JTC1/SC2/WG2/IRG
Date : 2017-07-14
IRGN2234_HKSAR_Feedback

ISO/IEC JTC1/SC2/WG2/IRG

Ideographic Rapporteur Group

(IRG)

Source/Contribution Identifier :	Hong Kong Special Administrative Region, China
Meeting :	IRG Meeting #49 in San Jose, US
Title :	Feedback to IRGN2234 on IRG Standing Document Summary Version 2
Status :	Member feedback for discussion

The HKSAR has reviewed IRGN2234 on IRG Standing Document Summary Version 2. Please refer to the parts highlighted in green in the marked-up draft attached to this document for details of proposed refinements.

END OF DOCUMENT

I made editorial changes from IRGN1648 which I did not highlight. Any text related to adding a new series (Series 5) is highlighted in yellow. You can help to revise the text of the whole document, not just the yellow part. Many thanks, LQ

Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC 1/SC 2/WG 2/IRGN2234 Draft
Date: 2017-06-26

Source:	IRG
Meeting:	After IRG Meeting #48, Seoul, Korea
Title:	IRG Standing Document Summary Version 2.
References	IRGN1648
Status:	Member’s submission
Actions required:	Feedback Requested
Distribution:	IRG
Medium:	Electronic
Pages:	6

Introduction:
IRG Working Document Series (IWDS) is a set of IRG maintained documents to keep up-to-date examples of CJK unification related ~~example~~ cases to supplement the published Annex S of ISO/IEC 10646 for IRG unification work. IRG also decided in its meeting #48 that a list of submitters’ Printed Character Normalization Guidelines are to be kept in IWDS for keeping track of the transformation rules used for handwritten character to printed character conversions.

The maintenance of the IRG Working Document Series should comply with the operational procedures established in Annex E of the *IRG Principles and Procedures*.

The Standing Document Series consists of the following documents:

Series 1: Summary of unification rules and sample examples.
Series 2: List of ~~UCV (~~Unifiable Component Variations~~)~~ (UCV) of Ideographs.
Series 3: List of ~~NUC (~~Non-~~U~~unifiable Components (NUC) of Ideographs and Overly-~~U~~unified Ideographs~~)~~
Series 4: List of Possibly Mis-~~U~~unified Ideographs (MUI).
Series 5: List of documents, each ~~is used to~~ describing~~e~~ ~~one~~a submitter’s

normalization guidelines (SNG) for conversion of one particular script to the printed form of ideographs~~(SNG)~~.

File format of IWDS:

Each of the IWDS file is named as IWDS_*SSS_II* where *SSS* is the name of the specific series (may be more than 3 letters for Series 5) and *II* refers to the IRG meeting where the list is confirmed. Thus, the Standing ~~d~~Document ~~s~~Series has 5 threads of documents. The first 4 threads are specified as follows:

IWDS_SUM: Summary document (Series 1) … This document.
IWDS_UCV: List of UCV~~ list~~ (Series 2)
IWDS_NUC: List of Non-~~U~~unifiable Components (Series 3)
IWDS_MUI: Possibly Mis-~~U~~unified Ideographs (Series 4)~~.~~ _

In Series 5, the guidelines are submitter dependent as well as script dependent. Thus, the file names need to include a submitter ~~id~~ID followed by a script ~~id~~ID. For example, the ROK's current guidelines ~~is~~ are for conversion of cursive Kai to printed Kai style. So the file name should be: IWDS_SNG_ROK_Kai_48.

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. Summary of IRG Standing Document Series (SUM)

The summary document (this document) is a definitive document giving detailed specification for each of the data files including specification on the nature of the data, the data format, and the examples.

2. List of Unifiable Component Variations (UCV)

The UCV list provides ~~the~~ a list of component variations ~~to be~~which are unifiable, as 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, these ideographs may be encoded differently. However, these 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 this entry for reference that is unique throughout the standardization works.

