL. 1.
HAERE atu ra e taku aroha i runga i nga hihi o te ra ki nga tangata katoa,—ahakoa tane, ahakoa wahine, ahakoa tamariki,—i aroha ki au, i atawhai ki au, i toku haerenga i roto o Waikato, o Rangiaohia, o Mokau. Haere e taku aroha ki nga tangata o Pukorokoro, ki nga tangata o Nakunaku, ki nga tangata o Takings-wairua, ki nga tangata o Meremere, ki nga tangata o Rangiriri, ki nga tangata o Paetai, ki nga tangata o Kupakupa, ki nga tangata o Te Whakapaku, ki nga tangata o Karakiriki, ki nga tangata o Whatawhata, ki nga tangata o Tiongahemo ki nga tangata o Kihikihiki, ki nga tangata o Rangiaohia ki nga tangata o Te Kopua, ki nga tangata o Hangatiki hui katoa.
E tai ma, tena ra ko koutou i runga i te atawhai o te Atua nana nei tatou i tiki, i tohutohu taenaoatia tenei takiwa o to tatou haerenga manenetanga, i runga i te mata o te whenua.
Na, kia rongu mai koutou ki taku, kahore aku utu ki a koutou mo ta koutou atawhai nui ki au. Heoi ano te utu ki a koutou ko te kupu o te Karaiti, e ki nei, "A, ko ia, e whakainu ana i tetahi o enei hunga nonohi ki to oko wai matao anake i runga i te ingoa o te akonga, he pono, e mea atu nei ahau ki a koutou, Kore rawa e kahore i a ia tona utu."—Matiu x 42.

10_1_1_2



TE WHETU O TE TAU.
Nimaru, Waikato
E hoa e Hare Reweti, tena koe Ka nui to matou
aroha ki a koe E hoa, kua tae mai to nupepa ki a
matou, kua kite matou. E pai ana nga korero o roto
E hoa he raruraru to matou, kua rongop pea koe, mo
Ngaruawahia otiia, kua oti He raruraru ano kei
Rangiaohia kei i a Hoani Papata raua ko Toma Ka
nui te raruraru ki Waikato ki Rangiriri, ki a te
Ngaungau, ki te Matetakahia E hoa, me tauri
mai hoki koe me mahi ki enei hara, ta te mea hoki
kua mo ko koe hei upoko mo enei tikanga pai. E
hoa, kia kaha koe ki te patu i enei raruraru, o to iwi,
ta te mea, ko koe hei kai whakakahi mo matou mahi
Heoi ano .Na o hoa aroha,
NA TAKEPEI Te Rau
NA PANAPA NOAUMU
NA TE REWETI
Ki a Hare Reweti, kei Akarana
HE PUPURU WHENUA
Whangape

E hoa e Hare Reweti tena koe i runge i te atawhai
o te Ariki, koia nei hoki te kai whakakahi a o nga wha
kawa Kua tae mai nei tau kupu ki ahau me tau aroha
ka taea mai noi ahau e koe te rapu mai i te whenua
tawhiti rawa E hoa, ko korero tenei naku u ki muri
whenua, Eppurua ana ahau i taua pihl ki ahau, ua
te mea, i matau ahau, no aku tupuna ia whenua
Kore au e pai ma te tangata e tuku i tena whenua
Tenei laku tikanga maku ano te ae te kahore ra nei
Te mea i penei ai au, pota noa aka whenua te hoko

10_1_1.item

<Title>Te Whetu o Te Tau

<Date>18580601

1:images/10_1_1_1_1.gif:text/10_1_1_1.txt:

2:images/10_1_1_2.gif:text/10_1_1_2.txt:

3:images/10_1_1_3.gif:text/10_1_1_3.txt:

4:images/10_1_1_4.gif:text/10_1_1_4.txt:

Metadata

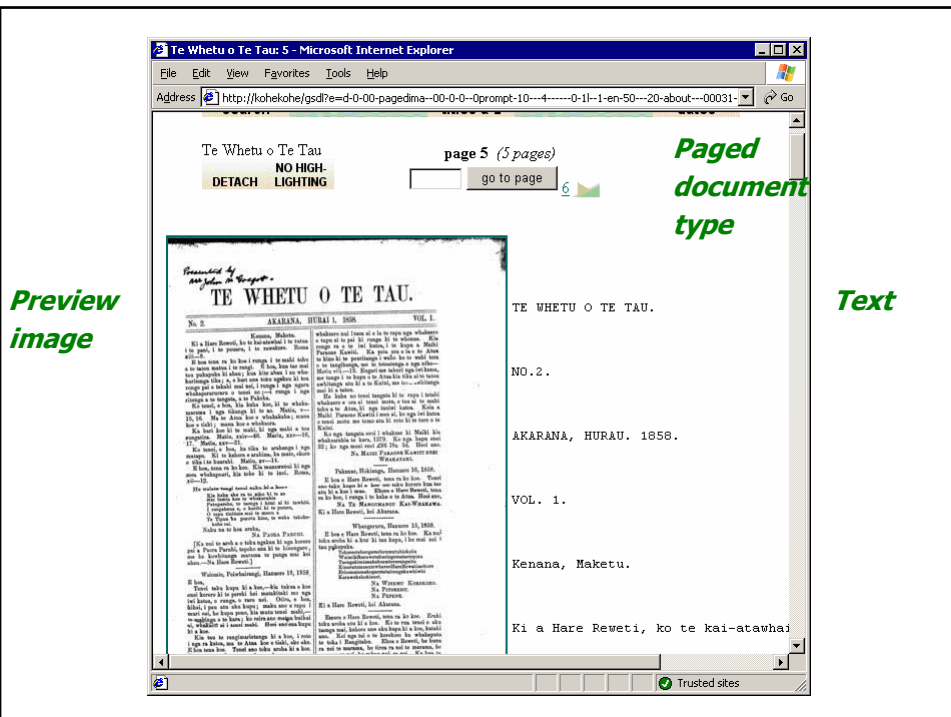
Page number

Image file

Text file

PagedImgPlug

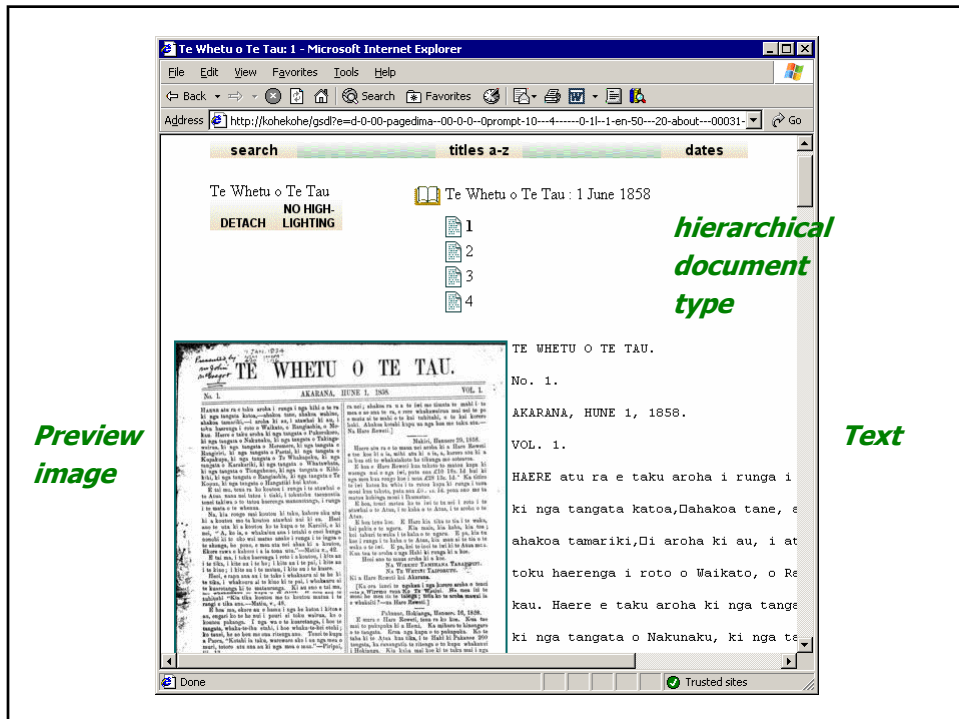
- ❖ Processes item files and their corresponding image and text files
- ❖ Options:
 - screenview (screenviewsize, screenviewtype) – produce a preview image
 - thumbnail (thumbnailsize, thumbnailtype) – produce a thumbnail image
 - documenttype – paged or hierarchical



