SC2/WG2 N4544R
L2/014-0xx
Date: 2014-02-26

Title: Representation of CJK ideograph glyphs
Source: Michel Suignard, ISO/IEC 10646 Project Editor
Distribution: UTC, WG2

Summary: This document proposes clarification for the status of the nominal glyph representations for CJK Ideographs and allowing these glyphs to be updated while maintaining the original source reference information.

1. Current status and issue statement

According to the text included in the clause 1 of ISO/IEC 10646, this International Standard 'defines a set of graphic characters used in scripts and written form of languages on a world-wide scale'.
This definition has been interpreted for most of the blocks shown in the code charts as making sure that the graphic symbols displayed in these charts represent the modern graphic representation of these characters. This obviously does not apply to historic repertoires.

There is however a glaring exception for the CJK Unified Ideographs where it has been accepted practice to allow showing the graphic symbols as they were when the characters were originally encoded. There is a note in sub-clause 23.1 List of source references that hints at the principle:

> NOTE 2 – Even if there is a new version of the source publication, the existing source reference information in the data files will not be updated. The updated source may only identify characters not previously covered by the older version.

Interestingly enough, although the note creates a 'principle', by being informative in nature, it has no 'teeth'. It is also largely ignored by many constituencies. Many of them have updated their source references either by defining new 'sources' or by updating the existing sources. It is important to realize that sources are not just a set of numerical data, but they also commonly specify graphic symbols for the characters included in the reference. Therefore, these updates have resulted in graphic symbols updates for characters referenced by these sources which have been reflected in recent ISO/IEC 10646 code charts.

Other constituencies have adhered to the 'Note 2' principle by preserving the historic nature of the standard. A good example is the set of 168 characters that are referenced in the Japanese JIS X 0213:2004 standard as having different prototypical glyphs from the 2000 version of the standard. But because the 2000 version is the one referenced by ISO/IEC 10646, the standard has not updated these glyphs. (In fact the situation is a bit more complicated, the characters were originally referenced by JIS X -208-1990 itself updated in 1997; JIS X 2013, while not formally containing a reference for these 168 characters nevertheless updated their graphic representations in its 2004 version, this is at least the understanding of the author).

Example for U+9022, ISO/IEC 10646 J column:  Modern Japanese representation: 

The problem with the historic representation principle is that it prevents the standard to be used as a modern up-to-date reference for these characters. Any time that the formal content of ISO/IEC 10646 is used to represent these characters it shows an obsolete version which is not anymore in use in modern computing platforms.

For example, ICANN (Internet Corporation for Assigned Names and Numbers) has launched a project to create Label Generation Rules for the Internet Domain Root Zone. The project involves creating PDF documents describing allowed characters for root domain labels. These documents reference Unicode and ISO/IEC 10646 repertoire containing these 168 characters with the Japanese source references. As of now, they do not represent their modern version. This is clearly less than optimal.

2. Proposed solution

The text of the standard should be modified to favor modern representation of the characters while allowing a clear description of the history that led to their encoding. A great resource is the set of previous versions of the standard which can naturally describe that history. The solution requires two part:

1) A modification of the Principles and Procedures to create a process to update graphic symbols when these are updated by new source references.
2) Implement text changes in the standard for cases where there is already a need to implement that process.

2.1 Change in Principles and Procedures

That document should clearly mention that encoded characters should be graphically represented following their latest version. Stability of source references should be reinforced by adding explicit terms in the standard. For example, the existing Note 2 in sub-clause 23.1 List of Source references should become a principle and documented in both the P&P document and the standard.
At the same time, it should be possible to documents cases when significant graphic updates have been specified for already encoded characters. If in those cases the sources references are not updated, terms should be added to the standard to document allowing graphic symbols to be updated while still using the previous source reference data. For example, a collection can be created to identify these updated characters.

