LAB 5:

Bibliographic Collections: MARC & CDS/ISIS

5.1. Bibliographic collection—Part A: MARC

This exercise looks at using fielded searching in a collection. Fielded searching is best used for metadata rich collections. Here we use bibliographic data in MARC format.

1. Start a new collection called **Papers Bibliography** which will contain a collection of example MARC records of the working papers published at the Computer Science Department, Waikato University. Enter the requested information and base it on **-- New Collection --**..

2. In the **Gather** panel, open the *sample_files* → *marc* folder, drag *CMSwp-all.marc* into the right-hand pane and drop it there. A popup window asks whether you want to add **MARCPlugin** to the collection to process this file. Click **<Add Plugin>**, because this plugin will be needed to process the MARC records.

3. Now select **Browsing Classifiers** within the **Design** panel and **remove** the default classifier for **Source** metadata.

4. In the **Search Indexes** section, **remove** the **ex.Source** index. In this collection all records are from the same file, so **ex.Source** metadata, which is set to the filename, is not particularly interesting or useful.

5. Switch to the **Create** panel, **build** the collection, and **preview** it. Browse through the *Titles* and view a record or two. Try searching—for example, find items that include **graphics**.

6. Back in the Librarian Interface, go to the **Browsing Classifiers** section of the **Design** panel. Select **AZCompactList** from the **Select classifier to add:** drop down menu, and click **<Add Classifier...>**. In the popup window, select **dc.Subjects and Keywords** as the metadata item. Click **<OK>**.

7. ***AZCompactList*** *is like* ***List****, except that terms that appear multiple times in the hierarchy are automatically grouped together and a new node, shown as a bookshelf icon, is formed.*

8. **Build** the collection and **preview** the result.

Using fielded searching

9. Collections built with MGPP (the default indexer) provide the option of fielded searching. In the browser, go to the *PREFERENCES* page. You will notice that there is a **Query style:** option which enables you to switch between "normal" and "fielded" search. Change to fielded search now and click on the **Search** button. The search form has changed to a fielded form.

10. You can specify which search form types are available for a particular collection, and which one is the default, using the **SearchTypes** format statement. In the **Format** panel select **Format Features** from the left-hand list. Select the **SearchTypes** format statement from the list of assigned formats, and set the contents to **form**. This will make only fielded searching available for this collection.

    *Search type options include **form** and **plain**. You can specify one or both separated by a comma. If both are specified, the first one is used as the default: this is the one that the user will see when they first enter the collection.*

11. **Preview** the collection again. Notice that the collection's home page no longer includes a query box. (This is because the search form is too big to fit here nicely.) To search, you have to click **Search** in the navigation bar. Note that the *PREFERENCES* page has changed so that the "normal" query style is no longer offered.

12. Look at the search form in the collection. There are two fields that can be searched: *text* and *Title*. Add some more fields to search on by going back to the Librarian Interface.

13. In the **Design** panel, go to the **Search Indexes** section. Add a new index based on **dc.Subjects and Keywords** by clicking **<New Index>**, selecting **dc.Subjects and Keywords** in the list of metadata, and clicking **<Add Index>**.

14. **Rebuild** the collection and **preview** the results. Notice the extra field in the **... in field** drop-down menus in the search form. You can do quite complicated queries by searching for words in different fields at the same time.

15. To change the text that is displayed in the drop-down menus of the search form, go to the **Search** section of the **Format** panel. Here you can change the display text for the indexes.

Exploding the database

16. Go to the **Enrich** panel and try to see the metadata. It doesn't appear! This is because the metadata is associated with records inside the file, not the file itself.

    Metadata file types, such as MARC, CDS/ISIS, BibTex etc. can be imported into Greenstone but their metadata cannot be viewed in the Librarian Interface. To edit any metadata you need to go back to the program that created the file.

Greenstone provides a way of *exploding* a metadata database so that each record appears as an individual document, with viewable and editable metadata. This process is irreversible: once this step has been done, the database is deleted and can no longer be used in its original program.