Preview
image

Paged
document
type

Text



Extended item format

<PagedDocument>

<Metadata name="Title">The Title of the entire document</Metadata>

<Metadata name="Subject">A Document level Subject</Metadata>

<Page pagenum="1" imgfile="image1.jpg" txtfile="page1.jpg">

<Metadata name="Title">The Title of this page</Metadata>

... more metadata

</Page>

... more pages

</PagedDocument>



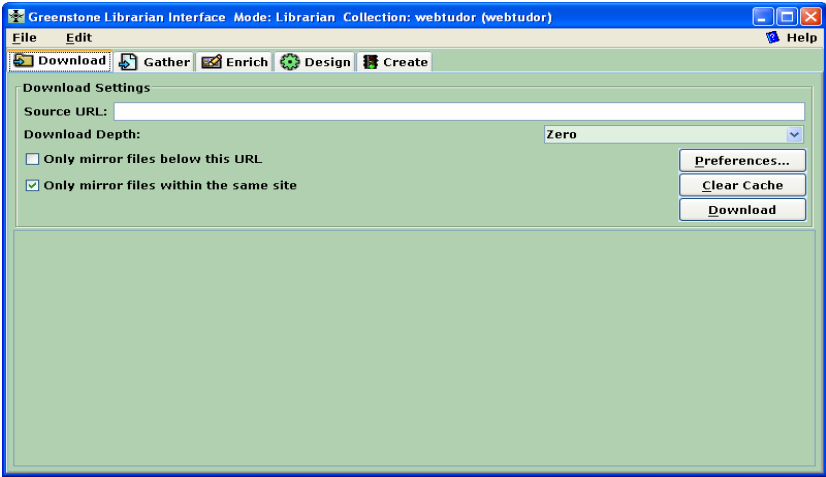
Agenda

- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ❖ PowerPoint documents
- ❖ CDS/ISIS

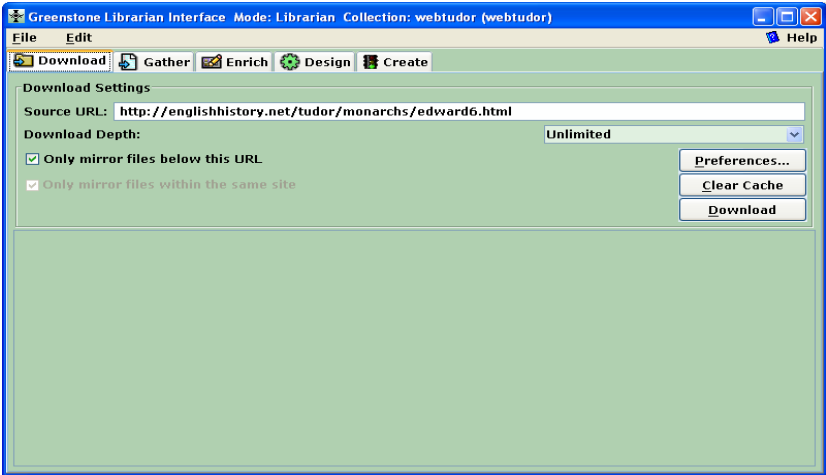
Downloading in GLI

- ❖ Can download, or “mirror”, web pages and web sites to local disk
- ❖ Options: within URL, within site, depth of links to follow
- ❖ Can be added into collection

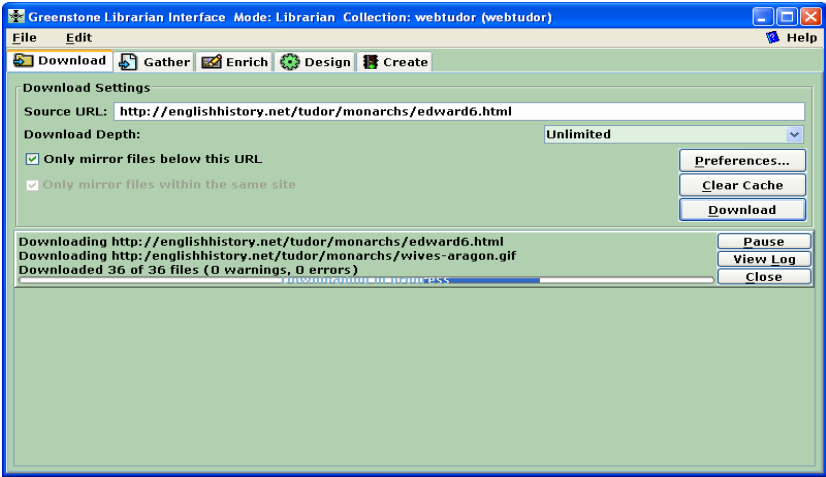
Download panel



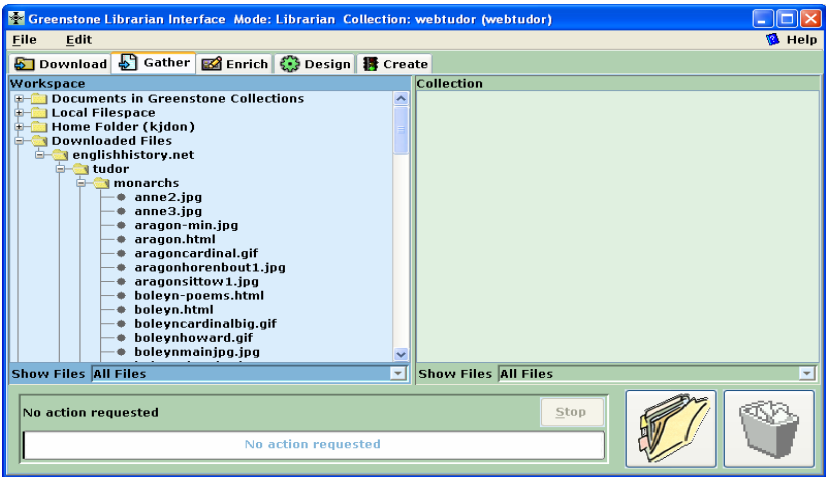
Setting up a download



Downloading in progress

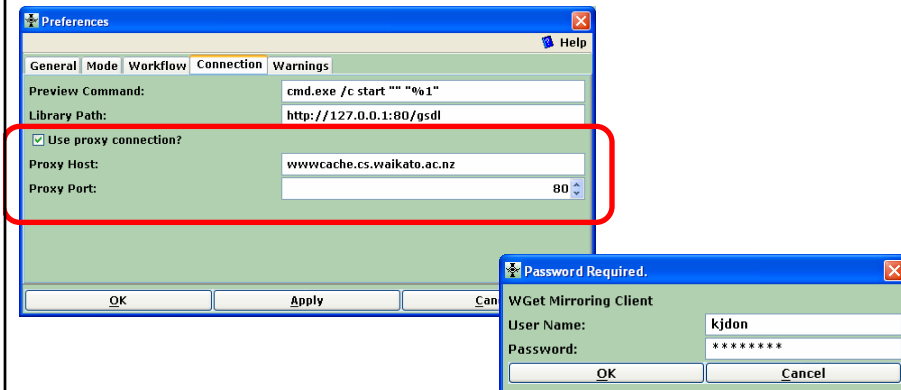


Downloaded files



Behind a firewall?

If you are behind a firewall or proxy server then you need to set this information in File->Preferences->Connection



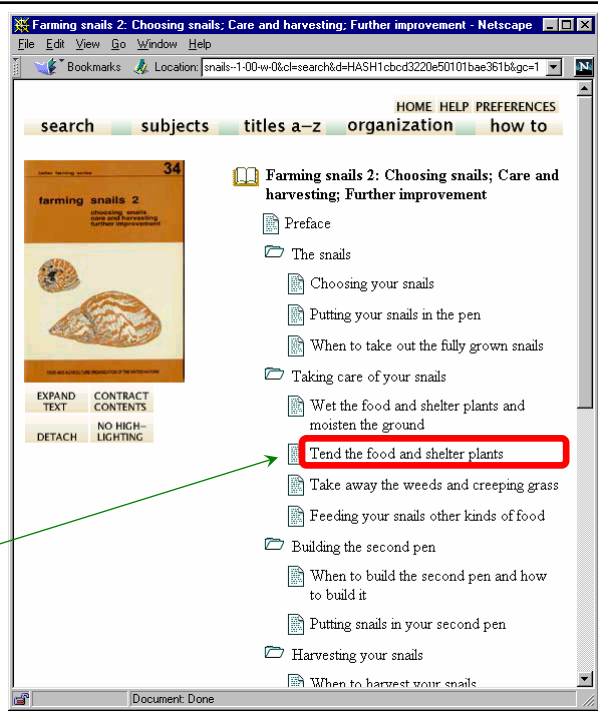
Downloaded files

- ❖ File hierarchy preserves site structure
- ❖ -file_is_url option to HTMLPlug adds URL metadata based on the file hierarchy
- ❖ [weblink][webicon][weblink] links to original if URL metadata has been set.
- ❖ So you can download web sites to index, then link back to the originals

