Writing an Index Using Microsoft Excel

By Seth A. Maislin

This article was written in 1999 and last published in the September 2000 issue of the STC Indexing SIG newsletter. Consequently, there are newer versions of Microsoft Excel available. Nevertheless, I believe that the techniques within are still accurate, though better methods may exist. Since writing it I've also discovered that some of my instructions are esily misunderstood by those not familiar with Excel, so I strongly recommend saving your work on a regular basis, in case you need to undo something. Finally, feel free to email me your suggestions, and I'll include them with this article.

When I first started indexing, the last thing I wanted to do was buy software. So instead I used what I had: Microsoft Excel.

Looking back, I'm astounded I indexed with Excel at all. An index hierarchy is nothing like a spreadsheet. Excel resisted me almost every step of the way: sorting page numbers, combining index entries, formatting cross references, alphabetizing, and alphabetizing overrides. I learned later than indexing-dedicated software like Sky Index, Macrex, and Cindex would have saved me from these agonies, and from regular expressions. Nevertheless, my desperation to not spend $600 drove me to adapt Excel for my purposes.

This article is a documentation of how Excel can be used for writing an index. But I warn you now: it's not pretty, and it's not recommended. There's only one halfway-decent reason to use Excel for indexing, and that's if you never plan on writing another index within the year. Otherwise, please spend the money. A list of indexing software packages is available at http://www.asindexing.org/site/software.shtml.

Writing Your Index Entries

Use one column for each level of the index, and an additional column for page numbers. For example, an index with sub-subentries requires at least four columns. Use the first column for page numbers, not the last; this speeds up input of index entries that don't use all sublevels.

When you input your entries, you must insert all levels for that entry. This is important: insert data into every column, starting with the page number, until your entry is complete. Empty fields affect sorting, so that null values for main entries will be missorted. The example below, which shows only two levels, represents a portion of an index. Notice how page numbers and page ranges are both inserted into the first column. Finally, working with your index is easier if you widen the columns as much as possible.

	A	B	C
407	95-97	color	
408	95	hexadecimal numbers for colors	
409	95	names of colors	
410	95	RGB color components	
411	95	color	representing with names and numbers
412	97-102	formatting HTML pages	
413	97-102	HTML	formatting in
414	98	fonts	in HTML pages

Page Numbers and Cross References

Page numbers are going to be the most troublesome element of using Excel. To simply this process later, I recommend using only one page number or page range per row. Excel thinks that page ranges are text (and not numeric) because of the hyphen character; this is why the ranges are left-justified in the above illustration, whereas all other numbers are right-justified. This distinction is not important, so you can ignore Excel's choice of alignment. In fact, having ranges "stand out" is an easy way to review your work later.

If you decide to use a symbol different from a hyphen for page ranges, you can make that change later, globally. If you are using complex page numbers, such as 12-3 or 12.3, make sure that you use different symbols for page ranges than you do to separate the parts of the page numbers. For example, use "12-3 to 12-7" or "12-3:12-7" instead of "12-3-12-7".

How cross references are written depends on your final format. Most formats require see-type cross references to be run in with the lowest-level, and see also–type cross references to be pushed down on level in the hierarchy (see below). Put see references in the same cell as your rightmost text, and put see also references in a column to the right of your rightmost text. Use a page number of zero (0). Don't concern yourself with formatting like italics. However, do use the correct syntax and capitalization.

alignment
    text
        See also text, formatting
text
    alignment
        See alignment

419	0	text	alignment	See alignment
420	0	alignment	text	See also text, formatting

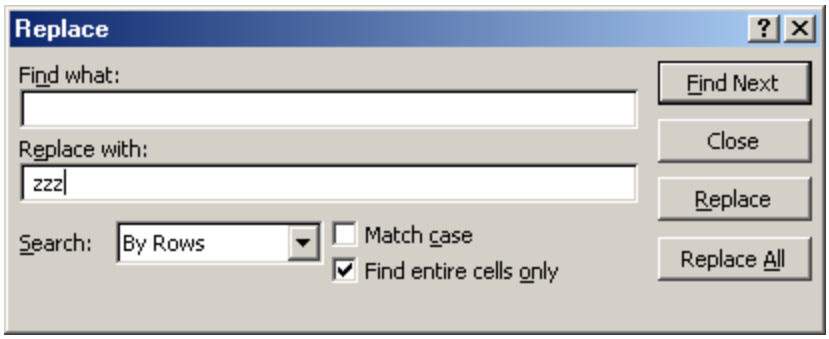
Saving Your Index

Now it's time to convert this spreadsheet into an index hierarchy. Every step from this point forward lends itself to serious accidents. You can accidentally delete entire columns, rows, or page numbers. To prevent this, save your index incessantly, using different filenames every time. Be particularly sure to save your "untouched" dataset, and do not delete this file ever. This is the only trustworthy source you have, and you'll want it every time you're unsure if you made a mistake.

