Subject: Re: Choosing the best way to go full UNICODE
Posted by cbpporter on Mon, 19 Jun 2017 08:22:08 GMT
View Forum Message <> Reply to Message

mirek wrote on Mon, 19 June 2017 11:03My 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).

That's why you read the spec!

Everything is convention based.

Compatibility decomposition and everything that is marked in compatibility in Unicode means that it would not be part of Unicode and has no reason to exist in a standalone standard, but it had to be added to be compatible with another standard.

Canonical is the only that is needed for comparing and search.

Compatibility decomposition and non-compatibility decomposition are separate entities with separate names and compatibility one should not be used unless you are trying to be compatible with another standard.

And the rest can be ignored. Like <font> substitutions:
http://www.fileformat.info/info/unicode/char/2102/index.htm

I really don't think that users expect that hollowed out C to return true when compared to a plain C.