Hierarchical document model

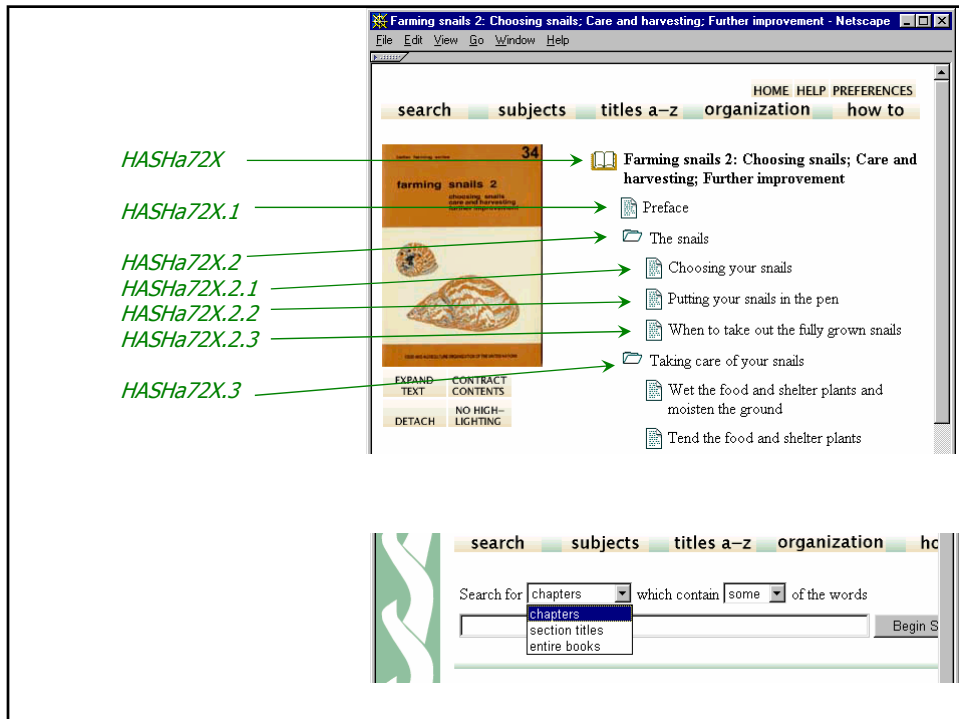
❖ Metadata specified at any level

Title metadata



Full Text Tagging

- ❖ While creating large digital collections:
 - the collection must be organized
 - the larger the collection the greater the need for organization
 - the larger the documents the greater the need for sections/subsections
- ❖ Greenstone lets you tag the full text of documents
- ❖ Then you can read them hierarchically ...
- ❖ ... and search them by section



Full Text Tagging...

To show the hierarchical structure, tag the source files like this:

```

<!--
<Section>
<Description>
<Metadata name="Title">Realizing human rights
  for poor people: Strategies for achieving the
  international development targets</Metadata>
</Description>
-->
  (text of section goes here)
<!--
</Section>
-->
  
```

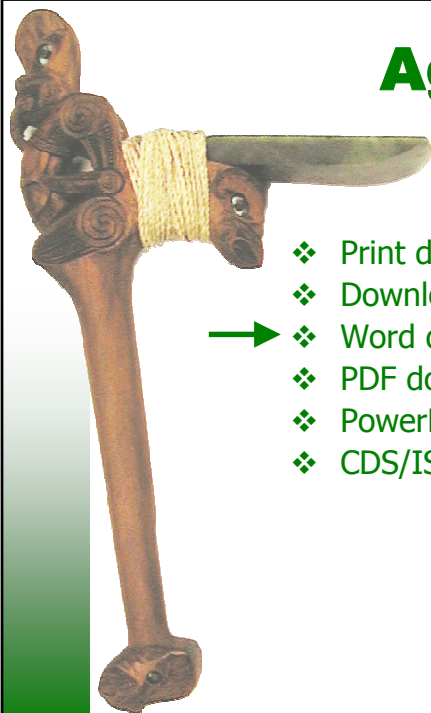
Full Text Tagging...

- ❖ Section tags define a hierarchical structure
- ❖ Sections can be nested within other sections
- ❖ All sections must be nested within a single enclosing section that encompasses the entire document
- ❖ In the collection configuration file, put

```
HTMLPlug -description_tags
```

- ❖ Mainly for HTML, but can be used in Word and PDF documents.

Agenda



- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ❖ PowerPoint documents
- ❖ CDS/ISIS

Word Document

❖ Word conversions in Greenstone

– Text

- ❖ Unix strings command

- ❖ use_strings option

– Flat format HTML => wvWare

– Styled format HTML => VB script

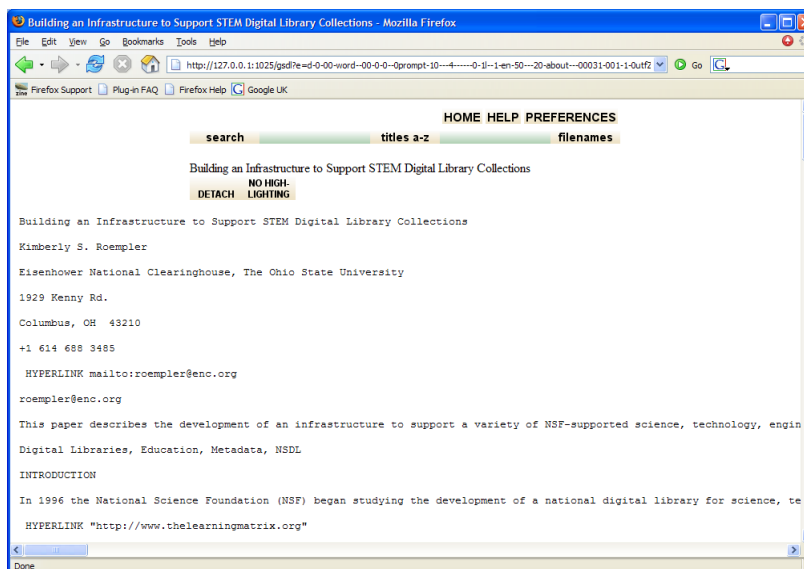
- ❖ windows_scripting option

- ❖ Heading setting

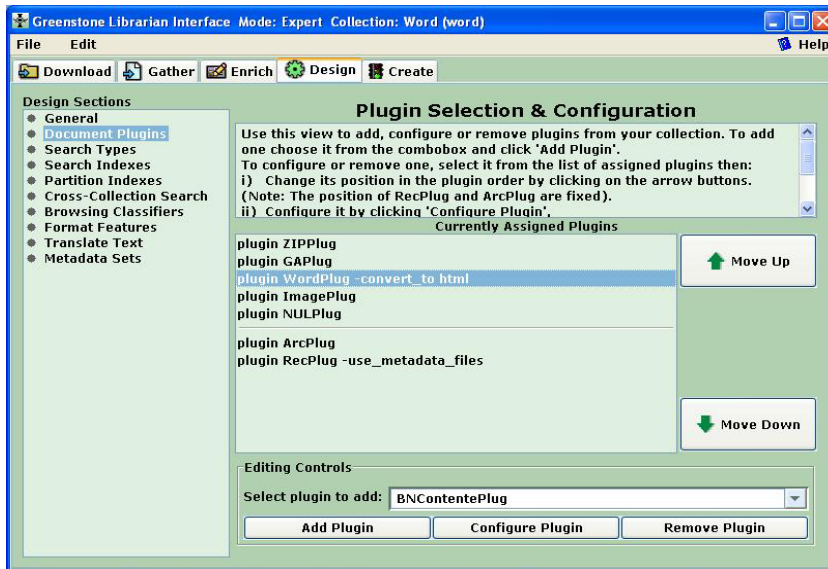
- <Heading 1>, <Heading 2>, <Heading 3>.....

- User-defined heading style

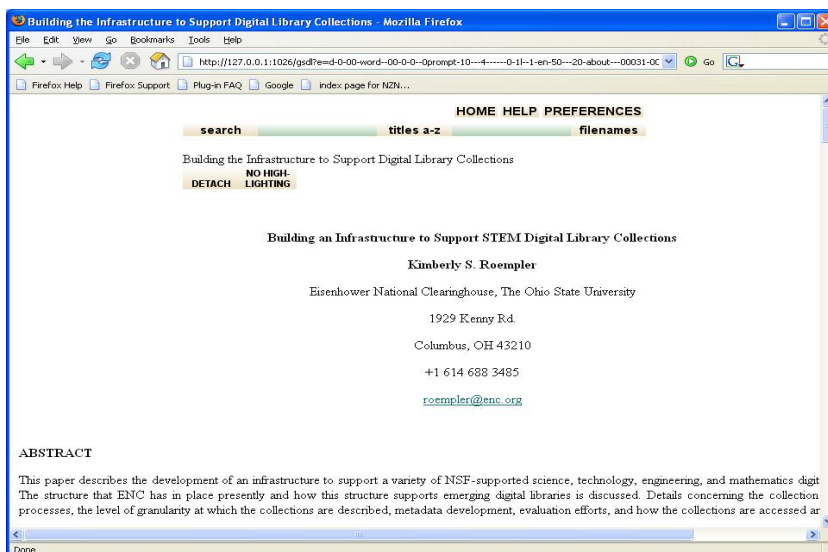
Word - Text



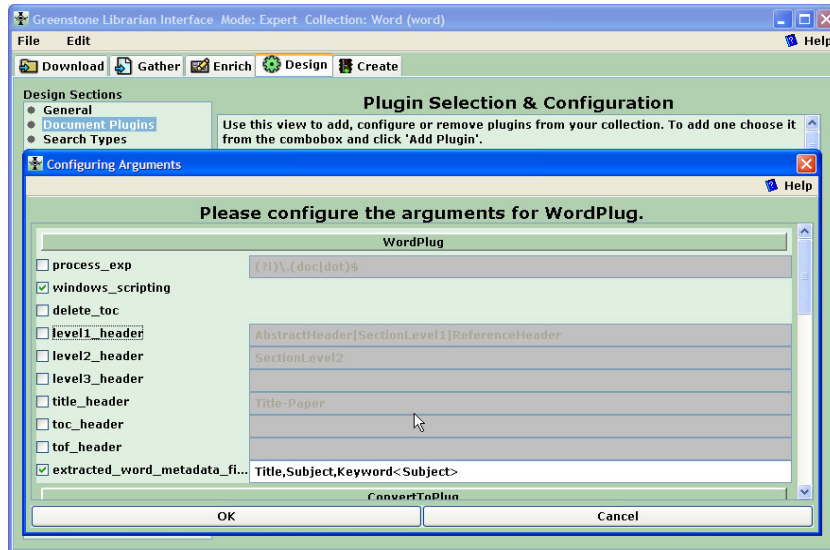
Word - HTML (wvWare)