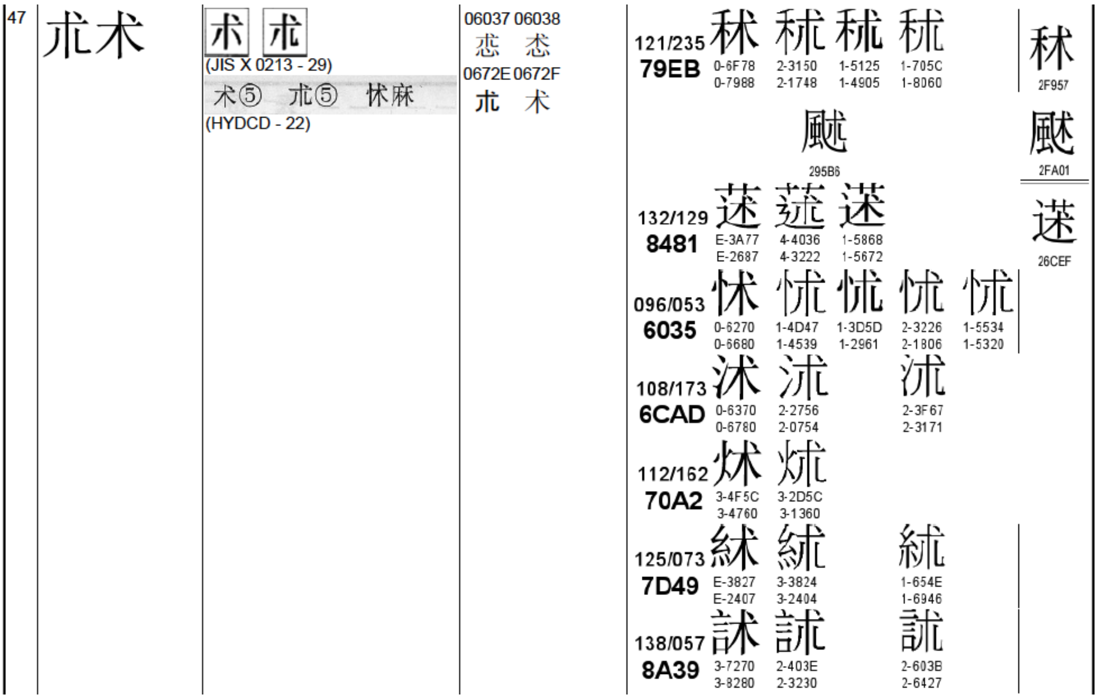
b. Criteria: List of actual glyphs.

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/Examples: The example list of unified ideographs and compatibility ideographs, and notes if necessary.

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



47	朮术	朮 朮 (JIS X 0213 - 29) 术⑤ 朮⑤ 怵麻 (HYDCD - 22)	06037 06038 怷 怸 0672E 0672F 朮 术	121/235 79EB	秫 0-6F78 0-7988	秫 2-3150 2-1748	秫 1-5125 1-4905	秫 1-705C 1-8060		秫 2F957
						𩖶 295B6				𩖶 2FA01
				132/129 8481	蒁 E-3A77 E-2687	蒁 4-4036 4-3222	蒁 1-5868 1-5672			𦳯 26CEF
				096/053 6035	怵 0-6270 0-6680	怵 1-4D47 1-4539	怵 1-3D5D 1-2961	怵 2-3226 2-1806	怵 1-5534 1-5320	
				108/173 6CAD	沭 0-6370 0-6780	沭 2-2756 2-0754		沭 2-3F67 2-3171		
				112/162 70A2	炢 3-4F5C 3-4760	炢 3-2D5C 3-1360				
				125/073 7D49	絉 E-3827 E-2407	絉 3-3824 3-2404		絉 1-654E 1-6946		
				138/057 8A39	訹 3-7270 3-8280	訹 2-403E 2-3230		訹 2-603B 2-6427		

