Lab 4: Greenstone formatting exercises

4.1. Formatting the HTML collection—Tudor

1. Open up your **tudor** collection, go to the **Format** panel (by clicking on its tab) and select **Format Features** from the left-hand list. Leave the editing controls at their default value, so that **Choose Feature** displays *All Features* and **VList** is selected as the **Affected Component**. The text in the **HTML Format String** box reads as follows:

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

   This displays something that looks like this:

   A discussion of question five from Tudor Quiz: Henry VIII
   (quizstuff.html)

   for a particular document whose *Title* metadata is **A discussion of question five from Tudor Quiz: Henry VIII** and whose *Source* metadata is **quizstuff.html**.

   This format appears in the search results list, in the **Titles** list, and also when you get down to individual documents in the **Subjects** hierarchy. This is Greenstone's default format statement.

   Greenstone's default format statement is complex because it is designed to produce something reasonable under almost any conditions, and also because for practical reasons it needs to be backwards compatible with legacy collections.

2. Delete the contents of the **HTML Format String** box and replace it with this simpler version:

   ```
   <td>[link][icon][/link]</td>
   <td>[ex.Title]<br>
       <i>([ex.Source])</i>
   </td>
   ```

   Preview the result (you don't need to build the collection, because changes to format statements take effect immediately). Look at some search results and at the **Titles** list. They are just the same as before! Under most circumstances this far simpler format statement is entirely equivalent to Greenstone's more complex default.

   *But there's a problem. Beside the bookshelves in the **Subjects** browser, beneath the subject appears a mysterious "()". What is printed for these bookshelves is governed by the same format statement, and though bookshelf nodes of the hierarchy have associated* Title *metadata—their title is the name of the metadata value associated with that bookshelf—they do not have **ex.Source** metadata, so it comes out blank.*

3. In the **Format Features** section of the **Format** panel, the **Choose Feature** menu (just above **Affected Component** menu) displays *All Features*. That implies that the same format is used for the search results, titles, and all nodes in the subject hierarchy—including internal nodes (that is, bookshelves). The **Choose Feature** menu can be used to restrict a format statement to a specific one of these lists. We

will override this format statement for the hierarchical *subject* classifier. In the **Choose Feature** menu, scroll down to the item that says

> CL2: Hierarchy -metadata dc.Subject and Keywords

and select it. This is the format statement that affects the second classifier (i.e., "CL2"), which is a **Hierarchy**classifier based on **dc.Subject and Keywords**metadata.

Click **<Add Format>**to add this format statement to the collection.

Edit the **HTML Format String**box below to read

```
<td>[link][icon][/link]</td>
<td>[ex.Title]</td>
```

4. **Preview** the **Subjects** list in the collection. First, the offending "()" has disappeared from the bookshelves. Second, when you get down to a list of documents in the subject hierarchy, the filename does not appear beside the title, because **ex.Source** is not specified in the format statement and this format statement applies to all nodes in the *subject* classifier. Note that the search results and titles lists have not changed: they still display the filename underneath the title.

5. Let's change the search results format so that **dc.Subject and Keywords** metadata is displayed here instead of the filename. In the **Choose Feature** menu (under **Format Features** on the **Format** panel), scroll down to the item **Search** and select it. Click **<Add Format>** to add this format statement to the collection. Change the **HTML Format String** box below to read

```
<td>[link][icon][/link]</td>
<td>[ex.Title]<br>
    [dc.Subject]
</td>
```

6. To insert the **[dc.Subject]**, position the cursor at the appropriate point and either type it in, or select it from the **Insert Variable...** drop down menu. This menu shows many of the things that you can put in square brackets in the format statement.

7. **Preview** the collection. Documents in the search results list will be displayed like this:

> A discussion of question five from Tudor Quiz: Henry VIII
> Tudor period|Others

(The vertical bar appears because this **dc.Subject and Keywords** metadata is hierarchical metadata. Unfortunately there is no way to get at individual components of the hierarchy. For most metadata, such as title and author, this isn't a problem.)

