

Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2/IRG N2550
Date: 2022-04-06

Source	IRG
Meeting:	After IRG#58
Title:	IRG Standing Document Summary Version 3.
References Status:	IRGN2426, IRGN2234, IRGN1648 (Only minor editorial changes)
Actions required:	Member's submission
Distribution:	Feedback requested
Medium:	IRG 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 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's 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 Unifiable Component Variations (UCV) of Ideographs.
- Series 3: List of Non-Unifiable Component Variations of ideographs and overly-unified ideographs (NUCV)
- Series 4: List of possibly Mis-Unified Ideographs (MUI).
- Series 5: List of documents, each describes one submitter's normalization guidelines(SNG) for conversion of one particular script to the printed form of ideographs.

File format of IWDS:

Each of the IWDS file is named as IWDS_SSS_// where SSS is the name of the specific series (may be more than 3 letters for Series 5) and // refers to the IRG meeting where the list is confirmed. Thus, the IRG working document 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 (Series 2)
IWDS_NUC: List of NUC (Series 3)
IWDS_MUI: List of MUI (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 followed by a script ID. For example, the ROK's current guidelines are for conversion of cursive Kai to printed Kai style. So the file name should be:
IWDS_SNG_ROK_Kai_49.

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 a list of component variations which are unifiable, as observed from existing UCS multi-column charts, or proposed and agreed among IRG members.

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 work.
- b. Criteria: List of actual glyphs.
- c. References: Excerpt from existing document (e.g. JISX0213 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.
- f. Level: Indicates is either of “should unify” (level 1) or “unifiable” (level 2). For detailed definition of level, see IRGN 2547.

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

228	朮 朮	<p>Level: 2</p> <p>朮 朮 (JIS X 0213 - 29)</p> <p>朮⑤ 朮⑤ 怵麻 (HYDCD - 22)</p> <p>Disunified Ideographs</p> <p>6037 怵 怵 怵 怵 6038 怵 怵 <small>G 913 G2297 H91-02F T2-046 K2-007 G 913 G2-00E T2-061</small></p> <p>672E 朮 朮 朮 朮 朮 672F 朮 朮 <small>G 701 G2-20F H91-02A T1-024 J2-022 K2-024 G2-407 G2-407 T2-022</small></p> <p>8481 迷 迷 迷 26CEF 迷 迷 <small>H 148 H 347 G2-347 T4-428 U2-020 G93-107 J4</small></p> <p>Compatibility Ideographs</p> <p>79EB 秫 秫 秫 秫 秫 2F957 秫 <small>G 155 G2-078 H92-02B T2-078 J2-79A K2-79C G 155 T2-061</small></p> <p>295B6 颯 颯 颯 2FA01 颯 <small>H 165 G2-020 G93-102 B T4-020 G 165 T2-020</small></p> <p>Unified Ideographs</p> <p>6035 怵 怵 怵 怵 怵 怵 <small>G 913 G2-078 H91-02E T1-027 J2-74E K2-02E V1-024</small></p> <p>6CAD 洑 洑 洑 洑 洑 <small>G 903 G2-078 H92-02F T2-078 J4-02F K2-07F</small></p> <p>70A2 煠 煠 <small>G 963 G2-07C T2-07C</small></p> <p>79EB 秫 秫 秫 秫 秫 <small>G 155 G2-078 H92-02B T2-078 J2-79A K2-79C</small></p> <p>7D49 紕 紕 紕 紕 <small>H 125 G2-02F T2-02F K2-02F</small></p> <p>8A39 誄 誄 誄 誄 <small>H 145 G2-77B H92-02B T2-42B K2-42B</small></p> <p>8FF0 述 述 述 述 述 <small>G 163 G2-407 H91-02A T1-030 J2-320 K2-078 V1-024</small></p> <p>9265 銖 銖 銖 銖 銖 <small>G 197 G2-79C H92-02A T2-42A J4-74E K2-07C</small></p> <p>3F81 瘥 瘥 <small>H 185 G2-020 T4-020</small></p> <p>4624 裌 裌 <small>G 145 G2-402 T4-022</small></p> <p>47E3 跣 跣 <small>G 197 G2-078 T4-078</small></p>
-----	-----	---

3. List of Non-Unifiable Component Variants (NUCV) of Ideographs

The NUC list provides 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 shape and can be confusing cognitively. In other words, this list should only contain components which are (possibly inappropriately) unified by precedence during the IRG working process, 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 (1) KangXi

radicals (such as 工 vs. 土) or (2) simplified vs. traditional components with no precedence of unification (such as 門 vs 冂).

