

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION  
ORGANISATION INTERNATIONALE DE NORMALISATION  
ISO/IEC JTC 1/SC 2/WG 2/IRG**

<b>Universal Coded Character Set (UCS)</b>
--

**ISO/IEC JTC 1/SC 2/WG 2/IRG N1975Draft (Revision of IRGN1942/  
RGN1772/IRGN1823/IRGN1920, **WG2N4511**)  
2013-11-22**

<b>Title:</b>	IRG Principles and Procedures Version 6
<b>Source:</b>	IRG P&P Drafting Group
<b>Action:</b>	For review by the IRG and WG2
<b>Distribution:</b>	IRG Member Bodies and Ideographic Experts
<b>Editor in chief:</b>	Lu Qin, IRG Rapporteur
<b>References:</b>	IRGN 1952 and feedback from HKSARG, Japan, ROK and TCA, IRG 1920 Draft(2012-11-15), Draft 2(2013-05-04) and Draft 3(2013-05-22); feedback from Japan(2013-04-23) and ROK(2013-05-16 and 2013-05-21); and IRG Meeting No. 40 discussions IRG 1823 Draft 3 and feedback from HKSAR, Korea and IRG Meeting No. 39 discussions IRGN1823 Draft2 feedback from HKSAR and Japan IRG N1823Draft_gimngs2_Feedback IRG N1781 and N1782 Feedback from KIM Kyongsok IRGN1772 (P&P Version 5) IRG N1646 (P&P Version 4 draft) IRG N1602 (P&P Draft 4) and IRG N1633 (P&P Editorial Report) IRG N1601 (P&P Draft 3 Feedback from HKSAR) IRG N1590 and IRGN 1601(P&P V2 and V3 draft and all feedback) IRG N1562 (P&P V3 Draft 1 and Feedback from HKSAR) IRG N1561 (P&P V2 and all feedback) IRG N1559 (P&P V2 Draft and all feedback) IRG N1516 (P&P V1 Feedback from HKSAR) IRG N1489 (P&P V1 Feedback from Taichi Kawabata) IRG N1487 (P&P V1 Feedback from HKSAR) IRG N1465, IRG N1498 and IRG N1503 (P&P V1 drafts)

**Table of Contents**

<b>1. Introduction .....</b>	<b>4</b>
1.1. Scope of IRG Work .....	4
1.2. Scope of this Document .....	4
<b>2. Development of CJK Unified Ideographs .....</b>	<b>4</b>
2.1. Principles on Identification of CJK Unified Ideographs .....	5
2.1.1. Principles on Encoding .....	5
2.1.2. Unification Procedure of CJK Ideographs .....	5
2.1.3. Non-cognate Rule .....	5
2.1.4. Maintaining Up-to-date Unification/Non-unification Examples .....	5
2.2. Principles on Submission of Ideographs to the IRG .....	5
2.2.1. Basic Rules of Submission and Required Data to be Submitted .....	5
a) Code position of the existing UCS CJK Unified Ideograph .....	8
b) Glyph(s) of the existing UCS CJK Unified Ideograph .....	8
c) Glyph of the CJK Unified Ideograph to be printed in the member body's column of CJK Unified Ideographs Code Table .....	8

d) New source reference (for detailed format, see 2.2.1.d.(5).a).....	8
e) Evidence showing why a new source reference for the CJK Unified Ideograph needs to be added to UCS (e.g., a national standard showing the relevant glyph) .....	8
2.2.2. Required Font to be Submitted .....	8
2.2.3. Required Evidence to be Submitted .....	8
2.2.4. Required Summary Form to be Submitted .....	8
2.2.5. Quality Assurance: The 5% Rule .....	9
2.3. Principles on Production of IRG Working Drafts .....	9
2.3.1. Principles on Submitted Ideographs .....	9
2.3.2. Principles on Assignment of Serial Numbers .....	9
2.3.3. Principles on Machine-checking of IDS of Submitted Ideographs.....	9
2.3.4. Production of IRG Working Drafts .....	10
2.4. Principles on Reviewing IRG Working Drafts .....	10
2.4.1. General Principles on Reviews .....	10
2.4.2. Principles on Manual Checking .....	10
2.4.3. Submission of Possibly Unifiable Ideographs.....	11
2.5. Principles on Discussions at IRG Meetings .....	11
2.5.1. Document-based Discussion .....	11
2.5.2. Discussion Procedure.....	11
2.5.3. Recording of Discussions .....	12
2.5.4. Time and Quality Management.....	12
2.6. Principles on Submission of Ideographs to WG2 .....	12
2.6.1. Checking of Stabilized M-Set.....	12
2.6.2. Preparation for WG2 Submission .....	12
<b>3. Procedures .....</b>	<b>12</b>
3.1. Call for Submission .....	13
3.2. Consolidation and Grouping of Submitted Ideographs .....	13
3.3. First Checking Stage .....	13
3.4. First Discussion and Conclusion Stage .....	13
3.5. Subsequent Checking Stage .....	14
3.6. Subsequent Consolidation and Conclusion Stage .....	14
3.7. Final Checking Stage .....	14
3.8. Approval and Submission to WG2 .....	14
<b>4. Guidelines for Comments and Resolutions on Working Sets .....</b>	<b>15</b>
4.1. Guidelines for M-Set .....	15
4.2. Guidelines for D-Set .....	16
<b>5. IRG Website.....</b>	<b>16</b>
<b>6. IRG Document Registration.....</b>	<b>17</b>
6.1. Registration Procedure .....	17
6.2. Contact for IRG Document Registration .....	17
<b>Annex A: Sorting Algorithm of Ideographs .....</b>	<b>18</b>
<b>Annex B: IDS Matching .....</b>	<b>20</b>
B.1. Guidelines on Creation of IDS .....	20
B.2. Requirements of IDS Matching .....	20
B.3. Limitation of IDS Matching .....	20
<b>Annex C: Urgently Needed Ideographs .....</b>	<b>21</b>
C.1. Introduction .....	21
C.2. Requirements .....	21
C.3. Dealing with Urgent Requests .....	21
<b>Annex D: Up-to-date CJK Unified Ideograph Sources and Source References .....</b>	<b>22</b>
<b>Annex E: Maintenance Procedure of the IRG Working Documents Series .....</b>	<b>25</b>
E.1 Introduction .....	25
E.2. IRG Working Documents Series .....	25
E.3. Maintenance Procedure .....	25
<b>Annex F: IRG Repertoire Submission Summary Form .....</b>	<b>27</b>
<b>Annex G: Examples of New CJK Unified Ideographs Submissions (i.e., Vertical Extension) .....</b>	<b>29</b>
G.1. Sample Data Files .....	29

G.2. Sample Evidence	29
G.3. Handling of Data with Privacy Concerns	29
G.4. Consideration for Acceptance of Characters that Cannot be Provided in Printed Form	30
<b>Annex H [Reserved for future use]</b> .....	<b>32</b>
<b>Annex I: Guideline for Handling of CJK Ideograph Unification or Dis-unification Errors</b> .	<b>33</b>
I.1 Guideline for “To Be Unified” Errors	<b>Error! Bookmark not defined.</b>
I.2 Guideline for “To Be Dis-unified” Errors	<b>Error! Bookmark not defined.</b>
I.3 Discouragement of New Dis-unification Requests	<b>Error! Bookmark not defined.</b>
<b>Annex J: Guideline for Correction of CJK Ideograph Mapping Table Errors</b> .....	<b>34</b>
<b>Annex K: List of First Strokes</b> .....	<b>35</b>
<b>Annex L: Guidelines for Forming Current Working Sets</b> .....	<b>36</b>
<b>with an Upper Limit</b> .....	<b>36</b>
<b>References</b> .....	<b>37</b>
<b>Glossary</b> .....	<b>38</b>

# 1. Introduction

This document is a standing document of the ISO/IEC JTC 1/SC 2/WG 2/IRG for the standardization of Chinese-Japanese-Korean (CJK) Unified Ideographs. It consists of a set of principles and procedures on a number of items relevant to the preparation, submission and development of repertoires of CJK Unified Ideographs extensions for addition to the [ISO/IEC 10646](#) standard. Submitters should check the standard documents (including all the amendments and corrigenda) before preparing new submissions.

For any issue that is not explicitly covered in this document, the IRG will follow the Principles and Procedures of WG2 and other higher level directives.

## 1.1. Scope of IRG Work

The IRG works on CJK ideograph-related tasks under the supervision of WG2 (SC2 Resolution M18-10). The following is a list of current and completed IRG projects:

- a. CJK Unified Ideograph Repertoire and its extensions
- b. Kangxi Radicals and CJK Radical Supplements(completed)
- c. Production of Ideographic Description Sequence
- d. International Ideographs Core (IICore)(completed)
- e. CJK Strokes(completed)
- f. Update of CJK Unification Rules

Work on new IRG projects requires the approval of WG2, and preparation of relevant documents for such approval is required before the IRG can officially launch any new projects.

## 1.2. Scope of this Document

The following sections are dedicated to the standardization of CJK Unified Ideographs, describing the set of principles and procedures to be applied in the development of new extensions to the CJK Unified Ideograph Repertoire as specified under work item a. in Section 1.1. In addition, the maintenance of the IRG website and registration procedure of IRG documents are detailed in Sections 5 and 6 respectively.

This document does not cover other IRG work items listed in Section 1.1. Standardizing CJK Compatibility Ideographs maintained in UCS for the purpose of round-trip integrity with other standards is out of the IRG scope. However, CJK compatibility characters submitted to WG2 must be reviewed by the IRG to avoid potential problems. For the handling of mis-unification and duplicate ideographs, WG2 Principles and Procedures Annexes I and J attached to this document should be referenced.

# 2. Development of CJK Unified Ideographs

All new extension work must be approved by WG2 before the actual consolidation and review can be formally carried out. There are no fixed rules for initiating a new extension. Any member body, international consortium, or individual experts can initiate a new extension by submitting a proposal which states the need of a required repertoire. Submission of such a proposal must follow the principles and procedures stated in this document. The IRG will vet and confirm if the proposal is within its scope of work.

Taking into consideration (1) the urgency and justifications of the proposal, (2) the proposed repertoire size, and (3) its current workload, the IRG may take one of the following actions:

- a. Endorse the proposal and submit it to WG2 for approval.
- b. Request other member bodies to submit characters of similar nature so as to estimate the real workload before submitting the proposal to WG2 for endorsement.
- c. Accept the proposal as a contribution to an ongoing IRG work item.
- d. Reject the proposal with justifications. A rejected proposal may be revised and re-submitted to the IRG.

## 2.1. Principles on Identification of CJK Unified Ideographs

### 2.1.1. Principles on Encoding

Ideographs that have the same abstract shape are unified under the unification rule (Annex S of ISO/IEC 10646) and assigned a single character code. A CJK ideographic character can take many actual forms depending on the writing style adopted. Examples of common writing styles include Song style and Ming style as typical print forms, Kai style as a handwritten form, and Cao style as a cursive form. Stylistically different forms of the same character may involve different numbers or different types of strokes or components, which may in turn affect identification of the abstract shape of the character. In order to reach a common ground for identifying abstract shapes to be encoded as distinct CJK Unified Ideographs, the IRG only accepts submissions using a print form of glyphs (usually Song style or Ming style).

