Lab 1: Installing, browsing, building

1.1. 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 and click **<OK>**.

3. On the welcome screen, 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 [CD-ROM] UNAIDS Library 2.0 (Standard Version).

   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 for 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:

    1. 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.

    2. 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?
- How many documents are there that mention *Australia* in the title?
- How many top-level subject categories are there?
- What does AAVP stand for?
- What does AIDS stand for?
- Considering lower case variants only, how many times does the word "condom" appear in the collection?
  How many times for "condoms"?
- If case sensitivity does not matter, how many times does the word "condom" appear in the collection?
  How many times for "condoms"?
- If word endings are ignored, how many times does "condom" and variants such as "condoms" appear in the collection?
- How many *chapters* contain some variations of the word "condom"?
  Does this make it a useful search term?
- What year saw the first reported case of AIDS in New Zealand?

1.2. Installing Greenstone

Installing Greenstone on a Windows system

There are various ways of getting Greenstone:

1. From a UNESCO CD-ROM (version 2.70) (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 tutorial exercises and a set of **sample files** to be used for these exercises. CD-ROMs with Greenstone version 2.62 or earlier also include 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. CD-ROMs with version 2.70 or later now come with reader's interfaces in all available languages. Textual images have been removed from the interface; they are now done using CSS (Cascading Style Sheets). The Greenstone Language Pack is no longer needed. Instead, these CD-ROMs come with the **Classic Interface Pack**, which contains the old text images for use with a backwards compatibility macro file.*

   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 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 (Greenstone 2.70 and earlier), the **Language Pack** (Greenstone 2.62 and earlier) and **Classic Interface Pack** (Greenstone 2.63 and later). 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 **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. (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.)

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>**.

Installing Ghostscript on a Windows system

If you wish to do advanced conversion of PDF and Postscript documents (as described in exercise **Enhanced PDF handling**), you will need to install Ghostscript. If you are installing from a Greenstone CD-ROM you will automatically be prompted for this; the procedure is analogous to that described above for ImageMagick. If not, you will need to download Ghostscript from http://www.cs.wisc.edu/~ghost/ (follow the link to the current stable release).

If you are not sure whether you will need Ghostscript or not, you might as well install it anyway—it will do no harm.

1.3. 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 in **Installing Greenstone**. 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 (2.62 and earlier)

*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 (glp-x.xx.exe/glp-x.xx-linux.bin/gli-x.xx-macOSx.command). 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. Run the executable file (double click it on Windows); 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.

Enabling other languages (2.63 and later)

If you have downloaded Greenstone from the web, then all the languages will be enabled by default. However, if you have installed Greenstone from a UNESCO CD-ROM, then only English, French, Spanish and Russian will be enabled.

10. To enable a new language, edit the file *greenstone etc main.cfg* . Look for the appropriate "Language" line, and uncomment it (i.e. remove the # from the start). Check that the required encoding is also enabled.

    For example, suppose that we want to enable Turkish. The "Language" line for Turkish looks like:

    ```
    #Language shortname=tr longname=Turkish default_encoding=windows-1254
    ```

    To enable it, we remove the #, i.e. make it look like:

    ```
    Language shortname=tr longname=Turkish default_encoding=windows-1254
    ```

    The default encoding for Turkish is windows-1254. So we look for the windows-1254 Encoding line:

    ```
    Encoding shortname=windows-1254 "longname=Turkish (Windows-1254)" map=win1254.ump
    ```

    This is already enabled (no # at the start) so we don't need to do anything else.

Installing the Classic Interface Pack (2.63 and later)

Greenstone now comes with all languages enabled. The generated HTML uses text + CSS rather than images for navigation bar, home, help, preferences buttons etc. The classic interface pack is not needed if you want to use Greenstone in another language. It is only needed if you want to revert back to the old style HTML with text images. This may be useful if you have customized your Greenstone, or if you require compatibility with Netscape 4.

