

Historic Pittsburgh Full-Text: Pre-Zoning for OCR Using PRView

Revision: 2.0

Last Revised: 2003-01

Note: Images that have a landscape orientation must be rotated prior to this processing.

Beginning the Batch Process

The pre-zoning is done with a program called *PRView*; but in order to speed the process, a Perl script is used to semi-automate the processing of each book. So in order to begin the process we must first run the Perl script:

Open a *MS-DOS* prompt. (*Start Menu* → *Programs* → *Command Prompt*)
At the command prompt type:

```
P:\ocr\bin\prview.pl
```

The script will request the directory you wish to process. You can actually either process an entire shipment directory or books, or a single book at a time. Most often, we work through one book at a time, and the following examples will be based upon processing one book. Type the *full path* to the directory, including the drive letter, the CD number, and the book number in the format specified in this example:

```
P:\ocr\cd9\00abc1234m
```

Include the ":\ " or it will fail to find the directory.

After a few moments the script will display the path to the book it expects to process, such as:

```
P:\ocr\cd9\00abc1234\v0000\i000
```

Type "y" to proceed.

The script will check the book to see if it has already been processed. If a book has already been completely processed, the script will display a message indicating this fact, and you should begin the process again with another book.

If the book has not been fully processed, the script will display the number of blank pages and the number of figures that have been processed of the total number of images with figures in the book.

You can answer:

```
yes to proceed,  
or exit to exit the program.
```

When the script starts to process a file it will display a message such as:

```
Processing P:\ocr\cd9\00aaa1224m\v0000\i000\03450567.tif
```

Configuring and Using the PRView Program

After a few moments the *PRView* interface should pop up with the image in it.

IMPORTANT: Each time you first run the script, check to make sure that the following settings are correct:

- *Image Process* → *Horizontal Line Removal*
- *Image Process* → *Vertical Line Removal*
- *Outputs* → *Formatted ASCII*

Note: Only these three settings should be checked; uncheck anything under the *View* and *Zoning* Menus.

- On the side menu Set *Accuracy Level* to 5
- Set *Lexical Check* to None
- Template job directories are correct

Capture the text to be OCR'd. If there is no text to be OCR'd capture an empty block and save both job and template.

After highlighting the region(s) to be processed use the *File* → *Save Template As* command and choose the directory *P:\ocr\template*.

Save the job by using the *OCR* → *Save the Job As* command to save the Job to the directory *P:\ocr\viewjob*.

IMPORTANT: If upon saving a template or a job the program says that one already exists **DO NOT** over write it. Save it by another name and append a few letters or numbers to make it unique.

After completing the process for the first image, close the program by choosing *File* → *Exit* (Alt F - x).

A new window will pop up with an image.

Capture the text to be OCRed.

File -> *Save Template* **OR** *Ctrl-S* to save the template

OCR-> *Save The Job* **OR** *Ctrl-J* to save the Job.

File -> *Exit* **OR** *Alt-F-X* to Exit the program.

The script will repeat this process until you get to the end of the directory (book). If there is another book to be processed you will be asked if you wish to proceed. Repeat the steps explained above for pre-zoning images.

Leaving the Program before Completing the CD

If you need to stop in the middle of the processing, type *Ctrl-C* at the prompt. One of the *PRView* interfaces will probably pop up before you are done. Just exit without saving anything.

If the CD has been partially processed the script will print several messages about completed files and directories before it pops up the *PRview* program.

Anomalies and Problems

All images should be rotated before getting to this phase. If you find an image that has not been rotated, skip the processing step and notify the system administrator with the full name and path of the image. Highlight a blank spot and save both template and job anyway.