Sorting Your Index

Sorting your spreadsheet-bound index is not straightforward. If your index has four columns, it is tempting to simply sort first by Column B, then by Column C, and then by Column D. (Excel allows you to sort rows based on three different columns.) However, the blank cells in Columns C and D sort below cells with text. In addition, sorting the text alone is not enough, because you also want the page number to appear in numerical order—and Excel is unable to sort page numbers and page ranges as equals, because ranges are “text” and not numeric.

Let’s address these problems individually. First, we need to give value to the empty cells. Highlight all the cells in the text columns (not the page number column) within the rows of your index. Use the Replace dialog (Edit>Replace) to find empty cells and replace them with the value “ zzz”. (Notice there is a space before the letters.) Be sure to check the “Find entire cells only” box. Select Replace All. This adds the space character plus “zzz” into every empty cell in your index.



Now, in the first empty cell to the right of your first index entry, add the concatenation formula below. This formula creates a single string of text that represents the sorting order of each index entry. Excel formulas always start with the equal sign (=). For this example, I am assuming that the index begins with Row 1, that Column A contains page numbers, and that Columns B through D contain text. (Thus this formula is inserted into cell E1.)

=CONCATENATE(B1,"zzz",C1,"zzz",D1)

Then fill the entire column with this formula by using Edit>Fill>Down. The numbers in the formula will change from one row to the next, so that the numbers in any particular version of this formula are equal to the row number. For example, the formula that appears in row 46 should look like this:

=CONCATENATE(B46,"zzz",C46,"zzz",D46)

Here is what my index sample looks like after using this formula. Save this file.

	A	B	C	D	E
407	95-97	color	zzz	zzz	colorzzzzzzzzzzzz
408	95	hexadecimal numbers for colors	zzz	zzz	hexadecimal numbers for colorszzzzz
409	95	names of colors	zzz	zzz	names of colorszzzzzzzzzzzz
410	95	RGB color components	zzz	zzz	RGB color componentszzzzzzzzzzzz
411	95	color	representing with names and numbers	zzz	colorzzzrepresenting with names and
412	97-102	formatting HTML pages	zzz	zzz	formatting HTML pageszzzzzzzzzzzz
413	97-102	HTML	formatting in	zzz	HTMLzzzformatting inzzzzzz
414	98	fonts	in HTML pages	zzz	fontszzzin HTML pageszzzzzz
415	99-101	<FONT> tags	zzz	zzz	<FONT> tagszzzzzzzzzzz
416	101	constant-width text	zzz	zzz	constant-width textzzzzzzzzzzz
417	101	text	constant-width	zzz	textzzzconstant-widthzzzzzz
418	101	preformatted text	zzz	zzz	preformatted textzzzzzzzzzzz
419	0	text	alignment (see alignment)	zzz	textzzzalignment (see alignment)zzzz
420	0	alignment	text	(see also text, formatting)	alignmentzzztextzzz(see also text, fo

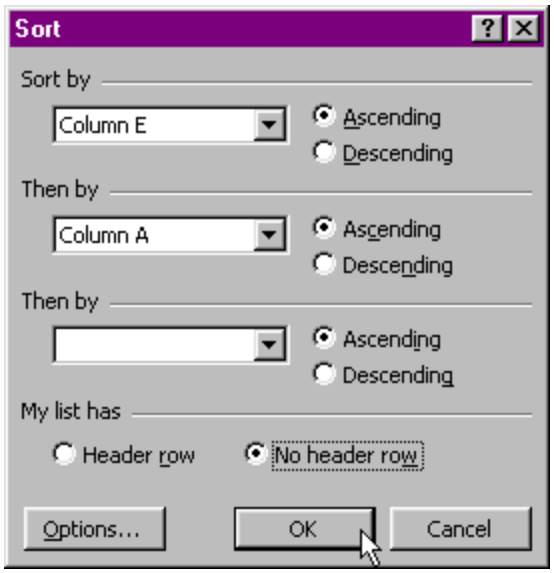
This new column is made of up formulas, not values. To get values, start by copying this column. Now highlight the next column over and select Paste Special from the Edit menu. In the resulting dialog, select the Values radio button and click OK. In our example, this copies the values from Column E into Column F. Now delete Column E entirely; we don't need it any more.

