Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type:	**Working Group Document**
Title:	**Comments on issues raised in N4021**
Source:	**UK**
Status:	**National Body Contribution**
Action:	**For consideration by JTC1/SC2/WG2 and UTC**
Date:	**2011-05-22**

1. Summary

N4021 (IRG Meeting #36 Resolutions) notes IRG concerns about a proposed Ideographic Variation Database (IVD) registration of 21 simplified CJK characters by the UTC (IRGN1757, L2/11-109), and having reviewed this document, we believe that the proposed IVD registration is effectively introducing a new encoding model for simplified CJK characters through the backdoor, fundamentally changing the way processes and users interact with CJK ideographs. We think that an issue of this nature should be addressed by WG2 as well as at IRG.

The UK believes that the proposed IVD registration will add an unnecessary burden to processes dealing with conversions between simplified and traditional form characters with no appreciable benefit; and will cause unnecessary confusion and inconvenience for the CJK user community. We further note that six of the 21 proposed IVS sequences correspond to simplified characters that have already been encoded in CJK-C. The UK NB therefore endorses the IRG request to the UTC not to proceed with this registration.

2. Scope of the Ideographic Variation Database

As admitted in IRGN1757, there is currently an expectation that the IVD is only used to register glyph variants that are unifiable under the Annex S rules, and as a precedent non-unifiable glyph variants in the original ADOBE-JAPAN1 proposal were encoded as characters. We believe that changing the scope of the IVD to allow registration of non-unifiable variants is dangerous as it allows for the possibility of accidentally registering IVS sequences corresponding to characters that are already encoded (as shown below, six of the twenty-one proposed IVS sequences in IRGN1757 correspond to existing characters) or for characters that may be proposed for encoding at a later date (ten of the twenty-one proposed IVS sequences in IRGN1757 correspond to G-source characters already scheduled for inclusion in Extension B). Having two different ways to represent the same character is undesirable, and should be avoided if at all possible, but allowing IVS registrations for non-unifiable variants increases the possibility of creating such duplicate representations, and imposes a further burden on quality control for CJK encoding. We therefore request that the UTC confirm that the IVD shall only be used for registering glyph variants that are unifiable under the Annex S rules.

3. Costs and Benefits of an IVS Solution

IRGN1757 argues that the use and maintenance of mapping tables for simplified/traditional characters is inefficient and has a performance impact on processes that do CJK searches or otherwise need to process CJK data. However, whilst we agree that an IVS solution to simplified characters may have been a viable and more efficient alternative to encoding had it been implemented at early stage in the history of the UCS (before CJK-A, B, C and D were encoded), we believe it is now far too late to introduce an alternative encoding model for CJK characters at this stage. Introducing an IVS solution for a handful of simplified CJK characters will not obviate the continuing need for processes to use and maintain mapping tables for existing simplified and traditional characters; however it will require such processes to rewrite the mapping tables and code that processes the mapping tables to deal with the few exceptional cases of traditional characters that map to an IVS sequence of two characters rather than to a single character. It will also not reduce the need for maintenance of such tables, as new IVS sequences for simplified characters may be introduced in the future, in exactly the same way that new characters may be encoded. Therefore, from an industry perspective, the existence of two competing models for representing simplified CJK characters brings no appreciable benefit, but potentially has very significant implementation costs.

From an end user perspective, the proposed IVS solution is also unhelpful. Users would have to become proficient in manipulating variation selector characters in their text, adding them after characters that they want to simplify, and deleting them from after characters that they want to convert to traditional characters. For the vast majority of users this would not be a simple exercise, and would inevitably lead to variation selector characters being inserted into text in the wrong places, resulting in invalid or unexpected IVS sequences. Handling of variation sequences is made even more complicated by the fact that some characters require VS-17 whereas other characters require VS-18, so users cannot simply treat a particular variation selector character as "simplifier"; and if they use the wrong variation selector character then the glyph they see may change in an unexpected way. This can only be a source of frustration for end users. Furthermore, users expect to be able to find obscure characters that they cannot enter using their preferred IME by means of code charts and character map applications, but as IVS representations of simplified characters would not be given in the code charts or most character map applications, most users will not even realise that an IVS representation of the character they are looking for exists, never mind how they can generate the required IVS sequence. All in all, the use of IVS sequences to represent a handful of simplified characters can only detract from the end-user experience.

4. Analysis of Proposed IVS Sequences

IRGN1757 proposes to register 21 IVS sequences, as shown in the table below. One of the proposed sequences has already been defined in the ADOBE-JAPAN1 collection, and six of the proposed sequences are for characters that have already been encoded in CJK-C. Ten of the other proposed IVS sequences are for characters already scheduled for inclusion in Extension E that have G-source references. The fact that almost one third of the proposed IVS sequences are for characters that are already encoded raises serious questions about the safety of IVD registration for characters that are not unifiable glyph variants.

IVS	UTC Ref	Trad	Simp	IDS	Comments
5D19 E0101	UTC-00668	崙	岑	⿱山仑	**Already encoded as U+2AA27**
7A68 E0100	UTC-00669	穩	稳	⿰禾贵	
7D41 E0101	UTC-00029	紲	绁	⿰纟世	**Already encoded as U+2B11F**
7D9D E0101	UTC-00914	綝	绡	⿰纟林	Ext.E O6431
8A0F E0100	UTC-00071	訏	讠	⿰讠于	Ext.E 07696 (G_CH501580) IVS already defined as Adobe-Japan1 CID+15137
8B30 E0101	UTC-00030	謰	谰	⿰讠连	Ext.E 07776 (G_CH301637)
8B46 E0101	UTC-00675	譆	谐	⿰讠喜	**Already encoded as U+2B37B**
8B54 E0101	UTC-00676	譔	馔	⿰讠巽	Ext.E 07850 (G_XC301692)
8F36 E0101	UTC-00024	輶	辎	⿰车酋	Ext.E 08253 (G_CH501859)
91B2 E0101	UTC-00038	醲	酴	⿰酉农	Ext.E 08520 (G_CH401963)
9265 E0101	UTC-00052	鈰	铈	⿰钅术	Ext.E 08748
96A4 E0101	UTC-00674	隤	隤	⿰阝贵	Ext.E 09079 (G_CH502228)
982B E0101	UTC-00677	頫	頫	⿰兆页	**Already encoded as U+2B5AF**
992C E0101	UTC-00678	餬	馎	⿰饣胡	**Already encoded as U+2B5EB**
99BC E0101	UTC-00842	駮	驳	⿰马文	**Already encoded as U+2B61C**
9A23 E0101	UTC-00679	騣	骔	⿰马叕	Ext.E 09650 (G_XC102508)
9D4F E0100	UTC-00117	鶻	鹘	⿰甫鸟	
9DB1 E0101	UTC-00680	鶱	鹫	⿱寅鸟	Ext.E 10287 (G_CH402876)
9DC3 E0101	UTC-00061	鷃	鹌	⿰晏鸟	
9DC7 E0101	UTC-00068	鷇	鹅	⿰⿱士冖一鸟殳	Ext.E 10312 (G_GJZ00295)
9F6E E0101	UTC-00013	齮	龃	⿰齿奇	Ext.E 10486 (G_CH302990)