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

Universal Coded Character Set (UCS)

ISO/IEC JTC 1/SC 2/WG 2/*IRGN2652 WG2N5258 SC2N4911*

(Revision of IRG N1503/N1772/N1823/N1920/N1942/N1975/N2016/N2092/N2153/N2222/N2275/N2310/N2345/N2408/N2427/N2474/N2515/N2582)

2024-03-20

Title:	**IRG Principles and Procedures (IRG PnP) Version 17**
Source:	**IRG Convenor**
Action:	**For review by IRG and WG2**
Distribution:	**IRG Reviewers and Ideographic Experts**
Editor in chief:	**Lu Qin, IRG Convenor**
References:	**Recommendations from IRG#61(IRGN2620, IRGN2612&Feedback, IRGN2639&Feedback, IRGN2646 and IRGN2622), IRG#59(IRGN2570, IRGN2551 and Feedback, IRGN2566 , IRGN2474 and feedback, IRGN2482, IRGN2506, IRGN2500), IRG #56(IRGN2365, IRGN2450,IRGN2463), IRG #53(IRGN2410&IRG2408,IRGN2412), IRG #52 (IRGN2360), IRGN2345 drafts, and feedback from Ken Lunde and HKSAR; IRG #51 (IRGN2329); IRG #49 (IRGN2275); IRG #48 (IRGN2220); IRG #47 (IRGN2180); IRG #45 (IRGN2150); IRG # 44 (IRGN2080), IRGN2016 and IRGN1975; IRG #42 (IRGN1952 and feedback from HKSAR, Japan, ROK and TCA, IRG1920 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); IRG #40 discussions, IRG1823 Draft 3 and feedback from HKSAR, Korea; IRG #39 discussions IRGN1823 Draft2 feedback from HKSAR and Japan, from KIM Kyongsok, IRGN1781 and N1782 Feedback from KIM Kyongsok, IRGN1772 (P&P Version 5), IRGN1646 (P&P Version 4 draft), IRGN1602 (P&P Draft 4) and IRGN1633 (P&P Editorial Report), IRGN1601 (P&P Draft 3 Feedback from HKSAR), IRGN1590 and IRGN1601(P&P V2 and V3 draft and all feedback), IRGN1562 (P&P V3 Draft 1 and Feedback from HKSAR), IRGN1561 (P&P V2 and all feedback), IRGN1559 (P&P V2 Draft and all feedback), IRGN1516 (P&P V1 Feedback from HKSAR), IRGN1489 (P&P V1 Feedback from Taichi Kawabata) IRGN1487 (P&P V1 Feedback from HKSAR), IRGN1465, IRGN1498 and IRGN1503 (P&P V1 drafts)**

Table of Contents

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, IRG will follow the Principles and Procedures of WG2[1] and other higher level directives.

1.1. Scope of IRG Work

IRG works on CJK ideograph-related tasks under the supervision of WG2 (SC2 Resolution M20-07). 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 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 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 IRG scope. However, CJK compatibility characters submitted to WG2 must be reviewed by IRG to avoid potential problems. For the handling of mis-unification and duplicate ideographs, Annexes I and J of WG2 Principles and Procedures 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. IRG can initiate a call for proposals once its current collection is near completion. Any WG2 member body, authoritative organization, international consortium, or individual expert 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. 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) IRG's current workload, IRG may take one of the following actions:

a. Endorse the proposal and submit it to WG2 for approval as an urgently needed repertoire[2].
b. Invite other reviewers 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.

[1] http://std.dkuug.dk/jtc1/sc2/wg2/docs/principles.html
[2] See Annex C for Guidelines on Urgently Needed Characters.

d. Reject the proposal with justifications. A rejected proposal may be revised and re-submitted to 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 rules (Annex S of ISO/IEC 10646 as well as IRG's additional examples in IRG's Working Document Series. See details in Section 2.1.4) 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, IRG only accepts submissions using a print form of glyphs (usually Song style or Ming style). Other styles of writing are generally not accepted unless an approved transcription normalization specification is accepted by IRG.

IRG further spells out its additional requirements for encoding of characters for all submissions in all its extensions (from IRG#53) as follows:

a. **Type of scripts (文種限制):** Encoding request must be for Han character scripts.
b. **Writing style (字體限制)**: The supporting evidence for submitted characters in printed form must be in regular scripts (**楷書**). Other styles cannot be used as evidence for encoding such as clerical style, small seal, etc..
c. **Text use evidence (文本限制)**: characters must be used in script as characters in text. Logos and images used separately from running text are not acceptable.

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 with different glyph shapes that are unrelated in historical derivation (non-cognate characters) are not unified no matter how similar their glyph shapes may be. 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.

"戊"(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).

The non-cognate rule does not apply to characters that have identical glyphs even if the characters are historically unrelated. For example ⿰木几 (wooden table) and ⿰木几 (c-simplified form of 機) shall not be separately coded because they have identical glyphs despite being unrelated in historical derivation.

Characters that are related in historical derivation may also be disunified as long the difference in glyph shape is sufficient for reflecting different semantics. An example is 間 and 閒 which are related in historical derivation but are disunified as long as they have different semantics and typically used in mutually exclusive context in present day use. For the purpose of IRG processes, these characters have been and are still considered applicable for the non-cognate rule even if they are related in historical derivation and technically cognate.

For disunification to a coded character under the non-cognate rul
e, information and supporting evidence provided by a submitter should include the pronunciation of the submitted character as well as the meaning of the submitted character. However, pronunciation alone is not sufficient information for separate encoding.

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

In Annex S of ISO/IEC 10646, 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, IRG must first have a formal discussion for agreement. In case there are worthy examples for recording, IRG will add them to its lists of unification/non-unification examples maintained as IRG working document series (IWDS) on 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 IRG

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

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 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 Fourth Edition (2014-09-01). 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 IRG before they are incorporated by IRG technical editor into the up-to-date IRG list of sources for the current IRG working sets. For sources of miscellaneous nature with reference to individual documents that are too tedious to enumerate by IRG, the submitter should group them together and make a permanent site available for reference.

c. **New CJK Compatibility Ideographs (Vertical extension).** Please be aware that WG2 accepts new CJK Compatibility Ideographs only under very extreme circumstances due to the effects of normalization and the need to add standardized variation sequences to accommodate them. The preferred method of treating unifiable characters whose distinctions are deemed important is by registering them in a new or existing Ideographic Variation Database (IVD) collection as new Ideographic Variation Sequences (IVSes). See Section 2.2.1.g.

   To add CJK Compatibility Ideographs, a submitter needs to supply the following information, which will be reviewed by 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 appropriate column of CJK Compatibility Ideographs Code Table
      d) Source reference (for detailed format, see 2.2.1.d.(5).
      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 the glyph to be printed in the appropriate 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) **Size for the Working Sets of an IRG Collection**: As all collections are defined by submitters according to their own criteria, 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 a collection or a part of an IRG collection, to be reviewed by IRG as a working set normally does not exceed 4,000 ideographs. Based on this principle, submitters should refrain from submitting more than 1,000 characters in each call for an IRG collection. Submitters

may also be asked to divide their submissions into subsets to be processed in different working sets of an IRG collection. The guidelines for forming the working set are 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 maintain high quality at the time of submission. Therefore, any character submission that does not fulfill all the requirements stipulated in 2.1.1 would be rejected. Furthermore, a character submission must be accompanied by evidence to satisfy at least one of the following conditions:

a) **Original Source (證據源限制):** The source of evidence must be considered authoritative by IRG, as validated by past literature and IRG experts. IRG has the right to reject characters from questionable sources.
b) **Multiple Sources (多源證據):** Supply character use evidence from multiple independence sources. IRG has the right to reject characters with evidence of use from only a single source, especially if the source is not considered authoritative by IRG.
c) **Semantics** (字理考證): Supply sufficient evidence on the meaning and phonetics. Supply of other information on its origin and evolution would be very helpful.
d) **Context (上下文信息):** Sufficient context in text to decipher the semantic meaning of the character. IRG has the right to reject characters that do not have sufficient evidence for IRG to decipher its semantics.
e) **Usage (需求限制): The use of characters must be for justifiable public interest. Examples of public use include evidence of: governmental needs; scientific use; digitization projects for public use; and working systems of significance as accepted by IRG. IRG has the right to reject characters that do not have** sufficient evidence for IRG of justifiable public interest.

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 IRG working sets including both M-set and D-set[3].
d) Ideographs mis-unified or over-unified with ideographs in the current standard based on the lists maintained by IRG in its working document series, namely IWDS_MUI and IWDS_NUC.
e) Ideographs from ancient documents that are rare and not in general use, along with variants from tombstone carvings that are not in circulation nor used in printed form, should have an appropriate base character identified through the use of authoritative dictionaries and other references, then be submitted as IVSes to be registered in a new or existing IVD collection. See Section 2.2.1g.
f) Nonce characters are not in general considered suitable for submission to IRG and evidence from the original publications alone of such characters is insufficient. Nonce characters should only be submitted if there is also evidence of significant wider usage.

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:

IRGN*nnnn_mmmm_pppp_*sub*_sss*

where *nnnn* indicates an IRG document number assigned by IRG Convenor, *mmmm* indicates the submitter's abbreviation (as listed in 2.2.1.d.(5), *pppp* indicates the collection year (such as 2015 for the 2015 collection), "sub" is a short hand for submission, and *sss* can be any submitter-designated indicator.

