Information technology—Universal Multiple-Octet Coded Character Set (UCS)—Part 1:
Architecture and Basic Multilingual Plane

AMENDMENT 8

Technologies de l'information—Jeu universel de caractères codés à plusieurs octets—
Partie 1: Architecture et table multilingue
AMENDMENT 8

ICS 35.040

Descriptors: data processing, information interchange, text processing, graphic characters, character sets, representation of characters, coded character sets, architecture

Information technology—Universal Multiple-Octet Coded Character Set (UCS)—
Part 1:
Architecture and Basic Multilingual Plane
AMENDMENT 8

Annex T
(informative)
Procedure for the unification and arrangement of CJK Ideographs

The graphic character collection CJK UNIFIED IDEOGRAPHS in ISO/IEC 10646-1:1993 contains 20,902 ideographs (see clause 26). They are derived from over 54,000 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 four groups according to their origins. The groups are identified as the G-, T-, J-, and K-sources.

G-source: GB2312-80, GB12345-90,
GB7589-87*, GB7590-87*,
GB8565-88*,
General Purpose Hanzi List for
Modern Chinese Language*

T-source: TCA-CNS 11643-1986/1st plane,
TCA-CNS 11643-1986/2nd plane,
TCA-CNS 11643-1986/14th plane*

J-source: JIS X 0208-1990, JIS X 0212-1990

K-source: KS C 5601-1989, KS C 5657-1991

(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”.

T.1. Unification procedure

T.1.1 Scope of unification

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

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.

T.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.

T.1.3 Procedure

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

a) Analysis of component structure;

b) Analysis of component features;

T.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

T.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.

T.1.4 Examples of differences of abstract shapes

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

T.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.

, ,

T.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.

,

T.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.

, ,

T.1.5 Differences of actual shapes

To illustrate the classification described in T.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

, , , , ,

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

, , , , , ,

c) Differences in contact of strokes

, , , ,

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

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 26 of this International Standard.

T.1.6 Source separation rule

To preserve data integrity through multiple stages of code conversion (commonly known as “round-trip integrity”), any ideographs that are separately encoded in any one of the source standards listed above have not been unified.

However, some ideographs encoded in two standards belonging to the same source group (e.g. GB2312-80 and GB12345-90) may have been unified during the process of collecting ideographs from the source group.

T.2. Arrangement procedure

T.2.1 Scope of arrangement

The arrangement of the CJK UNIFIED IDEOGRAPHS in the code table of clause 26 of this International Standard is based on the filing order of ideographs in the following dictionaries.

Priority	Dictionary	Edition
1	Kangxi Dictionary	Beijing 7th ed.
2	Daikanwa Jiten	9th ed.
3	Hanyu Dazidian	1st ed.
4	Daejaweon	1st ed.

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.

T.2.2 Procedure

T.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.

T.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.

T.3. Source code separation examples

The pairs (or triplets) of ideographs shown below are exceptions to the unification rules described in clause T.2 of this Annex. They are not unified because of the source code separation rule described in clause T.1.

NOTE - The particular source code group (or groups) that causes the source code separation rule to apply is indicated by the letter (G, T, J, or V) that appears to the right of each pair (or triplet) of ideographs. The source code groups that correspond to these letters are identified at the beginning of this Annex.