8. Finally, let's return to the *Subjects* hierarchy and learn how to do different things to the bookshelves and to the documents themselves. In the **Choose Feature** menu, re-select the item

> CL2: Hierarchy -metadata dc.Subject and Keywords

Edit the **HTML Format String**box below to read

```
<td>[link][icon][/link]</td>
<td>{If}{[numleafdocs],<b>Bookshelf title:</b> [ex.Title],
        <b>Title:</b> [ex.Title]}
</td>
```

Again, you can insert the items in square brackets by selecting them from the **Insert Variable...** drop down box.

*The **If** statement tests the value of the variable **numleafdocs**. This variable is only set for internal nodes of the hierarchy, i.e. bookshelves, and gives the number of documents below that node. If it is set we take the first branch, otherwise we take the second. Commas are used to separate the branches. The curly brackets serve to indicate that the **If** is special—otherwise the word "If" itself would be output.*

9. **Preview** the collection and examine the subject hierarchy again to see the effect of your changes. Bookshelves should say **Bookshelf title:** and then the title, while documents will display **Title:** and the title. Note that the number of documents in the bookshelf is not displayed: we are using `[numleafdocs]` to test what kind of item in the list we are at, but we are not displaying it.

4.2. Formatting the Word and PDF collection

In this exercise, we play around with the format statements in the Word and PDF collection.

1. Open the **reports** collection in the Librarian Interface and go to the **Format Features** section of the **Format** panel.

Tidying up the default format statement

2. In this part of the exercise, we make the format statement simpler without changing the resulting display.

   Greenstone's default format statement is complex because it is designed to produce something reasonable under almost any conditions, and also because for practical reasons it needs to be backwards compatible with legacy collections. For this collection, we don't need all of the complexity.

   Make sure that the **VList** format statement is selected in the list of formats.

   The default **VList** format statement looks like the following:

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

   This format statement is the default used for any vertical list, such as search results, classifiers, and document table of contents.

   {Or}{[ex.thumbicon],[ex.srcicon]} chooses *ex.thumbicon* metadata if its there, otherwise chooses *ex.srcicon* metadata. If neither are present, nothing is displayed. For this collection there is no *ex.thumbicon* metadata so the choice is not needed.

   Replace {Or}{[ex.thumbicon],[ex.srcicon]} (highlighted above) with [ex.srcicon].

   There is no *exp.Title* metadata, so remove that element from {Or}{[dc.Title],[exp.Title],[ex.Title],Untitled}.

   The resulting format statement looks like the following:

   ```
   <td valign=top>[link][icon][/link]</td>
   <td valign=top>[ex.srclink][ex.srcicon][ex./srclink]</td>
   <td valign=top>[highlight]
   {Or}{[dc.Title],[ex.Title],Untitled} [/highlight] {If}{[ex.Source],
   <br><i>([ex.Source])</i>}</td>
   ```

   Preview the collection to make sure the display hasn't changed. You shouldn't notice any difference when looking at search results, classifiers etc.

Linking to Greenstone version or original version of documents

3. For collections with documents that undergo a conversion process during importing (e.g. Word, PDF, PowerPoint documents, but not text, HTML documents), the original file is stored in the collection along with the converted version. The default **VList** format statement links to both versions:

   `[ex.link][icon][/ex.link]` links to the Greenstone HTML version, while `[ex.srclink][ex.srcicon][/ex.srclink]` links to the original.

   Choose **SearchVList** in **Format Features** by selecting **Search** from the **Choose Feature** drop down list, and **VList** from the **Affected Component** list. Click **<Add Format>** to add the **SearchVList** format statement into the list of assigned formats. Experiment with removing either of the two links from the format statement.

   To see the results of your changes, preview the collection and do a search. You are making changes to **SearchVList**, which means the changes will only apply to search results.

   Storing and displaying the original allows users to see the correct format, but requires the user to have the relevant program installed. It also increases the size of the collection. The Greenstone version can be viewed in a browser, but may not look as nice.

Making bookshelves show how many items they contain

4. Next, we'll customize the format for the *Creators* list. Classifier bookshelves have only a few pieces of metadata to display: `[ex.Title]` and `[numleafdocs]`. Whatever metadata the classifier has been built on, the bookshelf label is always stored as `[ex.Title]`. This is why a Creator is printed out for each bookshelf even though `[dc.Creator]` is not specified in the format statement. `[numleafdocs]` is only defined for bookshelves, so this metadata can be used in an `{If}` statement to make bookshelves and documents display differently in the list.

