Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC1/SC2/WG2 N

IRG N 379

Date: 1996-07-31

Source: ISO/IEC JTC1/SC2/WG2/**IRG**:

Ideographic Rapporteur Group

Title: Draft Text of CJK Annex

for Extension A

Action: To be submitted to WG2

for Inclusion in ISO/IEC 10646-1

Status: Supporting Paper for IRG #7 (Resolution 3)
Distribution: ISO/IEC JTC1/SC2/WG2 and WG2/IRG

JTC 1/SC2/WG2/IRG N379

DATE: **1996.07.12**

Supersedes:

ISO/IEC JTC 1/SC2/WG2/IRG Ideographic Rapporteur Group Secretariat: China

Source / Contribution Identifier:	tk9627			
Meeting:	@ / for	SC2/WG2 #31		
Issue Number:	Project Numb	per:		
Title: Draft text of CJK annex for Extension A				
Keywords:				
Status :ISO/IEC JTC1/SC2/WG2/I	IRG contribut	ion		
Short Description:				
This draft text is annex of IRG Un When Extension A is agreed the B merged into the IS/HS revision.		nified Ideograph Extension A. by SC2/WG2, this draft text should be		
Proposed Conclusion / Requested Ad	ction:			
To be submitted at SC2/WG2 #31	meeting			
Arguments / Text of Contribution :				
See attachment.				

Contact:

Mr. Zhang Zhoucai, Rapporteur of ISO/IEC JTC 1/SC2/WG2/IRG Center of Computer & Microelectronics Industry Development Research Apartment 620-621, DaTun Road, ChaoYang District, Beijing, 100101, China

Tel: +8610 6201 1118-300 Fax: +8610 6496 0428 E-mail: joezhg@public.bta.net.cn

Annex

(informative)

Procedure for the unification and arrangement of Unified CJK Ideograph Extension A

The graphic character collection CJK UNIFIED IDEOGRAPH EXTENSION A in this Edition of 10646-1 contains 6,585 ideographs (see clause xx). They are derived from over 12,000 ideographs which are found in various different national and regional standards for coded character sets (the "source codes").

This Annex describes how the ideographs in this standard are derived from the source codes by applying a set of unification procedures. It also describes how the ideographs in this standard are arranged in the sequence of consecutive code positions to which they are assigned.

The source code standards are shown below in five groups according to their origins. The groups are identified as the G-, T-, J-, K-, and V-sources.

G-source: GB7589-87*, GB7590-87*,

General Purpose Hanzi List

for

Modern Chinese Language*,

Singapore characters

T-source: TCA-CNS 11643-1992 3rd plane*,

TCA-CNS 11643-1992 4th plane*, TCA-CNS 11643-1992 5th plane*, TCA-CNS 11643-1992 6th plane*, TCA-CNS 11643-1992 7th plane*,

TCA-CNS 11643-1992 15th

plane*

J-source: Unified Japanese IT Vendors

Contemporary Ideographs 1993

K-source: PKS C5700-2 1994 V-source TCVN 5773:1993

(A " * " after the reference number of a standard indicates that some of the ideographs included in that standard are not introduced into the unified collection.)

For the purposes of ISO/IEC 10646-1 a unification process is applied to the ideographic characters taken from the codes in the source groups. In this process single ideographs from two or more of the source groups are associated together, and a single code position is assigned to them in this standard. The associations are made according to a set of procedures that are described below. Ideographs that are thus associated are described here as "unified"

1. Unification procedure

1.1 Scope of unification

Ideographs that are unrelated in historical derivation (non-cognate characters) have not been unified.

士, 土

Example:

NOTE - The difference of shape between the two ideographs in the above example is in the length of the lower horizontal line. This is considered an actual difference of shape. Furthermore these ideographs have different meanings. The meaning of the first is "Soldier" and of the second is "Soil or Earth".

An association between ideographs from different sources is made here if their shapes are sufficiently similar, according to the following system of classification.

1.2 Two level classification

A two-level system of classification is used to differentiate (a) between abstract shapes and (b) between actual shapes determined by particular typefaces. Variant forms of an ideograph, which can not be unified, are identified based on the difference between their abstract shapes.