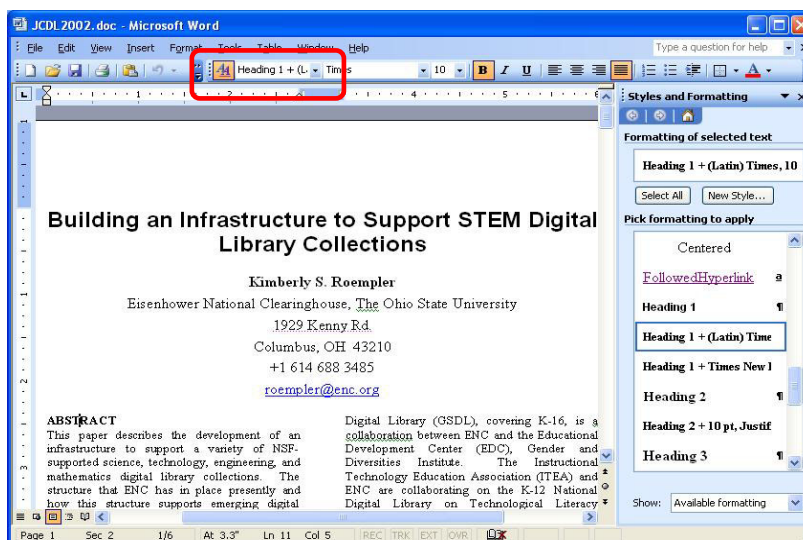
Word: Flat HTML format



Word - HTML (Windows Scripting)

