The Old New Thing

Why do NTFS and Explorer disagree on filename sorting?

17 Jun 2005 11:17 AM | **11**

Some people have noticed that NTFS automatically sorts filenames, but does so in a manner different from Explorer. Why is that?

For illustration purposes, I created files with the following names:

Name	Code point	Description
a	U+0061	Latin small letter A
b	U+0062	Latin small letter B
×	U+00D7	Multiplication sign
å	U+00E5	Latin small letter A with ring above
ø	U+00F8	Latin small letter O with stroke

And here's the sort order for various scenarios, at least on my machine. (You'll later see why it's important whose machine you test on.)

Plain "dir" command

a	U+0061	Latin small letter A
b	U+0062	Latin small letter B
å	U+00E5	Latin small letter A with ring above
×	U+00D7	Multiplication sign
ø	U+00F8	Latin small letter O with stroke

"dir /on"

×	U+00D7	Multiplication sign
a	U+0061	Latin small letter A
å	U+00E5	Latin small letter A with ring above
b	U+0062	Latin small letter B
ø	U+00F8	Latin small letter O with stroke

Explorer sorted by name

×	U+00D7	Multiplication sign
a	U+0061	Latin small letter A
å	U+00E5	Latin small letter A with ring above
b	U+0062	Latin small letter B
ø	U+00F8	Latin small letter O with stroke

First, notice that Explorer and "dir /on" agree on the alphabetic sort order. (Once you throw digits into the mix, things diverge.) This is not a coincidence. Both are using the default locale's word sort algorithm.

Why does the raw NTFS sort order differ?

Because NTFS's raw sort order has different goals.

The "dir /on" and Explorer output are sorting the items for humans. When sorting for humans, you need to respect their locale. If my computer were in Sweden, Explorer and "dir /on" would have sorted the items in a different order:

× U+00D7 Multiplication sign

a U+0061 Latin small letter A

b U+0062 Latin small letter B

å U+00E5 Latin small letter A with ring above

ø U+00F8 Latin small letter O with stroke

You can ask a Swede why this is the correct sort order if you're that curious. My point is that different locales have different sorting rules.

NTFS's raw sort order, on the other hand, is not for humans. As we saw above, sorting for humans can result in different results depending on which human you ask. But there is only one order for files on the disk, and NTFS needs to apply a consistent rule so that it can find a file when asked for it later.

In order to maintain this consistency, the NTFS raw sort order cannot be dependent upon such fickle properties as the current user's locale. It needs to lock in a sort algorithm and stick to it. As Michael Kaplan pointed out earlier, NTFS captures the case mapping table at the time the drive is formatted and continues to use that table, even if the OS's case mapping tables change subsequently. Once the string has been converted to uppercase, it then needs to be sorted. Since this is not for humans, there's no need to implement the complex rules regarding secondary and tertiary keys, the interaction between alphanumerics and punctuation, and all the other things that make sorting hard. It just compares the code points as binary values, also known as an ordinal sort.

In summary, therefore, Explorer sorts the items so you (a human) can find them. NTFS sorts the items so it (the computer) can find them. If you're writing a program and you want the results of a directory listing to be sorted, then sort it yourself according to the criteria of your choice.

Blog - Comment List MSDN TechNet

Comments

Andreas Haeber
17 Jun 2005 12:36 PM
 #
Just a little trivia, but I never knew that Swedes sort ø and å different then (us) Norwegians. In Norway the end of the alphabet is: x, y, z, æ, ø, å.

Btw. here is the complete Swedish variant of the latin alphabet http://sv.wikipedia.org/wiki/Latinska_alfabetet, if anybody else is curious :) Funny that it is the opposite of the Norwegian alphabet.

Will Sullivan
17 Jun 2005 1:27 PM
 #
What I wanna know is why, if I drag a sorted list of files from an explorer window to a

program, the files are sent not in order, but with the first five or so in the list appended to the end of the list. THAT'S what I wanna know.

Jonathan
17 Jun 2005 2:13 PM
#

Will Sullivan: From my experience (and cursory experimentation now), progs get the files by the order they appear. But, the file you drag with (the "active" file?) gets to be first. So, if you have 1,2,3,4,5 selected, and you drag 3 to a prog, it becomes 3,1,2,4,5.

(And Outlook Express reverses the order as well, so it becomes 5,4,2,1,3)

microbe
17 Jun 2005 3:07 PM
#

So NTFS organizes its directories by sorted trees (like a B+ tree)? For FAT, it's not sorted at all.

Michael S. Kaplan
17 Jun 2005 9:21 PM
#

One interesting point about the casing rules of NTFS is that the table is really not exposed to developers, so there is a ton of code that assumes the current OS tables are the right ones. So you end up with situations where you can do some things with files that you cannot do, asnd other things that you can. But only for people trying to use files that ought not to be allowed to co-exist....

Jonathan Wilson
18 Jun 2005 7:07 AM
#

What I want to know is why NTFS preserves case of files but yet is case insensitive.

Spire
18 Jun 2005 12:22 PM
#

>What I want to know is why NTFS preserves case of files but yet is case insensitive.

Probably for some degree of backward compatibility with FAT, which was also case-insensitive. (Or at least the DOS API calls were -- but only on the input side of things. If you used a disk editor to hand-edit non-LFN FAT directory entries to use lowercase characters, DOS would completely fail to find the files.)

DOS users were used to typing commands in all-lowercase, even though filenames were displayed and handled internally as all-uppercase. If the Windows command processor suddenly started requiring uppercase-only access for all old files, users would not be happy at all. Almost every batch file in existence would break. So would many applications.

Case preservation was a desirable new feature in NTFS, but they didn't have to do away with case-insensitivity to add this feature, so they didn't.

Lars Viklund
20 Jun 2005 3:38 AM
#

It's actually possible to enable case sensitivity on NTFS drives in Windows.

I'm unsure of how you do it manually, but the Services For Unix installer asks you if you want to enable it for greater compability with SFU applications.

Lars Olaussen
20 Jun 2005 5:45 AM
#

Actually, ø is the Norwegian and Danish equivalent character to the Swedish ö (U+00F6).

Anyway, an "interesting" feature with the Norwegian locale sorting is that aa is treated as å, since aa is used for å in many names of people and places.

Just a little tip when you can't find your files starting with aa and you're using the Norwegian locale.

CalvinH
28 Jun 2005 1:33 PM
#

If you have some folders named "1", "2", "3" and "10", and display them sorted by name using Windows Explorer, they will be sorted numerically rather than alphanumerically. This can lead to confusion if you have folders like "200008", 200009", "20000814", which will show August, then September, then August 14th.

余啊雷
19 Feb 2008 7:44 AM
#

PingBack from http://soci.hu/blog/index.php/2008/02/19/sql-server-2008-ujdonsagok-23-ordered-sqlclr-table-valued-functions/