# 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\_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 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.

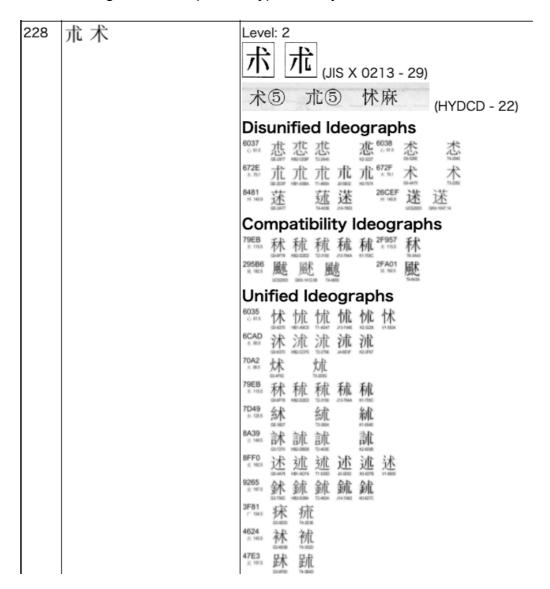
Unification is meant to be at the component level only. In other words, if the components themselves are also ideographs proper, this list does not imply that the corresponding ideographs proper are unifiable.

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.



## 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.

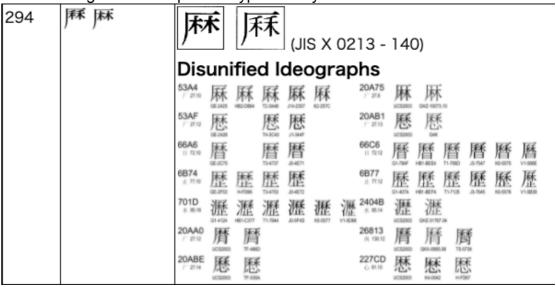
Furthermore, this list should not contain components which are (1) KangXi

radicals (such as  $\bot$  vs.  $\pm$  ) or (2) simplified vs. traditional components with no precedence of unification (such as  $\sqcap$  vs  $\sqcap$ ).

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.



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

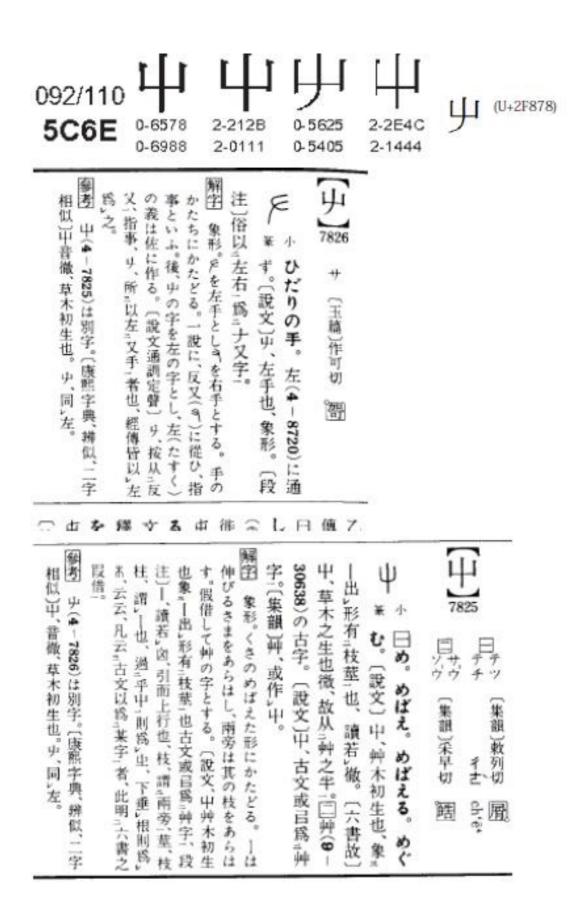
The MUI list provides the possibly mis-unified ideographs as pairs of CJK compatibility ideographs and their corresponding CJK unified ideographs, which have different meanings and pronunciations with the supplied related reference information in a single document (possibly a dictionary).

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.



## 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	<b>步</b> (少)	步( <b>少</b> )	頻- <b>頻</b> [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-改[A104:175d:6(丙)]	
1-3	J.	少	少 <b>-沙</b> [A50;30b;7(甲)]	
1-4	此	少	少- <b>火</b> [A40;445b;6(甲)]	
1-5	少	少	少- <b>少</b> [B061;322d;8(乙)]	
1-6	少	少	少 <b>-</b> 少[B094;516b;1(丙)]	
1-7	少	少	隲- <b>嚄</b> [B081;501a;4]	
1-8	<b>示</b> (小)	亚(少)	歲-歳[B069;097c;1(乙)]	* refer to UCV_N015.

(End of document)