[3] See Section 2.3.4 for the purposes of M-set and D-set.

(4) **Submission of Over-unified or Mis-unified Ideographs**: Submissions of ideographs that are found 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 IRG technical editor and made available for update in 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 disunified characters.[4]

(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 designated WG2 member body abbreviation (G, H, J, K, KP, M, MY, T, UK, UTC or V)[5] or an international consortium designation[6] followed by no more than 9 characters. It should contain only Latin capital letters and Arabic numbers to indicate the source. Numeric values to indicate the position in a specific source should only be followed by a hyphen ("-") (Please note that underscores ("_") must not be used for source reference). 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 IRG source reference abbreviations.

b) Glyph Image should have a unique codepoint in the PUA as TrueType font. The character glyph must have a unique U-code corresponding to its font. The font file should be named using the source reference followed by a submission date in the form of YYMMDD.

c) Kangxi Radical Code(primary radical) from 1 to 214[7] with an additional .0, .1 or .2 to indicate a traditional radical (0), simplified radical (1) or non-Chinese simplified radical (2). The selection of 0, 1, or 2 is based on the radical's glyph shape. The list of radicals with both traditional and simplified glyphs is given in Annex A.a. If the technically correct (aka semantic) radical for an ideograph hampers its discoverability, or is region-dependent, the primary radical shall be assigned as though made by an ideograph expert who is neither a specialist in the history of the Han script nor familiar with ideograph etymology. The technically correct radical can be assigned as a second radical. Both are shown in the code charts, though the primary one serves as the basis for ordering within a CJK Unified Ideographs block.

d) Stroke Count(primary radical) of components other than the radical. Assignment of stroke count should be based on IRG agreed rules (ref. IRGN954AR, IRGN1105, and IRGN2221) regardless of the actual shape with the final decision lies in IRG Chief Editor.

e) First Stroke(primary radical) for components other than the radical, from 1 to 5 as listed in Annex K. Assignment of first stroke should be based on IRG agreed rules (ref. IRGN954AR, IRGN1105, and IRGN2221), regardless of the actual shape with the final decision lies in IRG Chief Editor. If the technically correct (aka semantic) radical for an ideograph hampers its discoverability, or is region-dependent, the primary radical shall be assigned as though made by an ideograph expert who is neither a specialist in the history of the Han script nor familiar with ideograph etymology. The technically correct radical can be assigned as a secondary radical. Both are shown in the code charts, though the primary one serves as the basis for ordering within a CJK Unified Ideographs block.

f) Total stroke count is an integer indicating the total number of strokes of a character including that of its radical. Assignment of stroke count will be based on IRG

[4] It should be noted that the source separation rule described in Annex S of the ISO/IEC 10646 is confined to only those coding standards listed in that section and is not applicable to new IRG submissions.

[5] WG2 Member body abbreviations correspond to the source standard categories in Section 23 of ISO/IEC 10646 except ***MY***.

[6] Currently, the only one IRG has is SAT for the SAT project

[7] The corresponding code range for Kangxi radicals in ISO/IEC 10646 is from U+2F00 to U+2FD5.

agreed rules (ref. IRGN954AR, IRGN1105 and IRGN2221), regardless of the actual shape with the final decision lies in IRG Chief Editor[8].

~~***g) Flag*** is to show whether the ideograph is traditional (0) or simplified (1).~~

h) Ideographic Description Sequence(IDS) (ref. IRGN1183).

i) Similar Ideographs if available (identified by their code points in the standard in the form of U+xxxxx). If there are multiple ideographs, please separate them by comma. Enter "No" if no known variants; leave the column empty if not checked.

j) Pronunciation gives phonetic denotation. Multiple pronunciations can be supplied, separated by comma.

k) Nomalization reference gives the normalization rule(rule number) used if this character is a normalized form.

l) Total number of evidences indicates the number of evidences you supply. IRG encourages multiple evidences to show the meaning and use of this character.

A separate table should be supplied to show information of the evidences. Table should have the following items:

m).1 Character reference should be the source reference number, the same as that in the attribute table.

m).2 File name is for the evidence file. Accepted file types include 'png', `jpg`, `pdf`, `webp` so that they can be uploaded individually to IRG's ORT. Please ensure that all files must be under 1MB and preferably under 400 KB. The source reference is suggested as the file name of the evidence, optionally followed by a "-" (hyphen) and a multi-digit number especially if multiple pieces of evidence are supplied. Example would be GDM-00001-001.jpg.

***m).3* Source** of the evidence which should indicate Source and page number.

m).4 Source URL is optionally supplied if there are online information available.

m).5 Any other information can be included if submitter considers it useful for reviewers.

m) Notes can be used to input any useful information for IRG review.

***n) Additional* optional Information** in text format can be included in additional columns starting from n). If a character was submitted in previous working set, the information should be supplied with respective serial number. Examples of additional information include secondary radicals, secondary stroke count, secondary first stroke. Please add separate columns with appropriate column names after the Notes column.

o) Each submission should include excel file of data description by an assigned IRG document number for each submission. The glyph images should be supplied as a TrueType font. Evidences may be packed into one or more zip files, with the assigned document number, a hyphen, and the string "evidences", (followed by an optional hyphen and padded number for multiple zip files), as the zip file name(s). Examples of file names are given below:

IRGN1000-glyphs.ttf
IRGN1000-evidences-001.zip
IRGN1000-evidences-002.zip

Each submission must strictly follow the formats given above so that the data can be imported into the IRG Online Review Tool(ORT)[9]. 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.[10]

e. **Existing CJK Compatibility Ideographs (Horizontal extension).** To add new source references to existing CJK Compatibility Ideographs, a submitter needs to supply the following

[8] IRG will supply IRG total stroke count data to UTC for the Unihan database. But individual submissions for locale specific total stroke count will not be checked by IRG and IRG takes no responsibility for their correctness.

[9] Note that it was decided in IRG #56 that all future submissions of IRG working sets, all attributes and evidences should be readily imported into the IRG Online Review Tool(ORT).

[10] IRG is in the process of building an IRG Working Set collection system which will be used for automatic checking for the first submission of a new working set. Reviewers can give comments which is for the sole benefit of the submitters. Submitter can also withdraw characters that are problematic characters. Only those characters that pass the automatic checking system will be imported to the ORT for formal IRG review process.

information, which will be reviewed by IRG before submission to WG2 to avoid possible problems.

(1) Table showing the following data for each proposed horizontal extension of CJK Compatibility Ideographs
   a) Code position of the existing UCS CJK Compatibility Ideograph
   b) Glyph(s) of the existing UCS CJK Compatibility Ideograph
   c) Code position of the corresponding UCS CJK Unified Ideograph
   d) Glyph(s) of the corresponding UCS CJK Unified Ideograph
   e) Glyph of the Compatibility Ideograph in the source reference
   f) Glyph of the Compatibility Ideograph to be printed in the appropriate column of CJK Compatibility Ideographs Code Table
   g) New source reference (for detailed format, see 2.2.1.d.(5)
   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 the glyph to be printed in the appropriate 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 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 appropriate 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 the glyph to be printed in the appropriate column of CJK Unified Ideographs Code Table (for detailed format, see 2.2.2.b).

g. **Ideographic Variation Database (IVD)**. For unifiable characters, which may be present in an IRG working set and identified as such, members are strongly encouraged to register them as IVSes in a new or existing IVD collection according to the procedures described in Unicode Technical Standard #37 (UTS #37) (Unicode Ideographic Variation Database). If IRG approves and authorizes the registration of the IVSes in a new or an existing IVD collection, registration fee will not be charged[11].

   It should be noted that separate encoding of variant characters should be discouraged. The use of IVD registration is the more appropriate mechanism for encoding variants of already coded characters. This rule applies to submission of all new IRG working sets after IRG#53.

2.2.2. Required Font to be Submitted

IRG #56 has agreed that submitter to a new working set must supply TrueType Font. Glyph image files are no longer accepted. Font file names should follow the requirement given in 2.2.1.d.(5) b). General font specification can be found under point 5 of A.1. – Submitter's Responsibilities in Annex A of the Principles and Procedures of WG2.

2.2.3. Required Evidence to be Submitted

Supporting Evidence: Evidence of the proposed glyph shape, its usage and context with pronunciation(s), meaning(s), etc. should be supplied to convince IRG that it is actually in use or non-cognate with other similar ideographs. The appearance of a character as a head entry in a dictionary is generally considered evidence of actual use if the dictionary is listed in Annex D or is otherwise accepted by IRG as an authoritative source.

a. 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

[11] IRG does not consider it appropriate to maintain an IRG unified IVD collection. However, IRG is willing to help review individual submissions to UTC.

privacy issues, IRG has suggested some compromised provisions. Details are given in Annex G.3.

Note: To support e-government related initiatives, 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)i), 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[12]. 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 derived simplified ideographs require actual usage evidence. Providing only their corresponding traditional ideographs will not be considered as producing usage evidence. Derived simplified characters from a dictionary as a source should not be used as the sole evidence of actual use unless the dictionary is an IRG-accepted authoritative dictionary.

2.2.4. Required Summary Form to be Submitted

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

2.2.5a. 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,
- IVD, or
- any of the working sets of IRG.

