IRG N1489

Subject: IRG PnP feedback from Taichi
From: Yasuhiro Anan <yasuhira@microsoft.com>
Date: Tue, 1 Jul 2008 07:58:01 +0800
To: Lu Qin <csluqin@comp.polyu.edu.hk>

Dear Dr. Lu,

The attached is the feedback I received from Taichi on IRG N 1465 (IRG PnP v1 draft).
Would you give an IRG number and upload them to the IRG website so that people can review?

Best regards,
/anan

PnP_flow_chart.ppt

PnP_Taichi-30Jun2008.doc

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 *N1489R*
2008-7-03

Title:	**IRG Principles and Procedures Version 1**
Source:	**IRG PnP Drafting Group**
Action:	**For review by IRG**
Distribution:	**IRG Members and Ideographic Experts**

Table of Contents

Deleted: *1465Draft*

Deleted: 6-13

Formatted: Font: (Asian) MS Mincho, (Asian) Japanese

Deleted: 5

Deleted: 6

Deleted: 6

Deleted: 6

Deleted: 7

Deleted: 7

Deleted: 7

Deleted: 8

Deleted: 8

Deleted: 8

Deleted: 9

Deleted: 10

Deleted: 10

Deleted: 10

Deleted: 10

Deleted: 10

Deleted: 11

Deleted: 11

Deleted: 11

Deleted: 11

Deleted: 12

Deleted: 13

Deleted: 1. Introduction→ 2¶
2. IRG Scope→ 2¶
3. Development of CJK Unified Ideographs→ 2¶
3.1 General principles - Characters Not Glyphs→ 2¶
3.2 Preparation for submission to IRG→ 3¶
3.3 Production and review process of IRG working drafts→ 3¶
3.4 Preparation for discussions at IRG meetings→ 4¶
3.5 Recording of unification arguments and decisions→ 4¶
3.6 Preparation for submission to WG2→ 4¶
4. Handling defect reports→ 4¶
5. IRG web site→ 4¶
Annex A: Information accompanying submissions→ 5¶
Annex B: IDS matching→ 5¶
B.1 Simple IDS→ 5¶
B.2 Handling of complex or incomplete IDS→ 5¶
Annex C: Work flow and stages of progression→ 5¶
C.1 The IRG working drafts→ 5¶
C.2 Stages of progression→ 5¶
C.3 Dealing with urgent requests→ 5¶
C.3 Dealing with individual submission to WG2→ 5¶
WG2 PnP Annex I: Guideline for handling of CJK ideograph unification and/or disunification error→ 6¶
I.1 Guideline for “to be unified” errors→ 6¶
... [1]

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 repertoire of 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.

Deleted: The document also contains procedures and guidelines for ###. …

Deleted: Submitters are encouraged to visit the "Unihan Database" page on the Unicode web site for more information on checking whether a CJK Ideograph is already encoded in the standard.…

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

2. Scope of This Document

Deleted: IRG

IRG works for CJK ideograph-related tasks under the supervision of WG2 (SC2 Resolution M13-05). The following are the list of IRG projects.

- CJK Unified Ideographs and its extensions.
- Kangxi Radicals and CJK Radical Supplements
- Ideographic Description Characters
- IICORE
- CJK Strokes
- Old Hanzi

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.

Deleted: and

This document does not cover the standardization of other IRG activities described above. Standardizing CJK Compatibility Characters maintained in UCS for the purpose of round-trip integrity with other standards is out of IRG scope. For handling the mis-unification and duplicate ideographs, Appendix I and J of this document should be referenced.

Deleted:

3. Principles

Deleted: 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?

3.1 Principles on Identification of CJK Unified Ideographs

Deleted: ***General principles***

Deleted: ***- Characters Not Glyphs***

A. UCS encodes 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 a writing style such as Song or Ming style (typical print form), Kai style (hand written form), Cao style (cursive form), etc., and those 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 shapes. In order to facilitate a common ground to identify those abstract shapes to be encoded as distinct CJK Unified Ideographs, IRG accepts submissions consisting only from print form of glyphs (usually Song or Ming style).

B. 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 dictionary and possibly reflect the meaning of the character and a phonetic component which represents the reading of the character [to be revisited]. Basically, two submitted print form of glyphs with the same phonetic component are distinct characters if they have different radicals. For non trivial cases, further shape analysis will 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.