T T

4E1F 4E22 4FF1 5036

GT T

4E48 5E7A 5024 503C

GTJ T

4E89 722D 5077 5078

J TJ

4EDE 4EED 507D 50DE

T T

4F75 5002 514C 5151

T TJ

4FA3 4FB6 514E 5154

TJK T

4FC1 4FE3 5156 5157

T TJ

4FDE 516A 518A 518C

G TJ

51C0 51C8 5433 5434 5449
T T

51E2 51E3 5436 5450
TJ T

5203 5204 543F 544A
TJ T

520A 520B 5527 559E
T T

5220 522A 55A9 55BB
T T

5225 522B 5618 5653
TJ GTJ

5238 52B5 568F 5694
T T

5239 524E 56EF 56FD

T TJ

524F 5259 5708 570F
T T

525D 5265 570E 5713
J T

5292 5294 5716 5717
T T

52FB 5300 5759 5DE0
T J

5355 5358 57D2 57D3
TK T

5373 537D 5848 588D
TJ TJ

5377 5DFB 5861 586B
GT T

53C1 53C2 5897 589E
T GTJ

53C3 53C4 58EE 58EF
T T

5415 5442 58FD 5900
T T

541E 5451 5910 657B
GTJ J

5932 672C 5C02 5C08
J GTJ

5965 5967 5C06 5C07
TJ T

5968 596C 734E 5C13 5C14
GT T

5986 599D 5C19 5C1A
T T

598D 59F8 5C2A 5C2B
T T

59CD 59D7 5C36 5C37
GT T

59EB 59EC 5C4F 5C5B
T GT

5A1B 5A2F 5A31 5CE5 5D22
T T

5A55 5AAB 5DD3 5DD4
T T

5A7E 5AAE 5E21 5E32
TK TJ

5AAA 5ABC 5E2F 5E36
T T

5AAF 5B00 5E76 5E77
T T

5B0E 5B14 5EC4 5ECF
GT T

5B24 5B37 5F11 5F12
T T

5B73 5B76 5F37 5F3A
T T

5BAB 5BAE 5F39 5F3E
T TJ

5BDB 5BEC 5F50 5F51
T T

5BDC 5BE7 5F54 5F55
GTJ T

5BDD 5BE2 5F59 5F5A
J TJ

5F5B 5F5C 634F 63D1
T TJ

5F5D 5F5E 635C 641C
T T

5F65 5F66 63B2 63ED
T TJ

5FB3 5FB7
T T

5FB4 5FB5
TJ TJ

6075 60E0
T T

6085 60A6
T T

609E 60AE
T T

60B3 60EA
T T

6120 614D
TJ T

613C 614E
GT T

6229 622C
T J

622F 6231
T T

6236 6237 6238
T T

623B 623E
T T

629B 62CB
TJ T

629C 62D4
T T

6329 635D
TJ T

633F 63D2 63F7
T T

6985 69B2
T T

699D 6A27
JT T

69C7 69D9
TJ T

69D8 6A23
T T

6A2A 6A6B

63FA 6416 6447

63FE 6435

6483 64CA

654E 6559

6553 655A

65E2 65E3

6602 663B

665A 6669

66A8 66C1

66FD 66FE

67B4 67FA

67E5 67FB

67F5 6805

6961 6986

6982 69EA

6DF8 6E05

6E07 6E34

6E29 6EAB

6E88 6F59

6E89 6F11

T T

6B65 6B69 6EDA 6EFE
T GTJK

6B72 6B73 6F5B 6FF3
T T

6B7F 6B81 7028 702C
GTJ GTJ

6BBB 6BBC 70BA 7232
T GTJK

6BC0 6BC1 712D 7162
T J

6BCE 6BCF 7155 7199
T T

6C32 6C33 7174 7185
T GT

6C5A 6C61 72B6 72C0
TJ TJ

6C92 6CA1 7464 7476
TJ T

6D44 6DE8 74F6 7501
T T

6D89 6E09 7522 7523
T J

6D97 6D9A 75E9 7626
T T

6D99 6DDA 76A1 76A5
T TJ

6DE5 6E0C 771E 771F
TJK T

773E 8846 812B 8131
T T

7814 784F 817D 8183
TJ GT

797F 7984 8203 8204
T TJ

79BF 79C3 820D 820E
T J

7A05 7A0E 8216 8217
TJ TJ

7A42 7A57 8358 838A
J TJ

7B5D 7B8F 83D1 8458

T T

7BB3 7C08 8480 8495
TK J

7BE1 7C12 848B 8523
T T

7CA4 7CB5 848D 853F
T T

7D55 7D76 8570 8580
T T

7DA0 7DD1 85AB 85B0
T T

7DD2 7DD6 85F4 860A
T T

7DE3 7E01 865A 865B
T T

7DFC 7E15 86FB 8715
T TJK

7E48 7E66 885B 885E
TJ TK

7FAE 7FB9 886E 889E
T GJK

7FF6 7FFA 88C5 88DD
T T

80FC 8141 8A2E 8A7D
T TK

8AAA 8AAC 932C 934A
TJ TJ

8ACC 8AEB 93AD 93AE
J T

8B20 8B21 95B1 95B2
T G

8C5C 8C63 9667 9689
TJ T

8D70 8D71 9751 9752
TK GTJ

8EFF 8F27 9759 975C
J J

8F1C 8F3A 976D 9771
T T

8F3C 8F40 9839 983D
T TJ

8FBE 8FD6 984F 9854

TJ J

8FF8 902C 985A 985B
J J

9059 9065 98EE 98F2
T TJ

90A2 90C9 9905 9920
