Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Request for priority allocation in Latin Extended-D and -E blocks in BMP**
Source: **Lorna Priest (SIL International) and Deborah Anderson (Script Encoding Initiative)**
Status: **Liaison Contribution**
Action: **For consideration by UTC and JTC1/SC2/WG2**
Date: **2012-06-04**

An SIL linguist recently requested the addition of an upper case Latin character (not in Unicode) to be a case pair with a lower case Latin character that is already in Unicode. This is for a language group that is just settling their orthography and moving into the digital age. This raised the question as to whether there will even be space in the BMP for this character, and if not, whether there would be problems with the case pair being split across planes.

As we began looking at Latin lower case characters with no matching upper case character, we came up with the following statistics:

- There are 190 Latin characters in Unicode with a General Category property of "Ll" and no Upper Case Equivalent.
    - 41 of those are "Small Capitals", and it seems unlikely an upper case would be need for a Small Capital.
    - An additional 14 have a character decomposition (where the decomposition is to something that already has a case pair).
    - There are 3 upper case characters in PDAM2 which have a lower case already in Unicode
    - This leaves 132 Latin characters where there is the potential for requiring an upper case character.
- Additionally, there are 45 Latin lower case characters in DAM1 where there is no upper case character encoded (there are none in PDAM2).
- This brings us to a total of 177 Latin characters where there is the potential for requiring an upper case character.
- There are currently 169 "free" spaces in the Latin Extended-D and Latin Extended-E blocks.

(For reference, there are also 1 Coptic lower case character, 1 Cyrillic lower case character and 2 Greek lower case characters in the BMP where there is no matching upper case characters. The Coptic and Greek blocks seem less problematical as there is not a rush to fill those blocks with new characters.)

Some software implementations include programming interfaces for simple case mapping that assume that the byte length of the input and output are the same, and case mappings between BMP and supplementary planes could create significant problems for such software.

Because of potential problems for simple case pairs across planes (such as a lower case character in the BMP and the matching upper case character in the SMP), we request WG2 and UTC give high priority for the remaining spaces in Latin Extended-D and Latin Extended-E to be used for characters required for current/modern orthographies and eventual case pairs.