Deleted: ***procedures***

C. 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.

'戌'(U+620C) and '戍'(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 the shape analysis might not tell non-cognacy or semantic differences, it is submitter's responsibility to provide supporting evidences in order to invoke the non-cognate rule.

D. Enhancement to Annex S with new Submission:

The Annex S examples shall be continuously updated. In reviewing characters submission, the IRG shall consider whether or not a new submission is worth including in the Annex S as a new example for unification or disunification.

3.2 Principles on Submission of Ideographs to IRG

A. Basic Rule for Submission:

- A member body may submit the followings to the IRG. All the following submissions must accompany the documents described below.
  - ➢ ***New Sources to existing standard***. (After IRG check, new sources to existing standard should be submitted to WG2.)
  - ➢ ***New Sources to previous working sets*** (In case there are some remaining D-set characters in previous standardization stages. New sources to the previous working set should be incorporated to the current working sets by the IRG technical editor.)
  - ➢ ***New Compatibility Ideographs*** (After IRG check, they should be submitted to WG2.)
  - ➢ ***New Unified Ideographs***. (If not applicable to all the above. Scope of this document.)
- 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.
- 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 reviewer 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.
- All submission document should be registered as IRG N documents, whose file name should be "IRGN*nnnn*_*pppp*_*mmmm*_submission", where nnnn indicates document number, *pppp* indicates project name (such as "ExtC", "ExtD", etc), and *mmmm* indicates member body name.

B. Required Font to be submitted:

- ***A glyph image*** with a specified dimension and filename in bitmap format (128 x 128 image) for each proposed ideograph in Song or Ming style.
- (optional) ***TrueType font*** for the glyph of the proposed ideographs (as specified under point 5 of A.1 – Submitter's Responsibilities in Annex A, WG2N3452)

C. 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.

- ***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.
- ***Glyph Image file*** name or Truetype codepoint of submitted glyphs.
- ***KangXi Radical Code*** (R001-R214) with a flag (.0 or .1) to indicate simplified or traditional
- ***Stroke Count*** of the Non-radical Component
- ***Flag to show whether the ideograph is traditional (0) or simplified (1).***
- ***First Stroke Code*** of the Non-radical Component (ref. IRG N 954 AR)

Deleted: ***Preparation***

Deleted: ***for***

Deleted: ***submission***

Formatted: Font: Bold, Italic

Formatted

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: ***A***

Deleted: ***data to be submitted***

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Formatted: Bulleted + Level: 1 + Aligned at: 0.49" + Indent at: 0.78"

Formatted

Deleted: t

Deleted: the specified

Deleted: usually

Deleted: together with its glyph image

Formatted: Font: Bold, Italic

Deleted: <#>KXI (KangXi Index with a flag to indicate real or virtual)¶

Formatted: Font: Bold, Italic

Deleted: KangXi Radical

Deleted:)

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

- ***Ideographic Description Sequence*** (ref. Appendix C)
- (optional) ***Similar Ideographs and Variant Ideographs*** of submitted ideograph.

D. Required Evidences to be submitted:

- Evidences to support the proposed glyph shape and the usage and context with readings, meanings etc. of the proposed ideograph to convince it is actually being used and/or non-cognate with other similar ideographs should be provided with proposed ideographs.
- (optional) Especially for those candidates with possible unification questions, submitters are encouraged to supply more detail evidence of use from authoritative source and additional information to other related characters, variants and characters similar in shape or meaning encoded in UCS for review.

E. 5 % rule:

For any character encoding standard, a common general principle is to encode the same character once and only once. It is submitter's responsibility to filter out already encoded characters from her proposal. 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.

3.3 Principles on Productionof IRG working drafts

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

A. Principles on Submitted Ideographs

- All the original ideograph submissions, including glyphs, IDS, radicals, strokes and evidences, must be registered as IRG N document.
- If required information is missing, IRG technical editor may ask for the additional information to the submission member, or reject the submission.

B. Principles on Assignment of Serial Number:

- IRG technical editor should consolidate and sort the submitted ideographs in accordance with Annex A of this document.
- The *serial numbers* should be assigned to them after the sorting.
  - ➢ The serial numbers must be unique throughout the entire standardization works. They must not be changed, re-set, re-numbered nor re-assigned. This principle would make a reference to the past discussion easier.
- If multiple ideographs submitted by different member bodies are obviously unifiable, such ideographs may be unified and assigned a same serial number by IRG technical editor.

