Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 6. EXAMPLE OF DIGITAL LIBRARY SOFTWARE: GREENSTONE

LESSON 3. USING GLI: DESIGNING, CREATING AND PREVIEWING A COLLECTION

NOTE

Please note that this is a reference guide covering the main steps for using GLI as described in the corresponding lesson. It provides an excellent means for additional practice and for printing after you have completed the course.

We recommend that you first take the lesson using the courseware environment, as it contains interactive features such as exercises with feedback, pop-ups, animations etc.

© FAO and UNESCO, 2005

SCOPE NOTES

These practical exercises will help you build an operational digital library collection from a set of sample source documents. The collection will be built using the Greenstone Librarian Interface (GLI).

Building a new collection using GLI takes five steps:

1) Gathering
    - Copying the documents from the computers' file space, including existing collection/s, into the new collection. Any existing metadata remains attached to these documents. Documents may also be gathered from the web through a built-in mirroring facility
2) Enriching
    - Enrich the documents by adding further metadata to individual documents or groups of documents
3) Designing
    - Design the collection by establishing its appearance and the access facilities it will support
4) Creating
    - Build the collection
5) Previewing
    - Preview the newly created collection, which will be included in your Greenstone home page as one of the collections.

By using this document you will apply steps 3) to 5): Designing, Creating and Previewing.

GLI STEP 3: DESIGNING THE COLLECTION

The interface design mainly refers to defining collection specifications in terms of searching, browsing, and results display requirements. For the sample collection that you will be building, you will be guided in providing the following access features:

- Simple Searching by
    - full text, author and title
- Advanced, form-based searching by
    - full text, author and title
- Browsing by
    - author and title

You have full control over the way the records are displayed after a search or while a collection is being browsed. The format you will be guided to use to display the search records is:

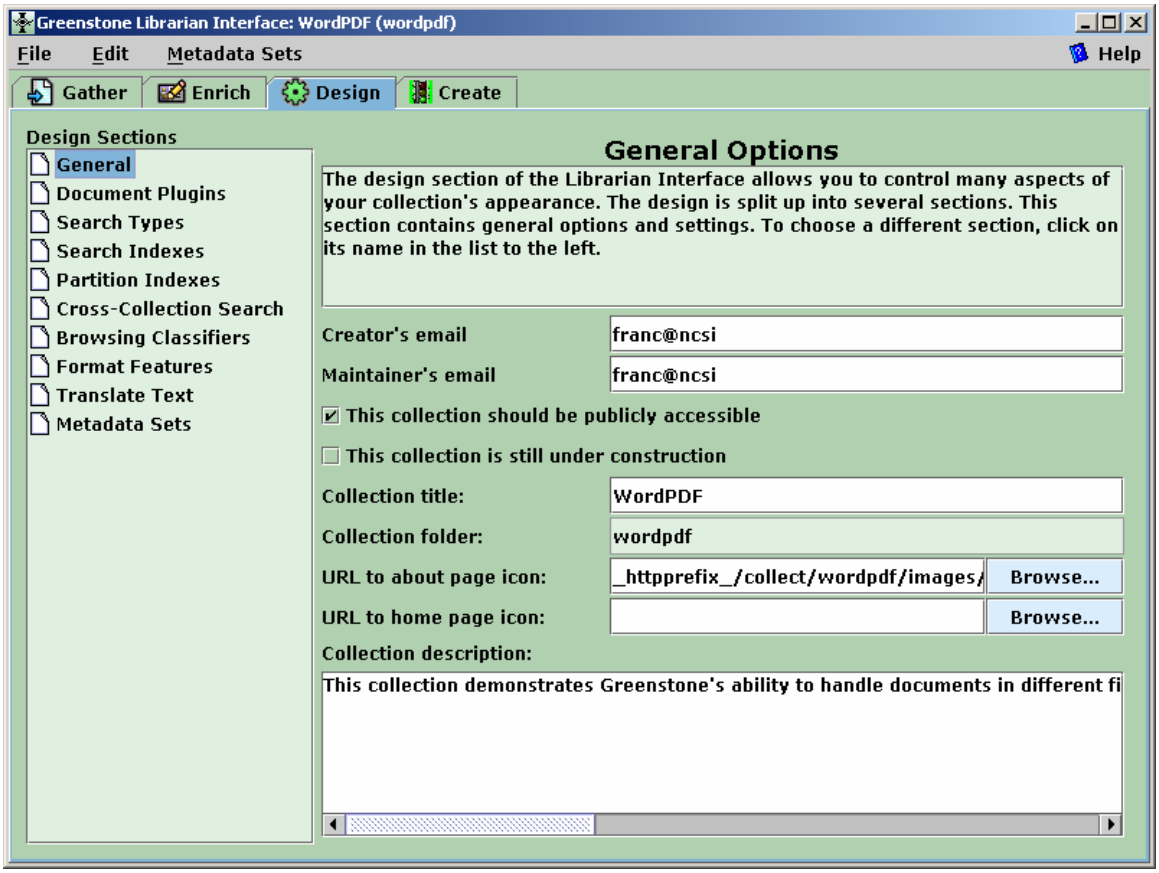
- Search results display: We shall use the default format which is: Link to HTML icon (for the HTML version of the document) followed by a link to the document type image icon (for the native format of the document), document title, and source file name.
- Title browse display: Default format described above.
- Author browse display: Grouping by individual authors using the bookshelf icon. Upon clicking the bookshelf icon, show document details as per the default format.

GLI STEP 3: DESIGNING THE COLLECTION

The General section

The Design step involves a series of separate interaction screens, each dealing with one aspect of the collection design. The design step is split up into several sections as shown in the figure below.

The first section contains general options and settings. In this section, values provided during the collection definition step (GLI step 1) can be modified if required.



Collection Design (general section)

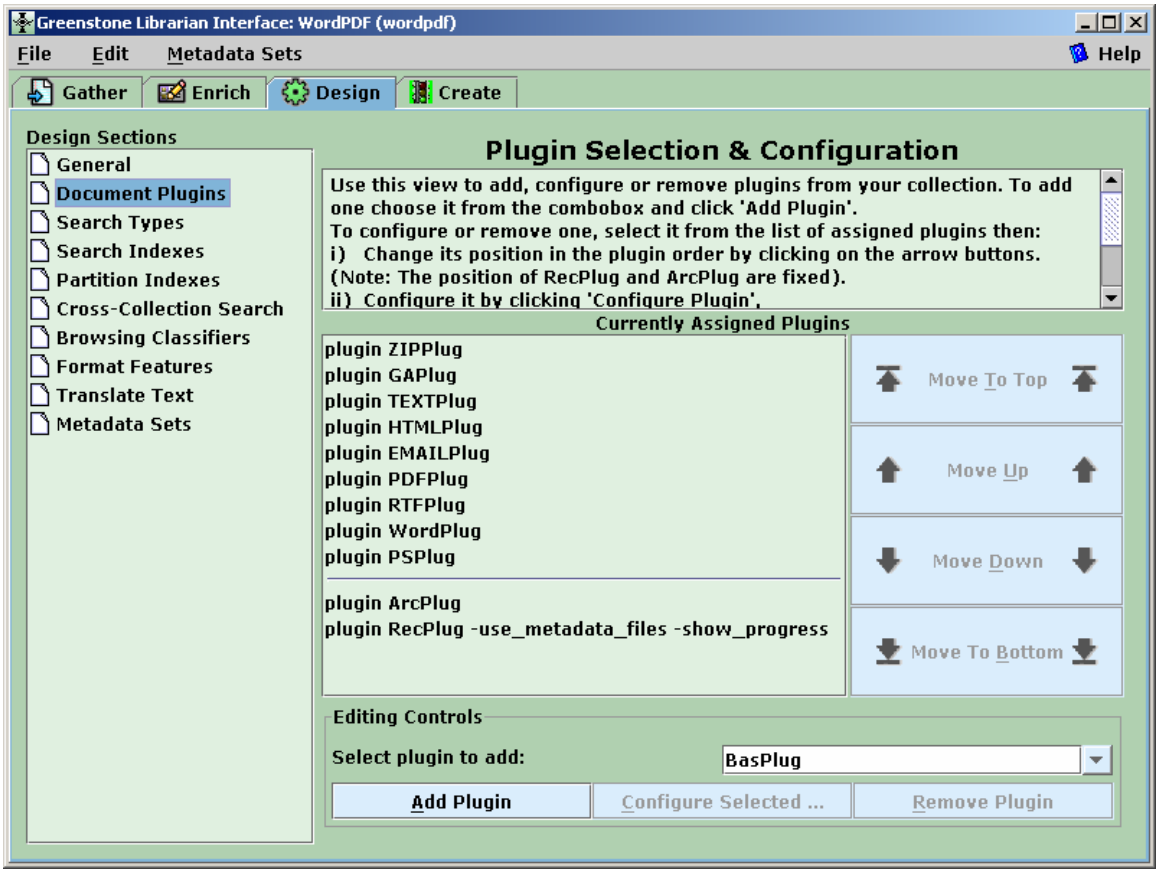
You can brand your collection using a suitable image. For example, you can use the wordpdf-e.gif (you will find it in the 'image' folder of the 'wordpdf-e' collection) and copy it to the 'image' folder within the 'wordpdf' collection folder created by Greenstone during collection definition (typically, c:\program files\gsdl\collect\wordpdf\images\ if you had installed Greenstone in the C: drive).

After this is done, go back to the GLI/General Options screen, use the 'Browse...' button associated with 'URL to about page icon:' and select this image file. GLI automatically generates the URL.

GLI STEP 3: DESIGNING THE COLLECTION

The Document Plugins section

The next section is 'Document Plugins'. Activate this section by clicking it.