2.2 Change in the standard itself

The Note 2 in sub-clause 23.1 (mentioned above) is removed, replaced by a new sub-clause inserted in the clause 23 Source reference for CJK Ideographs. It contains a general principle and an enumeration of special cases. Because there is only one case so far, a list is unnecessary at this point.

> **23.2 Revision and updating of source references**
>
> Even if there is a new version of the source publication, the existing source reference information in the data files is not updated. The updated source only identifies characters not previously covered by the older version.
>
> The collection 289 JAPANESE JISX2004 UPDATED IDEOGRAPHS contains 168 characters that were graphically updated by JIS X 0213:2004 but encoded in this International Standard as part of JIS X 0208-1990 and JIS X 0213-2000. The source reference data maintains the original

information, but the graphic symbols in the code chart contains the JIS X 0213:2004 updated representations. The graphic symbols corresponding to JIS X 208:1990 for these updated characters are available in the third and prior editions of this International Standard.

NOTE – These graphic symbols can also be found in version 7.0 and prior of the Unicode Standard (see Annex M).

Then in Annex A, the collection 289 should be introduced in clause A.1 and described in a new CJK collection sub-clause (A.4.4). The content is shown below:

U+5026	U+6062	U+6DEB	U+79E4	U+85F7	U+905C
U+50C5	U+6108	U+6EA2	U+7A17	U+8654	U+9061
U+5132	U+6241	U+6EBA	U+7A7F	U+86F8	U+912D
U+514E	U+633A	U+6F23	U+7AC8	U+8703	U+914B
U+51A4	U+633D	U+7015	U+7B08	U+8755	U+91DC
U+537F	U+6357	U+701E	U+7B75	U+87F9	U+9306
U+53A9	U+6372	U+7026	U+7BAD	U+8805	U+9375
U+53C9	U+63C3	U+7058	U+7BB8	U+8956	U+939A
U+53DB	U+647A	U+7078	U+7BC7	U+8A0A	U+9453
U+53DF	U+64B0	U+707C	U+7BDD	U+8A1D	U+9699
U+54AC	U+64E2	U+7149	U+7C3E	U+8A3B	U+9744
U+54E8	U+65A7	U+714E	U+7C7E	U+8A6E	U+9771
U+55B0	U+6666	U+717D	U+7C82	U+8AB9	U+9784
U+5632	U+6753	U+723A	U+7FEB	U+8AFA	U+9798
U+5642	U+6756	U+724C	U+7FF0	U+8B0E	U+97AD
U+564C	U+6897	U+7259	U+8171	U+8B2C	U+98F4
U+56C0	U+68D8	U+727D	U+817F	U+8C79	U+9905
U+5835	U+6962	U+72E1	U+818F	U+8CED	U+990C
U+5A29	U+696F	U+7337	U+8258	U+8FBB	U+9910
U+5C51	U+698A	U+7511	U+8292	U+8FBF	U+9957
U+5C60	U+6994	U+7515	U+82A6	U+8FC2	U+99C1
U+5C62 *	U+69CC	U+7526	U+8328	U+8FC4	U+9A19
U+5DF7	U+6A0B	U+75BC	U+845B	U+8FE6	U+9BAB
U+5E96	U+6A3D	U+77A5	U+84EC	U+9017	U+9BD6
U+5EDF	U+6A59	U+7941	U+8511	U+9019	U+9C2F
U+5EFB	U+6ADB	U+7947	U+853D	U+9022	U+9C52
U+5F98	U+6B4E	U+795F	U+85A9	U+903C	U+9D09
U+5FBD	U+6C72	U+79B0	U+85AF	U+9041	U+9D60

(* 167 characters came from JIS X 0208:1990, 1 character: U+5C62 came from JIS X 0213:2000.)

Finally, the code charts for the J column of these characters should have the updated graphic characters corresponding to JIS X 213:2004.

While these recommendations apply to the CJK Ideographs they could be extended to other repertoires if needed.