Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2/IRG N2008

Date: 2014-05-16

Source:	Japan
Title:	Concern on WG2 N4544
Meeting:	IRG #42
Status :	
Actions required	For Discussion
Distribution:	IRG
Medium :	Electronic
Page:	
References:	WG2 N4544

0. Background

On WG2 #62, held in San Jose last February, WG2 project editor suggested replacing 168 glyph shapes of J source characters with the latest version of JIS X0213. For this purpose he also proposed to establish the policy about handling CJK glyph shapes that is very different from the current practice.

Japan abstained from this resolution because this is big policy change of IRG work and have impact to the consistency. This is general issue for IRG group

1. Proposal in WG2 N4544

In WG2 N4544, WG2 project editor proposes changing Japan's 168 glyph shape as corresponding to the latest version of JIS X0213. For this purpose, he also proposes to establish the policy of CJK glyph in the standard.

There is a "NOTE" about CJK glyphs in 23.1 of ISO/IEC 10646 as:

> 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.

Although "note" in the standard is just an information, Japan believes that this is a basic policy in practice of IRG work since the first version of CJK.

Propose in WG2 N4544 is deleting above note and adding the text below as a normative text instead.

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

Small changes are made to the text and the meaning is totally different.

2. Expected side effects

Changing ISO/IEC 10646 as proposed in WG2 N4544, it may cause

a) Character inconsistency in shape with the old version

It is necessary to keep information about glyph changes to avoid confusion.

b) Increasing workload of reviewing code chart by IRG

Modification of code chart may cause unexpected error. Because IRG should be responsible on CJK code chart, it is necessary to the all code chart is correct. This is what IRG experienced when the 2nd and 3rd version of ISO/IEC10646 was developed.

3. Suggestion

It is understandable that there are some people who wants to know the latest glyph shape corresponding to the specific national standard, but that purpose is not general. It is enough to add information about glyph version of national standards to somewhere in ISO/IEC 10646. Japan guess "Annex P" is the good place for such purpose.

Changing the current practice is serious. Japan suggests not to change the text in ISO/IEC 10646 to keep the current practice.

Japan also expects IRG to understand this issue and to discuss for the solution.

(End of Document)

References:

- Clause 23 in ISO/IEC 10646:2012
- Annex P in the next version of ISO/IEC 10646 (Draft)
- WG2 N4544

23 Source references for CJK Ideographs

23.1 List of source references

A CJK Ideograph is always referenced by at least one source reference. These source references are provided in a machine-readable format that is accessible as links to this document. The content pointed by these links is also normative.

NOTE 1 – The referenced files are only available to users who obtain their copy of the standard in a machine-readable format. However, the file format makes them printable.

The source reference information establishes the character identity for CJK Ideographs. A source reference is established by associating a CJK Ideograph code point with one or several values in the source standards listed below in this clause. Such a source standard originates from the following categories:

- Hanzi G sources,
- Hanzi H sources,
- Hanzi M sources,
- Hanzi T sources,
- Kanji J sources,
- Hanja K sources,
- Hanja KP sources,
- ChuNom V sources, and
- Unicode U sources

For a given code point, only one source reference can be created for each of the source standard category (G, H, M, T, J, K, KP, V, and U). In order to provide a comprehensive coverage for a source standard category, when a source standard is referenced, all its unique associations with existing CJK Ideographs are documented.

The following list identifies all sources referenced by the CJK Ideographs in both the BMP and the SIP.

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 Hanzi G sources are

G0	GB2312-80
G1	GB12345-90
G3	GB7589-87 unsimplified forms
G5	GB7590-87 unsimplified forms
G7	General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi
GS	Singapore Characters
G8	GB8565-88
G9	GB18030-2000
GE	GB16500-95
GH	GB15564-1995 Code of Chinese Ideogram set for teltext broadcasting Hong Kong subset
GK	GB12052-89 Korean Character Coded Character Set for Information Interchange
G4K	Siku Quanshu （四庫全書）
GBK	Chinese Encyclopedia （中國大百科全書）
GCH	Ci Hai （辞海）
GCY	Ci Yuan （辭源）
GCYY	Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字）
GFZ	Founder Press System （方正排版系统）
GGH	Gudai Hanyu Cidian （古代汉语词典）

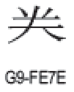
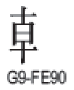
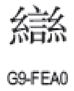
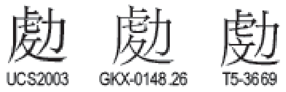
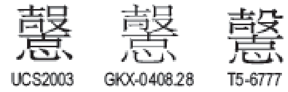
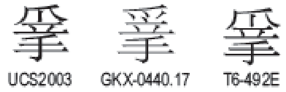
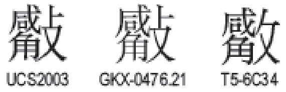
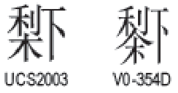
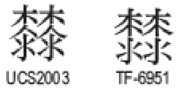
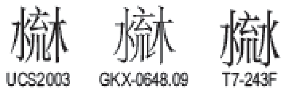
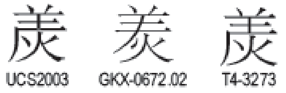
Annex P
(informative)
Additional information on CJK Unified Ideographs