Collection Design (document plugins)

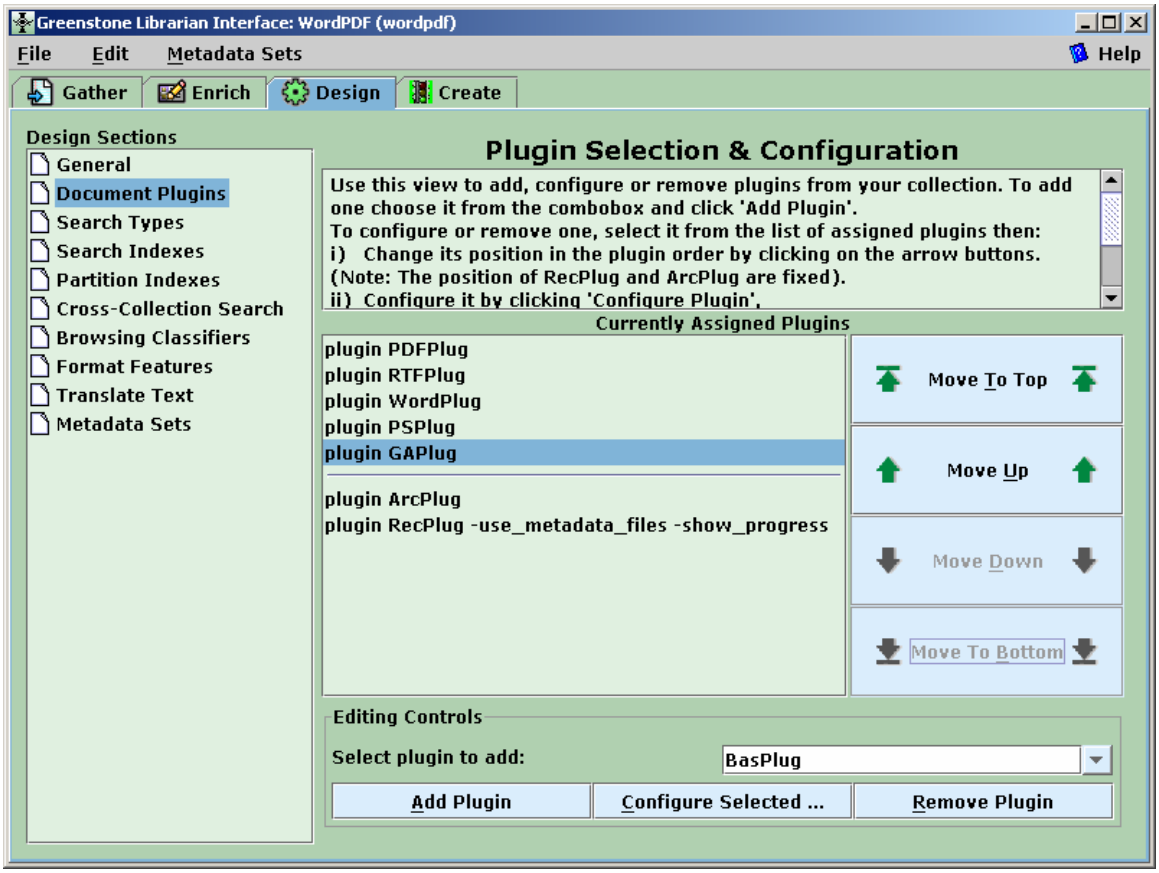
In this section you can add, configure or remove plugins to be used for your collection. Plugins extract text from source documents and convert these to the Greenstone Archive (GA) format. They also extract some of the metadata fields automatically.

For our source collection, which consists of PDF, Word, RTF, and Postscript documents, we do not require ZIPPlug, TEXTPlug, HTMLPlug, and EMAILPlug plugins. These can be deleted by selecting them one at a time and clicking on 'Remove Plugin'. By removing unwanted plugin/s, the collection building process will be faster.

The source documents will be processed by the plugins in the same order as listed in the 'Document Plugin' section.

GLI STEP 3: DESIGNING THE COLLECTION

The following figure shows the revised 'Document Plugin' section:



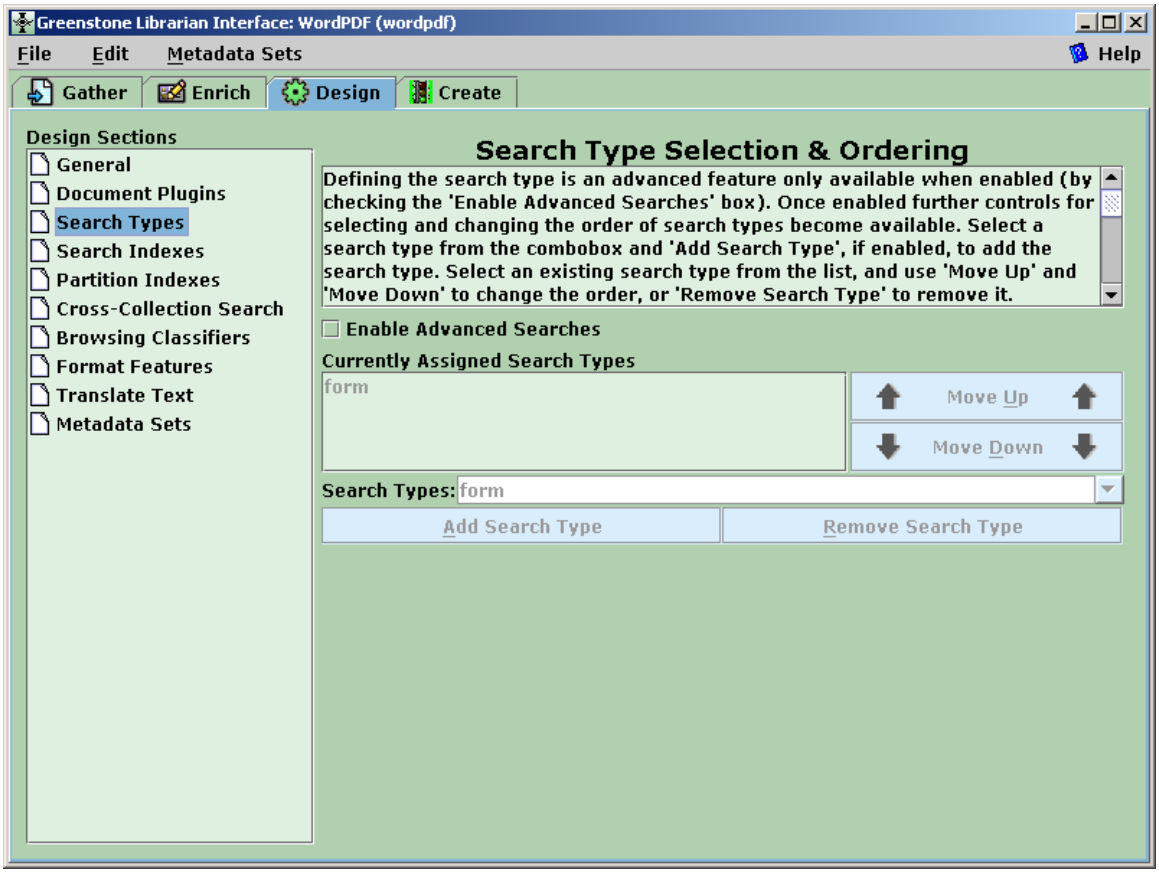
Revised Document Plugin Section

The Search Types section

The search types specify what kind of search interface should be provided for the collection:

- form for field level searching; and/or
- plain for regular searching.

The ordering is important, as the first one will be used for the initial search page by default.



Collection Design (search types)

- Check the 'Enable Advance Searches' box.

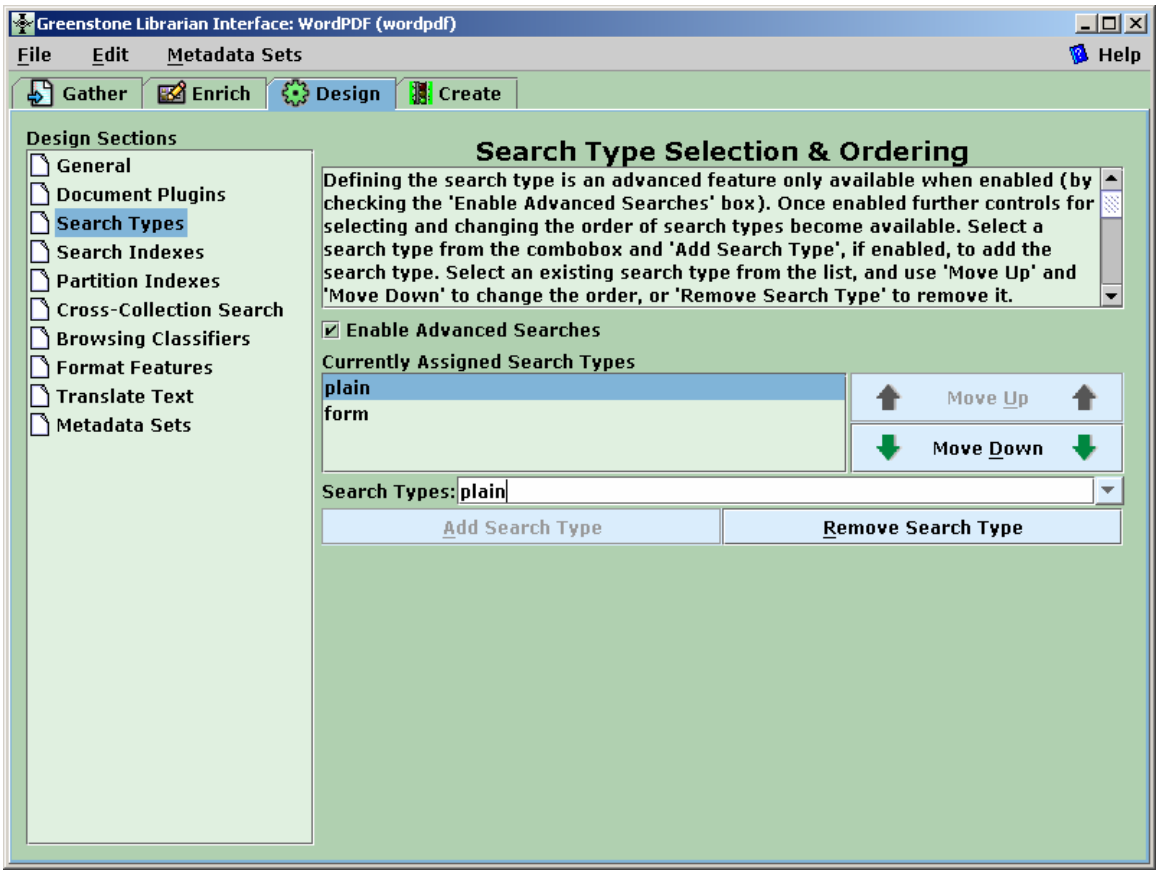
The other items in the screen get activated. For our collection let us use both the plain and form searches. To add a search type:

- select it from the 'Search types' list and click 'Add Search Type'.