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

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

D. Production of IRG working drafts:

-
- By the result of IDS checking, submitted ideographs may be grouped into the following two working sets.
  - ➢ M-set (main set), for the ideographs with proper IDS, and found not to be unifiable with current standardized ideographs nor previously discussed ideographs with proper IDS.
  - ➢ D-set (discussion set), for the ideographs with missing or incomplete IDS, or ideographs that might be unifiable with standardized or previously discussed

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Formatted: Font: Bold, Italic

Deleted: <#>Unique ID to indicate source and the name of the glyph image for keep tracking¶
<#>Evidences to support the proposed glyph shape and the usage and context with readings, meanings etc. of the proposed ideograph to convince it is actually being used and/or non-cognate with other similar ideographs.¶

Deleted: ***B***

Deleted: ***Optional data***

Deleted: For questionable characters e

Deleted: member bodies

Deleted: <#>TrueType font for the glyph of the proposed ideographs (as specified under point 5 of A.1 – Submitter's Responsibilities in Annex A, WG2N3452)¶

Deleted: ***C***

Formatted: Font: (Asian) MS Mincho, (Asian) Japanese

Deleted: ***and review process***

Formatted

Formatted: Indent: Left: 0"

Deleted: ***A***

Deleted: After IRG accepts all of submissions, IRG technical editor will produce a set of IRG working (TBD) Describe the following:¶
<#>All working drafts should register with an IRG document number¶
<#>All editors should request IRG document number from Rapporteur and comments should be submitted with the IRG document number assigned.¶
<#>Consolidated comments should be prepared with a IRG document number.¶
Unique Character id: once given, do not change across all versions of the same project.…M set, D set and
<#>M set, D set and other sets for the purpose of discussion. [(Note: add explanation about D Set, M set and other sets for reference.) Criteria for putting characters into these sets.]¶
<#>Machine generated dup lists according to IDS data.¶

Deleted: (TBD) Describe the following:¶
<#>All working drafts should register with an IRG document number¶
<#>All editors should request IRG document number from Rapporteur and comments should be submitted with the IRG document number assigned.¶
<#>Consolidated comments should be prepared with a IRG document number.¶
Unique Character id: once given, do not change across all versions of the same project.…M set, D set and
<#>M set, D set and other sets for the purpose of discussion. [(Note: add explanation about D Set, ... [2]

Deleted: <#>M set, D set and other sets for the discussion. [(Note: add explanation about D Set, ... [3]

ideographs. Ideographs with missing or incomplete IDS should be commented as is, and checked intensively by human eyeballs. Ideographs that might be unifiable with standardized or previously discussed ideographs should also be commented as so, and by evidence, their unifiabilities must be checked.

- The file name should follow the format of "IRGNnnnnXXXXX" where "n" is assigned document number and "X" are alphabets for easy identification. No spaces are allowed but use of underscore "_" for separation is allowed. Use short form "Vn", e.g. V3 for version 3. Use shorter form as far as possible for convenience use.
- Archive of consolidated glyph images whose image size is 128x128, file format is "png" and whose file name is Source-ID.
- No ideographs should be added to the working set once development process begins.
- If previously discussed D-set exist, new D-set ideographs should be merged to existing D-set.
- After the consolidation, IRG technical editor may ask members to review M-set and D-set.

3.4 Principles on Reviewing IRG Working Drafts

IRG technical editor may request and assign member bodies to review the working drafts.

A. General Principles on Reviewing

- Each member body should check the ideographs of working sets requested by IRG technical editor for the following viewpoint.
  - ➢ Correctness on KangXi radical, Strokes, Radicals, Fast Stroke and IDS.
  - ➢ Correctness on Glyphs and source information if necessary.
  - ➢ Any duplicate or unifiable ideographs based on Annex S guidelines.
- 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 eyeball (procedures described in next subsection), they should not be moved back to M-set.

B. Principles on Eyeball Reviewing

- For D-set ideographs, members should assure that ideographs may not be duplicated or unified with any ideographs in the standard or working set.
- Assurance is done by enumerating all possible radicals of 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 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 eyeball may assure that such ideograph do not have duplicate or unifiable ideographs.
- After eyeball review, member should put the comment of "Checked against all standardized and working ideographs with radical *X* and stroke of *Y*±2."