3. List of Non-Unifiable Components (NUC) of Ideographs~~(NUC)~~

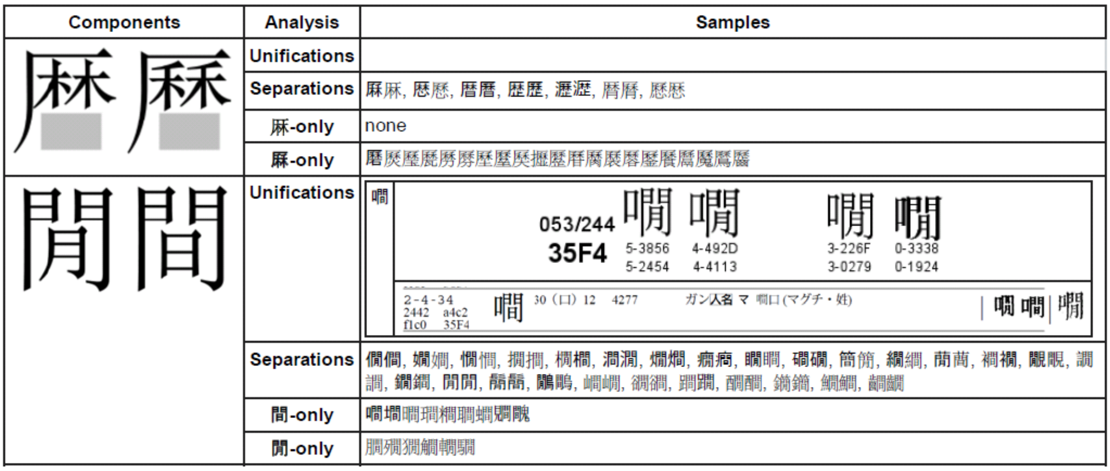
The NUC list provides ~~the~~ a list of component variations which are not to be unified. This list should be kept as minimal as possible. Components that are not obviously unifiable will not be listed here. That is, it should only list those that are close in glyph shapes and can be confusing cognitively. In other words, this list should only contain ~~the~~ components which are (possibly inappropriately) unified by precedence during the IRG working process, ~~or~~and 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 work~~s~~.
b. Analysis: Reasons for dis-unification and each reason will be listed separately. Typical ~~ones are~~reasons include: the ideographs in question are already separated ~~ones and ideographs which are~~or they are encoded by one~~-~~ side only.
c. ~~Ex~~Samples: List of exhaustive possibly overly-unified ideographs (if exists).

The following is an example of a typical entry in the ~~U~~NUC list.~~:~~



Components	Analysis	Samples
厤 厤	Unifications	
	Separations	厤厤, 歴歷, 暦曆, 壢壢, 瀝瀝, 曆曆, 歴歷
	厤-only	none
	厤-only	[illegible]
閒 間	Unifications	㗴 053/244 35F4 㗴 5-3856 5-2454 㗴 4-492D 4-4113 㗴 3-226F 3-0279 㗴 0-3338 0-1924 2-4-34 2442 a4c2 f1c0 35F4 㗴 30（口）12 4277 ガン人名 マ 㗴口(マグチ・姓) | 㗴 㗴 | 㗴
	Separations	僩僴, 嫺嫻, 憪憪, 撊撊, 橺橺, 澗澗, 燗燗, 癇癇, 瞯瞯, 礀礀, 簡簡, 繝繝, 蕑蕑, 襇襇, 覵覵, 譋譋, 鐗鐗, 閒閒, 鬝鬝, 鷳鷳, [illegible]
	間-only	㗴墹 [illegible]
	閒-only	[illegible]

