Creating a TEI Shell from OCR

This page is stable, but sections are still in progress and/or being updated. Certain links to specific DLP project information are restricted to Digital Library Program staff.

A Perl script has been developed to concatenate text files produced as a result of OCR for TEI encoding. The script creates a TEI "shell" file that contains a TEI Header with boilerplate information particular to a given project and full text contained in anonymous block <ab> tags. Page breaks tags <pb> with corresponding ID attributes are automatically inserted. Encoders then update the headers and encode the texts according to specific project encoding guidelines.

Step 1: Download the Concatenation Script

1. Download the Perl script.
2. Upload the Perl (.pl) file to your home account on http://bleu.dlib.indiana.edu or other Unix server.
3. Change the file permissions by executing the command >> chmod 755 filename.pl

Step 2: Customize Concatenation Script per Project

Handy Hint

The teiHeader information will likely change from project to project. Update subroutine *print_header* to reflect the project specific teiHeader. Make sure that the corresponding ending tags in the print_footer are correct. Use backslash (\ ) as an escape character every time you want to print special characters like forward slash (/), quotes ("), dot (.), etc. Perl is very sensitive about these special characters.

1. Copy the existing script for modifications. For example:

```
cp ocr2tei.pl newfile.pl
```

2. Update teiHeader information in the subroutine *print_header*. For example:

```
print OUTF "<biblScope type=\"issue\">$issue<\/biblScope>\n"; (double quotes with backslash)
print OUTF "<biblScope type=\"issue\">$issue<\/biblScope>\n"; (single quotes; necessary when using
variables)
```

3. Update VAA number regular expression
    a. The code snippet below shows the regular expression used to extract the VAA number from the folder path provided while executing the script:

```
if ($dirname =\~ /(VAA\[0-9\]+)/)
     \{
        $id = $1;
     \}
```

    b. For example, the regular expression to update the VAA# for the IMH is as follows:

```
if ($dirname =\~ /(VAA\d\d\d\d-\d\d\d-\d)/)
     \{
        $id = $1;
     \}
```

Step 3: Execute the Concatenation Script

1. Use the command perl, followed by the path to the perl file, followed by the path to the folder containing the OCR text files. The resulting xml file will be placed in the VAA directory. For example:

```
perl ocr2tei.pl VAA0001/OCR
```

Another example: Assuming that the user is logged into bleu and the perl file is in his/her home directory, the following code will concatenate the OCR files for an Indiana Authors book. The xml file will be named VAA3893.xml and will be saved in the VAA3893 directory.

```
perl ocr2tei.pl ../../digitize/data/Indiana/Indiana_Authors/TIFFs_for_unencoded_books/VAA3893/OCR
```

Concatenation Scripts for Ongoing Projects

Indiana Magazine of History (.pl, P4)
Indiana Authors and Their Books (.pl, P4)
Brevier Legislative Reports (.pl, P5)