C. Reviewing Possibly Unifiable Ideographs

- 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 to be possibly unifiable to existing UCS or other proposed ideographs in the same working drafts.
- The questioned ideographs with no counter arguments shall be automatically marked as unified and IRG will move on.

D. Submission of Comments

- Member bodies should prepare the comment with serial number of ideograph. The guidelines of comments are described in Section 5 of this document. Comments file should be CSV text file or Microsoft Excel format file.

Formatted

Formatted

Formatted: Bulleted + Level: 1 + Aligned at: 0.58" + Indent at: 0.88"

Deleted: ***B. Review process of IRG working drafts:***¶
(TBD) Describe the following:¶
how to split review work. Project editor can split and assign the review work to members depending on the amount of submission to be reviewed.¶
what to look for.¶
duplicate characters¶
data errors (glyph shape, KXI, Rad, SC, FS)¶
how to review¶
Use of known patterns in Annex S¶
Use of the updated list of characters of unification examples in the IRG standing document.¶
how to return feedback. Each review cycle has its schedule. Members missing the review deadline will not have their comments considered. ...

Formatted

Formatted: Normal

Formatted

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Normal, Bulleted + Level: 1 + Aligned at: 0.58" + Indent at: 0.88"

Formatted: Font: Italic

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

3.5 Principles on Discussions at IRG meetings

A. Document-based Discussion
For the efficient and smooth work, discussion items and evidences must be prepared as registered IRG N documents before the commencement of IRG meeting. Items or evidences not appeared in the IRG N documents should not be discussed. Any discussions on evidences or items raised after the commencement of IRG meeting may be postponed to next IRG meeting if any member body requests some longer time to examine such items or evidences.

B. Discussion Procedures

Discussion should be based on the review comments on working sets.

For non-unification issues, submitter should present the evidence document showing that suspected two ideographs are distinctively used as non-cognate character in the same region, or that these two characters should cannot be unified in accordance with Annex S. When IRG members agreed that the ideographs are unifiable, the submitter should do one of the following actions, and its decision must be recorded.

- Remove as duplicate ideographs with existing standardized or working set ideograph.
- Submit as compatibility character by original proposer.
- Adding new source to existing standardized or working set ideograph.

For data issues, when different choice of radical, strokes or first stroke code is possible for the same ideograph, IRG members should resolve the most appropriate one based on the most common abstract shape of the specific glyph. When KangXi radical or stroke count data are agreed to be incorrect, they should be moved to D-set, and should wait for another eye-ball review to prevent any unification check error caused by not reviewing such Ideographs with newly corrected KangXi radical or stroke count.

Guidelines for typical comments and resolutions are appeared in Section 5 of this document.

C. Recording the Discussion
Comments and Resolutions must be recorded to working drafts beside the target ideograph.

D. Time and Quality Management
Before the discussion begins, the number of comments should be counted and the schedule should be determined based on it. During the discussion, the number of comments resolved per hour should be counted and the schedule should be adjusted by this rate. If there are more than 600 comments to be resolved, resolution stage should be extended to multiple IRG meetings.

3.6 Principles on Submission of Ideographs to WG2

A. Final checking stage

- When M-set is consolidated, the ideographs of M-set should be intensively checked for the data and glyph integrities.
- With the approval of all member bodies, the ideograph set shall be prepared for WG2 submission.

B. Preparation for WG2 submission.
After the approval by the IRG member bodies, IRG technical editor should prepare the following.

Formatted: Normal, Bulleted + Level: 1 + Aligned at: 0.58" + Indent at: 0.88"

Formatted: Normal

Deleted: ***Preparation for***

Deleted: ***discussions***

Formatted: Indent: Left: 0"

Deleted: ***Unification issues:***

Deleted: After filtering out obvious cases from machine generated duplication report, submitters must prepare arguments with further evidences supporting the use, e.g. dictionaries, legal documents, publications, etc. for all of those proposed ideographs which have been questioned to be possibly unifiable to existing UCS or other proposed ideographs in the same working drafts. The questioned ideographs with no counter arguments shall be automatically marked as unified and IRG will move on.…¶
¶
For questionable characters, member bodies must supply more detail evidence of use from authoritative source and additional information to other related characters, variants and characters similar in shape or meaning encoded in UCS for review. ¶
¶
Further examples on the relationship with the other characters that are possibly unifiable can speed up the review and enhance quality of the work.¶

