IRG N934

International Basic Subsets of CJK Unified Ideographs

2002/09/30

Japanese members of the Interest Group

1. Purpose

The purpose of the International Basic Subsets of the CJK Unified Ideographs is to provide useful in daily life, interoperable and compact subsets of the CJK Unified Ideographs for those who are more concerned with indispensable smaller parts of modern characters rather than the whole repertoire of ideographs in ISO/IEC 10646 (UCS).

2. Needs

UCS today encodes over 70,000 ideographs to fulfill the needs of various applications and regions, including characters of obsolete, rare and/or non-standard forms in terms of contemporary usage. It is much larger than the number of characters taught in primary and middle schools. An excessive support for such a large number of characters in UCS can bring unnecessary difficulty to deal in the field of education or in a daily life.

While computers today are getting larger capability of processing even the whole UCS repertoire, there's still a need of the limited character repertoire useful for the low-end communication devices and consumer appliances with a small computer which have resource constraints (memory, storage, etc) but must be capable of dealing with the basic character repertoire for localized user interface and/or efficient information exchange.

Without suitable guidelines, the adopted subsets would diverge. It will cause interoperability problems. If a guideline is country/region specific, the interoperability beyond borders would be lost. It will effectively reduce the advantage of UCS over legacy encodings, and retrograde use of country/region specific encodings could revive.

3. Scope

The International Basic Subsets of the CJK Unified Ideographs specify two subsets of ideographs selected from CJK UNIFIED IDEOGRAPHS and CJK UNIFIED IDEOGRAPHS EXTENSION A.

1) LEVEL-1 SUBSET is intended to be the most basic repertoire which contains all characters in the official list for education and daily use issued by the government of each region, primarily for use in the field of education and consumer applications in everyday life.
2) LEVEL-2 SUBSET is a superset of LEVEL-1 SUBSET and extends the repertoire

from more practical point of view that meets to general business needs.

4. Criteria

1) LEVEL-1 SUBSET should contain all characters in the official list for education or daily use of each region. The number of characters in LEVEL-1 SUBSET should be targeted so that it accounts about 3,000 character classes†.

> †Character class is a group of characters, each member of which is 'associated' with (or considered as a variant of) each other. A single character class can be represented by one member character in KangXi Dictionary style.

LEVEL-2 SUBSET is a superset of LEVEL-1 SUBSET and it accounts 7,000 through 8,000 character classes. The key criterion for inclusion to this subset is a degree of functional importance‡.

> ‡The degree of functional importance is a priority of an ideograph by taking account of its use in social context in a specific language environment, word-forming capability as well as published frequency statistics.

2) Example sources for LEVEL-1 and LEVEL-2 subsets.

LEVEL-1:

(Japan) "常用漢字表 (1,945)" + "新人名用漢字 (284)"
+ "人名用漢字許容字体表 (#1: 195, #2: 10)" = 2,434

(China) "常用漢字 (2,500)" *

* The cover average of 2,500 is 97.97%.

(Taiwan) "国民常用字表 (2,408)"

(Korea) "漢文教育用基礎漢字 (1,800)"

LEVEL-2:

(Japan) "JIS X 0208 (6,355)"

(China) "通用漢字表 (7,000)"

(Taiwan) "常用国字標準字体表 (4,808)"

5. Proposed Timeline

LEVEL-1: 2003, Spring: Submit First version to IRG #21 meeting.
2003, Summer: Submit revised edition to WG2.

LEVEL-2: 2003, Autumn: Submit First version to IRG#22 meeting.
2003, Winter: Submit revised edition to WG2.

References

IPSJ-TS 0005:2002, Basic Subset of Coded Character Sets (BUCS)