Lab 1 Building collections and adding metadata

1.1. Building a small collection of HTML files

You will need some HTML files, such as those in the simple_html folder in sample_files.

Running the Greenstone Librarian Interface

1. Start the Greenstone Librarian Interface:

   Start → All Programs → Greenstone-2.81 → Librarian Interface (GLI)

   After a short pause a startup screen appears, and then after a slightly longer pause the main Greenstone Librarian Interface appears. (A command prompt is also opened in the background.)

Starting a new collection

2. Start a new collection within the Librarian Interface:

   File → New...

3. You will create a collection based on a few HTML web pages from the Tudor collection.

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

   Collection title: Small HTML Collection
   Description of content: A small collection of HTML pages.

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

4. Next you must gather together the files that will constitute the collection. A suitable set has been prepared ahead of time in *sample_files → simple_html*. Using the left-hand side of the Librarian Interface's **Gather** panel, interactively navigate to the *sample_files* folder.

Adding documents to the collection

5. Now drag the *simple_html* 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 collection panel.

   You can inspect the files that have been copied by double-clicking on the folder in the right-hand side.

6. 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 its tab.

Building the collection

7. To start building the collection, click the **<Build Collection>** button.

8. Once the collection has built successfully, a window pops up to confirm this. Click **<OK>**.

9. 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).

Viewing the extracted metadata

10. Back in the Librarian Interface, click the **Enrich** tab to view the metadata associated with the documents in the collection.

11. Presently there is no manually assigned metadata, but the act of building the collection has extracted metadata from the documents. Double click the *simple_html* folder to expand its content. Then single-click *aragon.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 for manually entered data.

12. 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 Greenstone determines by analyzing the document's text.

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

Viewing the internal links and external links

14. Hyperlinks in a Greenstone collection work like this. If the link is to a document that is also in the collection, clicking it takes you to that document in the collection. If the link is to a document that is *not* in the collection, clicking it takes you to that document on the web.

    Open *boleyn.html* and look for the link to *Katharine of Aragon* (in the 5th paragraph of the *Biography* section). This links to a document inside the collection--*aragon.html*. View this document by clicking the link. For an external link, click *letters written by Katharine* (in the *Primary Sources* section). This takes you out on to the web. If you want a warning message to be displayed first, you can open *Greenstone → etc → main.cfg* file and uncomment the line `cgiarg shortname=el argdefault=prompt`.

Setting up a shortcut in the Librarian interface

15. To set up a shortcut to the source files, in the **Gather** panel navigate to the folder in your local file space that contains the files you want to use—in our case, the *sample_files* folder. Select this folder and 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 *sample_files*. Click **<OK>**. Close all the folders in the file tree in the left-hand pane, and you will see the shortcut to your source files.

1.2. A simple image collection

1. In the Librarian Interface, start a new collection (**File → New...**) called **backdrop**. Fill out the fields with appropriate information. For **Base this collection on:**, select the item **Simple image collection (image-e)** from the pull-down menu.

   When you base a collection on an existing one, it inherits all the settings of the old one, including which metadata sets (if any) the collection uses.

2. Copy the images provided in *sample_files → images* into your newly-formed collection.

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

4. **Preview** the result.

5. Click on **Browse** in the navigation bar to view a list of the photos ordered by filename and presented as a thumbnail accompanied by some basic data about the image. The structure of this collection is the same as **Simple image collection (image-e)**, but the content is different.

6. Back in the Librarian Interface, change to the **Enrich** panel and view the extracted metadata for *Bear.jpg*.

Adding Title and Description metadata

7. We work with just the first three files (*Bear.jpg*, *Cat.jpg* and *Cheetah.jpg*) to get a flavour of what is possible. First, set each file's **dc.Title** field to be the same as its filename but without the filename extension:

   Click on *Bear.jpg* so its metadata fields are available, then click on its **dc.Title** field on the right-hand side. Type in **Bear**.

   Repeat the process for *Cat.jpg* and *Cheetah.jpg*.

8. Add a description for each image as **dc.Description** metadata.

   What description should you enter? To remind yourself of a file's content, the Librarian Interface lets you open files by double-clicking them. It launches the appropriate application based on the filename extension, Word for .doc files, Acrobat for .pdf files and so on.

