Proposal to include Cyrillic and Latin Extended Blocks into the Roadmap (related to the current Osage proposal L2/14-175)

2014-08-01 – Karl Pentzlin – karl-pentzlin@acssoft.de

While the “Latin Extended” blocks in the BMP are completely (-A, -B, -C) or nearly (-D, _E) full, there are several larger sets of unencoded Latin letters which are candidates for encoding.
See e.g.

- Landsmålsalfabetet (WG2 N3555 = L2/08-428)
- Letters used in the former Soviet Union (WG2 N4162 = L2/12-045
- The current Osage proposal (WG2 N4587 = L2/14-175).
- Other candidates may be the Initial Teaching Alphabet or several dialectology character sets.

Therefore, the SMP is to be used for encodings of such sets.
The current Osage proposal proposes a new block “Latin Extended-F” at U+104B0…104FF, nearly filled by the proposed Osage characters, leaving no room for other sets as large as the ones listed before.

If the same kind of block assignment is continued for upcoming proposals (including revisions of the existing proposals for the sets listed above), this will end up in having several “Latin Extended-G/H/I/J…” blocks randomly scattered across the SMP.

As the existing Cyrillic blocks are completely full also, also new Cyrillic Extended blocks in the SMP are to be considered.

Therefore, I propose to make the new blocks “Latin Extended-F” and “Cyrillic Extended-C large enough for foreseeable addtions from the beginning. In detail, change the Roadmap as follows:

- At "The SMP is tentatively mapped out to the following zones:" change:
  0001E000-0001E7FF unassigned
  to:
  0001E000-0001E5FF Alphabetic Script extensions
  0001E600-0001E7FF unassigned
  (This leaves some "unassigned" place for other unrelated future allocations.)
- In this "Alphabetic Script extensions" zone, allocate two blocks:
  1E000-1E0FF Cyrillic Extended-C
  1E100-1E3FF Latin Extended-F
  which leaves 512 places in this zone for possible future allocations for other extensions of alphabetic scripts.
- If the current Osage proposal (WG2 N4587 = L2/14-175) is accepted, change the character allocation accordingly from U+104B0… to U+1E100…

By the way, then an informal guideline for future Latin letter proposals can be applied as follows:

- Use the BMP (i.e., the few remaining places in Latin Extended-D and -E) only for:
  - capital counterparts for lowercase letters which already are encoded in the BMP,
  - single letters which are used in current orthographies of living languages,
  - single letters which complete a larger set which is already encoded and completely contained in the BMP.