ISO/IEC JTC1/SC2/WG2/IRG N1757

Preliminary Proposal for an Ideographic Variation Database Registration

John H. Jenkins 井作恆
Richard Cook 曲理查
Ken Lunde 小林劍
Unicode Technical Committee
6 April 2011

Summary

This documented is being presented to IRG members for information only. No action is required.

The Unicode Technical Committee maintains a database of Han ideographs which have been brought to its attention as potential candidates for encoding. This database is found in Unicode Technical Report #45, U-source Ideographs (UTR 45, found at http://www.unicode.org/reports/tr45/). All U-source indices refer to this document, but a fuller description of the sources for these ideographs is found in Unicode Standard Annex #38, Unicode Han Database (UAX 38, found at http://www.unicode.org/reports/tr38/).

The UTC is currently investigating the possibility of representing a number of the unencoded ideographs in UTR 45 via registered ideographic variation sequences (IVSs) as described in Unicode Technical Standard #37, Unicode Ideographic Variation Database (UTS 37, found at http://www.unicode.org/reports/tr37/). The actual database of registered variation sequences is referred to as the Ideographic Variation Database (IVD).

This document summarizes the current state of that investigation.

Rationale

The forms included in this proposal are all completely synonymous with the base characters in question. Usage of one form rather than another reflects locale-specific preference, and any differences in glyph shape are otherwise irrelevant to UCS encoding.

The proposed forms are all regular modern simplifications of traditional Chinese characters, produced by applying well-known simplification rules. The cases presented here involve only one-to-one mappings, and conversion from one form to the other is completely reversible.

There is currently an expectation that registered variation sequences cover only forms which are unifiable under Annex S rules (see UTS 37 §2). Registering the current collection would involve a broadening of current guidelines to explicitly allow unification of so-called “y-variants” under certain conditions.

(We note that Annex S is itself an informative part of ISO/IEC 10646 and not normative. The recommendation to use Annex S as a guideline for determining whether or not two ideographs can be unified is also informative within UTS 37.)

Annex S describes variation of ideographs using a three axis model: the x-axis relates to differences in meaning, the y-axis relates to relatively large differences in shape, and the z-axis relates to relatively minor differences in shape. The process of determining whether two synonymous ideographs are y- or z-variants of one another is described in detail in Annex S.

Characters with very similar shapes but very different usage (differing in meaning, along the x-axis) are not unifiable. For example, the characters 未 (U+672A) and 末 (U+672B) may be written with very similar forms, but have different usage. They are divergent o栢n the x-axis (clearly separated in dictionaries), and so are not unifiable. Even though they may easily be confused with one another, that alone has not prevented them from being separately encoded.

Ideographs with only very minor shape differences (such as 花 and 花—U+82B1 as drawn with two different “kai” typefaces) are z-variants as defined by Annex S and may be unified.

Characters with larger shape differences but identical meanings are termed “y-variants.” An example would be 貝 (U+8C9D) and 贝 (U+8D1D), which are a traditional Chinese/ simplified Chinese pair. The general approach in developing the UCS, applied to every script other than Han, would be to unify these two forms. Annex S currently includes no guidelines for how it could be done now for CJKV ideographs. Adding such guidelines to a future revision of Annex S seems desirable.

As a rule, CJKV y-variants have hitherto been separately encoded, largely for historical reasons. This unnecessarily complicates the encoding, overburdens the IRG and computing processes, and has no practical benefit for anyone. Indeed, the encoding of duplicate characters adds significant cost to implementers and frustrates end-users.

Example 1

To illustrate, let’s start by looking at the screen shot below.

This screen shot is from a Google search for the phrase “香港科幻小說書店” (“Hong Kong science fiction bookstores”). Two characters in the search phrase (說 and 書) have standard simplifications (说 and 书), and Google has accommodated this, with a search result containing the traditional forms on top and a search result containing the simplified forms on the bottom.

In order to do this, Google needs to have a table of character equivalents. Having to double-check equivalents via a table has a performance impact. It also means that the table of variants has to be continually updated whenever a new CJK Unified Ideographs Extension is encoded.

Moreover, since there is no authoritative source for variant data, different vendors will have different tables and different degrees of support for this feature. This detracts from the end-user experience.

If y-variants are represented via IVSs, however, the variation relationship is automatically encoded in the text itself. If you want one y-variant to match any of the others, you simply ignore the variation selector. If you want one y-variant *not* to match any of the others, you take the variation selector into account. This doesn’t even require that you explicitly add support for any particular variation sequence; it’s simply a matter filtering out variation selectors via a range check or not filtering them out.

Example 2

Searching is not the only process to benefit. Even though fonts do need to be updated whether you use separate encoding or a registered variation sequence, display of text also works better with the use of variation sequences.

