Greenstone tutorial exercises, Version 1 (May 2005)

1. Working with a pre-packaged collection (UNAIDS) 3

Installing a pre-packaged Greenstone collection 3
Browsing around a Greenstone collection 3
Searching within a Greenstone collection 4
Leaving the Greenstone digital library 4
Exercise: Use the UNAIDS collection to answer these questions 4

2. Working with a pre-packaged collection (Digital Libraries in Education) 5

Installing a pre-packaged collection 5
Browsing around a Greenstone collection 5
Exercise: Read the Help page; then answer these questions 6
Exercise: Use the How to build a digital library collection to answer these questions. 6

3. Installing Greenstone 7

Installing Greenstone on a Windows system 7
Installing ImageMagick on a Windows system 8

4. Updating a Greenstone installation 9

Removing Greenstone from a Windows system 9
Reinstalling Greenstone on a Windows system 9
Amalgamating different Greenstone collections 9
Installing the Greenstone language pack 9

5. Building a small collection of HTML files 10

Setting up a shortcut in the Librarian interface 11

6. A collection of Word and PDF files 12

Manually adding metadata to documents in a collection 12
Collection design; branding a collection with an image 12
Document plugins 12
Search types and fielded searching 13
Search indexes 13
Browsing classifiers 13

7. Difficult PDF documents 14

Modes in the Librarian Interface 14

8. A simple image collection 15

Adding a metadata set to the collection 15
Adding Title metadata 15
Change Format Features to display new metadata 15
Changing the size of image thumbnails 16
Adding Description metadata 16
Adding a browsing classifier based on Description metadata 17
Creating a searchable index based on Description metadata 17

9. A large collection of HTML files—Tudor 18

Looking at different views of the files in the Gather and Enrich panels 18

10. Exporting a collection to CD-ROM 19

11. Pointing to documents on the web 20

12. Downloading files from the web 21

13. Enhanced collection of HTML files 22

Adding hierarchically-structured metadata and a Hierarchy classifier 22
Partitioning the full-text index based on metadata values 22
Adding a hierarchical phrase index (PHIND) 23
Controlling the building process 23

14. Learning about formats and macros 24

Experimenting with format statements 24
Collection-specific macros 25
General macros 26

15. A bibliographic collection 27

Adding fielded searching 27
Branding a collection with an image 27

16. Looking at a multimedia collection 28

17. Building a multimedia collection 29

Manually correcting metadata 29
Browsing by media type 30
Suppressing dummy text 30
Using AZCompactList rather than AZList 31
Making bookshelves show how many items they contain 31
Using UnknownPlug 32
Cleaning up a title browser using regular expressions 32
Using non-standard macro files 33
Using different icons for different media types 33
Changing the collection's background image 34
Building a full-size version of the collection 34
Adding an image collage browser 34

18. Scanned image collection 35

Grouping documents by series title and displaying dates within each group 35
Suppressing dummy text 36

19. Open Archives Initiative (OAI) collection 37

Tweaking the presentation with format statements 37

20. Downloading over OAI 39

21. Exporting a collection as METS 39

22. Moving a collection from DSpace to Greenstone 40

Adding indexing and browsing capabilities to match DSpace's 40

23. Moving a collection from Greenstone to DSpace 41

Using Greenstone from the command line 41

1. Working with a pre-packaged collection (UNAIDS)

You will need the Greenstone UNAIDS *CD-ROM*

Installing a pre-packaged Greenstone collection

1. □ On inserting the **UNAIDS CD-ROM**, for many computers installation will begin automatically. If not, “auto-run”—a configurable setting under Windows—is disabled on your computer and you need to double-click *setup.exe* on the CD-ROM.

   My Computer → UNAIDS20 → setup.exe

2. □ The InstallShield Wizard begins to install the UNAIDS pre-packaged collection. Select the English language.

3. □ Click the **<next>** button.

4. □ Choose **Run from CD-ROM (standard)** as the setup type. This is the default and is already selected. Then click **<next>**.

5. □ Click **<next>** again to install the UNAIDS collection in the default folder, which is **C:\Program Files\UNAIDS Library 2.0 [CD-ROM]**.

   Installation Wizard copies the required files from CD-ROM to disk

6. □ Click **<OK>** to confirm completion of UNAIDS collection (twice).

   InstallShield quits—the UNAIDS Library is installed.

CD-ROMs like this one that contain pre-packaged Greenstone collections do not include the full Greenstone software. Instead they embody a mini version of Greenstone that allows you to view the collection but not to build new ones.

Browsing around a Greenstone collection

7. □ Launch the prebuilt library by clicking:

   Start → All Programs → UNAIDS Library 2.0.

To access Greenstone through the Local Library Server, it is sometimes necessary to turn off the proxy settings of the browser. Greenstone normally detects this and pops up a window alerting you to the problem.

8. □ Click **<Enter Library>** in the dialog box and your browser (typically Internet Explorer by default) will display the Greenstone home page.

9. □ Within the web browser, click **titles a-z** (in the centre of the navigation bar near the top of the page).

10. Access the **first book** in the list of titles by clicking the **book icon** next to the title: *About UNAIDS*.

11. Use the scroll bar to view the full length of the page.

12. In the table of contents near the top, click the **page icon** next to the heading *Guiding principles of UNAIDS* to view this section.

13. Click the **page icon** next to the heading *Global and local impact* to view the next section.

This style of interaction can be continued to further expand and contract folders and switch to a different section.

14. To fully expand the contents of this introduction chapter, click **Expand Document or Chapter** in the upper left portion of the page, under the picture of the document’s front cover.

15. You can return to the currently selected page of document titles by clicking the **book icon** next to the title of the book at the top of the table of contents (this signifies closing the book). You also get to the document titles using **titles a-z** in the navigation bar, in this case to the titles beginning with A–D.

If the table of contents is open at the top level—showing all the chapters—then clicking ***Expand Document or Chapter*** *expands the full document. For long documents, which take some time to load in, Greenstone seeks confirmation for this action: clicking 'continue' loads the full document.*

16. Browse around and peruse some other documents in the collection.

Searching within a Greenstone collection

17. Access the search page by clicking **search** in the navigation bar.
18. In the query box under **Search chapters in any language which contain some of the words**, enter the term **gender** then click **<Begin Search>**.

    After a short pause, the web browser loads a fresh page showing the results of the search.

19. Click the **page icon** for the **first matching document** in the result set (*Five Year Implementation Review of the Vienna Declaration and Programme of Action*) to view the document. Because the search was at the chapter level, you are taken directly to the matching chapter within the document.
20. Experiment further with searching, and with the interface in general. For example, there is a detailed **Help** page. It contains a **Preferences** section through which you can control some search settings.

    The Preferences options in the UNAIDS collection are intentionally minimalist. Most collections have a separate Preferences button that offers more features.

    The home page of the UNAIDS library collection cycles through a sequence of front cover images, updated every 5 seconds or so. Clicking a particular image takes you directly to that document.

Leaving the Greenstone digital library

21. There are two ways of leaving Greenstone:

    a. Exit from the Greenstone Software server. Click on the **Greenstone Software** in the task bar, then choose **Exit** from the **Browser Selection and Settings** menu (or click on the exit hotspot, the red cross at the top right). The Greenstone Software exits, but your web browser continues to run.

    b. Exit from your web browser. Leave your web browser in the usual way. The Greenstone server detects when you exit from the browser and generates a popup window that asks whether to close down the server as well. (The reason is that other people may be using Greenstone over the network, and should not be rudely terminated.)

Exercise: Use the UNAIDS collection to answer these questions

- How many publications are there in the collection? *900*
- How many documents are there that mention *Australia* in the title? *15*
- How many top-level subject categories are there? *21*
- What does AAVP stand for? *African Aids Vaccination Programme*
- What does AIDS stand for? *Acquired Immuno-Deficiency Syndrome (Search for "AIDS stands for")*
- Considering lower case variants only, how many times does the word "condom" appear in the collection? How many times for "condoms"? *6789, 5243*
- If case sensitivity does not matter, how many times does the word "condom" appear in the collection? How many times for "condoms"? *7905, 5571*
- If word endings are ignored, how many times does "condom" and variants such as "condoms" appear in the collection? *13477*
- How many *chapters* contain some variations of the word "condom"? Does this make it a useful search term?
- *2413 chapters.* No, since there are only 900 documents
-
- What year saw the first reported case of AIDS in New Zealand? *1983*

2. Working with a pre-packaged collection (Digital Libraries in Education)

You will need the Greenstone Digital Libraries in Education *CD-ROM*

Installing a pre-packaged collection

1. Insert your CD-ROM for the course *Digital libraries in education* into a Windows computer. If the installation process does not start up straightaway (because the AutoPlay feature is disabled on your computer), navigate to your CD-ROM/DVD drive (normally D:), open the folder *prebuilt*, and double click on **Setup.exe**.
2. During installation you are offered a choice of folder to install in: we recommend the default, which is *C:\GSDL*.
3. You are also presented with the option to run Greenstone from the CD-ROM or to copy the entire CD-ROM. We recommend the latter: please check the box that says **Install all collection files**. It will take at least a couple of minutes to copy the files across.
4. Finally, the installer offers to install the Netscape browser for you. Do *not* request this except in the unlikely event that you do not already have a web browser on your computer.

CD-ROMs like this one that contain pre-packaged Greenstone collections do not include the full Greenstone software. Instead they embody a mini version of Greenstone that allows you to view the collection but not to build new ones.

Browsing around a Greenstone collection

5. To run Greenstone, open the Windows Start menu, Programs, and select *Greenstone*, then the submenu item *Digital Libraries in Education*: then <***Enter Library***>.
6. Click the Digital libraries in Education collection's icon. This takes you to the collection's home page, often called the "about" page.

The home page contains an access bar with buttons called search, contents, authors a–z, modules, and acronyms. This access bar is the key to finding information in any Greenstone collection.

7. Click <**authors a–z**>. A list of bookshelf icons appears. Click the one called Marchionini, G. to see the two course readings by Gary Marchionini.
8. One of these items is a PDF file and the other is an HTML file. Click them both in turn to open up the documents.
9. Click the <**contents**> button in the access bar. This shows two bookshelves, one for this Study Guide and the other for the Course Readings. Choose one and look at what it contains.
10. Clicking a bookshelf that is open closes it. Close the bookshelf you have just opened and then choose the other one and examine its contents.
11. Click <**acronyms**> in the access bar and find the meaning of the acronym "LOM".
12. Click <**search**> and search for the word "LOM". Check out the difference between searching text and searching titles (use the pull-down box on the search page).
13. Click the collection icon **Digital Libraries in Education** at the top left. This takes you back to the collection's **about** page.

