Differences between Unicode NFD and MARC21

The MARC21 character repertoire for Unicode, generally uses decomposed character sequences, although there are a number of differences between the MARC21 repertoire and Unicode Normalization Form D (NFD).

Most character sequences are decomposed in MARC21, but a small set of Latin and Cyrillic base character and combining diacritics are considered as base characters in USMAR and retain their precomposed forms.

Codepoints in blue are the codepoints found in the MARC21 character repertoire, while the codepoints highlighted in red are the NFD equivalent representation.

Latin

Latin script normalisation differences between NFD and the MARC21 Character repertoire affect Vietnamese and a range of other languages spoken in Vietnam.

Uppercase		Lowercase	
Ơ		ơ	
01A0	004F 031B	**01A1**	008F 031B
Ư		ư	
01AF	0055 031B	**01B0**	0075 031B

Cyrillic

Cyrillic script normalisation differences between NFD and the MARC21 Character repertoire impact a range of languages including Belarusian, Bulgarian, Macedonian, Russian and Ukrainian.

Uppercase		Lowercase	
Й		й	
0419	0418 0306	**0439**	0438 0306
Ѓ		ѓ	
0403	0413 0301	**0453**	0433 0301
Ё		ё	
0401	0415 0308	**0451**	0435 0308
Ї		ї	
0407	0406 0308	**0457**	0456 0308
Ќ		ќ	
040C	041A 0301	**045C**	043A 0301
Ў		ў	
040E	0423 0306	**045E**	0443 0306

An example: https://hdl.handle.net/10079/bibid/13186283

In the title *Україна і Схід*, the word *Україна* uses a precomposed sequence: <U+0423 U+043A U+0440 U+0430 **U+0457** U+043D U+0430>. The decomposed form would be <U+0456 U+0308>.

While the romanised version of the title is, *Ukraïna i Skhid*, where Ukraïna uses the decomposed sequence: <U+0055 U+006B U+0072 U+0061 **U+0069 U+0308** U+006E U+0061>.