ISO/IEC JTC1/SC2/WG2 N4620
L2/14-149 *Date: 2014-07-15*

Title: CJK ideograph glyphs representation and sources references

Source: Michel Suignard, ISO/IEC 10646 Project Editor

Distribution: UTC, WG2

Summary: This document proposes to clarify the identification of CJK Ideographs, both in term of glyphs representation and sources references. If adopted, it would allow the Unicode Standard and ISO/IEC 10646 to use updated glyphs and source references when industry practice and National Standards update their own information. The document addresses issues with CJK Ideographs with Japanese sources (J source), but its recommendation can be applied to other constituencies.
The document is an update of the document WG2 N4544R. The previous document only covered the glyph representation aspect.

Acknowledgement: This work would not have been possible without the information provided by Dr. Ken Lunde from Adobe. His publications, including recent blogs posting mentioned in references, were extremely valuable in creating this document.

1. Current status and issue statement

According to the text included in the clause 1 of ISO/IEC 10646 (4th edition), 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 an 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.

The note refers only to the 'source reference', not the graphic representation. But the effect is the same, for CJK Ideographs, sources references always contain a graphic representation and by maintaining the historic source reference, the historic glyph representation is also preserved.

Interestingly enough, although the note creates a 'principle', it is only informative. It is also largely ignored by many constituencies which have updated their source references either by defining new 'sources' or by updating the existing sources. Therefore, these updates have resulted in graphic symbols updates and source 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 at the price of a complicated status. A good example concerns the Japanese source references.

In ISO/IEC 10646 and Unicode the Kanji J sources are specified as follows:

J0	JIS X 0208-1990
J1	JIS X 0212-1990
J3	JIS X 0213:2000 level-3
J3A	JIS X 0213:2004 level-3
J4	JIS X 0213:2000 level-4
JA	Unified Japanese IT Vendors Contemporary Ideographs, 1993
JH	Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009
JK	Japanese KOKUJI Collection
JARIB	Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

Glyphs are shown in the charts as originally specified by these sources.

However JIS X 0213:2004 has changed deeply the status of character encoding in Japan. It updated a total of 205 glyphs from previous standard and created 2828 new source references for characters encoded in previous standards, summarized as follows:

- 167 characters from J0 had their glyph changed (no source change)
- 31 character from J1 had their glyph changed and new sources created
- 7 characters from JA had their glyph changed and new sources created
- 2743 characters from J1 had new sources created (along for some a glyph change, covered above)
- 85 characters from JA had new sources created (along for some a glyph change, covered above)

Finally, the source designation J3A is a bit misleading because it actually refers to 8 level-3 characters that were added between JIS X 0213:2000 and JIS X 0213:2004 (there in fact a total of 10, but the 2 remaining were already specified in previous standards). The level-3 character set in the 2004 version is a superset of the 2000 version.

Needless to say, the current status has created some confusion.

- Many modern use Kanji characters, including 167 out of the 6536 J0 basic set, have outdated representation in the standard.

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

- The Japanese source reference data, because of the parallel glyph change, is obsolete in two fronts (source data and glyph). Furthermore, the complexity of the changes introduced by JIS X 0213:2004 which are only partially included in ISO/IEC 10646 makes the situation even worse.

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 graphic representation which is not anymore in use in modern computing platforms. Furthermore, some source references are partially obsoleted by being superseded in a more recent version JIS X 0213:2004.

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 many of these 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. While previous versions of the ISO/IEC 10646 standard would be the natural way to retrieve that history, the ISO process makes them unavailable. Therefore another alternative has to be designed. 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. In addition, when a new standard also clarifies and simplifies the source reference context it should be possible to use it.
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 CJK ideographs should be graphically represented following their latest version as published in national standards. Stability of source references should be explained in more details in the document, allowing them to be updated when it would result in a simplification of their description without technical change in the encoded characters.

2.2 Change in the standard itself

The Note 2 in sub-clause 23.1 (mentioned above) was already removed in Amendment 1 of the 4th edition of ISO/IEC 10646.
In addition, the description of the Kanji J sources can be updated in the same sub-clause by using JIS X 0213:2004 when it supersedes previous versions as follows:

J0	JIS X 0208-1990
J1	JIS X 0212-1990
J3	JIS X 0213:2004 level-3
J3A	JIS X 0213:2004 level-3 addendum from JIS X 0213:2000 level-3
J4	JIS X 0213:2004 level-4
JA	Unified Japanese IT Vendors Contemporary Ideographs, 1993
JH	Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009
JK	Japanese KOKUJI Collection
JARIB	Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

(all sources are preserved with a clarification for J3A and a date change for J3 and J4)

Two new notes can then be added to describe the status of the various Kanji J sources.

