Greenstone 2.85 as OAI Server

The new Greenstone 2.85 features which facilitate the creation of institutional repositories and other open access collections:

1. OAI server
Your collections can easily be made available for remote harvesting using OAI-PMH protocol, which works silently in parallel with normal web access to the collections. All that you have to do is to add a bit of configuration data in the oai.cfg text file in the etc subdirectory under the Greenstone home directory. The data to specify is explained in comment lines in the above file. If the collections to be made available through OAI-PMH do not all use Dublin Core metadata or one of the two other standard OAI metadata sets, the oai.cfg file will need to contain mapping data to translate your metadata into one of the Greenstone OAI-PMH metadata sets (also explained in the comments to the oai.cfg file).

Starting from version 2.83, OAI-PMH support was provided, but there were a few inconsistencies and bugs in that version and in 2.84. In version 2.85, all official OAI-PMH validation criteria have been tested and satisfied; you will be able to validate your own OAI-PMH server using instructions given in the release notes. If you don't specify the urls for the associated documents in the metadata, the system can autmatically generate internal urls so that users can access the full documents from the harvested OAI records. You will also now be able to harvest OAI-PMH records and the associated documents residing in external Greenstone collections (in 2.84 harvesting worked to access information in non-Greenstone collections, but there was a bug which caused problems in harvesting from other Greenstone collections).

Much information is put up on the web without clear specification of the concerned intellectual property rights. Although this is not good practice in general, when activating the OAI server special care should be taken to ensure that your documents are really available under open access conditions (in the public domain or freely distributable and re-distributable under an open access license such as Creative Commons). Greenstone can only take care of the technical access - for legal and organisational considerations, prospective open access providers may consult, for example, the resources links of the EIFL Open Access programme (http://www.eifl.net/eifl-oa-resources [1]).

Once your OAI server is operational, to provide maximal international visibility for your open access collections you should register them in at least one (and ideally all) of the following: the ROAR directory (http://roar.eprints.org/ [2]), the OAI directory (http://www.openarchives.org/Register/BrowseSites [3]) and the OpenDOAR directory (http://www.opendoar.org/ [4]). It would also be very nice if you could confirm to the list that your server is operational, providing the url base address.

2. PDF metadata
Prior to version 2.83, reliable import of, and metadata extraction from, pdf files was limited to PDF versions 1.4 and earlier. Starting with 2.84 a new "PDF Box extension" has been available as a separate download to handle all PDF versions. This extension file need only be placed in

the ext subdirectory of Greenstone for the improved PDF handling
facilities to be operational (see the release notes). The PDF Box
extension has been further improved in version 2.85, so please be sure
to download, unzip and insert an up-to-date PDF Box extension for this
version, replacing the version of the file which you may have downloaded
for version 2.84.

By using the PDF Box extension, you can extract any metadata entered in
standard manner in a pdf file, i.e. the traditional pdf metadata
(Author, Title, Subject, Keywords) and/or the newer XMP format metadata
(including user defined fields). In general, we recommend that for users
interested in extracting PDF metadata, it is better to use the PDF Box
extension, even for pdf files in version 1.4 or earlier.

Using the PDF metadata extraction facility means that for PDF files
generated by the users with metadata included (either directly with a
tool like Acrobat, or by generating a PDF file from a package like Word
which would transfer Word metadata), these metadata can be automatically
incorporated into a Greenstone collection (without having to enter it in
GLI or compile a metadata.xml file). This could clearly be of interest
to open access applications, particularly when decentralized input is
being submitted.

There is a catch: the metadata extraction procedure may not work
flawlessly on recent version PDF files which are not "linearised"
(called Fast Web View in Acrobat). So linearised PDF files should be
used; the open source QPDF program (http://qpdf.sourceforge.net/ [5]) claims
to be able to linarise non-linearised PDF files, but this remains to be
confirmed in so far as Greenstone treatment is concerned. Feedback from
users on the PDF metadata extraction facility is most welcome.

3. Section handling for PDF files
For several years Greenstone has proposed a facility to automatically
generate internal section (chapter) information from a Word or html
document, but not for PDF file - this allows for finer searching and
table of contents display of the document. In the special case of Word
files Word must be installed in the computer in which a collection is
built
(http://wiki.greenstone.org/wiki/gsdoc/tutorial/en/enhanced_word.htm [6]).

An example collection (see
http://www.nzdl.org/gsdlmod?a=p&p=about&c=assocext-e [7]) has now been
prepared to show how this can be extended to PDF files. Included is an
explanation of how to build the collection in the following steps:
a. develop a Word version and a PDF version of the document (conversion
of the Word version to pdf or vice-versa);
b. make sure that the heading formats in Word are consistent with what
you want for sections and subsections;
c. import the Word file into Greenstone specifying the PDF file as an
associated file;
d. use the format statement guidance in the worked example to be able to
search on the document subsections and also display the hit terms in the
original PDF file (Word no longer needed after building - the collection
could in the meantime have been transferred to Linux).

This process has the disadvantage of relying on proprietary software. An
alternative, but labour intensive, method without Word would be to
import the pdf file into Greenstone, right click in the Gather view and
convert it to html, call an html editor and ensure that the section

information is correctly introduced, add the pdf again but as an
associated file (by setting the assoc-files parameter in HTMLPlugin),
then build and display as per the worked example.

More complete documentation is being developed for all of the above
techniques, and we will keep you informed on its progress.

To switch to version 2.85 from an earlier Greenstone version with
minimal risks, you could i) back up your collections, ii) install 2.85
in a new home directory (to specified to the installer), and iii) copy
the collect sub-directory from the old to the new version. If you are
presently using a recent previous version of Greenstone (2.8x), the
collections should be immediately available for use; if not,
particularly for collections built under older versions of Greenstone,
it should suffice to rebuild the collections under the new version. Any
problems can be addressed to this list or the main Greenstone users list.

If you want to transfer information on users and user groups, the
corresponding databases (users.gdb, key.gdb) should be copied from the
etc sub-directory in the old collection to the new one. Of course if you
have customised your previous version (main.cfg, style.css, macros,
etc.), the old versions should also be copied to the new installation.
When all is working perfectly, the old installation can be deleted.

Author: John Rose of Greenstone team

Source: ADLSN email support list archive, November 2011 [8]

Share / Save
[9]
Source URL: http://adlsn.org/greenstone-285-as-oai-server

Links:
[1] http://www.eifl.net/eifl-oa-resources
[2] http://roar.eprints.org/
[3] http://www.openarchives.org/Register/BrowseSites
[4] http://www.opendoar.org/
[5] http://qpdf.sourceforge.net/
[6] http://wiki.greenstone.org/wiki/gsdoc/tutorial/en/enhanced_word.htm
[7] http://www.nzdl.org/gsdlmod?a=p&p=about&c=assocext-e
[8] http://www.eifl.net/pipermail/adlsngreenstone/2011-November/subject.html
[9] http://www.addtoany.com/share_save?linkurl=http%3A%2F%2Fadlsn.org%2Fgreenstone-285-as-
oai-server&linkname=Printer-friendly%20PDF%20%7C%20African%20Digital%20Library%20Su
pport%20Network