2.2.5b. It is the submitter's responsibility to supply sufficient evidence for its semantics and use. In addition to the requirement of clear images for all evidences, submitters are asked to supply evidences according the to rules suited for its submission:
(1) Complete page evidence showing the submitted character should be provided if possible(clipping image showing only a small incomplete section of the relevant text should be avoided);
(2) If text relating to the submitted character extends over more than one page, images of all relevant pages should be provided;
(3) Complete references for source evidence should be supplied (author, title, publisher, year, page number, etc.) for each evidence image.
(4)The red square used to highlight the submitted character should avoid touching any part of the submitted character;
(5) Characters found in quotations from classical or pre-20th-century texts in a modern typeset edition should also provide an image of an original edition of the text in order to be sure that the character form given in the modern edition is not an error form;
(6) Characters for which the evidence is a pre-modern woodblock printed text should if possible provide at least two evidence images from different sources (may comprise the same or different content) in order to demonstrate that the character is not a one-off error in a single edition.
(7) Characters used in captions and subtitles can be accepted as supporting evidence if agreed by IRG experts[13]. Acceptance of these evidences by IRG shall consider the authoritativeness of

[12] The new IRG PnP in 2.2.1g will consider submission of a variant of an encoded character with similar shape to be a mistake, unless other sufficient justifications for encoding are supplied.

[13] Currently, IRG mainly accepts evidence from printed material if they are accepted as IRG sources. In general, IRG does not accept multimedia material as IRG sources. Note: the acceptance of

the multimedia material, the popularity of the material, cultural influences, and other factors that warrants its acceptance.

In assessing the suitability of a proposed ideograph for encoding, IRG will evaluate the credibility and quality of the submitter's proposal. If IRG finds more than 5% of the submitter's source set are either duplicates of characters in the above mentioned proposals during IRG review process, the whole submission will be removed from the subsequent IRG working drafts for that particular IRG project. Suitable rules in 2.2.5b should also be followed. Evidence that are of poor quality or insufficient evidence for review will be rejected in the first instance and it is also counted as poor submission in the 5% rule. 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 reviewers 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. 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%).

2.2.5c. If a submission has many quality issues validated by analysis of submission data during the review process of the current working set, IRG can impose a capped submission size to the submitter in the next IRG working set. Quality issues include a high rate of unification issues, ill-formed IDSes, evidence quality, rejection rate, and so on.

2.3. Principles on Production of IRG Working Drafts

After IRG accepts submissions based on principles specified in Section 2.2, and follows the guidelines to form working sets of the current collection, the development process of the current working set begins[14]. IRG chief editor and the IRG ORT manager 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 glyph font, IDS, radical(primary), stroke count(primary), first stroke(primary), total stroke count, secondary radical(optional and stroke count and first stroke if secondary radical is provided), and evidence, must have registered IRG document numbers.
b. If any required information is missing, IRG Chief Editor and IRG ORT manager can ask for additional information from the submitter. Without timely supply of such information, the submission may be rejected by the chief editor in producing the working drafts. This is permitted provided the total number of such cases are extremely small. Chief editor and/or IRG ORT manager should report such cases to IRG for quality assurance purpose. Based on the quality report, IRG may apply the 5% rule for rejection.

2.3.2. Principles on Assignment of Serial Numbers

a. IRG chief 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 IRG chief editor.

---

evidence from captions and subtitles may warrant the acceptance of some multimedial materials as IRG source in the future.

[14] In case of multiple working sets in a collection, the review will be conducted for one working set at a time. Once a working set is completed and submitted to WG2, the review of the next working set will start. The process repeats until all working sets are finished in sequence.

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

a. IRG chief editor or his/her designate 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. The IDS checking algorithm would 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 subsets:
   (1) **M-set (main working set)**: for ideographs with proper IDS and found not to be unifiable with current standardized ideographs or previously discussed ideographs with proper IDS. The working procedure is that initially all ideographs in the current working set will be included in this set. For ideographs with questionable attribute data and/or validity of character and/or evidence identified by experts in the review process and the problems of which cannot be resolved during IRG discussions for further information, they can be moved to the D-set (discussion set). Characters that are unifiable with standardized characters or deemed problematic can be withdrawn by submitters.
   (2) **D-set (discussion set**): Ideographs with questionable attribute data and/or validity of character and/or evidence raised by experts in the review process and cannot be resolved during IRG discussions for further information, are moved from M-Set to D-set as decision of IRG as postponed characters for follow up actions to supply further information. Ideographs in the D-set should be withdrawn by submitters if further information cannot be supplied in the next IRG meeting.
b. **Naming of Working Drafts**: The file name should follow the format of "IRGN*nnnnWS####V#.#*[*XXX*]" where *nnnn* is 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 "IRGN2444WS2017V5.2", "IRGN2480WS2017V6.1Draft", etc.
c. **Glyph Images**: An archive of consolidated glyph images will be produced from the font file in the ORT so that glyph change in the font can be compared.
d. **Addition of Characters**: No ideographs should be added to a working set 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 either as a result of IRG recommendation or self initiated with justifications with explicit approval from IRG if the altered glyphs are unifiable with the character glyphs in the original submission. A change of glyph beyond the Annex S(and IRG UCV list) unification criteria is considered to be an addition of a 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. IRG, based on its evaluation, may decide to accept the alteration, reject the proposal or request the withdrawal of such a character by the submitter. If the submitter finds that the glyph of a character is wrong at any working stage, the character will be rejected by IRG and should be withdrawn by the submitter.
f. After consolidation, the IRG Chief Editor can ask IRG editors and contributing experts (collectively referred to as reviewers) 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 IRG instructs reviewers to review the working drafts (different portions may be assigned to different reviewers), reviewers should submit review result according to the agreed schedule, preferably using ORT. They should follow the principles set out below during the review process.

2.4.1. General Principles on Reviews

a. Each reviewer should check the ideographs of the current working set assigned by The IRG Chief Editor for the following issues:
   (1) Correctness of Kangxi radical, Kangxi index, stroke count, first stroke, and IDS.
   (2) Correctness and quality of glyphs, source information (as well as quality of evidence files at the initial stage for quality assurance purpose) if necessary.

(3) Presence of duplicate or unifiable ideographs based on Annex S guidelines as well as examples in IWDS.
(4) Consistency of submitted characters with the submitted evidence and documentary proof.

b. When any data of an ideograph, including IDS, Kangxi radical(primary), stroke count(primary) or first stroke(primary), total strokes, and secondary radical if available(and their stroke count an first stroke) are found to be incorrect, they should be corrected during IRG meeting. Questionable characters with respect to these attributes and to the evidence should be moved from M-set to D-set as their 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, reviewers should ensure that they are not duplicates of or unifiable with any ideograph in the standard, working set(s) submitted to WG2, or in the current working set.
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 ±2 stroke counts of ideographs in the standard, working set(s) submitted to WG2, and current working set. 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 the standard, working set(s) submitted to WG2, and current working set 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. If a secondary radical information is also available, the same process should be done for the secondary radical.
c. **Recording of Review Results**: The checking work should be recorded in the review comments as "Checked against all ideographs in the standard, working set(s) submitted to WG2, and current working set with radical *X* and stroke count of *Y*±2."

2.4.3. Submission of Possibly Unifiable Ideographs

a. **Preparation of Comments**: Reviewers should prepare comments and feedback quoting the assigned serial numbers of the ideographs in question. In the ORT, try to use pulldown menu to select any known issues. If no selected item is available, put comment in remark field. For off-line(with respect to ORT) reviews, comments writing should also be as standardized as possible. The guidelines on comments are described in Section 4 of this document. Comment files should be tabulated in CSV text format, Microsoft Excel, or Word file format. All off-line review comment files should use the pre-assigned IRG document number in the current version with Source of reviewer appended to the file name.
b. **Additional Evidence and Arguments**: For each proposed ideograph in the D-set that has been questioned for possible unification, the submitter should prepare response 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. When submitters disagree with a suggested unification, it is insufficient to simply point out that there is no unifiable component variations (UCV) examples. The UCV list and other working documents in the IWDS, which is updated as a result of IRG Working Set reviews, should also be used. As documents in IWDS will be updated during IRG reviews during IRG meeting, "No case in UCV list" is no longer a sufficient reason for dis-unification. IRG requests submitters to also provide unification and dis-unification examples from published versions of the standard as well as those accepted for publication. The additional information will help IRG to determine whether the ideographs in question should be dis-unified or unified (which may result in additional UCVs or other related documents).
c. **Submission Deadlines**: Each reviewer should submit review comments at least two months before the next IRG meeting. The IRG Chief Editor will consolidate them and register the results as IRG documents one month before the next IRG meeting.
d. **Written Responses from Submitters**: Submitters should examine the consolidated comments on their respective characters and send The IRG Chief Editor a written document containing their responses to the comments together with additional evidence at least one week before the next IRG meeting either using ORT or document using the current working document appended with "ResponseXXX" where XXX is the source submitter designation.
e. **Rejection**: Questioned ideographs with no counter arguments supplied to IRG meeting will either be moved to D-set or withdrawn.

f. **Revised font**: In case of a glyph mismatched to evidence or mismatched to normalization for consistency, the submitter needs to provide a revised font to IRG for review for acceptance/rejection. The revision is only accepted if it is a unifiable change. Or its change has no impact to other characters in the working set(including corresponding revised attribute data).

2.5. Principles on Discussions at IRG Meetings

2.5.1. Record-based Discussion

For efficient and smooth work, all discussion items and evidence must be presented as registered IRG documents or registered in ORT before the commencement of an IRG meeting. Items or evidence that are not contained in a registered IRG document (or on the ORT) will not be discussed or treated as evidence during IRG meetings.