For example, consider the line 訏謨定命 from the Chinese classic, the *Shi Jing*.[1] The first two characters both have potential simplifications: 訏 to 讦, and 謨 to 谟, but only the latter is currently encoded. The simplified Chinese form of the line is therefore

[1] From Ode 256 (蕩之什，抑); it’s translated as “He who takes counsel widely, is final in his commands” by Waley.

usually given as 訏谟定命. The form 讦 is attested, however, and a simplified Chinese book may want to write it 讦谟定命.

If 讦 is separately encoded and a particular user hasn't updated their fonts, what they will see is something like 〓谟定命. If 讦 is represented using a variation sequence and the user hasn't updated their fonts, they will see 訏谟定命, which is distinctly better.

(And a text-to-speech engine would know to pronounce 讦谟定命 as *heoi1 mou4 ding6 ming6* without needing to be specifically updated as well.)

Summary

In general, representation of y-variants with IVSs provides for a better user experience and a more robust and flexible representation of text. It goes without saying that it also reduces the IRG's workload since fewer characters need be submitted for consideration. It would also reduce the workload of specific IRG members—most notably China—because the amount of effort they expend producing and tracking their encoding proposals is reduced.

The characters in this proposal are not common, and use of one y-variant or the other is not mandatory in current practice. Whereas 書店 would be considered "wrong" in a simplified Chinese text, 訏谟定命 is perfectly acceptable, even though 讦谟定命 would be preferred if it's available. There is therefore no significant difficulty in using variation sequences to represent these forms—but there is a distinct benefit.

Sources

Detailed descriptions of the sources for these forms are found in UAX 38 and UTR 45. These sources include:

- *ABC Chinese-English Comprehensive Dictionary*. John DeFrancis, ed. University of Hawai'i Press, 2003. ISBN: 0-8248-2766-X.
- 《漢語大字典》, 湖北辭書出版社, Wuhan, 1988. ISBN: 7-5403-0030-2/H.16. [<http://www.unicode.org/reports/tr38/#kHanYu>.]
- 《现代汉语词典》 [Xiàndài Hànyǔ Cídiǎn 'Modern Chinese Dictionary']. 中国社会科学院语言研究所词典编辑室编 [Chinese Academy of Social Sciences, Linguisitics Research Institute, Dictionary Editorial Office, eds.]. 北京: 商务印书馆, 2007 [第 5 版; 2007 年 11 月北京第 377 次印刷. ISBN: 7-100-04385-9/H.1100.]
- 《现代汉语词典》 [Xiàndài Hànyǔ Cídiǎn 'Modern Chinese Dictionary']. 中国社会科学院语言研究所词典编辑室编 [Chinese Academy of Social Sciences, Linguisitics Research Institute, Dictionary Editorial Office, eds.]. 北京: 商务印书馆, 1983 [1978 年

12 月第 1 版; 1983 年 1 月第 2 版; 1984 年 1 月北京第 49 次印刷印张 54; 统一书号: 17017.91]. [<http://www.unicode.org/reports/tr38/#kXHC1983>.]

- 《形音義規範字典》 Xíng-yīn-yì Guīfàn Zìdiǎn. 主编:李行健. 臺北市:五南圖書出版股份有限公司;語文出版社 (《現代漢語規範字典》), 2003. ISBN: 957-11-3317-5. [Taiwanese stroke order; also contains PRC simplified characters, but without stroke diagrams; cp. XHG.]

- 《现代汉语规范字典》 Xiàndài Hànyǔ Guīfàn Zìdiǎn. 主编:李行健. 北京: 語文出版社 (《現代漢語規範字典》), 1998. ISBN: 7-80126-346-4/H.76. [1998 年 1 月; earlier PRC version of ROC 《形音義規範字典》.]

- 文林 Wénlín Software for Learning Chinese, Version 4.0.1. Wenlin Institute: Eureka, California. [includes electronic ABC English-Chinese/Chinese-English Dictionary, DeFrancis et al.]

References

IRG N1468 Recommendation For IRG To Use IVD Collections
http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg30/IRGN1468IVS_Recommendation.pdf

IRG N1646 Principles and Procedures
http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg34/IRGN1646Confirmed.doc

IRG N1765 Current Status of IVS Support in OSes & Application

UAX #38: Unicode Han Database (Unihan)
http://www.unicode.org/reports/tr38/

UTR #45: U-Source Ideographs
http://www.unicode.org/reports/tr45/

Appendix A
Text Data to Add to IVD Database Files

We append below a summary of the data to be added to the two text files included in the IVD.