    Beneath the access bar on the collection's about page is a search box (just the same as the one that appears on the search page), a description of the collection under the heading **About this collection**, and instructions on how to find information in this collection.

    Above the access bar is the collection's icon, saying **Digital Libraries in Education**. On the right is an icon saying **about**, above which are three buttons, **home**, **help**, and **preferences**.
14. Click <**home**>. This returns you to the Greenstone home page.

15. Return to the collection (by clicking its icon), and click <**help**>. This gives more information about how to access the collection.

16. Click <**preferences**>. This takes you to a page where you can change some of the settings.

17. Now explore the collection by navigating freely around it. Click liberally: all images that appear on the screen are clickable. If you hold the mouse stationary over an image, most browsers will soon pop up a brief "mouse-over" message that tells you what will happen if you click. Experiment! Choose common words like "the" or "and" to search for—that should evoke some response, and nothing will break. (Note: unlike many search systems, Greenstone indexes all words, including these ones.)

Exercise: Read the Help page; then answer these questions

- What does this collection contain?
- Name five ways to navigate to a target document in this collection.
- How many documents in the collection are written by Erik Duval?
- Compare the number of times the words "he" and "she" appear in the collection.
- How many times does the word "metadata" appear in titles? In the text itself?
- What's the difference between a *some* and an *all* search?
- What does "MODS" stand for?
- How do you switch the interface from English to Russian? Does it stay in Russian when you go to the Greenstone home page?
- Find a search term that yields different results depending on whether you have *ignore word endings* or *whole word must match* set on the Preferences page.
- What's the difference between Graphical and Textual interface format (on the Preferences page)?

Exercise: Use the How to build a digital library collection to answer these questions.

- How many sentences contain the word education?
- What story from the School Journal collection is featured in the book?
- How many acronyms used in the book begin with the word Standard?
- What does tapu mean?
- How many times does the word library appear? The word libraries?
- How many times does Library appear with an initial capital letter?
- How many times does some derivative of the word form appear?
- Name an English poem that was probably written in about 1000 A.D.
- Who is Alan Kay?
- On what page is the first mention of some aspect of Chinese culture?

Most of these questions would be rather difficult to answer from the printed book.

3. Installing Greenstone

Installing Greenstone on a Windows system

There are various ways of getting Greenstone:

1. From a UNESCO CD-ROM (version 2.60) (or FAO IMARK CD-ROM, but this is an earlier version 2.51)

   These CD-ROMs contain the **Greenstone software**, plus **documented example collections**, four **language interfaces** (English French Spanish Russian), the **Export to CD-ROM** package, the **ImageMagick** graphics package, the **Java runtime environment**, and an **installer** that installs all of these.

2. From the IITE Digital Libraries in Education CD-ROM, or a Greenstone workshop CD-ROM

   In addition to all the above software, these CD-ROMs contain the **Greenstone Language Pack**, which gives reader's interfaces in many languages (currently about 40). This has its own installer which you have to invoke separately, after you have installed Greenstone. They also contain a set of **sample files** to be used for exercises.

All these CD-ROMs contain the full Greenstone software, which allows you to view collections and build new ones. They are not the same as CD-ROMs that contain a pre-packaged Greenstone collection, which only allow you to view that collection.

3. From http://www.greenstone.org

   Most people download the Windows distribution from *http://www.greenstone.org*, which contains the latest version of the Greenstone. There are several optional modules that must be downloaded separately (to avoid a single massive download): **documented example collections**, the **Export to CD-ROM** package, and the **Language Pack**. There is also the set of **sample files** used in these exercises. (To reduce the download size the documented example collections are distributed in unbuilt form and need to be built.)

   You need **Java** to run Greenstone. You might already have it; otherwise download it from *http://java.sun.com*. To work with image collections, you need **ImageMagick** (from *http://www.imagemagick.org*).

Most Greenstone CD-ROMs start the installation process as soon as they are inserted into the drive, assuming that the AutoPlay feature is enabled on your computer. If installation does not begin by itself, locate the file *setup.exe* and double click it to start the installation process. (On the IMARK CD-ROM this file resides in the folder *software_tools→Greenstone*). If you download Greenstone over the web, what you get is the installer—just double-click it.

If Greenstone has been installed on your computer before, you should completely remove the old version before installing a new one. (However, you need not remove any pre-packaged collections that you may have installed.) To do this, see below under *Updating a Greenstone installation*.

Here is what you need to do to install Greenstone. Older versions of the installer follow much the same sequence but use slightly different wording.

- Select the language for this installation. We choose **English**
- Welcome to the InstallShield Wizard for the Greenstone Digital Library Software. Click <**Next**>
- License Agreement. Accept the agreement and then click <**Next**>
- Choose location to install Greenstone. Leave at the default and click <**Next**>
- Setup Type. Leave at the default (Local Library) and click <**Next**>
- (For older installers you must now select collections. Leave at the default, Documented Example Collections, and click <**Next**>)
- Set admin password. Choose a suitable password and click <**Next**> (If your computer will not be serving collections online, the password doesn't matter)
- Click <**Install**> to complete the installation
- Files are copied across

- Installation is complete. If you are installing from a CD-ROM, the installer will offer to install ImageMagick (see below), and Java, if necessary.

To invoke the Greenstone Reader's interface, go to the *Greenstone Digital Library Software* item under *Programs* on the Windows *Start* menu and select *Greenstone Digital Library*. To invoke the Librarian interface, go to the same item and select *Greenstone Librarian Interface*.

Installing ImageMagick on a Windows system

Once Greenstone has been installed, you should ensure that ImageMagick is installed on your computer if you wish to build any image collections. If you are installing from a Greenstone CD-ROM, you will be asked whether you want to install ImageMagick: say **Yes**. If you are not, you will need to download ImageMagick (from *http://www.imagemagick.org*). To install this program you must have Windows "Administrator" privileges.[1] The remaining steps are straightforward, and, as before, we recommend the default settings. Here is what you need to do.

- "This will install ImageMagick 5.5.7 Q8. Do you wish to continue?" **Yes**
- "Welcome to the ImageMagick Setup Wizard Click <**Next**>
- "Information: Please read the following …" Click <**Next**>
- "Select Destination Directory …" Leave at default and click <**Next**>
- "Select Start Menu Folder …" Leave at default and click <**Next**>
- "Select Additional Tasks …" Leave at default and click <**Next**>
- "Ready to Install". Click <**Install**>
- Files are copied across
- "You have now installed …" Click <**Next**>
- "Setup has finished …". Deselect "View index.html" and click <**Finish**>.

[1] If you do not have Windows Administrator privileges, the ImageMagick installer will give a cryptic error complaining that it failed to set a particular Windows registry value. If this happens you can continue your work with Greenstone, but you will not be able to build collections of images.

4. Updating a Greenstone installation

These tutorial exercises assume that you are using Greenstone 2.60 or above.

Before updating to a new version of Greenstone, ensure that the computer is not running the Greenstone Librarian Interface or the Greenstone local library server. Normally, quitting your web browser, or quitting the Librarian Interface, also quits the server.

Removing Greenstone from a Windows system

Completely remove the existing version before you install a new version of Greenstone.

1. Ensure that you are not running Greenstone.
2. Remove the old version by going to the Windows Control Panel (from the *Settings* item on the *Start* menu). Click **Add or Remove Programs**, select **Greenstone Digital Library Software**, and **Remove** it. (To do this you may need Windows "Administrator" privileges.)
3. At the end of this procedure you will be asked whether you would like all your Greenstone collections to be removed: you should probably say *No* if you wish to preserve your work.

Occasionally, problems are encountered if older Greenstone installations are not fully removed. To clean up your system, move your Greenstone collect folder, which contains all your collections, to the desktop. Then check for the folder C:\Program Files\gsdl or C:\Program Files\Greenstone, which is where Greenstone is usually installed, and remove it completely if it exists.

Reinstalling Greenstone on a Windows system

4. The reinstallation procedure is exactly the same as the original installation procedure, described above. If you already have ImageMagick, you do not need to install it again.

There have been some superficial changes to the installation procedure in moving to Greenstone Version 2.60, because it uses a different installer program.

There is another important difference that you should be aware of: Versions 2.60 and above are installed in the folder Program Files\Greenstone*, whereas prior versions were placed in the folder* Program Files\gsdl *(these are both default locations that you could have changed during installation.) When upgrading to Version 2.60, if you want to save existing collections you must explicitly move the contents of your collect folder from the old place to the new one. Future Greenstone versions will be installed in the new place,* Program Files\Greenstone*, so this problem will not happen again.*

Amalgamating different Greenstone collections

5. If you have previously installed the Greenstone Digital Library software in a non-standard place, you should amalgamate your collections by moving them from the *collect* folder in the old place into the folder *Program Files\Greenstone\collect*.
6. If you have installed collections from pre-packaged Greenstone CD-ROMs, they reside in a different place: *C:\GSDL\collect*. To amalgamate these with your main Greenstone installation, move them into the folder *Program Files\Greenstone\collect*. The mini version of Greenstone that is associated with the pre-packaged collections is no longer necessary. To uninstall it, select *Uninstall* on the Greenstone menu of the Windows *Start* menu.

Installing the Greenstone language pack

If you go to the Preferences page of any Greenstone collection, and look at the ***Interface language*** *menu, you will probably find that only English, Spanish, French and Russian interfaces are installed.*

7. Locate the Greenstone Language Pack. This may be on the CD-ROM from which you installed Greenstone, or you may have to download it from *http://www.greenstone.org*.
8. Double-click the *.exe* file; this will start the installer. Accept all the defaults
9. Restart the Greenstone Digital Library and look at the interface language menu again. Now you should see about 40 different languages.

5. Building a small collection of HTML files

You will need some HTML files, such as those in the hobbits *folder in* sample_files. *You can download the sample files that are used in these exercises from http://www.greenstone.org.*

1. □ Start the Greenstone Librarian Interface:

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

   After a short pause a startup screen appears, and then after a slightly longer pause the main Greenstone Librarian Interface appears.

