Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N935

Date: 2002-11-16

Source: Title: Status : Distribution: Medium :	China, HKSAR, MSAR, and TCA Proposal on Basic International Ideograph Subset (BIIS) Joint Proposal IRG Members and Ideographic Experts Electronic

Needs

There are over 70,000 CJK unified ideographs encoded in ISO/IEC 10646. Since there are different demands from vendors, implementers and users of CJK unified ideographs, it is necessary to specify a CJK subset containing daily use ideographs for CJK common use. The objectives for producing such a common set are:

a. to lower the cost for users, to provide conveniences to them and meet the day-to-day need;
b. to meet the demands for international information interchange electronically
c. to encourage countries/regions to apply international standards.

We recognize that different applications may need different subsets, a basic subset is needed currently.

Definition

The CJK international basic subset (hereafter abbreviated to Subset) is a coded character set containing basic and most frequently used ideographs from CJK Unified Ideographs and CJK Unified Ideographs Extension A of ISO/IEC 10646.

Acceptance Criteria

a. The repertoire of the Subset should be stable for a long period of time.
b. The repertoire of Subset should reflect the need of modern daily use. IRG members are required to provide statistics of ideograph-use frequency, which is based on modern publications, such as newspapers, and elementary education texts, with corpus size over 10,000,000 characters (not limited to ideographs only), while submitting their Subset proposals. Ideographs which are represented by phonetic symbols in daily use should carry less weight.

c. The repertoires submitted by IRG members should be generated based on, if any, basic ideograph lists issued by their governments or other authoritative institutions respectively.
d. Ideographs not in basic ideograph lists but nonetheless needed by IRG members will be considered for inclusion in the Subset. It is suggested that such ideographs from each IRG member's submission should be limited to 20 unless strong justifications with high frequency use are provided.
e. Variant forms of simplified ideographs that are in the basic ideograph lists can be included in the Subset too. Therefore, the simplified and unsimplified variants can be in the Subset if they are in any basic ideograph list such as the *General List of Simplified Hanzi*.
f. Only canonical forms of variants confirmed by their submitters can be contained in the Subset.

It is estimated that the Subset should contain less than 10,000 ideographs according to above criteria.

References: basic ideograph lists

Below are some references of China, HKSAR, MacauSAR and TCA. References from other IRG members are needed.

a. China: *General Purpose Hanzi List of Modern Chinese Ideograms* (National Language Committee, Administration of the Press and Publication)
b. China: *General List of Simplified Hanzi* (National Language Committee, 1986)
c. China: GB 2312-80 Chinese Ideograms Coded Character Set for Information Interchange — Basic Set
d. 香港《常用字字形表》(李學銘主編，香港教育學院 2000 年) ➔ English ?
e. 香港：《香港增補字符集-2001 版》(香港特區政府 2001) ➔ English ?
f. TCA-CNS 11643-1992 1st Plane & 2nd Plane