Title: Proposal to restrict the range of code positions to the values up to $10FFFF_{16}$

Source: Unicode Technical Committee

Status: Liaison

Action: For consideration by JTC1/SC2/WG2

The Unicode Consortium proposes an amendment to ISO/IEC 10646 to restrict the range of code positions to the values of 0 to $10FFF_{16}$.

Background and Justification

After the merger of ISO/IEC 10646 and Unicode, the Unicode Consortium had the goal to keep the Unicode standard conformant with international standard ISO/IEC 10646. One remaining source of confusion between the two is that ISO/IEC 10646 has a code space up to $7FFFFFFF_{16}$ but that Unicode has restricted itself to the code space of UTF-16, namely up to $10FFFF_{16.}$ This difference in the code space between UTF-16 and the UTF-8 and UCS-4 encoding forms causes continued confusion among developers and users, and an appearance of a schism between Unicode and ISO/IEC 10646. Moreover, the situation presents unnecessary interoperability problems for implementers.

To resolve this situation, Unicode proposes that SC2/WG2 restrict the code space of UCS-4 and UTF-8 to the same domain as UTF-16, namely 0 to $0010FFF_{16}$, by amending ISO/IEC 10646 to exclude values above $10FFFF_{16}$, much as values above $7FFFFFFF_{16}$ are currently excluded. (See UTR #19: [Mark, simply enclose it as an attachment.] http://www.unicode.org/unicode/reports/tr19/.)

Concerns with the Proposal

The Unicode Consortium sees two potential problems with this recommendation: limitation of the size of the code space could eventually prevent coding future characters, and the possibility of implementations using the private use characters above $10FFFF_{16}$. Let's consider these two issues in sequence. First, the design of UTF-16 permits addresses up to $10FFFF_{16}$, which represents over 1,000,000 characters. When SC2/WG2 decided on this size, it was four times the estimated number of characters to be encoded in ISO/IEC 10646. At this time, it appears that this range of code positions is sufficient for all foreseeable character allocations. Thus, the limitation in the character space should not be of concern. Second, some implementations may make use of the private use characters above $10FFFF_{16}$ from 60000000_{16} to $7F000000_{16}$. At this time, the Unicode Consortium is aware of no implementations that make use of these private use characters. Therefore, use of private use characters about $10FFFF_{16}$ should also not be of concern.

Details of the Changes to the ISO/IEC 10646 Standard

This proposal is to amend ISO/IEC 10646 to exclude values above $0010FFFF_{16}$. This affects the following specific areas of the standard:

- In Section 7, paragraph 1, the values of G-octets are restricted to being precisely zero, and the values of P-octets restricted to the values from 00 to 10_{16}.

- In Section 10.2, the statement reserving the code positions of the 32 groups from Group 60 to Group 7F for private use is withdrawn, replaced by a statement that the use of code positions of Group 60 through Group 7F is deprecated.
- In Annex D (UTF-8) appropriate deletions are made to limit the format to 4 bytes, with additional edits to correct byte ranges.