For our collection let us have the plain search as the default search type. To achieve this:

- click on 'plain' and then click on 'Move Up'.

The modified ‘Search Types’ screen will look like in the figure:



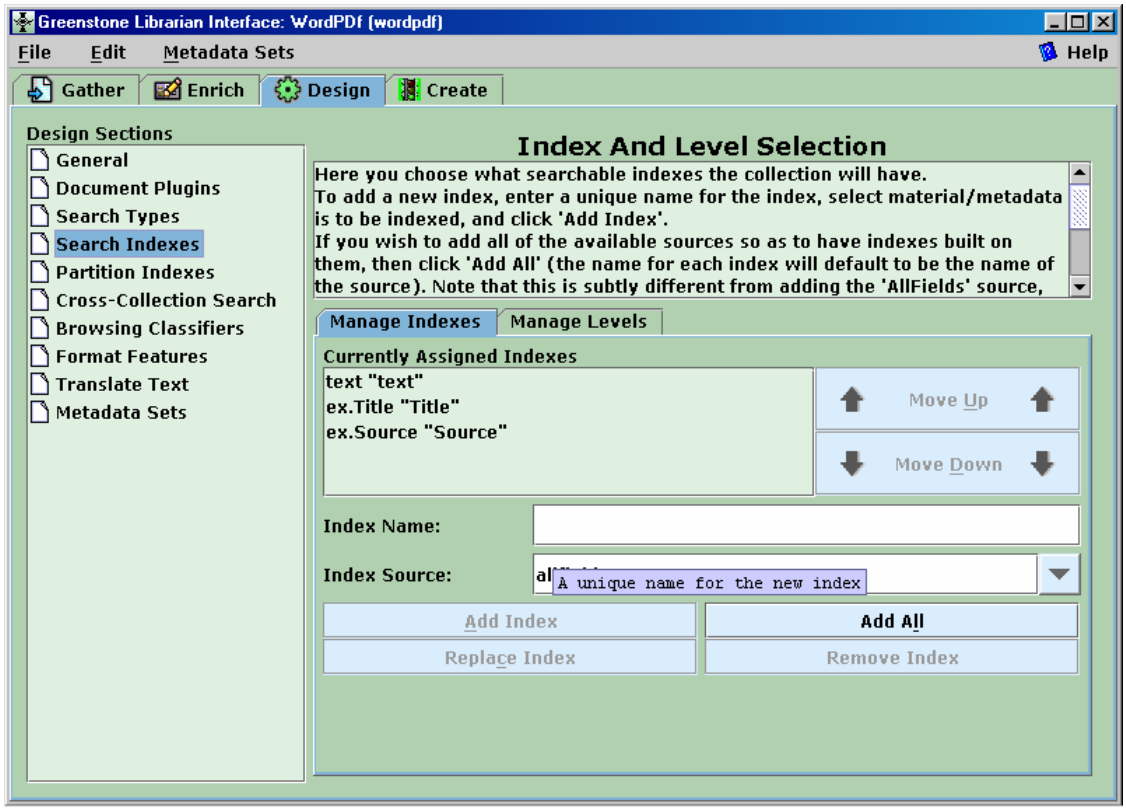
Collection Design (search types)

Have you been saving the collection building process periodically? If yes, good! If not, save it right now by clicking on File menu and then clicking on Save.

GLI STEP 3: DESIGNING THE COLLECTION

The Search Indexes section

The next step in the Design Section is 'Search Indexes'. Indexes specify what parts of the collection are searchable. The following figure shows a screen shot of this section.



Collection Design (search indexes)

By default Greenstone builds indexes and supports searching on automatically extracted metadata (title and source file names) and words extracted from document text. If you wish to support searching using the metadata set elements (in our case Dublin Core), you need to provide the required indexing and search specifications here. For the WordPDF collection that we are building, Greenstone has created three indexes by itself. These indexes are displayed in the 'Currently Assigned Indexes' box. A brief description of each of these three indexes follows.

The first index is text and is called 'text'. It is a full text index. This implies is that this index has been built using all the words from the documents of the WordPDF collection.

The second index is ex.Title and is named as 'Title". This index will be built by extracting the titles of the source documents. The titles are automatically extracted by the plugins, which handle the respective file types. The titles thus extracted are not reliable (as Greenstone may have difficulty in correctly extracting this information from some documents). The prefix, 'ex' in ex.Title indicates that the titles are automatically extracted by Greenstone from the source documents (remember the 12 documents that we have added?). \

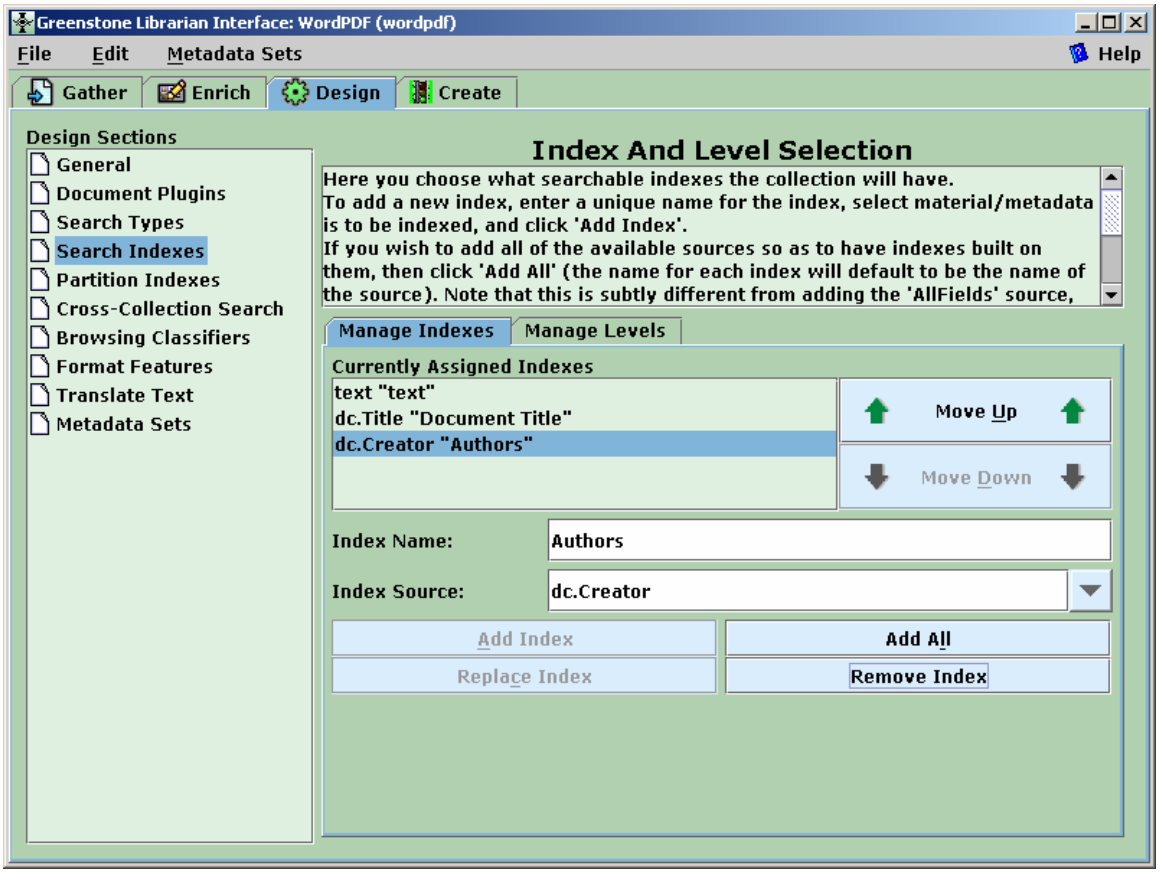
The third index is ex.Source named as 'Source'. This is an index of source file names.

We will delete the ex.title and ex.Source indexes. As mentioned above, ex.Title is not reliable and ex.Source (source file names) may not be of much use as an access point, for our collection. To delete these two indexes, select them one at a time and click 'Remove Index'. Once you have removed these two indexes only the 'text' index will remain.

Let us continue with our collection building process by adding two more indexes to our collection. The two indexes that we will be adding are dc.Title and dc.Creator.

Adding Title (dc.Title) index:
Provide an Index name. Let us name the dc.Title index 'Document Title'.
From the Index Source box, select dc.Title. Click on Add Index to add the dc.Title index to the Currently Assigned Indexes list.