2. □ Start a new collection within the Librarian Interface:

   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. This is discussed in other exercises. For now, select **Dublin Core Metadata Element Set Version 1.1** followed by **<OK>**.

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

6. □ Now drag the *hobbits* 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.

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

7. □ 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.

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

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

10. 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!

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

12. Presently there is no manually assigned metadata, but the act of building the collection has extracted metadata from the documents. Double click the *hobbits* 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.

13. 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.

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

Setting up a shortcut in the Librarian interface

15. To set up a shortcut to the source files, return to the Gather panel and 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. Follow the instructions to set up a shortcut. Close all the folders in the file tree and you will see the shortcut to your source files in the left-hand pane of the Gather panel.

6. 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**, fill out appropriate fields for it, and choose Dublin Core as the metadata set.
2. Copy the 12 files from *sample_files*→*Word_and_PDF*→*Documents* into the collection. 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.
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.". The PostScript documents (*cluster.ps* and *langmodl.ps* do not have extracted titles: what appears in the *titles a–z* list is just the first few characters of the document).

Manually adding metadata to documents in a collection

6. In the **Enrich** panel, manually add Dublin Core *dc.Title* metadata to one of these documents. Select *word03.doc* and double-click to open it in Word. Copy the title of this document ("Greenstone: A comprehensive open-source digital library software system") from Word, return to the Librarian Interface, click the *dc.Title* field, and paste the value into the Value box. Click <**Append**>.
7. Now add *dc.Creator* information for the same document. You can add more than one value for the same field, to accommodate multiple authors—just put in the next value and click <**Append**>.
8. Next add title and creator metadata for a few of the other documents.

If you build and preview your collection at this point, you will find that nothing has changed. You need to alter the collection design to use the new Dublin Core metadata instead of the original extracted metadata.

Collection design; branding a collection with an image

9. Change to the **Design** panel, which is split into several sections. The first section **General Options** appears. This allows you to modify the values you provided when defining the collection, if desired. You can also brand the collection using a suitable image.
10. Click on the <**Browse**> button associated with "URL to about page icon", and browse to the image *sample_files*→*Word_and_PDF*→*wrdpdf.gif* on your computer. When you select this image, Greenstone automatically generates an appropriate URL for the image.
11. If you are on the web, you can easily make your own Greenstone-style icon by going to

    http://www.greenstone.org/make-images.html

    and following the instructions there.

Document plugins

12. Now look at the **Document Plugins** section, 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 ZIPPlug, TEXTPlug, HTMLPlug, EMAILPlug ImagePlug and NULPlug plugins. To delete a plugin, select it and click <**Remove Plugin**>. GAPlug is required for any type of source collection and should not be removed.

Search types and fielded searching

13. Go to the **Search Types** section. This specifies what kind of search interface and what search indexes will be provided for the collection. Let's add a form search option. Click <**Enable Advanced Searches**>; this includes "form search" in the collection.

14. To include "plain search" as well, pull down the **Search Types** menu and select **plain**; then click <**Add Search Type**>.

15. To set plain search as the default type, click on **plain**, then click <**Move Up**>. The first one in the list will be the default.

Search indexes

16. 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.Title* and *ex.Source* indexes, which are not particularly useful, by selecting them one at a time and clicking <**Remove Index**>. Only the *text* index remains.

17. Now add a Title index based on *dc.Title* by providing an **Index Name** (e.g. "Document Title") and selecting *dc.Title* from the **Index Source** box. Then click <**Add Index**>.

18. You can add indexes based on any metadata. Add an index called "Authors" based on *dc.Creator* metadata.

*The next two sections are **Partition Indexes** and **Cross-Collection Search**. In this exercise, we will not make any changes to these.*

Browsing classifiers

19. 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. Remove both of these by selecting them in turn and clicking <**Remove Classifier**>.

20. Now we add an *AZList* classifier for *dc.Title* metadata. Select *AZList* from the **Select classifier to add** drop-down list and click <**Add Classifier**>

21. A popup window **Configuring Arguments** appears. Select *dc.Title* from the **metadata** drop-down list, and check the **button name** checkbox to provide a button name; call it "Title". Now click <**OK**>.

22. Now add an *AZCompactList* classifier. Click <**Add Classifier**> and configure it to use *dc.Creator* metadata, with button name "Creator". Check the **mingroup** box and select 1 (to group all publications by the same author under a single bookshelf icon). Click <**OK**>.

*The last three sections are **Format Features**, **Translate Text** and **Metadata Sets**. In this exercise, we will not make any changes to these.*

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

24. Check that all the facilities work properly. There should be three full-text indexes, called *text*, *Document Title*, and *Authors*. In the *titles a–z* list should appear all the documents to which you have assigned *dc.Title* metadata (and only those documents). In the *authors a–z* list should appear 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.

At this point you might like to publish the collection on CD-ROM, See below under "Exporting a collection to CD-ROM" for how to do this.

7. Difficult PDF documents

25. Build a fresh Greenstone collection from the two files in *sample_files\difficult_documents*. Use the default collection configuration: that is, simply gather the files into a new collection, and build it.

These files are called No extractable text.pdf *and* Weird characters.pdf—*their names hint at the problems they will cause!*

26. Now preview the collection. The titles and filenames lists show only one of the documents. When you click the "text" icon to look at the text extracted from that document, it's garbage. During the building process this message appeared: "One document was processed and included in the collection; one was rejected."

Modes in the Librarian Interface

The Librarian Interface can operate in different modes. So far, you have been using the default mode, called "Librarian."

27. Use the *Preferences* item on the *File* menu to switch to *Expert* mode and then build the collection again. The **Create** panel looks different in Expert mode because it gives more options: locate the **Build Collection** button, near the bottom of the window, and click it. Now a message appears saying that the file could not be processed, and why.

28. We recommend that you switch back to *Librarian* mode for subsequent exercises, to avoid confusion.

8. A simple image collection

1. 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.

Greenstone does not ask you to choose a metadata set because the new collection inherits whatever is used by the base collection.

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 <**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.□ Change to the **Enrich** panel and view the extracted metadata for *Ascent.jpg*.

We now add our own metadata and use it to give users a new way to browse the collection. We use the Dublin Core metadata set.

Adding a metadata set to the collection

7.□ The collection (image-e) on which **backdrop** is based uses only extracted metadata. To add another metadata set, go to the **Design** panel of the Librarian Interface and click <**Metadata Sets**> in the list on the left (the last one). Then click **<Add Metadata Set**> (lower left button).

8.□ In the window that pops up, select **dublin.mds** and click <**Add Metadata Set**>.

Adding Title metadata

9.□ Now switch to the **Enrich** panel by clicking this tab. The metadata for each file now shows the Dublin core *dc.* fields as well as the extracted *ex.* fields.

10. We work with just the first three files (*Ascent.jpg*, *Autumn.jpg* and *Azul.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.

11. Click on ***Ascent.jpg*** so its metadata fields are available, then click on its **dc.Title** field on the right-hand side. Click on the **Value** text box, enter **Ascent**, and click **<Append>**.

*The **All Previous Values** box will become more useful when more entries have been added.*

12. Repeat the process for **Autumn.jpg** and **Azul.jpg**.

Now we customize the collection's appearance. Building or previewing the collection at this point won't reveal 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

13. Go to the **Design** 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** remains blank 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]"

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

14. Next click **<Replace Format>**. The first of the above changes alters the fragment of text that appears to the right of the thumbnail image, the second alters the item of metadata that follows it.

15. Go to the **Create** panel and click **<Build Collection>**. Now **preview** the collection. When you click on **browse** in the navigation bar the presentation has changed to "Title: Ascent" and so on.

Because we only assigned metadata to the first three items, after this the title becomes blank because the subsequent items have no dc.Title metadata. To get a full listing, enter all the metadata.

For some design parameters the collection must be rebuilt before the effect of changes can be seen. However, changes to format statements take place immediately and you can see the result straightaway by clicking ***reload*** *in the web browser.*

Changing the size of image thumbnails

16. Thumbnail images are created by the *ImagePlug* 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 **plugin ImagePlug** to pop up a window that shows its settings. (Alternatively, select **ImagePlug** 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>**.

17. **Build** and **preview** the collection.

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

Now add metadata that describes the photos in the collection. Again, for illustration, we focus on the first three images (Ascent.jpg, Autumn.jpg and Azul.jpg).

Adding Description metadata

19. Switch to the **Enrich** panel and select *Ascent.jpg*. We'll store our description in the **dc.Description** metadata element, so select it now in the right-hand panel.

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 Ascent.jpg: the image will normally be displayed by Microsoft's Photo Editor (although this depends on how your computer has been set up).

20. Back in the Librarian Interface enter the text **Moon rising over mountain landscape** as the **dc.Description** field's value and click **<Append>** to have it added.

21. Repeat this process for *Autumn.jpg* and *Azul.jpg*, adding a suitable description for each.

22. Build the collection again, to incorporate the new metadata.

23. Now update the format statement to use the new **dc.Description** metadata. Switch to the **Design** panel and enter the **Format Features** section by selecting this from the list of names on the left-hand size and ensure **VList** is selected. In the **HTML Format String**, place your cursor after the text that says

```
[dc.Title]<br>
```

24. and add the following text:

```
Description: [dc.Description]<br>
```

25. Then click **<Replace Format>**.

26. **Preview** the result (you don't need to build the collection as was done in step 22 to incorporate the metadata, because changes to format statements take effect immediately). Each image's description should appear beside the thumbnail, following the title.

Adding a browsing classifier based on Description metadata

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

28. A window pops up to control the classifier's options. Set the menu item for metadata to **dc.Description** and click **<OK>**. Now switch to the **Create** panel, **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

29. Switch to the **Design** panel and select **Search Indexes** from the left-hand list. Enter the text "descriptions" as the **Index Name**, select **dc.Description** and click **<Add Index>**.

30. Switch to the **Create** panel, **build** the collection, then **preview** it. As an example, search for the term "mountain" in the *descriptions* index.

9. A large collection of HTML files—Tudor

1. Invoke the Greenstone Librarian Interface (from the Windows *Start* menu) and start a new collection called **tudor** (use the *File* menu). Fill out the pop-up dialog with appropriate values and leave **Dublin Core**, which is selected by default, as the metadata set.
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.
4. Switch to the **Create** panel and click **<Build Collection>**.
5. When building has finished, **preview** the collection.
6. The browsing facilities in this collection (*titles a–z* and *filenames*) are based entirely on extracted metadata. Return to the Librarian Interface and examine the metadata that has been extracted for some of the files.