11. Locate the Classic Interface Pack (gcip-x.xx.zip). This may be on the CD-ROM from which you installed Greenstone, or you may have to download it from http://www.greenstone.org.

12. The classic interface pack is a zip file containing the old text images, such as classifier buttons. Unzip the zip file into the images directory of your Greenstone installation.

13. Enable the use of the old-style macros by editing *greenstone etc main.cfg* : replace "nav_css.dm" with "nav_ns4.dm" in the "macrofiles" list.

14. Restart the Greenstone Digital Library. It should now be using the old text images.

1.4. Building a small collection of HTML files

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

Running the Greenstone Librarian Interface

1. Start the Greenstone Librarian Interface:

   Start All Programs Greenstone Digital Library Software v2.73 Greenstone Librarian Interface

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

Starting a new collection

2. Start a new collection within the Librarian Interface:

   File New...

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

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

   Collection title: Small Tudor
   Description of content: A smaller version of the Tudor collection.

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

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

Adding documents to the collection

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

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

6. Since this is our first collection, we won't complicate matters by manually assigning metadata or altering the collection's design. Instead we rely on default behaviour. So pass directly to the **Create** panel by clicking its tab.

Building the collection

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

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

9. Click the **<Preview Collection>** button to look at the end result. This loads the relevant page into your web browser (starting it up if necessary).

Viewing the extracted metadata

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

documents in the collection.

11. Presently there is no manually assigned metadata, but the act of building the collection has extracted metadata from the documents. Double click the *simple_html* folder to expand its content. Then single-click *aragon.html* to display all its metadata in the right-hand side of the panel. The initial fields, starting "dc.", are empty. These are Dublin Core metadata fields for manually entered data.

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

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

Setting up a shortcut in the Librarian interface

14. To set up a shortcut to the source files, in the **Gather** panel navigate to the folder in your local file space that contains the files you want to use—in our case, the *sample_files* folder. Select this folder and then right-click it, and choose **Create Shortcut** from the menu. In the **Name** field, enter the name you want the shortcut to have, or accept the default *sample_files*. Click **<OK>**. Close all the folders in the file tree in the left-hand pane, and you will see the shortcut to your source files.

1.5. A collection of Word and PDF files—Part A

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

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

2. Copy all the 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.

Viewing the extracted metadata

4. Again, this collection contains no manually assigned metadata. All the information that appears—title and filename—is extracted automatically from the documents themselves. Because of this the quality of some of the title metadata is suspect.

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

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

7. The extracted Title metadata for some documents is incorrect. For example, the Titles for *pdf01.pdf* and *word03.doc* (the same document in different formats) have missed out the second line. The Title for *pdf03.pdf* has the wrong text altogether.

In exercise 2.1 we correct some of this incorrect metadata by manually adding Dublin Core Title metadata.

1.6. 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), based on the default **-- New Collection --**.

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. (This material is from Marilee Hanson's Tudor England Collection at http://englishhistory.net/tudor.html, distributed with her permission.)

4. Switch to the **Create** panel and click **<Build Collection>**.

5. When building has finished, **preview** the collection.

Extracting more metadata from the HTML

6. The browsing facilities in this collection (*Titles* and *Filenames*) are based entirely on extracted metadata. Return to the **Enrich** panel in the Librarian Interface and examine the metadata that has been extracted for some of the files.

7. Many HTML documents contain metadata in `<meta>` tags in the `<head>` of the page. Open up the *englishhistory.net tudor monarchs boleyn.html* file by navigating to it in the tree on the left hand side, and double clicking it. This will open it in a web browser. View the HTML source of the page (**View Source** in Internet Explorer, **View Page Source** in Mozilla). You will notice that this page has *page_topic, content* and *author* metadata.

8. By default, **HTMLPlug** only looks for Title metadata. Configure the plugin so that it looks for the other metadata too. Switch to the **Design** panel and select the **Document Plugins** section. Select the **plugin HTMLPlug** line and click **<Configure Plugin...>**. A popup window appears. Switch on the **metadata_fields** option, and set the value to

   ```
   Title,Author,Page_topic,Content
   ```

   Make sure that you have copied this exactly, with no spaces. Click **<OK>**.