### 2.1.2. Unification Procedure of CJK Ideographs

Standard print forms of CJK ideographs are constructed with a combination of known components or stroke types. Many can be broken down into two components - a radical chosen to classify the character in dictionaries and possibly reflect the meaning of the character, and a phonetic component which represents the pronunciation of the character. Basically, two submitted print forms of glyphs with different radicals are distinct characters even if they have the same phonetic component, such as “嘆” (U+5606) and “歎” (U+6B4E). For non trivial cases, further shape analysis must be conducted. Similar glyphs should be decomposed into radicals, components or stroke types and evaluated by following the unification procedure described in Annex S of ISO/IEC 10646.

### 2.1.3. Non-cognate Rule

Ideographs that are unrelated in historical derivation (non-cognate characters) are not unified. The following gives examples of semantically different characters with very similar glyphs. They are considered to have different abstract shapes because they are non-cognate.

“戌”(U+620C) and “戍”(U+620D) differ only in rotated strokes or dots (S.1.5 a).

“于”(U+4E8E) and “干”(U+5E72) differ only in folding back at the stroke termination (S.1.5 f).

Because shape analysis alone may not tell non-cognateness or semantic differences, it is the submitter's responsibility to provide information and supporting evidence in order to invoke the non-cognate rule.

### 2.1.4. Maintaining Up-to-date Unification/Non-unification Examples

In Annex S, unification/non-unification examples are summarized from past practice and the lists are not exhaustive. If there is ambiguity in applying the unification/non-unification rules, the IRG must first have a formal discussion for agreement. In case there are worthy examples for recording, the IRG will add them to its lists of unification/non-unification examples maintained as IRG working document series (IWDS) on the IRG website. The lists will be reported to WG2 from time to time as an input for Annex S revisions. The detailed procedure of IWDS update is given in Annex E.

## 2.2. Principles on Submission of Ideographs to the IRG

### 2.2.1. Basic Rules of Submission and Required Data to be Submitted

The IRG accepts various types of submissions as specified below. Along with their submissions, the submitters are required to provide the necessary information for IRG's consideration.

- a. **New Sources to Standardized Ideographs.** For submissions specifying new sources (such as an existing or a new national standard) to existing standardized ideographs, the new sources must be reviewed and approved by the IRG before submission to WG2. Sources and source references in the current ISO/IEC 10646 standard can be found in clause 23 of ISO/IEC 10646 Third Edition (2012-06-01). See also Annex D for an up-to-date IRG list of sources.
- b. **New Sources to Working Sets.** For submissions specifying new sources to remaining characters in previous standardization stages, the new sources must be reviewed and approved by the IRG before they are incorporated by the IRG technical editor into the up-to-date IRG list of sources for the current IRG working sets.

- c. **New CJK Compatibility Ideographs (Vertical extension).** To add CJK Compatibility Ideographs, a submitter needs to supply the following information, which will be reviewed by the IRG before submission to WG2 to avoid possible problems of unification or dis-unification with other CJK Unified Ideographs.
- (1) Table showing the following data for each proposed CJK Compatibility Ideograph
    - a) UCS code position of the corresponding CJK Unified Ideograph
    - b) Glyph(s) of the corresponding CJK Unified Ideograph
    - c) Glyph of the CJK Compatibility Ideograph to be printed in the member body's column of CJK Compatibility Ideographs Code Table
    - d) Source reference (for detailed format, see 2.2.1.d.(5).a)
    - e) Evidence showing why the CJK Compatibility Ideograph needs to be added to UCS (e.g., a national standard showing two distinct code positions for two glyphs that are one and the same).
  - (2) TrueType font containing glyphs to be printed in the member body's column of CJK Compatibility Ideographs Code Table (for detailed format, see 2.2.2.b.)
- d. **New CJK Unified Ideographs (Vertical extension).** All CJK Unified Ideograph submissions are subject to the following rules:
- (1) **Collection Size for the Working Sets:** As all collections are defined by submitters according to their own criteria, the IRG does not impose a limit on the collection size. However, to rationalize the feasibility of a timely checking process and to achieve a high quality of work within a reasonably short period of time, the size of the collection to be reviewed by IRG member bodies as the current working sets normally does not exceed 4,000 ideographs. Based on this principle, submitters may be asked to divide their submitted collections into subsets to be processed in different IRG collections. The guideline for forming the current working sets is given in Annex L.
  - (2) **Pre-submission Unification Checking:** Submitters should be **EXTREMELY CAREFUL *not to submit CJK Unified Ideographs that are already standardized or previously discussed*** and recorded at IRG meetings. By the nature of ideographs, it is very difficult for IRG reviewers to find out all unifiable ideographs. Thus, it is important to achieve high quality at the time of submission. Submitters must make sure that the ideographs they submit do not fall into any of the following categories:
    - a) Ideographs already standardized in the ISO/IEC 10646 standard (including its amendments).
    - b) Ideographs currently in WG2's working drafts.
    - c) Ideographs currently in the IRG working sets including both M-set and D-set<sup>1</sup>.
    - d) Ideographs mis-unified or over-unified with ideographs in the current standard based on the lists maintained by the IRG in its working document series, namely IWDS\_MUI and IWDS\_NUC.

Low quality submissions may be rejected under the "5% Rule" defined in Section 2.2.5 below.
  - (3) **Document Registration:** All submission documents should be registered as IRG documents with an IRG document number (IRGN). The file names should be in the form of:

IRGNnnnn\_mmmm\_ppp\_sub\_sss

where *nnnn* indicates an IRG document number assigned by the IRG Rapporteur, *mmm* indicates the submitter's abbreviation (as listed in 2.2.1.d.(5).a), *ppp* indicates the working set or repertoire name (such as Ext. X labelled by "\_X"), "sub" is a short hand for submission, and *sss* can be any submitter-designated indicator.

- (4) **Submission of Over-unified or Mis-unified Ideographs:** Submissions of ideographs that are already mis-unified or over-unified within the current standard should follow the principles in Annex I of WG2 Principles and Procedures. Lists of over-unified and mis-unified ideographs should be maintained by the IRG technical editor and made available for update in the IRG working document series (i.e. IWDS\_NUC and IWDS\_MUI) according to the maintenance procedure defined in Annex E of this document. For mis-unified non-cognate characters, requests can be made to add new code points for dis-unified characters.<sup>2</sup>

<sup>1</sup> See Section 2.3.4 for the purposes of M-set and D-set.

<sup>2</sup> It should be noted that the source separation rule described in Annex S of the ISO/IEC 10646 is confined to coding standards listed in that section and is not applicable to new IRG submissions.

- (5) The following data items for each proposed ideograph must be submitted in CSV (Comma Separated Value) text format (in UTF-8) or Microsoft Excel file format:

\* Sequence number starting from 1 is required in the first column of each row.

- a) Source Reference** to indicate the source and the name of the glyph image for tracking. The source reference should begin with a member body abbreviation (G, H, J, K, KP, M, MY, T, U or V)<sup>3</sup> or an international consortium abbreviation (Z) followed by no more than 9 characters. It should contain only Latin capital letters, Arabic numbers, and hyphens. The purpose of source references and an exhaustive list of source references accepted by ISO/IEC 10646 are provided in Section 23 of ISO/IEC 10646. See Annex D for details of member body/international consortium abbreviations.
- b) Glyph Image File Name.** The file name of each glyph image must be the same as the source reference with file extension of .bmp in bitmap format or .png in PNG format.
- c) Kangxi Radical Code** from 1 to 214<sup>4</sup> with an additional .0 or .1 to indicate a traditional radical(0) or simplified radical(1). The selection of 0 or 1 is based on the radical's glyph shape. The list of radicals with both traditional and simplified glyphs are given in Annex A.a.
- d) Stroke Count** of components other than the radical. In case of unified characters, the assignment of stroke count will be based on IRG agreed rules (ref. IRGN954AR and IRGN1105).
- e) First Stroke** from 1 to 5 as listed in Annex K for components other than the radical. The IRG does not enforce a unified method of first stroke assignment as specified in IRGN954, IRGN954AR, and IRGN1105. It is however advisable for submitters to supply their assignment principles to the technical editor for spotting differences between different submissions. The final decision on first stroke selection rests with the technical editor for IRG editorial work. If the decision of the technical editor is different from that of the submitter, the submitter should be informed of the change.
- f) Flag to show whether the ideograph is traditional (0) or simplified (1).**
- g) Ideographic Description Sequence(IDS)** (ref. IRGN1183).
- h) Similar Ideographs** if available (identified by their code points in the standard in the form of U+xxxxx). Enter "No" if no known variants; leave the column empty if not checked.
- i) References to evidence documents** including document names and other information. Submitters may supply additional information related to evidence submission and additional columns can be added as subitems of i). For example, in addition to i) as the name of the original evidence document, member bodies may include i1) for page number in the document, i2) for row number in that page, i3) to indicate position in that row, and i4) as the name of the JPG file showing the page of an evidence document where the character in question appears.
- j) Optional Information** in text format can be included as additional columns starting from j).

Some sample submissions are provided in Annex G for reference.

A blank form in Excel format is available for submitters' use as a separate document.

- e. **Existing CJK Compatibility Ideographs (Horizontal extension).** To add new source references to existing CJK Compatibility Ideographs, a submitter needs to supply the following information, which will be reviewed by the IRG before submission to WG2 to avoid possible problems.

- (1) Table showing the following data for each proposed horizontal extension of CJK Compatibility Ideographs
- Code position of the existing UCS CJK Compatibility Ideograph
  - Glyph(s) of the existing UCS CJK Compatibility Ideograph
  - Code position of the corresponding UCS CJK Unified Ideograph
  - Glyph(s) of the corresponding UCS CJK Unified Ideograph
  - Glyph of the Compatibility Ideograph in the source reference
  - Glyph of the Compatibility Ideograph to be printed in the member body's column of CJK Compatibility Ideographs Code Table
  - New source reference (for detailed format, see 2.2.1.d.(5).a)

<sup>3</sup> Member body abbreviations correspond to the source standard categories in Section 23 of ISO/IEC 10646 except **MY**.

<sup>4</sup> The corresponding code range for Kangxi radicals in ISO/IEC 10646 is from U+2F00 to U+2FD5.

- h) evidence showing why a new source reference for the CJK Compatibility Ideograph needs to be added to UCS (e.g. a national standard showing two distinct code positions for two glyphs that are one and the same)
- (2) TrueType font containing glyphs to be printed in the submitter's column of CJK Compatibility Ideographs Code Table (for detailed format, see 2.2.2.b.)
- f. **Existing CJK Unified Ideographs (Horizontal extension).** To add new source references to existing CJK Unified Ideographs, a submitter needs to supply the following information. These characters must be reviewed by the IRG before submission to WG2 to avoid possible problems.
  - (1) Table showing the following data for each proposed horizontal extension of CJK Unified Ideographs
    - a) Code position of the existing UCS CJK Unified Ideograph
    - b) Glyph(s) of the existing UCS CJK Unified Ideograph
    - c) Glyph of the CJK Unified Ideograph to be printed in the member body's column of CJK Unified Ideographs Code Table
    - d) New source reference (for detailed format, see 2.2.1.d.(5).a)
    - e) Evidence showing why a new source reference for the CJK Unified Ideograph needs to be added to UCS (e.g., a national standard showing the relevant glyph)
  - (2) TrueType font containing glyphs to be printed in the submitter's column of CJK Unified Ideographs Code Table (for detailed format, see 2.2.2.b).

