Lab 7: Sharing collections with OAI-PMH

7.1. 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.

2. In the **Gather** panel, locate the folder *sample_files → oai → sample_small → 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.

   The files for this collection consist of a set of images (in *JCDLPICS → srcdocs*) and a set of OAI records (in *JCDLPICS*) which contain metadata for the images.

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. No plug-in in the list is capable of processing the OAI file records that are copied across (they have the file extension .oai), so the Librarian Interface prompts you to add the appropriate plug-in.

*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 **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.

   OAIPlug will process the OAI records, and assign metadata to the images, which are processed by **ImagePlug**.

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 set **mingroup** to **2**, **mincompact** to **1**, **maxcompact** to **10** and **buttonname** to **Captions**.

   Setting **mingroup** to 2 will mean that two or more documents with the same description will be grouped into a bookshelf; the default **mingroup** of 1 means that every document will get a bookshelf. **mincompact** and **maxcompact** control how many documents are grouped into each section of the horizontal A-Z list. In this case, each group can have as few as one document, and no more than ten.

9. In the **Search Indexes** section of the **Design** panel, delete all indexes and add a new one based on **ex.Description** metadata.

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

Tweaking the presentation with format statements

11. In the **Format** panel, select **Format Features**. First replace the **VList** format statement with the following (which can be copied from the file *vlist_tweak.txt* in the *sample_files → oai →format_tweaks* folder).

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

   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 change its format statement to:

```
<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 oai → JCDLPICS → srcdocs and check its **ex. Source** 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 (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. Click **<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.

7.2. Downloading over OAI

*The previous exercise did not obtain the data from an external OAI-PMH server. This missing step is accomplished either by running a command-line program or by using the **Download** panel in the Librarian Interface. This exercise shows you how to do this using both methods.*

Downloading using the Librarian Interface

1. In the Librarian Interface, switch to the **Download** panel. Select **OAI** from the list of download types on the left hand side.

2. In the **url** box, type in the following URL:

   http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI/jcdlpix.pl

3. We want to download the documents as well as the metadata, so tick the **get_doc** checkbox.

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.

6. Downloaded files are stored in a top-level folder called **Downloaded Files** that appears on the left-hand side of the **Gather** panel. These files can then be added to a collection.

Downloading using the command line

For command line downloading to work, your computer must have a direct connection to the Internet—being behind a firewall may interfere with the ability to download the information. You will need to use the Librarian Interface for downloading if you are behind a firewall.

7. Close the Librarian Interface.

   We will work with the OAI collection used in exercise **Open Archives Initiative (OAI) collection**. You may have noticed that its internal name is **oaiservi**.

8. In a text editor (e.g. WordPad), open the collection's configuration file, which is in *Greenstone → collect → oaiservi → etc → collect.cfg*. Add the following line (all on one line):

   ```
   acquire OAI -src http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI/
   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.

9. 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.

10. 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.

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

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

12. Type:

    ```
    setup.bat
    ```

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

13. Change directory into the folder containing the OAI Services Provider collection you built in the last exercise.

    ```
    cd collect\oaiservi
    ```

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

14. 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.

15. 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.