2.5.2. Discussion Procedure

Discussions will be based on the review comments of the current working set. This includes two parts. The first part is based on comments and response for questionable characters in M-Set. The second part is on response/feedback of characters in D-Set.

(1) **For unification issues**. 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(and IWDS). When IRG has 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 now-unified ideograph and add a new source reference to the existing standardized or working set ideograph. This is particularly important if the existing standardized or working set ideograph to which it is unified does not yet have an assigned source reference that corresponds to the submitted and now-unified ideograph.
b. Register the now-unified ideograph in a new or an existing IVD collection as a new IVS, particularly if the existing standardized or working set ideograph to which it is unified already has an assigned source reference that corresponds to the submitted and now-unified ideograph.
c. To avoid misunification of ideographs, IRG may specify etymological constraints to the application of a particular UCV rule, i.e. an etymological relationship must be proven between a proposed character and a coded or another Working Set character for the rules to be applicable. An etymological constraint will only be specified for a particular UCV rule when one or more of the suggested unifiable forms may typically be etymologically related to another radical or component, thus the rule is at high risk of causing misunification of unrelated character. The unification shall only apply when there is sufficient evidence to prove that the two characters in discussion are etymologically related, i.e. the proof of burden lies in the reviewer instead of the submitter. A proposed character will also not be postponed unless there is reasonable doubt that the character is etymologically related with another coded or Working Set character.
d. Even if an etymological relationship can be proven to exist between two characters, the non-cognate rule still applies, i.e. as long as there is sufficient evidence to show that the two characters are used with mutually exclusive semantics in a certain language or region, these characters will not be unified. Because shape analysis alone cannot indicate 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.

   IRG at its discretion can allow a character discussion be labeled as "pending" with specified time for response during meeting if submitter considers the additional information can be supplied quickly. This allows for some off-line discussion, and make the progress of the discussion more efficiently. If a response is not prepared at the specified time, the pending character will remain in the D-set or withdrawn by the submitter.

   Discussions on evidence or items raised after the commencement of an IRG meeting may be postponed to the next IRG meeting if any submitter (or reviewer) requests longer time to examine such evidence or items (Remain in the D-set).

(2) **For radical and related attributes.** 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 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 for 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.

IRG recognizes that radical assignments by different submitters may be different as radical use is locale dependent. As each character must be assigned a unique radical for sequencing in code chart production, IRG will determine the radical used for IRG working set sequence based on Kangxi custom or the most appropriate one determined at IRG if multiple assignments are supplied. Submitters are also given the option to provide secondary radicals(and its corresponding stroke count and first stroke). If there is no unifiable characters from multiple submissions, IRG only make suggestion for the appropriate radical. In this case, the submitter decides the radical to be used.

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 in ORT for each ideograph reviewed for reference and checking. Document should also be produced so that it can be made available to all reviewers.

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 the editorial group has to deal with individual characters during an IRG meeting, submitters and reviewers 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 reviewers is needed before the set can be prepared for WG2 submission.

2.6.2. Preparation for WG2 Submission

After the approval by IRG, The IRG Chief Editor with the help of the ISO/IEC 10646 project 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 all the attribute data and prepare for the WG2 submission summary form.
d. Make available the TrueType font by each submitter with mapping to assigned provisional UCS code positions provided by the chief editor with verification from submtters(fonts have to be available in accordance with the requirement stated in point 5 of A.1. – Submitter's Responsibilities in Annex A of WG2 Principles and Procedures). Each submitter should prepare and submit its own font to the ISO/IEC 10646 project editor for best font quality.
e. Prepare a list of source references.
f. Produce a packed Multi-column Ideograph Chart using the TrueType fonts.

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 submitter representing government, an organization or an international consortium requests a new project for CJK Unified Ideograph extension or IRG's current collection is near completion, IRG may decide at an IRG meeting to call for submission of new ideographs to form a new collection. IRG will specify the deadline for submission.
b. IRG will give a name to the new collection using the year of the call (in four-digit format)[15].
c. Each submitter with proposed ideographs must submit the ideographs before the specified deadline with the required data described in Section 2 of this document.
d. Submitters must ensure that the ideographs are submitted 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 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.
e. An initial review of all submissions for a new collection will be done during an IRG meeting to estimate the size of the collection. Each submitter is allowed to submit no more than 1,000 characters. As the normal work set size is set at 4,000, IRG will use the guidelines given in Annex L to estimate the number of working sets for the collection in case the total number of characters is much larger than 4,000. If multiple working sets are deemed necessary, IRG meeting will determine how the collection will be split so that IRG can work on one working set at a time. The decision on the number of working sets may be revised after the review of the first consolidation in an IRG meeting as stated in 3.4.f.

3.2. Consolidation and Grouping of Submitted Ideographs

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

a. IRG Chief Editor will sort and assign *serial numbers* to submitted ideographs as described in Section 2.3.2.
b. After serial numbers have been assigned, submitted ideographs must undergo IDS checking to detect any duplication and unification. Based on 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 contributing experts and made available on the official IRG website so that any other experts can have access to them. IRG Chief Editor may assign ORT editors and other contributing reviewers to check M-set and D-set ideographs for either the entire working set or certain portions of it depending on a reasonable estimation of the workload.
d. Once the consolidation is done, the IRG Chief Editor will sent the data file to the Manager of the IRG ORT. The consolidated data will be uploaded to ORT to be ready for online review by ORT editors. It needs to be emphasized that submissions must strictly follow submission rules to ensure data and evidences can be uploaded into ORT. ORT is a working platform, not meant as a learning platform. Only designated ORT editors who are either IRG editors or experts who are familiar with IRG review process and have made active contributions are granted access to ORT at the permission of the IRG Convenor. Reviews can still be conducted in the traditional paper based review and comments can also be submitted in the manner described in Section 3.3 to Section 3.7. Online reviews follow the same review process described in Section 3.3 to Section 3.7 except the feedbacks are written online and the ORT Manager will export the review comments to the IRG Chief Editor for consolidation if written reviews are submitted.

3.3. First Checking Stage

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

a. Each reviewer 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.

[15] The first IRG collection using this naming convention is IRG Collection 2015.

b. Off-line reviewers must submit their comments in registered IRG documents to IRG Chief Editor at least two months before the next IRG meeting or according to IRG approved working schedule.
c. IRG Chief Editor will consolidate the comments with the help of the ORT manager and produce a registered IRG document for circulation and discussion at least one month before the next IRG meeting or according to IRG approved working schedule.
d. Submitters must submit their response to questionable characters either in ORT or in an IRG registered document one week before the next IRG meeting. All experts are encouraged to prepare and submit supplementary documents (with IRG document numbers) so that they can be discussed at the next IRG meeting. For offline comments, the ORT manager can opt to import them into ORT to facilitate discussion in the next IRG.

3.4. First Discussion and Conclusion Stage

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

a. All participating experts should review the comments which are officially submitted before the meeting either in the ORT or with assigned IRG document numbers. The editorial group must reach conclusion for each commented ideograph with written records. Guidelines for typical conclusions are provided in Section 4.
b. All the conclusions must be agreed and endorsed by IRG plenary in its recommendations. As a result of the recommendations, some ideographs may be withdrawn or moved between M-set and D-Set.
c. IRG Chief Editor with the help of ORT manager will create a new version of the M-set and D-set and the list of withdrawn characters one month after IRG meeting.
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 ideographs, 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. IRG should also instruct the IWDS editor/co-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.
f. If decision on the number of working sets is not made in 3.1.e when the collection is first submitted, IRG will review the size of the collection and determine the number of working sets at this stage. The working sets will go through the review process one at a time from 3.2 to 3.8 until all working sets are completed.

3.5. Subsequent Checking Stage

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

a. Reviewers must check the newly created M-set and D-set for correctness and duplication.
b. Reviewers should submit their comments either in ORT or in registered IRG documents to IRG Chief Editor at least two months before the next IRG meeting or according to IRG approved working schedule.
c. IRG Chief 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 IRG approved working schedule.
d. Submitters must submit response to questionable characters one week before the next IRG meeting. No response to questioned characters will be postponed in the coming IRG meeting. All 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. Participating experts in the editorial group 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 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. IRG Chief Editor will create a new version of M-set and D-set one month after IRG meeting, and produce a registered IRG document. The same update will be made in ORT by the ORT manager.
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 ideographs, 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 final stage, which is between IRG meetings, is a decision of a previous IRG meeting. This stage involves the following tasks:

a. All reviewers are requested to check M-set intensively based on comments and conclusions made at all previous stages. At the final checking stage, all characters in the D-set of previous IRG meeting will be considered withdrawn characters. No ideographs are allowed to be moved from D-Set to M-Set although ideographs in the M-set can still be moved to D-set if problems are found.
b. Reviewers must submit their comments in ORT or in registered IRG documents to IRG Chief Editor at least two months before the next IRG meeting.
c. IRG Chief 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 reviewers can have time to review them before the next IRG meeting. ORT will also be updated by the ORT manager.
d. Submitters who have questionable ideographs in the consolidated comments should submit their written response one week 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. Participating experts 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 withdrawn.
c. With the approval of the majority of IRG reviewers, M-set is considered frozen as the new ideograph extension set to be submitted to WG2. IRG Chief Editor with the help of the ISO/IEC 10646 project editor will prepare the submission in accordance with Section 2.6 of this document.