17. In the **Gather** panel, you may notice that the MARC database has a different coloured icon to other files. This green icon indicates that a file is a metadata database that can be exploded. Right-click on the file and choose **Explode Metadata Database** from the menu. A new window opens, containing options for the exploding process. A description of each option can be obtained by hovering the mouse over the option.

    Turn on the **metadata_set** option by checking its box. This option indicates which metadata set to explode the metadata into. The default set is the "Exploded Metadata Set"—a metadata set which initially has no elements in it, but will receive a new element for each metadata field retrieved from the database.

18. Click **<Explode>** to start the exploding process. This may take a short while, depending on the size of the database.

19. Once exploding has finished, the MARC database file will have been deleted, and a folder created in its place. This folder contains an empty file for each record in the original database. The metadata for these records can be viewed and edited by switching to the **Enrich** panel.

20. Because the MARC file is no longer present, and the collection contains empty (.nul) files, we need to change the list of plugins. In the **Document Plugins** section of the **Design** panel, remove **MARCPlugin**.

21. **Rebuild** and **preview** the collection. You will notice that the *Titles* classifier displays the filename not the record title, the *Subjects* classifier is empty, searching no longer returns any results, and the document display is useless.

22. Although the *Titles* classifier was built on **ex.Title**, it still displays the correct titles, but in the **Enrich** panel you can see the **ex.Title** metadata are actually the filenames rather than titles of the MARC records. This is because the default **VList** format uses the **exp.Title** metadata. In the **Format Features** section of the **Format** panel, select **VList** in the list of assigned format statements. The resulting format statement looks like:

```
<td valign="top">[link][icon][/link]</td>
<td valign="top">[ex.srclink]{Or}{[ex.thumbicon],[ex.srcicon]}[ex./srclink]</td>
<td valign="top">[highlight]
{Or}{[dc.Title],[exp.Title],[ex.Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}
```

Since there is no **dc.Title** metadata and **exp.Title** comes before **ex.Title**, the exploded titles will be displayed.

Reformatting the collection to use the exploded metadata

The collection previously used extracted (ex.) metadata, but now it uses exploded (exp.) metadata. The classifiers and search indexes were built on ex metadata, which is why they no longer work properly.

There is also no longer any text in the documents. Previously, **MARCPlugin** stored the raw record as the "text" of each record. Now that the metadata is in the Librarian Interface, there is no longer the concept of raw record, and so there is no text.

We need to modify the collection design to take note of these changes.

23. In the **Search Indexes** section, change the Title index to use **exp.Title**: select the Title index in the **Assigned Indexes** list and click **<Edit Index>**. Deselect **dc.Title** and **ex.Title** in the list of metadata, and select **exp.Title**. Click **<Replace Index>**.

24. Remove the **dc.Subject and Keywords** index by selecting it in the **Assigned Indexes** list and clicking **<Remove Index>**. Add an index on **exp.Subject**: click **<New Index>**, select **exp.Subject** in the metadata list, and click **<Add Index>**.

The text index is no longer any use, so remove that index too.

25. To enable combined searching across all indexes at once, click **<New Index>**, tick the **Add combined searching over all assigned indexes (allfields)** checkbox, and click **<Add Index>**. Move this to the top of the list using the **<Move Up>** button, so that it appears first in the drop down list. Click **<Set Default Index>** on the right so that it becomes the default field for searching.

26. To explicitly use the **exp.Title** metadata, in the **Browsing Classifiers** section, change the **dc.Title;Title List** to use **exp.Title** metadata. Double click the **dc.Title;Title List** in the **Assigned Classifiers** list, and change the **metadata** option to use **exp.Title**. Click **<OK>**. Do the same thing for the Subject **AZCompactList**, changing **dc.Subject and Keywords** to **exp.Subject**.

27. **Rebuild** and **preview** the collection. The classifiers should be back to normal and searching should now work.

28. In the **Format Features** section of the **Format** panel, select **VList** in the list of assigned format statements.