#### 2.2.2. Required Font to be Submitted

- a. **Glyph Image:** Each proposed ideograph must be accompanied by a corresponding 128 x 128 bitmap file in Song or Ming style. The file name should be the same as the source reference (defined in Section 2.2.1.d.(5).a) with .bmp as its file extension.
- b. **TrueType Font** (optional): TrueType font availability is highly recommended although not necessary. Font specification can be found under point 5 of A.1. – Submitter's Responsibilities in Annex A of the Principles and Procedures for UCS provided by WG2<sup>5</sup>. The IRG at certain stage of project development will set a deadline for TrueType font submission.

#### 2.2.3. Required Evidence to be Submitted

- a. **Supporting Evidence:** Evidence of the proposed glyph shape, its usage and context with pronunciation(s), meaning(s), etc. should be supplied to convince the IRG that it is actually in use or non-cognate with other similar ideographs. Evidence for each character must be supplied as scanned images. The provision of evidence on character usage including uses for personal names should not be exempted. A declaration for character use without accompanying evidence is generally not acceptable. Considering privacy issues, the IRG has suggested some compromised provisions. Details are given in Annex G.3.  
**Note:** To support e-government related initiatives, the IRG may at its discretion accept submissions of characters that are used in computer systems administered by government bodies for public service with wide access by government agencies and citizens. Factors considered for such acceptance are further elaborated in Annex G.4.
- b. **Questionable Characters** (optional): For candidate ideographs with possible unification questions, in addition to listing the possible unifiable characters as required in 2.2.1.d.(5).h, submitters are encouraged to provide for review detailed evidence of use from authoritative sources, and evidence showing their relationship to other standardized ideographs or variants having similar shape or meaning. Characters with this information are not counted as problem characters for quality assurance assessment given in 2.2.5.
- c. **Avoidance of Derived Simplified Ideographs:** To avoid encoding derived simplified characters that are not in actual use, submissions of simplified ideographs require actual usage evidence. Providing only their corresponding traditional ideographs will not be considered as producing usage evidence.

#### 2.2.4. Required Summary Form to be Submitted

Each submission for an ideograph collection should be accompanied by a duly completed "Proposal

<sup>5</sup> <http://std.dkuug.dk/jtc1/sc2/wg2/docs/principles.html>



Summary Form for Addition of CJK Unified Ideographs to the Repertoire of ISO/IEC 10646” (see Annex F).

### 2.2.5. Quality Assurance: The 5% Rule

For all character encoding standards, a common general principle is to encode the same character once and only once. Before any submission, it is the submitter's responsibility to filter out ideographs that are already in the ISO/IEC 10646 international coding standard:

- the published standard,
- any of its published amendments,
- any of its amendments under ballot in JTC1/SC2, or
- any of the working sets of the IRG.

In assessing the suitability of a proposed ideograph for encoding, the IRG will evaluate the credibility and quality of the submitter's proposal. If the IRG finds more than 5% of the submitter's source set are duplicates of characters in the above mentioned collections during the IRG review process, the whole submission will be removed from the subsequent IRG working drafts for that particular IRG project. However, the 5% rule does not apply if the submitter explicitly raises questions about unification/dis-unification for concrete cases in the proposal of characters.

It should be noted that the 5% rule is a general yardstick to remind submitters to adhere to IRG submission requirements and do a good screening job before submission to reduce the workload of member bodies for quality review. In practice, most submissions should have problems within the 1% range. In this regard, submitters should not interpret the rule as submissions with problems within the 5% range will definitely be accepted. The IRG has the right to review the problem cases and decide not to accept a submission even if it has problems within the 5% range (especially when the figure is very close to 5%).

## 2.3. Principles on Production of IRG Working Drafts

After the IRG accepts submissions based on principles specified in Section 2.2, the development process of the review work begins. The IRG technical editor will first produce a set of IRG working drafts.

### 2.3.1. Principles on Submitted Ideographs

- a. All the original ideograph submissions, including submissions of glyphs, IDS, radicals, stroke counts and evidence, must have registered IRG document numbers.
- b. If any required information is missing, the IRG chief editor or technical editor can ask for additional information from the submitter. Without timely supply of such information, the submission may be rejected by the technical editor in producing the working drafts.

### 2.3.2. Principles on Assignment of Serial Numbers

- a. The IRG technical editor will consolidate and sort the submitted ideographs in accordance with Annex A of this document.
- b. A unique *serial number* will be assigned to each submitted ideograph after consolidation. The serial numbers must be unique throughout the standardization process. They must not be changed, re-set or re-assigned unless there is an agreed dis-unification during the process. This principle allows easy reference to past discussions. In case of a split, one ideograph will keep the original serial number and the other will be assigned a new serial number.
- c. If ideographs submitted by different submitters are obviously unifiable, such ideographs may be unified and assigned the same serial number by the IRG technical editor.

### 2.3.3. Principles on Machine-checking of IDS of Submitted Ideographs

- a. The IRG technical editor will check the submitted IDS with existing IDS data to detect possible unifiable or duplicate ideographs.
- b. Machine checking sometimes detects obviously non-unifiable pairs. In such cases, when detected, they will be noted and assigned with different serial numbers before proceeding to the next stage.
- c. IDS checking algorithm will satisfy the requirements described in Annex B.

#### 2.3.4. Production of IRG Working Drafts

- a. **Division of Character Subsets:** By the result of IDS checking, submitted ideographs will be grouped into the following two working sets:
  - (1) **M-set (main set):** for ideographs with proper IDS and found not to be unifiable with current standardized ideographs or previously discussed ideographs with proper IDS.
  - (2) **D-set (discussion set):** for ideographs with missing, incomplete, or inconclusive IDS, or ideographs of which the attribute data have been questioned by any member body during a review process, or ideographs that might be unifiable with standardized or previously discussed ideographs. Ideographs with missing or incomplete IDS will be commented as such, and checked intensively through manual checking. Ideographs that might be unifiable with standardized or previously discussed ideographs will also be commented as such, and their suitability for unification must be manually checked. Dis-unification must be supported by evidence.
- b. **Naming of Working Drafts:** The file name should follow the format of “IRGNnnnnVX[XXX]” where *nnnn* is the IRG assigned document number and *X* is the version number. No spaces are allowed. But, the use of underscore “\_” and period “.” for separation is permissible. Examples of version numbers are “ExtFV1.0”, “V1.0Draft”, etc.
- c. **Glyph Images:** An archive of consolidated glyph images will be produced. Each image should be in the size of 128 x 128 pixels with file name using the source reference and the extension .bmp.
- d. **Addition of Characters:** No ideographs should be added to the working sets once the development process begins.
- e. **Alteration of Characters:** Alteration of characters is generally not allowed because it indicates instability and may have impact on other characters in the collection. However, submitters may submit proposals of minor alterations of characters with justifications at the final stage or with explicit approval from the IRG as long as the altered glyphs are unifiable with the original characters. A change of glyph beyond the Annex S unification criteria is considered to be an addition of new character and is NOT acceptable during the development process. The submitter of any alteration proposal must provide the results of thorough checks and verification showing that the alteration does not affect other characters in existing standards and working sets. The IRG, based on its evaluation, may decide to accept the alteration, reject the proposal or request the removal of such a character by the submitter. If the submitter finds that the glyph of a character is wrong at any working stages, the character will be rejected by the IRG and should be withdrawn by the submitter.
- f. After consolidation, the IRG chief editor and technical editor may ask member bodies to review M-set and D-set based on an agreed IRG review schedule and task division.

#### 2.4. Principles on Reviewing IRG Working Drafts

If the IRG instructs member bodies to review the working drafts (different portions may be assigned to different member bodies), member bodies' editors should review it according to the agreed schedule. They should follow the principles set out below during the review process.

##### 2.4.1. General Principles on Reviews

- a. Each member body should check the ideographs of the working sets assigned by the IRG chief editor and technical editor for the following issues:
  - (1) Correctness of Kangxi radical, Kangxi index, stroke count, first stroke and IDS.
  - (2) Correctness and quality of glyphs and source information if necessary.
  - (3) Presence of duplicate or unifiable ideographs based on Annex S guidelines.
  - (4) Consistency of submitted characters with the submitted evidence and documentary proof.
- b. When any data of an ideograph, including IDS, Kangxi radical, or stroke count are found to be incorrect, the ideograph should be moved from M-set to D-set as its standing data are no longer valid. Until the ideograph is confirmed to be unique by manual checking (procedure described in Section 2.4.2. below), it should not be moved back to M-set.

##### 2.4.2. Principles on Manual Checking

- a. **Duplication and Unification:** For D-set ideographs, member bodies should ensure that they are not duplicates of or unifiable with any ideographs in the standard or in another working set (including the current one).
- b. **Radical Checking:** Assurance is done by enumerating all possible radicals of a target ideograph and looking for any duplicate or unifiable ideographs in the range of  $\pm 2$  stroke counts of

standardized and working set ideographs. For example, “聞” may have the radical of “門” with 6 strokes for the remaining component, or the radical of “耳” with 8 strokes for the remaining component. In such a case, checking standardized and working sets for ideographs with radical of “門” and 4-8 strokes, or ideographs with radical of “耳” and 6-10 strokes manually can better assure that the ideograph does not have duplicate or unifiable ideographs.

- c. **Recording of Review Results:** The checking work should be recorded in the review comments as “Checked against all standardized and working set ideographs with radical X and stroke count of Y±2.”

#### 2.4.3. Submission of Possibly Unifiable Ideographs

- a. **Preparation of Comments:** Member bodies should prepare comments and feedback quoting the assigned serial numbers of the ideographs in question. The guidelines on comments are described in Section 4 of this document. Comment files should be in CSV text format or Microsoft Excel file format. All comment files must have pre-assigned IRG document numbers.
- b. **Additional Evidence and Arguments:** For each proposed ideograph in the D-set that has been questioned for possible unification, the submitter should prepare arguments with further evidence of its use and documentary proof (for example, from dictionaries, legal documents or other publications) showing that it is not unifiable with any standardized ideograph or ideograph proposed in the same or another working draft.
- c. **Submission Deadline:** Each member body should send feedback comments at least two months before the next IRG meeting. The IRG chief editor and technical editor will consolidate them and register the results as IRG documents one month before the next IRG meeting.
- d. **Written Response from Submitters:** Submitters should examine the consolidated comments on their respective characters and send the IRG chief editor and technical editor a written document containing their responses to the comments together with additional evidence at least one week before the next IRG meeting.
- e. **Rejection:** Questioned ideographs with no counter arguments in support of dis-unification supplied to the IRG meeting will be automatically marked as unified.
- f. **Revised font:** In case of glyph mismatch to evidence or mismatch to normalization for consistency, submitter needs to provide revised font to IRG for review.

## 2.5. Principles on Discussions at IRG Meetings

### 2.5.1. Document-based Discussion

For efficient and smooth work, all discussion items and evidence must be presented as registered IRG documents before the commencement of an IRG meeting. Items or evidence that are not contained in a registered IRG document will not be discussed or treated as evidence during IRG meetings. Discussions on evidence or items raised after the commencement of an IRG meeting may be postponed to the next IRG meeting if any member body requests longer time to examine such evidence or items.

### 2.5.2. Discussion Procedure

Discussions will be based on the review comments on working sets. For unification issues, the submitters should present evidence documents showing that the suspected unifiable ideographs are distinctively used as non-cognate characters in the same region, or that they cannot be unified in accordance with Annex S. When IRG member bodies have reached a consensus that two ideographs are unifiable, the submitter concerned should take one of the following actions, and the decision must be recorded.

- a. Withdraw the duplicate ideograph and map it to the existing standardized or working set ideograph.
- b. Submit the ideograph as a compatibility ideograph character.
- c. Add a new source reference to the existing standardized or working set ideograph.

