The format of the translit.dat file suggested for possible use for transliteration

William J G Overington

17 March 2014

This is a thought experiment at present.

Automated transliteration would be by having a file translit.dat available. In the thought experiment the file is a UTF-16 text file, such as can be saved from the WordPad program by selecting saving as a Unicode Text Document.

The translit.dat file would consist of a number of lines of text.

A valid line of text would have one of three possible formats.

If the first character of the line is an asterisk, then the line is a comment.

If the first character of the line is a PERCENT SIGN then the line is the last line of the file.

Otherwise the line is intended to be a transliteration line, yet only is a transliteration line if it is of the correct structure.

The correct structure for a transliteration line is as follows.

One or more characters that are not the VERTICAL LINE character.

A VERTICAL LINE character.

One or more characters that are not the VERTICAL LINE character.

The possibility was considered that on some software platforms that there might be complications, while reading characters from the translit.dat file, regarding detecting the end of the translit.dat file.

If the first character of the line is a PERCENT SIGN then the line is the last line of the file.

In a translit.dat file produced as a Unicode Text Document saved from the WordPad program, lines are separated by two characters, namely CARRIAGE RETURN and LINE FEED, in that order. That is, pressing the return key on the keyboard produces two characters in a Unicode Text Document saved from the WordPad program.

The final five characters of the translit.dat file are here specified to be as follows.

CARRIAGE RETURN
LINE FEED
PERCENT SIGN
CARRIAGE RETURN
LINE FEED

This is achieved using WordPad by pressing the return key both before and after the PERCENT SIGN has been entered.

It is noted that a Unicode Text Document saved from the WordPad program stores the two bytes of each character with the lower byte before the higher byte.

It is noted that a Unicode Text Document saved from the WordPad program starts with a U+FEFF character, used as a BYTE ORDER MARK. Thus the first two bytes of a translit.dat file do not represent a character used in the automated transliteration process.

It is noted that for English and for some other languages that a Unicode Text Document saved from the WordPad program has many bytes that have a value of zero. However, the use of a

Unicode Text Document saved from the WordPad program is deliberately chosen for this system so as to make participation in producing a translit.dat file as straightforward as possible, and with the hope that software developed for automated transliteration using this system will work for all languages that can be represented using Unicode characters.