

Lab 2 (2 hours): Librarian Interface: Building collections and adding metadata

The Librarian Interface is the tool used to create, develop and maintain collections.

Part I – Building your first collection with the Librarian Interface

1. Start the Greenstone Librarian Interface:

Start → Programs → Greenstone Digital Library Software → Greenstone Librarian Interface

[After a short pause the main Greenstone Librarian Interface (GLI) appears.]

2. Start a new collection within GLI:

File → New

3. You will create a collection based on a few HTML web pages that describe some Hobbits in *Lord of the Rings*.

A window pops up. Fill it out with appropriate values—for example,

Collection Title: About Hobbits
Description of Content: A collection about hobbits.

Leave the setting for **Base this collection on:** at its default **New Collection**, and click **<OK>**.

4. Another window pops up, from which you select the metadata set (or sets) to use. **Dublin Core Metadata Element Set Version 1.1** is selected by default; click **<OK>**.

[A progress bar appears while the collection is being created.]

5. Next you must gather together the files that will constitute the collection. A suitable set has been prepared ahead of time in *workshop_files* in the folder *html_small*. Using the left-hand side of the Librarian Interface's **Gather** panel, interactively navigate to this directory:

Local Filespace → D → workshop_files

[This assumes that 'D' is the letter used for the drive where the workshop files are stored: this may vary depending on how the computer has been configured.]

6. Now drag the *html_small* folder from the left-hand side and drop it on the right. The progress bar at the bottom shows some activity. Gradually, duplicates of all the files will appear in the right-hand panel.

7. You can inspect the files that have been copied by double-clicking on the folder in the right-hand side.
8. Since this is our first collection, we won't complicate matters by manually assigning metadata or altering the collection's design. Instead we rely on default behaviour. So pass directly to the **Create** panel by clicking the **Create** tab.
9. To start building the collection, click the **<Build Collection>** button.
10. Once the collection has built successfully, a window pops up to confirm this. Click **<OK>**.
11. Click the **Preview Collection** button to look at the end result. This loads the relevant page into your web browser (starting it up if necessary). Look around the collection and learn about Hobbits!
12. Back in the Librarian Interface, click the **Enrich** tab to view the metadata associated with the documents in the collection.

Presently there is no manually assigned metadata, but the act of building the collection has generated some extracted metadata. Double click the *html_small* folder to expand its content. Then single-click *bilbo.html* to display all its metadata in the right-hand side of the panel. The initial fields, starting 'dc.', are empty. These are Dublin Core metadata fields (we asked you to include this metadata set when the collection was initially formed) for manually entered data.

Use the scroll bar on the extreme right to view the bottom part of the list. There you will see fields starting 'ex.' that express the extracted metadata: for example ex.Title, based on the text within the HTML Title tags, and ex.Language, the document's language (represented using the ISO standard 2-letter mnemonic) which is set by an algorithm that Greenstone uses to analyse the document's text.

13. Close the collection by clicking **File→Close**. This automatically saves the collection to disk.

Part II – Building a larger collection and GLI features

We now build a larger collection of HTML files based on the workshop's html_large folder:

1. Start a new collection called **Tudor** which will contain documents about this period of English history. Fill out the pop-up dialog with appropriate values and choose Dublin Core as the metadata set.

[The act of starting a new collection or opening an existing one will automatically close the current collection, saving it to disk, so the previous step of closing the collection was not strictly necessary.]

2. In the **Gather** panel, open the *html_large* folder in *workshop_files*.

3. Drag *englishhistory.net* from the left-hand side to the right to include it in your **Tudor** collection. Note: this is quite a large set of documents. If you are pressed for time, navigate to one of the folders inside *englishhistory.net*, e.g. *Monarchs*, and use it instead.
4. Switch to the **Create** panel and click **<Build Collection>**. This collection has more files and takes longer to build than the hobbit one. When it has finished building, preview the collection.

Now look at different views of the files in the Gather and Enrich panels:

5. Switch to the **Gather** panel and open in the right-hand side *englishhistory.net* → *tudor*.
6. 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.
7. Change the **Show Files** menu to **Images**. Again, the files shown above alter.

Setting up a shortcut in the Librarian Interface

8. As we build several collections that use files from the *workshop_files* folder, it can be handy to set up a shortcut to this folder. In the **Gather** panel navigate to the *workshop_files* folder in the **Workspace** tree. Select this folder then right-click it, and choose **Create Shortcut** from the menu. In the **Name** field, enter the name you want the shortcut to have, or accept the default, *workshop_files*. Click **<OK>**. Close all the folders in the file tree and you will see the shortcut to the workshop files.

Part III – Adding metadata to a collection

Similarly styled collections can be built from Word and PDF documents. We now experiment with a new collection based on these two formats.

1. Start a new collection called **reports**, fill out appropriate fields for it, and choose Dublin Core as the metadata set.
2. Copy all the files from *workshop_files* → *Word_and_PDF* into the new collection.
3. Switch to the **Create** panel, and **build** and **preview** the collection.
4. Again, this collection contains no manually assigned metadata. All the information that appears—title and filename—is extracted automatically from the documents themselves. Because of this, the quality of some of the Title metadata is suspect.
5. Back in the GLI, switch to the **Enrich** panel, and look at the extracted metadata for the documents. The empty Dublin Core metadata will be shown first: scroll down to see the extracted metadata, which begins with “ex.”

6. Check whether the **Title** metadata is correct for each document by opening it. You can open a document from the GLI by double clicking on it.
7. The extracted Title metadata for some documents is incorrect. For example, the Titles for pdf01.pdf and word03.doc (the same document in different formats) have missed out the second line. The Title for pdf03.pdf has the wrong text altogether.
8. For the documents where the extracted Title (**ex.Title**) is incorrect, add a new correct version as **dc.Title**. To add **dc.Title** metadata, select the appropriate document in the left hand panel. Scroll up or down in the metadata table until you can see **dc.Title**. Click in the value box, type in the metadata and press **Enter**.
9. You will notice that as you add more values, they appear in the **Existing values for dc.Title** box below the metadata table. If you are adding the same metadata value to more than one document, you can select it from this list. For example, pdf01.pdf and word03.doc share the same Title.
10. Once you have corrected all the wrong Title metadata, go to the **Create** pane and **build** the collection. **Preview** the collection and look at the *titles a-z* list. You will see that it shows your new Titles.
11. Many of the documents don't look very nice in Greenstone. One of them, *pdf05-notext.pdf*, could not be processed using the default configuration. Another, *pdf06-weirdchars.pdf*, was processed but looks very strange. We will revisit this collection in **Lab 6**, and see how to configure the plugins to handle these files better.

Part IV – Your custom collection

In each lab, we will have a short exercise where you get to apply what you have just learnt to a collection of your own material.

Create your own collection

1. Using the electronic documents which you have brought to the workshop, build a simple collection. Add your documents into the collection, then **build** it and **preview**. You may like to add some metadata at this stage, which you will use in **Lab 3**.