Deleted: ¶
For questionable characters, member bodies must supply more detail evidence of use from authoritative source and additional information to other related characters, variants and characters similar in shape or meaning encoded in UCS for review. ¶
¶
Further examples on the relationship with the other characters that are possibly unifiable can speed up the review and enhance quality of the work.¶

Formatted: Bulleted + Level: 1 + Aligned at: 0.49" + Indent at: 0.78"

Formatted: Font: (Asian) MS Mincho, (Asian) Japanese

Deleted: ***B. Data issues:***¶
(TBD) Describe the following: [pending Anan San to clarify the purpose of this section]¶
<#>Different choice of Rad, SC, FS etc, which may or may not affect KXI. In case of making different choice of the Radical, other attributes may be affected and should be changed accordingly.¶

Formatted: Indent: Left: 0"

Deleted: ***3.5 Recording of unification arguments and decisions***¶
IRG should maintain all record of unification arguments and decisions and publish it at the IRG website. Search engine will be adopted to facilitate the searching of these information for reference.¶
Recording format and useful indices for easy search.¶

Deleted: ***Preparation for s***

Deleted: (TBD) Describe the following:¶

- Sorting the final M-set ideographs with sorting algorithms described in Appendix A.
- Assignment of provisional UCS code to the finally sorted M-set ideographs.
- 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)
    - ➢ Each submitted member body is encouraged to prepare their own font for best font quality.
    - ➢ 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 submitted member body must be respected.
    - ➢ The IRG should at least conduct one round review of the table generated with TrueType font before submission.
- Source references
- Packed Multi-column format Ideograph Chart, made by the created TrueType fonts.

4. Procedures

This section describes the basic development procedures of CJK Unified Ideograph extension. 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 the standardized ideograph set.

4.1 Call for Submission

- When member bodies requests the 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.
- Each member bodies with proposed ideograph must submit the ideographs before the specified deadline with required data described in Section 3 of this document.
- After the submission, member bodies may check whether all required information is accompanied with submitted ideographs. If required information is missing, IRG technical editor may ask for the additional information to the submission member, or must reject the submission.

4.2 Consolidation and Grouping of Submited Ideographs.

- This stage will be held between IRG meetings.
- IRG technical editor should sort and assign *serial numbers* to submitted ideographs as described in the Section 3.
- After serial numbers are assigned, submitted ideographs must be checked with IDS to detect duplication or unification. By the result of IDS checking, submitted ideographs may be grouped into M-set and D-set as described in the Section 3.
- After the consolidation, working draft will be registered as IRG N document, and will be distributed to the members in IRG meeting. IRG technical editor may ask and assign members to check M-set and D-set ideographs. Number of assignment should be proportional to the number of submitted ideographs.

4.3 1st Checking Stage

- This stage will be held between IRG meeting.
- Each member body must check the assigned M-set and D-set for the data integrity, correctness and duplication. Checking for possible duplication or unification is not mandatory.

Deleted: Preparation of

Formatted: English (US)

Formatted

Formatted: Bullets and Numbering

Deleted: The IRG should at least conduct one round review of the table generated with TrueType font before submission…Members are encouraged to review and <#>Members are encouraged to review and comment on IRG submissions to WG2. The IRG Rapporteur will forward members' comments to WG2. ¶

Formatted: Indent: Left: 0.78", No bullets or numbering

Formatted: Bullets and Numbering

Deleted: <#>Members are encouraged to review and on IRG submissions to WG2. The IRG Rapporteur will forward members' comments to WG2. ¶

Formatted: Normal

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Indent at: 0.29"

Formatted: Font: Italic

Formatted: Superscript

Formatted: Normal

- Typical comment examples for each set are provided in the next section.
- Members must submit their comments to IRG technical editor at least two months before the next IRG meeting.
- IRG technical editor must consolidate the comments and register as IRG N document for the discussion at least a month before the next IRG meeting.
- Members are encouraged to prepare the document which may augment the discussion during the IRG meeting.