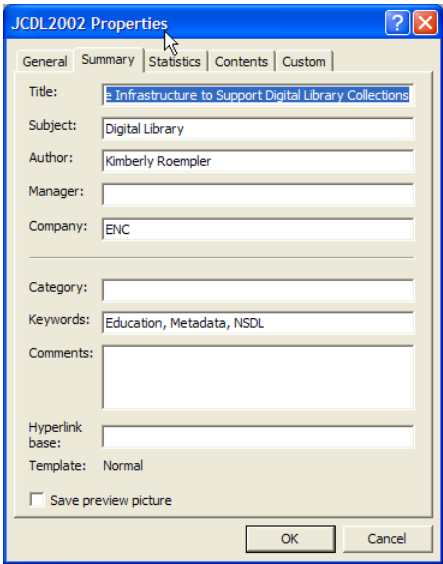


Word Document

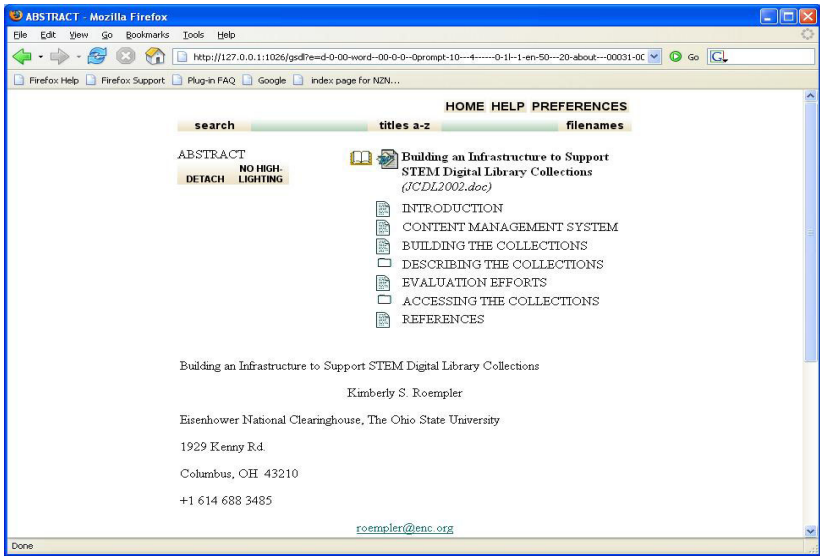


Word Document Properties

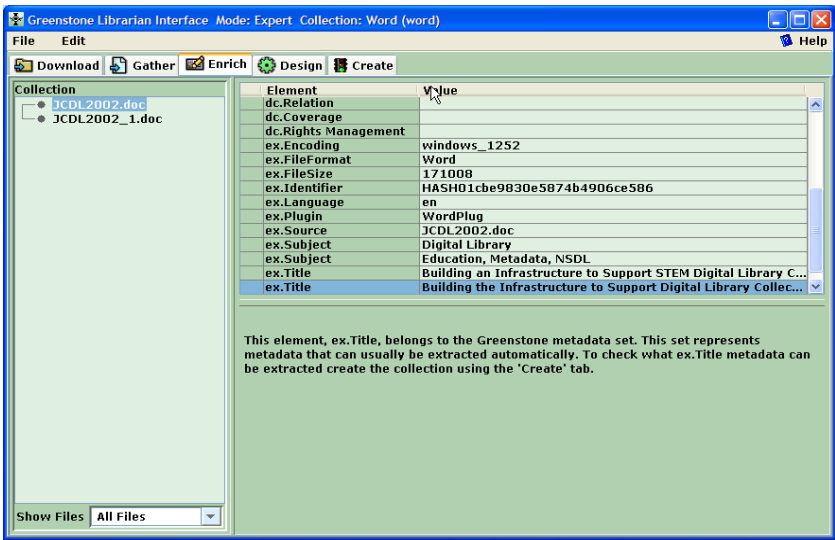
•File-> Properties



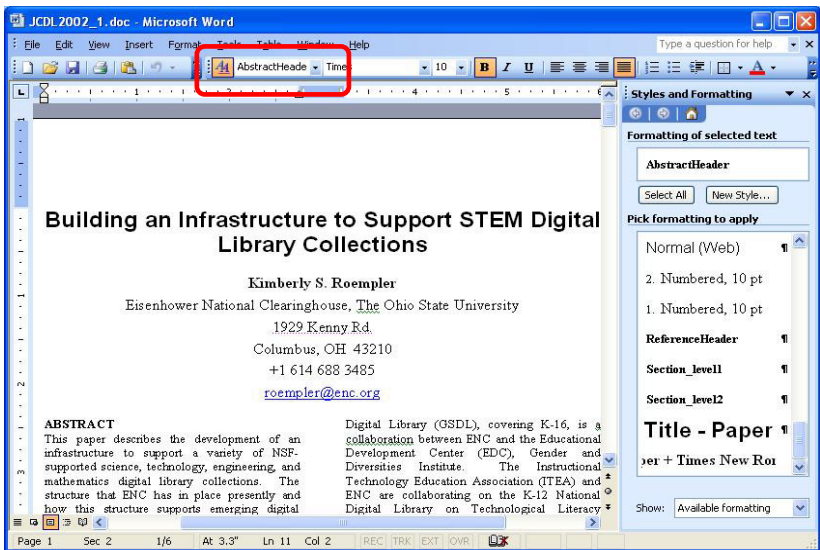
Word: Hierarchical HTML format



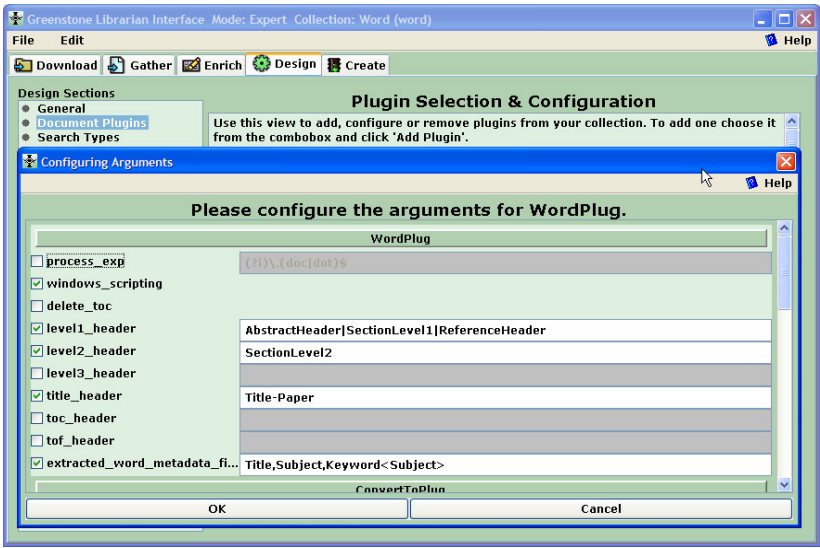
Extracted Word Document Properties



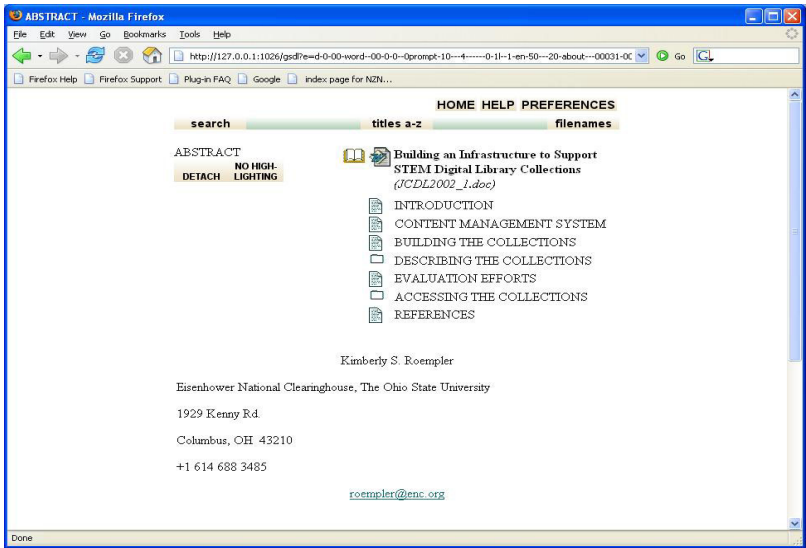
User-defined Style Formatting



WordPlug – User-defined Style



Word: Hierarchical HTML Format





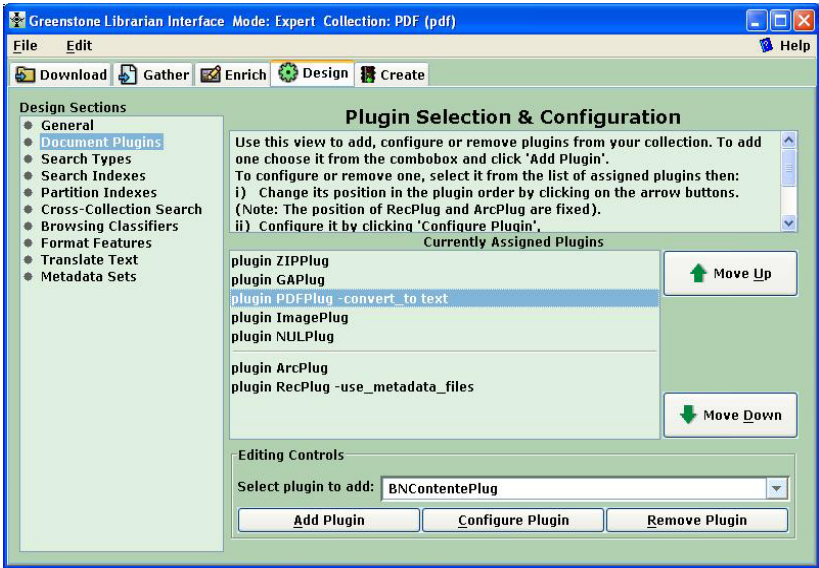
Agenda

- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ❖ PowerPoint documents
- ❖ CDS/ISIS

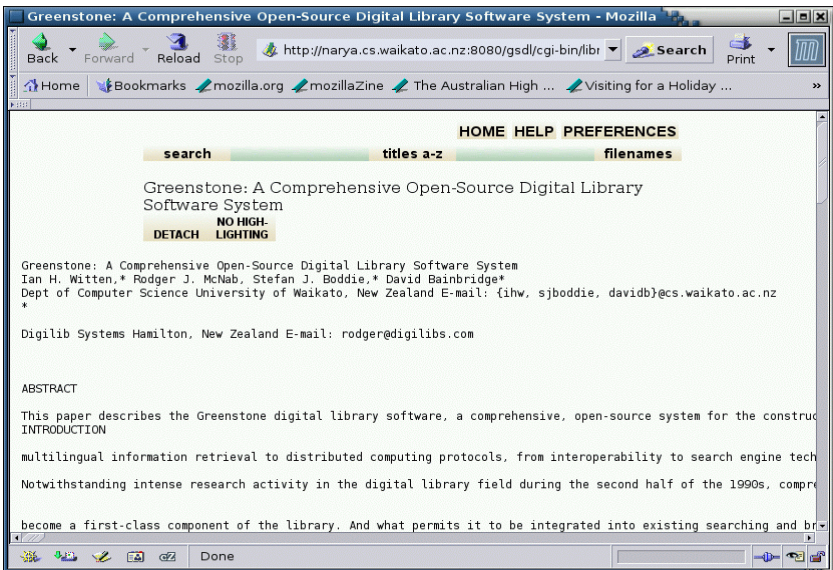
PDF Document

- ❖ PDF conversions in Greenstone
 - Text only for Unix system
 - HTML
 - ❖ use_sections option
 - ❖ complex option
 - Image
 - ❖ ImageMagick needs to be installed
 - ❖ Use of convert utility
 - ❖ Convert_to
 - pagedimg_jpg
 - pagedimg_gif
 - pagedimg_png