> NOTE 3 – When a more recent source such as JIS X 0213:2004 includes encoded characters already included in older sources, the former source references are used to reflect contemporary usage. This also reflects the fact that some of these more recent reference were accompanied by minor adjustment of the graphic representation of the related characters.
>
> NOTE 4 – J Sources that have been partially superseded can be found as provisional sources in the Unihan Data base in http://www.unicode.org/reports/tr38/ fully described as kJis1 for JIS X 0212-1990 and kJA for the JA source. The superseded graphic representations can also be found in version 7.0 and prior of the Unicode Standard (see Annex M)

In sub-clause 23.2 Source references file for CJK Ideographs, the linked file CJKSrc.txt should be updated with the updated 2828 source references. The updated values for these references is provided in a file as an attachment to this pdf file.

Finally, the code charts for the J column of these characters should have the updated graphic characters corresponding to JIS X 213:2004. Adobe has offered to provide the font 'Heisei Mincho StdN-W3' for the 205 glyph updates. The font was designed to precisely reflect the glyphs recommended by JIS X 0213:2004. To avoid further regression risk, the font would be only used for these 205 code points. It is possible to extend the range further if so desired.

Open to debate is whether it is necessary to identify the 205 characters with J sources which have updated graphic representations. If adopted, 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+382F [A]	U+5E96	U+6A59	U+75BC	U+82A6	U+9017
U+4105 [A]	U+5ECB [1]	U+6ADB	U+7608 [1]	U+8328	U+9019
U+42C6 [A]	U+5EDF	U+6B4E	U+7626 [1]	U+845B	U+9022
U+459D [A]	U+5EFB	U+6BA9 [1]	U+76D4 [1]	U+845C [1]	U+903C
U+484E [A]	U+5F98	U+6C72	U+77A2 [1]	U+84EC	U+9041
U+4B3B [A]	U+5FBD	U+6C74 [1]	U+77A5	U+8511	U+905C
U+4C17 [A]	U+6062	U+6DEB	U+7934 [1]	U+853D	U+9061
U+5026	U+6108	U+6EA2	U+7941	U+85A9	U+912D
U+50C5	U+6241	U+6EBA	U+7947	U+85AF	U+914B
U+50F2 [1]	U+633A	U+6F23	U+795F	U+85F7	U+91DC
U+5132	U+633D	U+6FF9 [1]	U+79B0	U+8654	U+9306
U+514E	U+6357	U+7015	U+79E4	U+86F8	U+9365 [1]
U+51A4	U+6372	U+701E	U+7A17	U+8703	U+9375
U+537F	U+63C3	U+7026	U+7A7F	U+8755	U+938B [1]
U+53A9	U+647A	U+7058	U+7AC8	U+87F9	U+939A
U+53C9	U+64B0	U+7078	U+7B08	U+8805	U+93A1 [1]
U+53DB	U+64E2	U+707C	U+7B75	U+8956	U+9453
U+53DF	U+65A7	U+7149	U+7BAD	U+8A0A	U+9699
U+54AC	U+6666	U+714E	U+7BB8	U+8A1D	U+96DA [1]
U+54E8	U+6677 [1]	U+717D	U+7BC7	U+8A3B	U+9744
U+55B0	U+6680 [1]	U+71B3 [1]	U+7BDD	U+8A6E	U+9771
U+5632	U+6753	U+723A	U+7C3E	U+8AB9	U+9784
U+5642	U+6756	U+724C	U+7C69 [1]	U+8AFA	U+9798
U+564C	U+6897	U+7259	U+7C7E	U+8B0E	U+97AD
U+56C0	U+68D8	U+727D	U+7C82	U+8B2C	U+98F4
U+5835	U+6962	U+72E1	U+7D5C [1]	U+8C79	U+9905
U+5A29	U+696F	U+7337	U+7FEB	U+8C9B [1]	U+990C
U+5ADA [1]	U+698A	U+7462 [1]	U+7FF0	U+8CED	U+9910
U+5C51	U+6994	U+74D8 [1]	U+8171	U+8FBB	U+9957
U+5C60	U+69CC	U+74EF [1]	U+817F	U+8FBF	U+99C1
U+5C62 [1]	U+69FE [1]	U+7511	U+818F	U+8FC2	U+9A19
U+5DB2 [1]	U+6A0B	U+7515	U+8258	U+8FC4	U+9B2D [1]
U+5DF7	U+6A3D	U+7526	U+8292	U+8FE6	U+9BAB

U+9BD6	U+9C52	U+9D09	U+9DBF [1]
U+9C2F	U+9CE6 [1]	U+9D60	

(167 characters came from JIS X 0208:1990, 7 characters followed by 'A' are originated from the JA source, 31 characters followed by '1' came from JIS X 0212:1990.)