   Double-click *Bear.jpg*: on Windows, the image will normally be displayed by Microsoft's Photo Editor (although this depends on how your computer has been set up).

   Back in the **Enrich** pane, make sure that *Bear.jpg* is selected in the collection tree on the left hand side. Enter the text **Bear in the Rocky Mountains** as the value for the **dc.Description** field.

   Repeat this process for *Cat.jpg* and *Cheetah.jpg*, adding a suitable description for each.

9. Go to the **Create** panel and click **<Build Collection>**. Once it has finished building, **preview** the

collection. You will not notice anything new. That's because we haven't changed the design of the collection to take advantage of the new metadata.

Change Format Features to display new metadata

10. Now we customize the collection's appearance. Go to the **Format** panel and select **Format Features** from the left-hand list. Leave the feature selection controls at their default values, so that **All Features** is selected for **Choose Feature**, and **VList** is selected as the **Affected Component**. In the **HTML Format String**, edit the text as follows:

    - Change `_ImageName_:` to `Title:`
    - Change `[Image]` to `[dc.Title]`
    - After `[dc.Title]<br>` add `Description: [dc.Description]<br>`

    Metadata names are case-sensitive in Greenstone: it is important that you capitalize "Title" and "Description" (and don't capitalize "dc").

11. The new format statement is displayed in the list of assigned format statements. The first substitution alters the fragment of text that appears to the right of the thumbnail image, the second alters the item of metadata that follows it. The addition displays the description after the Title.

12. Preview the collection by clicking the **<Preview Collection>** button. When you click on **Browse** in the navigation bar the presentation has changed to "Title: Bear" and so on. Each image's description should appear beside the thumbnail, following the title.

After the first three items, the Title and Description become blank because we have only assigned Dublin Core metadata to these first three. To get a full listing, enter all the metadata.

*Changes in the **Format** panel take place immediately and you can see the result straightaway by clicking the **Preview Collection**. If you modify anything in the **Gather**, **Enrich** or **Design** panels, you will need to rebuild the collection.*

Changing the size of image thumbnails

13. Lets change the size of the thumbnail image and make it smaller. Thumbnail images are created by the **ImagePlugin** plug-in, so we need to access its configuration settings. To do this, switch to the **Design** panel and select **Document Plugins** from the list on the left. Double-click **ImagePlugin** to pop up a window that shows its settings. (Alternatively, select **ImagePlugin** with a single click and then click **<Configure Plugin...>** further down the screen). Currently all options are off, so standard defaults are used. Select **thumbnailsize**, set it to **50**, and click **<OK>**.

14. **Build** and **preview** the collection.

15. Once you have seen the result of the change, return to the **Design** panel, select the configuration options for **ImagePlugin**, and switch the **thumbnailsize** option off so that the thumbnail reverts to its normal size when the collection is re-built.

Adding a browsing classifier based on Description metadata

16. Now we'll add a new browsing option based on the descriptions. In the **Design** panel, select

Browsing Classifiers from the left-hand list. Set the menu item for **Select classifier to add** to **AZList**; then click <**Add Classifier...**>.

17. A window pops up to control the classifier's options. Set the **metadata** option to **dc.Description** and click <**OK**>.

18. **Build** the collection, and **preview** it. Choose the new **Descriptions** link that appears in the navigation bar.

*Only three items are shown, because only items with the relevant metadata (**dc.Description** in this case) appear in the list. The original browse list includes all photos in the collection because it is based on **ex. Image**, extracted metadata that reflects an image's filename, which is set for all images in the collection.*

Creating a searchable index based on Description metadata

19. Now we'll add an index so that the collection can be searched by descriptions. Switch to the **Design** panel and select **Search Indexes** from the left-hand list. Click the <**New Index**> button. Select **dc.Description** from the list of metadata to include in the index, leave **Indexing level:** at its default, "document", and click <**Add Index**>.

20. Switch to the **Create** panel, **build** the collection, then **preview** it. There is now a **Search** button in the navigation bar. As an example, search for the term "bear" in the *document:dc.Description* index (which is the only index at this point).

21. To change the text that is displayed for the index (document:dc.Description), go to the **Format** panel back in the Librarian Interface. Select **Search** from the left-hand list. This panel allows you to change the text that is displayed on the search form. Change the **Display text** for the **document:dc.Description** index to "descriptions" (or other suitable text). Go back to the browser and reload the search page. Your new text will appear in the search form.