Sorting by Column E at this point would be possible, but premature. For example, characters such as the < character in row 415 and words like "in" in cell C414 get in the way of effective sorting. Unimportant nonalphanumeric characters and words should be ignored. To remove initial punctuation and prepositions from your index, use the Replace dialog in the rightmost column.

We can ignore the word "in" in cell C414 when sorting by searching Column E for the text "zzzin "—that's "zzz" plus the preposition plus a space—and replacing it with nothing. (This is why we added more "zzz" characters in the concatenation formula: to distinguish between prepositions that appear in Column B and those that appear in other columns.) We use the space to distinguish the word "in" from words that start with those letters. We can perform this search-and-replace process for every preposition, if we want. This is a repetitive process, and you may forget to remove certain prepositions along the way. Later you can clean up anything you can't fix now.

This is your last chance to adjust sorting globally. Remove (or replace with spaces) any hyphens, periods, and other nonalphanumeric characters that you want to ignore when sorting. If you want to sort entries letter-by-letter, remove *all* spaces and nonalphanumerics from the rightmost column.

As a final step, we also need to determine how cross references are to be sorted. If see also–type cross references are to sort at the bottom, add another set of "zzz" characters in front of your cross reference syntax. In the example above, which uses parentheses, I would globally replace "(see " with "zzz(see ". (Note the trailing spaces.) On the other hand, if you want to guarantee that see also entries sort to the top, add a space character in front of your cross reference syntax. (Remember that spaces sort above characters, but empty cells sort below occupied cells.) If you plan on adding spaces for cross references, however, I recommend performing this replacement absolutely last so that it doesn't interfere with the other replacements described above.



Now you are ready to sort your index. Sort by Column E and Column A. Although you could delete the rightmost column at this point, don't. Keep it in case you need to re-sort after an error. Highlight all rows of your index and use the Sort dialog. Select the "No header row" radio button when choosing your columns, and always sort in ascending order.

	A	B	C	D	E
407	99-101	<FONT> tags	zzz	zzz	<FONT> tagszzz zzzzzz zzz
408	0	alignment	text	(see also text, formatting)	alignmentzzztextzzz(see also text, fo
409	95-97	color	zzz	zzz	colorzzz zzzzzz zzz
410	95	color	representing with names and numbers	zzz	colorzzzrepresenting with names anc
411	101	constant-width text	zzz	zzz	constant-width textzzz zzzzzz zzz
412	98	fonts	in HTML pages	zzz	fontszzzin HTML pageszzz zzz
413	97-102	formatting HTML pages	zzz	zzz	formatting HTML pageszzz zzzzzz zz
414	95	hexadecimal numbers for colors	zzz	zzz	hexadecimal numbers for colorszzz z
415	97-102	HTML	formatting in	zzz	HTMLzzzformatting inzzz zzz
416	95	names of colors	zzz	zzz	names of colorszzz zzzzzz zzz
417	101	preformatted text	zzz	zzz	preformatted textzzz zzzzzz zzz
418	95	RGB color components	zzz	zzz	RGB color componentszzz zzzzzz zz
419	0	text	alignment (see alignment)	zzz	textzzzalignment (see alignment)zzz z
420	101	text	constant-width	zzz	textzzzconstant-widthzzz zzz

With the index in the proper order (above), the next challenge is remove repetitive information from one row to the next. That is, instead of having two entries with the same top-level text, we want one entry with main-level text and another entry with just the indentation. For example, we want to entry cells B410 and B420.

Verify that your sorting is exactly what you want. From this point forward, you will be unable to re-sort the document. It's a good idea to save what you have now under a different filename.

