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

Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC 1/SC 2/WG 2 *N*_____

ISO/IEC JTC 1/SC 2/WG 2/IRG *N1498*

2008-10-10

Title:	**IRG Principles and Procedures Version 1**
Source:	**IRG PnP Drafting Group**
Action:	**For review by the IRG**
Distribution:	**IRG Members and Ideographic Experts**
References:	**IRGN 1465(PnP Draft1), IRGN 1487(Feedbacks from HKSARG),**
	IRGN1489(Feedbacks from Taichi Kawabata)

Table of Contents

1. Introduction

This document is a standing document of ISO/IEC JTC 1/SC 2/WG 2/IRG for standardization of 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 Chinese-Japanese-Korean (CJK) Unified Ideographs extensions for additions to the standard (ISO/IEC 10646).   Submitters should check the standard documents (including all the amendments and corrigenda) before preparing new submissions. Submitters should check the standard documents (including all the amendments and corrigenda) before preparing new submissions.

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

1.1. Scope of IRG Work

The IRG works for CJK ideograph-related tasks under the supervision of WG2 (SC2 Resolution M13-05). 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
    c.   Ideographic Description Characters
    d.   IICORE (International Ideographs Core)
    e.   CJK Strokes
    f.   Old Hanzi

Work on new projects requires the approval of WG2 and preparation of documents for such approval is required before the project can be proceeded officially in IRG.

1.2 Scope of This Document

The following sections are dedicated for standardization of CJK Unified Ideographs, describing the set of principles and procedures to be applied in the development of a new repertoire of CJK Unified Ideographs as specified in Section 1.1.a.

This document does not cover the standardization of other IRG activities listed in Section 1.1 Standardizing CJK Compatibility Characters 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 in IRG to avoid potential problems. For handling the mis-unification and duplicate ideographs, Appendix I and J of this document should be referenced.

2. Development of CJK Unified Ideographs

(TBD) When and under which conditions will a new extension of CJK Unified Ideographs be developed as an IRG project?

2.1 Principles on Identification of CJK Unified Ideographs

2.1.1. Encoding of abstract characters :

A member of CJK Unified Ideographs is such an abstract character that should be determined by its own abstract shape. A CJK ideographic character can be written in many actual forms depending on the writing style adopted. Examples of common writing styles include Song style and Ming style as typical print form, Kai style as hand written form, and Cao style as cursive form. Stylistically different forms of the same character can be represented by different number of different type of strokes and/or components, which could affect identification of the same abstract shape. In order to reach a common ground to identify those abstract shapes to be encoded as distinct CJK Unified Ideographs, the IRG only accepts submissions using print form of glyphs (usually Song style or Ming style).

2.1.2. Unification procedures of CJK ideographs:

Standard print forms of CJK ideographs are constructed with a combination of known components and/or stroke types. Most of them are determined by 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 [to be revisited]. Basically, two submitted print forms of glyphs with different radicals are distinct characters even they have the same phonetic component. For non trivial cases, further shape analysis must be conducted. Two similar glyphs shall be decomposed into radicals, components and/or stroke types and evaluated by following the unification procedures described in Annex S of ISO/IEC 10646.

2.1.3. Non-cognate rule:

No matter how similar two ideographs are in actual shape, non-cognate or semantically different glyphs shall be considered to have different abstract shapes. The following gives examples of characters with very similar glyphs, yet the characters are semantically different thus considered having different abstract shapes because they are non-cognate.

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

'曰'(U+66F0) and '日'(U+65E5) differ only in contact of strokes (S.1.5 c). [TCA: to provide a relevant example for this case]

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

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

2.1.4. Enhancement of Annex S with new submission:

Examples in Annex S shall be continuously updated. In reviewing character submissions, the IRG shall consider whether or not a new submission is worthy of inclusion in an Annex S update as a new example for unification or disunification.

2.2 Principle on Submission of Ideographs to IRG

2.2.1. Basic Rules for Submission:

A member body may submit the followings to the IRG along with its repertoire. Different information may be handled differently as specified below.

a. **New Sources to existing Standard**. If the submission specifies new sources to some existing standards, they need to be reviewed and approved by IRG and submitted to WG2.

b. **New Sources to working sets**. In case there are some remaining D-set characters in previous standardization stages, new sources reviewed and approved by IRG shall incorporated into the current working sets by the IRG technical editor.

c. **New Compatibility Ideographs.** In case member body needs to add compatibility ideographs, these characters must be reviewed in IRG before submission to WG2 to avoid potential problems with unification and dis-unification of other CJK characters.

d. **New Unified Ideographs.** All ideograph submissions must be subjected to the following rules:

(1). **Collection Size**: A member body should not submit more than 4,000 ideographs at once. This is to minimize the burden of reviewers during eye-ball checking process and to accomplish faster, higher quality of standard.