When characters are reviewed by different people, different choices of Kangxi radical, stroke count or first stroke code are possible for the same ideograph. IRG member bodies should agree on the most appropriate ones based on the commonest abstract shape of the specific glyph. When the Kangxi radical or stroke count of an ideograph is found to be incorrect, the ideograph will be moved to D-set pending another manual review to prevent any unification errors caused by not having conducted the review with ideographs having the correct Kangxi radical or stroke count.

Guidelines on typical comments and resolutions are given in Section 4 of this document.

### 2.5.3. Recording of Discussions

Comments, rationales, and decisions must be recorded for each ideograph reviewed in a tabular format for reference and checking.

### 2.5.4. Time and Quality Management

Before a discussion begins, the number of ideographs under review will be counted and the schedule will be estimated based on it. During the discussion, the number of comments reviewed per hour will be noted and the schedule will be adjusted according to the progress (Note: It is recognized that some comments may take longer than others to discuss and resolve). If the comments cannot be handled in one IRG meeting, they may be partitioned and resolved in subsequent IRG meetings. Due to the limited time CJK Editorial Group has to deal with individual characters during an IRG meeting, member bodies can use emails to discuss and reach agreement on simple, straightforward cases before and after an IRG meeting.

## 2.6. Principles on Submission of Ideographs to WG2

### 2.6.1. Checking of Stabilized M-Set

- a. Once M-set is consolidated and stabilized, the ideographs in M-set will be checked intensively as a complete set at least once to ensure data and glyph integrity.
- b. Approval by a majority vote of IRG member bodies is needed before the set can be prepared for WG2 submission.

### 2.6.2. Preparation for WG2 Submission

After the approval by the IRG, the IRG technical editor will prepare the proposal to be forwarded to WG2. The preparation includes the following:

- a. Sort the final stable M-set ideographs by the sorting algorithm described in Annex A.
- b. Assign provisional UCS code positions to the sorted M-set ideographs (with agreement from the ISO/IEC 10646 project editor on block assignment).
- c. Make available the TrueType font for each member body with assigned provisional UCS code positions (fonts have to be available in accordance with the requirement stated in point 5 of A.1. – Submitter's Responsibilities in Annex A, WG2 Principles and Procedures).
  - (1) Each submitter is encouraged to prepare and submit its own font for best font quality.
  - (2) If a submitter has difficulty in creating the font, other member bodies or the IRG technical editor may help create the font. In this case, the glyph style of the submitter must be respected.
  - (3) If the submitter cannot provide the TrueType font by this time, its collection will be withdrawn from M-set.
- d. Prepare a list of source references.
- e. Produce a packed Multi-column Ideograph Chart using the TrueType fonts.

The IRG will conduct at least one round of review of the proposal and the chart generated using TrueType font before submission to WG2.

## 3. Procedures

This section describes the basic development procedure of CJK Unified Ideograph extensions. The ultimate purpose of the procedure is to realize the production of high quality CJK Unified Ideograph sets in an efficient manner.

The basic development procedure described in this section consists of 8 stages, and it may take two to three years to create a high quality ideograph set for standardization.

### 3.1. Call for Submission

- a. When a member body or an international consortium requests a new project for CJK Unified Ideograph extension and when the project is agreed upon at an IRG meeting, the IRG may call for submission of new ideographs. The IRG will determine the deadline for submission.
- b. Each member body/international consortium with proposed ideographs must submit the ideographs before the specified deadline with the required data described in Section 2 of this document.
- c. Submitters must check whether the submitted ideographs are accompanied with all the required information. If only minor problems are encountered, such as some required information is found missing or misplaced, the IRG chief editor or technical editor may ask the submitter to re-submit the information or supply additional information. Otherwise, the submission may be rejected because consolidation with other submissions cannot be carried out.

### 3.2. Consolidation and Grouping of Submitted Ideographs

Consolidation of submissions is normally done between IRG meetings. The consolidation involves the following tasks:

- a. The IRG technical editor will sort and assign *serial numbers* to submitted ideographs as described in Section 2.3.2.
- b. After serial numbers are assigned, submitted ideographs must undergo IDS checking to detect any duplication and unification. By the result of IDS checking as described in Section 2.3.3, submitted ideographs will be grouped into M-set and D-set as described in Section 2.3.4.
- c. After consolidation, the working drafts will be assigned an IRG document number with a version number. They will be distributed to member bodies' editors and made available on the official IRG website so that any other experts can have access to them. The IRG chief editor and technical editor may assign member bodies' editors to check M-set and D-set ideographs for either the entire collection or certain portions of it depending on their reasonable estimation of the workload.

### 3.3. First Checking Stage

This stage, which is between IRG meetings, involves the following tasks:

- a. Each member body's editor must check the assigned M-set and D-set for data integrity, correctness, missing data and duplication. Checking for unification is not mandatory, but desirable. Typical review comment examples for each set are provided in Section 4.
- b. Member bodies must submit their comments in registered IRG documents to the IRG chief editor and technical editor at least two months before the next IRG meeting or according to the IRG approved working schedule.
- c. The IRG chief editor and technical editor will consolidate the comments and produce a registered IRG document for circulation and discussion at least one month before the next IRG meeting or according the IRG approved working schedule.
- d. Submitters and outside experts are encouraged to prepare and submit supplementary documents (with IRG document numbers) so that they can be discussed at the next IRG meeting.

### 3.4. First Discussion and Conclusion Stage

This stage, which is during an IRG meeting, involves the following tasks:

- a. Member bodies and participating experts should review the comments which are officially submitted before the meeting with assigned IRG document numbers. The editorial group must reach conclusion for each commented ideograph in writing. Guidelines for typical conclusions are provided in Section 4.
- b. All the conclusions must be agreed and endorsed by the IRG plenary in its resolutions. As a result of the resolutions, some ideographs may be removed or moved between M-set and D-Set.
- c. The IRG technical editor will create a new M-set and D-set one month after the IRG meeting, and register them as IRG documents with version information.
- d. If more than 5% of the ideographs submitted by a specific submitter are removed as a result of duplication or unification with existing standardized sets, the entire submission of this submitter will be removed to ensure high quality of the project. This is known as the 5% rule described in Section 2.2.5 above.

- e. If new unification or non-unification cases or rules are agreed upon, such decisions must be recorded in a separate editorial report document. The IRG should also instruct the IWDS editor to modify and update the IWDS according to the guideline set out in Annex E of this document. This update will be reviewed and confirmed in the following IRG meeting so that it can be used in all future work. The 5% rule will not be applicable to unifications based on newly agreed unification rules.

### 3.5. Subsequent Checking Stage

This stage, which is between IRG meetings, involves the following tasks:

- a. Each member body's editor must check the newly created M-set and D-set for correctness and duplication.
- b. Member bodies should submit their comments in registered IRG documents to the IRG chief editor and technical editor at least two months before the next IRG meeting or according to the IRG approved working schedule.
- c. The IRG chief editor and technical editor will consolidate the comments and produce a registered IRG document for circulation and discussion at least one month before the next IRG meeting or according to the IRG approved working schedule.
- d. Submitters and outside experts are encouraged to prepare and submit supplementary documents to facilitate discussion during the next IRG meeting.

### 3.6. Subsequent Consolidation and Conclusion Stage

This stage, which is during an IRG meeting, involves the following tasks:

- a. Member bodies must review the comments and draw conclusion for each ideograph. Typical comment and conclusion examples for each set are provided in Section 4.
- b. All the conclusions must be agreed and endorsed by the IRG plenary in its resolutions. As a result of the resolutions, some ideographs may be removed or moved between M-set and D-set.
- c. The IRG technical editor will create a new M-set and D-set one month after the IRG meeting, and produce a registered IRG document.
- d. If more than 5% of the ideographs submitted by a specific submitter are removed as a result of duplication or unification with existing standardized sets, the entire submission of this submitter will be removed to ensure high quality of the project. This rule will not be applicable to new unifications based on rules added after the first checking stage.

### 3.7. Final Checking Stage

This stage, which is between IRG meetings, involves the following tasks:

- a. All member bodies' editors are requested to check M-set intensively based on comments and conclusions made at all previous stages. At the final checking stage, no ideographs are allowed to be moved from D-Set to M-Set.
- b. Member bodies' editors must submit their comments in registered IRG documents to the IRG chief editor and technical editor at least two months before the next IRG meeting.
- c. The IRG chief editor and technical editor will consolidate the comments and produce a registered IRG document for circulation and discussion at least one month before the next IRG meeting so that member bodies' editors can have time to review them before the next IRG meeting.

### 3.8. Approval and Submission to WG2

This stage, which is during an IRG meeting, involves the following tasks:

- a. Member bodies should review the comments on M-set and reach conclusion for each ideograph.
- b. If there is no positive decision on an M-set ideograph, it will be moved to D-set. No character will be moved from D-set to M-set at this stage. Ideographs may only be moved from M-set to D-set.
- c. With the approval of the majority of IRG member bodies, M-set will be frozen as the new ideograph extension set to be submitted to WG2. The IRG technical editor will prepare the submission in accordance with Section 2.6 of this document.

Once M-set is completed for submission to WG2, records of characters in the D-set will no longer be maintained by the IRG. Characters remained in the D-set can be re-submitted in future extensions if pending problems are solved.


## 4. Guidelines for Comments and Resolutions on Working Sets

Generally speaking, reviewers should put down their comments for any problems they want to alert other reviewers. For comments related to glyph shape, the relevant component(s) of the problem glyph and the referenced glyph(s) should be marked in red circles/boxes in the comment files. Similarly, for comments concerning identical or different components of two or more ideographs, the corresponding components should be indicated in red circles/boxes in the comment files.


All comments must be accompanied with date (in YY-MM-DD format) and member body/international consortium abbreviation (G, H, J, K, KP, M, MY, T, U, V or Z). All conclusions must be dated.

### 4.1. Guidelines for M-Set

The ultimate target of M-set is a standardized ideograph set. As such, it must be carefully examined. If any suspicious characters are found, they will be moved to D-set or removed from the working sets altogether. For comments on glyph shape, the relevant components of the ideographs should be marked in red circles/boxes in the comment file as shown below.

201C3		Wrong Glyph	T glyph seems wrong: 化 in T glyph
-------	---	-------------	-----------------------------------

Similarly, for comments concerning identical or different components of two or more ideographs, the corresponding components should be marked in red circles/boxes in the comment file as shown below.

