LAB 2:
Greenstone: Adding metadata - and using it

2.1 A collection of Word and PDF files – Part B

In the Librarian Interface, open up the reports collection that you have created in exercise 1.4 Remember that the extracted Title metadata for some documents was incorrect.

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, **PowerPointPlugin**, **ExcelPlugin**, **ImagePlugin**, **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. By default the titles index (**dc.Title,ex.Title**) includes **dc.Title** and **ex.Title**. 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 this index 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 *List* classifiers, based on **dc.Title;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. 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>**.

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

20. Check that all the facilities work properly. There should be three full-text indexes, called *Text*, *Titles*, and *Creators*. The *Titles* list should display all the document Titles. 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.

*The Titles list shows all documents which have been assigned **dc.Title** metadata, or have automatically extracted **ex.Title**. For many documents, extracted Titles may be fine, and it is impractical to add the same metadata again as **dc.Title**. Specifying a list of metadata names in the classifier allows us to use both.*

21. 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 the **metadata** field read `dc.Creator,ex.Creator`.

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

2.2 A simple image collection

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

1. Copy the entire folder

   sample_files → image-e

   (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 → Greenstone → collect

   Put ***image-e*** in there.

2. 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** from the pull-down menu.

   *This will only be available if the documented example collections are installed. If you don't have this collection, select **-- New Collection --**. You can still build an image collection, but some of the tutorial will not match exactly. 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.*

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

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

5. **Preview** the result.

6. 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**, but the content is different.

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

Adding Title and Description metadata

8. We work with just the first three files (*Bear.jpg*, *Cat.jpg* and *Cheetah.jpg*) to get a flavour of what is possible. First, we need to add the Dublin Core metadata set which is not used in the **Simple image collection** collection. Click the **<Manage Metadata Sets...>** button beneath the Collection file tree. A new window pops up showing the metadata sets used by current collection. Click the **<Add...>** button to bring up another window showing the available metadata sets. Select the "Dublin Core Metadata Element Set" from the list and click **<Add>**. Click **<Close>** to return to the **Enrich** panel.

   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*.

9. 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.

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

11. 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

12. 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").

13. 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.

14. 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

15. 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 most options are off, so standard defaults are used. Select **thumbnailsize**, set it to **50**, and click **<OK>**.

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

17. 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

18. 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 **List**; then click **<Add Classifier...>**.

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

20. **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

21. 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>**.

22. 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 *Descriptions* index (which is the only index at this point).

23. To change the text that is displayed for the index (*Descriptions*), 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 "dc.Description" index to "image descriptions" (or other suitable text). Go back to the browser and reload the search page. Your new text will appear in the search form.

Note that if you use text instead of macros in the search metadata display text, you will need to do any translations yourself.

2.3 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, 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**.

12. For versions before 2.82, this feature is disabled because you are operating in **Librarian** mode (this is indicated in the title bar at the top of the window). 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 either checking their boxes or press the **Select All** button, 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. For versions before 2.82, now return to **Librarian** mode in the Librarian Interface, 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.

2.4 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.

2.5 Section tagging for HTML documents

1. In a browser, take a look at the Greenstone demo collection. Browse to one of the documents. This collection is based on HTML files, but they appear structured in the collection. This is because these HTML files were tagged by hand into sections.

2. Using a text editor (e.g. WordPad) open up one of the HTML files from the demo collection: *Greenstone → collect → demo → import → fb33fe →fb33fe.htm*. You will see some HTML comments which contain section information for Greenstone. They look like:

```
<!--
<Section>
  <Description>
    <Metadata name="Title">Farming snails 1: Learning about snails;
    Building a pen; Food and shelter plants</Metadata>
  </Description>
-->

<!--
</Section>
<Section>
  <Description>
    <Metadata name="Title">Dew and rain</Metadata>
  </Description>
-->
```

   When Greenstone encounters a `<Section>` tag in one of these comments, it will start a new subsection of the document. This will be closed when a `</Section>` tag is encountered. Metadata can also be added for each section—in this case, **Title** metadata has been added for each section. In the browser, find the **Farming snails 1** document in the demo collection (through the *Titles* browser). Look at its table of contents and compare it to the `<Section>` tags in the HTML document.

3. Add a new Section into this document. For example, lets add a new subsection into the **Introduction** chapter. In the text editor, add the following just after the Section tag for the **Introduction** section:

```
<!--
<Section>
  <Description>
    <Metadata name="Title">Snails are good to eat.</Metadata>
  </Description>
-->
```

   Then just before the next section tag (**What do you need to start?**), add the following:


```
<!--
</Section>
-->
```

The effect of these changes is to make a new subsection inside the **Introduction** chapter.

4. Open the Greenstone demo collection in the Librarian Interface. In the **Document Plugins** section of the **Design** panel, note that **HTMLPlugin** has the **description_tags** option set. This option is needed when `<Section>` tags are used in the source documents.

5. **Build** and **preview** the collection. Look at the **Farming snails 1** document again and check that your new section has been added.

****

