

ISO/IEC JTC1/SC2/WG2/IRG IRG N1468

Title: Recommendation For IRG To Use IVD Collections*Richard Cook & Ken Lunde**11 June 2008***Background**

There is a pending UTC/US proposal that *CJK Compatibility Ideographs* henceforth be handled by IVSes. CJK Compatibility Ideographs served as a mechanism for effectively encoding otherwise unifiable characters with the intention to preserve distinctions. However, given the fact that the distinctions that are associated with CJK Compatibility Ideographs can easily be lost, due to the fact that *Normalization* may be applied at any time by any client, thus rendering them as their *Canonical Equivalents*, this decision is both prudent and practical.

Ideographic Variation Sequences (IVSes) are now a recognized part of the standard, as of Unicode Version 5.1. Furthermore, the infrastructure to support IVSes is rapidly being developed. Although not yet fully supported in Mac OS X, the *Variation Selector* (VS) component of an IVS is now handled correctly: it is not displayed, but it is also not discarded. In addition, Adobe Systems has provided IVS support in two of its key technologies, Acrobat (Version 9.0) and Flash Player (Version 10).

Although the infrastructure to support IVSes is still being developed, the “plain text” nature of IVSes allows them to survive and persist in almost every conceivable environment. Worst case, the IVS displays using only its *Base Character* (BC) component, which retains important properties of the character.

Recommendation Details

UTS #37 (*Ideographic Variation Database*) establishes a process for registering and thus standardizing IVSes, which has already be exercised by Adobe Systems for its “Adobe-Japan1” IVD collection.

Current IRG procedures effectively allow for only two paths for characters that are submitted for consideration as *CJK Unified Ideographs*, as follows

1. To disunify and thus encode as a new character
2. To unify and thus not to encode

IVSes allow for a third choice, as follows:

3. To unify, but to establish an IVS for the character

This latter treatment may be desirable if the unification rules and principles dictate or suggest unification, but the submitting entity feels strongly about including the character in Unicode in a fashion that allows distinctions to be preserved in a “plain text” environment.

The IRG itself could submit IVD collections, on a regular basis, as a way to handle otherwise unifiable characters. This would serve to better adhere to unification rules and principles, and at the same time satisfy the encoding needs of national body members and other entities that submit new characters. After all, submitting a new character naturally implies a desire to have it encoded.

In addition, national body members and other entities could submit new characters to be considered only as IVSes, as opposed to submitting them as CJK Unified Ideographs, if it is obvious or clear that they would otherwise be unified.

Base Character & IVS Process

The table on the following page details the proposed decision-making process for IRG's use of IVSes in its process of accepting ideographs.

"First Project" Recommendation

As a way to initiate the use of IVSes by the IRG, we recommend that the first project be to establish IVSes for the existing CJK Compatibility Ideographs. This project would serve two purposes, detailed as follows:

- *As a learning exercise*, the IRG members will discover first-hand how IVSes can be used and applied as part of its important work.
- *As a practical result*, the distinctions that were intended to be preserved by CJK Compatibility Ideographs can now be preserved through the use of IVSes.

Note that because a small number of CJK Compatibility Ideographs have more than one source, they may require more than one IVS in order to preserve multiple distinctions, if the distinctions that were meant to be preserved were different. Below are some examples for which the likelihood of multiple IVSes is high:

U+F907	kIRG_HSource	8BF8
U+F907	kIRG_KSource	0-5022
U+F929	kIBMJapan	FAE0
U+F929	kIRG_JSource	3-754E
U+F929	kIRG_KSource	0-5228
U+F936	kIRG_JSource	3-7B4F
U+F936	kIRG_KPSource	KP1-70DC
U+F936	kIRG_KSource	0-5249
U+F9DC	kIBMJapan	FBE9
U+F9DC	kIRG_JSource	3-7D5D
U+F9DC	kIRG_KSource	0-6B58
U+FA10	kIBMJapan	FA9C
U+FA10	kIRG_JSource	3-2F57
U+FA15	kIBMJapan	FB58
U+FA15	kIRG_JSource	3-775A

Ideographic Base and VS encoding process

Is the candidate UCS ideograph:

(1) already *encoded* in UCS (as base or VS)? ⇒

standard formal *descriptions* of candidate ⇒

// IDS, CDL, etc.

standard *determination* of formal encoding status ⇒

result:

A: yes ⇒ go to #3;

// form is *identical* to one or more encoded ideographs

B: maybe ⇒ go to #2;

// form is *similar* to one or more encoded ideographs

C: no ⇒ go to #2;

// form is *unlike* any encoded ideograph

(2) a *variant* of an encoded UCS character? ⇒

authoritative lexical *analyses* of variant status ⇒

// dictionary or other authority asserts variant relation

authoritative *determination* of variant status ⇒

result:

A: yes ⇒ encode by VS;

// form is a *member* of a known *varclass*

B: maybe ⇒ encode by VS;

// form is *similar to a member* of a known *varclass*

C: no ⇒ encode as **new base**;

// form is *unlike any member* of any known *varclass*

(3) *separated* in a lexical or encoding source? ⇒

authoritative *analyses* of source separation status ⇒

// justify preservation of separation: lexical distinction, round-tripping, etc.?

authoritative *determination* of source separation status ⇒

result:

A: yes ⇒ encode by VS;

// form is duplicated *yet distinct* in a lexical or other source

B: maybe ⇒ encode by VS;

// form is duplicated and *possibly distinct* in a lexical or other source

C: no ⇒ stop;

// source separation is *non-distinctive* and need not be preserved