Adding Creator (dc.Creator) index:
Provide an Index name. Let us name the dc.Creator index as 'Authors'.
From the Index Source box, select dc.Creator. Click on Add Index to add the dc.Creator index to the Currently Assigned Indexes list. The 'search indexes' screen will appear as shown in this figure:



Collection Design (search indexes)

GLI STEP 3: DESIGNING THE COLLECTION

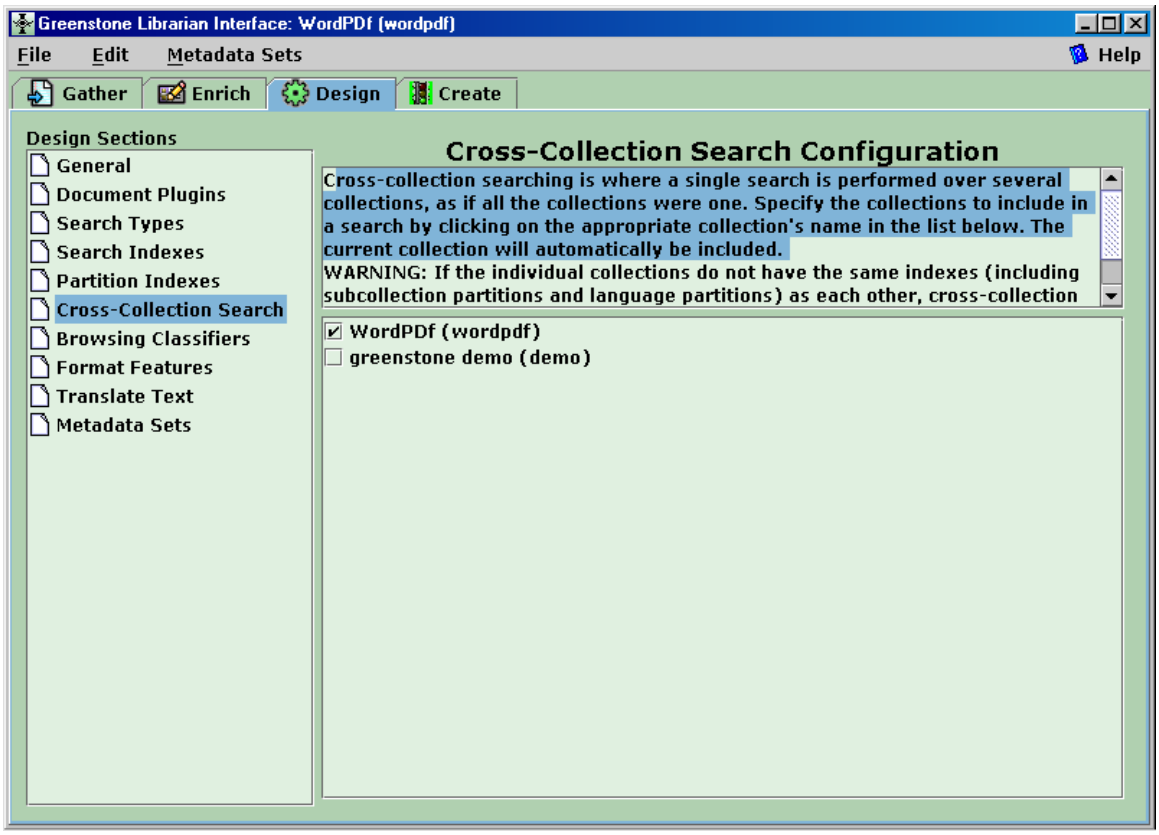
The Partition Indexes section

The next Design Section is 'Partition Indexes'. A collection index can be partitioned into sub collections based on metadata values and languages. This will facilitate controlling the search space. This partition can be achieved by using the 'Partition Indexes' feature. For our collection we will not create any Partition Indexes.

The Cross-Collection Search Configuration section

The next section is the 'Cross-Collection Search Configuration'.

Cross-collection searching is where a single search is performed over several collections, as if all the collections were one. Specify the collections to include in a search by clicking on the appropriate collection's name in the list below. The current collection will automatically be included. Let us not make any changes here. The default view of this section is shown in this figure:



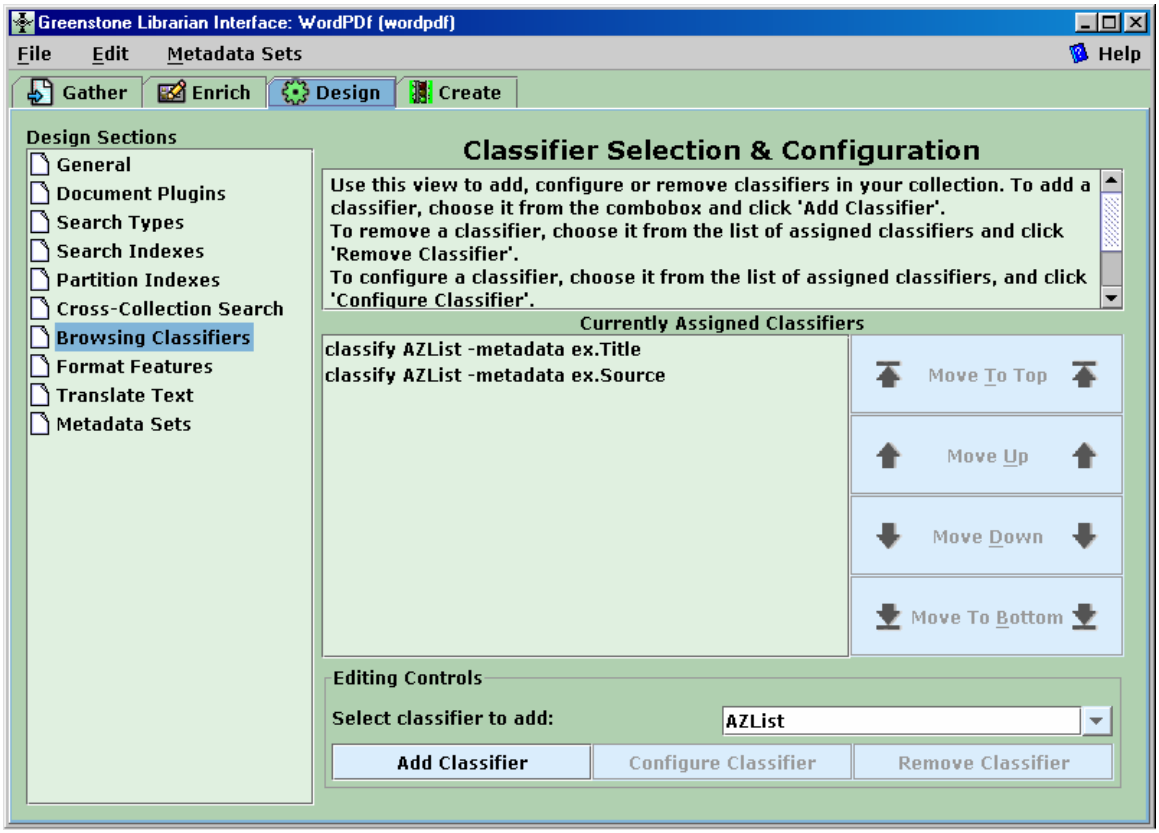
Collection Design (cross-collection search configuration.)

GLI STEP 3: DESIGNING THE COLLECTION

The Browsing Classifiers section

The next section is classifier selection and configuration.

This section explains how to assign 'classifiers'. Classifiers are used to provide browsing functionality to a collection. For this collection, Greenstone has provided us with two classifiers, namely, AZList for ex.Title and ex.Source as shown in the figure below.



Collection Design (classier selection and configuration)

What this means is that Greenstone has already provided an alphabetical browsing list for extracted titles and extracted source file names for our collection. Let us remove the default classifiers and add our own classifiers for the Title and Creator metadata fields.

To remove the default classifiers:

- select them one at a time and
- click 'Remove Classifier'.

Let us now add two classifiers, namely, AZList for dc.Title and AzcompactList for dc.Creator, for our collection.

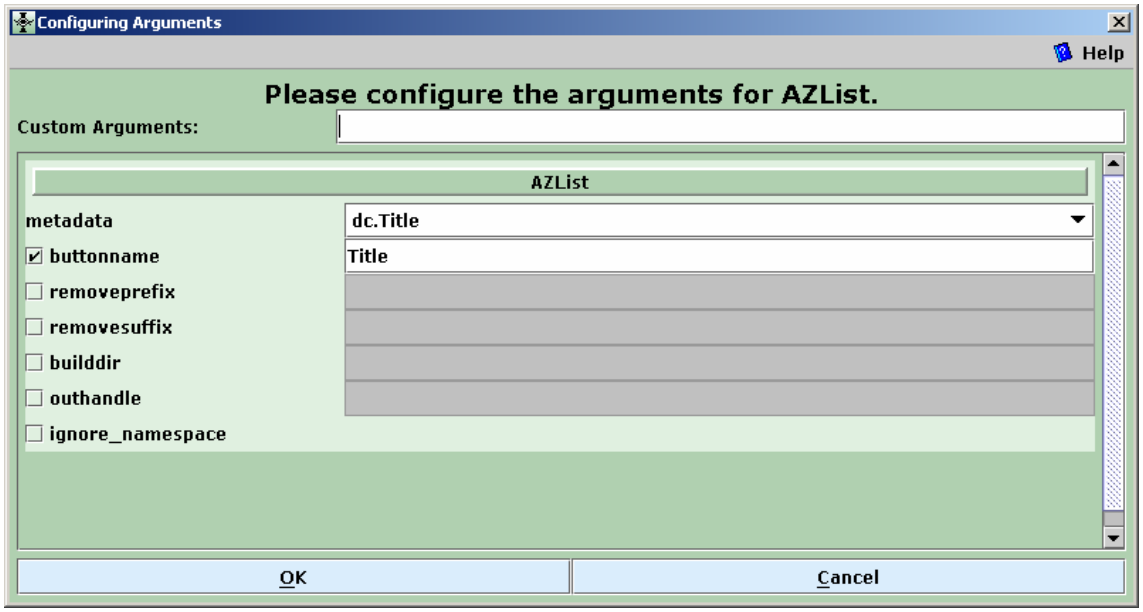
To add a classifier:

- select one from the 'Select classifier to add' drop-down list and
- click on Add Classifier.