1.3 Procedure

A unification procedure is used to determine whether two ideographs have the same, or a different, abstract shape. The unification procedure has two stages, applied in the following order:

- a) Analysis of component structure;
- b) Analysis of component features;

1.3.1 Analysis of component structure

In the first stage of the procedure the component structure of each ideograph is examined. A component of an ideograph is a geometrical combination of primitive elements. Alternative ideographs can be configured from the same set of components. Components can be combined to create a new component with a more complicated structure. An ideograph, therefore, can be defined as component tree, where the top node is ideograph itself, and the bottom nodes are the primitive elements. This is shown in Figure 1.

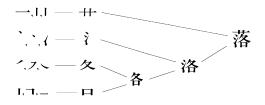


Figure 1 - Component structure

1.3.2 Analysis of component features

In the second stage of the procedure, the components located at corresponding nodes of two ideographs are compared, starting from the most superior node, as shown in Figure 2.

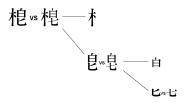


Figure 2 - The most superior node of a component

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

a: the number of components,

b : the relative position of the components in each complete ideograph,

c: the structure of corresponding components.

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

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

1.4 Examples of differences of abstract shapes

To illustrate rules a: to c: in 1.3.2, some typical examples of ideographs that are not unified, owing to differences of abstract shapes, are shown below.

1.4.1 Different number of components

The examples below illustrate rule a: since the two ideographs in each pair have different numbers of components.

崖•厓, 肱•厷, 降•夈

1.4.2 Different relative positions of components

The examples below illustrate rule b:. Although the two ideographs in each pair have the same number of components, the relative positions of the components are different.

峰・峯 荊・荆

1.4.3 Different structure of a corresponding component

The examples below illustrate rule c:. The structure of one (or more) corresponding components within the two ideographs in each pair is different.

1.5 Differences of actual shapes

To illustrate the classification described in 1.2, some typical examples of ideographs that are unified are shown below. The two or three ideographs in each group below have different actual shapes, but they are considered to have the same abstract shape, and are therefore unified.

The differences are further classified according to the following examples.

a) Differences in rotated strokes/dots

半•半,与•勺,羽•羽•羽,酋•酋, 兼•兼,益•益

b) Differences in overshoot at the stroke initiation and/or termination

身·身,雪·雪,拐·拐,不·不, 非·非,周·周,告·告

c) Differences in contact of strokes

奥•奥,酉•酉,児•児,查•查, 奔•奔 d) Differences in protrusion at the folded corner of strokes

e) Differences in bent strokes

f) Differences in folding back at the stroke termination

g) Differences in accent at the stroke initiation

h) Differences in "rooftop" modification

j) Combinations of the above differences

 $\ensuremath{\mathsf{NOTE}}$ - The similar differences listed source code separation

examples in annex S are unified by actual shape differences.

These differences in actual shapes of a unified ideograph are presented in the corresponding source columns for each code position entry in the code table in clause **xx** of this International Standard.

2. Arrangement procedure

2.1 Scope of arrangement

The arrangement of the CJK UNIFIED IDEOGRAPH EXTENSION A in the code table of clause **xx** of this International Standard is based on the filing order of ideographs in the following dictionaries.

riority	Dictio	nary	Edition
1	Kangxi Dictionary	-	Beijing 7th edition
2	Daikanwa Jiten	大漢和辞典	9th edition
3	Hanyu Dazidian	汉语大字典	1st edition
4	Daejaweon	大字源	1st edition

The dictionaries are used according to the priority order given in the table above. Priority 1 is highest. If an ideograph is found in one dictionary, the dictionaries of lower priority are not examined.

2.2 Procedure

2.2.1 Ideographs found in the dictionaries

- a) If an ideograph is found in the Kangxi Dictionary, it is positioned in the code table in accordance with the Kangxi Dictionary order.
- b) If an ideograph is not found in the Kangxi Dictionary but is found in the Daikanwa Jiten, it is given a position at the end of the radical-stroke group under which is indexed the nearest preceding Daikanwa Jiten character that also appears in the Kangxi dictionary.
- c) If an ideograph is found in neither the Kangxi nor the Daikanwa, the Hanyu Dazidian and the Daejaweon dictionaries are referred to with a similar procedure.

2.2.2 Ideographs not found in the dictionaries

If an ideograph is not found in any of the four dictionaries, it is given a position at the end of the radical-stroke group (after the characters that are present in the dictionaries) and it is indexed under the same radical-stroke count.