Lab 2: Designing collections

2.1. A large collection of HTML files—Tudor

You will need the files in the sample_files → tudor folder.

1. Invoke the Greenstone Librarian Interface (from the Windows *Start* menu) and start a new collection called **tudor** (use the **File** menu), based on the default **-- New Collection --**.

2. In the **Gather** panel, open the *tudor* folder in *sample_files*.

3. Drag *englishhistory.net* from the left-hand side to the right to include it in your **tudor** collection. (This material is from Marilee Hanson's Tudor England Collection at http://englishhistory.net/tudor.html, distributed with her permission.)

4. Switch to the **Create** panel and click **<Build Collection>**.

5. When building has finished, **preview** the collection.

Extracting more metadata from the HTML

6. The browsing facilities in this collection (*Titles* and *Filenames*) are based entirely on extracted metadata. Return to the **Enrich** panel in the Librarian Interface and examine the metadata that has been extracted for some of the files.

7. Many HTML documents contain metadata in `<meta>` tags in the `<head>` of the page. Open up the *englishhistory.net* → *tudor* → *monarchs* → *boleyn.html* file by navigating to it in the tree on the left hand side, and double clicking it. This will open it in a web browser. View the HTML source of the page (**View** → **Source** in Internet Explorer, **View** → **Page Source** in Mozilla). You will notice that this page has *page_topic*, *content* and *author* metadata.

8. By default, **HTMLPlug** only looks for Title metadata. Configure the plugin so that it looks for the other metadata too. Switch to the **Design** panel and select the **Document Plugins** section. Select the **plugin HTMLPlug** line and click **<Configure Plugin...>**. A popup window appears. Switch on the **metadata_fields** option, and set the value to

   ```
   Title,Author,Page_topic,Content
   ```

   Click **<OK>**.

9. Switch to the **Create** panel and **rebuild** the collection. Go back to the **Enrich** panel and look at the extracted metadata for some of the HTML files in *englishhistory.net* → *tudor* → *monarchs*. The new metadata should now be visible.

Blocking the stray images

You've probably noticed that the collection contains a few stray image files, as well as the HTML *documents. This is a mistake. The issue is that many of the HTML documents include images, and although Greenstone attempts to determine which images belong to HTML pages and only considers other images for inclusion in the collection, in this case it hasn't been completely successful. (This is because the web site from which these files were downloaded occasionally departs from the usual convention of hierarchical structuring.)*

10. Switch back to the **Document Plugins** section of the **Design** panel. Beside **plugin HTMLPlug** you will see **-smart_block**. This is the option that attempts to identify images in the HTML pages and block them from inclusion—in this case, it's not smart enough! Configure **plugin HTMLPlug** again, scroll down the page to locate the **smart_block** option, and switch it off.

11. **Rebuild** and **preview** the collection. The collection is exactly as before except that these stray images are suppressed. What is happening is that plug-ins operate as a pipeline: files are passed to each one in turn until one is found that can process it. By default (i.e. without **smart_block**) the HTML plug-in blocks *all* images, which is appropriate for this collection.

Looking at different views of the files in the Gather and Enrich panels

12. Switch to the **Gather** panel and in the right-hand side open *englishhistory.net* → *tudor*.

13. Change the **Show Files** menu for the right-hand side from **All Files** to **HTM & HTML**. Notice the files displayed above are filtered accordingly, to show only files of this type.

14. Change the **Show Files** menu to **Images**. Again, the files shown above alter.

15. Now return the **Show Files** setting back to **All Files**, otherwise you may get confused later. Remember, if the **Gather** or **Enrich** panels do not seem to be showing all your files, this could be the problem.

2.2. Enhanced collection of HTML files—Tudor

We return to the Tudor collection and add metadata that expresses a subject hierarchy. Then we build a classifier that exploits it by allowing readers to browse the documents about Monarchs, Relatives, Citizens, and Others separately.

Adding hierarchically-structured metadata and a Hierarchy classifier

1. Open up your **tudor** collection (the original version, not the **webtudor** version), switch to the **Enrich** panel and select the *citizens* folder (a subfolder of *englishhistory.net* → *tudor*). Set its **dc. Subject and Keywords** metadata to **Tudor period|Citizens**. The vertical bar ("|") is a hierarchy marker. Selecting a *folder* and adding metadata has the effect of setting this metadata value for all files contained in this folder, its subfolders, and so on. A popup alerts you to this fact. Click **<OK>** to close the popup.

2. Repeat for the *monarchs* and *relative* folders, setting their **dc.Subject and Keywords** metadata to **Tudor period|Monarchs** and **Tudor period|Relatives** respectively. Note that the hierarchy appears in the **Existing values for dc.Subject and Keywords** area.

   If you don't want to see the popup each time you add folder level metadata, tick the **Do not show this warning again** checkbox; it won't be displayed again.

3. Finally, select all remaining files—the ones that are not in the *citizens*, *monarchs*, or *relative* folders—by selecting the first and shift-clicking the last. Set their **dc.Subject and Keywords** metadata to **Tudor period|Others**: this is done in a single operation (there is a short delay before it completes).