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.)

7. Switch to the **Design** panel and select the **Document Plugins** section. 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! Select the **plugin HTMLPlug** line and click <**Configure Plugin**>. A popup window appears. Scroll down the page to locate the **smart_block** option and switch it off. Click <**OK**>.
8. Switch to the **Create** panel and **build** 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

9. Switch to the **Gather** panel and in the right-hand side open *englishhistory.net* → *tudor*.
10. 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.
11. Change the **Show Files** menu to **Images**. Again, the files shown above alter.
12. 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.

10. Exporting a collection to CD-ROM

To publish a collection on CD-ROM, Greenstone's Export to CD-ROM export *module must be installed (see above, under "Installing Greenstone").*

1. Launch the Greenstone Librarian Interface if it is not already running.
2. Choose **File→Write CD/DVD image**, and in the popup window select the **tudor** collection as the collection to export. You can optionally name the CD-ROM; otherwise the default "collections" is used. Do so now, entering "Tudor collection" in the field for **CD/DVD name**; then click **<Write CD/DVD image>**.

   The necessary files for export are written to:

   C:\Program Files\Greenstone\tmp\exported_Tudorcollection

   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_Tudorcollection* into the folder that represents the disk.

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

11. Pointing to documents on the web

1. Open up your **tudor** collection, and in the **Gather** panel inspect the files you dragged into it. The first folder is *englishhistory.net*, which opens up to reveal *tudor*, and so on. The files represent a complete sweep of the pages (and supporting images) that constitute the Tudor section to the *englishhistory.net* web site. They were downloaded from the web in a way that preserved the structure of the original site. This allows any page's original URL to be reconstructed from the folder hierarchy.

2. □ In the **Design** panel, select the **Document Plugins** section, then select the **plugin HTMLPlug** line and click <**Configure Plugin**>. A popup window appears. Locate the **file_is_url** option (about halfway down the first block of items) and switch it on. Click <**OK**>.

   Setting this option to the HTMLPlug means that Greenstone sets an additional piece of metadata for each document called URL, which gives its original URL.

   It is important that the files gathered in the collection start with the web domain name (*englishhistory.net* in this case). The conversion process will not work if you dragged over the *tudor* folder, because this will set URL metadata to something like

   > http://tudor/englishistory.net/tudor/...

   rather than

   > http://englishhistory.net/tudor/...

   If you have copied over the tudor folder previously, delete it and make a fresh copy. Drag the *tudor* folder in the right-hand side of the **Gather** panel on to the trash can in the lower right corner. Then obtain a fresh copy of the files starting with the *englishhistory.net* folder, by opening *tudor* on the left-hand side and dragging its contents across.

3. □ To make use of the new URL metadata, the icon link must be changed to serve up the original URL rather than the copy stored in the digital library. Go to the **Design** panel, select the **Format Features** section and edit the **VList** format statement by replacing

   > [link][icon][/link]

   with

   > [weblink][webicon][/weblink]

   Click <**Replace Format**> to commit the change.

4. □ Switch to the **Create** panel and **build** and **preview** the collection. The collection behaves exactly as before, except that when you click a document icon your web browser retrieves the original document from the web (assuming it is still there by the time you do this exercise!). If you are working offline you will be unable to retrieve the document.

12. Downloading files from the web

The Greenstone Librarian Interface's Download panel allows you to download individual files, parts of websites, and indeed whole websites, from the web.

1. Start a new collection called **webtudor**, and base it on the **tudor** collection.

2.□ In a web browser, visit *http://englishhistory.net*, follow the link to *Tudor England*, and click <**enter**>. You should be at the URL

    http://englishhistory.net/tudor/contents.html

    This is where we started the downloading process to obtain the files you have been using for the **tudor** collection.

3.□ You could do the same thing by copying this URL from the web browser, pasting it into the **Download** panel, and clicking the <**Download**> button. However, several megabytes will be downloaded, which might strain your network resources—or your patience! For a faster exercise we focus on a smaller section of the site. In the **Download** panel, enter this URL

    http://englishhistory.net/tudor/monarchs/edward6.html

    into the **Source URL** box. There are several options that govern how the download process proceeds. To copy the *monarchs* section of the website, select **Only mirror files below this URL**. If you don't do this, the downloading process will follow links to other areas of the *englishhistory.net* website and grab those as well.

4.□ Now click <**Download**>. A progress bar appears in the lower half of the panel that reports on how the downloading process is doing.

    *More detailed information can be obtained by clicking <**View Log**>. The process can be paused and restarted as needed, or stopped altogether by clicking <**Close**>. Downloading can be a lengthy process involving multiple sites, and so Greenstone allows additional downloads to be queued up. When new URLs are pasted into the Source URL box and <**Download**> clicked, a new progress bar is appended to those already present in the lower half of the panel. When the currently active download item completes, the next is started automatically.*

5.□ Downloaded files are stored in a top-level folder called **Downloaded Files** that appears on the left-hand side of the **Gather** panel. You may not need all the downloaded files, and you choose which you want by dragging selected files from this folder over into the collection area on the right-hand side, just like we have done before when selecting data from the *sample_files* folder. In this example we will include everything that has been downloaded.

    Select the **englishhistory.net** folder within **Downloaded Files** and drag it across into the collection area.

6.□ Switch to the **Create** panel to **build** and **preview** the collection. It is smaller than the previous collection because we included only the *monarchs* files. However, these now represent the latest versions of the documents. Since you based your **webtudor** collection on **tudor** it includes the modified *[weblink][webicon][/weblink]* format, so the new collection also links back to the original web documents.

13. Enhanced collection of HTML files

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 *monarchs* folder (a subfolder of *tudor*). Set its **dc.Subject and Keywords** metadata to **Tudor period|Monarchs**. (For brevity, we refer to this metadata element in future simply as **dc.Subject**.) The vertical bar ("|") is a hierarchy marker. Selecting a *folder* and using the **Append** button to set its 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.

2. Repeat for the *relative* and *citizens* folder, setting their **dc.Subject** metadata to **Tudor period|Relatives** and **Tudor period|Citizens** respectively. Note that the hierarchy appears in the **All Previous Values** area.

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

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 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.

*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** metadata. Then an index is assigned to each subcollection.*

Partitioning the full-text index based on metadata values

8. 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).

9. Switch to *Library Systems Specialist* mode by going to **Preferences** (on the *File* menu) and clicking <**Mode**>. Read about the other modes too. Note that the mode appears in the title bar.

10. 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** metadata contains the word *Monarchs*.

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

12. Having defined the subcollections, we partition the index into corresponding parts. Click the <**Assign Partitions**> tab. Select the first subcollection and give it the name **monarchs**; click <**Add Partition**>. Repeat for the other three subcollections, naming their partitions **relatives**, **citizens** and **others**. **Build** and **preview** the collection.

13. 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.

14. 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. Type **all** into the **Partition Name** and select all four subcollections by checking their boxes.

15. To ensure that the *all* 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.

16. Search for a common term (like *the*) in all five index partitions, and check that the numbers add up.

17. Return to *Librarian* mode, using **Preferences** (on the *File* menu).

Adding a hierarchical phrase index (PHIND)

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

19. 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>**.

20. **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**.

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.

Controlling the building process

21. 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**. In fact, you will find that the collection now contains 5 documents (not 3 as you specified: for technical reasons the number you give to **maxdocs** is an approximate value.)

22. Preview the newly rebuilt collection's **titles a–z** 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.

14. Learning about formats and macros

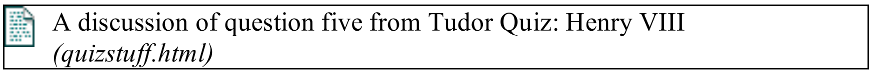
Format statements and macro files allow you to customize the appearance of Greenstone collections. They are very powerful, but complex and hard to learn. This tutorial exercise gives an introduction to the facilities they provide.

Experimenting with format statements

1. Open up your **tudor** collection, go to the **Design** 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** remains blank 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}{[dls.Title],[dc.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 a–z* 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—even baroque—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 a–z* 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 hierarchy browser, beneath the subject appears a mysterious "()". What is printed on these bookshelf nodes 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 **Design** panel, the **Choose Feature** menu (just above **Affected Component** menu) is blank. 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; when it's blank, the **VList** specification applies throughout. 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.

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

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

and click <**Add Format**>.

4. Now go to the **Create** panel and click <**Preview**>. 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.

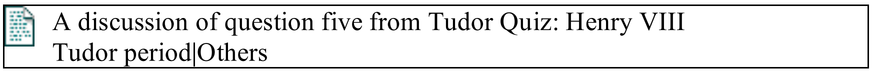
5. Let's change the search results format so that *dc.Subject and Keywords* metadata is displayed here instead of the filename. In the **Choose Feature** menu (under **Format Features** on the **Design** panel), scroll down to the item **Search** and select it. Change the **HTML Format String** box below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

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

and click <**Add Format**>.

6. To insert the **[dc.Subject]**, position the cursor at the appropriate point and investigate the **Variables** dropdown menu below—the one that says **[Text]**. Make it say **[dc.Subject]** and click **Insert** to insert this into the **HTML Format String**. This menu shows all the things that you can put in square brackets in the format statement. The only exception is **[Text]**, which gives the full text of the document, and can only be used when **DocumentText** is the **Affected Component**.