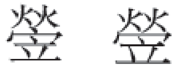
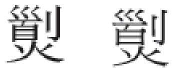
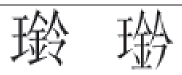
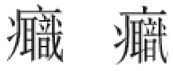
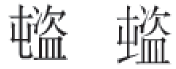
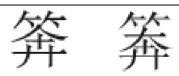
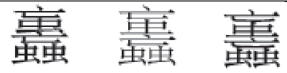
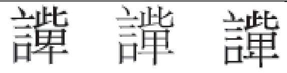
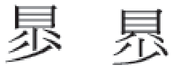
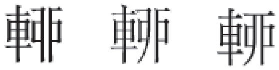
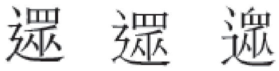
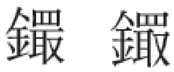
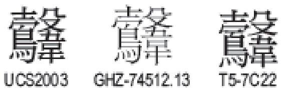
Annex P contains additional information on CJK Unified Ideographs.

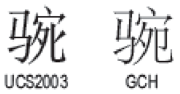
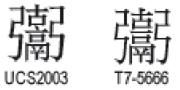
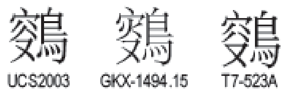
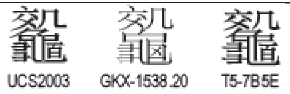
NOTE – The first edition of this International Standard (ISO/IEC 10646:2003 and amendments 1 to 5) used this annex to provide additional information on all characters. This edition of the standard includes most of that information in the code charts. Because the code charts for CJK unified ideographs do not include any name list, the information about these characters is still included here.

Each entry in the table P.1 consists for each row of an extract of the CJK Unified Ideograph code point entry in the code chart, followed in the next column by the related additional information. Entries are arranged in ascending sequence of code point.

Table P.1: Additional information on CJK Unified Ideographs

UCS / Glyph		Additional information
9FB9 八 12.4		These three characters are intended to represent a component at a specific position of a full ideograph. The ideographs representing the same structure without a preferred positional preference are encoded at 20509 𠔉, 2099D 𠦝, and 470C 䜌 respectively.
9FBA 十 24.6		
9FBB 言 149.12		
20885 力 19.10		T5-3669 source glyph was mistakenly unified to this code point.
22936 心 61.16		T5-6777 source glyph was mistakenly unified to this code point.
22BA3 手 64.8		GKX-0440.17 source glyph was mistakenly unified to this code point.
23023 攴 66.16		T5-6C34 source glyph was mistakenly unified to this code point.
235F1 木 75.10		UCS2003 glyph for this code point was mistakenly designed.
2382C 木 75.18		UCS2003 glyph for this code point was mistakenly designed.
23EE4 水 85.11		T7-243F source glyph was mistakenly unified to this code point.
24229 火 86.7		GKX-0672.02 source glyph was mistakenly unified to this code point.

UCS / Glyph	Additional information
24369 火 86.12  UCS2003 TF-5024	UCS2003 glyph for this code point was mistakenly designed.
243BE 火 86.12  UCS2003 T7-2F4B	The source glyph for T7-2F4B should have been unified with 24381 𤎁 but was allocated here by a mistake. The UCS2003 glyph for this code point should have been based on T7-2F4B but showed different shape by a mistake. For consistency with TCA CNS standards, 243BE's source reference to T7-2F4B is kept as in this International Standard.
24A8A 王 96.13  UCS2003 V2-7C66	UCS2003 glyph for this code point was mistakenly designed.
24F15 疒 104.18  UCS2003 V2-7D5A	UCS2003 glyph for this code point was mistakenly designed.
25089 皿 108.10  UCS2003 V2-7D6B	UCS2003 glyph for this code point was mistakenly designed.
25B88 竹 118.8  UCS2003 V3-364B	UCS2003 glyph for this code point was mistakenly designed.
27555 虫 142.17  UCS2003 GKX-1103.29 T5-7649	UCS2003 glyph for this code point was mistakenly designed.
27B1F 言 149.12  UCS2003 GHZ-64018.09 T7-5035	GHZ-64018.09 source glyph was mistakenly unified to this code point.
27D41 貝 154.4  UCS2003 TF-385F	UCS2003 glyph for this code point was mistakenly designed.
28321 車 159.8  UCS2003 GKX-1244.18 T6-632A	T6-632A source glyph was mistakenly unified to this code point.
28599 辵 162.11  UCS2003 T5-516D V4-5565	V4-5565 source glyph was mistakenly unified to this code point.
28B75 金 167.13  UCS2003 TF-686D	The glyph of TF-686D in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, TF-686D glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point.
293FB 韋 178.20  UCS2003 GHZ-74512.13 T5-7C22	The glyph of T5-7C22 in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, T5-7C22 glyph needs to be as in this International Standard, although the glyph is not usually unified with GHZ-74512.13 glyph and/or UCS2003 glyph of this code point.