2.2 Change in the Unicode Standard

The update in Unicode should parallel the changes made in ISO/IEC 10646 in term of source references and charts. And because Unicode does not remove access to former versions, when a new one is created it is easier to document past practice.

To that effect, to make the text in the Note 4 in sub-clause 23.1 from ISO/IEC 10646 accurate, the Unihan database should have a new provisional kJA field created to include all JA sources used in the standard (660 entries).

3. References

CJKV Information Processing, 2nd Edition, 2008, Ken Lunde, http://shop.oreilly.com/product/9780596514471.do

Blog articles from Ken Lunde:
JIS X 0212 versus JIS X 0213: http://blogs.adobe.com/CCJKType/2014/04/jisx0212-vs-jisx0213.html
JIS X 0213 versus kIRG_JSource: http://blogs.adobe.com/CCJKType/2014/04/jisx0213-vs-jsource.html
JIS X 0213 versus kIRG_JSource—Redux : http://blogs.adobe.com/CCJKType/2014/04/jisx0213-vs-jsource-redux.html

Appendix
The following tables show the previous and new graphic representation for these characters.

7 Characters originally from JA sources (now J3 and J4)

UCS	Old glyph	New glyph	Old source	New source
382F	㕣	㕣	JA-2256	J4-286F
4105	䄅	䄅	JA-2537	J4-7264
42C6	䋁	䋁	JA-2567	J4-742B
459D	䖒	䖒	JA-2657	J3-7B51
484E	䠏	䠏	JA-272F	J4-795C
4B3B	䢅	䢅	JA-2772	J4-7C55
4C17	䱽	䱽	JA-2822	J3-7E3E

31 Characters originally from J1 sources (now J3, J3A and J4)

UCS	Old glyph	New glyph	Old source	New source
50F2	傰	傰	J1-3251	J3-2E46
5ADA	嫚	嫚	J1-3A25	J4-256A
5C62	屢	屢	J1-3A7A	J3-4F60
5DB2	嵩	嵩	J1-6674	J4-2864
5ECB	廋	廋	J1-3C52	J3-742F
6677	晷	晷	J1-4247	J3-7540
6680	晬	晬	J1-424B	J4-2E2A
69FE	槾	槾	J1-4471	J4-2F4A
6BA9	瘵	瘵	J1-462B	J3-7647
6C74	汴	汴	J1-4664	J3-7654
6FF9	澴	澴	J1-4927	J3-7739
71B3	熳	熳	J1-4A2F	J4-7021
7462	瑢	瑢	J1-4C42	J3-7833
74D8	瓏	瓏	J1-4C6F	J4-712A

UCS	Old glyph	New glyph	Old source	New source
74EF	甌	甌	J1-4C78	J4-712C
7608	瘈	瘈	J1-4D68	J3-7852
7626	瘦	瘦	J1-4D77	J3A-7E7D
76D4	盬	盬	J1-4E4F	J3-786A
77A2	瞢	瞢	J1-4F34	J4-7230
7934	礴	礴	J1-5049	J3-7932
7C69	籩	籩	J1-5324	J4-736F
7D5C	絜	絜	J1-5368	J3-7A24
845C	葜	葜	J1-585B	J4-764D
8C9B	貜	貜	J1-5F27	J4-792A
9365	鍥	鍥	J1-6474	J3-7D3D
938B	鎈	鎈	J1-6529	J4-7B39
93A1	鎷	鎷	J1-6531	J4-7B35
96DA	雚	雚	J1-6676	J4-7B73
9B2D	鬬	鬬	J1-6A32	J3-7E3F
9CE6	鴕	鴕	J1-6B59	J4-7E21
9DBF	鷙	鷙	J1-6C37	J4-7E3E