7. Now go to the **Create** panel and click <**Preview**>. 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>
```

and click **Replace Format**. Again, you can insert the items in square brackets by selecting them from the **Variables** dropdown box (don't forget to click **Insert**).

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. Go to the **Create** panel, click <**Preview**>, and examine the subject hierarchy again to see the effect of your changes.

Collection-specific macros

The appearance of all pages produced by Greenstone is governed by macro files, which reside in the folder C:\Program Files\Greenstone\macros. *The* garish *example collection is a version of* *the* demo *collection with bizarre layout and coloring. Now we apply the same bizarre layout and coloring to the* tudor *collection.*

10. Go to the folder *C:\Program Files\Greenstone\macros*. Copy the file *garish.dm* (do not *cut* it; it must remain in this folder, otherwise Greenstone will not start up). Now go to your collection folder *C:\Program Files\Greenstone\collect\tudor* and create a new folder in there called *macros*. Paste *garish.dm* into that new folder, and change its name to *extra.dm*. The overall effect is that you have created a new file *C:\Program Files\Greenstone\collect\tudor\macros\extra.dm* with the contents of *garish.dm*. Make sure that the new file is called *extra.dm* and not *extra.dm.dm*, which Windows sometimes tends to do.

11. Now edit this file *extra.dm* using WordPad or Notepad. Search for the string *[c=garish]* and remove this string wherever it appears. Lines starting with # are comment lines; you don't have to remove the string from these (though it won't do any harm). There are 11 other occurrences of *[c=garish]*: you must remove them all.

12. Go to the **Create** panel and click <**Preview**>. The content of your collection remains the same, but its appearance has changed completely—for example, all the pages are pink! To learn about how to control these changes, go to the documented example collection called *Garish version of demo collection*, and read about it.

A small but important enhancement to Greenstone has been made since the garish collection was written. Instead of using the [c=garish] *macro argument to restrict the macros to apply to a certain collection, you can now put collection-specific macros in the* macros *directory of the collection, in a file called* extra.dm. *In fact, this is what you have just done.*

General macros

You can also use macros to completely change the appearance of your Greenstone site. Like the above exercise, what follows is just a lead-in to illustrate what is possible and show you where to look to achieve different kinds of effects.

13. Exit from the Librarian Interface, since it is concerned with individual collections and we are now dealing with the site as a whole.

14. Go to the folder *C:\Program Files\Greenstone\etc* and edit the file called *main.cfg*. This is Greenstone's main configuration file, and contains a list of the macros that will be loaded in on startup. One of them, *home.dm*, dictates how the Greenstone home page will look, which is specified in the file *C:\Program Files\Greenstone\macros\home.dm*. This *macros* folder contains an alternative version, called *yourhome.dm*, which is not currently being used. To use it instead, in *main.cfg* change the string *home.dm* to *yourhome.dm*.

15. Now restart Greenstone (just the Greenstone Digital Library will do, rather than the Greenstone Librarian Interface). You will find that the appearance of the home page has changed completely.

16. Instead of substituting *yourhome.dm* for *home.dm* in the file *main.cfg*, you could have simply edited *home.dm* and left *main.cfg* as it is. However, we wanted to preserve *home.dm* so that you could revert to your original Greenstone home page! Do this now by editing *main.cfg* and changing the string *yourhome.dm* back to *home.dm*. You will need to re-start Greenstone for this to take effect.

To learn how more about macros, read Customizing the Greenstone User Interface*, an illustrated guide to customizing the user interface, by Allison Zhang of the Washington Research Library Consortium, available at http://www.wrlc.org/dcpc/UserInterface/interface.htm.*

15. A bibliographic collection

1. Start a new collection called **Beatles Bibliography**. Enter the requested information and make it a **New Collection**. There is no need to include any metadata sets because the metadata extracted from the MARC records will appear as extracted metadata.

2. □ In the **Gather** panel, open the *marc* folder, drag **locbeatles50.marc** into the right-hand pane and drop it there. A popup window asks whether you want to add **MARCPlug** to the collection to process this file. Click <**Add Plugin**>, because this plugin will be needed to process the MARC records.

3. □ Remove the plugins **TextPlug** to **PSPlug** (**ZIPPlug**, **GAPlug** and **MARCPlug** remain). It is not strictly necessary to remove these redundant plugins, but it is good practice to include only plugins that are needed, to avoid accidentally including stray documents.

4. □ Now select **Browsing Classifiers** within the **Design** panel and **remove** the default classifier for **Source** metadata. In this collection all records are from the same file, so **Source** metadata, which is set to the filename, is not particularly interesting.

5. □ Switch to the **Create** panel, **build** the collection, and **preview** it. Browse through the **titles a–z** and view a record or two. Try searching—for example, find items that include **George Martin**.

6. □ Add an **AZCompactList** classifier for the **Subject** metadata. Select this item from the relevant menu of the **Browsing Classifiers** section of the **Design** panel and click **<Add Classifier>**. In the popup window, select **ex.Subject** as the metadata item, activate the **mingroup** option and set its field to **1**.

AZCompactList *is like* ***AZList,*** *except that terms that appear multiple times in the hierarchy are automatically grouped together and a new node, shown as a bookshelf icon, is formed. Setting* ***mingroup*** *to 1 means that the bookshelf appears even when there is just one item, and is done here to provide a more uniform display.*

7. □ **Build** the collection and **preview** the result.

8. □ Make each bookshelf node show how many entries it contains by appending this to the **Format Features** for **VList** format statement in the **Design** panel:

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

9. □ Click **<Replace Format>**, switch to the **Create** panel, and click **<Preview Collection>** (no need to build the collection again).

Adding fielded searching

10. In the **Design** panel select **Search Types** from the left-hand list and activate the **Enable Advanced Searches** options.

11. **Build** the collection once again, and **preview** the results. Notice that the collection's home page no longer includes a query box. (This is because the search form is too big to fit here nicely.) To search, you have to click **search** in the navigation bar. Note that the *Preferences* page has changed to control the advanced searching options.

To finish off the collection, brand it with an image that will be used to represent the collection on the Greenstone page, and appear at the top of each page of the collection

Branding a collection with an image

12. From the **General** section of the **Design** panel, click the **<Browse>** button next to the label **URL to 'about page' icon** and use the resulting popup file browser to access the folder *sample_files\marc*. Select *beatles_logo.jpg* and click <**Open**>.

    Greenstone copies the image into your collection area, so the collection will still work when the CD-ROM is removed from the drive.

13. Repeat this process for the **URL to 'home page' icon**, selecting the same image.

14. Now **build** the collection and **preview** it.

16. Looking at a multimedia collection

1. Copy the entire folder

   sample_files→beatles→advbeat_large

   (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 *advbeat_large* in there.

2. □ If the Greenstone Digital Library Local Library Server is already running, re-start it by clicking the CD icon on the task bar and then pressing *Restart Library*. If not, start it up by selecting *Greenstone Digital Library* from the *Start* menu.

3. □ Explore the Beatles collection. Note how the *browse* button divides the material into seven different types. Within each category, the documents have appropriate icons. Some documents have an audio icon: when you click these you hear the music (assuming your computer is set up with appropriate player software). Others have an image thumbnail: when you click these you see the images.

4. □ Look at the *titles a–z* browser. Each title has a bookshelf that may include several related items. For example, *Hey Jude* has a cover image, MP3 audio and MIDI versions, lyrics, and a discography item.

5. □ Observe the low quality of the metadata. For example, the four items under *A HARD DAY'S NIGHT* (under "H" in the *titles a–z* browser) have different variants as their titles. The collection would have been easier to organize had the metadata been cleaned up manually first, but that would be a big job. Only a tiny amount of metadata was added by hand—fewer than ten items. The original metadata was left untouched and Greenstone facilities used to clean it up automatically. (You will find below that this is possible but tricky.)

6. □□ In the Windows file browser, take a look at the files that makes up the collection, in the

   sample_files→beatles→advbeat_large→import

   folder. What a mess! There are over 450 files under seven top-level sub-folders. Organization is minimal, reflecting the different times and ways the files were gathered. For example, *html_lyrics* and *discography* are excerpts of web sites, and *cover_images* contains album covers in JPEG format. For each type, drill down through the hierarchy and look at a sample document.

17. Building a multimedia collection

We will proceed to reconstruct from scratch the Beatles collection that you have just looked at. We develop the collection using a small subset of the material, purely to speed up the repeated rebuilding that is involved.

1. Start a new collection (*File→New*) called **small_beatles**, basing it on the default "New Collection." (Basing it on the existing Advanced Beatles collection would make your life far easier, but we want you to learn how to build it from scratch!) Fill out the fields with appropriate information. Use the Dublin Core metadata set (set by default).

2. Copy the files provided in

   sample_files→beatles→advbeat_small

   into your new collection. Do this by opening up *advbeat_small*, selecting the eight items within it (from *cover_images* to *beatles_midi.zip*), and dragging them across. Because some of these files are in MP3 format you will be asked whether to include the **MP3 Plugin** in your collection. Click <**Add Plugin**>.

3. Change to the **Enrich** panel and browse around the files. There is no metadata—yet. Recall that you can double-click files to view them.

   (There are no MIDI files in the collection: these require more advanced customisation because there is no MIDI plugin. We will deal with them later.)

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

5. **Preview** the result.

Manually correcting metadata

6. You might want to correct some of the metadata—for example, the atrocious misspelling in the titles "MAGICAL MISTERY TOUR." These documents are in the discography section, with filenames that contain the same misspelling. Locate one of them in the **Enrich** panel. Notice that the extracted metadata element **ex.Title** is now filled in, and misspelt. You cannot correct this element, for it is extracted from the file and will be re-extracted every time the collection is re-built.

7. Instead, add **dc.Title** metadata for these two files: "Magical Mystery Tour." Change to the **Enrich** panel, open the discography folder and drill down to the individual files. Set the **dc.Title** value for the two offending items.

*Now there's a twist. The **dc.Title** metadata won't appear in titles a–z because the classifier has been instructed to use **ex.Title**. But changing the classifier to use **dc.Title** would miss out all the extracted titles! Fortunately, there's a way of dealing with this by specifying a list of metadata names in the classifier.*

8. Change to the **Design** panel and select the **Browsing Classifiers** section. Double-click the **Title** classifier (the first one) to edit its configuration settings.

   - Type "dc.Title," before the *ex.Title* in the metadata box—i.e. make it read

     dc.Title,ex.Title

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

   Extracted metadata is unreliable. But it is very cheap! On the other hand, manually assigned metadata is reliable, but expensive. The previous section of this exercise has shown how to aim for the best of both worlds by using extracted metadata but correcting it when it is wrong. While this may not satisfy the professional librarian, it could provide a useful compromise for the music teacher who wants to get their collection together with a minimum of effort.

Browsing by media type

9. First let's remove the **AZList** classifier for filenames, which isn't very useful, and replace it with a browsing structure that groups documents by category (discography, lyrics, audio etc.). Categories are defined by manually assigned metadata.

   - Change to the **Enrich** panel, select the folder *cover_images* and set its **dc.Format** metadata value to "Images". Setting this value at the folder level means that all files within the folder inherit it.
   - Repeat the process. Assign "Discography" to the *discography* folder, "Lyrics" to *html_lyrics*, "MARC" to *marc*, "Audio" to *mp3*, "Tablature" to *tablature_txt*, and "Supplementary" to *wordpdf.*
   - Switch to the **Design** panel and select the **Browsing Classifiers** section.
   - Delete the **ex.Source** classifier (the second one).
   - Add an **AZCompactList** classifier. Select **dc.Format** as the metadata field and specify "Category" as the **buttonname**.

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

10. Greenstone has no pre-defined button for "Category", so it appears in the navigation bar as text. It does, however, have a button for *browse* (it's used in the Beatles collection you looked at in Part I).

    - Go back to the **AZCompactList** classifier for **dc.Format**. Click the **sort** checkbox, and leave **Title** in the adjacent text box: this will make the classifier display documents in alphabetical order of title. Also, specify "Browse" as the **buttonname**.

    You will need to build the collection for this to take effect.

Suppressing dummy text

11. Alongside the Audio files there is an MP3 icon, which plays the audio when you click it, and also a text document that contains some dummy text. This isn't supposed to be seen, but to suppress it you have to fiddle with a format statement.

    - Change to the **Design** panel and select the **Format Features** section.
    - Ensure that **VList** is selected, and make the changes that are highlighted below. You need to insert three lines into the first line, and delete the second line.

      Change:

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

      to this:

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

    - Then click <**Replace Format**>:

    To make this easier for you we have prepared a plain text file that contains the new text. In WordPad open the following file:

    sample_files→beatles→format_tweaks→audio_tweak.txt

    (Be sure to use WordPad rather than Notepad, because Notepad does not display the line breaks correctly.) Place it in the copy buffer by highlighting the text in WordPad and selecting Edit→Copy. Now move back to the Librarian Interface, highlight all the text that makes up the current VList format statement, and use Edit→Paste to transform the old statement to the new one. Remember to press <**Replace Format**> when finished.

Preview the result. If you are using the Greenstone Local Library server, change to the **Create** panel and click <**Preview Collection**>, which causes the local library server to rescan the format statements. You do not need to build the collection again because format statements are only used by the runtime system.

However, you may need to click the browser's <**Reload**> button to force it to re-load the page.

12. While we're at it, let's remove the source filename from where it appears after each document.

    - In the VList format feature, delete the text that is highlighted below:

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

Don't forget to click <**Replace Format**> after all this work! **Preview** the result (you don't need to build the collection.)

Using AZCompactList rather than AZList

13. There are sometimes several documents with the same title. For example, *All My Loving* appears both as lyrics and tablature (under *ALL MY LOVING*). The *titles a–z* browser might be improved by grouping these together under a bookshelf icon. This is a job for an **AZCompactList**.

    - Change to the **Design** panel and select the **Browsing Classifiers** section.
    - Remove the **Title** classifier (at the top)
    - Add an **AZCompactList** classifier, and enter **dc.Title,ex.Title** as its metadata.
    - Activate **mingroup** and set it to 1. This gives a uniform appearance by creating a bookshelf for every title.
    - Finish by pressing <**OK**>.
    - Move the new classifier to the top of the list (**Move Up** button).

Build the collection again and **preview** it. Both items for *All My Loving* now appear under the same bookshelf. However, many entries haven't been amalgamated because of non-uniform titles: for example *A Hard Day's Night* appears as four different variants. We will learn below how to amalgamate these.

Making bookshelves show how many items they contain

14. Make the bookshelves show how many documents they contain by inserting a line in the VList format statement in the **Design** panel:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
```

