



## Historic Pittsburgh Full-Text: Using the Quality Control Tool for Books

---

Revision: August 12, 2002

### Preparing a New Shipment of Images for Quality Control

1. When a new shipment of digitized books arrives in the DRL, load all the new CD-ROMs into

```
P:\ocr\shipment_no
```

You will need to change the permission on all of the files copied from the CDs, as they are "read only" by default. To do this, right-click on the shipment folder (in the location above), choose "Properties", and click the "read only" checkbox so that it is **unchecked**. Click the "Apply" button, and then "OK" to close the box. This permission should be inherited by all of the files and subdirectories within the directory.

2. Fill out the tracking sheet for each book. Make sure to write in the shipment number, the CD-ROM number and their location in the file system.

3. Although spreadsheets should have been checked for correctness before this stage of the process, check the document types on the spreadsheets before moving into the book directories.

a. Open a Command Prompt.

b. Change directories to the directory of tab delimited text versions of the spreadsheets that you want to evaluate. (For Historic Pittsburgh these are generally located on the *P:*/ drive in the directories called by the shipment name.)

c. Run

```
p:\ocr\bin\finddoctype.pl
```

in the shipment directory that is named according to the shipment number. This will run for a couple of minutes. It will evaluate all of the files with the **.txt** extension. A report called *doctype.report* will be deposited in the shipment directory.

- d. Open the *doctype.report* in a text editor. This lists suspicious doctypes that might not be processed correctly. If there are errors, the report lists the spreadsheet, the accession and native page number, and the doctype that is wrong. Correct the *Excel* spreadsheet in *hpbooksheets* and export as a tab delimited text file back into the shipment directory.
- e. When fully corrected, move to step 4.
4. Run the following script to move spreadsheets to the book directories and place them in a file named *document.txt* (This command can be run from anywhere in the file system):

```
p:\ocr\bin\movespreadsheet
```

The script will ask for the directory in which the tab delimited text versions of the spreadsheets resides. It will then ask for all of the directories that contain books in the shipment.

**Example:** When working on shipment 10, first specify the directory of spreadsheets *P:\shipment10*. Then specify the cd level directories one at a time at the prompt. **p:\ocr\cd30**

5. Check that files are accounted for by running the following script:

```
p:\ocr\bin\checktiffiles.pl
```

and reviewing the results. This script checks to make sure that the spreadsheets have been placed in the directories. It then checks to make sure that the .tif files and *document.txt* correspond. If they do, a message saying, "all is correct" will be printed on the screen. If not, a message will state that they are inconsistent. In that case, you should look at the file called *report*. It will appear in the directory that you were in when you ran the program. For instance, if you were in the directory *p:\ocr\* there will now be a file called *p:\ocr\report*. If all is correct, you can proceed to the next phase of QC. If there is an inconsistency, the book, images, and spreadsheet must be evaluated and corrections made before you can proceed.

6. Place tracking sheets in the bin for Image QC.

### Set Up for Image QC of Individual Books

1. Pick up a tracking sheet from the tracking bin. Get both the original book and the facsimile reprint that are listed on the tracking sheet. Compare all of the pages in the facsimile reprint to the pages in the original book. Inform the

Production Librarian about extremely dark images in the reprints. Make sure that all pages are in the reprint and in order.

## Running the Quality Control Program

1. Open the MS-DOS Prompt under Programs in the Start Menu.

a. At the prompt, type:

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

b. The script will request the directory that you wish to process. Type the drive letter and the number of the CDROM or Batch # and the ID number of the book that you are going to review. This information can be found on the tracking sheet.

```
Example: p:\ocr\jan02\00abc1234m
```

c. Include the :\ or it will fail to find the directory. You may also process a batch of books by typing the parent directory of the books you wish to process.

```
Example: If you wanted to process all of the books in cd9, you would type:
```

```
p:\ocr\cd9
```

2. After a few moments, it will present a list of all the directories (books) it expects you to process in this session.

```
Example:
```

```
p:\ocr\cd9\00abc1234\v0000\i000.  
Type "y" to proceed.
```

3. The script will check the directory (book) to see if it has been processed. If it has been fully processed, a message that the directory has been completed will appear, and the script will move to the next book.

4. When the script gets to a point where a directory has not been fully processed, the script will list the number of images that have been processed out of the number of images that are to be checked in the book.

You should answer one of three ways:

```
yes - to proceed  
no - to skip to the next directory (book), or  
exit - to exit the program
```

5. When the script starts to process a file, it will print a message such as:

```
Processing p:\ocr\cd9\00aaa1224m\v0000\i000\03450567.tif -- TPG001 --
text that is in the title field of the spreadsheet
```

6. After a few moments, *Imaging* should pop up with the corresponding image for that file.

Examine and compare each digital print image:

- a. Check that the document structure and title of the page displayed at the prompt are correct, as compared with the image itself.
- b. Ensure the digital image is a faithful representation of the print image.
- c. Make certain the pagination is correct and all figures (illustrations, maps, etc.) are configured correctly in both the digital and print forms.
- d. If an image is skewed more than the original, it will have to be corrected. Make a note on the tracking sheet.
- e. If one image appears smaller than the other images, it may have been scanned at a lower dpi, and it may have to be rescanned. Note this on the QC portion of the tracking sheet.
- f. Look for clean edges, clear contrast, and legible text.
- g. There should be nothing on a digital or print image that is not on the original image.
- h. Note any broken figures (illustrations, maps, etc.) in the digital files.
- i. Make sure that all blank pages contain the following statement: "This page in original is blank."
- j. Make sure that all images have been rotated for online viewing. LDRs have been rotated right. LDLs have been rotated left.
- k. Check the binding of the facsimile reprint to see that the hinges are tight and that there are no loose pages.
- l. If there are any problems, note the file name and the problem on the tracking sheet in the Quality Control Section. Make sure you include the name of the file and a complete description of the problem.

7. Leaving the Program before Completing the CD

If you need to stop in the middle of the processing, type *Ctrl-C* at the prompt.

*Imaging* will pop up another image. Just close it.

### **A Few Notes**

1. If you are processing multiple books at once, a CD that has been partially processed will produce several messages about completed files and directories before *Imaging* reappears the next time you start the program.
2. If an error message states that there is no corresponding image, it means that the spreadsheet and the actual files do not correspond. This should be reported

to the Digital Projects Librarian, and work on the book should be stopped until the error is corrected.

3. The scripts for this program reside on 'Ulsdr100' on the *P: drive* in the \ocr\bin directory.