4.4 1st Discussion and Resolution Stage

Formatted: Superscript

- This stage will be held during IRG meeting.
- Members should review the comments and provide the resolution for each commented ideographs.
- Guidelines for typical resolution are provided in the next section.
- As a result of resolution, some ideographs would be removed, or moved between the sets.
- IRG technical editor should create the newly created M-set and D-set a month after the IRG meeting, and register them as IRG N document.
- If more than 5% of ideographs submitted by the specific member body are removed as a result of duplication or unification with existing standardized set, then the entire submission should be removed to ensure the higher quality of standard.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Indent at: 0.29"

Formatted: Not Superscript/ Subscript

4.5 2nd Checking Stage

Formatted: Superscript

- This stage will be held between IRG meeting.
- Each member must check the newly created M-set and D-set for the correctness and duplication.
- Members should submit their comments to IRG technical editor at least two months before the next IRG meeting.
- IRG technical editor should consolidate the comments and register as IRG N document for the discussion at least a month before the next IRG meeting.
- Members are encouraged to prepare the document which may augment the discussion during the IRG meeting.

Formatted: Bulleted + Level: 1 + Aligned at: 0" + Indent at: 0.29"

4.6 2nd Consolidation and Resolution Stage

Formatted: Superscript

- This stage will be held during IRG meeting.
- Members must review the comments and provide the resolution for each ideograph.
- Typical resolution examples for each set are provided in the next section.
- As a result of resolution, some ideographs would be removed, or moved among the sets.
- IRG technical editor should create the newly created M-set and D-set a month after the IRG meeting, and register them as IRG N document.
- If more than 5% of ideographs submitted by the specific member are removed as a result of unification with existing standardized set, then the entire submission should be removed to ensure the higher quality of standard.

4.7 Final Checking Stage

Formatted: Normal

- This stage will be held between IRG meeting
- All members are requested to check M-set intensively, among the comments and resolutions made by previous stages. In the final checking stage, no ideographs are allowed to move from D-set to M-set.
- Members should submit their comments to IRG technical editor at least two months before the next IRG meeting.
- IRG technical editor should consolidate the comments and register as IRG N document for the discussion at least a month before the next IRG meeting.

4.8 Approval and Submission to WG2

- This stage will be held during IRG meeting
- Members should review the comments on M-set and provide the resolution for each ideograph.
- No character should be moved from D -set to M-set at this stage. Ideographs may only moved from M-set to D-set.
- With the approval of the majority of IRG member bodies, M-set will be the new ideograph extension set to be submitted to WG2. IRG technical editor should prepare the document in accordance with Section 3.6 of this document.
- Remaining D-set should not be removed. They should be kept and used in next standardization work to maitain the discussion record and avoid repetition of the discussion.

5. Guidelines for Comments and Resolutions on Working Sets

The following tables show guidelines of typical comments and resolutions during 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 resolutions must be accompanied with date, too.

5.1 Guidelines for M-set

Formatted: Normal

M-set is the subject of the standardized ideograph set. As of it, it must be carefully examined and if any suspicious characters are found, they should be moved to D-sets or removed.

Possible Comment by a Reviewer	Possible Resolution
Worng/Missing Glyph	● Glyph is corrected/supplied and move to D-set for eyeball revewing.
Wrong KangXi radical / strokes / first stroke	● Data will be corrected and this Ideograph will be moved to D-set. ● Proposal to correct data is rejected, as it is ambiguous case and IRG agreed that the previous choice of *XX* is more appropriate.
Wrong IDS	● IDS will be corrected and move to D-set until they will be 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.

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Normal

5.2 Guidelines for D-set

Formatted: Normal

D-set ideographs are the ones that cannot be checked automatically by the IDS, or the ones that are suspected to be unifiable with other standardized or working ideographs. For the ideographs that cannot be machine-checked by the IDS, 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 the evidence to show whether such ideograph should be separately encoded or not.

Possible Comment by IDS checker	**Possible Resolution**
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 Resolution**
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.

6. 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 in 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 documentation, the documents should not be compressed and the search engine window should be attached.

Formatted: Normal

Formatted: Left

Formatted: Font: Not Bold

Formatted: Left, Bulleted + Level: 1 + Aligned at: 0" + Indent at: 0.29"

Formatted: Font: Not Bold

Formatted: Font: Italic

Formatted: Font: Not Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Normal

Deleted: 4. Handling defect reports¶
<#>IRG will follow WG2 procedures on reporting of defect according to Annex I and J of WG2 P&P document.¶

Deleted: 5

Annex A: Sorting Algorithm of Ideographs

The ideographs must be sorted by the following order.

