Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: ISO/IEC JTC1/SC2/WG2/IRG

Title: Horizontal Extension Proposal

Source: Dr. Ken Lunde (小林劍), Adobe Systems Incorporated

Status: Individual Contribution

Action: For consideration by the IRG

Date: 2014-11-24

The purpose of this brief document is to propose a method for submitting horizontal extensions in a way that does not involve the publishing of new national standards and assigning new character codes, which are processes that can be associated with bureaucratic hurdles, logistical overhead, and other obstacles.

Background documents include IRG N1964 (aka L2/13-192) from IRG #41 (*Continued National Standards Development* & *Horizontal Extensions*) and IRG N2050 from IRG #43 (*Recommendations of IRG Meeting #43*), specifically referring to Recommendation IRG M43.2 (*Horizontal Extension of H Source for existing characters*).

The Problem

The number of CJK Unified Ideographs in Unicode and ISO/IEC 10646—when Extension E is factored in—now exceeds 80,000 characters (the exact figure is 80,379). The sheer number of characters poses the following two real-world problems to implementers of the standard, particularly font developers:

1. A single font resource cannot include more than 64K glyphs (though there is now a standard, ISO/IEC 14496-28:2012, that provides a work-around via Composite Font Representation objects, but it is not yet universally supported).

2. For reasons of practicality, most fonts—including Pan-CJK ones—are intended for use in one or more regions, locales, or languages, and thus there is no benefit in including in them glyphs for all 80,379 CJK Unified Ideographs.

Therefore, the fundamental problem is in determining which particular CJK Unified Ideographs require glyphs in fonts that are intended for use in a particular region, locale, or language.

Horizontal Extensions

When a new CJK Unified Ideograph is proposed by member bodies—and ultimately accepted into the standard—it includes a source reference that can be used to tie the character to a particular region, locale, or language via one or more *kIRG_{G,T,H,M,J,K,KP,V,U}Source* references.

However, it is not uncommon for an existing CJK Unified Ideograph to later be deemed useful for additional regions, locales, or languages. This is performed by submitting a "horizontal extension" whose effect is to tag, flag, or otherwise identify a CJK Unified Ideograph as being useful for an additional region, locale, or language.

When submitting horizontal extensions, a unique *kIRG_{G,T,H,M,J,K,KP,V,U}Source* reference is necessary. These unique source references correspond to code points in published national standards.

The Proposed Solution

As a way to eliminate the overhead involved in publishing new national standards for the purpose of horizontal extensions, I propose the use of the "U" identifier immediately following the source prefix, and followed by a zero-padded five-digit hexadecimal character code of the UCS code point to which it corresponds.

This solution has the following benefits:

- Explicitly indicates that no corresponding national standard exists, so font developers need not bother searching for one.
- Eliminates the need to publish new national standards to accommodate horizontal extensions.
- Explicitly indicates that the character has been deemed useful for a particular region, locale, or language.
- Provides the representative glyph for the region, locale, or language.
- Promotes the code charts as the most up-to-date and authoritative reference for representative glyphs across all regions, locales, and languages.

Others have suggested putting this information into an appendix or other document, but doing so is less useful, and easily overlooked. Having the information closely tied to the character, such as in the multicolumn code charts and as the *kIRG_{G,T,H,M,J,K,KP,V,U}Source* reference, is more convenient for developers and less likely to be overlooked.

Of course, horizontal extensions that use this method would still require that representative glyphs be provided for use in the corresponding column of the code charts.

Specific Use Cases

Hong Kong SAR is the primary beneficiary of this proposal because it has thus far identified several hundred CJK Unified Ideographs that have been deemed useful for Hong Kong, yet they lack a *kIRG_HSource* reference. Font developers have no convenient way of knowing what these additional characters are.

ROK (South Korea) submits horizontal extensions on a somewhat regular basis, which are tied to obscure KS standards that seem to exist for the sole purpose of submitting horizontal extensions.

Japan would be in a position to replace the existing "JA-*XXXX*" source references—whose corresponding standard cannot be found—with "JU-*HHHHH*" ones, and could also propose "JU-20BB7" as the kIRG-JSource reference for U+20BB7 (𠮷).