Subject: Re: Rename files using IDL - is it possible?
Posted by Heinz Stege on Fri, 13 Aug 2010 22:15:36 GMT
View Forum Message <> Reply to Message

Hi Chris.

On Thu, 12 Aug 2010 14:12:05 -0700 (PDT), Chris Torrence wrote:

> Hi all,
>
...
>
> Now, regarding Heinz's problem - this is unfortunately a pathological
> case, where the 195+133 happens to be the UTF-8 encoding for the
> extended ASCII character for 197 (the A with the ring).
>
> Internally, IDL tries to recognize UTF-8 strings by the rules for
> UTF-8 encoding. If the first byte is > 193, and the second byte is
> between 128-191, then IDL assumes that this is a UTF-8 string. The
> file I/O routine then converts the filename to UTF-8 encoding, which
> converts the 195+133 to 197, and therefore ends up with the same
> filename.
>
> Our UTF-8 conversion routines are based upon the UTF-8 standard, and
> are designed to the best of our ability to work with multiple
> languages and encodings, including European and Asian languages.
> Unfortunately, given a random string of bytes, there is no way to
> definitively determine that this is a UTF-8 versus native string.
> Hopefully, the case given by Heinz will occur rarely if ever.
>
> Cheers,
> Chris
> ITTVIS

Thank you very much for this explanation.

I share your view, that this "pathological case" under normal
circumstances will occur rarely if ever. On the other hand it is a
very serious thing to delete wrong files or even wrong directories. I
wonder if it is the concept of UTF-8 to convert strings to extended
ASCII - as far as possible. In this way we get a mix of filenames,
some encoded in extended ASCII and others in UTF-8.

I am afraid that this approach will make trouble for IDL users again
and again. Aren't there better solutions?

Would it be possible to use UTF-8 filenames for the I/O routines in
principle? Or at least if a corresponding keyword is set? I myself

felt fine with the IDL 7 approach.

You could add the conversion routines separately to the builtin functions. This would give users the option to do the conversion if wanted. This furthermore would not break old code.

Regards, Heinz