(2). **Pre-submission Unification Checking**: A member body should be EXTREMELY CAUTIOUS *not to submit the unified ideographs that are already standardized or previously discussed* and recorded at IRG meetings. By nature of the ideographs, it is very difficult for reviewers to find out all unifiable ideographs. Thus, it is important to keep high quality at the time of submission. Low quality submission may become a subject of "5% rule" described below.

(3). **Document Registration**: All submission documents should be registered as IRG N documents, whose file name should be in the form of:

IRGN*nnnn_mmmm*_sss[_*ppp*]_submission

where nnnn indicates an IRG rapporteur assigned document number, *mmmm* indicates member body name, ppp indicates the working set or repertoire name(such as ExtC).

2.2.2. Required Font to be submitted:

a. *Glyph image:* Each proposed ideograph must be accompanied by a corresponding 128 x 128 bitmap file in Song or Ming style. The file name should be the same as the source ID (defined below in Section 2.2.3) with .png as its file extension.
b. *TrueType font* ((optional)): TrueType Font availability is highly recommended although not necessary. Font specification can be found under point 5 of A.1 – Submitter's Responsibilities in Annex A, WG2N3452)

2.2.3. Required Data to be submitted:

 The following data for each proposed ideograph must be submitted with CSV (Comma Separated Value) text format (in UTF-8) or Microsoft Excel format file.
a. *Source ID* to indicate source and the name of the glyph image for keep tracking.   ID should begin with member body code (C,T,J,K,V,KP,H,M,or U) and should be less than 9 characters. ID should contain only Latin capital letter, Arabic number, and hyphen.
b. *Glyph Image file* name or Truetype codepoint of submitted glyphs.
c. *KangXi Radical Code* (R001-R214) with a flag (.0 or .1) to indicate simplified or traditional
d. *Stroke Count* of the Non-radical Component
e. *Flag to show whether the ideograph is traditional (0) or simplified (1).*
f. *First Stroke Code* of the Non-radical Component (ref. IRG N 954 AR)
g. *Ideographic Description Sequence* (ref. Appendix C)
h. *Similar Ideographs and Variant Ideographs*(optional) of submitted ideograph.

2.2.4. Required Evidences to be submitted:

a. **Supporting Evidences**: Evidences should supplied to support the proposed glyph shape and the usage and context with readings, meanings etc. to convince IRG that it is actually being used and/or non-cognate with other similar ideographs.
b. **Questionable Characters**(optional): For those candidates with possible unification questions, submitters are encouraged to supply more detailed evidence of use from authoritative sources and additional information to other related characters, variants and characters similar in shape or meaning encoded in UCS for review.

2.2.5. Quality Assurance: The 5 % rule:

For any character encoding standard, a common general principle is to encode the same character once and only once. It is the submitter's responsibility to filter out already encoded characters before submission. In assessing the suitability of a proposed ideograph for encoding, IRG shall evaluate the credibility and quality of the submitter's proposal. If IRG should find more than 5 % of duplicated characters in the latest UCS from the submitter's source set during the IRG review process, the whole submission will be removed from the subsequent IRG working drafts for that particular IRG project.

2.3 Principles on Production IRG working drafts

After IRG accepts all of submissions, IRG technical editor will produce a set of IRG working drafts.

2.3.1. Principles on Submitted Ideographs

a. All the original ideograph submissions, including glyphs, IDS, radicals, strokes and evidences, must have registered IRG document numbers.
b. If any required information is missing, the IRG technical editor can ask for additional information to the submitter. Without timely supply of such information, the submission can be rejected by the technical for production of working draft.

2.3.2. Principles on Assignment of Serial Number:

a. IRG technical editor should consolidate and sort the submitted ideographs in accordance with Annex A of this document.
b. A unique *serial number* should be assigned to each submission after consolidation.
   The serial numbers must be unique throughout the entire standardization work process.   They must not be changed, re-set, re-numbered, nor re-assigned. This principle makes reference to past discussion easier.

c. If multiple ideographs submitted by different member bodies are obviously unifiable, such ideographs may be unified and assigned the same serial number by IRG technical editor.