Addition to IVD_Collections.txt

```
UTC; UTC-[0-9]{5}; http://www.unicode.org/reports/tr45/
```

Additions to IVD_Sequences.txt

```
5D19 E0101; UTC; UTC-00668
7A68 E0100; UTC; UTC-00669
7D41 E0101; UTC; UTC-00029
7D9D E0101; UTC; UTC-00914
8A0F E0100; UTC; UTC-00071
8B30 E0101; UTC; UTC-00030
8B46 E0101; UTC; UTC-00675
8B54 E0101; UTC; UTC-00676
8F36 E0101; UTC; UTC-00024
91B2 E0101; UTC; UTC-00038
9265 E0101; UTC; UTC-00052
96A4 E0101; UTC; UTC-00674
982B E0101; UTC; UTC-00677
992C E0101; UTC; UTC-00678
99BC E0101; UTC; UTC-00842
9A23 E0101; UTC; UTC-00679
9D4F E0100; UTC; UTC-00117
9DB1 E0101; UTC; UTC-00680
9DC3 E0101; UTC; UTC-00061
9DC7 E0101; UTC; UTC-00068
9F6E E0101; UTC; UTC-00013
```

Appendix B
Mappings and Glyphs

We append below a summary of the proposed IVSs, showing the glyphs for the composed form as shown in UTR 45 and, in parentheses, the glyph for the base form.

<U+5D19 U+E0101> == UTC-00668 崙󠄁 (崙)
<U+7A68 U+E0100> == UTC-00669 穨󠄀 (穨)
<U+7D41 U+E0101> == UTC-00029 絁󠄁 (絁)
<U+7D9D U+E0101> == UTC-00914 綝󠄁 (綝)
<U+8A0F U+E0100> == UTC-00071 訏󠄀 (訏)
<U+8B30 U+E0101> == UTC-00030 謰󠄁 (謰)
<U+8B46 U+E0101> == UTC-00675 譆󠄁 (譆)
<U+8B54 U+E0101> == UTC-00676 譔󠄁 (譔)
<U+8F36 U+E0101> == UTC-00024 輶󠄁 (輶)
<U+91B2 U+E0101> == UTC-00038 醲󠄁 (醲)
<U+9265 U+E0101> == UTC-00052 鉥󠄁 (鉥)
<U+96A4 U+E0101> == UTC-00674 隤󠄁 (隤)
<U+982B U+E0101> == UTC-00677 頫󠄁 (頫)
<U+992C U+E0101> == UTC-00678 餬󠄁 (餬)
<U+99BC U+E0101> == UTC-00842 馼󠄁 (馼)
<U+9A23 U+E0101> == UTC-00679 騣󠄁 (騣)
<U+9D4F U+E0100> == UTC-00117 鵏󠄀 (鵏)
<U+9DB1 U+E0101> == UTC-00680 鶱󠄁 (鶱)
<U+9DC3 U+E0101> == UTC-00061 鷃󠄁 (鷃)
<U+9DC7 U+E0101> == UTC-00068 鷇󠄁 (鷇)
<U+9F6E U+E0101> == UTC-00013 齮󠄁 (齮)

ISO/IEC JTC1/SC2/WG2/IRG N1757_ROKFeedback
Doc. #: Korea JTC1/SC2 K2051

Korea JTC1/SC2, Committee on Character Codes

KIM, Kyongsok, Chairperson of Korea JTC1/SC2
email: gimgs0 AT gmail DOT com, phone: +82-51-510-2292
address: Division of Computer Science and Engineering,
Pusan National University
2 Busandaehagro 63 Beon Gil, Geumjeonggu, BUSAN 609-735, Rep. of KOREA

Author: KIM, Kyongsok
Date: 2011.06.01.
Status: NB position
Subject: R.O.Korea's comments RE: IRG N1757 (UTC Preliminary Proposal for an IVD Registration)
Relevant documents: IRG N1757,

1. A related IRG resolution (M36.2):

Resolution IRG M36.2: UTC Preliminary Proposal for an IVD Registration (IRGN1757)

Unanimous

The IRG has reviewed IRGN1757 from the UTC and is concerned that an IVD registration from the UTC might be treated as a de facto encoding. The IRG requests that the UTC not proceed with the registration until this issue is addressed.

The IRG encourages its members to do further review of IRGN1757 and provide feedback to the UTC at any time.

2. R.O.Korea's feedback:

R.O.Korea discussed this issue within domestic committee and concluded as follows:

- In CJKU, y-variants such as simplified Chinese characters were encoded separately from the traditional ones (base characters).
- If y-variants should be registered with IVD, some y-variants (simplified) will be using the same code position as traditional ones while other y-variants (simplified) will be using code position different from traditional ones.
- R.O.Korea is concerned that this will cause confusion to the users.
- Furthermore, in the future, there will be confusion as to whether y-variants should be given a separate code position or registered with IVD.

- As a conclusion, R.O.Korea suggests that y-variants be given a separate code position instead of registering y-variants with IVD.

＊＊＊