

## Issue with UTF8 and the encoding...

Posted by msa - 2003/03/17 07:13

---

(Just to let you know)

Today I found an issue with the automatic UTF8-hyphen-conversion added in 2.12 that caused me alot of problems... The issue relates to when cutting and pasting from help output from a perl man-module. The UNICODE-hyphens are shown in the help-file, but cutting/pasting them to the command line does not work (UNICODE-hyphens versus "real" hyphens)...

Also the editing of command-lines gets a bit confused (I am able to go before the actual command starts).

To reproduce on RedHat linux 8.0 xterm:

- 1) Execute the command "man man"
- 2) Cut the "-C config\_file" text.
- 3) Press "Q" to quit the man program.
- 4) Paste to the command line and press left arrow
- 5) The cursor will move two chars before "-C".

/msa

---

## Re: Issue with UTF8 and the encoding...

Posted by bpence - 2003/03/17 12:46

---

Mattias,

It comes down to this...

When replacing the UNICODE hyphen, the ASCII dash character is substituted for display only. To reduce the possibility of accidentally stomping on the hyphen for those who \*really\* want to have a hyphen, it is still treated as a hyphen internally. So, when you copy and paste, the full UNICODE hyphen is used. The ASCII dash is substituted for display only.

What you noticed was that at the command line, the cursor position can become confused when cursoring over the hyphen. This is caused by the shell not interpreting UTF8 correctly. The shell is interpreting bytes instead of characters. With UTF8, a single character may be several bytes long (the hyphen for example). After some experimentation, I have noticed that some shells handle this better than others. '/bin/bash' - the default RedHat shell has this problem, but '/bin/bash2' does not. '/bin/sh' doesn't have the problem either.

Still, if you use a proper shell where the command editing is working correctly with the UTF8 characters, you'll still run into the problem that a hyphen simply doesn't work as a command-line flag. Only an ASCII dash will do. This, of course, begs the question of why the heck they used hyphens in the man page when the really should have used a dash!!

Go figure!

Still, I'm not sure there's anything I can do about this on the client side. I'm doing everything I can to make sure things display properly without being overly restrictive. For example, I \*could\* internally convert all hyphens to dashes, but that would almost certainly break things for other people. It would mean that a \*real\* hyphen would never be recognized.

You could always remove the 'utf8' designation from your LANG variable to disable utf8 altogether. On US-English based systems it's generally not necessary anyway.

=====

## Re: Issue with UTF8 and the encoding...

Posted by msa - 2003/03/17 13:39

---

Brian,

I understand the complexity of this and I must say that I admire the way AT currently handles UNICODE. As you pointed out; why have hyphen instead of a dash?!? Better looking display?!? Please consider my posting as a FYI if there are other users with similar problems.

Thanks,  
/msa

=====