Experiment specification

William Overington

Saturday 20 May 2017

Please write an application program to run on a PC that does the following, capable of using all characters in the Basic Multilingual Plane of ISO/IEC 10646 / Unicode.

Ask the end user for the name of a PDF (Portable Document Format) file that contains source text.

Ask the end user for the name of a PDF file that contains a localization table.

Produce a new PDF file, the result text, by the following method. Please note that when copying a character through to the result text that the font used to display that character is the same as the font used to display it in the place from where it was copied.

Make one pass through the source text and for each character act according to whether it is or is not a Private Use Area character.

If the character is not a Private Use Area character, then copy it through to the result text.

If the character is a Private Use Area character, then act as follows. Look in the localization table and if and only if the Private Use Area character is listed in the localization table followed by a | character, then do not copy it to the result text yet do copy all text after the said | character through to the end of that paragraph in the localization table through to the result file. If the Private Use Area character followed by a | character is not available in the localization table then copy the Private Use Area character through to the result text.

The following test data is supplied.

source_text_for_an_experiment.pdf

localization_table_for_an_experiment_English.pdf

The following is the desired result from the above test data.

desired_result_for_an_experiment_English.pdf

Please note that the test data is in English. The intention is that if the localization table were in another language, for example, French, Latvian, Japanese then the program would still produce a good result, so it is important that 16-bit characters are used. This specification is for using 16-bit characters and not the full character range of ISO/IEC 10646 / Unicode so as not to complicate this experiment.