

Naming Files and folders...Simpler is not always better

By Manuel Bulwa

Integrated Scanning of America, Inc.

September, 2007

Simpler is only better if you choose smartly from all possible “simple” options. Naming your files and folders using keywords and values (a.k.a. “metadata” (*)) contained in your documents is **not** necessarily your smartest choice. I know, this practice is quite common, but so is drinking and driving. You may be tempted to mimic your manual filing system, with drawers and folder tabs labeled chronologically or alphabetically. Back then, that simple choice was a smart one, because there were not many other choices anyway. With today’s technology, much smarter options are, indeed, available.

This practice is popular because modern Operating Systems (*) provide you with a graphical tree interface that helps visualizing the folder structure and file names. However, this simplistic searching process may be somehow adequate for a handful of files (dozens, maybe hundreds), but definitely not good enough for larger business environments and serious use.

So, what’s wrong with it?:

Metadata is subject to changes. These changes, whether they are natural or caused by data maintenance needs, often affect just the way you search and find your documents, which remain intact. You may be forced to move or rename a file, only because you corrected a misspelled keyword or added a new one. It may even be impossible to move or rename files if you have them stored on read-only media or under restricted permissions, which are often necessary.

In a multiuser environment (*) you are exposing a vulnerable file system to misuse and abuse by others. You don’t really want others to drag, drop, move and rename files and folders.

Folder and file names may be too long or contain special symbols that are illegitimate in current or future operating systems (*). Many such document sets failed when upgraded to a new version.

Duplicates can become a nightmare. You may permanently lose valuable files that were overwritten by a different document named identically. Or you may be caught in the trap of naming files “-001, -002”, etc. which makes maintenance and searching less efficient.

Abbreviating and concatenating (*) words in a folder or file name creates trouble as well. It also restricts your ability to experiment with different hierarchical views (*) of your data when presented. I have witnessed convoluted naming with trailing and leading blanks and underscores, etc, just to improve on

the visual presentation.

Microsoft™ Windows Explorer is particularly risky when not used carefully. It makes “mousing around”, dragging and dropping, moving columns and files unsafe. The slightest quiver can cause files to be lost or misplaced without notice.

Your backups are ineffective or unnecessarily complicated. You are forced to create lengthy backups with the exact same intact files, but now relocated elsewhere or named differently. If you distribute duplicates, they may become prematurely obsolete.

Which, “simpler” option is better or smarter?:

Name your files after a unique value such as a zero-filled sequential number. (If you feel brave you may even use a GUID (*)). If you deal with thousands of files or more, you may want to group them in folders (also named numerically) of approximately 1,000 files each. However, this is entirely optional and even unnecessary if properly handled.

Create and maintain a separate “Table of Contents” (TOC) (*), using a spreadsheet or a data base. You may even use a text or HTML editor if you don’t need to get too fancy.

The TOC will contain columns with each search index (*), one of which will contain a hyperlink (*) to the actual document. Hyperlinks are very simple to create, maintain and use.

Search and rearrange presentation views using the columns in the TOC, then click on the hyperlink when you need to view your selected documents.

Keep your TOC read-only, except for an administrator to perform maintenance on it.

If you use sequential numbers as file names, do not impose the obligation of sequentiality across folders, i.e. do not expect the first file of a folder to be one higher than the last number on the previous folder. In fact, numbers can and should be totally arbitrary and meaningless.

Backup your files only once and when truly added or changed file contents, not just names.

In essence, all we did is:

We separated metadata from document files, and

We stopped using a “poor man’s database” (the file system).

You do not need to be an expert or ask for expert assistance to implement any of the simple techniques I suggested. In fact, you will need such expertise to repair (when repairable) the consequences of using a flawed (although extremely popular) methodology.

(*)Glossary

Metadata, Index: Collection of words and values used to describe a document for cataloguing and/or searching purposes.

Concatenate: Append or link values together.

Hierarchy, structure: The ordering, sorting and grouping of a set of values.

Operating System: Software used to run a computer, such as Microsoft Windows™, Unix, etc.

File System: The methodologies and formats used for storing, naming and retrieving files on a disk.

Globally Unique Identifier (GUID): Unique reference key that has no duplicate or repeated value in any context, hence globally.

Database, DBMS: Collection of data or information organized for rapid search and retrieval. Databases are structured to facilitate storage, retrieval, modification, and deletion of data. A database consists of a file or set of files that can be broken down into records, each of which consists of one or more fields. Using keywords and sorting commands, users can rapidly search, rearrange, group, and select records to retrieve or create reports according to the rules of the database management system (**DBMS**) being used.

Permissions: The ability in a file system to control and restrict file access (read, write, execute, traverse, etc.) by certain users and groups of users.

Globally Unique Identifier (GUID): A unique value that is generated by an application to unambiguously identify an object.

Hyperlink: A convenient way of linking to a file or a section of a document by simply clicking on a value. Frequently blue and underlined.

Multi user environment: A computer scenario where more than one user may have access to the same files or other computer resources.