2.3.3. Principles on Machine-Checking of IDS of Submitted Ideographs:

a. IRG technical editor should check the submitted IDS with existing IDS data to detect possible unifiable ideographs.
b. Machine checking sometimes detect obviously non-unifiable pairs. Such cases should be filtered out before proceeding to the next stage.
c. IDS checking algorithm should satisfy the requirement described in Appendix B.

2.3.4. Production of IRG working drafts:

a. **Division of Character subsets:** By the result of IDS checking, submitted ideographs shall be grouped into the following two working sets.
   i. **M-set(main set)**: for ideographs with proper IDS, and found not to be unifiable with current standardized ideographs nor previously discussed ideographs with proper IDS.
   ii. **D-set (discussion set)**:,for ideographs with missing or incomplete IDS, or ideographs that might be unifiable with standardized or previously discussed ideographs. Ideographs with missing or incomplete IDS should be commented as such, and checked intensively through manual checking. Ideographs that might be unifiable with standardized or previously discussed ideographs should also be commented as such, and their unifiabilities must be manually checked and supported by evidence for disunification.
b. **Naming of Working Drafts**: The file name should follow the format of "IRGNnnnnVX[XXX]" where "n" is the IRG assigned document number and "X" are version numbers. No spaces are allowed but use of underscore "_" for separation is allowed. Example of version numbers are "V1.0", "V1.0Draft", etc.
c. **Glyph Images**: Archive of consolidated glyph images whose image size should be 128x128 with file name using the Source I with extension .png.
d. **Addition of characters**: No ideographs should be added to the working set once development process begins.
e. **Previous D-Set:** If a previously discussed D-set exist, new D-set ideographs should be merged to the previous existing D-set.
f. After the consolidation, IRG technical editor may ask members to review M-set and D-set based on IRG scheduled review and task division.

2.4 Principles on Reviewing IRG Working Drafts

If review of a working draft is commanded by IRG to member bodies, designated member body editors with assigned portions should review the working draft according to schedule following the following principles.

2.4.1. General Principles on Reviewing

a. Each member body should check the ideographs of the working sets requested by IRG technical editor for the following issues.
   i. Correctness on KangXi radical, Strokes, Radicals, Fast Stroke and IDS.
   ii. Correctness on Glyphs and source information if necessary.
   iii. Any duplicate or unifiable ideographs based on Annex S guidelines.
b. When any data, including IDS, KangXi radical, or strokes is found incorrect, such M-set ideograph should be moved to D-set as its standing data showing uniqueness is no longer valid. Until such ideograph is assured to be unique by manual checking (procedures described in Section 2.4.2), they should not be moved back to M-set.

2.4.2. Principles on Manual Checking (Eyeball Review):

a. **Duplication and Unification:** For D-set ideographs, members should assure that ideographs may not be duplicated or unified with any ideographs in the standard or another working set(including the current one).
b. **Radical Checking:** Assurance is done by enumerating all possible radicals of a target ideograph and look for any duplicate or unifiable ideographs in the range of ±2 strokes of standardized and working ideographs by eyeballs. For example, "聞" may have the radical of "門" with 6 strokes, or radical of "耳" with 8 strokes. In such a case, checking standardized and

working set ideographs with radical of "門" and 4-8 strokes, or ideographs with radical of "耳" and strokes of 6-10 by eyeballs can have much better assurance that such an ideograph does not have duplicate or unifiable ideographs.

c. **Recording of Review**: After eyeball review, reviewing member should put the comment of "Checked against all standardized and working ideographs with radical *X* and stroke of *Y*±2."

2.4.3. Submission of Possibly Unifiable Ideographs

a. **Comments Preparation**: Member bodies should prepare comments and feedbacks with reference to the assigned serial number of the ideograph in question. The guidelines of comments are described in Section 4 of this document. Comments file should be in CSV form as a text file or a Microsoft Excel format file

b. **Additional Evidence and Arguments**: For D-set ideographs that might be duplicated with other standardized or working ideographs, *submitter member body* should prepare arguments with further evidences supporting the use, evidence document showing that suspected ideographs are not unifiable e.g. dictionaries, legal documents, publications, etc. for all of those proposed ideographs which have been questioned for possibly unification with existing UCS or other proposed ideographs in the same working draft or another draft.

c. **Submission deadline:** Each member body should be requested to send comment feedbacks at least two months before the next IRG meeting. IRG technical editor should consolidate them and register the result as IRG N documents a month before the next IRG meeting so that each member body can examine the comments and prepare any additional documents for discussion on IRG meeting.

d. **Rejection**: Questioned ideographs with no counter arguments shall be automatically marked as unified.