Once M-set is frozen as completed for submission to WG2, records of characters in the D-set will no longer be maintained by IRG. Characters remained in the D-set can be re-submitted in future extensions if pending problems are solved. Reference to the serial number of the previous working set should be supplied in the new working set.

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[16].

All comments must be accompanied with date (in YY-MM-DD format) and the designated IRG abbreviation (G, H, J, K, KP, M, MY, SAT, T, UK, UTC, 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 current working set 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.

[16] For editors and experts using ORT to review data, most of the comments are standardized as selections and thus this can speed up the review process and easier to consolidate.

201C3		Wrong Glyph	T glyph seems wrong: 𠂉 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 *KX Dictionary* 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	● The wrong glyph is corrected, or the missing glyph supplied if evidence is provided. The ideograph can also be moved to D-set for manual checking if insufficient information is provided by the submitter.
Wrong Kangxi radical / stroke count / first stroke	● Data will be corrected if agreement is made. Otherwise, the ideograph will be moved to D-set for further manual checking.
Wrong IDS	● IDS will be corrected and the character will be moved to D-set for checking by the IDS checker. ● Move to D-set (in case IDS cannot be corrected).
May be unifiable with U+xxxxx (standardized ideograph)	● Unified with U+xxxxx and the submitter will request a new source reference to U+xxxxx. ● Unified with U+xxxxx and the submitter will request that this character be treated as a Compatibility Ideograph. ● Unified to U+xxxxx and this entry will be removed. (May consider to register it as IVS.) ● Not unifiable.
May be unifiable with xxxxx (M-set ideograph)	● Unified with xxxxx and this source reference will be attached to xxxxx. ● Unified with xxxxx and the submitter may consider registering it as a Compatibility Ideograph Character or IVS. ● Not unifiable.

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 reviewers or whose unification with other ideographs in the standard, working set(s) submitted to WG2 or current working set has been proposed. For those ideographs that cannot be machine-checked by IDS matching, at least two non-submitter reviewers must check them manually to ensure that they are not unifiable with any ideographs in the standard, working set(s) submitted to WG2, or current working set. 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
● Incomplete IDS IDS with extra character ● Component is not an ideograph	● IDS will be corrected and the character will be moved to M-set when next IDS-checking is done. ● Proper IDS cannot be generated and manual

	checking is needed.
Possible Comment by a Reviewer	**Possible Conclusion**
● Wrong Kangxi radical ● Wrong stroke count ● Wrong first stroke	● Data will be corrected. ● Proposal to correct data is not accepted, as it is an ambiguous case and IRG agrees that the original data are more appropriate.
Wrong IDS	● IDS will be corrected and checked by the IDS checker again. ● Correct IDS cannot be generated and manual checking is needed.
May be unifiable with U+xxxxx (standardized ideograph)	● Unified with U+xxxxx and a new source will be added to U+xxxxx. The new candidate entry should be deleted. ● Not unifiable, as shown by the evidence *IRGN xxxx*. Move to M-set.
May be unifiable with *xxxxx* (M-set or D-set ideograph)	● 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. ● 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. ● Not unifiable, as shown by the evidence *IRGN xxxx*. Move to M-set.
Checked against all ideographs in the standard, working set(s) submitted to WG2 and current working set with radical *X* and stroke count of *Y*±2 for characters that cannot be described by IDS for automatic checking.	● Move to M-set, as two non-submitter reviewers (*XX* and *YY*) confirmed that this ideograph is not unifiable with any existing ideographs in the standard, working set(s) submitted to WG2, or current working set. ● Checking against ideographs with radical *X* may not be enough. This ideograph will also be checked against ideographs with radical *Z*.
Evidence does not match glyph	● New evidence must be supplied or the character will be moved to D-set.
Evidence not clear	● Character will be moved to D-set unless a clear evidence is supplied.

5. IRG Website

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, recommendations, document register, documents and standing documents are made available on this site. Hyperlinks to WG2 websites are provided for reviewers' 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 be in either WinZip format with .zip extension or RAR format with .rar extension.

6. IRG Document Registration

All documents to be formally discussed by IRG must be registered with IRG document numbers assigned by IRG Convenor and contain the submission date, title, name of the submitter or author, purpose (or summary), and the "IRG 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 IRG must be given a document number. The number is to be assigned by IRG Convenor. The submitter should first contact IRG Convenor for a document number with a document title. Once the document number is assigned, the information will be posted on 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 IRG Convenor. 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. Note: Feedback to and comments on a given document will not be assigned a new document number. Rather, they will use the same document number with extensions to facilitate tracking for the same topic.

c. **File size:** Documents larger than 4MB must be split into multiple files for easy uploading, downloading and searching. The compressed files can be in either zip format with .zip extension or RAR format with .rar extension. Files that are much larger than 4MB, and not easy for splitting, IRG has both Google drive and Weiyun to store them. However, the submitters must inform IRG Convenor at the time of file submission. A short cover page to describe the content of the large file should also be submitted. IRG Convenor will upload the cover page to the IRG server and the actual large files to Google Drive and Weiyun (for access in China) using accounts managed by the IRG convenor. Links to these large files will be made available on the IRG website.

d. **Posting of Documents**: Properly submitted documents are then posted by IRG Convenor on IRG website as official documents and the submitters will be notified by IRG Convenor 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. In case of a large file, the submitter must first provide a method (e.g., web download, ftp) for IRG Convenor to obtain the file. IRG Convenor will then post it on IRG large-file posting sites and provide the link(s) on IRG website for download.

e. **Disqualified Documents**: Documents with certain basic information missing such as the submitter's name, title, purpose or files that are not in the appropriate size may be rejected by IRG Convenor for posting. All other documents that fail to comply with the above registration process and the preliminary review by IRG Convenor for basic information will not be treated as IRG documents. As such, issues contained in such documents will not be discussed by IRG formally.

f. **Document format**: Documents submitted by submitters should use the most commonly-used document format for easy reading by members on all platforms. Static documents should use PDF. Data files intended for consolidation, revision and processing can use other appropriate document formats depending on the nature of data, such as Microsoft Excel, CSV, plain text, or PNG.

6.2. Contact for IRG Document Registration

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

Professor Qin Lu
Retired from Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong
Tel. (852) 9684 0623
Email: csluqin@comp.polyu.edu.hk

Annex A: Sorting Algorithm of Ideographs

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, 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: Ideographs with the simplified radicals listed below must be placed after ideographs with the corresponding traditional radicals.

Traditional Radicals		Simplified Radicals		Non-Chinese Simplified Radicals	
R090.0	月	R090.1	刂		
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	R182.2	⿻二
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	黾		
				R208.2	鼻
R210.0	齊	R210.1	齐	R210.2	齐
R211.0	齒	R211.1	齿	R211.2	齿
R212.0	龍	R212.1	龙	R212.2	竜
R213.0	龜	R213.1	龟	R213.2	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 submitter should consult IRGN1183 on IDS. In addition to the Character Description Components (CDC) defined in IRGN1183, all CJK Unified Ideographs accepted by ISO/IEC 10646 in its amendments are also qualified as CDC in constructing IDS.

The use of “overlapping” Ideographic Description Characters (IDC) 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 IRG should support the following features:

1. Handling different split points.
   (e.g. ⿰亻顷 and ⿰化頁 should be matched.)
2. Handling different split levels.
   (e.g. ⿰亻悉 and ⿰亻⿱釆心 should be matched.)
3. Matching different glyphs of the same abstract shape.
   (e.g. ⿰礻申 and ⿰示申 should be matched.)
4. Matching similar glyphs.
   (e.g. ⿰忄生 and ⿰小生 should be matched.)
5. Matching IDS with different orderings of overlapping IDC.
   (e.g. ⿻三丨 and ⿻丨三 should be matched.)
6. Matching unifiable IDC patterns.
   (e.g. ⿰麥离 and ⿺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 WG 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, IRG may submit the ideographs, independent of the current IRG working set 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 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. Evidence of current use is not in and of itself evidence of urgent need. The type of use also needs to be taken into account. For example, requirements of government, industry, science, or scholarship will generally be taken as evidence of urgent need.
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 set 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 set 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 set. 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 standardized CJK ideographs, but also the working set(s) submitted to WG2 for consideration and IRG working set for any unifiable characters against its submission. If a submitter fails to do the above, the submission will not be approved by 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 IRG for correctness, duplication and unification against the latest published ISO/IEC 10646 as well as the current IRG working set. When an urgently needed ideograph is found to be identical or unifiable with any ideograph in the current IRG working set, the latter must be noted and removed from the current IRG working set.

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

IRG tracks the sources of the CJK Unified Ideographs. The past practice was that the character sources were tracked based on submissions by submitters. Thus every submitter was assigned a WG2 member body abbreviation as defined in ISO/IEC 10646. In recognizing that submitters might be an international consortium not affiliated with any WG2 member body, IRG decided at IRG Meeting No. 39 to add a new abbreviation "Z" for this type of submissions. It should be noted that Z source may include submissions from different projects and additional letters may be used to indicate each repertoire.

As described in Section 2.2.1.d.(5)a): "The source reference should begin with a WG2 member body abbreviation (G, H, J, K, KP, M, MY, T, UK, UTC or V) or an IRG recognized submitter designation [17]followed by no more than 9 characters. It should contain only Latin capital letters and Arabic numbers to indicate the source. Numeric values to indicate the position in a specific source should be followed by a hyphen ("-")."