1.3. A collection of Word and PDF files

You will need some source files like those in the sample_files → Word_and_PDF folder.

1. Start a new collection called **reports** (**File → New...**) and base it on **-- New Collection --**.

2. Copy all the .doc, .rtf, .pdf and .ps files from *sample_files → Word_and_PDF → Documents* into the collection. There are 9 files in all: you can select multiple files by clicking on the first one and shift-clicking on the last one, and drag them all across together. (This is the normal technique of multiple selection.)

3. Switch to the **Create** panel, and **build** and **preview** the collection.

Viewing the extracted metadata

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 Librarian Interface, click the **Enrich** tab to view the automatically extracted metadata. You will need to scroll down to see the extracted metadata, which begins with "ex.".

6. Check whether the **ex.Title** metadata is correct for some of the documents by opening them. You can open a document from the Librarian Interface 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.

Manually adding metadata to documents in a collection

8. In the **Enrich** panel, manually add Dublin Core **dc.Title** metadata to those documents which have incorrect **ex.Title** metadata. Select *word03.doc* and double-click to open it. Copy the title of this document ("Greenstone: A comprehensive open-source digital library software system") and return to the Librarian Interface. Scroll up or down in the metadata table until you can see **dc.Title**. Click in the value box and paste in the metadata.

9. Now add **dc.Creator** information for the same document. You can add more than one value for the same field: when you press **Enter** in a metadata value field, a new empty field of the same type will be generated. Add each author separately as **dc.Creator** metadata.

10. Close the document (in Microsoft Word) when you have finished copying metadata from it. External programs opened when viewing documents must be closed before building the collection, otherwise errors can occur.

11. Next add **dc.Title** and **dc.Creator** metadata for a few of the other documents.

12. You will notice as you add more values, they appear in the **Existing values for ...** 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; and many documents have common authors.

*If you build and preview your collection at this point, you will see that the **Titles** list now shows your new Titles. However, the **dc.Creator** metadata is not displayed. You need to alter the collection design to use this metadata.*

Document Plugins

13. In the Librarian Interface, look at the **Document Plugins** section of the **Design** panel, by clicking on this in the list to the left. Here you can add, configure or remove plugins to be used in the collection. There is no need to remove any plugins, but it will speed up processing a little. In this case we have only Word, PDF, RTF, and PostScript documents, and can remove the **ZIPPlugin**, **TextPlugin**, **HTMLPlugin**, **EmailPlugin**, **ImagePlugin**, **PowerPointPlugin**, **ExcelPlugin**, **ISISPlug** and **NULPlugin** plugins. To delete a plugin, select it and click **<Remove Plugin>**. **GreenstoneXMLPlugin** is required for any type of source collection and should not be removed.

Search indexes

14. The next step in the **Design** panel is **Search Indexes**. These specify what parts of the collection are searchable (e.g. searching by title and author). Delete the **ex.Source** index, which is not particularly useful, by selecting it and clicking **<Remove Index>**.

15. Modify the **ex.Title** index to include **dc.Title** by selecting the index in the **Assigned Indexes** box and clicking **<Edit Index>**. Select **dc.Title** from the list of metadata, and click **<Replace Index>**. Searching this index will search both **dc.Title** and **ex.Title** metadata. If you want to restrict searching to just the manually added **dc.Title** metadata, edit the index again and deselect **ex.Title** from the list of metadata.

16. You can add indexes based on any metadata. Add a new index based on **dc.Creator** by clicking **<New Index>**. Select **dc.Creator** in the list of metadata, and click **<Add Index>**.

Browsing classifiers

17. The **Browsing Classifiers** section adds "classifiers," which provide the collection with browsing functions. Go to this section and observe that Greenstone has provided two classifiers, *AZLists* based on **ex.Title** and **ex.Source** metadata. These correspond to the *Titles* and *Filenames* buttons on the collection's access bar.

    Remove the **ex.Source** classifier by selecting it and clicking **<Remove Classifier>**.

18. Modify the **ex.Title** classifier to use **dc.Title** instead. Select the classifier and click **<Configure Classifier...>**. In the **metadata** box, select **dc.Title** instead of **ex.Title**. Click **<OK>**.