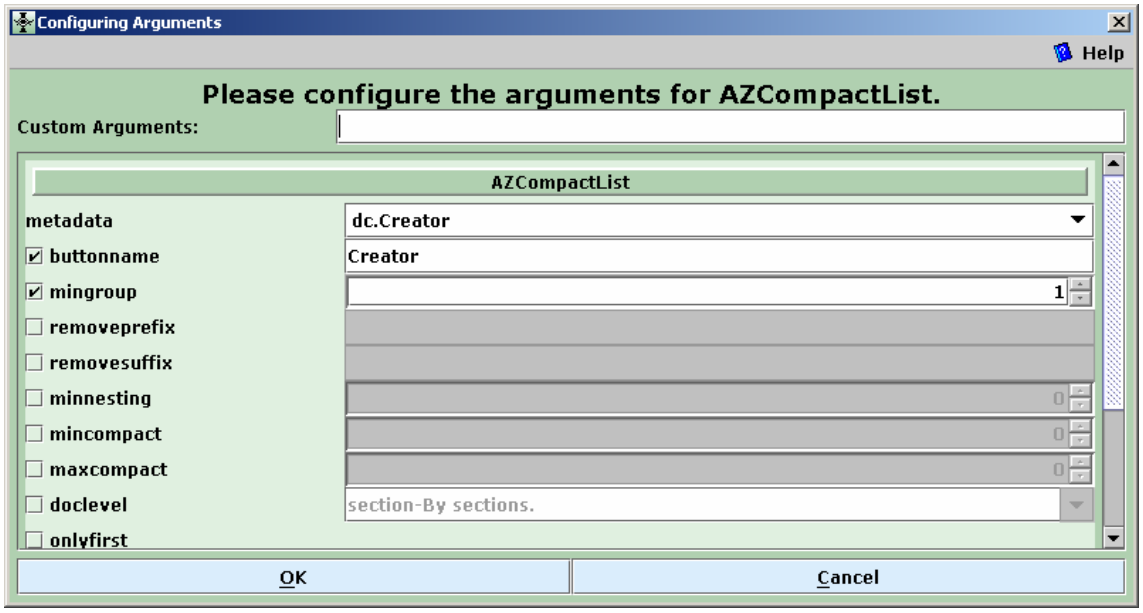
GLI STEP 3: DESIGNING THE COLLECTION

Let us now add an AZList classifier for our collection. Select AZList from the 'Select classifier to add' drop-down list and click on Add Classifier. A pop-up screen as shown in figure 17 should come up. Select dc.Title metadata for this classifier and provide a buttonname by checking the 'buttonname checkbox. Let the name be 'Title'.

Similarly add AZCompactList for dc.Creator. Let the button name be 'Creator'. Furthermore, to group all publications by the same author and generate a bookshelf icon, select the 'mingroup' checkbox and increase this number to 1. The screen shot is shown in this figure:



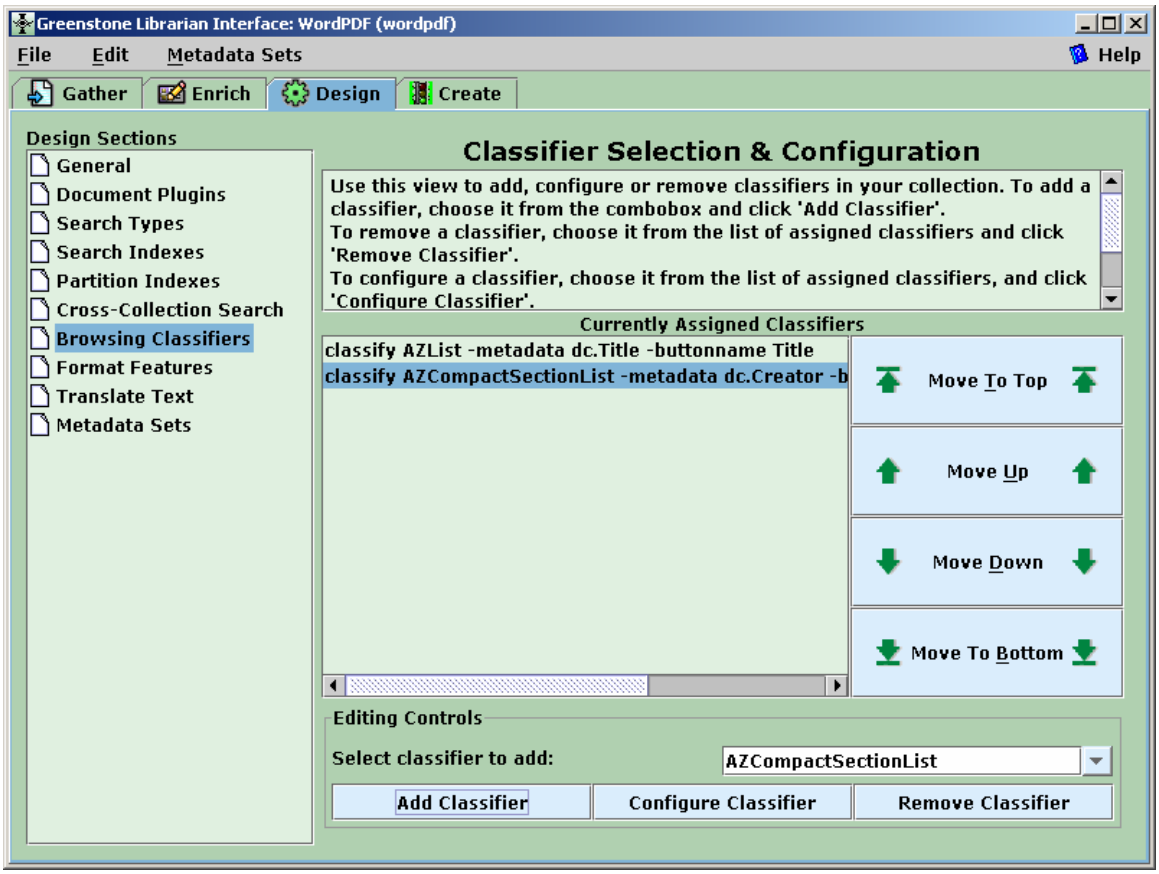
Collection Design (adding a classifier - Title)



Collection Design (adding a classifier - Creator)

GLI STEP 3: DESIGNING THE COLLECTION

After adding the two classifiers the ''browsing classifiers' section looks as shown in the figure below.



Collection Design (classifier selection and design)

The Format Features section

The next section is the 'Format Features'. Format commands control the structure and appearance of search and browse results and the display of selected documents. Format commands are not easy to develop. You should read chapter 2 of the Greenstone developer's guide. We are not making any changes in this section as far as our collection is concerned.

The Translate Text section

The 'Translate Text' section is used to review and assign translations of text fragments in a collection. We are not going to do any translation for our collection.

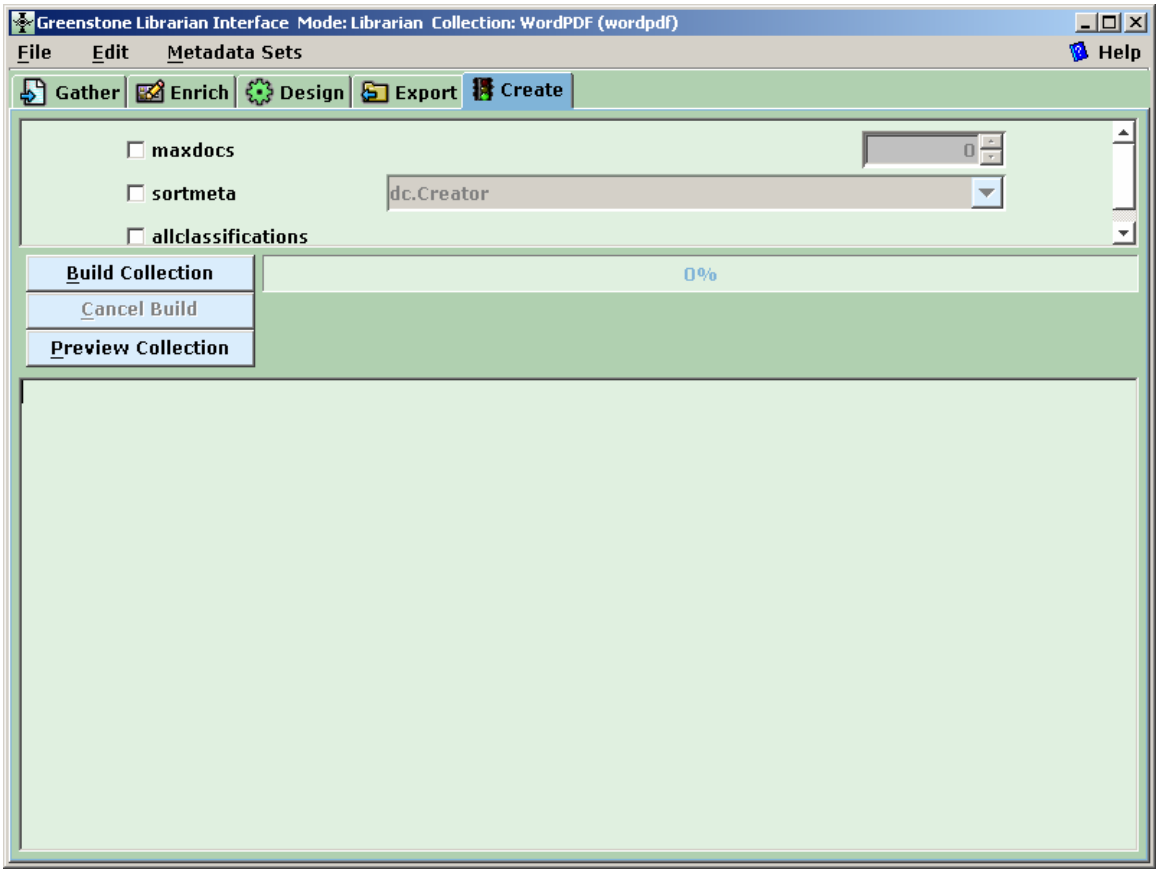
The Metadata Sets section

The 'Metadata Sets' is the last part of the Design Section. In this read-only design view you can review the Metadata Sets in your collection, identify what Elements they contain, and see how they will appear in the collection configuration file. We are now all set to create and build the 'WordPDF' collection!