2.5 Principles on Discussions at IRG Meetings

2.5.1. Document-based Discussion

For efficient and smooth work, all discussion items and evidences must be prepared with registered IRG documents before the commencement of IRG meeting. Items or evidences not appeared in the IRG document registry are not treated as evidence and will not be discussed during IRG. Any discussions on evidences or items raised after the commencement of IRG meeting may be postponed to the next IRG meeting if any member body requests some longer time to examine such items or evidences.

2.5.2. Discussion Procedures

Discussion should be based on the review comments on working sets. For non-unification issues, submitter should present evidence document(s) showing that suspected unifiable ideographs are distinctively used as non-cognate character in the same region, or that these two characters cannot be unified in accordance with Annex S. When IRG members have consensus that the ideographs are unifiable, the submitter should do one of the following actions, and its decision must be recorded.

a. Withdraw the duplicate ideographs and map submitter character to the existing standardized or working set ideograph.

b. Change the submission as compatibility character by the original proposer.

c. Add this character as a new source to the existing standardized or working set ideograph.

When characters are reviewed by different people, different choice of radical, strokes or first stroke code are possible for the same ideograph. IRG members should resolve to the most appropriate one based on the most common abstract shape of the specific glyph. When KangXi radical or stroke count are agreed to be incorrect, the ideographs should be moved to D-set and wait for another manual review to prevent any unification error caused by not having covered the reviewi with Ideographs with the correct KangXi radical or stroke count.

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

2.5.3. Recording of Discussions:

Comments, rationales, and decisions must be recorded for each of the ideograph reviewed in a tabular way for each reference and check.

2.5.4. Time and Quality Management

Before discussion begins, the number of ideographs under review should be counted and the estimated schedule should be determined based on it. During the discussion, the number of comments reviewed per hour should be noted and the schedule should be adjusted by this rate. If there are more than 600 comments to be reviewed, they may be partitioned and resolved in subsequent IRG meetings if one IRG meeting does not have enough time.

2.6 Principles on Submission of Ideographs to WG2

2.6.1. Stabled M-Set Checking

a.   Once M-set is consolidated and stablized, the ideographs of M-set should be checked at least once as a complete set for intensive checking to assure data and glyph integrities.
b.   Approval by all member bodies is needed before the collection shall be prepared for WG2 submission.

2.6.2. Preparation for WG2 submission.

After the approval by majority of IRG member bodies, IRG technical editor should prepare the following.
a. Sort the final stable M-set ideographs by the sorting algorithm described in Appendix A.
b. Assign provisional UCS code to the finally sorted M-set ideographs (with agreement from ISO 10646 project editor on block assignment).
c. Make available the TrueType fonts for each member body with assigned provisional UCS code (fonts have to be available in accordance with the requirement stated in point 5 of A.1 – Submitter's Responsibilities in Annex A, WG2N3452)
  i.   Each submitter is encouraged to prepare their own font for best font quality.
  ii. If a member body has difficulty creating the font, other member bodies or the IRG technical editor may help creating the font.   In this case, the glyph style of the submitter must be respected.
d. List source references
e. Produce packed Multi-column format Ideograph Chart, made by the created TrueType fonts.
IRG should conduct at least one round of review of the table generated with TrueType font before submission to WG2.

3. Procedures

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

Development procedures described in this section consists of 8 stages, and it may take two to three years to create a good quality ideograph set for standardization.

3.1 Call for Submission

a.   When a member body requests for a new project for CJK Unified Ideograph extension and agreed upon IRG meeting, the IRG may call for submission of new ideographs.   The IRG must also determine the deadline for the submission.
b.   Each member body with proposed ideograph must submit the ideographs before the specified deadline with required data described in Section 2 of this document.
c.   After the submission, member bodies must check whether all required information is accompanied with submitted ideographs.   If some required information is missing or misplaced, IRG technical editor may ask the submitter to resubmit or supply the additional information if only minor problems are encountered. Otherwise, the submission can be rejected because consolidation to other member submissions cannot be carried out.

