Lab 6: Multimedia

6.1. Looking at a multimedia collection

1. Copy the entire folder

   sample_files → beatles → advbeat_large

   (with all its contents) into your Greenstone *collect* folder. If you have installed Greenstone in the usual place, for Greenstone version 2.81 and above, this is

   My Computer → Local Disk (C:) → Users → <Username> → collect

   where <Username> is the username under with Greenstone is installed. For versions below 2.81, this is

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

   Put *advbeat_large* in there.

2. On Windows, if the Greenstone Digital Library Local Library Server is already running, re-start it by clicking the CD icon on the task bar and then pressing *Restart Library*. If not, start it up by selecting *Greenstone Digital Library* from the *Start* menu. On Linux and Mac, just do a forced reload/refresh of the web browser.

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. Some documents have an audio icon: when you click these you hear the music (assuming your computer is set up with appropriate player software). Others have an image thumbnail: when you click these you see the images.

4. Look at the *Titles* browser. Each title has a bookshelf that may include several related items. For example, *Hey Jude* has a MIDI file, 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 *Titles* 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. Only 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 in **Building a multimedia collection** that this is possible but tricky.)

6. In the file browser, take a look at the files that makes up the collection, in the

   sample_files → beatles → advbeat_large → import

   folder. What a mess! There are over 450 files under seven 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 *images* contains various images in

JPEG format. For each type, drill down through the hierarchy and look at a sample document.

6.2. Building a multimedia collection

We will proceed to reconstruct from scratch the Beatles collection that you have just looked at. We develop the collection using a small subset of the material, purely to speed up the repeated rebuilding that is involved.

1. 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!)

2. Copy the files provided in

   sample_files → beatles → advbeat_small

   into your new collection. Do this by opening up *advbeat_small*, selecting the eight items within it (from *discography* to *beatles_midi.zip*), and dragging them across. Because some of these files are in MP3 and MARC formats you will be asked whether to include **MP3Plugin** and **MARCPlugin** in your collection. Click **<Add Plugin>**.

3. 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.

   (There are no MIDI files in the collection: these require more advanced customisation because there is no MIDI plugin. We will deal with them later.)

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

5. **Preview** the result.

Manually correcting metadata

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

7. 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 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.*

8. Change to the **Design** panel and select the **Browsing Classifiers** section. Double-click the **ex. 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
  ```

  and click <**OK**>.

Build the collection again, and **preview** it.

Extracted metadata is unreliable. But it is very cheap! On the other hand, manually assigned metadata is reliable, but expensive. The previous section of this exercise has shown how to aim for the best of both worlds by using extracted metadata but correcting it when it is wrong.

Browsing by media type

9. 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 *discography* and set its **dc.Format** metadata value to "Discography". Setting this value at the folder level means that all files within the folder inherit it.
   - Repeat the process. Assign "Lyrics" to the *html_lyrics* folder, "Images" to *images*, "MARC" to *marc*, "Audio" to *mp3*, "Tablature" to *tablature_txt*, and "Supplementary" to *wordpdf*.
   - Switch to the **Design** panel and select the **Browsing Classifiers** section.
   - Delete the **ex.Source** classifier (the second one).
   - Add an **AZCompactList** classifier. Select **dc.Format** as the **metadata** field and specify "browse" as the **buttonname**. Click the **sort** checkbox, and select **ex.Title** in the drop-down list: this will make the classifier display documents in alphabetical order of title.

   Build the collection again and **preview** it.

*Note how we assigned **dc.Format** metadata to all documents in the collection with a minimum of labour. We did this by capitalizing on the folder structure of the original information. Even though we complained earlier about how messy this folder structure is, you can still take advantage of it when assigning metadata.*

Suppressing dummy text

10. Alongside the Audio files there is an MP3 icon, which plays the audio when you click it, and also a text document that contains some dummy text. Image files also have dummy documents. These dummy documents aren't supposed to be seen, but to suppress them you have to fiddle with a format statement.

    - Change to the **Format** panel and select the **Format Features** section.
    - Ensure that **VList** is selected, and make the changes that are highlighted below. You need to insert five lines into the first line, and delete the second line. (Note, the changes are available in a text file, see below.) Change:

      ```
      <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}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

to this:

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

To make this easier for you we have prepared a plain text file that contains the new text. In WordPad open the following file:

sample_files → beatles → format_tweaks → audio_tweak.txt

(Be sure to use WordPad rather than Notepad, because Notepad does not display the line breaks correctly.) Place it in the copy buffer by highlighting the text in WordPad and selecting **Edit → Copy**. Now move back to the Librarian Interface, highlight all the text that makes up the current **VList** format statement, and use **Edit → Paste (ctrl-v)** to transform the old statement to the new one.

Preview the result. You may need to click the browser's <**Reload**> button to force it to re-load the page.

11. 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],
{If}{[dc.Format] eq 'Images',
[srclink][thumbicon][/srclink],
[link][icon][/link]}}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

Preview the result (you don't need to rebuild the collection.)

Using AZCompactList rather than AZList

12. 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** browser might be improved by grouping these together under a bookshelf icon. This is a job for an **AZCompactList**.

- Change to the **Design** panel and select the **Browsing Classifiers** section.
- Remove the **ex.Title** classifier (at the top)
- Add an **AZCompactList** classifier, and enter **dc.Title,ex.Title** as its metadata.
- Finish by pressing **<OK>**.
- Move the new classifier to the top of the list (**<Move Up>** button).

Build the collection again 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.

Making bookshelves show how many items they contain

13. Make the bookshelves show how many documents they contain by inserting a line in the **VList** format statement in the **Format Features** section of the **Format** panel. The added line is shown highlighted below. The complete format statement can be copied from *sample_files → beatles → format_tweaks → show_num_docs.txt*.