Note: IRG internal documents allowed underscore ("_") to be used in source references in submissions from China and SAT until IRG Meeting No. 46. To be consistent with the ISO/IEC 10646 published standard[18], the use of underscore symbol will be disallowed from any new working set submissions starting from IRG Meeting No. 47. Sources accepted in the past using underscores are thus corrected in the list below. The original reference formats are retained for cross reference to older documents only.

D.1. WG2 Member body abbreviations:

G	China
H	Hong Kong Special Administrative Region, China
J	Japan
K	Republic of Korea[19]
KP	Democratic People's Republic of Korea
M	Macao Special Administrative Region, China
MY	Malaysia (added in Nov. 2008 at IRG Meeting No. 31)
SAT	SAT Int. Project
T	Taipei Computer Association
UK	United Kingdom
UTC	Unicode Consortium
V	Vietnam

Note: all WG2 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. Hanzi G sources

The format of the accepted sources is now consistent with the published standard. Code used in pointed brackets are original source references no longer supported after IRG Meeting No. 47.

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
G4K	Siku Quanshu (四庫全書) < G_4K >
GBK	Chinese Encyclopedia (中國大百科全書) < G_BK >

[17] Currently, only one designation SAT for the SAT International Project

[18] Pages 29-33 in Section 23 of ISO/IEC 10646 4th edition, 2014-09-01.

[19] A new KU source will be used from IRG#53 to indicate ghost K source based on disposition comments of UCS ed. 5.

GCH Ci Hai (辞海) < G_CH >
GCY Ci Yuan (辭源) < G_CY >
GCYY Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字) < G_CYY >
GDM Place name characters from the Public Order Administration, Ministry of Public Security, People's Republic of China
GDZ Geographic Publishing House Ideographs (地质出版社用字) < G_GDZ>
GFC Modern Chinese Standard Dictionary (现代汉语规范词典第二版。主编：李行健。北京：外语教学与研究出版社) 2010, ISBN：978-7-5600-9518-9
GFZ Founder Press System (方正排版系统) < G_FZ >
GGFZ Tongyong Guifan Hanzi Zidian (通用规范汉字字典)
GGH Gudai Hanyu Cidian (古代汉语词典) < G_GH >
GHC Hanyu Dacidian (漢語大詞典) < G_HC >
GHZ Hanyu Dazidian ideographs (漢語大字典) < G_HZ >[20]
GIDC ID system of the Ministry of Public Security of China, 2009 < G_IDC >
GJZ Commercial Press Ideographs (商务印书馆用字) < G_GJZ >
GKX GKX Kangxi Dictionary ideographs (康熙字典) 9th edition (1958) including the addendum (康熙字典) 補遺 < G_GKX >
GLGYJ Zhuang Liao Songs Research,(壮族嘹歌研究》2008年广西民族出版社, ISBN 78-7-5363-5069-4
GOCD Oxford English-Chinese Chinese-English Dictionary (牛津英汉汉英词典。主编：Julie Kleeman, 于海江。牛津：牛津大学出版社。2010年。ISBN：978-0-19-920761-9)
GPGLG Zhuang Folk Song Culture Series - Pingguo County Liao Songs (壮族民歌文化丛书•平果嘹歌) 2004-2006, ISBN 7-5363-[4820-7 | 5012-0 | 5013-9 |5014-7 | 5015-5]
GRM People's Daily Ideographs (人民日报用字) < G_GRM >
GXC Xiandai Hanyu Cidian (现代汉语词典) < G_GXC >
GXH Xinhua Zidian (新华字典) < G_XH >
GXHZ Xinhua Big Dictionary（新华大字典：彩色本修订本。《新华字典》编委会编。北京：商务印书馆国际有限公司）2011, ISBN：978-7-80103-718-3
GWZ Hanyu Dacidian Publishing House Ideographs (漢語大詞典出版社用字) < G_WZ >
GXM Characters for use in personal names in China
GZ Ancient Zhuang Character Dictionary, (古壮字字典) 1989, ISBN 7-5363-0614-8
GZA-1 A Vibrant and Unbroken Transmission - Filial Piety and Zhuang Funeral Songs生生不息的传承•孝与壮族行孝歌之研究
GZA-2 Annotated Long Zhuang Morality Songs 壮族伦理道德长诗传扬歌译注
GZA-3 Compendium of Old Zhuang Folksong Texts - Wooing Songs vol. 1 – Liao Songs 壮族民歌古籍集成•情歌（一）嘹歌
GZA-4 Compendium of Old Zhuang Folksong Texts - Wooing Songs vol. 2 – Fwen Nganx 壮族民歌古籍集成•情歌（二）欢榨
GZA-6 Zhuang Proverbs from China 中国壮族谚语
GZA-7 Ancient Remembrance - Zhuang Creation Myth Songs 远古的追忆•壮族创世神话古歌研究
GZFY Hanyu Fangyan Dacidian (汉语方言大词典) < G_ZFY >
GZH Zhonghua Zihai (中华字海) < G_ZH >
GZJW Yinzhou Jinwen Jicheng Yinde (殷周金文集成引得) < G_ZJW >
GZYS Chinese Ancient Ethnic Characters Research (中国民族古文字研究), 1984
SJT11239 SJ/T 11239-2001 Information Technology - Chinese ideograms coded character set for information interchange - The 8th supplementary set (信息技术 信息交换用汉字编码字符集 第八辅助集)

D.3. Hanzi H sources

H Hong Kong Supplementary Character Set (HKSCS)

Note: The ISO/IEC 10646 published standard also included Big-5 characters as H source characters as HKSCS is considered a supplement to Big-5. The respective Big-5 characters under their categories are listed here for reference:

HB0 Big-5: Computer Chinese Glyph and Character Code Mapping Table, Technical Report C-26, 電腦用中文字型與字碼對照表，技術通報 C-26, 1984, Symbols

[20] As this is a modern dictionary which may be extended/revised from time to time, IRG accepts its 1st and 2nd versions. Future extensions will be subjected to review before acceptance as appropriate IRG sources.

HB1 Big-5, Level 1
HB2 Big-5, Level 2

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 addendum from JIS X 0213:2000 level-3
J13A JIS X 0213:2004 level-3 addendum from JIS X 0213:2000 level-3 replacing J1 characters
J13 JIS X 0213:2004 level-3 characters replacing J1 characters
JA3 JIS X 0213:2004 level-3 characters replacing JA characters
J4 JIS X 0213:2004 level-4
J14 JIS X 0213:2004 level-4 characters replacing J1 characters
JA4 JIS X 0213:2004 level-4 characters replacing JA characters
JA Unified Japanese IT Vendors Contemporary Ideographs, 1993
JARIB Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007
JH Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009
JMJ Character Information Development and Maintenance Project for e-Government "Moji-Joho-Kiban Project" (文字情報基盤整備事業), 2010-
JK Japanese KOKUJI Collection

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 X1027-4:2011)
KC Korean History On-Line (한국 역사 정보 통합 시스템)

D.7. Hanja KP sources

KP0 KPS 9566-97
KP1 KPS 10721-2000 and KPS 10721: 2003

D.8. ChuNom V sources

V0 TCVN 5773:1993
V1 TCVN 6056:1995
V2 VHN 01:1998
V3 VHN 02: 1998
V4 Kho Chữ Hán Nôm Mã Hoá (Hán Nôm Coded Character Repertoire), Hà Nội, 2007
VN Vietnamese horizontal extensions

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 Information System Character Set （澳門資訊系統字集）
MA Same as Hong Kong Supplementary Character Set -2008
MB1 Big-5, level 1
MB2 Big-5, level 2
MC Macao Supplementary Character Set (MSCS) – 2020
MD, MDH MSCS -2020, horizontal extension

D.11. Unicode sources

UTC Unicode Standard Annex #45 (Unicode Version 10.0.0), June 14, 2017 (Used to be called Unicode Technical Report #45)

D.12. U sources (renamed from Z sources since IRG Meeting No. 45)

This is a new source created since IRG Meeting No. 39 to accommodate submissions from international groups working on CJK ideographs. Originally, it was named the Z source. As at the end of IRG Meeting No. 44, only one designation of SAT is used as Z_SAT for the international project of SAT for Daizōkyō. However, due to a request from WG2, this designation is renamed the U sources from IRG Meeting No. 45.

