Lab 3 (3/4 hour): Heterogeneous Documents and Multimedia

All the collections built so far handle a single document type. Now we examine a complex heterogeneous collection that includes multimedia elements of sound and pictures as well as textual and bibliographic documents. It centers around the Beatles *pop group, and source documents include lyrics and discography information in the form of HTML files, guitar tablature as text files, audio in MP3 and MIDI formats; images of album covers in JPEG format, and metadata in MARC format.*

Part I – Look at what can be done!

1. Copy the entire folder *workshop_files→heterogeneous→advbeat_large* (with all its contents) into your Greenstone *collect* folder. If you have installed Greenstone in the usual place, this is *My Computer→Local Disk (C:)→Program Files→gsdl→collect*. Put *advbeat_large* in there.

2. Re-start the Greenstone Digital Library Local Library Server by clicking the CD icon on the task bar and then pressing *Restart Library*.

3. Explore the Beatles collection. Note how the *browse* button divides the material into seven different types. Within each category, the documents have appropriate icons. When you click the audio icons you hear the music (assuming your computer is set up with appropriate player software). When you click the image thumbnails you see the images.

4. Look at the *Titles* browser. Each title has a bookshelf that often includes several different related items. For example, *Hey Jude* has a cover image, MP3 audio and MIDI versions, lyrics, and a discography item.

5. Observe the low quality of the metadata. For example, the four items under *A Hard Day's Night* (under "h" in the title browser) have different variants as their titles. The collection would have been easier to organize had the metadata been cleaned up manually first, but that would be a big job. A tiny amount of metadata was added by hand—fewer than ten items. The original metadata was left untouched and Greenstone facilities used to clean it up automatically. (You will find below that this is possible but tricky.)

6. In the Windows file browser, take a look at the files that makes up the collection, in the *workshop_files→heterogeneous→advbeat_large→import* folder. What a mess! There are over 450 files under eight top-level sub-folders. Organization is minimal, reflecting the different times and ways the files were gathered. For example, *html_lyrics* and *discography* are excerpts of web sites, and *cover_images* contains album covers in JPEG format. For each type, drill down through the hierarchy and look at a sample document.

Part II – Building a basic collection

7. Start a new collection (*File→New*) called **small_beatles**, basing it on the default "New Collection." (Basing it on the existing Advanced Beatles collection would make your life far easier, but we want you to learn how to build it from scratch!) Fill out the fields with appropriate information. Select the Dublin Core metadata set.

8. Copy the files provided in *workshop_files →heterogeneous→advbeat_small* into your new collection. Do this by opening up *advbeat_small*, selecting the eight items within it (from *cover_images* to *beatles_midi.zip*), and dragging them across. We use this small subset to speed up the repeated re-building that will be required during collection development.

9. Change to the **Enrich** panel and browse around the files. There is no metadata—yet. Recall that you can double-click files to view them.

10. Switch to the **Design** panel and look at the **document plugins**. The default list will process all HTML, PostScript, PDF, and text files, but the others will be omitted from the collection. To rectify this:

    - add MARCPlug with default settings;
    - add MP3Plug, again with default settings;
    - add ImagePlug.

    (The MIDI files require more advanced customization because there is no MIDIPlug. We will deal with them later.)

11. Change to the **Create** panel and build the collection.

12. Preview the result. Relying on default settings, as we have done here, works moderately well, but we can improve the organization and presentation of the collection by iterating over the design process. This is the focus of the remaining exercises.

13. You might want to correct some of the metadata manually—for example, the atrocious mis-spelling in the titles "MAGICAL MISTERY TOUR." These documents are in the discography section, with filenames that contain the same mis-spelling. Locate one of them in the Enrich panel. Notice that the extracted metadata element **ex.Title** is now filled in, and mis-spelt. You cannot correct this element, for it is extracted from the file and will be re-extracted every time the collection is re-built.

    - Instead, add **dc.Title** metadata for these two files: "Magical Mystery Tour." Change to the **Enrich** panel, open the discography folder and drill down to the individual files. Set the **dc.Title** value for the two offending items.

    Now there's a twist. The **dc.Title** metadata won't appear in Titles A–Z because the classifier has been instructed to use **ex.Title**. But changing the classifier to

use **dc.Title** would miss out all the extracted titles! Fortunately, there's a way of dealing with this by specifying a list of metadata names in the classifier.

- Change to the Design panel and select the browsing classifiers zone. Double-click the Title classifier (the first one) to edit its configuration settings.
- Type "dc.Title," before the *ex.Title* in the metadata box—i.e. make it read

  ```
  dc.Title,ex.Title
  ```

- Set Buttonname to *Title*

Rebuild the collection and preview it.

Part III – Simple customization

14. Now we will tidy up the collection and add some new features. First let's remove the AZList classifier for filenames, which isn't very useful, and replace it with a browsing structure that groups documents by category (discography, lyrics, audio etc.). Categories are defined by manually assigned metadata.