   Make each bookshelf in the Creator classifier show how many entries it contains. In the **Format Features** section of the **Format** panel, select the **CL2 AZCompactList** classifier which is based on **dc.Creator** metadata from the **Choose Feature** drop down list, and **VList** from the **Affected Component** list. Click the **<Add Format>** button to add this format into the list of assigned formats. Note that it gets added as **CL2VList** in this list: it is the **VList** format for the second (**CL2**) classifier.

   Append the following text to the bottom of the format statement:

   `{If}{[numleafdocs],<td><i>([numleafdocs])</i></td>}`

   Preview the collection. Click on the *Creators* list and notice that the bookshelves now display how many documents they contain.

   This revised format statement has the effect of specifying in brackets how many items are contained within a bookshelf. Since only bookshelves define `[numleafdocs]`, only they will display this. By modifying **CL2VList** instead of **VList**, the change will only apply to the second classifier (Creators).

Displaying multi-valued metadata

5. Next we modify the document entries in the Creator classifier to display all authors. Back in **Format Features**, select the **CL2VList** format in the list of assigned formats. After `{If}{[ex.Source], <br>` in the format statement, add `[sibling:dc.Creator]`.

   `[ex.Source]` is not defined for bookshelves, so can also be used to differentiate bookshelves and documents.

The resulting format statement looks like:

```
<td valign=top>[link][icon][/link]</td>
<td valign=top>[ex.srclink][ex.srcicon][ex./srclink]</td>
<td valign=top>[highlight]
{Or}{[dc.Title],[ex.Title],Untitled}[/highlight]
{If}{[ex.Source],<br>[sibling:dc.Creator]
<i>([ex.Source])</i>}</td>
{If}{[numleafdocs],<td><i>([numleafdocs])</i></td>}
```

This will display the Greenstone link, the link to the original, then the Title. For bookshelves, it will also display how many documents the bookshelf contains. For documents, it will display all the Authors (Creators), and the source document. [sibling:dc.Creator] displays all the Creator metadata for the document, separated by a space (" "), while [dc.Creator] displays only the first author. Preview the *Creators* list and make sure that all authors are displayed for documents.

6. You can change the separator between the authors. Modify the format statement, and replace [sibling:dc.Creator] with [sibling(All'
'):dc.Creator]. This will add a new line after each author (
 specifies a line break in HTML). Preview the *Creators* list.

If you have done exercise **Enhanced Word document handling**, the collection will have both dc.Creator and ex.Creator metadata. To display both, you can use

```
[sibling:dc.Creator] [sibling:ex.Creator]
```

To display dc.Creator if it is present, otherwise display ex.Creator, use

```
{Or}{[sibling:dc.Creator],[sibling:ex.Creator]}
```

Advanced multi-valued metadata

7. You may notice that the **AZCompactList** classifier has two options after the **metadata** option: **firstvalueonly** and **allvalues**. Manually added metadata can be used to replace or enhance automatically extracted metadata, and these options control exactly which pieces of metadata a document is classified by.

For example, say we have two documents. Document 1 has four Creators specified (dc.Creator = dcA, dc.Creator = dcB, ex.Creator = exA, ex.Creator = exB), while document 2 has three (ex.Creator = exA, ex.Creator = exB, ex.Creator = exC). The following table shows which metadata values each document is classified by, for the different classifier options:

AZCompactList options	Document 1	Document 2
-metadata dc.Creator,ex.Creator	dcA, dcB	exA, exB, exC
-metadata dc.Creator,ex.Creator -firstvalueonly	dcA	exA
-metadata dc.Creator,ex.Creator -allvalues	dcA, dcB, exA, exB	exA, exB, exC

8. Now we set the **firstvalueonly** option for the *Creators* classifier. Switch to the **Browsing Classifiers** section of the **Design** panel, select the **AZCompactList** for **dc.Creator** metadata in the **Assigned Classifiers** box and click **<Configure Classifier...>**. Select the **firstvalueonly** option.

Rebuild and **preview** the collection. Now the *Creators* list classifies documents based on the first author appearing in the **dc.Creator** metadata.

If you set the **metadata** field of **AZCompactList** to dc.Creator,ex.Creator in the **A collection of**

Word and PDF files exercise, now the *Creators* list will classify based on the first author appearing in either the **dc.Creator** metadata or the **ex.Creator** metadata.