3.2 Consolidation and Grouping of Submitted Ideographs.

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

a. IRG technical editor should sort and assign *serial numbers* to submitted ideographs as described in the Section 2.3.2.
b. After serial numbers are assigned, submitted ideographs must be checked with IDS to detect duplication and unification. By the result of IDS checking as described in 2.3.3, submitted ideographs will be grouped into M-set and D-set as described in the Section 2.3.4.
c. After the consolidation, the working draft will be assigned a IRG N document with a version number, and will be distributed to member editors and made available so that any other IRG experts can have access. IRG technical editor may ask and assign member editors to check M-set and D-set ideographs either for the entire collection or certain portions depending on reason estimation of workload by the technical editor.

3.3 1st Checking Stage

This stage will be held between IRG meetings. The checking involves the following tasks.

a. Each member editor must check the assigned M-set and D-set for data integrity, correctness, missing data, and duplication. Checking for unification is not mandatory, but desirable. Typical review comment examples for each set are provided in Section 4.
b. Members must submit their comments to IRG technical editor at least two months before the next IRG meeting.
c. IRG technical editor must consolidate the comments and produce an IRG registered document for circulation and discussion at least one month before the next IRG meeting.
d. Submitters are encouraged to prepare supplementary documents (with IRG document numbers) to be submitted so that it can be discussed in the next IRG meeting.

3.4 1st Discussion and Conclusion Stage

This stage will be held during an IRG meeting and the tasks include

a. Members should review the comments which are officially submitted before the meeting with assigned IRG document numbers and the editorial group must supply conclusions for each commented ideographs in writing. Guidelines for typical conclusion are provided in Section 4.
b. All the conclusions must be endorsed/agreed by IRG plenary in its resolution. As a result of resolution, some ideographs would be removed or moved between M-set and D-Set.
c. IRG technical editor should create the newly created M-set and D-set a month after the IRG meeting, and register them as IRG registered document with version information.
d. If more than 5% of ideographs submitted by a specific submitter is removed as a result of duplication or unification with existing standardized set, the entire submission of this submitter should be removed to ensure that high quality of standard project.

3.5 2nd Checking Stage

This stage will be held between IRG meetings with the following tasks.

a. Each member editor must check the newly created M-set and D-set for correctness and duplication.
b. Members should submit their comments with registered IRG document to IRG technical editor at least two months before the next IRG meeting.
c. IRG technical editor should consolidate the comments and produce a registered IRG document for the discussion at least a month before the next IRG meeting.
d. Members are encouraged to prepare supplementary documents which help with the discussion during the next IRG meeting.

3.6 2nd Consolidation and Conclusion Stage

This stage will be held during an IRG meeting with the following tasks.

a. Members must review the comments and provide conclusion for each ideograph. Typical comment and conclusion examples for each set are provided in Section 4.

b. All the conclusions must be endorsed/agreed by IRG plenary in its resolution. As a result of resolution, some ideographs would be removed or moved between M-set and D-Set.
c. IRG technical editor should create the newly created M-set and D-set a month after the IRG meeting, and produce an IRG registered document.
d. If more than 5% of ideographs submitted by a specific submitter is removed as a result of duplication or unification with existing standardized set, the entire submission of this submitter should be removed to ensure that high quality of standard project.

3.7 Final Checking Stage

This stage will be held between IRG meetings with the following tasks.
a. All member editors are requested to check M-set intensively using comments and conclusions made by all previous stages. In the final checking stage, no ideographs are allowed to move from D-set to M-set.
b. Member editors should submit their comments to IRG technical editor at least two months before the next IRG meeting.
c. IRG technical editor should consolidate the comments and produce an IRG registered document for discussion at least a month before the next IRG meeting so that member editors have time to review before the next IRG meeting.

3.8 Approval and Submission to WG2

This stage will be held during an IRG meeting with the following tasks.
a. Members should review the comments on M-set and provide conclusion for each ideograph.
b. If there is no positive decision on an ideograph in the M-set, it should be moved to D-set. No character should be moved from D -set to M-set at this stage.   Ideographs may only be moved from M-set to D-set.
c. With the approval of the majority of IRG member bodies, M-set frozen and be the new ideograph extension set to be submitted to WG2. IRG technical editor should prepare the document in accordance with Section 2.6 of this document.
d. The remaining D-set should not be removed. They should be kept and used in next standardization work to maintain the discussion record and avoid repetition of discussion.