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.
- b. Analysis: Reasons for dis-unification and each reason will be listed separately. Typical reasons include: the ideographs in question are already separately encoded by one-side.
- c. Samples: List of exhaustive possibly overly-unified ideographs (if exists).

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

294	𦉳 𦉳	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; padding: 2px;">𦉳</div> <div style="border: 1px solid black; padding: 2px;">𦉳</div> </div> <p>(JIS X 0213 - 140)</p> <p>Disunified Ideographs</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; vertical-align: top;"> <p>53A4 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F10 GE-2425 HB2-2846 T3-2946 J14-2387 K2-257C</small></p> <p>53AF 𦉳 𦉳 𦉳 <small>U+2F12 GE-2426 T4-3240 J1-344F</small></p> <p>66A6 𦉳 𦉳 𦉳 <small>U+2F16 GE-2427 T3-4737 J3-4671</small></p> <p>6B74 𦉳 𦉳 𦉳 <small>U+2F18 GE-2F53 H-F086 T3-4753 J3-4672</small></p> <p>701D 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1A G1-4124 HB1-C377 T1-7844 J3-8F42 K5-8877 V1-9238</small></p> <p>20AA0 𦉳 𦉳 <small>U+2F1C UCS2000 TF-4960</small></p> <p>20ABE 𦉳 𦉳 <small>U+2F1E UCS2000 TF-535A</small></p> </td> <td style="width: 50%; vertical-align: top;"> <p>20A75 𦉳 𦉳 <small>U+2F18 UCS2000 GNC-10073.10</small></p> <p>20AB1 𦉳 𦉳 <small>U+2F1E UCS2000 GAK</small></p> <p>66C6 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1C G1-734F HB1-8E34 T1-7380 J3-7347 K5-8575 V1-886E</small></p> <p>6B77 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F18 G1-401A HB1-8E7A T1-7125 J3-7945 K5-8576 V1-8839</small></p> <p>2404B 𦉳 𦉳 <small>U+2F1A UCS2000 GNC-21767.2A</small></p> <p>26B13 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 GNA-0965.06 T3-6728</small></p> <p>227CD 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 KA-0842 H-F067</small></p> </td> </tr> </table>	<p>53A4 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F10 GE-2425 HB2-2846 T3-2946 J14-2387 K2-257C</small></p> <p>53AF 𦉳 𦉳 𦉳 <small>U+2F12 GE-2426 T4-3240 J1-344F</small></p> <p>66A6 𦉳 𦉳 𦉳 <small>U+2F16 GE-2427 T3-4737 J3-4671</small></p> <p>6B74 𦉳 𦉳 𦉳 <small>U+2F18 GE-2F53 H-F086 T3-4753 J3-4672</small></p> <p>701D 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1A G1-4124 HB1-C377 T1-7844 J3-8F42 K5-8877 V1-9238</small></p> <p>20AA0 𦉳 𦉳 <small>U+2F1C UCS2000 TF-4960</small></p> <p>20ABE 𦉳 𦉳 <small>U+2F1E UCS2000 TF-535A</small></p>	<p>20A75 𦉳 𦉳 <small>U+2F18 UCS2000 GNC-10073.10</small></p> <p>20AB1 𦉳 𦉳 <small>U+2F1E UCS2000 GAK</small></p> <p>66C6 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1C G1-734F HB1-8E34 T1-7380 J3-7347 K5-8575 V1-886E</small></p> <p>6B77 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F18 G1-401A HB1-8E7A T1-7125 J3-7945 K5-8576 V1-8839</small></p> <p>2404B 𦉳 𦉳 <small>U+2F1A UCS2000 GNC-21767.2A</small></p> <p>26B13 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 GNA-0965.06 T3-6728</small></p> <p>227CD 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 KA-0842 H-F067</small></p>
<p>53A4 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F10 GE-2425 HB2-2846 T3-2946 J14-2387 K2-257C</small></p> <p>53AF 𦉳 𦉳 𦉳 <small>U+2F12 GE-2426 T4-3240 J1-344F</small></p> <p>66A6 𦉳 𦉳 𦉳 <small>U+2F16 GE-2427 T3-4737 J3-4671</small></p> <p>6B74 𦉳 𦉳 𦉳 <small>U+2F18 GE-2F53 H-F086 T3-4753 J3-4672</small></p> <p>701D 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1A G1-4124 HB1-C377 T1-7844 J3-8F42 K5-8877 V1-9238</small></p> <p>20AA0 𦉳 𦉳 <small>U+2F1C UCS2000 TF-4960</small></p> <p>20ABE 𦉳 𦉳 <small>U+2F1E UCS2000 TF-535A</small></p>	<p>20A75 𦉳 𦉳 <small>U+2F18 UCS2000 GNC-10073.10</small></p> <p>20AB1 𦉳 𦉳 <small>U+2F1E UCS2000 GAK</small></p> <p>66C6 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F1C G1-734F HB1-8E34 T1-7380 J3-7347 K5-8575 V1-886E</small></p> <p>6B77 𦉳 𦉳 𦉳 𦉳 𦉳 𦉳 <small>U+2F18 G1-401A HB1-8E7A T1-7125 J3-7945 K5-8576 V1-8839</small></p> <p>2404B 𦉳 𦉳 <small>U+2F1A UCS2000 GNC-21767.2A</small></p> <p>26B13 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 GNA-0965.06 T3-6728</small></p> <p>227CD 𦉳 𦉳 𦉳 <small>U+2F1E UCS2000 KA-0842 H-F067</small></p>			