T TJK

90CE 90DE 99B1 99C4
T TK

90F7 9109 9115 99E2 9A08
T T

9196 919E 9AA9 9AAB
J T

91A4 91AC 9AD8 9AD9
T TJ

9203 9292 9AEA 9AEE
T T

92B3 92ED 9B2C 9B2D
T TJ

9304 9332 9C1B 9C2E
T T

9CEF 9CF3 9EBC 9EBD
J T

9D87 9DAB 9EC3 9EC4
J T

9DC6 9DCF 9ED1 9ED2
T

9EAA 9EAB

In accordance with the unification procedures described in T.1 of this Annex the pairs (or triplets) of ideographs shown below are not unified. The reason for non-unification is indicated by the reference which appears to the right of each pair (or triplet). For “non-cognate” see T.1.1

NOTE - The reason for non-unification in these examples is different from the source code separation rule described in clause T.1.6.

non cognate non cognate

5191 80C4 6710 80CA
T1.4.3 non cognate

51B2 6C96 6713 8101
T1.4.3 non cognate

51B3 6C7A 6718 8127
T1.4.3 non cognate

51B5 6CC1 6723 81A7
T1.4.3 T1.4.3

579B 579C 6735 6736
T1.4.2 T1.4.3

5B7C 5B7D 7054 7067
T1.4.3 T1.4.3

5BF3 5BF6 7A32 7A3B
T1.4.1 T1.4.3

5EF0 5EF3 7FF1 7FF6
T1.4.1 T1.4.3

61D0 61F7 8007 8008 8009
T1.4.3 T1.4.1

6560 656A 8074 807C 807D
non cognate T1.4.2

670C 80A6 8346 834A
non cognate T1.4.3

670F 80D0 8EB1 8EB2

Report on JTC1 Letter Ballot on DAM No. 8 to ISO/IEC 10646-1 (New Annex on CJK ideographs)

Disposition of Comments
Responses to the letter ballot on DAM No. 8 appear in JTC1/SC2/N 2791. National bodies that submitted comments are listed below. Where a national body has submitted a negative vote, the indication (N) appears after its name in the list.

Canada (N)

Not accepted. This draft has been approved by a large majority, 19-2; if the negative vote from Japan can be easily resolved, there will be a 20-1 vote in favor. Accordingly, it is not necessary or desirable at this time to carry out "much more work and rewriting" as proposed in paragraph 3 of the comment from Canada. Furthermore, any substantial changes would require a re-ballot, so that other National Bodies could review them, a delay of at least a further nine months.

Only minor editorial or wording improvements can be accepted at this stage, such that there is no risk of objections from other National Bodies who have already approved the text.

Japan (N)

1. Accepted.

2. Accepted.

3. Accepted in principle. The HS/IS repertoire is at present subject to pDAM ballot in JTC1/SC2. The new Annex within DAM.8 will be due for publication at least one year before the HS/IS repretoire has completed the JTC1 ballot stage. Revisions of the new Annex can be considered as a part of that ballot process.

Korea
Accepted.

U.K.
Accepted.

U.S.A.
Accepted.

Note: The text of this Annex has not been revised in accordance with Resolution 9 of IRG meeting 8. Resolution 9 does not reflect the comments of any Member Body voting on DAM.8; although the Editor agrees with the content of IRG Resolution 8-9 it would be improper to edit the text of a Draft Amendment except to accomodate the comments accompanying a Member Body担 vote.

[End of Resolution of Comments]