GLI STEP 4: CREATE – BUILD THE COLLECTION

After having defined a new collection, collected documents for the collection, annotated it with metadata and designed its appearance, you are now ready to produce the collection using Greenstone. Do not get confused with Greenstone and GLI. GLI is just one of the tools which can be used to build a collection using Greenstone.

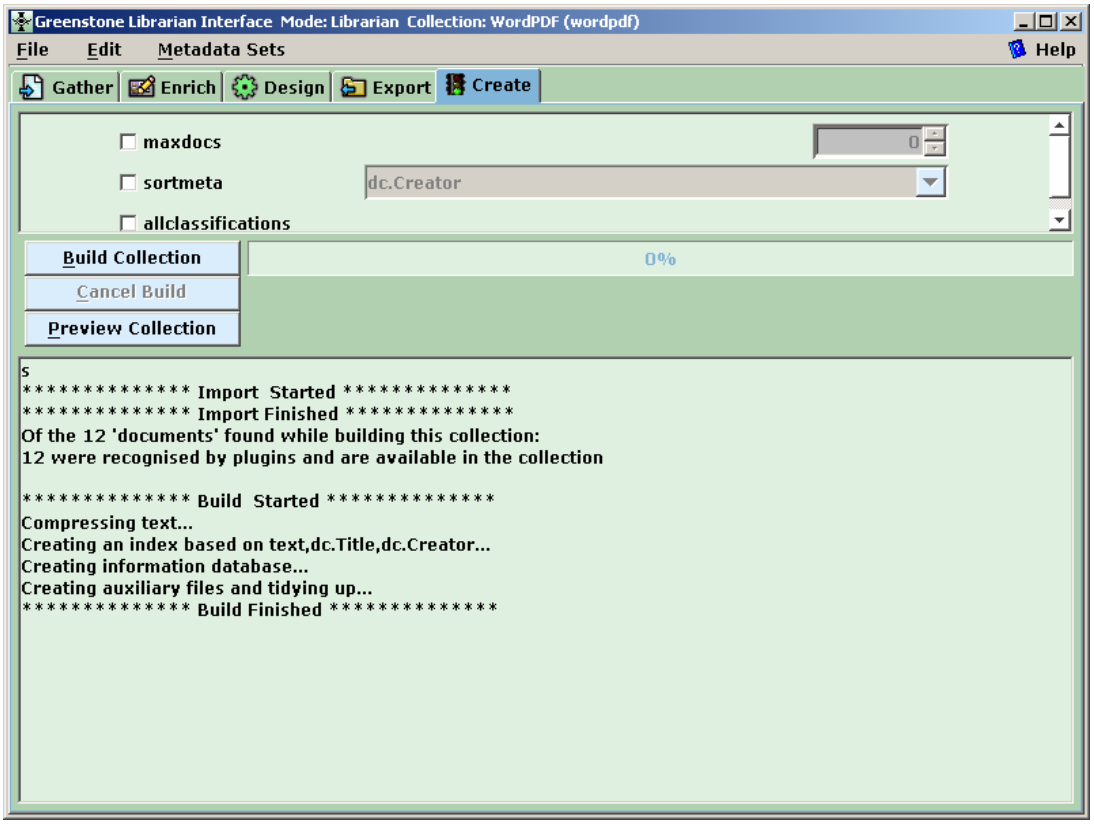
Click the 'Create' tab. The create screen as shown in the following figure comes up.



The Create View

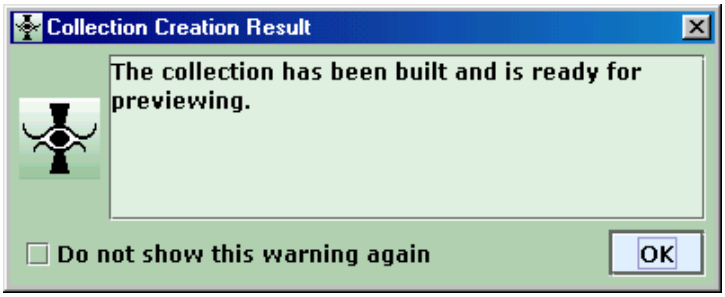
GLI STEP 4: CREATE – BUILD THE COLLECTION

Let us not make any changes to 'maxdocs', 'sortmeta' and 'allclassifications' options. Click on 'Build Collection' to import and then build the 'WordPDF' collection. A progress bar will indicate the status of importing and building. A few messages will be displayed in the text area. The following figure shows a screen shot of the collection building process.



Build Collection

The time taken for collection building process depends on the number and types of source documents, and of course, on your PC's configuration. If you have been correctly following the GLI collection building procedure, a small pop-up window as shown in this figure should come up once the collection has successfully been built:

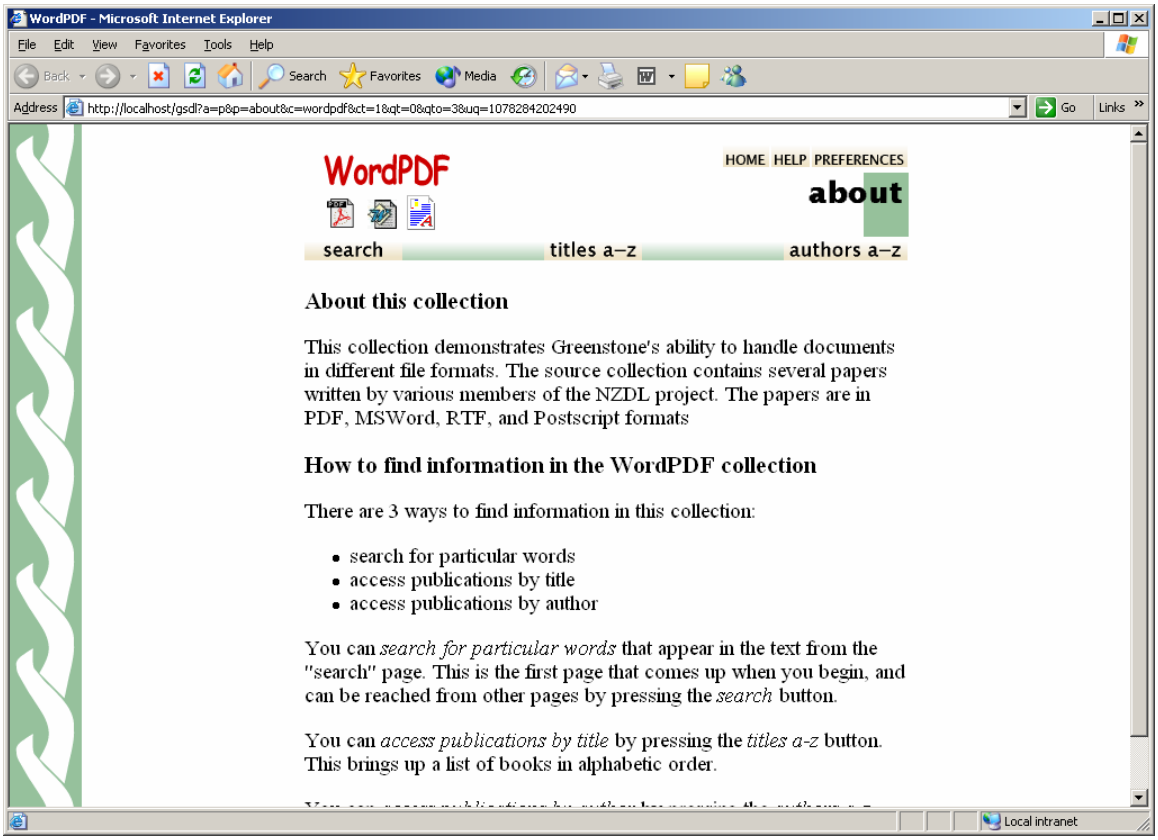


Collection Creation Result

Did it? Congratulations! You have successfully built your fist Greenstone collection using GLI. You can now preview the collection.

GLI STEP 5: PREVIEW THE COLLECTION

The Preview step is used to view the collection that has been built. In practice, previewing often shows deficiencies in the collection design or in the individual metadata values. The user frequently returns to earlier stages to correct these. The preview button and the 'Create' tab become active once the collection has been created, and take you to the collection in your regular Greenstone installation, as shown in the figure:



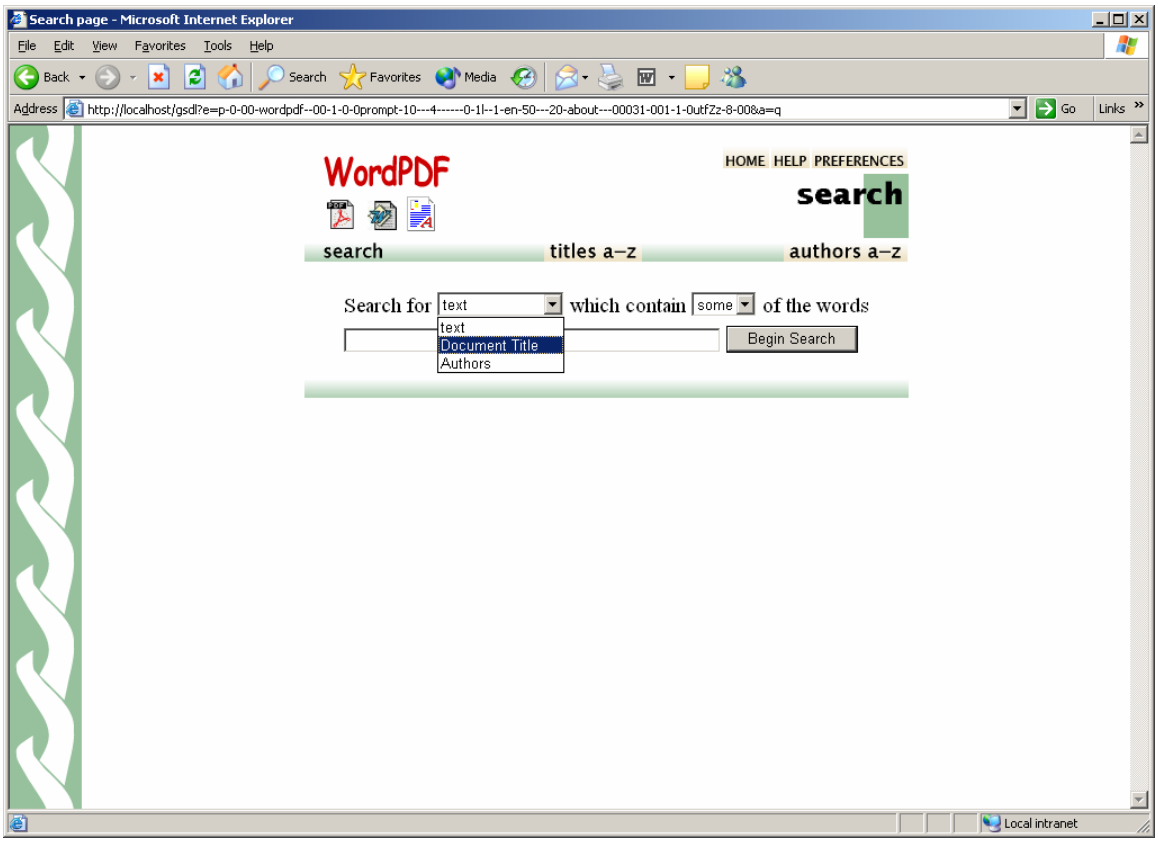
Greenstone Homepage for WordPDF Collection

You can now search and/or browse the WordPDF collection. Two types of searching are supported – plain and form. Click on the 'Search' button to do a plain search.

No collection can ever be an ideal collection. There is always a scope for further refinement or improvement in terms of additional indexes, appearance, or navigation. For the WordPDF collection, we created two additional indexes, namely 'Titles' and 'Creators'. These indexes correspond to the titles and author/s of the source collection you have keyed-in during the 'Enrich' section.

GLI STEP 5: PREVIEW THE COLLECTION

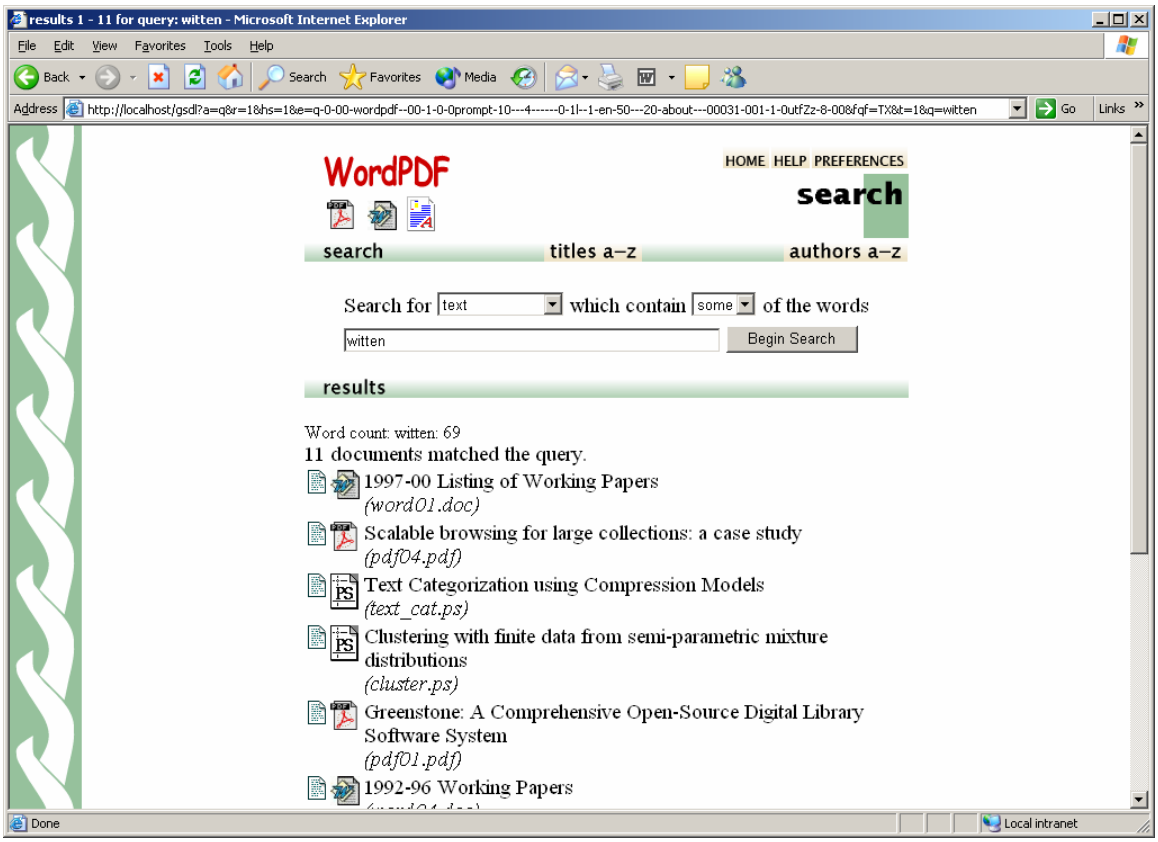
The WordPDF collection will have three indexes as shown in the following figure. The first index, namely, text was automatically created by Greenstone. This index corresponds to full text. Depending on our requirements, we have the option of either retaining the default indexes or not. This can be controlled during the design section (search indexes) stage.



Plain Search View

GLI STEP 5: PREVIEW THE COLLECTION

Enter a search term, 'witten', and click on Begin Search. Titles of the matched records with the corresponding filenames will be displayed as shown in this figure:

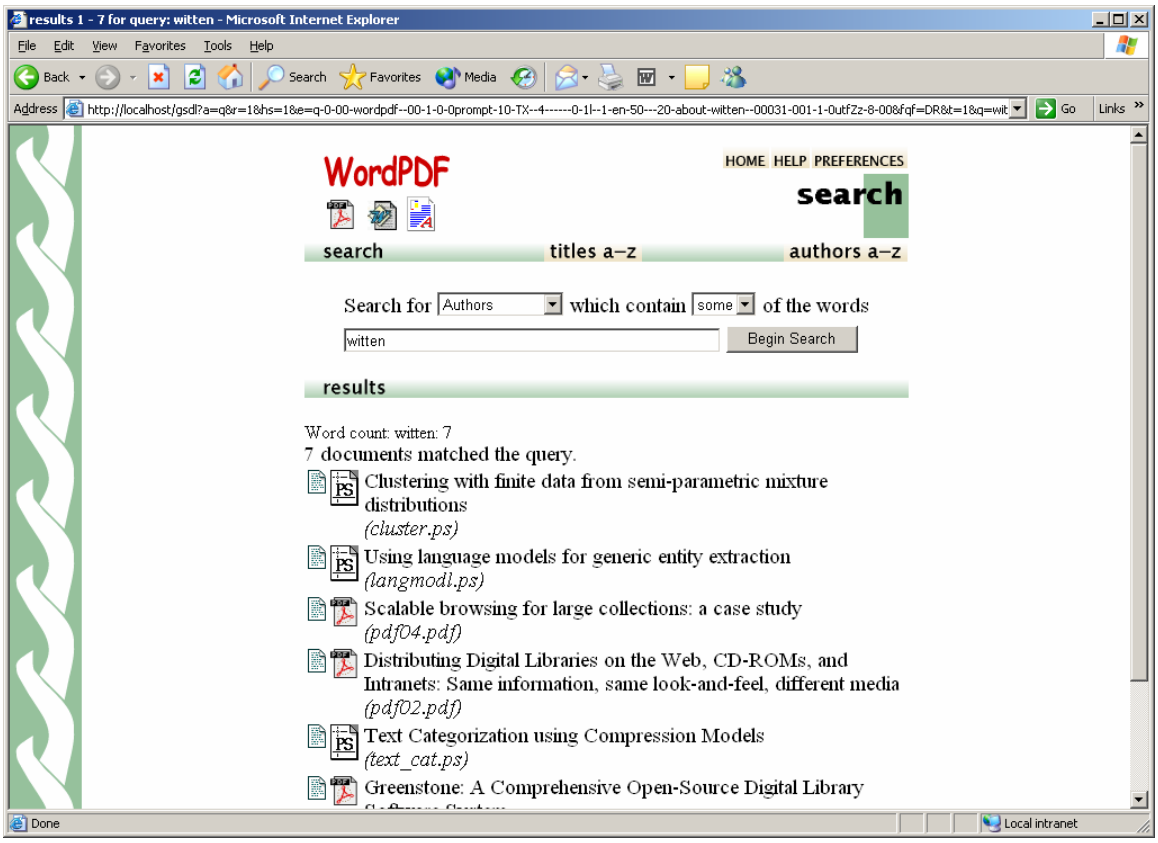


Search Results

GLI STEP 5: PREVIEW THE COLLECTION

You can restrict the search to a specific field.