    - Change to the **Enrich** panel, select the folder *cover_images* and set its **dc.Format** metadata value to "Images". Setting this value at the folder level means that all files within the folder inherit it.
    - Repeat the process. Assign "Discography" to the *discography* folder, "Lyrics" to *html_lyrics*, "MARC" to *marc*, "Audio" to *mp3*, "Tablature" to *tablature_txt*, and "Supplementary" to *wordpdf*.
    - Switch to the **Design** panel and select the **browsing classifiers** zone.
    - Delete the **ex.Source** classifier (the second one).
    - Add an **AZCompactList** classifier. Select **dc.Format** as the metadata field and specify "Category" as the **buttonname**.

    Rebuild the collection and preview it. Notice that the Category browser contains no "Images" entry. This is because the image files have not yet been processed by ImagePlug due to its order in the list of plugins, as explained below.

15. Greenstone has no pre-defined button for "Category", so it appears in the navigation bar as text. It does, however, have a button for *Browse* (it's used in the Beatles collection you looked at in Part I).

    - Go back to the **AZCompactList** classifier for **dc.Format** and specify "Browse" as the **buttonname**.

    You will need to rebuild the collection for this to take effect.

16. Alongside the Audio files there is a MP3 icon, which plays the audio when you click it, and also a text document that contains some dummy text. This isn't supposed to be seen, but to suppress it you have to fiddle with a format statement.

    - Change to the **Design** panel and select the Format Features zone.
    - Ensure that **VList** is selected, and make the changes that are highlighted below. You need to insert three lines into the first line, and delete the second line.

      Change:

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

      to this:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
```

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

- Then click <**Replace Format**>:

Preview the result. If you are using the Greenstone Local Library server, change to the **Create** panel and click <**Preview Collection**>. You do not need to rebuild the collection because format statements are only used by the runtime system.

17. While we're at it, let's remove the source filename from where it appears after each document.

    - In the VList format feature, delete the text that is highlighted below:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[Source],<br><i>([Source])</i>}</td>
```

    Don't forget to click <**Replace Format**> after all this work! Preview the result (you don't need to rebuild the collection.)

18. There are sometimes several documents with the same title. For example, *All My Loving* appears both as lyrics and tablature (under *ALL MY LOVING*). The *Titles A–Z* browser might be improved by grouping these together under a bookshelf icon. This can be done with an AZCompactList.

    - Change to the **Design** panel and select the browsing classifiers zone.
    - Remove the **Title** classifier (at the top)
    - Add an **AZCompactList** classifier, and select **dc.Title,ex.Title** as its metadata.
    - Activate **min_group** and set it to 1. This gives a uniform appearance by creating a bookshelf for every title.
    - Finish by pressing <OK>
    - Move the new classifier to the top of the list (*Move Up* button).

    Rebuild the collection and preview it. Both items for *All My Loving* now appear under the same bookshelf. However, many entries haven't been amalgamated because of non-uniform titles: for example *A Hard Day's Night* appears as four different variants. We will learn below how to amalgamate these.