4. Guidelines for Comments and Resolutions on Working Sets

The following tables list guidelines of typical comments and conclusions during the development process. All comments must be accompanied with date (YY-MM-DD format) and member identifier (C, H, M, J, K, KP, U and V).   All conclusions must be accompanied with date, too.

4.1 Guidelines for M-set

M-set is the ultimate target of the standardized ideograph set.   As such, it must be carefully examined. If any suspicious characters are found, they should be moved to D-sets or removed from the working sets all together.

Possible Comment by a Reviewer	Possible Resolution
Worng/Missing Glyph	● Glyph is corrected/supplied and move to D-set for eyeball reviewing.
Wrong KangXi radical / strokes / first stroke	● Data will be corrected and this Ideograph will be moved to D-set. ● Proposal to correct data to remain in M-set cannot take immediate effect in the current round of consolidation as it is an ambiguous case and its change may affect others in the set.
Wrong IDS	● IDS will be corrected and the character will be moved to D-set until they are

Commented [A1]: Taichi: I am not sure I understood your comments in your document IRGN 1489. Please double check this part.

	• machine-checked again. • Move to D-set (in case IDS can't be corrected.)
May be unifiable to U+xxxxx (standardized ideograph)	• Unified to U+xxxx and submitter will request new Source ID to U+xxxx. • Unified to U+xxxx and submitter will request this character as Compatibility Character. • Unified to U+xxxx and this entry will be removed. (May consider to register it to IVS.) • Not unifiable.
May be unifiable to xxxxx (M-set ideograph)	• Unified to xxxxx and this source ID will be attached to xxxxx. • Unified to xxxxx and the submitter may consider it to register as Compatibility Character or IVS. • Not Unifiable.

4.2 Guidelines for D-set

D-set ideographs are the ones that either cannot be checked automatically by IDS checking algorihtm or the ones that are suspected to be unifiable with other standardized or working ideographs. For the ideographs that cannot be machine-checked by IDS matching, at least two non-submitter members must check by human eyeballs to ensure that the ideographs are not unifiable with any standardized ideograph or working ideograph.   For the ideographs that might be unifiable with other ideographs, a submitter is 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. DC is not ideograph	• IDS will be corrected and it will be moved to M-set when next IDS-check is done. • Proper IDS can't be generated and eyeball checking is needed.
Possible Comment by a Reviewer	**Possible Conclusion**
Wrong KangXi radical / strokes / first stroke	• Data will be corrected. • Proposal to correct data is not accepted, as it is ambiguous case and IRG agreed that the previous choice of XX is more appropriate.
Wrong IDS	• IDS will be corrected and will be machine-checked again. • Correct IDS can't be generated and human eyeball check is needed.
May be unifiable to *U+xxxxx* (standardized ideograph)	• Unified to *U+xxxxx* and new source is added to *U+xxxxx*. Entry is no longer used. • Not unifiable, as shown by the evidence *IRG N xxxx*. Move to M-set.
May be unifiable to *xxxxx* (M-set or D-set Ideograph)	• Unified to *xxxxx* and this entry is no longer used. • Unified with *xxxxx*. (*xxxxx* is removed.) • Not Unifiable, as shown by the evidence *IRG N xxxx*. Move to M-set
Checked against all standardized and working ideographs with radical *X* and stroke of *Y*±2.	• Move to M-set, as two non-submitter members (*XX* and *YY*) ensured that this ideograph is not unifiable with any existing standardized or working ideographs. • Checking against ideographs with radical *X* may not be enough. This ideograph should also be checked against ideographs with radical *Z*, too.

5. IRG web site

The IRG maintains its own web site 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, minutes, resolutions, document register, documents and standing documents are made available at this site. Hyperlinks to WG2 websites will be provided for members' easy access. For faster retrieval of the documents and searching, documents should not be compressed and the site search engine window should be available. Documents longer than 4Mb must be split into multiple files for easy upload, download and searching.

5. IRG Document Registration

All documents to be discussed in IRG formally must be registered with IRG with assigned IRG document numbers.