2010D		Glyph design	The T-glyph is different from the <i>KX Dictionary</i> glyph.
-------	---	--------------	---

The table below gives examples of review comments and possible actions associated with these problems:

Possible Comment by a Reviewer	Possible Resolution
Wrong or Missing Glyph	<ul style="list-style-type: none"> <li>The wrong glyph is corrected, or the missing glyph supplied. The ideograph will be moved to D-set for manual checking.</li> </ul>
Wrong Kangxi radical / stroke count / first stroke	<ul style="list-style-type: none"> <li>Data will be corrected and the ideograph will be moved to D-set for further manual checking.</li> </ul>
Wrong IDS	<ul style="list-style-type: none"> <li>IDS will be corrected and the character will be moved to D-set for checking by the IDS checker.</li> <li>Move to D-set (in case IDS cannot be corrected).</li> </ul>
May be unifiable with U+xxxxx (standardized ideograph)	<ul style="list-style-type: none"> <li>Unified with U+xxxxx and the submitter will request a new source reference to U+xxxxx.</li> <li>Unified with U+xxxxx and the submitter will request that this character be treated as a Compatibility Ideograph.</li> <li>Unified to U+xxxxx and this entry will be removed. (May consider to register it as IVS.)</li> <li>Not unifiable.</li> </ul>
May be unifiable with xxxxx (M-set)	<ul style="list-style-type: none"> <li>Unified with xxxxx and this source reference</li> </ul>

ideograph)	<p>will be attached to xxxxx.</p> <ul style="list-style-type: none"> <li>Unified with xxxxx and the submitter may consider registering it as a Compatibility Ideograph Character or IVS.</li> <li>Not unifiable.</li> </ul>
------------	---

#### 4.2. Guidelines for D-Set

Ideographs in D-Set are either the ones that cannot be checked automatically by the IDS checking algorithm or the ones whose attribute data have been questioned by member bodies or whose unification with other standardized or working set ideographs have been proposed. For those ideographs that cannot be machine-checked by IDS matching, at least two non-submitter member bodies must check them manually to ensure that they are not unifiable with any standardized or working set ideographs. For those ideographs that might be unifiable with other ideographs, the submitters are requested to prepare arguments and evidence to show that such ideographs should be separately encoded.

Possible Comment by IDS Checker	Possible Conclusion
<ul style="list-style-type: none"> <li>Incomplete IDS IDS with extra character</li> <li>Component is not an ideograph</li> </ul>	<ul style="list-style-type: none"> <li>IDS will be corrected and the character will be moved to M-set when next IDS-checking is done.</li> <li>Proper IDS cannot be generated and manual checking is needed.</li> </ul>
Possible Comment by a Reviewer	Possible Conclusion
<ul style="list-style-type: none"> <li>Wrong Kangxi radical</li> <li>Wrong stroke count</li> <li>Wrong first stroke</li> </ul>	<ul style="list-style-type: none"> <li>Data will be corrected.</li> <li>Proposal to correct data is not accepted, as it is an ambiguous case and the IRG agrees that the previous choice of XX is more appropriate.</li> </ul>
<ul style="list-style-type: none"> <li>Wrong IDS</li> </ul>	<ul style="list-style-type: none"> <li>IDS will be corrected and checked by the IDS checker again.</li> <li>Correct IDS cannot be generated and manual checking is needed.</li> </ul>
May be unifiable with U+xxxxx (standardized ideograph)	<ul style="list-style-type: none"> <li>Unified with U+xxxxx and a new source will be added to U+xxxxx. The new candidate entry should be deleted.</li> <li>Not unifiable, as shown by the evidence <i>IRG N xxxx</i>. Move to M-set.</li> </ul>
May be unifiable with xxxxx (M-set or D-set ideograph)	<ul style="list-style-type: none"> <li>Unified with xxxxx in M-set and a new source will be added to xxxxx. The new candidate entry should be deleted from D-Set.</li> <li>Unified with xxxxx in D-Set and a new source will be added to xxxxx. The new candidate entry should be removed from D-Set.</li> <li>Not unifiable, as shown by the evidence <i>IRG N xxxx</i>. Move to M-set.</li> </ul>
Checked against all standardized and working set ideographs with radical X and stroke count of Y±2 for characters that cannot be described by IDS for automatic checking.	<ul style="list-style-type: none"> <li>Move to M-set, as two non-submitter member bodies (XX and YY) confirmed that this ideograph is not unifiable with any existing standardized or working set ideographs.</li> <li>Checking against ideographs with radical X may not be enough. This ideograph will also be checked against ideographs with radical Z.</li> </ul>

## 5. IRG Website

The IRG maintains its own website at <http://www.cse.cuhk.edu.hk/~irg/>, hosted by the Department of Computer Science and Engineering at The Chinese University of Hong Kong. IRG meeting notices, resolutions, document register, documents and standing documents are made available on this site.



Hyperlinks to WG2 websites are provided for member bodies' easy access. For faster retrieval of documents and searching, documents should not be compressed as far as possible and the site search engine window should be made available. Documents larger than 4MB must be split into multiple files for easy uploading, downloading and searching. The compressed files can either be in WinZip format with .zip extension or RAR format with .rar extension.

## 6. IRG Document Registration

All documents to be formally discussed by the IRG must be registered with IRG document numbers assigned by the IRG Rapporteur and contain the submission date, title, name of the submitter or author, purpose (or summary), and the "IRG Ideographic Repertoire Submission Summary Form" (when applicable).

### 6.1. Registration Procedure

The following gives the registration procedure:

- a. **Request for Document Number:** All documents submitted to the IRG must be given a document number. The number is to be assigned by the IRG Rapporteur. The submitter should first contact the IRG Rapporteur for a document number with a document title. Once the document number is assigned, the information will be posted on the IRG website. Document numbers can be pre-assigned during IRG meetings for activities between IRG meetings.
- b. **Submission of Documents:** All registered documents must be submitted to the IRG Rapporteur. The submitted documents must contain an assigned IRG document number in text form (except files of pure tables to avoid interfering with the data presented in the table) so that searching can be supported.
- c. **Posting of Documents:** Properly submitted documents are then posted by the IRG Rapporteur on the IRG website as official documents and the submitters will be notified by the IRG Rapporteur by email. The submitters should double check the posted documents upon receiving the emails to ensure that the intended documents are properly posted for viewing by the public.
- d. **Disqualified Documents:** Documents with certain basic information missing such as the submitter's name, title and purpose may be rejected by the IRG Rapporteur for posting. All other documents which fail to comply with the above registration process and the preliminary review by the IRG Rapporteur for basic information will not be treated as IRG documents. As such, issues contained in such documents will not be discussed by the IRG formally.

### 6.2. Contact for IRG Document Registration

The current IRG Rapporteur is Prof. Qin LU and her contact information is as follows:

Professor Qin Lu  
Department of Computing  
The Hong Kong Polytechnic University  
Hung Hom, Hong Kong  
Tel. (852) 2766 7247  
Fax. (852) 2774 0842  
Email: csluqin@comp.polyu.edu.hk

## Annex A: Sorting Algorithm of Ideographs

The IRG recognizes that the choice of radicals, the sequence of strokes, and the stroke counting methods are locale dependent. Submitters may have different preferences of character orders. However, for the convenience of IRG editorial work, the IRG must adopt a sorting order which may be different from the submitters' preferences. Thus the principles of sorting of ideographs given below are internal for IRG editing purposes only. Ideographs consolidated for unification review must be sorted according to the following order.

### a. Kangxi Radical Order

**Note:** When radical glyphs are in simplified forms given below, ideographs with the simplified radical glyphs must be placed after ideographs with the corresponding traditional radicals.

Traditional Radicals		Simplified Radicals	
R090.0	月	R090.1	丩
R120.0	糸	R120.1	纟
R147.0	見	R147.1	见
R149.0	言	R149.1	讠
R154.0	貝	R154.1	贝
R159.0	車	R159.1	车
R167.0	金	R167.1	钅
R168.0	長	R168.1	长
R169.0	門	R169.1	门
R178.0	韋	R178.1	韦
R181.0	頁	R181.1	页
R182.0	風	R182.1	风
R183.0	飛	R183.1	飞
R184.0	食	R184.1	饣
R187.0	馬	R187.1	马
R195.0	魚	R195.1	鱼
R196.0	鳥	R196.1	鸟
R197.0	鹵	R197.1	卤
R199.0	麥	R199.1	麦
R205.0	黽	R205.1	龟
R210.0	齊	R210.1	齐
R211.0	齒	R211.1	齿
R212.0	龍	R212.1	龙
R213.0	龜	R213.1	龟

### b. Stroke Count

**Note:** Simplified characters must be placed after traditional characters within the same stroke-number group.

**c. First Stroke**

The technical editor will assign the first stroke based on IRGN954AR and IRGN1105. In case of previously unseen components, the technical editor will take the conventions of Kangxi for first stroke assignment without regard to the submitters' locale conventions.

## Annex B: IDS Matching

### B.1. Guidelines on Creation of IDS

Each member body should consult IRG N1183 on IDS. In addition to the CDC (Character Description Components) defined in IRG N1183, all CJK Unified Ideographs accepted by ISO/IEC 10646 in its amendments are also qualified as CDC in constructing IDS.

The use of “overlapping” IDC (Ideographic Description Characters) or more than four IDCs is considered to be “inappropriate” and may not be a subject of IDS comparison.

### B.2. Requirements of IDS Matching

The IDS matching algorithm used by the IRG should support the following features:

1. Handling different split points.  
(e.g. 𠄎𠄎頃 and 𠄎𠄎化頁 should be matched.)
2. Handling different split levels.  
(e.g. 𠄎𠄎𠄎悉 and 𠄎𠄎𠄎采心 should be matched.)
3. Matching different glyphs of the same abstract shape.  
(e.g. 𠄎𠄎𠄎申 and 𠄎𠄎𠄎示申 should be matched.)
4. Matching similar glyphs.  
(e.g. 𠄎𠄎𠄎生 and 𠄎𠄎𠄎小生 should be matched.)
5. Matching IDS with different orderings of overlapping IDC.  
(e.g. 𠄎𠄎𠄎三 | and 𠄎𠄎 | 三 should be matched.)
6. Matching unifiable IDC patterns.  
(e.g. 𠄎𠄎麥离 and 𠄎𠄎麥离 should be matched.)
7. Handling any combinations of the above.
8. Detecting any inappropriate IDS, such as IDS being too long, IDS with non-ideographic CDC, or missing or extra CDC or IDC.

### B.3. Limitation of IDS Matching

It should be noted that IDS matching cannot detect unification or duplication if a component cannot be encoded by an IDS, or if the glyph itself is very complex. IDS matching is done algorithmically. It is not versatile in detecting unifiable ideographs unless rules are explicitly given to the algorithm. Thus, it is not meant to be the replacement of manual checking. Rather, it is an assistive tool for quality assurance to identify duplication and known cases of unification. Therefore, it is very important for submitters to make sure that their submitted ideographs are not going to be unified with any standardized or previously discussed ideographs or working set ideographs.

## Annex C: Urgently Needed Ideographs

### C.1. Introduction

When a member body or an internationally recognized organization, consortium, or individual, as a submitter, demonstrates an urgent need for a small number of ideographs to be standardized for justifiable reasons, such as ideographs in a recently developed regional or national standard that must be implemented by a particular deadline, the IRG may submit the ideographs, independent of any of the current IRG working sets to WG2. Each urgently-needed submission will be treated as a separate urgently-needed repertoire, and a submitter can have no more than one active urgently-needed submission at a time. The process will be started only sparingly with demonstrated need.

### C.2. Requirements

Each submission should include no more than 30 ideographs. Submissions of more than 30 characters will be accepted at the sole discretion of the IRG. A submitter of urgently-needed ideographs must prepare the following:

- a. All the documents required for normal ideograph submissions.
- b. Justifications of the submission. A submission is deemed urgently-needed only if the submitter demonstrates urgency or a rationale for rapid standardization.
- c. A document that indicates whether, among the submitted urgently-needed ideographs, there are any ideographs that can be unified with ideographs in the current IRG working sets in addition to those in the standard or its amendments. When a particular urgently-needed repertoire is accepted by WG2, any unifiable ideographs in the current working sets will be removed as explained in C.3 below.
- d. For the rest of the submitted urgently-needed ideographs, the document must prove that they are not unifiable with any ideographs in the current working sets. The proof may be provided by listing the documents the submitter has checked, and for each proposed ideograph, a list of ideographs whose radicals and strokes have been checked against. It is an important responsibility of the submitter to check with not only the current standardized CJK ideographs, but also the IRG working sets for any unifiable characters against its submission. If a submitter fails to do the above, the submission will not be approved by the IRG as an IRG-endorsed independent submission to WG2.