4. List of Possibly Mis-~~U~~unified Ideographs (MUI).

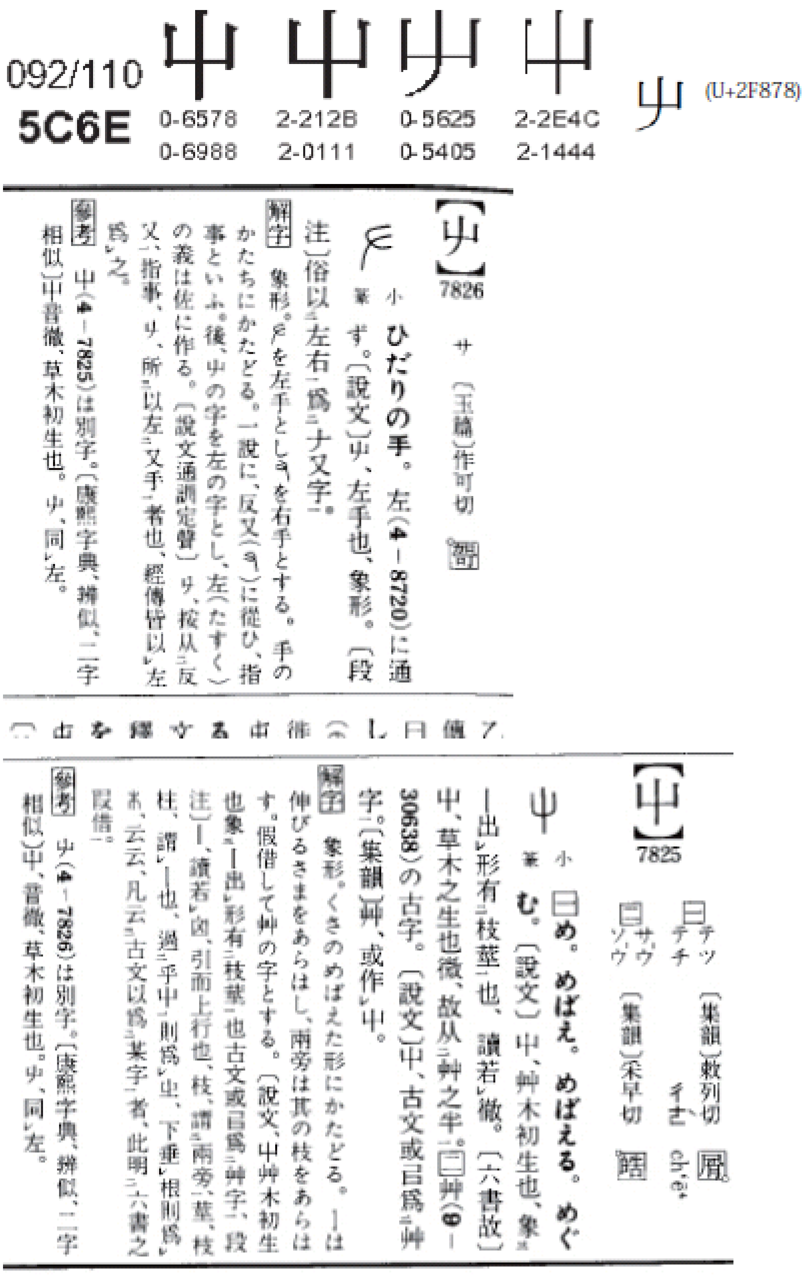
The MUI list provides the possibly mis-unified ideographs as pairs of CJK compatibility ideographs~~s~~ and their corresponding CJK unified ideographs, which have different ~~semantics~~ meanings and pronunciations with the supplied related reference information in a single document (possibly a dictionary).

It is possible that the coded CJK compatibility ideographs listed in this document will be proposed as new CJK unified ideographs. However, extreme care must be taken to ~~assure the~~ensure 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 code point.
b. Characters: List of possibly non-unifiable ideographs.
c. References: Excerpts of their usage from a single document source.

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



092/110
5C6E

屮	屮	屮	屮	屮 (U+2F878)
0-6578 0-6988	2-212B 2-0111	0-5625 0-5405	2-2E4C 2-1444	

【𠂇】7826 サ 〔玉篇〕作可切 哿

小篆 ひだりの手。左(4-8720)に通ず。〔説文〕𠂇、左手也、象形。〔段注〕俗以二左右一爲二ナ又字一。

解字 象形。𠂇を左手とし、ヨを右手とする。手のかたちにかたどる。一説に、反又(ヨ)に從ひ、指事といふ。後、𠂇の字を左の字とし、左(たすく)の義は佐に作る。〔説文通訓定聲〕𠂇、按从二反又一、指事、𠂇、所三以左二又手一者也、經傳皆以レ左爲レ之。

参考 屮(4-7825)は別字。〔康熙字典、辨似〕二字相似)屮音徹、草木初生也。𠂇、同レ左。

[illegible]

【屮】7825

曰 テツ チ 〔集韻〕敕列切 屑 ㄔㄜˋ chʻè

曰 サウ ソウ 〔集韻〕采早切 皓

小篆 曰め。めばえ。めばえる。めぐむ。〔説文〕屮、艸木初生也、象二|出レ形有二枝莖一也、讀若レ徹。〔六書故〕屮、草木之生也微、故从二艸之半一。曰艸(9-30638)の古字。〔説文〕屮、古文或以爲二艸字一。〔集韻〕艸、或作レ屮。

解字 象形。くさのめばえた形にかたどる。|は伸びるさまをあらはし、兩旁は其の枝をあらはす。假借して艸の字とする。〔説文、屮艸木初生也象二|出レ形有二枝莖一也古文或以爲二艸字一、段注〕|、讀若レ囟、引而上行也、枝、謂二兩旁一莖、枝柱、謂レ|也、過二乎屮一則爲レ屮、下垂レ根則爲レ木、云云、凡云二古文以爲二某字一者、此明二六書之叚借一。