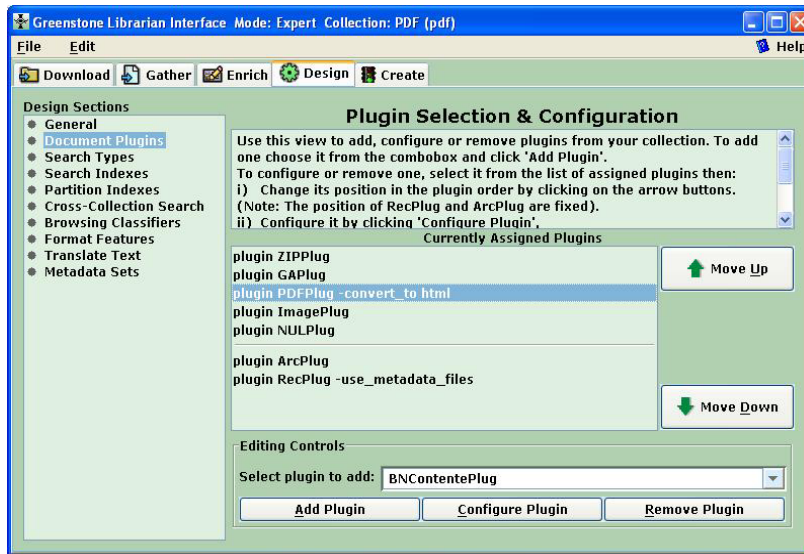
PDF - Text



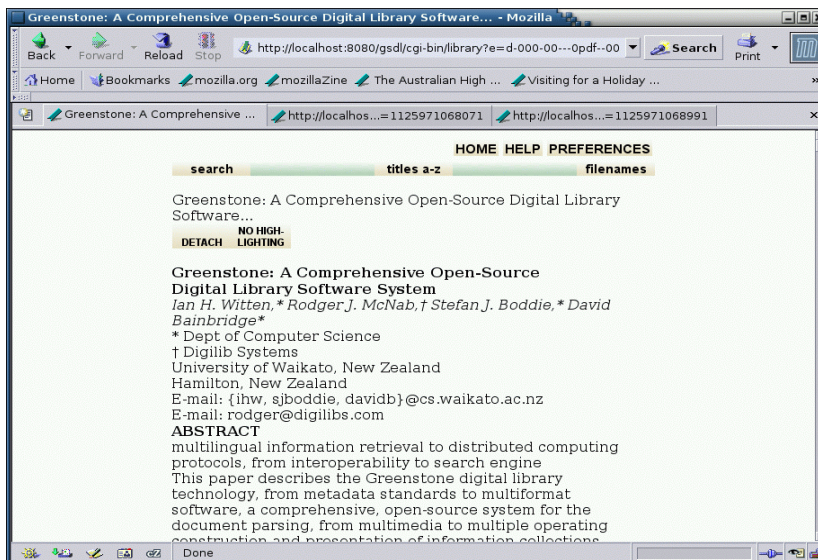
PDF: Text Document Display



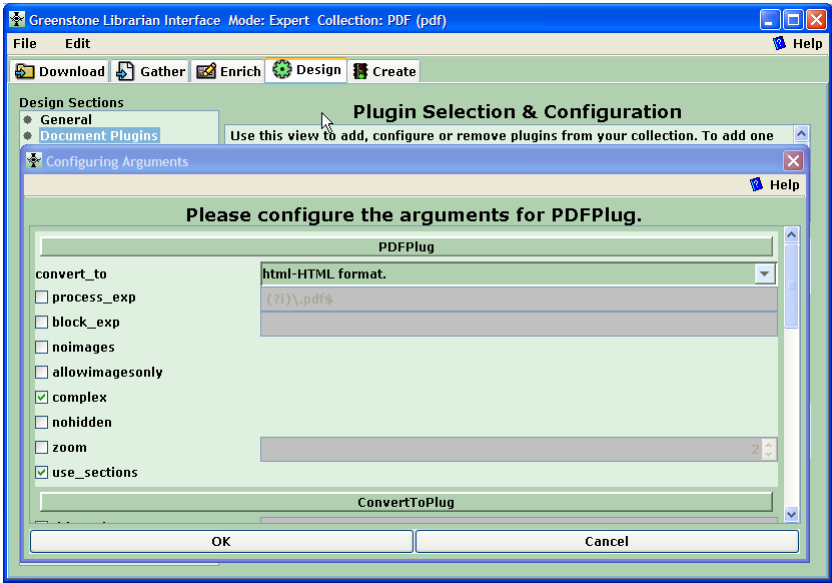
PDF - HTML



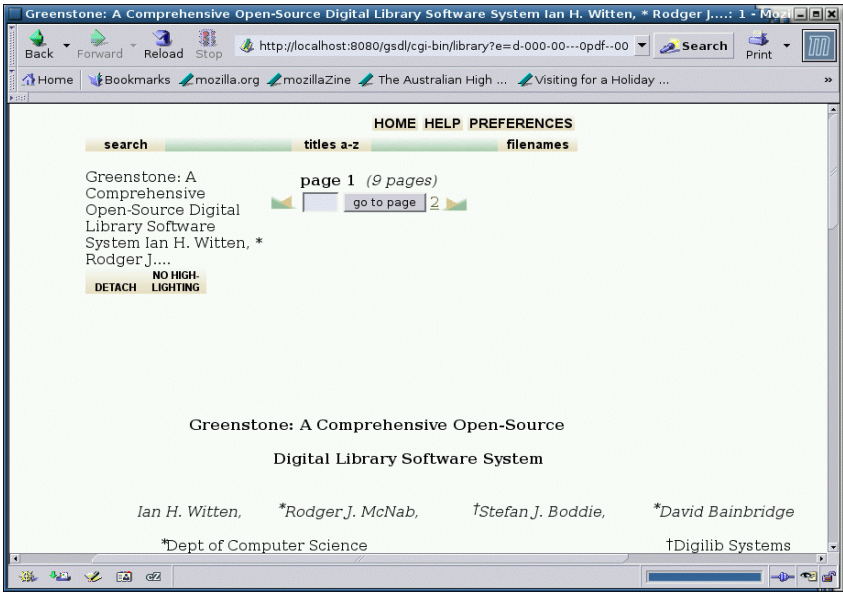
PDF: HTML Document Display 1



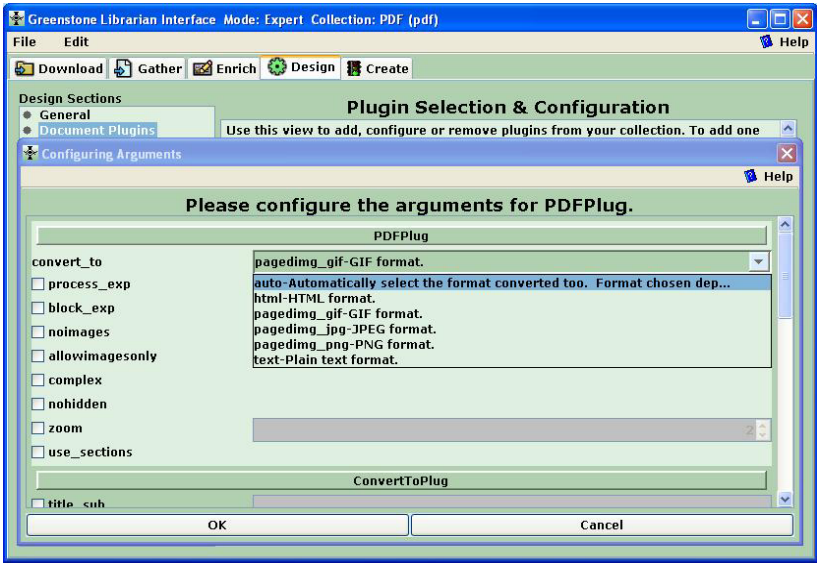
PDF – use_sections



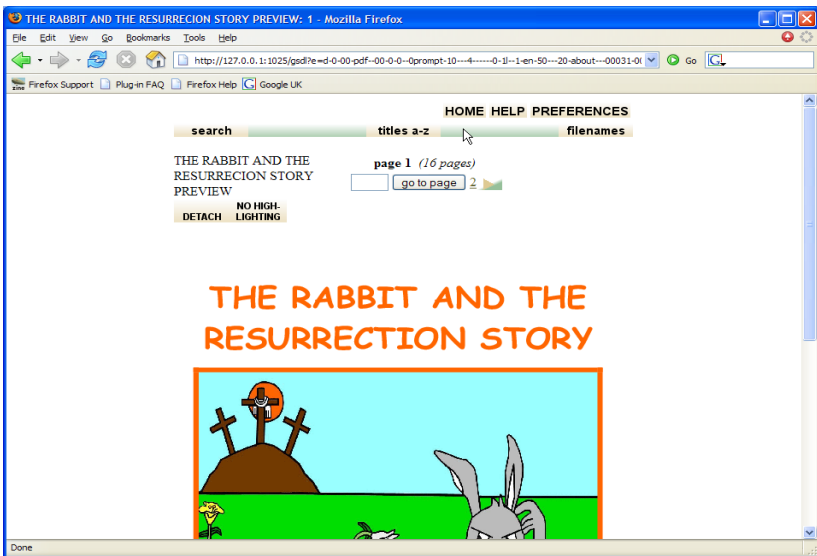
PDF: HTML Document Display 2



PDF - Image



PDF - Image Document Display





Agenda

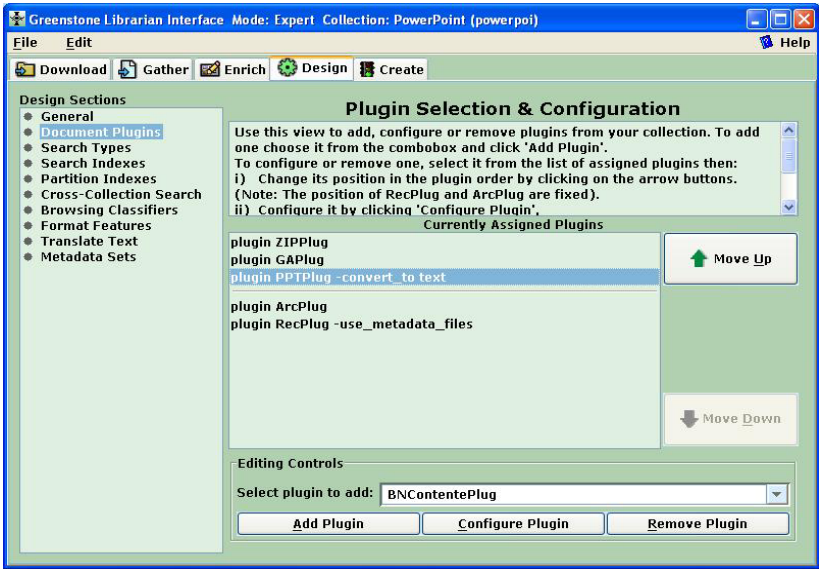
- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ➔ ❖ PowerPoint documents
- ❖ CDS/ISIS

PowerPoint Document

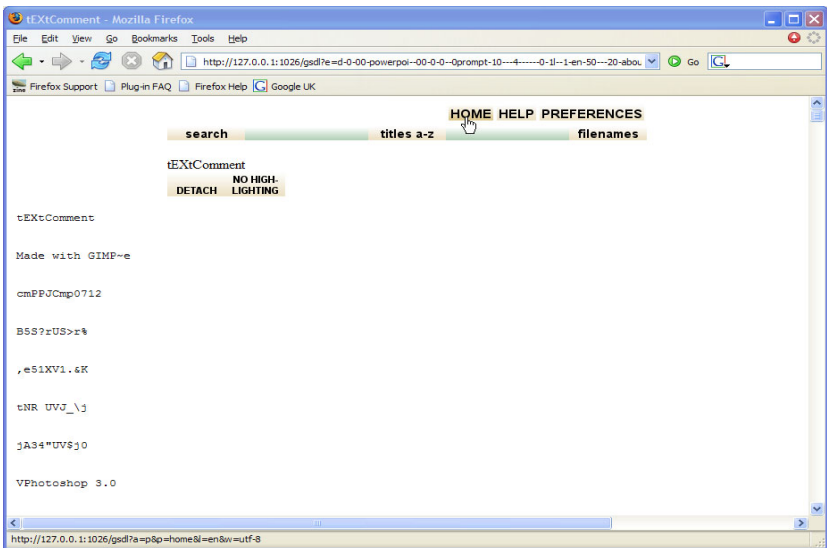
❖ PPT conversions in Greenstone

- Text
 - ❖ use_strings option
- HTML
- Image (JPEG, GIF, PNG)
 - ❖ windows_scripting option
 - ❖ convert_to
 - pagedimg_jpg
 - pagedimg_gif
 - pagedimg_png