### C.3. Dealing with Urgent Requests

Accepted urgently-needed ideographs as independent submissions must be checked by the IRG for correctness, duplication and unification against the latest published ISO/IEC 10646 as well as the current IRG working sets. When an urgently-needed ideograph is found to be identical or unifiable with any ideograph in the current IRG working sets, the latter must be noted and removed from the current IRG working sets.

## Annex D: Up-to-date CJK Unified Ideograph Sources and Source References

The IRG tracks the sources of the CJK Unified Ideographs. The past practice is that the character sources were tracked based on submissions by member bodies. Thus every member body has been assigned a member body abbreviation as defined in ISO/IEC 10646. In recognizing that submitters might be an international consortium not affiliated with any member body, the IRG decided at IRG Meeting No. 39 to add a new abbreviation “Z” for this type of submissions. It is noted that Z source may include submissions from different projects and additional letters may be used to indicate each repertoire.

### D.1. Member body abbreviations:

G	China
H	Hong Kong Special Administrative Region, China
J	Japan
K	Republic of Korea
KP	Democratic People’s Republic of Korea
M	Macao Special Administrative Region, China
MY	Malaysia (added in Nov. 2008 at IRG Meeting No. 31)
T	Taipei Computer Association
U	Unicode Consortium
V	Vietnam

Note: all member body abbreviations except MY are currently used in ISO/IEC 10646 Section 23. MY is defined and used by IRG internally only.

### D.2. The Hanzi G sources

G0	GB2312-80
G1	GB12345-90 with 58 Hong Kong and 92 Korean “Idu” characters
G3	GB7589-87 unsimplified forms
G5	GB7590-87 unsimplified forms
G7	General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi
GS	Singapore Characters
G8	GB8565-88
G9	GB18030-2000
GE	GB16500-95
GH	GB15564-1995 Code of Chinese Ideogram set for teltext broadcasting Hong Kong subset
GK	GB12052-89 Korean Character Coded Character Set for Information Interchange
G_4K	Siku Quanshu (四庫全書)
G_BK	Chinese Encyclopedia (中國大百科全書)
G_CH	Ci Hai (辭海)
G_CY	Ci Yuan (辭源)
G_CYY	Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字)
G_GDZ	Geographic Publishing House Ideographs(地质出版社用字)
G_FZ	Founder Press System (方正排版系统)
G_GH	Gudai Hanyu Cidian (古代汉语词典)
G_HC	Hanyu Dacidian (漢語大詞典)
G_HZ	Hanyu Dazidian ideographs (漢語大字典)
G_IDC	ID system of the Ministry of Public Security of China, 2009
G_GJZ	Commercial Press Ideographs (商务印书馆用字)
G_GKX	GKX Kangxi Dictionary ideographs (康熙字典) 9th edition (1958) including the addendum (康熙字典) 補遺
G_GRM	People’s Daily Ideographs(人民日报用字)
G_GXC	Xiandai Hanyu Cidian (现代汉语词典)
G_XH	Xinhua Zidian (新华字典)
G_WZ	Hanyu Dacidian Publishing House Ideographs (漢語大詞典出版社用字)
G_ZFY	Hanyu Fangyan Dacidian (汉语方言大辞典)
G_ZH	Zhonghua Zihai (中华字海)

G\_ZJW Yinzhou Jinwen Jicheng Yinde (殷周金文集成引得)

D.3. Hanzi H sources

H Hong Kong Supplementary Character Set (HKSCS)

D.4. Hanzi T sources

T1 TCA-CNS 11643-1992 1st plane  
T2 TCA-CNS 11643-1992 2nd plane  
T3 TCA-CNS 11643-1992 3rd plane with some additional characters  
T4 TCA-CNS 11643-1992 4th plane  
T5 TCA-CNS 11643-1992 5th plane  
T6 TCA-CNS 11643-1992 6th plane  
T7 TCA-CNS 11643-1992 7th plane  
TB TCA-CNS 11643-2007 11th plane  
TC TCA-CNS 11643-2007 12th plane  
TD TCA-CNS 11643-2007 13th plane  
TE TCA-CNS 11643-2007 14th plane  
TF TCA-CNS 11643-2007 15th plane

D.5. Kanji J sources

J0 JIS X 0208-1990  
J1 JIS X 0212-1990  
J3 JIS X 0213:2000 level-3  
J3A JIS X 0213:2004 level-3  
J4 JIS X 0213:2000 level-4  
JA Unified Japanese IT Vendors Contemporary Ideographs, 1993  
JH Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009  
JK Japanese KOKUJI Collection  
JARIB Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

D.6. Hanja K sources

K0 KS C 5601-1987(Now known as KS X 1001:2004)  
K1 KS C 5657-1991(Now known as KS X 1002:2001)  
K2 PKS C 5700-1 1994 (Reedited and standardized as KS X 1027-1:2011)  
K3 PKS C 5700-2 1994(Reedited and standardized as KS X 1027-2:2011)  
K4 PKS 5700-3:1998(Reedited and standardized as KS X 1027-3:2011)  
K5 Korean IRG Hanja Character Set  
5th Edition: 2001(Reedited and standardized as KS X 1027-4:2011)

D.7. Hanja KP sources

KP0 KPS 9566-97  
KP1 KPS 10721-2000

D.8. ChuNom V sources

V0 TCVN 5773:1993  
V1 TCVN 6056:1995  
V2 VHN 01:1998  
V3 VHN 02: 1998  
V4 Dictionary on Nom 2006, Dictionary on Nom of Tay ethnic 2006, Lookup Table for Nom in the South 1994

D.9. MY sources

MY "Dictionary Of Chinese Rustic Language In South-East Asia", written by Xu Yunqiao, published by Singapore Shjie Publisher, 1961. 《南洋华语俚俗辞典》，新加坡世界书局有限公司，1961年8月

D.10. Macao sources

MAC Macao Supplementary Character Set

D.11. Unicode sources

UTC Unicode Standard Annex #45 (Used to be called Unicode Technical Report #45), U-source Ideographs, Sept. 2013 and its future extensions

#### D.12. Z sources

This is a new source created since IRG Meeting No. 39 to accommodate submissions from international groups working on CJK ideographs. Currently, there is only one such submission from the SAT project.

Z\_SAT SAT Daizōkyō Text Database Committee



## Annex E: Maintenance Procedure of the IRG Working Documents Series

### E.1 Introduction

The IRG Working Documents Series (IWDS) is a set of IRG maintained documents which keep the up-to-date examples of CJK unification related cases to supplement the published Annex S of ISO/IEC 10646 for IRG unification work.

### E.2. IRG Working Documents Series

The formats of the IWDS and the specific lists of examples are maintained as a separate set of documents as follows.

Series 1: Summary of unification rules and examples (File name: IWDS\_SUM.pdf)

Series 2: List of UCV (Unifiable Component Variations) of Ideographs (File name: IWDS\_UCV.pdf)

Series 3: List of Non-unifiable Components of Ideographs and Overly-unified Ideographs (File name: IWDS\_NUC.pdf)

Series 4: List of Possibly Mis-unified Ideographs (File name: IWDS\_MUI.pdf).

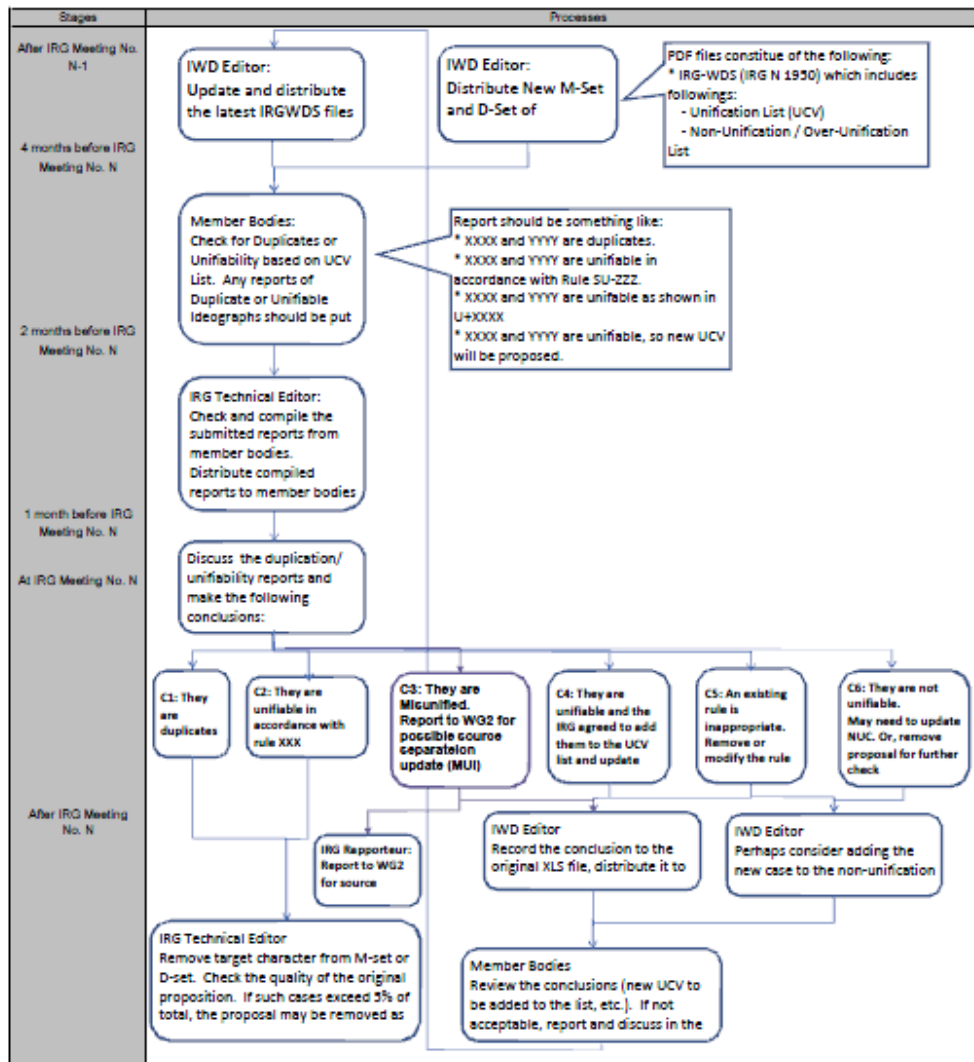
### E.3. Maintenance Procedure

The maintenance procedure describes how entries in the IWDS are added, removed, or changed. The IRG has an appointed IWDS editor (currently, Mr. Taichi Kawabata) who is in charge of the maintenance of the IWDS.

In principle, all update requests are results of IRG unification review work. A review cycle between two IRG meetings is needed. Every update must be discussed in at least one IRG meeting and confirmed in writing. An update normally starts from the unification review work assigned to member bodies in the past IRG meeting (Meeting No. N-1). During the review work before the next IRG meeting (Meeting No. N), if member bodies find duplicates, unifiable cases or mistakes which warrant a change in the IWDS, they need to report these cases in a specific form attached to IWD Series 1. These reported cases will then be consolidated by the IRG technical editor before IRG Meeting No. N. During IRG meeting No. N, time must be allocated to discuss these reported cases and conclusions must be recorded during this IRG meeting. Based on the confirmed conclusions on IWDS updates, the IWDS editor will update the IWDS. Any unclear conclusions will be further discussed in future meetings.

