# 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 (Revised on 2015-05-07)

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 or updated 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 80,388 for Unicode Version 8.0). 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 little or no benefit to include in them glyphs for all 80,000+ 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. A secondary problem is conveying to font developers the corresponding representative glyphs.

## 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, which are normative.

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 or updated national standards for the purpose of horizontal extensions, I propose the use of the "U" (an abbreviated form of "UCS") identifier immediately following the source prefix, and followed by a four- or five-digit hexadecimal character code of the UCS code point to which it corresponds.

This solution has the following benefits:

- Implicitly indicates that no corresponding national standard exists, so font developers need not bother searching for one.
- Eliminates the need to publish new or updated national standards to accommodate horizontal extensions.
- Explicitly indicates that the character has been deemed useful for a particular region, locale, or language.
- Provides a 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 is very easily overlooked. Having the information closely tied to the character, such as in the multiple-column 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, precisely because their glyphs may differ from the representative glyphs in the other columns.

Nothing in this proposal would prevent national bodies to continue issuing new or updated national standards for the purpose of horizontal extensions, if they choose to do so.

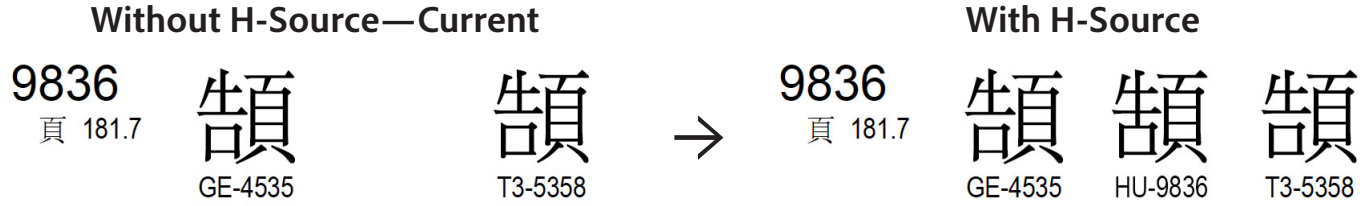
## Specific Use Cases & Examples

**Hong Kong SAR** is the primary beneficiary of this proposal because it has thus far identified a couple hundred existing CJK Unified Ideographs that have been deemed useful for Hong Kong SAR, yet they currently lack a *kIRG_HSource* reference (these are listed in the Not-Accepted.pdf file that is associated with the Hong Kong SCS standard). Font developers have no convenient way of knowing what these additional characters are, and to what extent their representative glyphs differ from those in the other columns in the code charts.

Attached to this proposal is a text file, *unihan-h-221.txt*, that represents the data portion of a complete horizontal extension that incorporates the 221 characters that are outside the scope of Big Five and Hong Kong SCS, but are already encoded in Unicode and ISO/IEC 10646, and deemed useful for Hong Kong SAR. Below is an excerpt of what would be included in the Unihan Database, using U+9836 as an example:

```
U+9836      kIRG_HSource      HU-9836
```

The example below shows the current code chart entry for U+9836 that lacks a representative glyph and source reference for the H column, along with a mockup that shows the same entry after the horizontal extension is applied:




**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. This proposal could potentially eliminate the need to publish additional KS standards simply for this purpose.

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

```
U+20BB7     kIRG_JSource      JU-20BB7
```

That is all.