You will find this text in *format_tweaks→show_num_docs.txt*, which can be copied and pasted in as before. Don't forget to click <**Replace Format**>.

Preview the result (you don't need to build the collection.)

15. Now turn to the images. Dummy documents are displayed here too. First change to the **Enrich** panel, open the folder *cover_images* and add **dc.Title** metadata, assigning to each of the ten documents the title of the corresponding album. Remember, you can double-click a file to view it.

16. To suppress the dummy documents, change the **VList** format statement in the **Design** panel again by adding the two highlighted lines, and the close curly bracket:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  {If}{[dc.Format] eq 'Images',
    [srclink][thumbicon][/srclink],
    [link][icon][/link]}}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
```

17. Add a **Phind** browsing classifier that sources its phrases from *Title* and *text* (the default setting).

18. To complete the collection, use the browse button of **URL to 'about page' icon** in the **General** section of the **Design** panel to select the following image: *advbeatles_large→images→flick4.gif.*

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

Note how we assigned dc.Format *metadata to all documents in the collection with a minimum of labour. We did this by capitalizing on the folder structure of the original information. Even though we complained earlier about how messy this folder structure is, you can still take advantage of it when assigning metadata.*

In the next exercise we incorporate the MIDI files. Greenstone has no MIDI plugin (yet). But that doesn't mean you can't use MIDI files! We also clean up the titles a–z *browser.*

To do this we must put the Librarian Interface into a different mode. The interface supports four levels of user: Library Assistants*, who can add documents and metadata to collections, and create new ones whose structure mirrors that of existing collections;* Librarians*, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions);* Library Systems Specialists*, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl programs); and* Experts*, who can perform all functions.*

So far you have mostly been operating in Librarian *mode. We switch to* Library Systems Specialist *mode for the next exercise.*

Using UnknownPlug

19. To switch modes, click *File→Preferences→Mode* and change to **Library Systems Specialist**. Note from the description that appears that you need to be able to formulate regular expressions to use this mode fully. That is what we do below.

20. **UnknownPlug** is a useful generic plugin. It knows nothing about any given format but can be tailored to process particular document types—like MIDI—based on their filename extension, and set basic metadata.

    - add **UnknownPlug**;
    - activate its *process_extension* field and set it to *mid* to make it recognize files with extension *.mid*;
    - Set *file_format* to "MIDI" and *mime_type* to "audio/midi".

    In this collection, all MIDI files are contained in the file *beatles_midi.zip*. **ZIPPlug** (already in the list of default plugins) is used to unpack the files and pass them down the list of plugins until they reach UnknownPlug.

21. **Build** the collection and **preview** it. Unfortunately the MIDI files don't appear as Audio under the *browse* button. That's because they haven't been assigned **dc.Format** metadata.

    - Back in the **Enrich** panel, click on the file *beatles_midi.zip* and assign its **dc.Format** value to "Audio"—do this by clicking on "Audio" in the **All Previous Values** list. All files extracted from the Zip file inherit its settings.

Cleaning up a title browser using regular expressions

22. Next we return to our *titles a–z* browser and clean it up. The aim is to amalgamate variants of titles by stripping away extraneous text. For example, we would like to treat "ANTHOLOGY 1", "ANTHOLOGY 2" and "ANTHOLOGY 3" the same for grouping purposes. To achieve this:

- Go to the Title **AZCompactList** under **Browsing Classifiers** on the **Design** panel;
- Activate **removesuffix** and set it to:

```
(?i)(\\s+\\d+)|(\\s+[[:punct:]].*)
```

Build the collection and **preview** the result. Observe how many more times similar titles have been amalgamated under the same bookshelf. Test your understanding of regular expressions by trying to rationalize the amalgamations. (Note: *[[:punct:]]* stands for any punctuation character.) The icons beside the Word and PDF documents are not the correct ones, but that will be fixed in the next format statement.

The previous exercise was done in Librarian Systems Specialist *mode because it requires the use of regular expressions, something librarians are not normally trained in.*

One powerful use of regular expressions in the exercise was to clean up the titles a–z *browser. Perhaps the best way of doing this would be to have proper title metadata. The metadata extracted from HTML files is messy and inconsistent, and this was reflected in the original titles a–z browser. Defining proper title metadata would be simple but rather laborious. Instead, we have opted to use regular expressions in the* AZCompactList *classifier to clean up the title metadata. This is difficult to understand, and a bit fiddly to do, but if you can cope with its idiosyncrasies it provides a quick way to clean up the extracted metadata and avoid having to enter a large amount of metadata.*

Using non-standard macro files

To put finishing touches to our collection, we add some decorative features

23. Using your Windows file browser outside Greenstone, locate the folder

    sample_files→beatles→advbeat_large

24. Copy the *images* and *macros* folders located there into your collection's top-level folder. (It's OK to overwrite the existing *images* folder: the image in it is included in the folder being copied.) The *images* folder includes some useful icons, and the *macros* folder defines some macro names that use these images. To see the macro definitions, take a look by using a text editor to open the file *extra.dm* in the *macros* folder.

Using different icons for different media types

25. Re-Edit your **VList** format statement to be the following

