JTC1/SC2/WG2/IRG N947

Date : 2002-11-19

ISO/IEC JTC1/SC2/WG2/IRG

Ideographic Rapporteur Group

(IRG)

Source/Contribution Identifier :	Ideograph subset ad hoc
Meeting :	20th IRG Meeting in Hanoi, Vietnam
Title :	Report to WG2 on Creating an Ideographic Subset Within ISO/IEC 10646
Status :	For submission from the IRG to WG2

1. The need for an ideographic subset

Ideographs have always been an integral part of ISO/IEC 10646, and the encoded repertoire continues to be extended. The goal for encoding ideographs is to be as complete as possible because of the standard's "universal" nature, so that different applications can be developed within the framework of ISO/IEC 10646. However, because a functional set for day-to-day use is only a few thousand characters in size, a system supporting the complete set can create problems for many users and developers. Users must confront large lists of generally unfamiliar characters when using input methods, for example. There are also large numbers of non-unifiable but potentially synonymous variants that need to be supported. Other forms of input such as pen-based input or OCR become considerably more complex when they are expected to support tens of thousands of characters. It is no longer possible for TrueType fonts to provide complete coverage of the entire set of ideographs within a single font file. In practice, it is important that we provide a functional subset whose implementation provides a good balance between efficiency as well as comprehensiveness for daily use.

Within the ISO/IEC 10646 framework, named subsets are not new concepts. Many countries and regions already have specified subsets of 10646. In East Asia, small government lists of ideographs for elementary education or computer use are common.

Having an ideographic subset would not in any way limit the implementation and use of the complete set of ISO/IEC 10646 Unified Ideographs. The intention is rather to provide for hardware and software optimization, not to exclude support for ideographs beyond the basic set.

The IRG believes that one subset is better than two for the moment. If two subsets are identified, there seems little motive for anyone to implement the smaller set.

This is intended to be a general-purpose subset. Specialized needs may require other, perhaps smaller subsets which are not covered in this proposal.

2. The need for a common subset

The main reason for having a common subset of core ideographs is the same as that for ISO/IEC 10646 itself: in a global computing environment, a common, standardized set means lower development and production costs. This includes work done by hardware manufacturers, system software designers, and application developers.

A common set of basic characters also means that as users travel from one region to another, they can be assured of consistent local support for the characters most important to them. Data also travels world-wide, and a common subset would mean that it does so consistently across platforms and software packages.

It will also be possible for font vendors to provide a single font file with locale-appropriate glyphs for all the major locales.

3. Why the basic ideographic subset should be a part of ISO/IEC 10646

If such a subset is to be developed and is to be non-proprietary, the IRG is the best organization to do the work of generating the list of characters to include. Other organizations do not have as much international scope, experience, or expertise as the IRG.

ISO/IEC 10646 is seen as the basis for character repertoires to be used in international products. As such, a subset of ideographs would seem a natural part of the data provided as part of the standard. This is particularly true since the basic problem this proposal seeks to solve—the large and unmanageable set of encoded ideographs—is the result of ISO/IEC 10646's universal nature, and the solution to the problem would therefore be best coming from the same source.

Users and developers are worldwide so the subset should come from an international body. That the subset comes from an international body gives developers more assurance and incentive to support it, particularly where governments use international standards for national procurement. This maximizes interoperability and minimizes cost.

4. Considerations for the contents of the common subset

The common subset should be geared at providing support for uses such as:

1. Elementary education
2. Unspecialized, general publications such as newspapers, magazines, and novels
3. Textbooks for primary and middle schools
4. Frequently used colloquial (spoken) characters
5. The most common personal names and place names

This will not cover 100% of all usage in these categories but should give sufficient coverage for everyday needs.

Names of people and places can be very difficult to support, as there are not authoritative lists of all the names in current use throughout East Asia, and personal names are a common source for rare, unusual, or variant forms. At the same time, major places (provinces, prefectures, and other large administrative regions) should be accommodated, with the understanding that not every village or street will have its name included in the collection.

Similarly, the personal names included should provide coverage for the large majority of the population, with an understanding that the basic subset will be insufficient for specialized uses that require complete coverage, such as insurance or census databases, phone books, and so on.

5. Size considerations

The work on the subset should be aimed at an overall collection size, and not a predetermined percentage of current usage to cover.

Among the national and regional standards in current use which provide for the bulk of common use, we find the following sizes:

G0 (level 1):	3755
T1:	5412
HKSCS:	4818
J0 (levels 1 and 2):	6356
K0:	4620

From the government lists for primary and secondary education, we find the following character counts:

China:	~3500
Taiwan:	4808
HKSAR:	4759
Japan:	1945
Korea:	1800

The intersection of these sets consists of about 2500 characters and their union of about 6300. It is therefore anticipated that a collection of 6500–7000 characters would be adequate to cover the needs of the standard subset.

Frequency alone is a dangerous criterion to use for inclusion of characters. Some important characters in primary education, for example, are rare in general publications. Moreover, the actual frequency of a character in general use is highly sensitive to the nature of the corpus being considered. Characters common in newspapers may be uncommon in secondary school textbooks, for example.

Nonetheless, from frequency considerations of simplified Chinese newspapers, we find that 2500 characters covers 97.97% of current needs and 3500 raises this to 99.48%. A collection of 6500—7000 characters should therefore cover the vast majority of needs for most modern text. It is not anticipated that the subset will go beyond 10,000 characters.

6. Development Plan

The ideographic subset should not be generated by simply incorporating existing encoded character sets wholesale. Important characters may be left out by this approach, and less useful characters included.

Each IRG member will start with two collections: Level 1 of its encoded character set, and any standard list of characters used in education, preferably government-issued. The union of the two will form the first approximation for the member's submission. The IRG member will then subtract from and add to this list.

Only characters currently encoded in ISO/IEC 10646 may be included. IRG members are strongly discouraged from including characters from Extension B. Such characters will be subject to much stricter review than characters from the BMP.

Justifications for inclusions to the basic list should be available. The justifications need not be detailed. For example, a member may justify characters using the following format:

Place names: U+6C39 氹

People names: U+8340 荀

Colloquial: U+4E5C 乜

By March 2003, the editorial committee representative for each IRG member will send to the rapporteur a list of the sources it will be using for its list of ideographs, together with an estimate for the size of their final list. The final list must be submitted at least one month before IRG #21 so that a merged list can be produced. The IRG will then review the result at the meeting.

The submission should consist of a text file, with an entry for one character per line. The line consists of seven bytes:

Bytes 1 through 4 are the ISO/IEC 10646 code point.

Byte 5 is the source ID (G, T, J, K, V, D, H, M, S, U)

Byte 6 should indicate source subID, such as "0" for G-0, "7" for T-7, and so on. Sources which are not subdivided should use "1".

Byte 7 should be one of the following, depending on where the character comes from:

A for level 1 of the source encoded character set
B for education
C for level 2 of the source encoded character set
D for personal names

E for place names
F for colloquial characters
G for anything else

(Any submissions including characters from Extension B should be in a separate file with eight-byte lines, the first five being used for the ISO/IEC 10646 code point.)