- KangXi Radical order.
  - Note: when radicals are the following simplified character, 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

- Number of Strokes
- Whether they are simplified or not. (Simplified characters must be put after the non-simplified characters within the same stroke-number groups.)
- First stroke.

Annex B: IDS matching

B.1 Guidelines on creation of the IDS

Each member body should consult IRG N 1155 for 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⿱采心 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.)

Deleted: Information accompanying submissions
Formatted ... [4]
Formatted ... [5]
Formatted ... [6]
Formatted ... [7]
Formatted ... [9]
Formatted Table ... [8]
Formatted ... [10]
Formatted ... [11]
Formatted ... [12]
Formatted ... [13]
Formatted ... [14]
Formatted ... [15]
Formatted ... [16]
Formatted ... [17]
Formatted ... [18]
Formatted ... [19]
Formatted ... [20]
Formatted ... [21]
Formatted ... [22]
Formatted ... [23]
Formatted ... [24]
Formatted ... [25]
Formatted ... [26]
Formatted ... [27]
Formatted ... [28]
Formatted ... [29]
Formatted ... [30]
Formatted ... [31]
Formatted ... [32]
Formatted ... [33]
Formatted ... [34]
Formatted ... [35]
Formatted ... [36]
Formatted ... [37]
Formatted ... [38]
Formatted ... [39]
Formatted ... [40]
Formatted ... [41]
Formatted ... [42]
Formatted ... [43]
Formatted ... [44]
Formatted ... [45]
Formatted ... [46]
Formatted ... [47]
Formatted ... [48]
Formatted ... [49]
Formatted ... [50]
Formatted ... [51]
Formatted ... [52]
Formatted ... [53]
Formatted ... [54]
Formatted ... [55]
Formatted ... [56]
Formatted ... [57]
Formatted ... [58]
Formatted ... [59]
Formatted ... [60]
Formatted ... [61]
Formatted ... [62]

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 the IDS, or the glyph itself is very complex. IDS matching is not versatile on detection of the unifiable ideographs. 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: Urgent ly Needed Ideographs

C.1 Introduction

When a member body urgently needs very few ideographs to be standardized for the some reason (such as they are Regional or National Standard ideographs) and IRG members approved, the member body may submit the ideographs independent of currently working set to the WG2. ***C.2 Requirements***

An urgently needed ideograph submitter must prepare the following documents.

- ● All the document required by the normal ideograph submissions.
- ● In addition to the above, the document to show any unifiable ideographs in currently working sets against the submitted ideographs
- ● For the ideographs not mentioned above, the doucument to prove that their submitted ideographs are not unifiable with any ideographs in the currently working set. (Proof 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 urgently needed ideograph submitter to check the working set for any unifiable characters against their submission. Failure of doing so may revert the IRG decision for independent submission.

C.3 Dealing with urgent requests

- ● IRG members must check the documents prepared by an urgenly needed ideograph submitter, and if any submitted ideograph is agreed to be unifiable with an ideograph in currently working sets, then such ideograph must be removed from the currently working sets as it will be standardized as an urgently needed ideograph.

Formatted: Normal, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Indent at: 0.25"

Formatted: Normal, Indent: Left: 0.25"

Formatted: Normal, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Indent at: 0.25"

Deleted: (TBD)

Deleted: Work flow and stages of progression

Deleted: (TBD)¶

Deleted: *The IRG working drafts*

Deleted: (TBD)¶

Deleted: *Stages of progression*

Deleted: (TBD)

Deleted: (TBD)¶
<#>For submission with the status of "National" or "Regional" standards, IRG will consider to give priority for processing with consideration of work load incurred.

Deleted: *C.3 Dealing with individual submission to WG2*¶
(TBD)¶
Guideline to deal with individual submission to WG2:¶
<#>small enough set¶
<#>urgent¶
<#>the proposal is sound and stable after exercising due diligence¶
<#>Members' submissions to WG2 for encoding characters in compatibility zone require to go through same unification review of CJK ideographs by IRG¶
<#>the same proposal should be submitted to IRG, with additional information if it might introduce any potential conflicts with IRG working projects.¶

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 will 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