167 Characters with J0 sources (source unchanged)

UCS	Old glyph	New glyph
5026	倦	倦
50C5	僅	僅
5132	儲	儲
514E	兎	兎
51A4	冤	冤
537F	卿	卿

UCS	Old glyph	New glyph
53A9	厩	厩
53C9	叉	叉
53DB	叛	叛
53DF	叟	叟
54AC	咬	咬
54E8	哨	哨

UCS	Old glyph	New glyph
55B0	喰	喰
5632	嘲	嘲
5642	噂	噂
564C	噌	噌
56C0	囀	囀
5835	堵	堵

UCS	Old glyph	New glyph
5A29	娩	娩
5C51	屑	屑
5C60	屠	屠
5DF7	巷	巷
5E96	庖	庖
5EDF	廟	廟
5EFB	廻	廻
5F98	徘	徘
5FBD	徽	徽
6062	恢	恢
6108	愈	愈
6241	扁	扁
633A	挺	挺
633D	挽	挽
6357	捗	捗
6372	捲	捲
63C3	揃	揃
647A	摺	摺
64B0	撰	撰
64E2	擢	擢
65A7	斧	斧
6666	晦	晦
6753	杓	杓
6756	杖	杖
6897	梗	梗

UCS	Old glyph	New glyph
68D8	棘	棘
6962	楢	楢
696F	楯	楯
698A	榊	榊
6994	榔	榔
69CC	槌	槌
6A0B	樋	樋
6A3D	樽	樽
6A59	橙	橙
6ADB	橛	橛
6B4E	歎	歎
6C72	汲	汲
6DEB	淫	淫
6EA2	溢	溢
6EBA	溺	溺
6F23	漣	漣
7015	瀕	瀕
701E	瀞	瀞
7026	瀦	瀦
7058	灘	灘
7078	灸	灸
707C	灼	灼
7149	煉	煉
714E	煎	煎
717D	煽	煽

UCS	Old glyph	New glyph
723A	爺	爺
724C	牌	牌
7259	牙	牙
727D	牽	牽
72E1	狡	狡
7337	猷	猷
7511	甑	甑
7515	甕	甕
7526	甦	甦
75BC	疼	疼
77A5	瞥	瞥
7941	祁	祁
7947	祇	祇
795F	崇	崇
79B0	禰	禰
79E4	秤	秤
7A17	稗	稗
7A7F	穿	穿
7AC8	竈	竈
7B08	笈	笈
7B75	筵	筵
7BAD	箭	箭
7BB8	箸	箸
7BC7	篇	篇
7BDD	簗	簗

UCS	Old glyph	New glyph
7C3E	簾	簾
7C7E	籾	籾
7C82	粂	粂
7FEB	翫	翫
7FF0	翰	翰
8171	腱	腱
817F	腿	腿
818F	膏	膏
8258	艘	艘
8292	芒	芒
82A6	芦	芦
8328	茨	茨
845B	葛	葛
84EC	蓬	蓬
8511	蔑	蔑
853D	蔽	蔽
85A9	薩	薩
85AF	薯	薯
85F7	藷	藷
8654	虔	虔
86F8	蛸	蛸
8703	蜃	蜃
8755	蝕	蝕
87F9	蟹	蟹
8805	蠅	蠅

UCS	Old glyph	New glyph
8956	襖	襖
8A0A	訊	訊
8A1D	訝	訝
8A3B	註	註
8A6E	詮	詮
8AB9	誹	誹
8AFA	諺	諺
8B0E	謎	謎
8B2C	謬	謬
8C79	豹	豹
8CED	賭	賭
8FBB	辻	辻
8FBF	辿	辿
8FC2	迂	迂
8FC4	迄	迄
8FE6	迦	迦
9017	逗	逗
9019	這	這
9022	逢	逢
903C	逼	逼
9041	遁	遁
905C	遜	遜
9061	遡	遡
912D	鄭	鄭
914B	酋	酋

UCS	Old glyph	New glyph
91DC	釜	釜
9306	錆	錆
9375	鍵	鍵
939A	鎚	鎚
9453	鑓	鑓
9699	隙	隙
9744	靄	靄
9771	靭	靭
9784	鞄	鞄
9798	鞘	鞘
97AD	鞭	鞭
98F4	飴	飴
9905	餅	餅
990C	餌	餌
9910	餐	餐
9957	饗	饗
99C1	駁	駁
9A19	騙	騙
9BAB	鮫	鮫
9BD6	鯖	鯖
9C2F	鰯	鰯
9C52	鱒	鱒
9D09	鴉	鴉
9D60	鵠	鵠