5.1 Registration Procedure

The following gives the registration Procedure.
    a. **Request for Document Number:** All documents submitted to IRG must be given a registered document number. The assignment is done by IRG Rapporteur. A member body shall first contact the IRG rapporteur for a document number with a supplied document title. Once the document number is assigned, the information will be posted on the IRG Website. Some document numbers can be pre-assigned during IRG meetings for activities between IRG meetings.
    b. **Submission of documents:** All registered documents must be submitted to the IRG rapporteur. The submitted document must also contain its IRG document number in text form so that searching can be supported.
    c. **Posting of documents**: Properly submitted documents in a timely manner are then posted by the IRG rapporteur on the IRG site as an official document.
    d. **Disqualified documents**: Documents with certain basic information missing such as submitter name, title, purpose can be rejected by IRG rapporteur for posting. All other documents which did not go through this registration process from a. to c. and preliminary review by IRG rapporteur for basic information will not be treated as IRG documents and thus corresponding issues to be addressed will not be discussed in IRG formally.

5.2 Contact for IRG document Registration

The Current IRG rapporteur is Dr. Qin LU and the following is her contact information:

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

Annex A: Sorting Algorithm of Ideographs

The ideographs must be sorted by the following order.

a. **KangXi Radical order**.
   Note: When radicals are in simplified forms given below, ideographs with simplified radicals must be placed after the ideographs with corresponding traditional radicals.

Simplified Radicals		Traditional Radicals	
R119. 1	纟	R119. 0	糸
R146. 1	见	R146. 0	見
R148. 1	讠	R148. 0	言
R153. 1	贝	R153. 0	貝
R158. 1	车	R158. 0	車
R166. 1	钅	R166. 0	金
R167. 1	长	R167. 0	長
R168. 1	门	R168. 0	門
R177. 1	韦	R177. 0	韋
R180. 1	页	R180. 0	頁
R181. 1	风	R181. 0	風
R182. 1	飞	R182. 0	飛
R183. 1	饣	R183. 0	食
R186. 1	马	R186. 0	馬
R194. 1	鱼	R194. 0	魚
R195. 1	鸟	R195. 0	鳥
R196. 1	卤	R196. 0	鹵
R198. 1	麦	R198. 0	麥
R204. 1	黾	R204. 0	黽
R209. 1	齐	R209. 0	齊
R210. 1	齿	R210. 0	齒
R211. 1	龙	R211. 0	龍

b. **Number of Strokes**.
   Note: Simplified characters must be put after the non-simplified characters within the same stroke-number groups.

c. **First stroke**.

Annex B: IDS Matching

B.1 Guidelines on Creation of IDS

Each member body should consult IRG N 1155 for the creation of IDS.

B.2 Requirements on IDS Matching.

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

1. IDS matching should be able to handle different split point.
   (e.g. ⿰亻頃 and ⿰化頁 should be matched.)
2. IDS matching should be able to handle different split level.
   (e.g. ⿰亻悉 and ⿰亻⿱釆心 should be matched.)
3. IDS matching should match different glyphs of the same abstract shape.
   (e.g. ⿰礻申 and ⿰示申 should be matched.)
4. IDS matching should match similar glyphs.
   (e.g. ⿰忄生 and ⿰小生 should be matched.)
5. IDS matching should match IDS with different ordering of overlapping IDC.
   (e.g. ⿴三丨 and ⿴丨三 should be matched.)
6. IDS matching should match unifiable IDC patterns.
   (e.g. ⿰麥离 and ⿲麥离 should be matched.)
7. IDS matching should be able to handle the combination of all the above.
8. IDS matching should be able to detect any inappropriate IDS, such as too long IDS, IDS with non-ideographic DC, or missing or extra DC or IDC.

B.3 Limitation on IDS Matching.

It should be noted that IDS matching cannot detect the unification or duplication if the component cannot be encoded by an IDS, or if the glyph itself is very complex.   IDS matching is done by strict programming logics. It is not versatile on detection of the unifiable ideographs unless rules are explicitly given to the algorithm. Thus, it is not meant to be the replacement for manual checking. Rather, it is an assistive tool for quality assurance to identify duplication and known cases of unification. Therefore, it is very important that each submitter should carefully check their submitted ideographs not to be unifiable with any standardized or previously discussed ideographs.

Annex C: Urgently Needed Ideographs

C.1 Introduction

When a member body urgently needs very few ideographs to be standardized for some good reasons (such as they are Regional or National Standard ideographs), with the approval of IRG, the member body may submit the ideographs independent of currently working set to the WG2.

