

The format of the readouts.dat file suggested for possible use in the application of localized read-out labels

William J G Overington

16 April 2014

This is a thought experiment at present.

Automated localization would be by having a file readouts.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 readouts.dat file could either be a standardization file, which would use English with en-gb-oed spelling, or could be a copy of that file that has been translated into some other language, with due consideration for localization issues.

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

A valid line of text would have one of four 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.

If the first character of the line is an EQUALS SIGN then the line is a heading for a cascading menu for semi-automated message construction; and also the rest of the line is intended to be a localization line as below.

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

The correct structure for a localization 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.

As an example of use, a localization line with one pictograph character before the VERTICAL LINE character; and several characters, including spaces if needed, in a script used for expressing a language, after the VERTICAL LINE character.

The possibility was considered that on some software platforms that there might be complications, while reading characters from the readouts.dat file, regarding detecting the end of the readouts.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 readouts.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 readouts.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 readouts.dat file do not represent a character used in the automated localization 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 localized version of a readouts.dat file as straightforward as possible, and with the hope that software developed for automated localization of read-out labels will work for all languages that can be represented using Unicode characters.

A localized file may be produced in any language, preferably by a native speaker of that language, and a Unicode Text Document saved from the WordPad program, and the file published, keeping the file name as readouts.dat as the idea is that software developed for automated localization of read-out labels will hopefully work successfully with whatever version, in whatever language, of a readouts.dat file with which it is supplied at any particular time.