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 *N1465Draft*

2008-6-13

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

1. Introduction

This document is a standing document of ISO/IEC JTC 1/SC 2/WG 2/IRG. 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). The document also contains procedures and guidelines for ###. Submitters should check the standard documents (including all the amendments and corrigenda) before preparing new submissions. 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. IRG Scope

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 CJK Unified Ideographs and the set of principles and procedures to be applied in the development of a new repertoire of CJK Unified Ideographs. Standardizing CJK Compatibility Characters maintained in UCS for the purpose of round-trip integrity with other standards is out of IRG scope.

3. 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 General principles - 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.

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.

'戊'(U+620C) and '戉'(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 Preparation for submission to IRG

A. Required data 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.

The following tdata for each proposed ideograph must be submitted with the specified text format (usually in UTF-8) together with its glyph image.

- KXI (KangXi Index with a flag to indicate real or virtual)
- KX Radical Code (KangXi Radical with a flag to indicate simplified or traditional)
- Stroke Count of the Non-radical Component
- First Stroke Code of the Non-radical Component
- Ideographic Description Sequence
- 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.

B. Optional data:

- For questionable characters especially for those candidates with possible unification questions, member bodies 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.
- 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. 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 Production and review process of IRG working drafts

A. Production of IRG working drafts:

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

(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 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.
- 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.

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.

3.4 Preparation for discussions at IRG meetings

A. Unification issues:

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.

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.

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.

3.6 Preparation for submission to WG2

(TBD) Describe the following:

- Preparation of TrueType fonts (fonts have to be available in accordance with the requirement stated in point 5 of A.1 – Submitter's Responsibilities in Annex A, WG2N3452)
- Source references
- Packed Multi-column format
- The IRG should at least conduct one round review of the table generated with TrueType font before submission
- Members are encouraged to review and comment on IRG submissions to WG2. The IRG Rapporteur will forward members' comments to WG2.

4. Handling defect reports

- IRG will follow WG2 procedures on reporting of defect according to Annex I and J of WG2 P&P document.

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 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.

Annex A: Information accompanying submissions

Annex B: IDS matching

(TBD)

B.1 Simple IDS

(TBD)

B.2 Handling of complex or incomplete IDS

(TBD)

Annex C: Work flow and stages of progression

(TBD)

C.1 The IRG working drafts

(TBD)

C.2 Stages of progression

(TBD)

C.3 Dealing with urgent requests

(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.

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