It takes time to update the IWDS, and sometimes it is difficult to find appropriate examples. The IRG has, therefore, requested the IWDS editor to keep a log of the actions carried out based on IRG instructions so that better tracking of changes can be carried out.

Below is the description of the maintenance procedure as a flow chart.



An attachment file in Excel form is more clear

# Annex F: IRG Repertoire Submission Summary Form

**ISO/IEC JTC 1/SC 2/WG 2/IRG**  
**PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS**  
**FOR ADDITION OF CJK UNIFIED IDEOGRAPHS TO THE REPERTOIRE OF ISO/IEC 10646**  
 Please fill in all the sections below.  
 Please read the Principles and Procedures Document (P & P) from  
<http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg41/IRGN1975PnPv6.doc>  
 for guidelines and details before filling in this form.  
 Please ensure that you are using the latest Form from  
[http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg41/IRGN1975\\_PnP\\_BlankDataFile.xls](http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg41/IRGN1975_PnP_BlankDataFile.xls)  
 See also <http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html> for the latest *Unifiable Component Variations*.

**A. Administrative**

1. **IRG Project Code:** \_\_\_\_\_ e.g. Extension-F

2. **Title:** \_\_\_\_\_

3. **Requester's Region/Country Name:** \_\_\_\_\_

4. **Requester Type (National Body/Individual Contribution):** \_\_\_\_\_

5. **Submission Date:** \_\_\_\_\_

6. **Requested Ideograph Type (Unified or Compatibility Ideographs)** \_\_\_\_\_  
 If Compatibility, does the requester have the intention to register them as IVS (See UTS #37) with the IRG's approval? (Registration fee will not be charged if authorized by the IRG.) \_\_\_\_\_

7. **Request Type (Normal Request or Urgently Needed)** \_\_\_\_\_

8. **Choose one of the following:**  
 This is a complete proposal \_\_\_\_\_  
 (or) More information will be provided later. \_\_\_\_\_

**B. Technical – General**

1. **Number of ideographs in the proposal:** \_\_\_\_\_

2. **Glyph format of the proposed ideographs: (128x128 Bitmap files or TrueType font file)** \_\_\_\_\_  
 If Bitmap files, are their file names the same as their source references? \_\_\_\_\_  
 If TrueType font file, are all the proposed glyphs put into BMP PUA area? \_\_\_\_\_  
 If TrueType font file, are data for source references vs. character codes provided? \_\_\_\_\_

3. **Source references:**  
 Do all the proposed ideographs have a unique, proper source reference (member body/international consortium abbreviation followed by no more than 9 alphanumeric characters)? \_\_\_\_\_

4. **Evidence:**  
 a. Do all the proposed ideographs have a separate evidence document which contains at least one scanned image of printed materials (preferably dictionaries)? \_\_\_\_\_  
 b. Do all the printed materials used for evidence provide enough information to track them by a third party (ISBN numbers, etc.)? \_\_\_\_\_

5. **Attribute Data Format: (Excel file or CSV text)** \_\_\_\_\_

### C. Technical - Checklist

#### Understanding of the Unification Policy

1. Has the requester read ISO/IEC 10646 Annex S and does the requester understand the unification policy?
2. Has the requester read the “Unifiable Component Variations” (contact IRG technical editor through the IRG Rapporteur for the latest version) and does the requester understand the unifiable variation examples?
3. Has the requester read the P&P document and does the requester understand the 5% Rule?

#### Character-Glyph Duplication (<http://www.itscj.ipsj.or.jp/sc2/open/pow.htm> contains all the published ones and those under ballot)

4. Has the requester checked that the proposed ideographs are **not unifiable** with any of the unified or compatibility ideographs of the latest version of ISO/IEC 10646?  
If the checking has been done against an earlier version of ISO/IEC 10646, please specify the version? (e.g. 10646:2011)
5. Has the requester checked that the proposed ideographs are **not unifiable** with any of the ideographs in the amendments, if any, of the latest version of ISO/IEC 10646?  
If yes, which amendment(s) has the requester checked?
6. Has the requester checked that the proposed ideographs are **not unifiable** with any of the ideographs in the proposed amendments, if any, of ISO/IEC 10646?  
If yes, which draft amendment(s) has the requester checked?
7. Has the requester checked that the proposed ideographs are **not unifiable** with any of the ideographs in the current working M-set and D-set of the IRG? (Contact IRG chief editor and technical editor through the IRG Rapporteur for the newest list)  
If yes, which document(s) has the requester checked?
8. Has the requester checked that the proposed ideographs are **not unifiable** with any of the over-unified or mis-unified ideographs in ISO/IEC 10646? (See Annex E of the P&P document).
9. Has the requester checked whether the proposed ideographs have any **similar ideographs** in the current standardized or working sets mentioned above?
10. Has the requester checked whether the proposed ideographs have any **variant ideographs** in the current standardized or working sets mentioned above?

#### Attribute Data

11. Do all the proposed ideographs have attribute data such as the Kangxi radical code and stroke count?
12. Are there any simplified ideographs (ideographs that are based on the policy described in 簡化字總表) among the proposed ideographs?  
If yes, does the proposal include proper simplified/traditional indication flag for each proposed ideograph in the attribute data?
13. Do all the proposed ideographs have the document page number of evidence documents in the attribute data?
14. Do all the proposed ideographs have the proper Ideographic Description Sequence (IDS) in the attribute data?  
If no, how many proposed ideographs do not have the IDS?
15. If the answer to question 9 or 10 is yes, do the attribute data include any information on similar/variant ideographs for the proposed ideographs?

# Annex G: Examples of New CJK Unified Ideographs Submissions (i.e., Vertical Extension)

## G.1. Sample Data Files

All submitted characters must follow the submission format given in Section 2.2.1. The following gives a sample list of characters submitted by Japan for consideration in CJK Ext. F work (IRGN 1887).

Seq.No.	a) Source reference	b) Glyph Image file name	c) KangXi Radical Code	d) Stroke Count	e) First Stroke	f) T/S (traditional 0, or simplified 1)	g) IDS (Ideographic Description Sequence)	h) Similar/Variants	i) Ref. to Evidence doc	i1) file name of evidence image
1	JMJ-058706	JMJ-058706.bmp	145.0	10	3	0	𠄎鬼衣	No	SD-pp2105-i15349	SD(Kodansha's Shin Dai Ziten).pdf
2	JMJ-058707	JMJ-058707.bmp	145.0	10	2	0	𠄎+𠄎三𠄎マ木衣	U+8943	SD-pp2105-i15352	SD(Kodansha's Shin Dai Ziten).pdf
3	JMJ-048391	JMJ-048391.bmp	145.0	11	2	0	𠄎广𠄎𠄎+?比衣	U+2772B	MD-pp10706-i34559	MD(Dai Kanwa Ziten).pdf

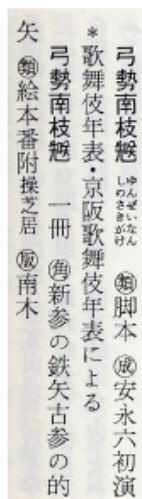
## G.2. Sample Evidence

All character submissions must include evidence of use as specified in Section 2.2.3. The following shows an example of a Japanese submission with reference to the use of the character in ancient books (IRG N1225 Part2).

JMK66114

𠄎

【補訂版国書総目録】(1969年4月30日第1刷発行, 2002年7月5日補訂版第4刷発行, 岩波書店)  
第7巻 870 ページ 4段



## G.3. Handling of Data with Privacy Concerns

The IRG understands that privacy laws and practices in some countries and regions can make the submission of complete records as evidence related to personal information difficult. As a compromise, the IRG suggests member bodies to provide evidence in such a way that it would not reveal complete personal/internal information. However, the character information itself must be shown in the supplied evidence. In other words, partial document images should be supplied with certain sensitive information



already supported in computer systems that are maintained by a designated government body with wide use by government bodies and citizens for administrative public service.

The IRG recognizes that some of the characters included in e-government systems cannot be provided with supporting evidence of actual use according to Section 2.2.3.a, and yet it is technically and administratively not practical to remove them from the systems. Thus the IRG is willing to consider their acceptance without actual evidence provided that they are from already implemented working systems only. However, the IRG requires the submitter to provide information on the quality assurance process for the maintenance of the character collection concerned. The submitter must supply information on the accessibility of the character collection and the working system, the stability and traceability of the collection, and the kind of evidence/information needed for approval of character removal, modification and addition by the administrative body of the collection.

## **Annex H [Reserved for future use]**

Annex H is purposely left out for the time being so that IRG Annex numbers tally with WG2 Annex numbers where the subjects are the same.



# Annex I: Guideline for Handling of CJK Ideograph Unification or Dis-unification Errors

(Same as WG2 Principles and Procedures Annex I)

Source: [www.dkuug.dk/jtc1/sc2/wg2/docs/principles.html](http://www.dkuug.dk/jtc1/sc2/wg2/docs/principles.html)

Below is the text extracted from Annex I of WG2 PnP (WG2N4102) dated 2012-01-04:

There are two kinds of errors that may be encountered related to coded CJK unified ideographs.

Case 1: *to be unified* error - Ideographs that should have been unified are assigned separate code points.

Case 2: *to be disunified* error - Ideographs that should not have been unified are unified and assigned a single code point. An example of this is the request from TCA in document [N2271](#).

When such errors are found, the following guidelines will be used by WG 2 to deal with them.

## I.1 Guideline for “to be unified” errors

- A. The “*to be unified*” pair will be left disunified. Once a character is assigned a code position in the standard, it will not be removed from the standard.
- B. If necessary, an additional note may be added to an appropriate section in the standard.

## I.2 Guideline for “to be disunified” errors

(Source: [ISO/IEC JTC 1/SC 2/WG 2 N3859](#) – 2010-06-28)

- A. The ideographs to be disunified should be disunified and should be given separate code positions as soon as possible (disunification in some sense, and character name change in some sense also). These ideographs will have two separate glyphs and two separate code positions. One of these ideographs will stay at its current encoded position. The other one will have a new glyph and a new code position.
- B. For the ideographs that are encoded, the code charts in ISO/IEC 10646 are presented in multiple columns, with possibly differing glyph shapes in each column. The question of which glyph shall remain in the current code point will be resolved by the IRG on a case by case basis.
- C. The disunified ideograph will have a glyph that is different from the one that retains the current code position.
- D. The net result will be an addition of new ideograph character and a correction and an additional entry to the source reference table.

## I.3 Discouragement of new disunification request

There is a possibility of “pure true disunification” request. This is almost like the new source code separation request. This kind of request shall not be accepted disregarding the reasoning behind. Key difference between “TO BE DISUNIFIED” and “SHALL NOT BE DISUNIFIED” is as follows.

- a. If character pair is non-cognate (meanings are different), that pair of characters is TO BE DISUNIFIED.
- b. If a character pair is cognate (means the same but different shape), that pair of characters SHALL NOT BE DISUNIFIED.

Disunification request with reason of mis-application (over-application usually) of unification rule should NOT be accepted due to the principle in resolution [M41.11](#).

# Annex J: Guideline for Correction of CJK Ideograph Mapping Table Errors

(Same as WG2 P&P Annex J)

Source: [ISO/IEC JTC 1/SC 2/WG 2 N2577](#) – 2003-09-02

In principle, the mapping table or reference to code points of an existing national/regional standard (in the source reference tables) must not be changed. But once a fatal error is found, it should be corrected as early as possible, under the following guidelines:

## J.1 Priority of Error Correction Procedure

- A. Consider adding a new code position and source-reference mapping for the character in question rather than changing the mapping table.
- B. If the change of mapping table is unavoidable, correction should be done as soon as possible.