    There is no dc metadata for this collection, so replace `{Or}{[dc.Title],[exp.Title],[ex.Title],Untitled}` with `{Or}{[exp.Title],[ex.Title],Untitled}`.

    There are no source or thumb icons, so remove the second line: `<td valign="top">[ex.srclink]{Or}{[ex.thumbicon],[ex.srcicon]}[ex./srclink]</td>`.

The ex.Source metadata is set to the nul filename, so remove that from the display:
remove `{If}{[ex.Source],<br><i>([ex.Source])</i>}`

The resulting format statement looks like:

```
<td valign="top">[link][icon][/link]</td>
<td valign="top">[highlight]
{Or}{[exp.Title],[ex.Title],Untitled}
[/highlight]</td>
```

29. Clear the **DocumentHeading** format statement by selecting it in the list of assigned format statements and deleting the contents in the **HTML Format String**. The record Title will be displayed as part of the **DocumentText** format, so we don't need it here.

30. Next, edit the **DocumentText** format statement. Delete the contents and replace it with

```
<table>
<tr><td>Title:</td><td>[exp.Title]</td></tr>
<tr><td>Subject:</td><td>[exp.Subject]</td></tr>
<tr><td>Publisher:</td><td>[exp.Publisher]</td></tr>
</table>
```

    The *DETACH* and *NO HIGHLIGHTING* buttons are not very useful for this collection, so lets get rid of them. Edit the **DocumentButtons** format statement to make it empty.

31. Press the **<Preview Collection>** button to preview the collection.

5.2. Bibliographic collection —Part B: CDS/ISIS

*This exercise is similar to the **Bibliographic collection** exercise, except that a CDS/ISIS database is used instead of a MARC database, and we do not explode the database.*

1. Start a new collection called **ISIS Collection**.

2. Drag the files from *sample_files* → *isis* (excluding the *format_tweaks* folder and README.txt file) into the collection.

3. **Build** and **preview** the collection. The default indexes, classifiers and formats are not very useful for this data. There is no metadata searching, and the *Titles* classifier is completely empty. The filenames classifier is useless because all records come from the same file.

4. In the **Search Indexes** section of the **Design** panel, remove the useless Source and Title indexes, and add new indexes for **Photographer^all, Country^all** and **Notes^all** metadata.

    CDS/ISIS metadata has subfields, and these are represented using ^.

5. In the **Browsing Classifiers** section, remove the existing (useless) classifiers for **Title** and **Source**, and add a new **List** for **Photographer**.

6. **Rebuild** and **preview** the collection.

7. In the **Format Features** section of the **Format** panel, change the **VList** format statement to display **Photographer** and **Notes** metadata. Change it to look like:

```
<td valign=top>[link][icon][/link]</td>
<td valign=top><b>[ex.Photographer^all]</b><br/>[ex.Notes^all]</td>
```

8. Make fielded searching the default by changing the **SearchTypes** format statement to **form,plain** (instead of **plain,form**).

ISISPlugin stores a nicely formatted version of the record as the document text, and this is what is displayed when we view a record. Lets tidy it up a little more.

9. Remove the *DETACH* and *NO HIGHLIGHTING* buttons by setting the **DocumentButtons** format statement to empty.

10. Remove the "Untitled" at the top of the document by setting the **DocumentHeading** format statement to empty.

11. Finally, lets link to the raw record, which is stored as **ISISRawRecord** metadata. Edit the **DocumentText** format statement to look like the following. (This format can be copied from *sample_files* → *isis* → *format_tweaks* → *document_text.txt*.)

```
<p>[Text]</p>
{If}{_cgiargshowrecord_,
<p><b>CDS Record:</b><br/><tt>[ISISRawRecord]</tt></p>
<center><a href=\'_gwcgi_?e=_cgiarge_&a=d&d=_cgiargd_\'>Hide CDS
Record</a></center>,
<center><a
href=\'_gwcgi_?e=_cgiarge_&a=d&d=_cgiargd_&showrecord=1\'>Show CDS
Record</a></center>
}
```

Preview the collection.

****