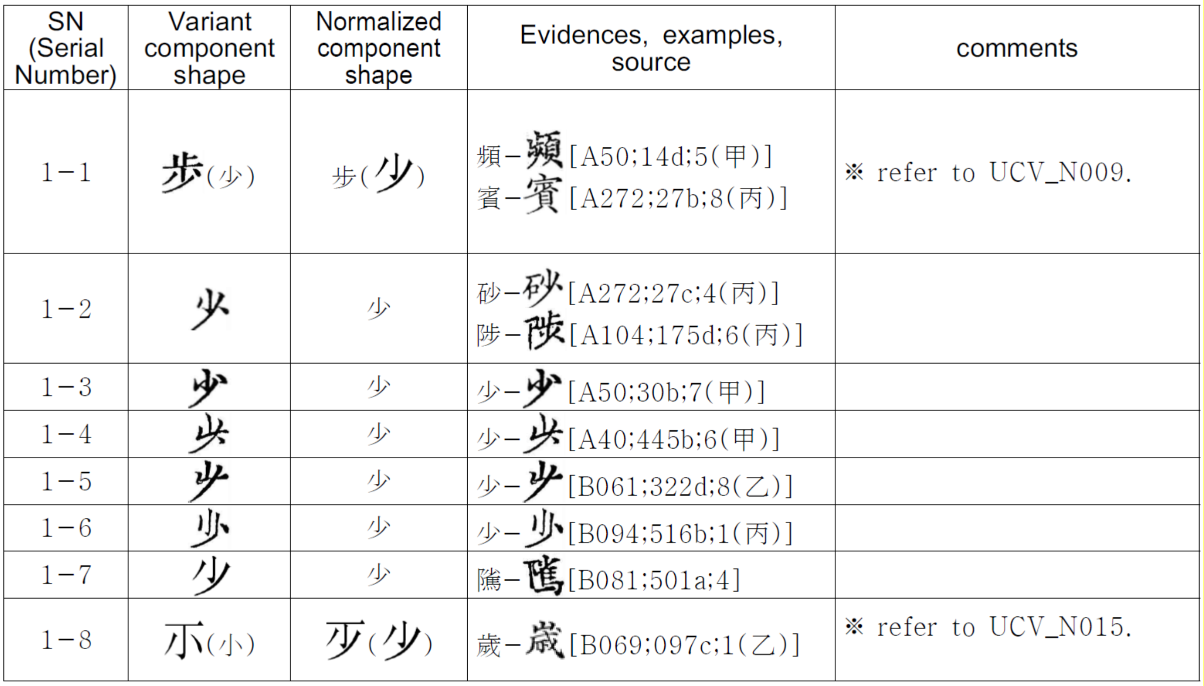
参考 𠂇(4-7826)は別字。〔康熙字典、辨似〕二字相似)屮、音徹、草木初生也。𠂇、同レ左。

5. Set of Printed Character Normalization Guidelines

As the normalization guidelines are submitter/culture/language dependent as well as script dependent, each document should provide~~have~~ an overview of the scope of the guidelines, the major references and authoritative document sources from where the guidelines ~~is~~ are derived. A set of rules can be described using text descriptions with ample examples for people in other language/culture environments to follow and to help with the review and acceptance of evidence~~s~~. It can also serve as a possible unification/dis-unification guide ~~to characters from~~ for other submitters. Content of the normalization table should include, but not limited to,~~-~~ the following data for each entry:

a. Serial numbers: A numbering system used internally for indexing and searching.
b. Variant glyph(s): Actual glyph shape~~s~~ of the component~~s~~ in the source script.
c. Normalized glyph: The corresponding normalized glyph shape of the component ~~glyph~~.
d. Evidence~~s~~: Examples of actual character glyphs with reference to their sources.
e. Comments: Any remark that may be helpful to IRG review.

The following is an example of some typical entries in a normalization table (from ROK: IRGN2154V1.1).



SN (Serial Number)	Variant component shape	Normalized component shape	Evidences, examples, source	comments
1－1	步(少)	步(少)	頻－頻[A50;14d;5(甲)] 賓－賓[A272;27b;8(丙)]	※ refer to UCV_N009.
1－2	少	少	砂－砂[A272;27c;4(丙)] 陟－陟[A104;175d;6(丙)]	
1－3	少	少	少－少[A50;30b;7(甲)]	
1－4	少	少	少－少[A40;445b;6(甲)]	
1－5	少	少	少－少[B061;322d;8(乙)]	
1－6	少	少	少－少[B094;516b;1(丙)]	
1－7	少	少	隲－隲[B081;501a;4]	
1－8	不(小)	歹(少)	歲－歲[B069;097c;1(乙)]	※ refer to UCV_N015.

(End of document)