Create a new column for each of your text columns; in our case we're going to create Columns F, G, and H to mimic columns B, C, and D. In the second row of each column, use these formulas:

Cell F2	=IF(B2=B1,"",B2)
Cell G2	=IF(C2=C1,"",C2)
Cell H2	=IF(D2=D1,"",D2)

Then fill these values down for each index entry. (There should be no formulas in Row 1.) Thus for row 408, these are the formulas:

Cell F408	=IF(B408=B407,"",B408)
Cell G408	=IF(C408=C407,"",C408)
Cell H408	=IF(D408=D407,"",D408)

These formulas create columns that are identical to columns B–D, but with duplicate cells replaced with empty cells. Much of the "zzz" content will disappear, and that's okay. Save your spreadsheet file.

Copy the new columns (for our example, highlight Columns F–H and select Edit>Copy) and use Paste Special to put the values into the columns that hold your index text (for us, Columns B–D). This effectively deletes (permanently) all repetitive information.

We also need to prepare the page numbers for later manipulation. Create a new column with the formula =CONCATENATE("pagenumbers"+A#) where # is the row number (and Column A has the page numbers). Fill down the entire column with this formula. Then copy and paste the values (using Paste Special) over your page numbers in Column A. This effectively adds the text "pagenumbers" to the beginning of your numbers.

Finishing the Job

There's one last step. Move the column with page numbers to the column to the right of the column containing your lower-level entries. In our example, then, Columns A through C have text, and column D contains page number data.

If you want en dashes for your page numbers, replace every hyphen in the page number column with the text "XXENDASHXX" or something similar.

We'll make replacements later.

Save your Excel file under yet another name. Delete every column that doesn't contain data you want to keep (the sorting information and the formulas). Save the remaining columns—your index (Columns A–D)—as a tab-delimited file using File>Save As. Now we're done with Excel.

Open the file using word processing software, such as Microsoft Word. You should see a document that looks something like the one at right, where the arrows represent tab characters. Notice how Excel adds quotation marks around certain cells.

```
alignment → text → "(see also text, formatting)" → pagenumbers0¶
color → zzz → zzz → pagenumbers95-97¶
   →   representing with names and numbers →   →   pagenumbers95¶
constant-width text   →   zzz →   →   pagenumbers101¶
<FONT> tags →   →   →   pagenumbers99-101¶
fonts in HTML pages   →   →   pagenumbers98¶
formatting HTML pages → zzz →   →   pagenumbers97-102¶
hexadecimal numbers for colors   →   →   →   pagenumbers95¶
HTML → formatting in   →   →   pagenumbers97-102¶
names of colors → zzz →   →   pagenumbers95¶
preformatted text →   →   →   pagenumbers101¶
RGB color components   →   →   →   pagenumbers95¶
text → alignment (see alignment)   →   →   pagenumbers0¶
   →   constant-width   →   →   pagenumbers101¶
```

Delete all every occurrence of "zzz" using the Replace dialog. Then delete extra spaces around tabs by searching for space+^t and for ^t+space and replacing them with just ^t. (The "caret-t" syntax is the code for a tab character.) You can globally delete quotations marks as well, but be careful not to delete quotation marks you intended to keep. If you're using en dashes, replace "XXENDASHXX" with en dashes. If you're using italics, replace the words "see" and "see also" with italicized versions. (Be sure to use a space after the word "see" to avoid italicizing words that coincidentally start with those letters.

Now finalize your cross references. How you do this is unique to your chosen syntax. In our case, we don't have to do anything except remove the quotation marks around see-alsos. If you are running in see-also references with periods, you might want to replace "^p^t^tSee also" with ". See also" so that your references abut the page numbers of the previous lines. (The ^p symbol represents a paragraph mark.) Use only the number of tabs you need.

```
alignment → text → pagenumbers56¶
   →   →   See also text, formatting¶
```

Only page numbers need fixing now. Globally delete "^tpagenumber0" to remove the phony page numbers used with cross references. Globally replace "^tpagenumbers" with "pagenumbers" again and again until you no longer can. Then replace "^ppagenumbers" with ", " (comma+space). (This assumes you are using commas to separate page numbers. If you are using semicolons, replace with semicolon+space.) This last step combines multiple page references for identical entries. Now replace every occurrence of

“pagenumbers” with comma+space.

Believe it or not, you’re done! Save your final index, and correct any mistakes manually.

Top