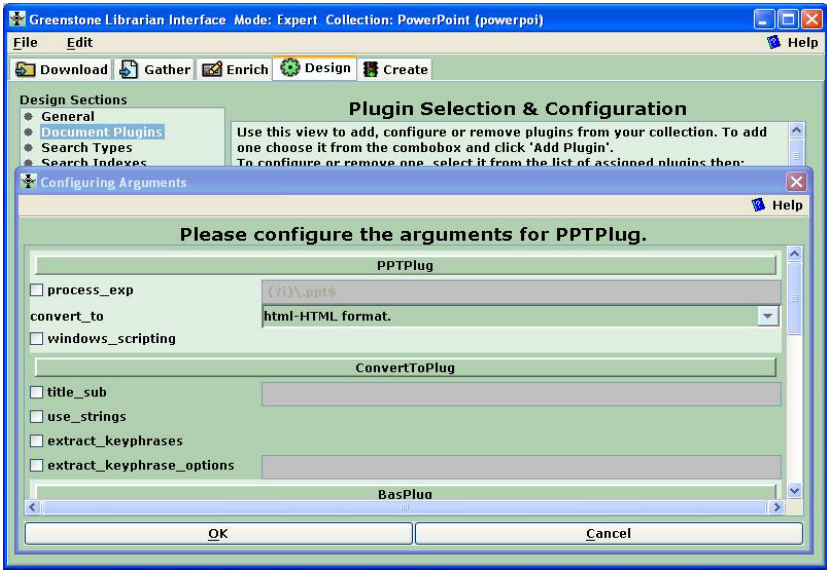
PPT - Text



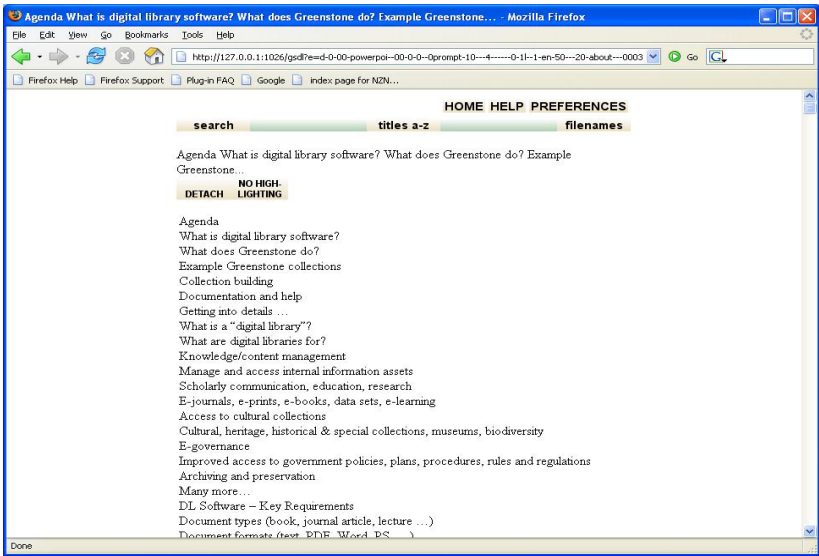
PPT: Text Document Display



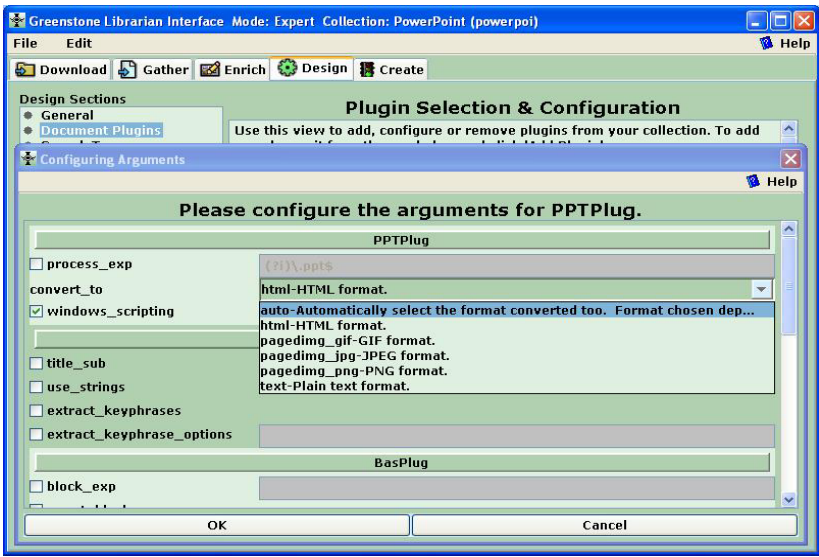
PPT - HTML



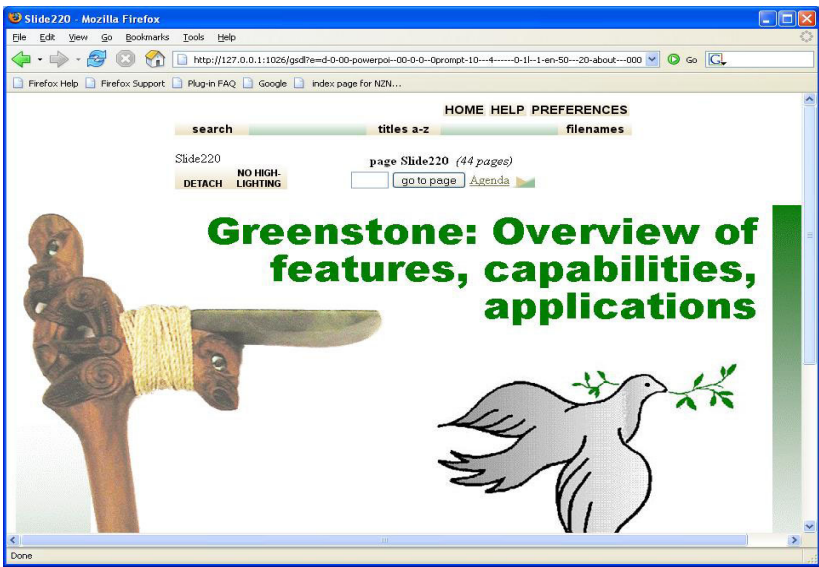
PPT: HTML Document Display



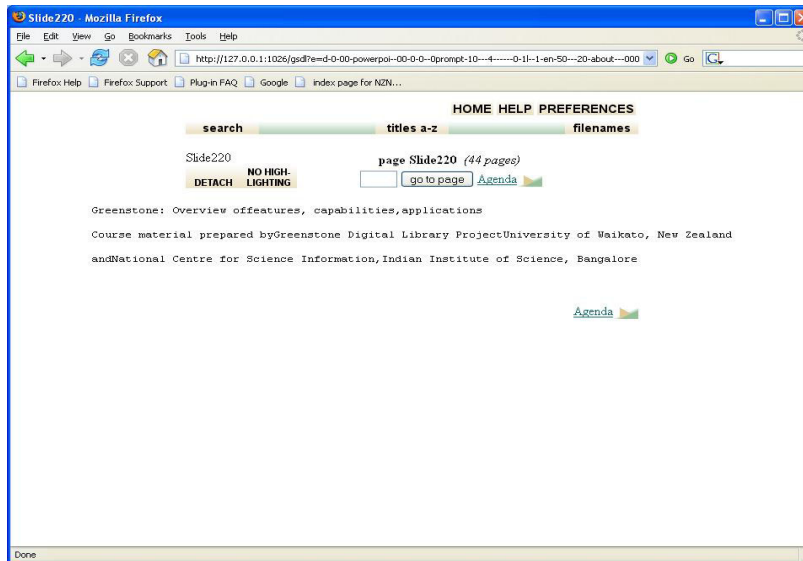
PPT - Image



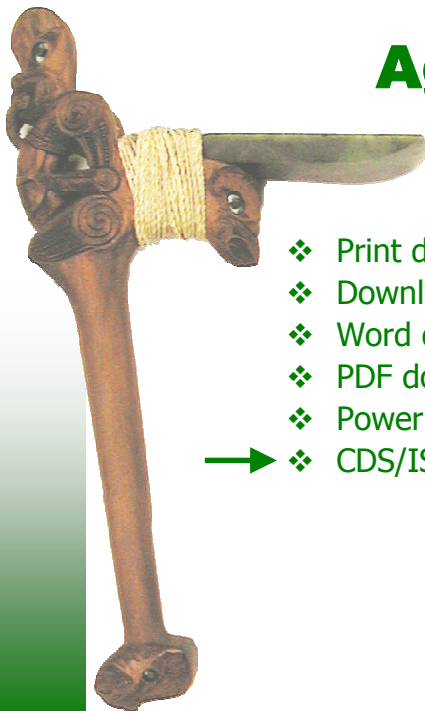
PPT Image: Image View



PPT Image: Text View



Agenda



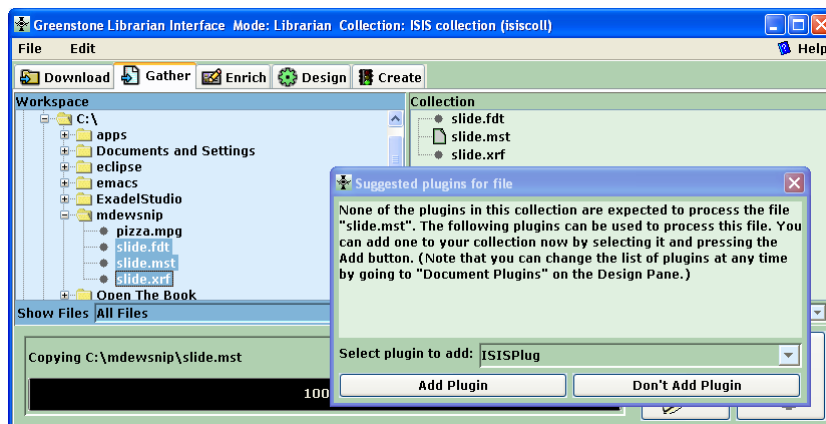
- ❖ Print documents
- ❖ Downloading HTML & full text tagging
- ❖ Word documents
- ❖ PDF documents
- ❖ PowerPoint documents
- ❖ CDS/ISIS

