Subject: Re: Choosing the best way to go full UNICODE
Posted by mirek on Mon, 19 Jun 2017 08:03:40 GMT
View Forum Message <> Reply to Message

My undestanding is that if decomposition sequence starts with "<", it is 'compatibility', if not, it is 'canonical'.

I believe that you should use compatibility seqences e.g. for comparing, but you should never 'recompose' these into single codepoint - one of reasons is that canonical compositions are unique, but there can be the same compatibility decompositions for multiple codepoints (found out that hard way during testing).

In either case, i have added a bool

int   UnicodeDecompose(dword codepoint, dword t[MAX_DECOMPOSED], bool& canonical);

to 'decompose' API and Compose is now not using noncanonical decompositions.

I believe that my "Unicode INFO" code is now complete. In the end, it is about 12KB of data (6KB compressed and 6KB of 'fast tables' for the first 2048 codepoints).

Documentation needs updating. Then the next part would be updating / deprecating those ToLower/ToUpper routines for Strings, and most importantly, implementing "apparent character logic".