L2/09-193

2009-05-04

Title: Allocating additional space for right-to-left characters in the SMP

Authors: Roozbeh Pournader and Anshuman Pandey

Date: 2009-05-04

Introduction

During research on a set of proposals for encoding various right-to-left (RTL) scripts and characters in Unicode, the authors realized that the current allocation for RTL characters in the Supplementary Multilingual Plane (SMP) is insufficient.

In light of proposals submitted for the encoding of RTL characters and scripts, possible disunification of RTL scripts allocated in the SMP, and potential RTL mathematical and technical symbols not yet allocated or proposed, the authors believe it is necessary to dedicate additional space to RTL characters in the SMP.

The limited availability of dedicated allocations for RTL characters is not simply a roadmap issue. It has ramifications for implementation in terms of the Bidi Class of unassigned characters. Currently, the range of unassigned code points with default RTL directionality is limited (see below). It is likely that this space will be exhausted before the encoding of RTL scripts and characters is completed.

The authors recommend to remedy this situation quickly to allow implementers enough time to put into place default assignments.

Recommendation

Allocating extra space now for RTL scripts will establish default directionality assignments for unassigned code-points and will facilitate ease of implementation in the future.

In order to accommodate the requirement to encode additional RTL scripts and characters, another set of rows should be allocated in the SMP for RTL characters. A likely allocation would be U+1D800..1DFFF or U+1E000..1E7FF. This range should be reserved for RTL mathematical, technical, and numeric systems, with the possibility of being used for historical scripts as needed. The present allocations for Persian Siyaq Numerals and Arabic Mathematical Alphabetic Forms should be moved to that space, while retaining the RTL properties of the allocations so that historical RTL scripts may be encoded at those locations.

This recommendation will require that the Bidi Class of the unassigned code points in the RTL block be changed from L to R, as specified in the Unicode data file `DerivedBidiClass.txt`.

The authors believe that this allocation should happen sooner rather than later, in order to make sure implementations of Unicode could be ready in advance, in case they need architectural changes to be able to handle RTL characters encoded outside the previously reserved range.

Background

Presently, RTL characters are restricted to four zones, namely:[1]

- U+0590..08FF (BMP): right-to-left scripts in common modern use: Hebrew, Arabic, Syriac, Thaana, N’Ko, Samaritan, and Mandaic
- U+FB1D..FB4F (BMP): Hebrew Presentation forms

1 See also L2/09-120, “Allocating Unicode Characters” by Mark Davis.

- U+FB50..FDFF and U+FE70..FEFF (BMP): Arabic Presentation Forms
- U+10800..10F00 (SMP): historical RTL scripts and mathematical and numerical systems

The remainder of the document discusses certain character repertoires that the current allocations may not fully accommodate. This information is only meant to be exemplary: the authors believe that it is not exhaustive.

Siyaq numerals

According to the "Roadmap to the SMP", version 5.1.4 (dated 2009-02-04), eight columns are reserved for "Persian Siyaq numerals". Unfortunately, that accounts only for one of the four Siyaq traditions:

- Diwani numerals, used in the Arabic-speaking world (see L2/09-141)
- South Asian (Raqm) Siyaq numerals (see L2/09-148)
- Turkish Siyaq numerals, used in the Ottoman empire (see L2/09-166)
- Persian Siyaq numerals (proposal forthcoming).

These for Siyaq systems were originally proposed for unification (see L2/07-414, "Proposal to Encode Siyaq numerals in ISO/IEC 10646" by Anshuman Pandey).

Additional research has indicated that each Siyaq system possesses different shaping requirements, character typologies, and orthographic conventions for the notation of numbers. These differences suggest implementation difficulties associated with unification. They support the independent encoding of the numerals in each Siyaq tradition. The encoding model for each Siyaq system is dependent upon the above factors and is still to be determined.

Assuming the independent encoding for the Siyaq systems, the authors expect that six columns would be needed for each tradition, for a total of 24 columns (one and half rows). There is, unfortunately, not enough space allocated in the RTL area of the SMP for these characters.

Other scripts and symbols

Research suggests that additional RTL historical scripts and specialist characters may require encoding in the SMP. These include scripts that may be encoded independently if they are disunified from scripts presently allocated in the SMP, but for which formal proposals have not been submitted. For example, early draft proposals suggested a unified encoding for Book Pahlavi, Psalter Pahlavi, and Avestan; or similarly, for Imperial Aramaic, Inscriptional Parthian, and Inscriptional Pahlavi. These scripts were ultimately encoded independently.

Similarly, if evidence suggests that Sogdian should be disunified from Uighur or if more varieties of Aramaic are proposed, then space in the RTL block should be available to accommodate these scripts.

There also exist several traditions in the RTL writing world for mathematical and "scientific" symbols, which have not yet been fully investigated. These symbols are used for writing notations in geometry, geography, astronomy, astrology, alchemy, numerology, and charm-writing. The authors expect research on these symbols to begin after the more basic RTL requirements are fulfilled (eg. the encoding of historical scripts like Book Pahlavi).

Acknowledgments

The authors are indebted to Asmus Freytag for encouraging us to write this proposal, and later providing us with very valuable suggestions to significantly improve an earlier draft of this document.