9. Switch to the **Create** panel and **rebuild** the collection. Go back to the **Enrich** panel and look at the extracted metadata for some of the HTML files in *englishhistory.net tudor monarchs*. The new metadata should now be visible.

Blocking the stray images

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.)

10. Switch back to the **Document Plugins** section of the **Design** panel. 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! Configure **plugin HTMLPlug** again, scroll down the page to locate the **smart_block** option, and switch it off.

11. **Rebuild** 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

12. Switch to the **Gather** panel and in the right-hand side open *englishhistory.net tudor* .

13. 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.

14. Change the **Show Files** menu to **Images**. Again, the files shown above alter.

15. 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.

1.7. 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 **-- New 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.html

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

3. Go to the **Download** panel by clicking its tab. There are four download types listed on the left hand side. For this exercise, we only use the **Web** type. Make sure this is selected in the list.

   Enter this URL

   http://englishhistory.net/tudor/citizens/

   into the **url** box. There are several other options that govern how the download process proceeds. To see a description of an option, hover the mouse over it and a tooltip will appear. To copy just the *citizens* section of the website, switch on the **below** option by checking its box and set the **depth** option to 1. If you don't do this (or if you miss out the terminating "/" in the URL), the downloading process will follow links to other areas of the *englishhistory.net* website and grab those as well.

4. If your computer is behind a firewall or proxy server, you will need to edit the proxy settings in the Librarian Interface. Click the **<Preferences...>** button. Switch on the **Use proxy connection?** checkbox. Enter the proxy server address and port number in the **Proxy Host:** and **Proxy Port:** boxes. Click **<OK>**.

5. Now click **<Download>**. If you have set proxy information in **Preferences...**, a popup will ask for your user name and password. Once the download has started, 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 **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.*

6. 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.

7. Switch to the **Create** panel to **build** and **preview** the collection. It is smaller than the previous collection because we included only the *citizens* files. However, these now represent the latest versions

of the documents.

1.8. Enhanced Word document handling

The standard way Greenstone processes Word documents is to convert them to HTML format using a third-party program, wvWare. This sometimes doesn't do a very good job of conversion. If you are using Windows, and have Microsoft Word installed, you can take advantage of Windows native scripting to do a better job of conversion. If the original document was hierarchically structured using Word styles, these can be used to structure the resulting HTML. Word document properties can also be extracted as metadata.

1. In your digital library, preview the **reports** collection. Look at the HTML versions of the Word documents and notice how they have no structure-they have been converted to flat documents.

Using Windows native scripting

2. In the Librarian Interface, open up the **reports** collection. Switch to the **Design** panel and select the **Document Plugins** section on the left-hand side. Double click the **WordPlug** plugin and switch on the **windows_scripting** option.

   In the **Search Indexes** section, check the **section** checkbox to build the indexes on section level as well as document level.

3. **Build** the collection. You will notice that the Microsoft Word program is started up for each Word document—the document is saved as HTML from Word itself, to get a better conversion. **Preview** the collection. In the **Titles** list, notice that *word03.doc* and *word06.doc* now have a book icon, rather than a page icon. These now appear with hierarchical structure.

   The default behaviour for **WordPlug** with **windows_scripting** is to section the document based on "Heading 1", "Heading 2", "Heading 3" styles. If you open up the *word03.doc* or *word06.doc* documents in Word, you will see that the sections use these Heading styles.

   Note, to view style information in Word, you can select **Format Styles and Formatting** from the menu, and a side bar will appear on the right hand side. Click on a section heading and the formatting information will be displayed in this side bar.

4. Some of the documents do not use styles (e.g. *word01.doc*) and no structure can be extracted from them. Some documents use user-defined styles. **WordPlug** can be configured to use these styles instead of Heading 1, Heading 2 etc. Next we will configure **WordPlug** to use the styles found in *word05.doc*.

