ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC2/WG2
Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC JTC1/SC2/WG2 N 1943

Date: 1999-01-08

TITLE: Revision of 10646-1 Annex T for CJK Unified Ideographs Extension A (Draft)

SOURCE: Bruce Paterson, project editor, and Takayuki K. Sato

STATUS: Response to Disposition of Comments on FPDAM.17

ACTION: For review by SC2/WG2

DISTRIBUTION: JTC1/SC2/WG2

Introduction

The comments from National Bodies on the ballot on FPDAM.17, CJK Unified ideographs Extension A, included a comment from Japan that Annex T (Procedure for the unification and arrangement of CJK unified ideographs) in Amd.8 of ISO/IEC 10646-1 should be amended to recognise the particular source standards which were used in the preparation of Extension A. These source standards are different from those used in preparing the original set of CJK Unified ideographs in the First Edition of 10646-1.

The Disposition of Comments on FPDAM.17, approved at the London meeting of WG2 in September 1998, stated that an amendment for Annex T would be prepared in future for that purpose (i.e. after consultation with experts from the IRG).

This paper now proposes a set of amendments for Annex T based on those consultations.

Principles of amendments to Annex T

The list of sources of the original set of CJK unified ideographs is shown in clause 26 of 10646 First Edition, and also on page 1 of Annex T. The list of sources for Extension A was previously submitted to WG2 in WG2 N1426. The list of sources for all CJK unified ideographs (original set + Extension A) now appears in clause 26 of FPDAM.17, which will soon replace clause 26 of First Edition.

In preparing the amendments (overleaf) for Annex T, the following principles have been adopted:
- to avoid duplicate lists of sources between clause 26 and Annex T,
- to ensure, as far as possible, that Annex T will not require further amendment if additional blocks of CJK unified ideographs are added to the proposed Part 2 of 10646 in future.

The proposed amendments are shown in the form of extracts from Annex T, showing the subclauses that would be changed, and from Annex L, Sources of characters. Clause 26 of FPDAM.17 is also provided for reference purposes.

The exact number of CJK unified ideographs in Part 1 is 20,902 + 6,582 + 2 = 27,486, but we propose to state only the approximate number, 27,500, in Annex T.

Questions regarding the sources listed in Annex L, and how they relate to the sources listed in clause 26, appear at the end of page 3.

Annex T

(informative)

Procedure for the unification and arrangement of CJK Ideographs

The graphic character collections of CJK unified ideographs in ISO/IEC 10646-1 are specified in clause 26. They contain almost 27,500 ideographs, and ~~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~~ 66,000 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 in clause 26 ~~below~~ in five ~~four~~ groups according to their origins. The groups are identified as the G-, T-, J-, ~~and~~ K-, and V-sources.

[Editor's note: The following deleted text is moved to T1.6 below without change.]

~~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 toT.1.5 remain unchanged.]

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 below ~~above~~ have not been unified.

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.)

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

[T.2 and T.3 remain unchanged.]

Extract from

Annex L, Sources of characters

GB2312-80 *Code of Chinese Graphic Character Set for Information Interchange: Jishu Biaozhun Chubanshe* (Technical Standards Publishing).

NOTE: For additional sources of the CJK unified ideographs in this part of ISO/IEC 10646-1, refer to clause 26.

.....

JIS X 0201-1976 Japanese Standards Association. *Jouhou koukan you fugou (Code for Information Interchange).*

JIS X 0208-1990 Japanese Standards Association. *Jouhou koukan you kanji fugoukei (Code of the Japanese Graphic Character Set for Information Interchange).*

JIS X 0212-1990 Japanese Standards Association. *Jouhou koukan you kanji fugou-hojo kanji (Code of the supplementary Japanese graphic character set for information interchange).*

KS C 5601-1992 Korean Industrial Standards Association. *Jeongbo gyohwanyong buho (Code for Information Interchange).*

[*Editor's notes:*

1. Is this list of sources, taken in conjunction with those in clause 26, sufficient ?

2. The following reference was deleted at Amd.5, Hangul syllables. However it is still listed as source K1 in clause 26. Is this correct?

KS C 5657-1991 Korean Industrial Standards Association. *Jeongho gyohwanyong buho hwakjang saten (Code of the supplementary Korean graphic character set for information interchange).*]

26 CJK unified ideographs

Detailed code tables for:
- CJK Unified Ideographs Extension A (starting at code position 3400), and
- CJK Unified Ideographs (starting at code position 4E00),
are shown on the following pages.

Entries in the code tables for both CJK (Chinese / Japanese / Korean) Unified Ideographs and its Extension A are arranged as follows.

Row/Cell	G		J	K	V
Hex code	G- Hanzi	-T	Kanji	Hanja	ChuNom
078/000	——	——	——	——	——
4E00	0-523B	1-4421	0-306C	0-6C69	1-2121
	0-5027	1-3601	0-1676	0-7673	1-0101

NOTE - Under each ideograph the two lines of numbers indicate the source code positions; the first line shows hexdecimal values, the second line shows decimal values.

The leftmost column of an entry shows the code position in ISO/IEC 10646, giving the code representation both in decimal and in hexadecimal notation.

Each of the other columns shows the graphic symbol for the character, and its coded representation, as specified in a source standard for character sets that is also identified in the table entry. Each of these source standards is assigned to one of five groups indicated by G, T, J, K, or V as shown in the lists below. In each table entry, a separate column is assigned for the corresponding character (if any) from each of those groups of source standards.

An entry in any of the G, T, J, K, or V columns includes a sample graphic symbol from the source character set standard, together with its coded representation in that standard. The first line below the graphic symbol shows the coded representation in hexadecimal notation. The second line shows the coded representation in decimal notation which comprises two digits for section number followed by two digits for position number. Each of the coded representations is prefixed by a one-character source code identification followed by a hyphen. This source code character identifies the coded character set standard from which the character is taken as shown in the lists below.

Hanzi G sources are

G0 GB2312-80

G1 GB12345-90 with 58 Hong Kong and 92 Korean “Idu” characters

G3 GB7589-87 unsimplified forms

G5 GB7590-87 unsimplified forms

G7 General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi

GS Singapore Characters

G8 GB8565-88

GE GB16500-95

Hanzi T sources are

T1 TCA-CNS 11643-1992 1st plane

T2 TCA-CNS 11643-1992 2nd plane

T3 TCA-CNS 11643-1992 3rd plane with some additional characters

T4 TCA-CNS 11643-1992 4th plane

T5 TCA-CNS 11643-1992 5th plane

T6 TCA-CNS 11643-1992 6th plane

T7 TCA-CNS 11643-1992 7th plane

TF TCA-CNS 11643-1992 15th plane

Kanji J sources are

J0 JIS X 0208-1990

J1 JIS X 0212-1990

JA Unified Japanese IT Vendors Contemporary Ideographs, 1993

Hanja K sources are

K0 KS C 5601-1987

K1 KS C 5657-1991

K2 PKS C 5700-1 1994

K3 PKS C 5700-2 1994

ChuNom V sources are

V0 TCVN 5773:1993

V1 TCVN 6056:1995

For CJK (Chinese/Japanese/Korean) Ideographs in the BMP, the names shall be algorithmically constructed by appending their two-octet coded representation in hexadecimal notation to “CJK UNIFIED IDEOGRAPH-”. For example, the first CJK ideograph character in the BMP has the name “CJK UNIFIED IDEOGRAPH-3400”