Making Digital Facsimiles of Documents
Part 6: Image File Naming

It's important to have a good system for naming the image files that result from scanning. Such a system can make the assembly of digital facsimiles easier and faster. And it can minimize confusion, which can waste time and produce flawed results. Resist the temptation to use *ad hoc* names to be changed later.

There are many possible naming systems. In designing a naming system, it is important that it be simple, easy to use, and able to handle unusual situations. One basic point is that it's much easier to place images in a page-layout program if the names of the files are in the order of the pages of the document. Order means the order the computer lists them by name. In this ordering, digits come before letters, and digits and letters themselves are in their ordinary order. Other characters fall in various places in the order, but digits and letters are all that are needed for naming image files. An extension, such as **.tif** is produced by default in Windows. On the Macintosh, an extension is not necessary; skipping it saves time and effort.

Most documents are paginated; that is, they have page numbers. Books and similar documents, however, usually are not numbered consecutively from beginning to end. They typical have front-matter (a title page, foreword, contents, and so on), a body, which contains most of the content, and possibly back matter, such as advertisements.

Front matter typically is paginated with Roman numerals, omitted but implied on some pages. The body is usually paginated in Arabic numerals starting with 1 but in some cases continuing the numbering of the front matter pages, changing from Roman to Arabic numerals. The back matter may or may not be paginated. If it is, the numbers may or may not continue the body page numbering. And of course, there are the inevitable exceptions.

The pagination used in a document needs to be considered when setting up a method for naming its image files. One method is described in the following section. It accommodates the pagination of most documents, names body image files as they appear in the body, and keeps all the pages in order as described above.

A System for Naming Image Files

If the document has distinct sections with separate page numbering (or no numbering at all), such as front matter and body, it is advisable to have separate folders for each section. The folders should be named so that they are in alphabetical order corresponding to the order of sections, such as **a-front** and **b-body**.

Once the folders are created, the next step is to scan the document, section by section, putting the image files in their respective folders.

To determine how to name the pages in a section, determine the number of digits needed for the last page. For example, if there are 93 body pages, two digits are required. If there are 287, three digits are required, and so on. Then the image files for the pages in the section will be named the same as the page numbers in the sections, but with leading zeros so that all names have the same length (this preserves their order). So for a section with 93 pages, numbered starting at 1, the image file names would be **01**, **02**, …, **09**, **10**, **11**, …, **93**.

Note: You can always use more digits than are necessary. For example, three digits will accommodate all the sections of documents you are likely to scan, and you can use that number for all sections in all documents so you won't have to figure out how many are needed and keep track of different requirements for different documents.

Pagination Exceptions and Problems

All kinds of exceptions and problems occur in page numbering. Here are some of the most common exceptions and ways of dealing with them:

Blank pages. By convention, some printed pages in a document may be left blank. For example, it is common practice for the first page of a chapter to be on an odd-numbered page. If the preceding page is odd-numbered, a blank page (side) is inserted. It is a matter of taste whether or not to include blank pages in a digital facsimile. Since they add to document size, require more paper when printed, and may give the impression something is omitted, omitting blank pages is recommended. If blank pages are retained, there is no

January 12, 2003; last modified May 19, 2003

need to have files for them — the corresponding pages in the digital facsimile can be left blank.

Physically missing pages. If pages of a document are missing, the names for the corresponding pages can be omitted (file names only need to be in order; they need not be consecutive). As in other cases of document problems, a note mentioning the missing pages should be added to the digital document, perhaps on an otherwise blank page inserted at the appropriate place.

Pages without numbers. Numbers for some kinds of front-matter pages may be omitted but implied (that is, they figure in the numbering but are not printed). Such pages should be treated as if the page number was printed. The same rule applies to other kinds of pages. For example, in some books the numbers for the opening pages of chapters are implicit.

A different kind of unnumbered page occurs when plates or similar material are inserted between numbered pages but not given page numbers of their own. For example, a book may have a page numbered 250 followed by an unnumbered plate followed by a page numbered 251. Since the image file names for the two numbered pages would be **250** and **251**, the plate needs to be given a name that falls between these in order. This can be accomplished by appending a letter to the name of the file preceding the plate, such as **250m**. Thus the names appear in order as **250**, **250m**, **251**. A letter from the middle to the end of the alphabet is used because letters from the beginning of the alphabet have another use that is explained later. Think of **m** as starting the middle.

Of course, if there are several unnumbered pages between consecutively numbered pages, letters in order can be used, as in **250m**, **250n**, **250o**, **250p**, **250q**. The specific letters do not matter; order is all that counts.

Pagination errors. Errors in the pagination of documents, especially old documents, sometimes occur. For example, the page numbers in a book might be given as 257, 258, 257, 259 even though nothing is missing or duplicated. In such cases, the scheme given above for pages without numbers can be used. For the example given here, the following image file names could be used: **257**, **258**, **258m**, **259**, where **258m** is used to name the second page numbers 257. In a case like this, a note should be added to the digital facsimile to explain the problem.

Pages with More Than One Image File

As described in the article on scanning [1], a page with both text and pictures require separate scans for the different kinds of material. This results in more than one image file for the page.

This situation can be accommodated by using the same basic name for all the files and appending letters from the beginning of the alphabet to distinguish them. For example, if document page 143 has a color picture in addition to text, the file names for the two image file names could be **143** and **143a**. Additional letters can be used if there are more image files for a page. To be able to distinguish image files for picture, give the regular name to a B/W scan and names with appended letters to other scans.

Note that letters from the beginning of the alphabet identify multiple image files for the same document page, while letters from the middle to the end of the alphabet identify unnumbered pages.

Ralph E. Griswold
Department of Computer Science
The University of Arizona
Tucson, Arizona

Reference

1. Griswold, Ralph E. “Making Digital Facsimiles of Documents, Part 4: Scanning”, 2003: http://www.cs.arizona.edu/patterns/weaving/webdocs/gre_dd03.pdf