19. Make the bookshelves show how many documents they contain by inserting a line in the VList format statement in the **Design** panel:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
```

Don't forget to click <**Replace Format**>. Preview the result (you don't need to rebuild the collection.)

20. Add a Phind browsing classifier that sources its phrases from Title and text (the default setting).

21. To complete the collection, use the browse button of "URL to 'about page' icon" in the **General** zone of the **Design** panel to select the following image: *advbeatles_large→images→flick4.*

    Rebuild the collection and preview it.

Part IV – Advanced customization

To go further, the Librarian Interface must be in a more advanced mode. Click *File→Preferences→Mode* and change to Library Systems Specialist. Note from the description that appears that you need to be able to formulate regular expressions to use this mode fully. That is what we do below.

22. All the cover images are missing from the collection. This is because HTMLPlug gobbles up images on the assumption that they belong within web pages, and so ImagePlug, which is further down the plugin list, never sees them.

    - Move ImagePlug up above HTMLPlug in the list of plugins in the **Design** panel.

    But now ImagePlug will process all the images, including all those that belong to HTML files—and there are lots!

    - Restrict the images seen by ImagePlug by changing its *process_exp* to:

    ```
    (?i)(cover_images).*(\.jpe?g|\.gif|\.png|\.bmp|\.xbm|\.tif?f)$
    ```

      Do this by checking the box and then inserting the highlighted text "(cover_images).*" at the appropriate place in the existing string.

    Do you understand regular expressions? What we are doing is restricting ImagePlug to process image files whose file pathnames contain the string "cover_images"—which works because all the images we want are in the *cover_images* folder—and conclude with a filename extension of *.jpg*, *.jpeg*, *.gif*, *.png*, *.bmp*, *.xbm*, *.tif*, or *.tiff*. The "(?i)" at the beginning makes the whole match case-insensitive.

Rebuild the collection and preview it. Now the images appear under the Browse button. But all that is displayed are their titles, which have been derived from the filenames and are sometimes uninformative, and a dummy text document.

23. Change to the **Enrich** panel, open the folder *cover_images* and manually add dc.Title metadata, assigning to each of the ten documents the title of the corresponding album. Remember, you can double-click a file to view it.

24. To suppress the dummy document, change the VList format statement in the **Design** panel again by adding the two highlighted lines, and the close curly bracket:

    ```
    <td valign=top>
    {If}{[dc.Format] eq 'Audio',
      [srclink][srcicon][/srclink],
      {If}{[dc.Format] eq 'Images',
        [srclink][thumbicon][/srclink],
        [link][icon][/link]}}</td>
    <td>{If}{[numleafdocs],([numleafdocs])}</td>
    <td valign=top>[highlight]
    {Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
    ```

Now let's incorporate the MIDI files. Greenstone has no MIDI plugin (yet). But that doesn't mean you can't use MIDI files!

25. UnknownPlug is a useful generic plugin that knows nothing about any given format. It can be tailored to process particular document types based on their filename extension, and to set some basic metadata.

    - add UnknownPlug;
    - activate its *process_exp* field and set it to *\.mid$* to make it recognize files ending in .mid;
    - Set *file_format* to "MIDI" and *mime_type* to "audio/midi".

    In this collection, all MIDI files are contained in the file *beatles_midi.zip*. ZIPPlug (already in the list of default plugins) is used to unpack the files and pass them down the list of plugins until they reach UnknownPlug.

26. Rebuild the collection and preview it. Unfortunately the MIDI files don't appear as Audio under the Category browser. That's because they haven't been assigned dc.Format metadata.

    - Back in the **Enrich** panel, click on the file *beatles_midi.zip* and assign its dc.Format value to "Audio"—do this by clicking on "Audio" in the **All Previous Values** list. All files extracted from the Zip file inherit its settings.

27. Next we return to our Title browser and clean it up. The aim is to amalgamate variants of titles by stripping away extraneous text. For example, we would like to treat "ANTHOLOGY 1", "ANTHOLOGY 2" and "ANTHOLOGY 3" the same for grouping purposes. To achieve this:

    - Go to the Title AZCompactList under Browsing Classifiers on the **Design** panel;
    - Activate removesuffix and set it to:
      ```
      (?i)(\\s+\\d+)|(\\s+[[:punct:]].*)
      ```

    Rebuild the collection and preview the result. Observe how many more times similar titles have been amalgamated under the same bookshelf. Test your understanding of regular expressions by trying to rationalize the amalgamations. (Note: *[[:punct:]]* stands for any punctuation character.) The icons beside the Word and PDF documents are not the correct ones, but that will be fixed in the next format statement.

Part V – Expert customization

28. To put finishing touches to our collection, we add some decorative features.

- Using a file browser, locate the folder *workshop_files*→*heterogeneous*. Copy the *images* and *macros* folders located there into your collection's top level folder. (It's OK to overwrite the existing *images* folder: the image in it is included in the folder being copied.) The *images* folder includes some useful icons, and the *macros* folder defines some macro names that use these images. To see the macro definitions, take a look by using a text editor to open the file *extra.dm* in the macros folder.
- Re-Edit your VList format statement to be the following

```
<td valign=top>
 {If}{[numleafdocs],[link][icon][/link]}
 {If}{[dc.Format] eq 'Lyrics',[link]_iconlyrics_[/link]}
 {If}{[dc.Format] eq 'Discography',[link]_icondisc_[/link]}
 {If}{[dc.Format] eq 'Tablature',[link]_icontab_[/link]}
 {If}{[dc.Format] eq 'MARC',[link]_iconmarc_[/link]}
 {If}{[dc.Format] eq 'Images',[srclink][thumbicon][/srclink]}
 {If}{[dc.Format] eq 'Supplementary',[srclink][srcicon][/srclink]}
 {If}{[dc.Format] eq 'Audio',[srclink]{If}{[FileFormat] eq
'MIDI',_iconmidi_,_iconmp3_}[/srclink]}
</td>
<td>
{If}{[numleafdocs],([numleafdocs])}
</td>
<td valign=top>
[highlight]
{Or}{[dc.Title],[Title],Untitled}
[/highlight]
</td>
```

If you find it too painful to type all this into the panel that the GLI provides, you could enter it in a separate editor like Windows Notepad and cut and paste it in.

Save your collection and preview it as before. The collection now uses different icons for discography, lyrics, tablature, and MARC metadata. It even distinguishes between MP3 and MIDI audio file types. If you let the mouse hover over one of these images a "tool tip" appears explaining what file type the icon represents in the current interface language (note: *extra.dm* only defines English and French).

29. To finish, let's now build a larger version of the collection. To do this:

- Close the current collection.
- Start a new collection called *advbeat_large.*
- Base this new collection on *advbeatles*.
- Copy the content of *workshop_files*→*heterogeneous*→*advbeat_large*→*import* into this newly formed collection. Since there are considerably more files in this set of documents the copy will take longer.

- Build the collection and preview the result. (If you want the collection to have an icon, you will have to add it manually from the **Design** panel.)