Modes in the Librarian Interface

5. The Librarian Interface can operate in four modes. Go to **File Preferences... Mode** and see the four modes and what functionality they provide access to. **Librarian** is the default mode.

6. Change the mode to **Library Systems Specialist** because you will need to use regular expressions to set up the style options in the next part of the exercise.

Defining styles

7. Open up *word05.doc* in Word (by double-clicking on it in the **Gather** pane), and examine the title and section heading styles. You will see that various user-defined header styles are set such as:

   - *ManualTitle*: Title of the manual
   - *ChapterTitle*: Level 1 section heading
   - *SectionHeading*: Level 2 section heading
   - *SubsectionHeading*: Level 3 section heading
   - *AppendixTitle*: Appendix section title

8. In the **Document Plugins** section of the **Design** panel, select **WordPlug** and click **<Configure Plugin...>**. Four types of header can be set which are:

   - `level1_header (level1Header1|level1Header2|...)`
   - `level2_header (level2Header1|level2Header2|...)`
   - `level3_header (level3Header1|level3Header2|...)`
   - `title_header (titleHeader1|titleHeader2|...)`

   These header options define which styles should be considered as title, level 1, level 2 and level 3 styles.

   Set the options as follows (spaces in the Word styles are removed when converting to HTML styles, and these options must match the HTML styles):

   ```
   level1_header:(ChapterTitle|AppendixTitle)
   level2_header: SectionHeading
   level3_header: SubsectionHeading
   title_header: ManualTitle
   ```

   *If you can't see these options in the **WordPlug** configuration pane, check that you are in **Library Systems Specialist** mode as described above.*

   Once these are set, click **<OK>**.

9. Close any documents that are still open in Word, as this can prevent the build process from completing correctly.

10. **Build** the collection and **preview** it. Look in particular at *word05.doc*. You will see that this document is now also hierarchically structured.

    If you have documents with different formatting styles, you can use `(...|...)` to specify all of the different styles.

Removing pre-defined table of contents

11. If you look at *word05.doc* and *word06.doc* you will see that it now has two tables of contents. One is generated by Greenstone based on the document's styles, the other was already defined in the Word document. WordPlug can be configured to remove predefined tables of contents and tables of figures. The tables must be defined with Word styles in order for this to work.

12. To remove the tables of contents and figures from *word06.doc* and the table of contents from *word05.doc*, switch on the **delete_toc** option in **WordPlug**. Set the **toc_header** option to `(MsoToc1|MsoToc2|MsoToc3|MsoTof|TOA)`. In this document, the table of contents and list of figures use these four style names. Click **<OK>**.

13. **Build** and **preview** the collection. Both *word05.doc* and *word06.doc* should now have only one table of contents.

14. Switch the Librarian Interface back to **Librarian** mode (**File Preferences... Mode**).

Extracting document properties as metadata

15. Word document properties can be extracted as metadata. By default, only the Title will be extracted. Other properties can be extracted using the **metadata_fields** option.

16. In the **Enrich** panel, look at the metadata that has been extracted for *word05.doc* and *word06.doc*. Now open the documents in Word and look at what properties have been set (**File Properties**). They have Title, Author, Subject, and Keywords properties. **WordPlug** can be configured to look for these properties and extract them.

17. In the **Design** panel, under **Document Plugins**, configure **WordPlug** once again. Switch on the configuration option **metadata_fields**. Set the value to

```
Title,Author<Creator>,Subject,Keywords<Subject>
```

This will make **WordPlug** try to extract Title, Author, Subject and Keywords metadata. Title and Subject will be saved with the same name, while Author will be saved as Creator metadata, and Keywords as Subject metadata.

18. Make sure you have closed all the documents that were opened, then **rebuild** the collection.

19. Look at the metadata for the two documents again in the **Enrich** panel. You should now see ex.Creator and ex.Subject metadata items. This metadata can now be used in display or browsing classifiers etc.