   When multiple files are selected in the left hand collection tree, all metadata values for all files are shown on the right hand side. Items that are common to all files are displayed in black—e.g. **dc.Subject and Keywords**—while others that pertain to only one or some of the files are displayed in grey—e.g. any extracted metadata.

   Metadata inherited from a parent folder is indicated by a folder icon to the left of the metadata name. Select one of the files in the *relative* folder to see this.

4. Switch to the **Design** panel and select **Browsing Classifiers** from the left-hand list. Set the menu item for **Select classifier to add** to **Hierarchy**; then click **<Add Classifier...>**.

5. A window pops up to control the classifier's options. Change the **metadata** to **dc.Subject and Keywords** and then click **<OK>**.

6. 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.

7. Now switch to the **Create** panel, **build** the collection, and **preview** it. Choose the new **Subjects** link that appears in the navigation bar, and click the bookshelves to navigate around the four-entry hierarchy that you have created.

Adding a hierarchical phrase browser (PHIND)

Next we'll add an interactive hierarchical phrase browsing classifier to this collection.

8. Switch to the **Design** panel and choose the **Browsing Classifiers** item from the left-hand list.

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

10. **Build** the collection again, **preview** it, and try out the new **Phrases** option in the navigation bar. An interesting PHIND search term for this collection is "king". Note that even though it is called a phrase browser, only single terms can be used as the starting point for browsing.

Partitioning the full-text index based on metadata values

*Next we partition the full-text index into four separate pieces. To do this we first define four subcollections obtained by "filtering" the documents according to a criterion based on their **dc.Subject and Keywords** metadata. Then an index is assigned to each subcollection. This will enable users to restrict a search to a subset of the documents.*

11. Switch to the **Design** panel, and click **Partition Indexes**. This feature is disabled because you are operating in **Librarian** mode (this is indicated in the title bar at the top of the window).

12. Switch to **Library Systems Specialist** mode by going to **Preferences...** (on the **File** menu) and clicking **<Mode>**. Read about the other modes too.

13. Return to the **Partition Indexes** section of the **Design** panel. Ensure that the **Define Filters** tab is selected (the default). Define a subcollection filter with name **monarchs** that matches against **dc. Subject and Keywords**, and type **Monarchs** as the regular expression to match with. Click **<Add Filter>**. This filter includes any file whose **dc.Subject and Keywords** metadata contains the word *Monarchs*.

14. Define another filter, **relatives**, which matches **dc.Subject and Keywords** against the word **Relatives**. Define a third and fourth, **citizens** and **others**, which matches it against the words **Citizens** and **Others** respectively.

15. Having defined the subcollection filters, we partition the index into corresponding parts. Click the **Assign Partitions** tab. Select the citizens subcollection and click **<Add Partition>**. Next select monarchs, and click **<Add Partition>**. Repeat for the other two subcollections, so that you end up with four partitions, one based on each subcollection filter.

    The order they appear in the **Assigned Subcollection Partitions** list is the order they will appear in the drop down menu on the search page. You can change the order by using the **<Move Up>** and **<Move Down>** buttons.

16. **Build** and **preview** the collection.

17. The search page includes a pulldown menu that allows you to select one of these partitions for searching. For example, try searching the *relatives* partition for *mary* and then search the *monarchs* partition for the same thing.

18. To allow users to search the collection as a whole as well as each subcollection individually,

return to the **Partition Indexes** section of the **Design** panel and select the **Assign Partitions** tab. Select all four subcollections by checking their boxes and click **<Add Partition>**.

19. To ensure that the combined index appears first in the list on the reader's web page, use the **<Move Up>** button to get it to the top of the list here in the **Design** panel. Then **build** and **preview** the collection.

20. Search for a common term (like *the*) in all five index partitions, and check that the numbers of words (not documents) add up.

21. The text in the drop down box on the search page is based on the filters each partition was built on. To change the text that is displayed, go to the **Search** section of the **Format** panel. The single filter partitions have sensible default text, but the combined partition does not. Set the **Display text** for the combined partition to "all". **Preview** the collection.

22. In the Librarian Interface, return to **Librarian** mode, using **Preferences...** (on the **File** menu).

Controlling the building process

Finally we look at how the building process can be controlled. Developing a new collection usually involves numerous cycles of building, previewing, adjusting some enrich and design features, and so on. While prototyping, it is best to temporarily reduce the number of documents in the collection. This can be accomplished through the ***maxdocs*** *parameter to the building process.*

23. Switch to the **Create** panel and view the options that are displayed in the top portion of the screen. Select **maxdocs** and set its numeric counter to **3**. Now **build**.

24. Preview the newly rebuilt collection's **Titles** page. Previously this listed more than a dozen pages per letter of the alphabet, but now there are just three—the first three files encountered by the building process.

25. Go back to the **Create** panel and turn off the **maxdocs** option. **Rebuild** the collection so that all the documents are included.