USAT Taishō Shinshū Daizōkyō (大正新脩大藏經), 1924-1934 <U_SAT>

Annex E: Maintenance Procedure of IRG Working Document Series

E.1 Introduction

IRG Working Document 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 Document 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 Unifiable Component Variations (UCV) of Ideographs (File name: IWDS_UCV.pdf)
Series 3: List of Non-Unifiable Components (NUC) of Ideographs and Overly-unified Ideographs (File name: IWDS_NUC.pdf)
Series 4: List of Possibly Mis-Unified Ideographs (MUI) (File name: IWDS_MUI.pdf).
Series 5: Locale based Glyph Normalizations

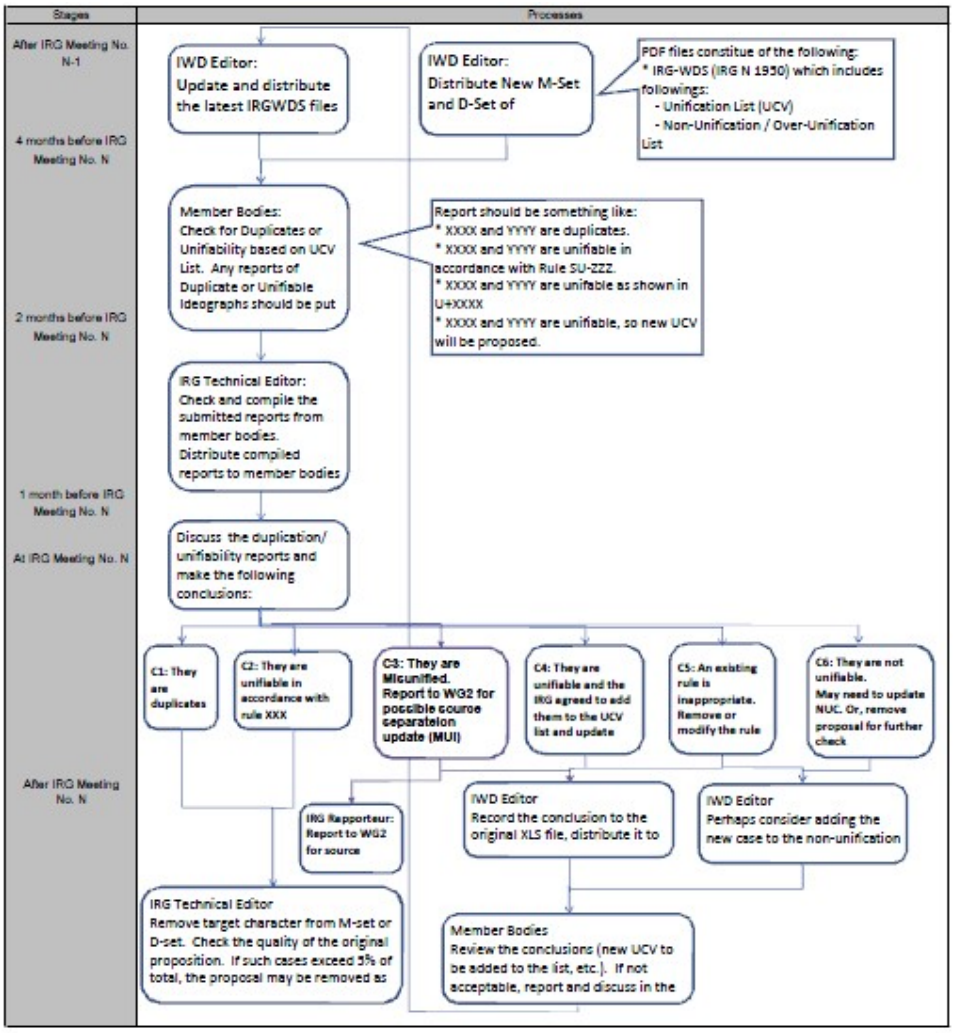
E.3. Maintenance Procedure

The maintenance procedure describes how entries in the IWDS are added, removed, or changed. IRG has an appointed IWDS editor (currently Mr. Taichi Kawabata), and IWDS co-editor(Mr. Yi Bai, appointed since IRG #56), who are 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 reviewers in the past IRG meeting (Meeting No. N-1). During the review work before the next IRG meeting (Meeting No. N), if reviewers 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 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. 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
Submitters are reminded to:
1.Fill in all the sections below.
2. Read the Principles and Procedures Document (P & P) available at for guidelines and details before filling in this form.
3. Use the latest Form from
https://appsrv.cse.cuhk.edu.hk/~irg/irg/irg56/IRGN2424SubmissionForm.xlsx
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. IRG Collection 2021
2. **Title:**
3. Submitter's Region/Country Name:
4. Submitter Type (National Body/Individual Contribution):
5. Submission Date:
6. Requested Ideograph Type (Unified or Compatibility Ideographs)
   If Compatibility, the submitter is strongly encouraged to instead register them as IVS in a new or an existing IVD collection(See UTS #37) with the IRG's approval (Registration fee will not be charged if authorized by the IRG.).
7. Proposal Type (Normal Proposal 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 is in TrueType?
   Are all the proposed glyphs put into PUA area?
   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 Principles

1. Has the submitter read ISO/IEC 10646 Annex S and does the submitter understand the unification principles?
2. Has the submitter read the "Unifiable Component Variations" (contact the IRG technical editor through the IRG Convenor for the latest version) and does the submitter understand the unifiable variation examples?
3. Has the submitter read the IRG PnP document and does the submitter 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 submitter 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. 10646:2012)
5. Has the submitter 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 submitter checked?
6. Has the submitter 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 submitter checked?
7. Has the submitter 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 Convenor for the newest list)
   If yes, which document(s) has the submitter checked?
8. Has the submitter 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 IRG PnP document)
9. Has the submitter checked whether the proposed ideographs have any ***similar ideographs*** in the current standardized or working sets mentioned above?
10. Has the submitter 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 including the Kangxi radical code, stroke count, and first stroke(primary)?
12. Do the proposed ideographs contain secondary radical code and their stroke count and first stroke are also provided?
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?
16. Do all the proposed ideographs contain the total stroke count (kTotalStrokes)[21]?

[21] The IRG understands that kTotalStrokes can be ambiguous and subject to different interpretations. The IRG takes no responsibility to check the correctness of the submitted attribute data.

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 UK for consideration in IRG Collection 2021.

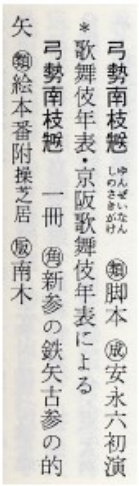
A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Seq. No.	a) Source reference	b) PUA Code of TTF	c) KangXi Radical Code(Primary)	d) Stroke Count (Primary)	e) First Stroke(Primary)	f)Secondary KX Radical Code(only if needed)	f)a. Secondary Stroke Count (only if needed)	f)b. Secondary First Stroke(only needed)	g) Total stroke count	i) IDS (Ideographic Description Sequence)	j) Similar/ Variants	k) Ref. to Evidence doc	l) Optional info	m1) Previous IRG WS	m2) Sequence No of Previous IRG WS	n) Evidence type(if applicable)
001	KC-00010	EE00	R005.0	3	2				4	㇐小乙	N	Seoul大學校 奎章閣 顯隆園園所都監儀軌 2冊				
002	KC-00009	EE01	R005.0	3	3				4	㇐久乙	N	韓國學中央研究院 大王大妃殿王大妃殿尊 崇都監都廳儀軌, 堂上手決内依				
003	KC-00012	EE02	R005.0	4	3				5	㇐介乙	N	Seoul大學校 奎章閣 『樂器造成廳儀軌』 1冊 手本秩				

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 (IRGN1225 Part2).

JMK66114

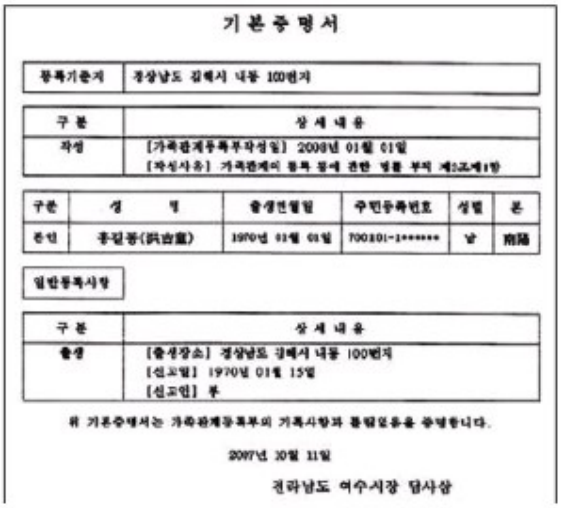
『補訂版国書総目録』(1969年4月30日第1刷発行, 2002年7月5日補訂版第4刷発行, 岩波書店)
第7巻870ページ4段



G.3. Handling of Data with Privacy Concerns

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, IRG suggests submitters to provide evidence in such a way that it will 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 blocked.

As different departments/organizations may have different types of documents, IRG suggests that, for each type of document, a submitter should provide a sample document with private information deleted. A good example is the original Basic Certificate of Family Relation Register in Korea as shown in Fig. G1. The evidence can be submitted as partial data in the form shown in Fig. G2.



기 본 증 명 서

등록기준지	경상남도 김해시 내동 100번지

구 분	상 세 내 용
작성	[가족관계등록부작성일] 2008년 01월 01일 [작성사유] 가족관계의 등록 등에 관한 법률 부칙 제2조제1항

구분	성 명	출생연월일	주민등록번호	성별	본
본인	홍길동(洪吉童)	1970년 01월 01일	700101-1******	남	南陽

일반등록사항

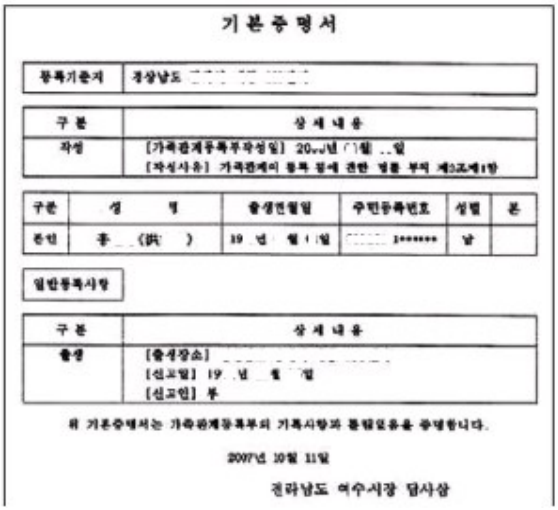
구 분	상 세 내 용
출생	[출생장소] 경상남도 김해시 내동 100번지 [신고일] 1970년 01월 15일 [신고인] 부

위 기본증명서는 가족관계등록부의 기록사항과 틀림없음을 증명합니다.