4. List of possibly Mis-Unified Ideographs(MUI).

The MUI list provides the possibly mis-unified ideographs as pairs of CJK compatibility ideographs and their corresponding CJK unified ideographs, which have different 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 ensure 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 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

中

0-6578
0-6988

中

2-212B
2-0111

中

0-5625
0-5405

中

2-2E4C
2-1444

中 (U+2F878)

【中】
7826

サ (玉篇)作可切

𠄎

𠄎

小ひだりの手。左(4-8720)に通ず。(説文)中、左手也、象形。(段

注)俗以左右爲ナ又字。

【解字】

象形。𠄎を左手とし、𠄎を右手とする。手のかたちにかたどる。一説に、反又(𠄎)に従ひ、指事といふ。後、𠄎の字を左の字とし、左(たすく)の義は依に作る。(説文通訓定聲)𠄎、按从反又、指事、𠄎、所以左、又手、𠄎者也、經傳皆以左爲之。

【参考】中(4-7825)は別字。(康熙字典、辨似、二字相似)中音徹、草木初生也。𠄎、同左。

7. 備日し(中) 律中(五) 立(中) 中(中) 中(中)

【中】
7825

𠄎

(集韻)敕列切

𠄎

𠄎

𠄎

(集韻)采早切

𠄎

𠄎

小曰め。めばえ。めばえる。めぐむ。(説文)中、𠄎木初生也、象

一出形有枝莖也、讀若徹。(六書故)

中、草木之生也微、故从𠄎之半。𠄎(𠄎)

30638)の古字。(説文)中、古文或曰爲𠄎

字。(集韻)𠄎、或作𠄎。

【解字】

象形。くさのめばえた形にかたどる。一は伸びるさまをあらはし、兩旁は其の枝をあらはす。假借して𠄎の字とする。(説文、中𠄎木初生也象一出形有枝莖也古文或曰爲𠄎字、段注)一、讀若𠄎、引而上行也、枝、謂兩旁、莖、枝柱、謂一也、過𠄎乎中、則爲𠄎、下垂、相則爲𠄎、云云、凡云古文以爲某字者、此明六書之假借。

【参考】中(4-7825)は別字。(康熙字典、辨似、二字相似)中、音徹、草木初生也。𠄎、同左。

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 an overview of the scope of the guidelines, the major references and authoritative document sources from where the guidelines 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. It can also serve as a possible unification/disunification guide 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 of the component in the source script.
- c. Normalized glyph: The corresponding normalized component glyph.
- d. Evidence: 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	尒(小)	尒(少)	歲-歲 [B069;097c;1(乙)]	※ refer to UCV_N015.

(End of document)