Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC 1/SC 2/WG 2/IRG N 1648Confirmed
Date: 2011-11-10
~~2010-03-16~~

Source:	IRG
Meeting:	~~After IRG#33, Danang, Vietnam~~
Title:	After IRG#36, Mountain View, CA, USA
Status:	
Actions required:	~~IRG Standing Document Summary~~
Distribution:	IRG Working Document Series(IRGWDS)
Medium:	Member's submission
Pages:	Feedback Requested IRG Electronic 5

Introduction:

IRG Working Document Series(~~IWDS~~ IRGWDS) is a set of IRG maintained documents which keeps the up-to-date ~~example~~ lists of CJK unification related example cases to supplement the published Annex S of ISO/IEC 10646 for IRG unification work. The maintenance of the IWDS should comply with the operational procedures established in Annex E of the *IRG Principles and Procedures*.

In addition to this document(labeled as SUM) which explains the IWDS with a summary of the examples, there are also 3 document series as listed below:

Series 1: List of UCV (Unifiable Component Variations) of Ideograhs(labeled as UCV)
Series 2: List of Non-Unifiable Components of Ideograph and Overly-Unified Ideographs(Labeled as UCV)
Series 3: List of Possibly Mis-Unified Ideographs(Labeled as MUI).

~~File format of IWDS:~~ Names of IRGWDS files:

Each ~~of IWDS~~ IRGWDS file is named as ~~IWDS~~ IRGWDS_*SSS*_*II* where

Formatted: Highlight
Formatted: Strikethrough
Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean
Formatted: Strikethrough
Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean
Formatted: Strikethrough
Formatted: Font: 12 pt
Formatted: body, Line spacing: single
Formatted: Strikethrough
Formatted: Strikethrough
Formatted: Strikethrough
Formatted: Font: Bold
Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean
Formatted: Font: (Default) Times New Roman
Formatted: Font: (Default) Times New Roman, Strikethrough
Formatted: Font: (Default) Times New Roman
Formatted: Font: (Default) Times New Roman, Strikethrough
Formatted: Font: (Default) Times New Roman
Formatted: body, Line spacing: single

SSS is the name of the specific series and *II* refers to the IRG meeting where the list is confirmed. Thus, ~~unification example lists in the series~~ file names of lists of unification/separation examples which are updated from time to time will ~~be taking the following format~~ look like:

IRGWDS_SUM_nn: a summary of IRGWDS documents
IRGWDS_UCV_nn: UCV list (Series 1)
IRGWDS_NUC_nn: Non-Unifiable Components (Series 2)

IRGWDS_MUI_nn: Possibly Mis-Unified Ideographs (Series 3).

// one more comment: Currently, an IRG N number is included in IRGWDS file name. However, according to SUM document, IRG N number should not appear in IRGWDS file name.
– Suggestion: include IRG N number to IRGWDS file name.

// modify file names according to the decision.

Detailed Specification of the Standing Document Series
This section explains the nature of each series as well as the format and the information contained in each series.

1. List of Unifiable Component Variations (UCV)

The UCV list provides the list of component variations to be unifiable observed from existing UCS multi-column charts, or proposed and agreed among IRG members to be unifiable.

If two ideographs differ only in terms of the components in the UCV list, but satisfy the requirement for dis-unification according to dis-unification rules, such ideographs may be encoded differently. However, such cases are exceptional and should be exhaustively listed in this document under the related components to avoid confusion for consideration of other characters.

Unification is meant to be at the component level only. In other words, if the components themselves are also ideographs proper, this list does not imply that the corresponding ideographs proper are unifiable.

The following is the format for each entry in the file:

a. No.: The serial number of the entry for reference that is unique

Formatted: Font: (Default) Times New Roman, Strikethrough
Formatted: Font: (Default) Times New Roman
Formatted: Font: (Default) Times New Roman, 14 pt
Formatted: Font: (Default) Times New Roman, 14 pt
Formatted: Font: (Default) Times New Roman
Formatted: Font: (Default) Times New Roman, Strikethrough
Formatted: Font: (Default) Times New Roman
Formatted: Font: 12 pt
Formatted: body, Indent: First line: 2 ch, Line spacing: single
Deleted: _
Formatted: Highlight
Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean

throughout the standardization works.

b. ~~Criteria~~ UCV: List of actual glyphs of UCV

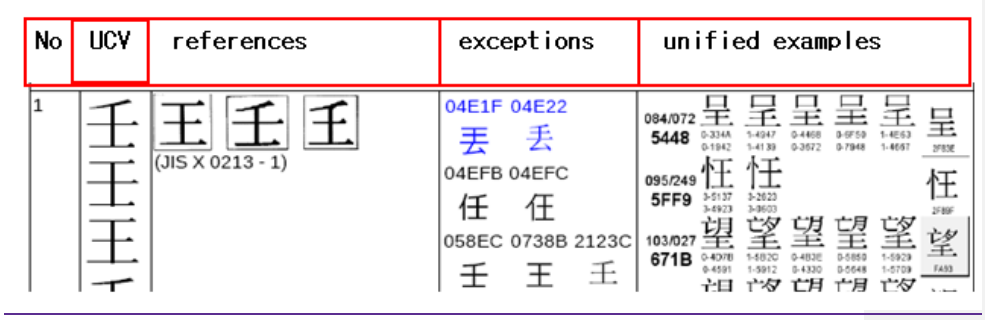
c. References: Excerpt from existing document (e.g. JIS X 0213 and HYDZD)

d. Exceptions: The exhaustive list of dis-unification examples.