UCS / Glyph	Additional information
299FB 马 187'.8 	The GCH glyph for this code point has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646. The GCH glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point.
29C52 鬲 193.12 	The glyph of T7-5666 in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646. For consistency with TCA CNS standard, T7-5666 glyph needs to be as in this International Standard, although the glyph is not usually unified with UCS2003 glyph of this code point
2A0B8 鳥 196.9 	The source glyph of T7-523A in TCA CNS standard has been changed after the original publication of CJK UNIFIED IDEOGRAPHS EXTENSION B in ISO/IEC 10646-2. For consistency with TCA CNS standard, T7-523A glyph needs to be as in this International Standard, although the glyph is not usually unified with GKX-1494.15 glyph and/or UCS2003 glyph of this code point.
2A6C0 龜 213.8 	GKX-1538.20 source glyph was mistakenly unified to this code point.

SC2/WG2 N4544
L2/014-0xx
Date: 2014-02-20

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: 逢 Modern Japanese representation: 逢

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 may not be updated. The updated source should only identify 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. 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.

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

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.

Subject: RE: This just in from Japan...
From: Michel Suignard <michel@suignard.com>
Date: 20/5/2014 1:54 AM
To: Ken Lunde <lunde@adobe.com>
CC: "John H. Jenkins" <jenkins@apple.com>, Lu qin <csluqin@comp.polyu.edu.hk>

(Adding Lu Qin)
Please John and Lu Qin, make sure my point of view is presented at this IRG:

I want to alert IRG that the version of WG2 N4544 appended to the IRGN2008 is obsolete in some important ways, especially because the quoted paragraph from N4544 is different now.
They should refer to N4544 as posted in std.dkuug.dk as http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4544.pdf

There are many errors, or at least questionable assertions in IRG2008.

1) I am supposedly 'proposing to establish the policy about handling CJK glyph shapes that is very different from the current practice'. I don't think there is a consistent practice among IRG members. Some (a majority) update their glyphs, some don't.

2) Related to that, the note in sub-clause 23.1 is by definition informative and cannot be construed as setting a basic policy for 10646. It has possibly established a guideline for Japan, but has not been followed by other IRG members.

3) at the end of page 1, IRG2008 says the 'propose in WG2 N4544 is deleting above note and adding the text below as a normative text'. Many issues here:
a) the 'added text' is from an earlier version of N4544, it was updated 2/26/2014. The new text is in fact closer to what Japan wants.
b) the description of added text is very incomplete. N4544 adds much more text than that 3 lines paragraph.

I still believe that the current version of 10646 by showing obsolete glyphs for 168 Japanese sources CJK Ideographs is failing users need. 10646 is a modern and living standard and should be usable as a current reference for all its glyphs. That is the major reason for having up to date glyphs. If Japan had a rationale for changing these glyphs in the first place, that should be reflected in the Universal Standard.

4) The argument of increased workload is also weak. We are talking of glyph changes for 168 characters which would be done by a font using a modern representation of these characters which is in broad use in Japan. The glyphs changes are subtle but very well known by Japanese experts. It would take less than a day to verify that change.

5) There is a larger debate about updating the sources themselves which was approached by the initial version of N4544, but as project editor I am willing to process slower on that one. However I also think that there is nothing wrong in using more up to date sources references while preserving the original sources that were used to create the character in a secondary repository. JIS X2013-2004 is a much better source of information for these Japanese sourced characters and if anything the current status is very confusing to many experts because it uses many old and obsolete references.

Ken Lunde has written extensively on that subject. He also discovers that an additional 30 glyphs have had a change in their prototypical glyph when they got into JIS X 0213: 2004. This would in essence require a source reference change because unlike the previous 168 glyphs, the original source was not 'updated'. The scope is 2743 source changes from J1 sources and 85 from JA sources.

This is a bit complex, but the devil is in the transition. The end result would be much cleaner, with less references and an easy access to these modern ones. We could document the legacy in Annex P and possible some collections in Annex A, which seems a much better approach than keeping the old complicated and obsolete state of affair in the main part of the standard and documenting the modern usage in Annexes. I am willing to prepare a document describing the whole scope for the next WG2 meeting in Colombo, but it is important for IRG to use up to date documents and not take decision limiting the use of the standard that would be in contradiction with the clear majority decision that was reflected in resolution M62.13.

Michel

-----Original Message-----
From: Ken Lunde [mailto:lunde@adobe.com]
Sent: Sunday, May 18, 2014 8:22 PM
To: Michel Suignard
Cc: John H. Jenkins

Subject: This just in from Japan...

http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg42/
IRGN2008_Glyphshape.pdf