Lab 10: Miscellaneous

10.1. 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 five 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 **Source 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 **Only filese below URL** option by checking its box and set the **Download 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 **<Configure Proxy...>** 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.

10.2. Editing metadata sets

GEMS (Greenstone Editor for Metadata Sets) can be used to modify existing metadata sets or create new ones. GEMS is launched from the Librarian Interface when you want to create a new metadata set, or edit an existing one. In this exercise, we run GEMS outside of the Librarian Interface.

Running GEMS

1. Start the Greenstone Editor for Metadata Sets (GEMS), for versions 2.81 and greater:

   Start → **All Programs** → **Greenstone-2.81** → **Metadata Set Editor**

   and for versions below 2.81:

   Start → **All Programs** → **Greenstone Digital Library Software** → **Greenstone Editor for Metadata Sets**

2. GEMS starts up with no metadata set loaded. You can start a new set, or open an existing one, from the **File** menu.

Creating a new metadata set

3. In this exercise, we will create a new metadata set. In order to save time, we will base it on an existing one: Development Library Subset. From the **File** menu, select **File** → **New...**. A popup window appears: **New Metadata Set**. Fill in the fields. Use "My Metadata Set" for the **Metadata set title:**, "my" for the **Metadata set namespace:**, and select "Development Library Subset Example Metadata" from the **Base this metadata set on:** drop down list. Click **<OK>**.

4. The new metadata set will be displayed. The left hand side list the elements (and sub-elements, if any) for the set, and the right hand side displays the set or element attributes. Since the new set was based on the Development Library Subset metadata set, it already contains all the elements from that set.

Adding a new element to a metadata set

5. Right click on **My Metadata Set** in the left hand tree (or in the blank space in the left hand side) and choose **Add Element** from the menu that appears. In the popup window, type "Category" for the new element name, and click **<OK>**. The new element will appear in the list.

6. In the right hand side, the default attributes will appear for the new element. "Label" and "definition" are used in the Librarian Interface when displaying metadata elements and their descriptions (the "definition" is shown as additional text for the element). These attributes can be set in multiple languages.

7. Save the new metadata set by **File** → **Save**, then close the GEMS by **File** → **Exit**.

10.3. Building and searching with different indexers

Greenstone supports three indexers **MG**, **MGPP** and **Lucene**.

MG is the original indexer used by Greenstone which is described in the book **"Managing Gigabytes"**. It does section level indexing and compression of the source documents. **MG** is implemented in C.

MGPP is re-implementation of **MG** that provides word-level indexes and enables proximity, phrase and field searching. **MGPP** is implemented in C++ and is the default indexer for new collections.

Lucene (http://lucene.apache.org/) is java-based full-featured text indexing and searching system developed by Apache. It provides a similar range of search functionality to MGPP with the addition of single-character wildcards and range searching. It was added to Greenstone to facilitate incremental collection building, which **MG** and **MGPP** can't provide.

Build with Lucene

1. Start a new collection (**File** → **New...**) called **Demo Lucene** and base it on the **Greenstone demo (demo)** collection, fill out its fields appropriately.

2. In the **Gather** panel, click **Documents in Greenstone Collections** and click **Greenstone demo (demo)**, it will show the documents in the **Greenstone demo** collection. Drag all 11 folders underneath *Greenstone demo (demo)* into the new collection.

   *If you haven't installed the **Greenstone demo (demo)** collection yet, you can download the demo. zip file from the link above, unzip it and put it into the collect folder in your Greenstone installation.*

3. Go to the **Enrich** panel, look at the metadata that associated with each directory. Go to the **Search Indexes** section in the **Design** panel. The **MG indexer** is in use because the original **Greenstone Demo** collection, which this collection is based on, uses **MG indexer**.

4. Click the **Change...** button at the right top corner of the panel. A new window will pop up for selecting the Indexers. After selecting an indexer, a brief description will appear in the box below. Select Lucene and click **OK**. Please note that the **Assigned Indexes** has changed accordingly.

5. **Build** and **preview** the collection.

Search with Lucene

6. Lucene provides single letter and multiple letter wildcards and range searching. The query syntax could be quite complicated (for more information please see http://lucene.apache.org/java/docs/queryparsersyntax.html). Here we will learn how to use the wildcards while constructing queries.

7. ***** is a multiple letter wildcard. To perform a a multiple letter wildcard search, append ***** to the end of the query term. For example, *econom** will search for words like *econometrics*, *economist*, *economical*, *economy*, which have the common part *econom* but different word endings.