CDS/ISIS

- ❖ Bibliography collections are typically fairly complex:
 - Form searching
 - Customised query result and browse lists
 - Customised document display
- ❖ Let's work through creating a simple collection using a small CDS/ISIS database describing a set of film slides
 - (More information in the "Bibliography collection" and "CDS/ISIS" documented example collections)

CDS/ISIS

- ❖ Add the CDS/ISIS files to a new collection:
 - The GLI will suggest adding ISISPlug: yes please!



CDS/ISIS

- ❖ After building, let's view the collection:
 - No metadata searching is available:



- The titles classifier is completely empty!

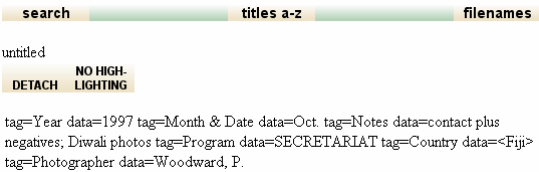


CDS/ISIS

- ❖ More problems:
 - The filenames classifier is useless!

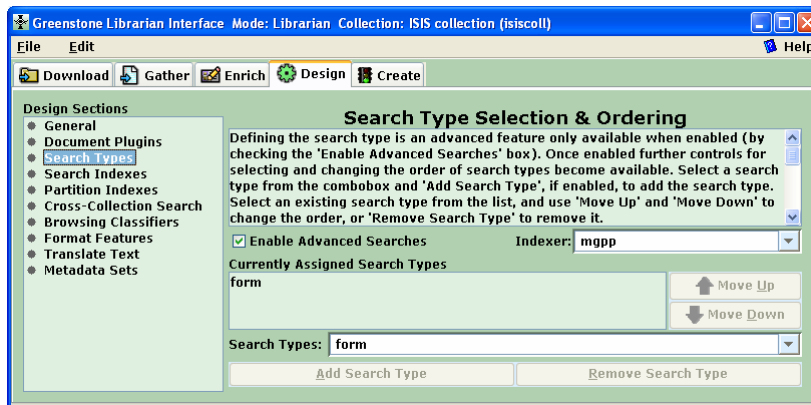


- The document display isn't very pretty:



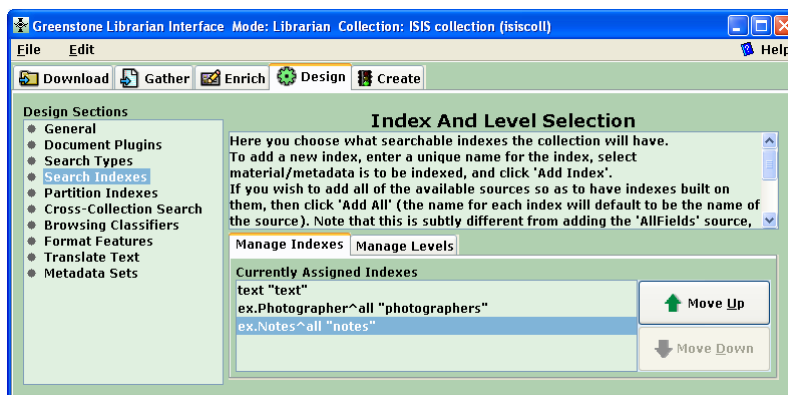
CDS/ISIS: Metadata searching

- ❖ To enable form searching, go to the "Search Types" area in the GLI's Design pane
 - Tick "Enable Advanced Searches" on
 - Add the "form" search type, and remove "plain"



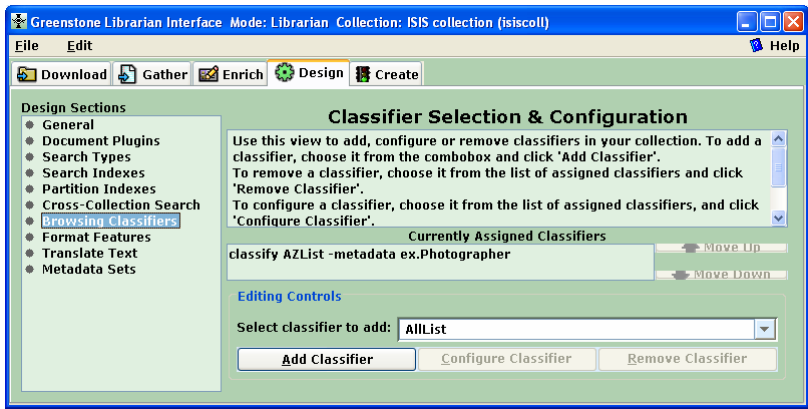
CDS/ISIS: Metadata Searching

- ❖ Add metadata indexes in the "Search Indexes" part of the GLI's Design pane
 - Add indexes for Photographer and Notes metadata
 - Remove the useless Source and Title indexes



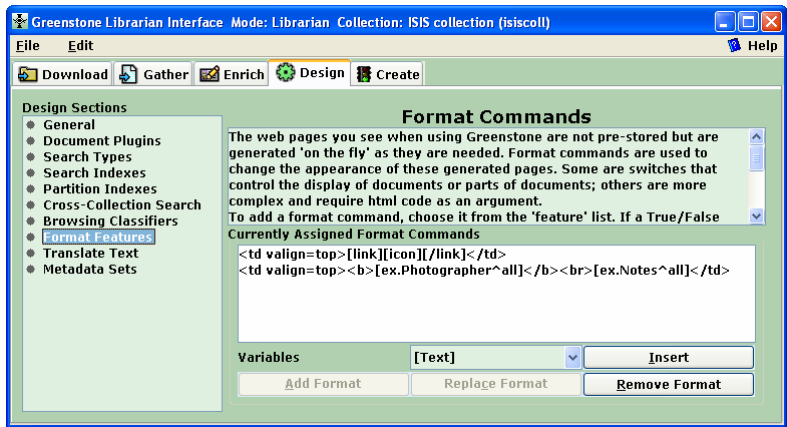
CDS/ISIS: Better browsing

- ❖ Remove the existing (useless) classifiers for Title and Source metadata, and add a new one for Photographer



CDS/ISIS: Better browsing

- ❖ Change the VList format statement to display the Photographer and Notes metadata:



CDS/ISIS: Document display

- ❖ Next, let's change the DocumentText format statement to show the Photographer and Notes metadata:

```
<center><table width=_pagewidth_><tr><td>Photographer:
</td><td>[ex.Photographer^all]</td></tr><tr><td>Notes:
</td><td>[ex.Notes^all]</td></tr></table></center>
```
- ❖ Then, let's remove those annoying "Detach" and "Highlight" buttons by setting DocumentButtons to empty
- ❖ Lastly, clear DocumentHeading to remove the "untitled" at the top of the document

CDS/ISIS: Finished!

- ❖ Metadata searching now available:

ISIS collection

search

Photographer

Search for

some

 of

Word or phrase

... in field

	text
	photographers
	notes
	text

Clear Form

Begin Search

- ❖ Better browsing facilities:

ISIS collection

Photographer

search

Photographer

A-E F-G-H-L M-R S T-X

Australia. Bureau of Mineral Resources

United States Geological Survey

Slides for presentations on the petroleum potential of the SOPAC region (New Ireland (PNG), Solomon Islands, Tonga and Vanuatu), obtained from N.F. Exxon

Barclay, W.

10 col. slides of hydrocarbon potential of the Southwest Pacific

CDS/ISIS: Finished!

❖ Document display improved:

search	Photographer
Photographer:	Australia. Bureau of Mineral Resources United States Geological Survey
Notes:	Slides for presentations on the petroleum potential of the SOPAC region (New Ireland (PNG), Solomon Islands, Tonga and Vanuatu), obtained from N.F. Exxon

❖ What could still be improved?

- More metadata indexes, classifiers
- Display all fields in the document display
- Nice images for classifiers
- ...?

**Questions?
Comments?
Discussion?
Feedback form!**