C.2 Requirements

The submitter of urgently needed ideographs must prepare the following documents.

b. All the documents required by the normal ideograph submissions.
c. In addition to the above, the document to show any unifiable ideographs in currently working sets against the submitted ideographs.
d. For the ideographs not mentioned above, the document must prove that their submitted ideographs are not unifiable with any ideographs in the currently working set. Proofs may be provided by showing which document the submitter checked, ideographs of which radicals and strokes they checked against each of submitted ideographs. It is an important responsibility of the submitter to check with not only current standardized CJK ideographs, but also the working set for any unifiable characters against their submission.   Failure to do so, will not be approved by IRG for endorsement of independent submission.

C.3 Dealing with Urgent Requests

IRG may at its discretion accept the document from he submitter of urgently needed ideographs for discussion of the amount of work is considered to be reasonably small for IRG review without unreasonable disruption to its on-going projects. Accepted submissions must be checked by IRG for correctness, duplication and unfication. All accepted ideographs as independent submission must be checked with current working set. Any one found to be identical or unifiable with the ones in the current working sets, such ideograph must be noted and removed from the currently working sets if approval by WG2 is given.

WG2 PnP Annex I: Guideline for handling of CJK ideograph unification and/or disunification error

(Source: ISO/IEC JTC 1/SC 2/WG 2 N2576R – 2003-10-21)

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

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 in the BMP, 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 be used for the currently encoded ideograph shall be resolved as follows. In the interest of synchronization between ISO/IEC 10646 and the Unicode standard, the ideograph with the glyph shape that is similar to the glyph that is published in the "Unicode Charts" will continue to be associated with its current code position. For the ideographs outside the BMP, the glyph shape in ISO/IEC 10646 and the Unicode Charts are identical and will be used with its current code position.
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.

WG2 PnP Annex J: Guideline for correction of CJK ideograph mapping table errors

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

In principle, mapping table or reference to code point of 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 following guidelines:

J.1 Priority of error correction procedure
    A. Consider adding new code position and source-reference mapping for the character in question rather than changing the mapping table.
    B. If change of mapping table is unavoidable, correction should be done as soon as possible.

J.2 Announcement of addition or correction of mapping table
Once any addition or correction of 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 WG 2 meeting, followed by subsequent process resulting in an appropriate amendment to the standard.

J.3 Collection and maintenance of mapping tables that are not owned by WG 2
There are many mapping tables, which are included in national/regional standards or developed by third parties.   These are out of WG 2'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.

References

Document numbers in the first column in the following table refer to IRG working documents (ISO/IEC JTC 1/SC 2/WG 2/IRG Nxxxx), except where noted otherwise.   For those documents for which a link is not given, you may try http://www.cse.cuhk.edu.hk/~irg/ ; some of the older documents are available only in paper form (contact the IRG Rapporteur of JTC1/SC 2/WG 2/IRG – Prof. Lu Qin).

Doc. No.	Title	Source	Date
WG2 N3201	Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names	WG2	2007-03-14
N681	Annex S	Bruce Peterson and IRG Rapporteur	1999-11-18
N881	CJK Extension C Submission Format	IRG	2001-12-04
N953	Minutes of the Adhoc meeting on submitted documents: N941, N942, N944, N945, N948, N949	CJK ad hoc group	2002-11-22
N954	Report on first stroke/stroke count by ad hoc group	CJK ad hoc group	2002-11-22
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 IRG N954AR	Macao	2005-01-03
N1183	IDS decomposition principles(Revised by IRG)	KAWABATA, Taichi	2005-12-28
N1197	Sample evidences for CJK C1 candidates	Japan	2006-05-22
N1372	On Better use of IDS on IRG development process	KAWABATA, Taichi	2007-11-09

Glossary:[to be updated later]

Source: A reputable published document such as a dictionary, a standardization document, or a well published and widely read or referenced book which the IRG would consider as authoritative such that the characters in this source are considered reliable and stable for consideration of inclusion.

Abstract shape:

D-set:

M-set:

Working set:

Compatibility characters:

Ideographic Description Sequence(IDS):

Flow Chart of Section 4 (Procedures) of IRG PnP Document proposal.

Call for Submission

submission

Consolidation

IDS checking

Old D-set

Standard

(merge)

M-set

D-set

1st reviewing & conclusion stage

unification

possibly unifiable
non-unification

wrong typeface / data
eyeball-checking

New M-set

New D-set

2nd reviewing & conclusion stage

unification

possibly unifiable
non-unification

wrong typeface / data
eyeball-checking

New M-set

New D-set

final reviewing & resolution stage

unification

possibly unifiable

wrong typeface / data

New M-set

New D-set

WG2
submission

Standard

D-set

They will be used in next standardization stages.