8. To perform a single letter wildcard search, use **?** instead. For example, search for *economi??* will only match words that have two and only two letters left after *economi*, such as *economist*, *economics*, and *economies*.

9. Please note that stopwords are used by default with Lucene indexer, so search for words like *the* will match 0 document. There is also a message on the search page saying that such words are too common and were ignored.

Build with MGPP

10. Start a new collection called **Greenstone Demo MGPP** and also base it on the **Greenstone demo (demo)**.

11. In the **Gather** panel, drag all the 11 folders from → *Greenstone demo (demo)* into the new collection.

12. Go to the **Search Indexes** section in the **Design** panel, click the **Change...** button and select **MGPP**. Click **OK**. Check the **Assigned Indexes** has changed accordingly.

13. There are three options at the bottom of the panel — **Stem**, **Casefold** and **Accent fold**. Notice that **Stem** and **Casefold** are enabled. Once an option is enabled, it will also appear in the collection's **PREFERENCES** page.

14. In the **Indexing Levels** section, also select **section**.

15. **Build** and **preview** the collection.

Search with MGPP

16. MGPP supports stemming and casefolding. By default search in collections built with MGPP indexer is set to **whole word must match** and **ignore case differences**. So search *econom* will return 0 document. Search for *fao* and *FAO* return the same result — 78 word counts and 9 matched documents.

    Go to the **PREFERENCES** page by click the **PREFERENCES** button at the top right corner. You can see that the **Word endings:** option is set to **whole word must match** and the **Case differences:** option is set to **ignore case differences**.

17. Sometimes we may want to ignore word endings while searching so as to match different variations of the term. Go to the **PREFERENCES** page and change the **Word endings:** option from **whole word must match** to **ignore word endings**. Click the **set preferences** button. Click **Search**. This time try search for *econom* again, 9 documents are found.

    Please note that word endings are determined according to the third-party stemming tables incorporated in Greenstone, not by the user. Thus the searches may not do precisely what is expected, especially when cultural variations or dialects are concerned. Besides, not all languages support stemming, only English and French have steming at the moment.

    Go to the **PREFERENCES** page and change back to **whole word must match** to avoid confusion later on. Click the **set preferences** button.

18. Sometimes we may want to search the exact term, that is, differentiate the upper cases from lower cases. Set the **Case differences:** option from **ignore case differences** to **upper/lower case must match**. Click the **set preferences** button. Click **Search**. Now try search for *fao* and *FAO* respectively this time, notice the difference in the results?

    Go back to the **PREFERENCES** page and change the **Case differences:** option back to **ignore case differences** to avoid confusion later on. Click **set preferences** button.

Use search mode hotkeys with query term

MGPP have several hotkeys to set search modes for a query term. These hotkeys explicitly set the ***Word endings:*** *option and the* ***Case differences:*** *option for the query being constructed.*

19. **#s** and **#u** are hotkeys for the **Word endings:** option. Appending **#s** to a query term will specifically enable the **ignore word endings** function. For example, try search for *econom#s*, 7 documents are found, which is the same as in step 17. Remember that we have set it back to **whole word must match**. This means using hotkeys will override the current preference settings.

20. Appending **#u** to a query term will explicitly set the current search to **whole word must match**.

    Note that using hotkeys will only affect that query term. That is, hotkeys are used per term. For example, if a query expresssion contains more than one terms, some tems can have hotkeys and others not, and the hotkeys can be different for different terms. This provides a fine-grained control of the query, whereas changing settings in the **PREFERENCES** page will affect the query as a whole.

21. Hotkeys **#i** and **#c** control the case sensitivity. Appending **#i** to a query term will explicitly set the search to **ignore case differences** (ie. case insensitive).

22. On the contrary, appending **#c** will specifically turn off the casefolding, that is, **upper/lower case must match**. For example, search for *fao#c* returns 0 document.

23. Finally, the hotkeys can also be used in combination. For example, you can append *#uc* to a query term so as to match the whole term (without stemming) and in its exact form (differentiate upper cases and lower cases).

A quick reference of the search mode hotkeys in MGPP

```
Word endings:

    #s    ignore word endings

    #u    whole word must match

Case differences:

    #i    ignore case differences

    #c    upper/lower case must match
```

Copyright © 2005 2006 2007 2008 2009 by the New Zealand Digital Library Project at the University of Waikato, New Zealand

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License."