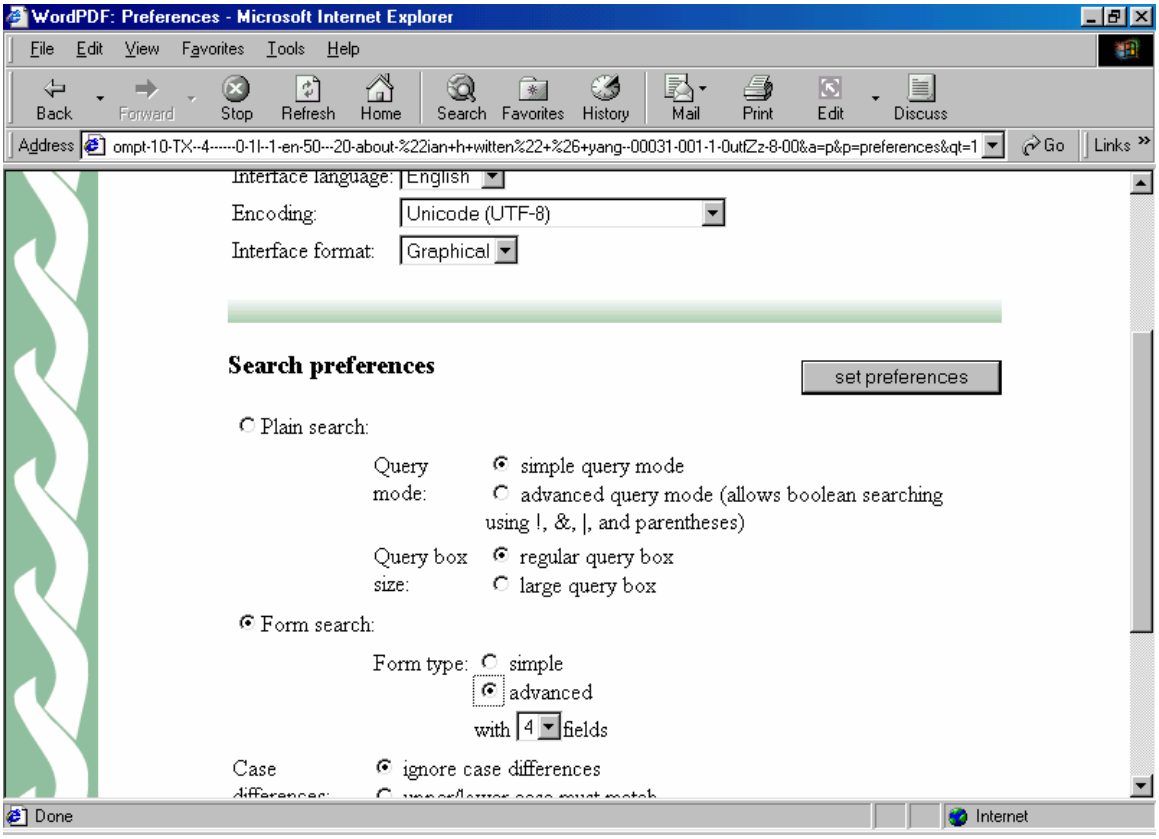
The same search term, 'witten', when restricted to the 'Creators' index retrieves only four documents as shown in the following figure:



Index specific Search

GLI STEP 5: PREVIEW THE COLLECTION

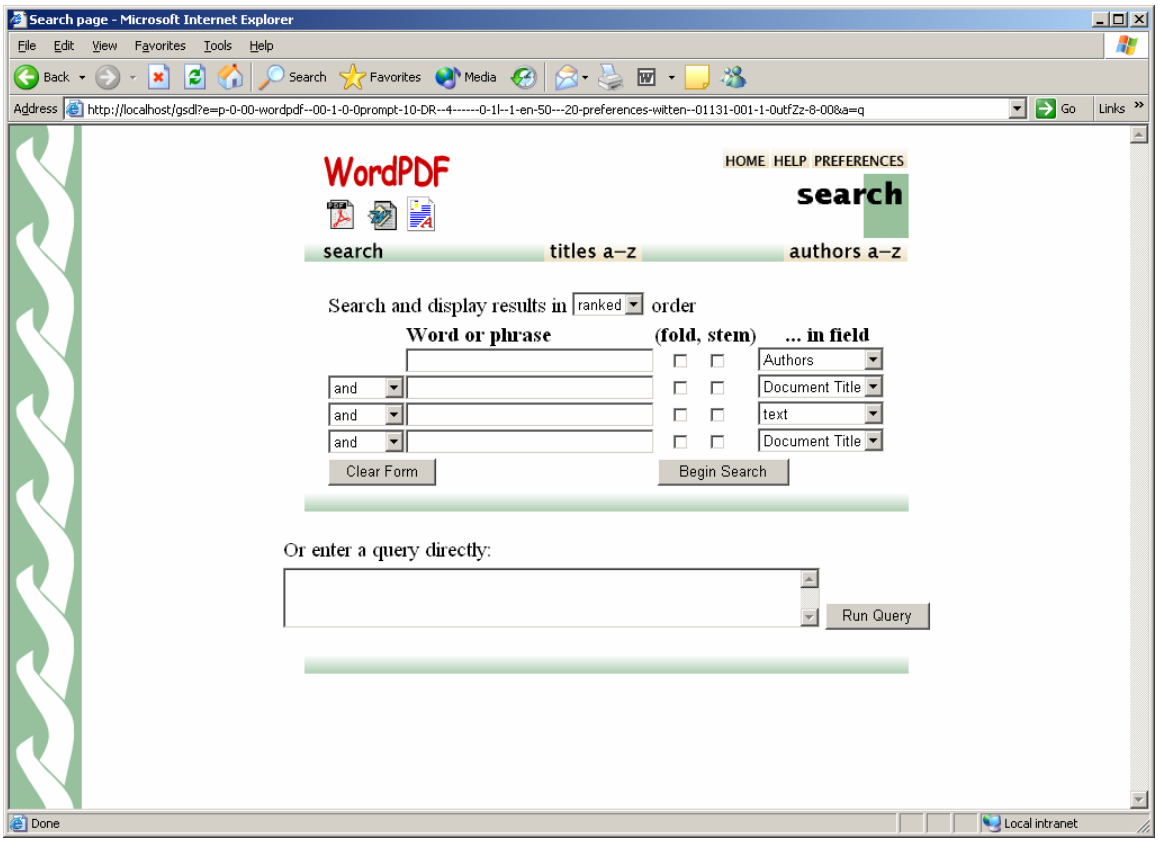
You can do a form-based search which facilitates searching across fields. During the design section (search types), we chose 'plain search' as the default search type. To do a form based search, click on 'PREFERENCES". Check the 'Form Search' option and check the 'advanced' 'Form Type'. Also, ensure that 'ignore case differences' and 'ignore word endings' are checked. Save the changes by clicking on 'set preferences'. A partial view of this screen is shown in the figure below.



Search Preference Settings

GLI STEP 5: PREVIEW THE COLLECTION

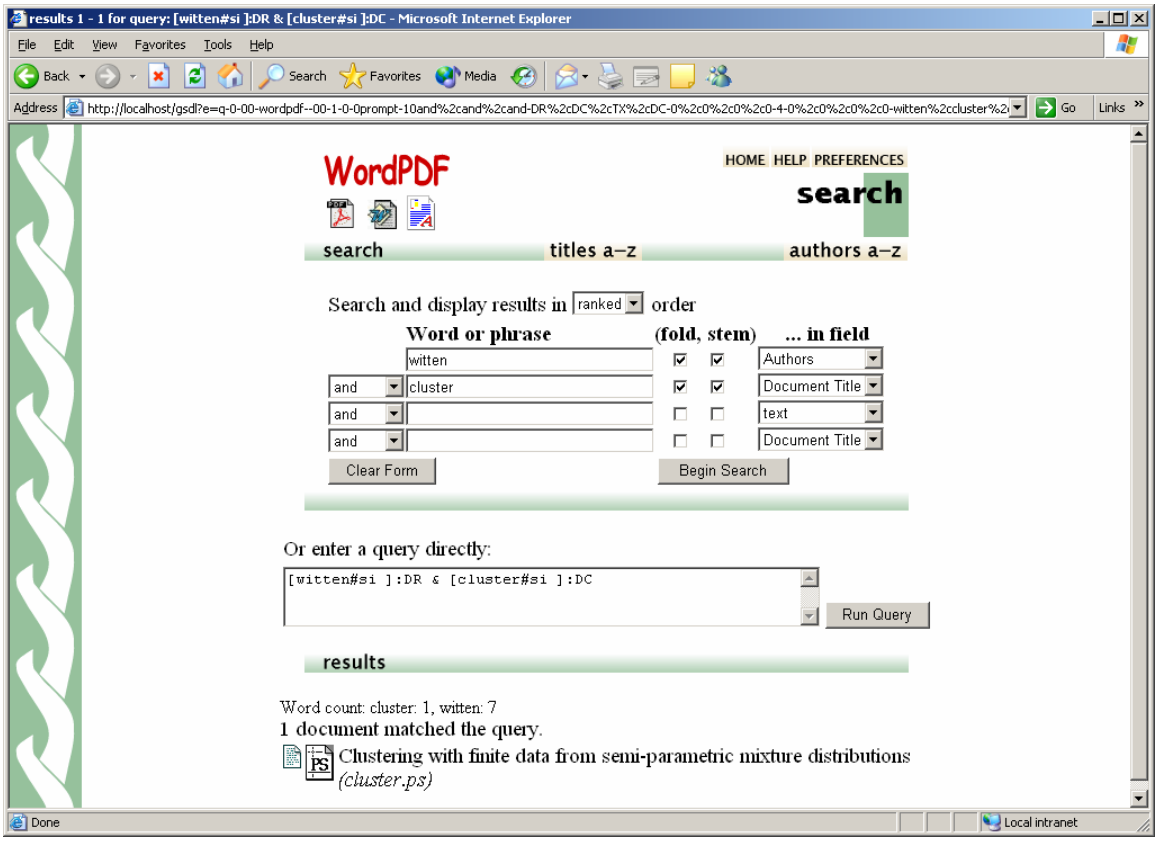
If you now click on the 'Search' button, you should get the advanced form based search interface as shown in the figure below.



Advanced Form Based Searching

GLI STEP 5: PREVIEW THE COLLECTION

A sample field-based search is shown in the figure below. In this search, the first search term 'witten' is searched for in 'Creators' and is combined, using the Boolean operator & (AND), with ' the search term 'cluster' in the 'titles' field. The search results in one document. Ensure that the 'fold' and 'stem' check boxes are checked.



Advanced Form Based Searching