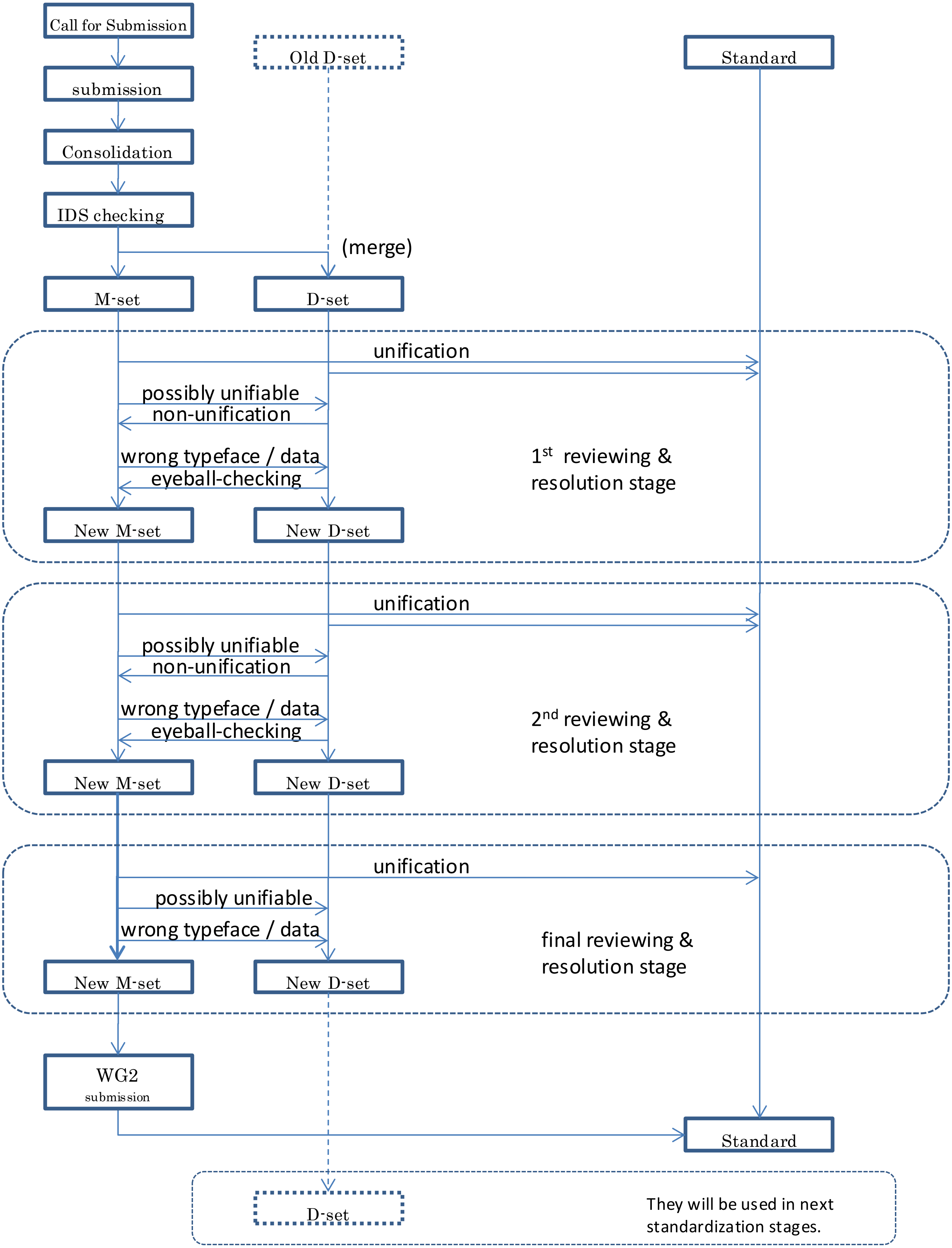
There are many mapping tables, which are included in national/regional standards or developed by third parties. These are out of WG 2's scope. Any organization (such as Unicode Consortium) that collects mapping information, maintains it consistently and makes this information widely available is invited 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

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
unification
possibly unifiable
non-unification
wrong typeface / data
eyeball-checking
1st reviewing & resolution stage
New M-set
New D-set
unification
possibly unifiable
non-unification
wrong typeface / data
eyeball-checking
2nd reviewing & resolution stage
New M-set
New D-set
unification
possibly unifiable
wrong typeface / data
final reviewing & resolution stage
New M-set
New D-set
WG2 submission
Standard
D-set
They will be used in next standardization stages.