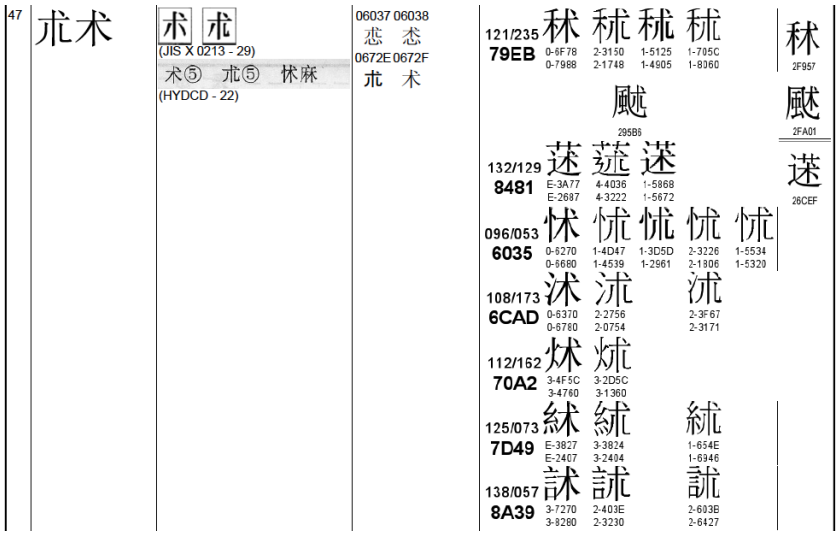
e. ~~Compatible/Duplicate/~~ Unified Examples: The example list of unified ideographs and compatibility ideographs, and notes if necessary

// Add actual column headings in N1648_UCV file; and

// add a note (in yellow) in N1648_SUM file for easier understanding:



The following is an example of a typical entry in UCV list.



Deleted: ,

Formatted: Strikethrough

Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean

Formatted: Strikethrough

Formatted: Highlight

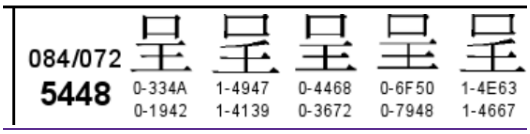
Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean

Formatted: Indent: Left 0 ch, First line: 0 ch

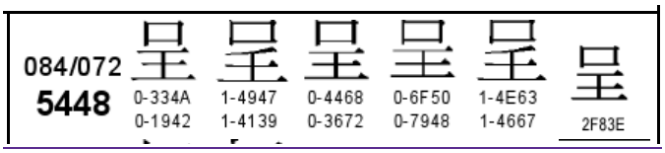
// The above example showing entry #47 seems incorrect.

// The current #47 is shown below:

// If appropriate, need to update the above example.

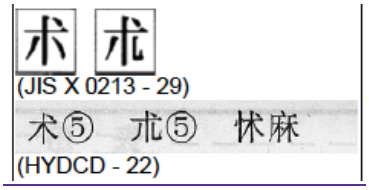


These five glyphs are unifiable and they are well unified



The glyph of comptb. char U2F957 is unifiable with glyphs of U5448 and they are well unified.

// could add some more explanation
= Currently, it does not seem to give enough info. in the context of UCV.
= comments are shown below:



RE: JIS example:
1) For example, does each variant have its own code position in JIS?
2) Or, does JIS include UCV list?
3) What does "29" mean?

Nearly the same comments apply to HYDCD example:
1) Does HYDCD list UCV?
2) Why are 忕麻 included here? (These two chars do not seem to include UCV)
3) What does "22" mean?

// How about numbering headings?

- An example Is shown below:

Differences in rotated Strokes/dots
Differences in stroke initiation direction
Differences in open direction of two strokes
....
Differences in intersection and extension of two strokes or dots
--->

1. Differences in rotated Strokes/dots
1.1 Differences in stroke initiation direction
1.2 Differences in open direction of two strokes

....
2. Differences in intersection and extension of two strokes or dots

--

// How about putting table of contents at the beginning?

Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean

2. List of Non-Unifiable Components of Ideograph(NUC)

The NUC list provides the list of component variations which are not to be unified. This list should be kept as minimum as possible. Components that are not obviously unifiable will not be listed here. That is, it should only list those that are close in glyph shapes which cognitively can be confusing. In other words, this list should only contain the components which are (possibly inappropriately) unified by precedence during the IRG working processing, or components that are stated to be unifiable by some local national standards, but not in the UCS.

Furthermore, this list should not contain components which are either (1) KangXi radicals (such as 工 vs. 土) or (2) simplified vs. traditional components with no precedence of unification (such as 門 vs 门).

