Lab 4 (1 3/4 hours): Configuring collections

Part I – Working with full text data

The first exercise focuses on additional features that can be added to text-based collections such as HTML, Word and PDF.

Revision:

1. Start a new collection called **tudorX** (the 'X' is for extra). Fill out the requested fields with appropriate information. Choose Dublin Core as the metadata set to use with this collection.

2. In the **Gather** panel, open the *html_large* folder in *test_files* on the CD-ROM and copy *englishhistory.net* into your new collection.

3. Build the collection, preview it and check the extracted metadata.

Now add some extra features:

4. To add a PHIND classifier (an interactive hierarchical phrase index), switch to the **Design** panel and choose the **Browsing Classifiers** item from the left-hand list.

   Choose **Phind** from the **Select classifier to add** menu. Click **Add Classifier**. A window pops up asking for configuration options: leave the values at their preset defaults (this will base the phrase index on the full text) and click **<OK>**.

   For tidiness' sake, **remove** the **classifier** for **Source** metadata (included by default) from the list of currently assigned classifiers, because this adds little to the collection.

   Build the collection again, **preview** it and try out the new **phrases** option in the navigation bar. An interesting search term for Phind is **king**.

We now experiment with restricting searching to different areas of a collection:

5. To partition the collection into subcollections, select **Partition Indexes** from the **Design** panel. **Define** the following three **filters**:
   - Subcollection filter name: "monarchs"; match against Filename; regular expression "monarchs"; Include.
   - Subcollection filter name: "relatives"; match against Filename; regular expression "relative"; Include.
   - Subcollection filter name "other"; match against Filename; regular expression "(monarchs|relative)"; Exclude.

6. Move to the **Assign Partitions** section of **Partition Indexes** (it's a tabbed panel *within* the **Design** panel). Add three partitions, one for each partition: *monarchs*,

relatives and *other*. This might seem redundant, having already entered these terms in the previous step; but the use of these partitions is connected with multilingual support.

7. Rebuild the collection and preview it to see the changes in the search page. Try searching for documents relating to ‘mary’ in the *relatives* partition of the collection.

We now manually add metadata that explicitly expresses a subject hierarchy, and build a classifier that exploits it:

8. Switch to the **Enrich** panel and open the *tudor* folder in the left-hand panel.

9. Click on the *citizens* folder and then select **dc.Subject** from the right-hand side as the value that will be manually assigned. In its **value** box enter:

   Tudor Period\Citizens

   and click **Append**. An alert appears stating that all documents within this folder will inherit this metadata item. Click **<OK>**.

10. Repeat the process for the *monarchs* and *relatives* folders, assigning the terms **Tudor Period\Monarchs** and **Tudor Period\Relatives** respectively.

11. Now switch to the **Design** panel and add a Hierarchy classifier in the **Browsing Classifiers** part. Set its configuration option for **hfile** to **dc.Subject.**txt, its **buttonname** to **Subject**, and its **metadata** to **dc.Subject**.

12. Finally rebuild your collection and preview it. There will be a “subject” button in the navigation bar that presents documents in the *citizens*, *monarchs* and *relatives* folders as a hierarchy under the shared root term **Tudor Period**.

Part II – Working with bibliographic data

We now turn attention to configuring collections based on bibliographic content. We use MARC as the sample input format:

1. Start a new collection called **Beatles Bibliography** which will contain a collection of MARC records from the U.S Library of Congress on the Beatles. Enter the requested information. There is no need to include any metadata sets.

2. In the **Gather** panel, open the **sample_marc** folder on the CD-ROM and include it in the collection by dragging **locbeatles50.marc** into the right-hand pane and dropping it there.

3. In the **Document Plugins** section of the **Design** panel, add **MARCPlug** to the list. Leave its options at their preset defaults.

4. Remove the plugins **TextPlug** to **PSPlug** (**ZIPPlug**, **GAPlug** and **MARCPlug** remain above the dividing line). It is not strictly necessary to remove these

redundant plugins, but it is good practice to include only plugins that are needed, to avoid unwanted (and unexpected) side effects.

[Note: you must delete plugins one at a time. Alt-R is a hot key for this, and you can work faster by clicking on a plugin, deleting it with Alt-R, and continuing in this fashion.]

5. Now select **Browsing Classifiers** from within the **Design** panel and **remove** the default classifier for **Source** metadata. In this collection all records are from the same file, so Source metadata, which is set to the filename, is not particularly interesting.

6. Switch to the **Create** panel, **build** the collection, and **preview** it. Browse through the **titles a-z** and view a record or two. Try searching—for example, find items that include **George Martin**.

7. Add an **AZCompactList** classifier for the **Subject** metadata. Select this item from the relevant menu and click **<Add Classifier>**. In the popup window, select **ex.Subject** as the metadata item, activate the **mingroup** option and set its field to **1**.

   AZCompactList is like **AZList,** 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. Setting **mingroup** to 1 means that the bookshelf appears even when there is just one item, and is done here to provide a more uniform display.

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

9. Make each bookshelf node show how many entries it contains by appending this to the **Format Features** for **VList** format statement in the **Design** panel:

   ```
   {If}{[numleafdocs],<td><i>([numleafdocs])</td>}
   ```

   Click **<Replace Format>**, switch to the **Create** panel, and click **<Preview>** (no need to rebuild).

10. Next add fielded searching. In the **Design** panel select **Search Types** from the left-hand list and activate the **Enable Advanced Searches** options.

11. **Rebuild** the collection and **preview** the results. 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 to control the advanced searching options.

To finish the collection off, brand it with an image that will be used to represent the collection on the Greenstone page, and appear at the top of each page of the collection:

12. Using the Windows file manager, open

    My Computer → GSDL Fiji → test_files → sample_marc

    (you will need to right-click *GSDL Fiji* to avoid autostarting it). Then open another window that shows the files in:

    My Computer → Local Disk (C:) → Program Files → gsdl → collect → beatlesb

    Drag *beatles_logo.jpeg* from the *sample_marc* folder and drop it into *beatlesb*.

13. Return to the Librarian Interface, switch to the **Design** panel, and select General from the left-hand list. In the box for **URL to about page icon** enter:

    ```
    _httpprefix_/collect/beatlesb/beatles_logo.jpeg
    ```

    Alternatively you can use the **browse** button to navigate to this file in

    My Computer → Local Disk (C:) → Program Files → gsdl → collect → beatlesb

    and select it.

14. Repeat this process for the **URL to home page icon**, using the same text (copy and paste it using the edit menu).

15. Now **rebuild** the collection and **preview** it.

Part III – Extra work

1. Add textual documents (downloaded from the web) to your "tourism" collection