2007년 10월 11일

전라남도 여수시장 담사삼

Figure G1. The original Basic Certificate of Family Relation Register



기 본 증 명 서

등록기준지	경상남도 [illegible]

구 분	상 세 내 용
작성	[가족관계등록부작성일] 20..년 ..월 ..일 [작성사유] 가족관계의 등록 등에 관한 법률 부칙 제2조제1항

구분	성 명	출생연월일	주민등록번호	성별	본
본인	홍_ (洪)	19 년 월 일	[illegible] 1******	남	

일반등록사항

구 분	상 세 내 용
출생	[출생장소] [illegible] [신고일] 19..년 ..월 ..일 [신고인] 부

위 기본증명서는 가족관계등록부의 기록사항과 틀림없음을 증명합니다.

2007년 10월 11일

전라남도 여수시장 담사삼

Figure G2. A modified Basic Certificate of Family Relation Register (private information such as full name and date of birth has been deleted).

G.4. Consideration for Acceptance of Characters that Cannot be Provided in Printed Form

The consideration for acceptance of characters that cannot be provided with evidence in printed form is not meant to relax IRG requirement in the provision of evidence of modern use. Rather, it is meant to facilitate e-government initiatives for computerized processing of information. It is under this presumption IRG is willing to consider acceptance of characters 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.

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 IRG is

willing to consider their acceptance without actual evidence provided that they are from already implemented working systems only. However, 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 and/or Dis-unification Error

(Same as WG2 Principles and Procedures Annex I)
Source: www.dkuug.dk/jtc1/sc2/wg2/docs/principles.html

Below is the text extracted from Annex I of WG2 Principles and Procedures (WG2N4502) dated 2014-01-28:

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 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 Principles and Procedures 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

Note that if a character has no residue stroke besides the radical, the value 0 should be used.

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

As stated in Section 2.2.1.d, IRG sets an upper limit for the number of ideographs in a working set to ensure sufficient time for delivering quality output in a timely manner. The current limit ($Limit_{IRG}$) is set to about 4,000 ideographs. Also, each submission should not go beyond 1,000 ideographs. Since the number of submissions and their repertoire sizes may differ each time when a new collection is formed, IRG needs some basic guidelines on how the working sets can be formed in a fair manner to accommodate various needs. This Annex serves for this purpose.

At the start of the development work, submitters 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 will be used to form the working set of this collection.

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

> **Scenario 1:** The simplest case solution is that the number of ideographs from each submission should not exceed $Limit_{single\_submission}$, where $Limit_{single\_submission} = Limit_{IRG} / N$. This works especially well when all submission sizes are larger than $Limit_{single\_submission}$.
> **Scenario 2:** In case there is a submission with a total number of ideographs ($TOTAL_{single}$) less than $Limit_{single\_submission}$, the spare quota, $Spare_{single\_submission} = Limit_{single\_submission} - TOTAL_{single}$, can be equally divided among those submissions which exceed $Limit_{single\_submission}$. If spare quota remains afterwards, it can be divided under the same principle (recursively) among submissions which can still take the unused quota.

Even though the above mathematical method can set a quick and undisputed limit to each submission, it may not be the best solution when considering the practical needs of the submitters for different applications. Submitters are encouraged to subdivide their submission and give them priority levels with explanation and justifications. IRG can consider these justifications and agree on a division of $Limit_{IRG}$ close to the one given in the mathematical model above with minor modifications.

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

If the current collection is too large to form a single working set, IRG will use the above principle to split the current collection into multiple working sets, and work on each of them in sequence. In this case, the subdivided sets will be named using the original call name with a subset name assigned in alphabetic sequence, eg. IRG Collection 2015A, IRG Collection 2015B, etc.

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 IRG Convenor Prof. Qin LU).

Doc. No.	Title	Source	Date
WG2 N4502	Principles and Procedures for Allocation of New Characters and Scripts and Handling of Defect Reports on Character Names	WG2	2014-01-28
N681	Annex S https://standards.iso.org/ittf/PubliclyAvailableStandards/c069119_ISO_IEC_10646_2017.zip	Bruce Peterson and IRG Convenor	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
N954AR	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
N1105	Amendments to IRGN954AR	Macao	2005-01-03
N1183	IDS decomposition principles (Revised by IRG)	KAWABATA, Taichi	2005-12-28
N1197	Sample evidence for CJK C1 candidates	Japan	2006-05-22
N1372	On Better use of IDS on IRG development process	KAWABATA, Taichi	2007-11-09
SC2 N3933	ISO/IEC JTC 1 Directives, 5th 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. IRG maintains an up-to-date Unification Examples List.

Character Description Component (CDC): It refers to any symbols that can be used with the Ideograph Description Characters to form an 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 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 term continues to be used in the standardization process and has not changed.

Compatibility Ideographs: Compatibility ideographs are defined in Section 18 of ISO/IEC 10646. Below is a direct quote from ISO/IEC 10646:2017:

> *"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 reviewers 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:

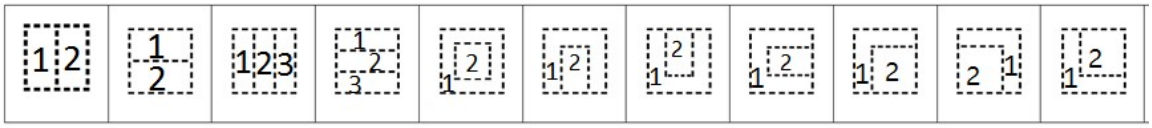
Let G = {Σ, N, P, S}, where

- Σ: 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::=<Binary_Symbol><IDS1><IDS1>|<Ternary_Symbol>
  <IDS1><IDS1><IDS1>
- <IDS1> ::= <IDS> | <CDC>
- <CDC>::= coded_ideograph | coded_radical | coded_component
- <Binary_Symbol> ::=⿰ | ⿱ | ⿴ | ⿵ | ⿶ | ⿷ | ⿸ | ⿹ | ⿺ | ⿻
- <Ternary_Symbol> ::= ⿲ | ⿳

Note1: Even though the IDCs are terminal symbols, they are not part of the CDCs.
Note2: Other than the binary symbol ⿻ (embedment which indicate overlay of two components), all the other 11 IDCs takes the IDS components (either 2 or 3) in a specific order. The order is indicated in the following table:



Ideographic Variation Database (**IVD):** A database maintained by the Unicode Consortium to keep registered Ideographic Variation Sequences as standardized ideographs. See Unicode Technical Report #37 for details[22].

Ideographic Variation Sequence (**IVS**): is a sequence of two coded characters, the first being a character with the Ideographic property that is not canonically nor compatibly decomposable, the second being a variation selector character in the range from U+E0100 to U+E01EF[23]. The first character is referred to as a base character. The purpose of IVSes is to define specific variant glyphs that are unifiable and otherwise cannot be coded according to CJK unification rules. The sequence is a representation of a unifiable character glyph that is not identical to the base character. A registered IVS in IVD has a specific glyph shape defined. Once registered, an IVS can be used as other standardized ideographs.

IRG Collection: An IRG Collection refers to a collection of ideographs consolidated from submissions received upon IRG's call for proposals. The collection is named using the year of the call in four-digit format. A collection called in 2015 is named IRG Collection 2015. An IRG collection may be split into multiple working sets to warrant effective reviews.

IRG Working Document Series (**IWDS**): 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.

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

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

[22] http://unicode.org/reports/tr37/
[23] Extracted directly from UTC Technical Report #37.

Nonce characters/ideographs

A nonce character is a character created specifically for use in a literary work, or series of works, and not intended for general usage. In some cases these characters may even be copyrighted or registered in some way. Nonce words are common in literature like science fiction as the name of some imaginary gadget or alien race, or a character created for literature, commercial or fiscal use. Nonce words in any language are usually quite short lived and so are not included in ordinary dictionaries unless over many years they have gained wider use.

Normalization

Regular Script (楷書**):** Regular script refers to text written in Song style (宋體) and Ming style (明體) which are considered printed forms. It also refers to writing in Kai style (楷體) which is a formal brush style in hand written form. Other more ancient text written formally such as clerical style (隸書) and small seal (小篆) are quite different and are not considered regular script by IRG. Informal writing and artistic expressions written in semi-cursive (行書) and cursive (草書) styles are not considered regular scripts.

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

Urgently Needed Characters: Urgently Needed Characters are a collection of ideographs submitted by a WG2 member body or an ISO recognized organization with a size no bigger than 30 (normally). If the submitter can demonstrate an urgent need for the ideographs to be standardized for justifiable reasons, IRG will accept them for review and endorsement to WG2 for acceptance independent of the current working set. IRG will not initiate any call for urgently needed characters.

Working set: A working set is the set of characters accepted by IRG as a collection or part of a collection to work on for extension to ISO/IEC 10646. Characters accepted in a working set are subject to review by IRG reviewers for inclusion in a particular extension. IRG uses the year of the call in four-digit format to name its new collections. If a new collection is split into multiple working sets, an additional alphabet letter will be used to name these working sets.