The following is the format for each entry in the file:

a. Components: List of non-unifiable glyphs that is unique throughout the standardization works

b. Analysis: Reasons for dis-unification and each reason will be listed separately. Typical ones are already separated ones and ideographs which are encoded by one-side only.

c. Examples: List of exhaustive possibly over-unified ideographs (if exists)

The following is an example of a typical entry in ~~UNC~~ NUC:

Formatted: Strikethrough

Components	Analysis	Samples
厤 厤	Unifications	
	Separations	厤厤, 厯厯, 厝厝, 厯厯, 瀝瀝, 厝厝, 厯厯
	厤-only	none
	厤-only	厝厥厯厯厮厯厯厥厥摭厯厝厲厥厝厯厝厲厲厲
閒 間	Unifications	嚪 053/244 35F4 嚪 5-3856 5-2454 嚪 4-492D 4-4113 嚪 3-226F 3-0279 嚪 0-3338 0-1924 2-4-34 2442 a4c2 f1c0 35F4 嚪 30 (口) 12 4277 ガン人名 マ 嚪口 (マグチ・姓) 嚪 嚪 嚪
	Separations	僩僩, 嫺嫺, 憪憪, 撊撊, 橍橍, 澗澗, 爓爓, 癇癇, 瞯瞯, 礀礀, 簡簡, 纈纈, 蕑蕑, 襉襉, 覯覯, 譋譋, 鐧鐧, 閒閒, 鬅鬅, 鷳鷳, 嵧嵧, 徟徟, 蹋蹋, 醨醨, 鏑鏑, 鱱鱱, 齵齵
	間-only	嚪嚪嚪珂椆聃蜩嚪䖍
	閒-only	膶殰獧觸轛騆

= Analysis - Unifications examples.

- Are these chars dis-unified by mistake?? Or are they exceptional?

- could add references to dis-unification rule?

- Reasons are not given why they are dis-unified.

Unifications	壳 088/243 58F3 壳 0-3F47 0-3139 壳 3-282D 3-0813 壳 1-385E 1-2462 壳 2-2B3E 2-1130 鬳 鬳 293FB

- The above example seems to come from UCS.

could add some explanations

Separations	殼殼, 毂毂, 穀穀, 觳觳, 鷇鷇

It is not obvious from where the above examples came. (probably not from UCS?).

could add some explanations

Formatted: Font: (Asian) Malgun Gothic, (Asian) Korean

3. List of Possibly Mis-Unified Ideographs(MUI).

MUI list provides the possibly mis-unified ideographs as pairs of CJK compatibility ideographs and their corresponding CJK unified ideographs, which have different semantics and pronunciations with the supplied related reference information in a single document (possibly a dictionary).

It is possible that the coded CJK compatibility ideographs listed in this document need to be proposed as new CJK unified ideographs. However, extreme care must be taken to assure the compatibility with existing standards in accordance with Annex I of WG2's Principles and Procedures.

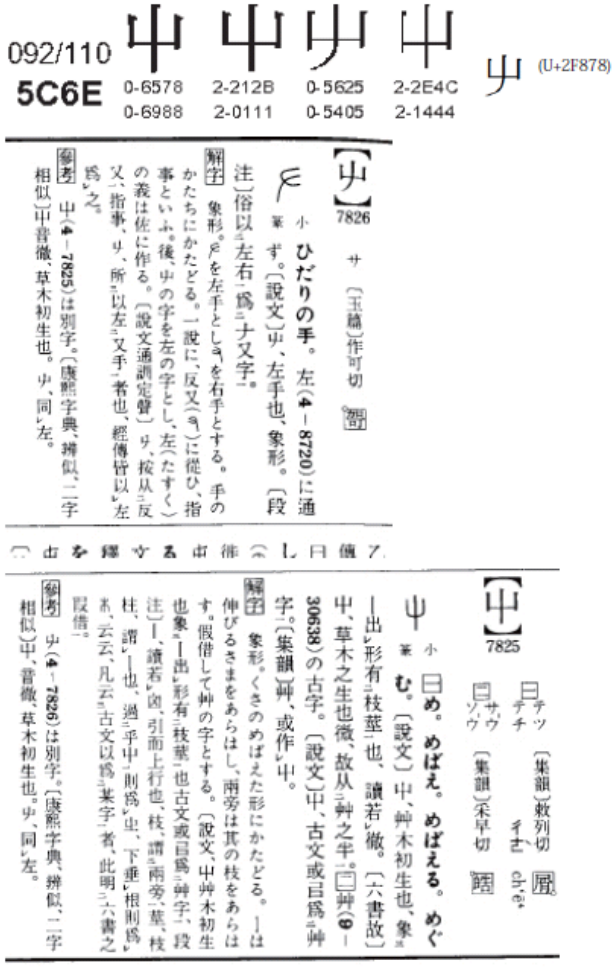
The following is the format for each entry in the file:

a. U-code: The UCS codepoint

b. Characters: List of possibly non-unifiable ideographs.

b. References: Excerpts of their usage from a single document source .

The following is an example of a typical entry in MUI:



(End of document)