19. Now add an **AZCompactList** classifier for **dc.Creator**. Select **AZCompactList** from the **Select classifier to add** drop-down list and click **<Add Classifier...>**. A popup window **Configuring Arguments** appears. Select **dc.Creator** from the **metadata** drop-down list and click **<OK>**.

    AZCompactList is like **AZList**, except that values that appear multiple times in the hierarchy are automatically grouped together and a new node, shown as a bookshelf icon, is formed.

20. Switch to the **Create** panel, and **build** and **preview** the collection.

21. Check that all the facilities work properly. There should be three full-text indexes, called *text*, *dc.Title* (or *dc.Title,Title* if you didn't deselect **ex.Title** in the search indexes step above), and *dc.Creator*. The *Titles* list should display all the documents to which you have assigned **dc.Title** metadata (and only those documents). The *Creators* list should show one bookshelf for each author you have assigned as **dc.Creator**, and clicking on that bookshelf should take you to all the documents they authored.

Renaming the search indexes

22. The default display text for the indexes in the drop-down list on the search page contains the content of the index. Now we will change this display text to make it nicer. Go to the **Format** panel by clicking its tab. This panel is split into several sections, each controlling some aspect of collection presentation.

23. Select **Search** in the left hand list. This section allows you to modify what text is displayed for the drop-down lists in the search form (indexes, subcollections, levels etc). Set the **Display text** for the **dc.Title** (or **dc.Title,Title** if you didn't deselect **ex.Title** in the search index) index to be "titles", and that for the **dc.Creator** index to be "creators". Preview the collection by clicking the **Preview Collection**. The search form should display the new text.

Classifying on multiple metadata

24. The new *Titles* list shows only those documents which have been assigned **dc.Title** metadata. For many documents, extracted Titles may be fine, and it is impractical to add the same metadata again as **dc.Title**. Fortunately there is a way we can use both metadata types in one classifier: specify a list of metadata names in the classifier.

25. In the **Browsing Classifiers** section of the **Design** panel, select the **AZList** for **dc.Title** in the **Assigned Classifiers** box and click **<Configure Classifier...>**. Note you can achieve the same result by double clicking on the classifier.

26. In the **metadata** field, type ",ex.Title" after the "dc.Title"—i.e. make it read

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

27. If you have already done the **Enhanced Word document handling** exercise, some of the documents will have extracted ex.Creator metadata, and some will have dc.Creator. To use both of these in the Creators classifier, make a similar change to the **AZCompactList**: make the **metadata** field read `dc.Creator,ex.Creator`.

    Build the collection again and **preview** it. Now all of the documents should appear in the *Titles* list (and extracted Creators should appear in the *Creators* list).

We will play around with the format statements and customize the outlook of this collection in the ***Formatting the Word and PDF collection*** *exercise.*

1.4. Exporting a collection to CD-ROM/DVD

*To publish a collection on CD-ROM or DVD, Greenstone's Export to CD-ROM export module must be installed. This is included with CD-ROM distributions, and all distributions 2.70w and later. It must be installed separately for non-CD-ROM versions of Greenstone, version 2.70 and earlier (see **Installing Greenstone**).*

1. Launch the Greenstone Librarian Interface if it is not already running.

2. Choose **File → Write CD/DVD image...**. In the resulting popup window, select the collection or collections that you wish to export by ticking their check boxes. You can optionally enter a name for the CD-ROM: this is the name that will appear in the menu when the CD-ROM is run. If a name is not entered, the default **Greenstone Collections** will be used. You can also specify whether the resulting CD-ROM will install files onto the host machine when used or not. Click **<Write CD/DVD image>** to start the export process.

   The necessary files for export are written to:

   Greenstone → tmp → exported_xxx

   where xxx will be similar to the name you have entered. If you didn't specify a name for the CD-ROM, then the folder name will be *exported_collections*.

   You need to use your own computer's software to write these on to CD-ROM. On *Windows XP* this ability is built into the operating system: assuming you have a CD-ROM or DVD writer insert a blank disk into the drive and drag the *contents* of *exported_xxx* into the folder that represents the disk.

   The result will be a self-installing Windows Greenstone CD-ROM or DVD, which starts the installation process as soon as it is placed in the drive.