## J.2 Announcement of Addition to or Correction of Mapping Table

Once any addition to or correction of the mapping table is made, an announcement of the change should be made immediately. Usually this will be in the form of a resolution of a WG2 meeting, followed by a subsequent process resulting in an appropriate amendment to the standard.

## J.3 Collection and Maintenance of Mapping Tables that are not Owned by WG2

There are many mapping tables, which are included in national/regional standards or developed by third parties. These are out of WG2's scope. Any organization (such as Unicode Consortium) that collects mapping information, maintains it consistently and makes this information widely available is invited and encouraged to do so.

## Annex K: List of First Strokes

Below gives the list of first strokes including their glyphs and names in English and Chinese (with pinyin provided).

Glyph	Stroke No.	Name	Name in Chinese	Pinyin
一	1	Horizontal bar	橫	heng2
丨	2	Vertical bar	豎	shu4
丿	3	Slash	撇	pie3
丶	4	Dot	點	dian3
乙	5	Turn	折	zhe2

# Annex L: Guidelines for Forming Current Working Sets with an Upper Limit

As stated in Section 2.2.1d, the IRG sets an upper limit when forming the working set for review to ensure sufficient time for quality output can be delivered in a timely manner. The current limit is set to 4,000 ideographs ( $Limit_{IRG}$ ). Since the number of submissions and their repertoire sizes can be different each time a new working set is formed, IRG need some basic guidelines on how the working set will be formed with consideration of both fairness and accommodation for various needs. This Annex serves for this purpose.

At the start of the development work, member bodies submit their proposals. Let us assume that the number of submissions is  $N$ .

If the total number of ideographs is less than  $Limit_{IRG}$  (or reasonably close to  $Limit_{IRG}$ ), all submissions are then used to form the working set.

If the total number of ideographs is much larger than  $Limit_{IRG}$ , setting an upper limit to each submission is needed to decrease the total number. The general principle based on simple mathematic calculation is given below:

**Scenario 1:** The simplest case solution is all submissions should not be bigger than  $Limit_{single\_submission} = Limit_{IRG} / N$ . This works especially well if all submission sizes are larger than  $Limit_{single\_submission}$ .

**Scenario 2:** In case that there are some submissions which size ( $TOTAL_{single}$ ) is less than  $Limit_{single\_submission}$ . In other words, there are some spare quota,  $Spare_{single\_submission} = Limit_{single\_submission} - TOTAL_{single}$ . In this case, the total number of spare quota can be equally divided to other submissions whose total exceeds  $Limit_{single\_submission}$ . In case the remaining quotas cannot be equally divided, the remaining quota can be split under the same principle (recursively) for the remaining submissions who still can take the unused quota.

Even though using the above mathematical method can give a quick and undisputed limit to each submission, it may not be the best solution when considering the practical need of the submitters for different applications. Submitters are encouraged to divide their collection and give them priority levels with explanation and justifications. IRG can consider these justifications to arrive at an agreed division of  $Limit_{IRG}$  to be close to the numbers given in the mathematic model with minor modifications.

It should be noted that the upper limit,  $Limit_{IRG}$ , is indicative and set based on IRG's experience from past review that targeted for a one year review cycle. Minor modification to this limit can be allowed because unification among submissions and the withdrawal of characters by submitter can potentially reduce the total number of characters eventually included as a repertoire for WG2 submission.

For those ideographs that are in the current submissions, but not included in the current working set decided by IRG, submitters can submit them in the future with adjusted priority as needed.

## References

Document numbers in the first column in the following table refer to IRG working documents (ISO/IEC JTC 1/SC 2/WG 2/IRGNxxxx), except where noted otherwise. For documents with no link, one may try <http://www.cse.cuhk.edu.hk/~irg/> ; some older documents may only be available in paper form (contact the IRG Rapporteur Prof. Qin LU ).

Doc. No.	Title	Source	Date
<a href="#">WG2 N4102</a>	Principles and Procedures for Allocation of New Characters and Scripts and Handling of Defect Reports on Character Names	WG2	2012-01-04
<a href="#">N681</a>	Annex S <a href="http://standards.iso.org/ittf/PubliclyAvailableStandards/c039921_ISO_IEC_10646_2003(E).zip">http://standards.iso.org/ittf/PubliclyAvailableStandards/c039921_ISO_IEC_10646_2003(E).zip</a>	Bruce Peterson and IRG Rapporteur	1999-11-18
N881	CJK Extension C Submission Format	IRG	2001-12-04
N953	Minutes of the Adhoc meeting on submitted documents: N941, N942, N944, N945, N948, N949	CJK ad hoc group	2002-11-22
N954	Report on first stroke/stroke count by ad hoc group	CJK ad hoc group	2002-11-22
<a href="#">N954AR</a>	N954 Appendix: First Stroke / Stroke Count Chart	CJK ad hoc group	2002-11-21
N955	IRG Radical Classification	Ideograph Radical Ad Hoc	2002-11-21
N956	Ideograph Unification	Ideograph Radical Ad Hoc	2002-11-21
<a href="#">N1105</a>	Amendments to IRG N954AR	Macao	2005-01-03
<a href="#">N1183</a>	IDS decomposition principles(Revised by the IRG)	KAWABATA, Taichi	2005-12-28
<a href="#">N1197</a>	Sample evidence for CJK C1 candidates	Japan	2006-05-22
<a href="#">N1372</a>	On Better use of IDS on IRG development process	KAWABATA, Taichi	2007-11-09
<a href="#">SC2 N3933</a>	ISO/IEC JTC 1 Directives, 5 <sup>th</sup> Edition, Version 3.0	SC2	2007-04-06

## Glossary

**Abstract shape:** Ideographic characters are used as symbols to represent different entities and used for different purposes. The same character conceptually can sometimes be written in different actual shapes with minor stroke differences, due to preference, which do not affect the recognition of the character as a unique symbol. These characters having the same abstract shapes are not coded separately because ISO/IEC 10646 is a character (symbol) standard, not a glyph standard. In other words, character glyphs (actual shapes) that are considered to have the same abstract shapes are to be unified under the CJK unification rules (defined in Annex S of ISO/IEC 10646).

As ideographs are formed by both the components and the relative positioning of the components, the examination of glyph difference is observed by taking into consideration the meaning, components, and their relative positions. Characters having different meanings and different actual shapes are not considered to have the same abstract shapes. Characters having the same components yet different in relative positions are generally considered to have different abstract shapes. However, component difference is subjected to examination by experts to see if they have influenced the recognition of the character as a whole with consideration of the character's origin and use. Annex S of ISO/IEC 10646 has defined the examination procedure which is given below:

*“The following features of each ideograph to be compared are examined:*

- a) the number of components,*
- b) the relative position of the components in each complete ideograph,*
- c) the structure of corresponding components.*

*If one or more of the features a) to c) above are different between the ideographs in the comparison, the ideographs are considered to have different abstract shapes and are therefore not unified.*

*If all of the features a) to c) above are the same between the ideographs, the ideographs are considered to have the same abstract shape and are therefore unified.”*

Please also refer to Annex S in ISO/IEC 10646 for examples of characters and components that are considered to have the same abstract shape. The IRG maintains an up-to-date Unification Examples List.


**Character Description Component:** It refers to any symbols that can be used with the Ideograph Description Characters to form a Ideograph Description Sequence. It includes all coded CJK unified ideographs, Kangxi Radicals, CJK Radical Supplements, and coded CJK Compatibility ideographs.

**CJK Unified Ideographs:** It refers to the collection of unified Han characters in ISO/IEC 10646 standard. CJK stands for Chinese, Japanese and Korean. The term CJK Unified Ideographs was adopted in the earlier years of the IRG to reflect the development work of the Han character unification from the three languages at that time. It is obvious today that Han unification covers far beyond the scripts used in China, Japan and Korea. However, the name is consistently being used in the standardization process and is not changed.

**Compatibility Ideographs:** Compatibility ideographs are a kind of compatibility characters defined in Section 18 of ISO/IEC 10646. Below is a direct quote from ISO/IEC 10646-2012:

*“The CJK compatibility ideographs are ideographs that should have been unified with one of the CJK unified ideographs, per the unification rule described in Annex S. However, they are included in this International Standard as separate characters, because, based on various national, cultural, or historical reasons for some specific country and region, some national and regional standards assign separate code points for them.”*

**D-set (discussion set):** D-set is the set of characters that have been reviewed by IRG member bodies with pending issues which need further discussion/evidence for inclusion in the M-set of a working set.

**Ideographic Description Characters (IDC):** The 12 characters defined in ISO/IEC 10646 starting from the code point U+2FF0: 

**Ideographic Description Sequence (IDS):** IDS describes a character using its components and indicating the relative positions of the components. IDCs are considered operators to the components.

IDSs can be expressed by a context free grammar through the Backus Naur Form (BNF). The grammar G has four components:

Let  $G = \{\Sigma, N, P, S\}$ , where

- $\Sigma$ : the set of terminal symbols including all coded radicals and coded ideographs (referred to as CDC, Character Description Components), and the 12 IDCs.
- N: the set of 5 non-terminal symbols  
 $N = \{IDS, IDS1, Binary\_Symbol, Ternary\_Symbol, CDC\}$
- $S = \{IDS\}$ , which is the start symbol of the grammar
- P: a set of rewrite rules

The following is the set of rewriting rules P:

- $IDS ::= \langle Binary\_Symbol \rangle \langle IDS1 \rangle \langle IDS1 \rangle | \langle Ternary\_Symbol \rangle \langle IDS1 \rangle \langle IDS1 \rangle \langle IDS1 \rangle$
- $\langle IDS1 \rangle ::= \langle IDS \rangle | \langle CDC \rangle$
- $\langle CDC \rangle ::= coded\_ideograph | coded\_radical | coded\_component$
- $\langle Binary\_Symbol \rangle ::= \text{[A]} | \text{[B]} | \text{[C]} | \text{[D]} | \text{[E]} | \text{[F]} | \text{[G]} | \text{[H]} | \text{[I]} | \text{[J]} | \text{[K]}$
- $\langle Ternary\_Symbol \rangle ::= \text{[L]} | \text{[M]}$

Note that even though the IDCs are terminal symbols, they are not part of the Character Description Components.

**M-set (main set):** M-set is the set of characters that have been reviewed and accepted by IRG member bodies without pending questions in the current working set.

**New Source:** Any CJK source that is newly submitted by IRG member bodies which is not yet accepted by ISO/IEC 10646, thus is not present in Section 23 of ISO/IEC 10646. Member bodies may first submit their new source to the IRG for acceptance. Once accepted, the characters in that source can be accepted by the IRG for consideration for inclusion in future extensions. The IRG will also submit the source to WG2 for approval and inclusion in Section 23 of ISO/IEC 10646.

**Source:** A reputable published document such as a dictionary, a standardization document, or a well published and widely read or referenced book which the IRG would consider as authoritative such that the characters in this source are considered reliable and stable for consideration of inclusion. A set of ISO/IEC 10646 accepted sources is listed in Section 23 of the ISO/IEC 10646 document.

**Working set:** A working set is the set of characters accepted by the IRG as a collection to work on for extension to ISO/IEC 10646. Characters accepted in a working set are subject to review by IRG member bodies for inclusion in a particular extension.