```
<td valign=top>
 {If}{[numleafdocs],[link][icon][/link]}
 {If}{[dc.Format] eq 'Lyrics',[link]_iconlyrics_[/link]}
 {If}{[dc.Format] eq 'Discography',[link]_icondisc_[/link]}
 {If}{[dc.Format] eq 'Tablature',[link]_icontab_[/link]}
 {If}{[dc.Format] eq 'MARC',[link]_iconmarc_[/link]}
 {If}{[dc.Format] eq 'Images',[srclink][thumbicon][/srclink]}
 {If}{[dc.Format] eq 'Supplementary',[srclink][srcicon][/srclink]}
 {If}{[dc.Format] eq 'Audio',[srclink]{If}{[FileFormat] eq
'MIDI',_iconmidi_,_iconmp3_}[/srclink]}
</td>
<td>
{If}{[numleafdocs],([numleafdocs])}
</td>
<td valign=top>
[highlight]
{Or}{[dc.Title],[Title],Untitled}
[/highlight]
</td>
```

26. The complete statement is in the file *format_tweaks→multi_icons.txt*.

27. **Preview** your collection as before. Now different icons are used for discography, lyrics, tablature, and MARC metadata. Even MP3 and MIDI audio file types are distinguished. If you let the mouse hover over one of these images a "tool tip" appears explaining what file type the icon represents in the current interface language (note: *extra.dm* only defines English and French).

Changing the collection's background image

28. Open your collection's *macros* folder and locate the *extra.dm* file within it. **Right-click** on it. If prompted, select **WordPad** as the application to open it with.

29. The file content is fairly brief, specifying only what needs to be overridden from the default behaviour for this collection. In WordPad, near the top of the file you should see:

    ```
    _httpiconchalk_ {_httpcimages_/beat_margin.gif}
    _widthchalk_ {1800}
    _heightchalk_ {68}.
    ```

    Use copy and paste on these three lines to make this part of the file look like:

    ```
    # Original statements
    #_httpiconchalk_ {_httpcimages_/beat_margin.gif}
    #_widthchalk_ {1800}
    #_heightchalk_ {68}
    _httpiconchalk_ {_httpcimages_/tile.jpg}
    _widthchalk_ {22}
    _heightchalk_ {22}
    ```

    A hash (#) at the start of line signals a comment, and Greenstone ignores the following text. We use this to comment out the original three statements and replace them with modified lines. It is useful to retain the original version in case we need to restore the original lines at a later date. These three lines relate to the background image used. The new image *tile.jpg* was also in the *images* folder that was copied across previously.

30. Within **WordPad**, save *extra.dm*.

31. **Preview** the collection's home page. The page background is now the new graphic.

    Other features can be altered by editing the macro files—for example, the headers and footers used on each page, and the highlighting style used for search terms (specify a different colour, use bold etc.).

32. If you want to you can reverse the most recent change you made by commenting out the three new lines added (add #) and uncommenting the original three (delete # character). Remember to save the file. To undo all the customized changes made, delete the content of the *macros* and *images* folders.

Building a full-size version of the collection

33. To finish, let's now build a larger version of the collection. To do this:

    - Close the current collection.
    - Start a new collection called *advbeat_large.*
    - Base this new collection on *small_beatles.*
    - Copy the content of *sample_files*→*beatles*→*advbeat_large*→*import* into this newly formed collection. Since there are considerably more files in this set of documents the copy will take longer.
    - **Build** the collection and preview the result. (If you want the collection to have an icon, you will have to add it from the **Design** panel.)

Adding an image collage browser

34. Switch to the **Design** panel and select the **Browsing Classifiers** section. Pull down the **select classifier to add** menu and select **Collage**. Click <**Add Classifier**>. There is no need to customize the options, so click <**OK**> at the bottom of the resulting popup.

35. Now change to the **Create** panel and **build** and **preview** the collection.

18. Scanned image collection

Here we build a small replica of Niupepa, the Maori Newspaper collection, using five newspapers taken from two newspaper series. It allows full text searching and browsing by title and date. When a newspaper is viewed, a preview image and its corresponding plain text are presented side by side, with a goto page *navigation feature at the top of the page.*

The collection involves a mixture of plug-ins, classifiers, and format statements. The bulk of the work is done by PagedImgPlug*, a plug-in designed precisely for the kind of data we have in this example. For each document, an "item" file is prepared that specifies a list of image files that constitute the document, tagged with their page number and (optionally) accompanied by a text file containing the machine-readable version of the image, which is used for full text searching. Three newspapers in our collection (all from the series* Te Whetu o Te Tau*) have text representations, and two (from* Te Waka o Te Iwi*) have images only. Item files can also specify metadata. In our example the newspaper series is recorded as* ex.Title *and its date of publication as* ex.Date*. This metadata is extracted as part of the building process.*

1. Start a new collection called **Paged Images** and fill out the fields with appropriate information: it is a collection sourced from an excerpt of Niupepa documents; the only metadata used is document title and date, and these are extracted from the "item" files included in the source documents so no metadata set need be stipulated.

2. Add **PagedImgPlug** and switch on its **screenview** configuration option by checking the box. The source images we use were scanned at high resolution and are large files for a browser to download. The *screenview* option generates smaller screen-resolution images of each page when the collection is built.

3. In the **Gather** panel, open the *niupepa\sample_items* folder in *sample_files* and drag it into your collection on the right-hand side.

4. Some of the files you have just dragged in are text files that contain the text extracted from page images. We want these to be processed by **PagedImgPlug**, not **TEXTPlug**. Switch to the **Design** panel and delete **TEXTPlug**. While you are at it, you could tidy things up by deleting **HTMLPlug**, **EMAILPlug**, **PDFPlug**, **RTFPlug**, **WordPlug**, and **PSPlug** as well, since they will not be used.

5. Now go to the **Create** panel, **build** the collection and **preview** the result. Search for *waka* and view one of the titles listed (all three appear as *Te Whetu o Te Tau*). Browse by *titles a–z* and view one of the *Te Waka o Te Iwi* titles.

This collection was built with Greenstone's default settings. You can locate items of interest, but the information is less clearly and attractively presented than in the full Niupepa collection.

Grouping documents by series title and displaying dates within each group

Under titles a–z *documents from the same series are repeated without any distinguishing features such as date. It would be better to group them by series title and display dates within each group. This can be accomplished using an* AZCompactList *classifier rather than* AZList*, and tuning the* VList *format statement.*

1. In the **Design** panel, under the **Browsing Classifiers** section, delete the **AZList** classifiers for *ex.Source* and *ex.Title*.

2. Now add **AZCompactList** for *ex.Title* and **DateList** for *ex.Date*.

3. **Modify** the format statement for **VList**. Find the part of the default statement that says

   ```
   {If}{[ex.Source],<br><i>([ex.Source])</i>}
   ```

   and change it to

   ```
   {If}{[ex.Date],: [ex.Date]}
   ```

   This has the effect of displaying the extracted date information, if present.

4. At the end of this format statement, where is says:

   ```
   </td>
   ```

append

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

As a consequence of using the AZCompactList *classifier, bookshelf icons appear when titles are browsed. This revised format statement has the effect of specifying in brackets how many items are contained within a bookshelf. It works by exploiting the fact that only bookshelf icons define* [numleafdocs] *metadata.*

Suppressing dummy text

When you reach a newspaper, only its associated text is displayed. When either of the Te Waka o Te Iwi *newspapers is accessed, the document view presents the message* This document has no text. *No scanned image information (screen-view resolution or otherwise) is shown, even though it has been computed and stored with the document. This can be fixed by a format statement that modifies the default behaviour for* DocumentText.

5. Staying within the **Format Features** section of the **Design** panel, under "Choose Feature" select **DocumentText**. Its HTML format string is empty, triggering the default behaviour of displaying the document's plain text, or, if there is none, "This document has no text". Change this to:

   ```
   <center>
     <table width=_pagewidth_>
       <tr>
         <td valign=top>[srclink][screenicon][/srclink]</td>
         <td>[Text]</td>
       </tr>
     </table>
   </center>
   ```

   (available as niupepa\doc_tweak.txt)

   Including [screenicon] has the effect of embedding the screen-sized image generated by switching the screensize option on in PagedImgPlug. It is hyperlinked to the original image by the construct [srclink]...[/srclink].

6. Switch to the **Create** panel**; build** and **preview** the revised collection.

7. If you like, add a logo and change the background as you have done before. You will find a suitable image in the file *niupepa\images*, that is activated through *macros\extra.dm*.

In the collection you have just built, newspapers are grouped by series title, and dates are supplied alongside each one to distinguish it from others in the same series. Users can browse chronologically by date, and when a newspaper page is viewed a preview image is shown on the left that displays the original high-resolution version when clicked, accompanied on the right by the plain-text version of that newspaper (if available).

19. Open Archives Initiative (OAI) collection

This exercise explores service-level interoperability using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). So that you can do this on a stand-alone computer, we do not actually connect to the external server that is acting as the data provider. Instead we have provided an appropriate set of files that take the form of XML records produced by the OAI-PMH protocol.

One of Greenstone's documented example collections is sourced over OAI. This exercise takes you through the steps necessary to reconstruct it. (Note: this example is a collection of images: you will not be able to build it unless ImageMagick is installed on your computer.) You may wish to take a look at the documented example collection OAI demo *now to see what this exercise will build.*

1. Start a new collection called **OAI Service Provider**. Fill out the fields with appropriate information. You can leave the default metadata set as Dublin Core, although we do not make use of it.
2. In the **Gather** panel, navigate to the *sample_small* folder in *sample_files/oai*. Drag this folder into the collection and drop it there.
3. During the copy operation, a popup window appears asking whether to add **OAIPlug** to the list of plug-ins used in the collection, because the Librarian Interface has not found an existing plug-in that can handle this file type. Press the **<Add Plugin>** button to include it.

When files are copied across like this, the Librarian Interface studies each one and uses its filename extension to check whether the collection contains a corresponding plug-in. Up until now, the answer has always been yes, so all file transfers have proceeded without interruption. This time, however, no plug-in in the list is capable of processing the OAI file records that are copied across (they have the file extension .oai).

Sometimes there is more than one plug-in that could process a file—for example, the .xml extension is used for many different XML formats. The popup window, therefore, offers a choice of all possible plug-ins that matched. It is normally easy to determine the correct choice. If you wish, you can ignore the prompt (click <Don't Add Plugin>), because plug-ins can be added later, in the Document Plugins section of the Design panel.

4. You need to configure the Image plug-in. In the **Design** panel, select the **Document Plugins** section, then select the **plugin ImagePlug** line and click <**Configure Plugin**>. In the resulting popup window locate the **screenviewsize** option, switch it on, and type the number 300 in the box beside it to create a screen-view image of 300 pixels. Click <**OK**>.
5. Now switch to the **Create** panel and **build** and **preview** the collection.

Like other collections we have built by relying on Greenstone defaults, the end result is passable but can be improved. The next steps refine the collection using the metadata harvested by OAI-PMH into the .oai files.

6. In the **Browsing Classifiers** section of the **Design** panel, delete the two **AZList** classifiers (*ex.Title* and *ex.Source*).
7. Add an **AZCompactList** classifier based on **ex.Subject** metadata.
8. Now add an **AZCompactList** classifier based on **ex.Description** metadata. In its configuration panel select **mincompact** = **1**, **maxcompact** = **10** and **buttonname** = **Captions**.
9. In the **Search Indexes** section of the **Design** panel, delete all indexes and add a new one called "captions" based on *ex.Description* metadata.
10. **Build** collection and **preview** it.

Tweaking the presentation with format statements

11. In the **Design** panel, select **Format Features**. First replace the **VList** format statement with this:

```
<td>
{If}{[numleafdocs],[link][icon][/link],[link][thumbicon]
     [/link]}
</td>
<td valign=middle>
  {If}{[numleafdocs],[Title],<i>[Description]</i>}
</td>
```

You will find this text in the file *vlist_tweak.txt* in the *oai* folder of *sample_files.* Remember to press <**Replace Format**> when finished

This format statement customizes the appearance of vertical lists such as the search results and captions lists to show a thumbnail icon followed by Description metadata. Greenstone's default is to use extracted metadata, so [Description] is the same as [ex.Description].

12. Next, select **DocumentHeading** from the **Choose Feature** pull-down list and make its format statement (which is currently blank) read

    ```
    <h3>[Subject]</h3>
    ```

    The document heading appears above the detach *and* no highlighting *buttons when you get to a document in the collection. By default* DocumentHeading *displays the document's* ex.Title *metadata. In this particular set of OAI exported records, titles are filenames of JPEG images, and the filenames are particularly uninformative (for example,* 01dla14*). You can see them in the* ***Enrich*** *panel if you select an image in* sample_small→oai→ JCDLPICS→srcdocs *and check its filename and* ex.Title *metadata. The above format statement displays* ex.Subject *metadata instead.*

13. Finally, you will have noticed that where the document itself should appear, you see only *This document has no text.* To rectify this, select **DocumentText** in the **Choose Feature** pull-down list and use the following as its format statement (which is currently blank) (this text is in *doctxt_tweak.txt* in the *format_tweaks* folder mentioned earlier):

    ```
    <center><table width=_pagewidth_ border=1>
    <tr><td colspan=2 align=center>
    <a href=[OrigURL]>[screenicon]</a></td></tr>
    <tr><td>Caption:</td><td> <i>[Description]</i> <br>
    (<a href=[OrigURL]>original [ImageWidth]x[ImageHeight] [ImageType]
    available</a>)
    </td></tr>
    <tr><td>Subject:</td><td> [Subject]</td></tr>
    <tr><td>Publisher:</td><td> [Publisher]</td></tr>
    <tr><td>Rights:<td> [Rights]</td></tr>
    </table></center>
    ```

    This format statement alters how the document view is presented. It includes a screen-sized version of the image that hyperlinks back to the original larger version available on the web. Factual information extracted from the image, such as width, height and type, is also displayed.

14. Format statements are processed by the runtime system, so the collection does not need to be rebuilt for these changes to take effect. Switch to the **Design** panel and press <**Preview Collection**> to see the changes.

To expedite building, this collection contains fewer source documents than the pre-built version supplied with the Greenstone installation. However, after these modifications, its functionality is the same.

20. Downloading over OAI

The previous exercise did not obtain the data from an external OAI-PMH server. This missing step is accomplished by running a command-line program. To do this, your computer must have a direct connection to the Internet—being behind a firewall may interfere with the ability to download the information.

15. **Save** your collection. Note its directory name, which should be *oaiservi* (it appears in the title bar of the Librarian Interface), and **quit** the Librarian Interface.

16. Perform the first four steps of the “Moving a collection from Greenstone to DSpace” exercise: open a command window, change directory to where Greenstone is installed, run *setup.bat*,and change directory once again, this time into *collect\oaiservi*, the folder containing the OAI Service Provider collection you built in the last exercise.

17. In a text editor, open the collection’s configuration file, which is in *oaiservi\etc\collect.cfg*. Add the following line (all on one line):

```
acquire OAI -src rocky.dlib.vt.edu/~jcdlpix/
    cgi-bin/OAI1.1/jcdlpix.pl -getdoc
```

Although the position of this line is not critical, we recommend that you place it near the beginning of the file, after the public and creator lines but before the index line. Save the file and quit the editor.

18. Delete the contents of the collection’s *import* folder. This contains the canned version of the collection files, put there during the previous exercise. Now we want to witness the data arriving anew from the external OAI server.

19. Back at the DOS prompt, run *perl –S importfrom.pl oaiservi*

Greenstone will immediately set to work and generate a stream of diagnostic output. The importfrom.pl *program connects to the OAI data provider specified in collection configuration file (it does this for each “acquire” line in the file) and exports all the records on that site.*

20. The downloaded files are saved in the collection’s import folder. Once the command is finished, everything is in place and the collection is ready to be built. Confirm you have successfully acquired the OAI records by rebuilding the collection.

21. Exporting a collection as METS

1. In the Greenstone Librarian Interface, open the **Tudor** collection.

To be able to substitute METSPlug *for* GAPlug *you need to be in* Expert *mode.*

2.□ Click *File→Preferences→Mode* and change to *Expert* mode.

3.□ Switch to the **Design** panel select **Document Plugins**. Remove **GAPlug** from the list of plug-ins and add **METSPLug**.

4.□ Now change to the **Create** panel, locate the options for the import process and set *–saveas* to *METS*. Import options are not available unless you are in *Expert* mode.

5.□ Rebuild the collection.

6.□ In your Windows file browser, locate the *archives* folder for the Tudor collection. For each document in the collection, Greenstone has generated two files: *docmets.xml*, the core METS description, and *doctxt.xml*, a supporting file. (Note: unless you are connected to the Internet you will be unable to view *doctxt.xml* in your web browser, because it refers to a remote resource.) Depending on the source documents there may be additional files, such as the images used within a web page. One of MET’s many features is the ability to reference information in external XML files. Greenstone uses this to tie the content of the document, which is stored in the external XML file *doctxt.xml*, to its hierarchical structure, which is described in the core METS file *docmets.xml*.

22. Moving a collection from DSpace to Greenstone

7. If you have just done the previous exercise the Greenstone Librarian Interface will already be in *Expert* mode. Otherwise, change to *Library System Specialist* (or *Expert*)mode (using File→Preferences), because you will need to change the order of plug-ins in the **Design** panel.

8. Start a **new collection** called **StoneD** and fill out its fields appropriately. Leave the metadata set at Dublin Core, the default.

9. Switch to the **Design** panel and select the **Document Plugins** section on the left-hand side. **Remove TEXTPlug**, **EMailPlug** and **HTMLPlug**. Strictly speaking we do not need to remove these, however it reduces clutter.

10. Now add **DSpacePlug**. Leave the plugin options at their defaults and press <**OK**>.

11. Using the up and down arrows, **Move** the position of **DSpacePlug** to above **PDFPlug** and below **GAPlug**.

12. Now add **MP3Plug**, with the default configuration options. Its position in the plug-in pipeline need not be changed.

13. In the **Gather** panel, locate the folder **sample_files\dspace\exported_docs**. It contains five example items exported from a DSpace institutional repository. Copy them into your collection by dragging them over to the right-hand side of the panel.

14. **Build** the collection and **preview** it to see the basic defaults exhibited by a DSpace collection.

If you browse by titles a–z*, you will find 7 documents listed, though only 5 items were exported from DSpace. Two of the original items had alternative forms in their directory folder. DSpace plug-in options control what happens in such situations: the default is to treat them as separate Greenstone documents.*

*Below we use a plug-in option (*first_inorder_ext*) to fuse the alternative forms together. This option has the effect of treating documents with the same filename but different extensions as a single entity within a collection. One of the files is viewed as the primary document—it is indexed, and metadata is extracted from it if possible—while the others are handled as "associated files."*

The first_inorder_ext *option takes as its argument a list of file extensions (separated by commas): the first one in the list that matches becomes the primary document.*

15. Select **DSpacePlug** and click <**Configure Plugin**>. Switch on its configuration option **first_inorder_ext**. Set its value to *pdf,doc,mp3* in the popup window that appears and press **<OK>**.

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

There are now only 5 documents, because only one version of each document has been included—the primary version.

The DSpace exported files contain Dublin Core metadata for title and author (amongst other things).

Adding indexing and browsing capabilities to match DSpace's

17. In the **Design** panel, select **Search Indexes**. Delete the *ex.Title* and *ex.Source* indexes, and add one for **dc.Title** called "titles" and another for **dc.Contributor** called "authors".

18. Staying within the **Design** panel, select **Browsing Classifiers** and **delete** both **AZList** classifiers (*ex.Title* and *ex.Source*). Add an **AZList** classifier for **dc.Title** and another for **dc.Contributor**.

19. Now select the **Format Features** section of the **Design** panel and replace the **VList** format statement with this:

```
<td valign=top>
  [srclink]{or}{[thumbicon],[srcicon]}[/srclink]
</td>
<td valign=top>
[highlight]{or}{[dls.Title],[dc.Title],[ex.Title],Untitled}[/highlight]
  {If}{[ex.Source],<br>
  <i>([ex.Source])</i>}{If}{[equivlink],<br>
  Also available as:[equivlink]}
</td>
```

You will find this text in the file *format_tweak.txt* in the *dspace* folder of *sample_files*, and you can copy and paste this. Remember to press <**Replace Format**> when finished.

20. **Build** collection once again and **preview** it.

There are still only 5 documents, but against some of the entries—for example, Interview with Bob Dylan*—appears the line "Also available as," followed by icons that link to the alternative representations.*

23. Moving a collection from Greenstone to DSpace

In this exercise you export a Greenstone collection in a form suitable for Dspace. It is possible to do this from the Librarian Interface's File menu, which contains an item called Export … *that allows you to export collections in different forms. However, to gain a deeper understanding of Greenstone, we perform the work by invoking a program from the Windows command-line prompt. This requires some technical skill; if you are not used to working in the command-line environment we recommend that you skip this exercise.*

Using Greenstone from the command line

21. Open a DOS window to access the command-line prompt. This facility should be located somewhere within your Start→Programs menu, but details vary between different Windows systems. If you cannot locate it, select Start→Run and enter *cmd* in the popup window that appears.

22. In the DOS window, move to the home directory where you installed Greenstone. This is accomplished by something like:

    ```
    cd C:\Program Files\Greenstone
    ```

23. Type:

    ```
    setup.bat
    ```

    to set up the ability to run Greenstone command-line programs.

24. Change directory into the collection you built in the last exercise:

    ```
    cd collect\stoned
    ```

    Even though the collection name used capital letters the directory generated by the Librarian Interface is all lowercase.

25. Run the following command to export the collection using the DSpace import/export format:

    ```
    perl –S export.pl –saveas DSpace –removeold stoned
    ```

Exporting in Greenstone is an additive process. If you ran the export.pl command once again, the new files exported would be added—with different folder names—to those already in the export folder. For the kind of explorations we are conducting we might re-run the command several times. The –removeold option deletes files that have previously been exported.

26. This command has created a new subfolder, *collect\stoned\export*. Use the file browser to explore it. In it are the files needed to ingest this set of documents into Dspace.

You could equally well run the export.pl command on a different Greenstone collection and transfer the output to a DSpace installation by using DSpace's batch-import facility.