ISO/IEC JTC1/SC2/WG2 N 3229

Date: **2007-04-03**

ISO/IEC JTC1/SC2/WG2
Coded Character Set
Secretariat: Japan (JISC)

Doc. Type: **Input to ISO/IEC 10646**

Title: **Proposal for a new edition of ISO/IEC 10646**

Source: **Project Editor**
Project: JTC1 02.18
Status: For review by WG2
Date: 2007-04-03
Distribution: WG2
Reference: WG2 N3230
Medium:

This document describes a proposal for a new edition of ISO/IEC 10646 and should be evaluated with the accompanying document WG2 N3230.

With now four amendments being processed against the last edition of ISO/IEC 10646:2003, it becomes increasingly difficult to read the standard. This on itself would justify the creation of a new edition to reflect the consolidated content.

In addition, the synchronization with the Unicode Standard presents a mix of challenges and opportunities:

Terminologies used by the two standards are not well aligned

ISO/IEC 10646:2003 uses the concept of UCS form which is sometimes interpreted as abstract notation (canonical form), memory representation, or serialized representation. Some of these forms use the 'UCS Transformation Format' moniker without clear description what it entails.

On another hand, the Unicode Standard has separated these concepts using code point mapped into a codespace to describe abstract characters, encoding forms for in-memory representation, and encoding scheme for serialization.

To describe the coding of characters (clause 6), ISO/IEC 10646:2003 uses a segmented view of characters into multi-octet sequences (G-P-R-C for Group-Plane-Row-Cell) which is unnecessarily complicated for what is in essence a 32-bit code unit. In addition with all the coding space beyond 10FFFF permanently reserved, it would be much easier to describe the coding space as a range extending from 0000 to 10FFFF.

In addition, serialized aspects of the coded representations are intermixed in various parts of the standard without clear separation between when a character coding needs to be serialized or not. If anything, amendment 3 with the introduction of new serialized encoding (UTF-32, UTF-32LE and UTF32-BE) makes the matter worse.

Essential part of the standards separated in annexes

ISO/IEC 10646:2003 describes UTF-8 and UTF-16 in annex D and C respectively which may give the impression that these forms are not as important. In fact, UTF-8 is the preferred encoding forms for many

applications and protocols in IETF and UTF-16 is widely implemented by operating systems. It would seem wiser to bring back these two UTFs into the main body of the standard.

Data set expressed through non machine readable list

Several key concepts of the standard (Combining characters and mirrored characters) are maintained by enumerated lists of characters that are not machine readable and are hard to maintain. For these concepts, the Unicode Standard simply maintains a set of properties available through machine readable files.

Lack of details in the name list

The ISO/IEC 10646:2003 in its clause 34 describes the character glyphs and names in a format that does not allow for extra comments and references short of a simple terse annotation. This creates the need for annex P which contains additional information about character. The solution used by the Unicode Standard is more flexible as it allows the same information and much more to be presented in the chart section.

It would also simplify the production of both standards to have a common format for this part. It would also remove the need for annex P, except for maybe some few entries such as the CJK entries which btw are incomplete (only 2 characters where 12 would be needed).

Proposal

The document WG2 N3230 represents a prototype of what could be a new edition of ISO/IEC 10646. It preserves the overall current structure of the standard but modifies the terminology along the principles mentioned above. It incorporates text from the four amendments (although obviously the amendment 3 and 4 are still ongoing so these parts are only tentative).

There are at least two different ways to make this happens:

- Treat this is as a regular amendment incorporated into other text changes. It may be however productive to provide both editing instructions and the final result because the changes are important. This would probably be the least disruptive.
- Create a CD or FDIS (preferred) with this text and process it in parallel with current amendments.

There are also other minor issues to be resolved, such as preserving the dual column presentation as opposed to the single column which may be more suitable for most of the content (in other words treating dual column as an exception instead of the default).

Finally, although the proposed revision may look quite different from the original, it is interesting to note that only three terms have completely disappeared from the new document: group, RC-element, and zone. Other names have been slightly modified or preserved while new terms have been introduced for clarification purpose.