```
<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>
```

Preview the result (you don't need to build the collection.) Bookshelves in the titles and browse classifiers should show how many documents they contain.

Adding a Phind phrase browser

14. In the **Browsing Classifiers** section on the **Design** panel, add a **Phind** classifier. Leave the settings at their defaults: this generates a phrase browsing classifier that sources its phrases from *Title* and *text*.

Build the collection again and **preview** it. Select the new **Phrases** option from the navigation bar. Enter a single word in the text box, such as **band**. The phrase browser will present you with phrases found in the collection containing the search term. This can provide a useful way of browsing a very large collection. Note that even though it is called a phrase browser, only single terms can be used as the starting point for browsing.

Branding the collection with an image

15. To complete the collection, lets give it a new image for the top left corner of the page. Go to the **General** section of the **Format** panel. Use the browse button of **URL to 'about page' image:** to select the following image:

sample_files → beatles → advbeat_large → images → beatlesmm.png

Preview the collection, and make sure the new image appears.

Using UnknownPlugin

In this section we incorporate the MIDI files. Greenstone has no MIDI plugin (yet). But that doesn't mean you can't use MIDI files!

16. **UnknownPlugin** is a useful generic plugin. It knows nothing about any given format but can be tailored to process particular document types—like MIDI—based on their filename extension, and set basic metadata.

    In the **Document Plugins** section of the **Design** panel:

    - add **UnknownPlugin**;
    - activate its **process_extension** field and set it to "mid" to make it recognize files with extension *.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*. **ZIPPlugin** (already in the list of default plugins) is used to unpack the files and pass them down the list of plugins until they reach **UnknownPlugin**.

17. **Build** the collection and **preview** it. Unfortunately the MIDI files don't appear as Audio under the *browse* button. 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 **Existing values for dc. Format** list. All files extracted from the Zip file inherit its settings.

Cleaning up a title browser using regular expressions

*We now clean up the **Titles** browser.*

*To do this we must put the Librarian Interface into a different mode. The interface supports four levels of user: **Library Assistant**, who can add documents and metadata to collections, and create new ones whose structure mirrors that of existing collections; **Librarian**, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions); **Library Systems Specialist**, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl programs); and **Expert**, who can perform all functions.*

*So far you have mostly been operating in **Librarian** mode. We switch to **Library Systems Specialist** mode for the next exercise.*

18. To switch modes, 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.

19. Next we return to our **Titles** 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:]].*)
```

Build 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.

*The previous exercise was done in **Library Systems Specialist** mode because it requires the use of regular expressions, something librarians are not normally trained in.*

*One powerful use of regular expressions in the exercise was to clean up the **Titles** browser. Perhaps the best way of doing this would be to have proper title metadata. The metadata extracted from HTML files is messy and inconsistent, and this was reflected in the original Titles browser. Defining proper title metadata would be simple but rather laborious. Instead, we have opted to use regular expressions in the **AZCompactList** classifier to clean up the title metadata. This is difficult to understand, and a bit fiddly to do, but if you can cope with its idiosyncrasies it provides a quick way to clean up the extracted metadata and avoid having to enter a large amount of metadata.*

Using non-standard macro files

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

20. Close the collection in the Librarian Interface (**File → Close**).

21. Using your Windows file browser outside Greenstone, locate the folder

    sample_files → beatles → advbeat_large

22. Open up another file browser, and locate the small beatles collection in your Greenstone installation:

    Greenstone → collect → smallbea

    smallbea is the folder name generated by Greenstone for this collection. You can determine what the folder name is for a collection by looking at the title bar of the Librarian Interface: the folder name is displayed in brackets after the collection name.

23. Using the file browser, copy the *images* and *macros* folders from the *advbeat_large* folder into the *smallbea* 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, open the collection in the Librarian Interface (**File → Open...**) and view the **Collection Specific Macros** section in the **Format** panel.

Using different icons for different media types

24. Re-edit your **VList** format statement to be the following (in **Format Features** on the **Format** panel). You can copy this text from the file *sample_files → beatles → format_tweaks → multi_icons.txt*.

```
<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>
```

25. **Preview** your collection as before. Now different icons are used for discography, lyrics, tablature, and MARC metadata. Even MP3 and MIDI audio file types are distinguished. 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).

Changing the collection's background image

26. Go to the **Collection Specific Macros** section in the **Format** panel.

27. The content is fairly brief, specifying only what needs to be overridden from the default behaviour for this collection. Near the top you should see:

```
_collectionspecificstyle_ {
<style>
body.bgimage \{ background-image: url("_httpcimages_/beat_margin.
gif"); \}
\#page \{ margin-left: 120px; \}
</style>
}
```

    Replace the text **beat_margin.gif** with **tile.jpg**.

    This line relates to the background image used. The new image *tile.jpg* was in the *images* folder that was copied across previously.

28. **Preview** the collection's home page. The page background is now the new graphic.

    Other features can be altered by editing the macros—for example, the headers and footers used on each page, and the highlighting style used for search terms (specify a different colour, use

bold etc.).

Building a full-size version of the collection

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

    - Close the current collection (**File → Close**).
    - Start a new collection called *large beatles* (**File → New...**).
    - Base this new collection on *small beatles*.
    - Copy the content of *sample_files → beatles → 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 from the **Format** panel.)

Adding an image collage browser

30. Switch to the **Design** panel and select the **Browsing Classifiers** section. Pull down the **Select classifier to add** menu and select **Collage**. Click **<Add Classifier...>**. There is no need to customize the options, so click **<OK>** at the bottom of the resulting popup.

31. Now change to the **Create** panel and **build** and **preview** the collection.
