From: Michel Suignard **L2/010-205**

To: UTC/L2

Date: May 13 2010

Subject: IRG Source format change

This document contains an excerpt of the new FDIS text (not yet final) where the format changes for the IRG sources have been applied. The modifications are as follows:

All sources use as a prefix the name of their source reference as indicated in the clause 1.1 List of source reference (below). That prefix is always separated by a hyphen from the numeric value when the latter exists. This introduces a syntax change for many G sources that had no '-' in their value, as well as M source (MAC) and U source (UTC).

The prefix for the G source 'Gudai Hanyu Cidian（古代汉语词典）' is stabilized as 'GGH'. It was originally described as 'G_GH' or 'GH'.

The previous source G_GFHZB has been replaced by 'GZH ZhongHua ZiHai (中华字海) ' which only covers part of the original content covered by G_GFHZB. The rest of that original content is now covered by the sources GCH, GIDC, and GXC.

All G sources that were using a 'G_' sequence within that prefix had that sequence replaced by 'G' except for 'G_GJZ' which was replaced with 'GJZ'.

The prefix for the Japanese J_ARIB source is consistently changed into 'JARIB'.

The GHZ Hanyu Dazidian source has now always a numeric value expressed as 'ddddd.dd'. The missing values were extracted from the Unihan database. The only sources that may have no numeric values are: G4K, GBK, GCH, GCY, GFZ, and GHC. The intent is to decrease the number of those numberless references.

Some six digits numeric references were incorrectly described as 5 digits before: GBKddddd, GCHddddd, GHCddddd, GXCddddd are now documented as follows: GBK-dddd.dd, GCH-dddd.dd, GHC-dddd.dd, GXC-dddd.dd. The two last digit separated by a '.' from the previous digits typically indicate the character serial index in the respective dictionary.

1.1 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 in ... 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.

The Hanzi G sources are

G0	GB2312-80
G1	GB12345-90 with 58 Hong Kong and 92 Korean “Idu” characters
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
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）
GHC	Hanyu Dacidian（漢語大詞典）
GHZ	Hanyu Dazidian ideographs（漢語大字典）
GIDC	ID system of the Ministry of Public Security of China, 2009

GJZ	Commercial Press Ideographs（商务印书馆用字）
GKX	Kangxi Dictionary ideographs（康熙字典）9th edition (1958) including the addendum（康熙字典）補遺
GXC	Xiandai Hanyu Cidian (现代汉语词典）
GZFY	Hanyu Fangyan Dacidian（汉语方言大辞典）
GZH	ZhongHua ZiHai (中华字海)
GZJW	Yinzhou Jinwen Jicheng Yinde（殷周金文集成引得）

The Hanzi H source is

H	Hong Kong Supplementary Character Set – 2008

The Hanzi M source is

MAC	Macao Information System Character Set（澳門資訊系統字集）

The Hanzi T sources are

T1	TCA-CNS 11643-1992 1st plane
T2	TCA-CNS 11643-1992 2nd plane
T3	TCA-CNS 11643-1992 3rd plane with some additional characters
T4	TCA-CNS 11643-1992 4th plane
T5	TCA-CNS 11643-1992 5th plane
T6	TCA-CNS 11643-1992 6th plane
T7	TCA-CNS 11643-1992 7th plane
TB	TCA-CNS Ministry of Education, Hakka dialect, May 2007
TC	TCA-CNS 11643-1992 12th plane
TD	TCA-CNS 11643-1992 13th plane
TE	TCA-CNS 11643-1992 14th plane
TF	TCA-CNS 11643-1992 15th plane

The Kanji J sources are

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

The Hanja K sources are

K0	KS X 1001:2004 (formerly KS C 5601-1987)
K1	KS X 1002:2001 (formerly KS C 5657-1991)
K2	PKS C 5700-1 1994
K3	PKS C 5700-2 1994
K4	PKS 5700-3:1998
K5	Korean IRG Hanja Character Set 5th Edition: 2001

NOTE 2 – The content of the repertoire covered by the K2, K3, K4, and K5 sources is in the process of being reedited in new Korean standards.

The Hanja KP sources are

KP0	KPS 9566-97
KP1	KPS 10721:2000 and KPS 10721:2003

The ChuNom V sources are

V0	TCVN 5773:1993
V1	TCVN 6056:1995
V2	VHN 01:1998
V3	VHN 02: 1998
V4	Dictionary on Nom 2006, Dictionary on Nom of Tay ethnic 2006, Lookup Table for Nom in the South 1994

The Unicode U source is

UTC	The Unicode Technical Report #45, U-source Ideographs, June 2008

NOTE 3 – Even if source references get updated, the source reference information is not updated. The updated source references may only identify characters not previously covered by the older version.

1.2 Source references for CJK Unified Ideographs

The procedures that were used to derive the unified ideographs from the source character set standards, and the rules for their arrangement in the code charts in .., are described in ….

NOTE 1 – The source separation rule described by the clause of that annex only apply to CJK Unified Ideographs within the BMP.

The content linked to is a plain text file, using ISO/IEC 646-IRV characters with LINE FEED as end of line mark, that specifies, after a 13-lines header, as many lines as CJK Unified Ideographs in the sum of the two planes; each containing the following information organized in fields delimited by ';' (empty fields use no character):

- 1st field: BMP or SIP code point `(0hhhh)`, `(2hhhh)`
- 2nd field: Radical Stroke index `(d{1,3}'.d{1,2})`. This informative field contains radical index (one to three digits), optionally followed by an apostrophe for alternate index, followed by a full stop, and ending by one or two digits for the stroke count.

  NOTE 2 – All ideographs are classified following radical/stroke indexes in various CJK dictionaries. The primary value provided in this field is the most common one, while alternate indexes provide variant values also in use. More information is available in the Unicode Standard UAX#38 Unicode Han Database at http://www.unicode.org/reports/tr38/.

- 3rd field: Hanzi G sources `(G0-hhhh)`, `(G1-hhhh)`, `(G3-hhhh)`, `(G5-hhhh)`, `(G7-hhhh)`, `(GS-hhhh)`, `(G8-hhhh)`, `(G9-hhhh)`, `(GE-hhhh)`, `(G4K)`, `(GBK)`, `(GBK-dddd.dd)`, `(GCH)`, `(GCH-dddd.dd)`, `(GCY)`, `(GCYY-ddddd)`, `(GFZ)`, `(GFZ-ddddd)`, `(GGH-ddddd.dd)`, `(GHC)`, `(GHC-dddd.dd)`, `(GHZ-ddddd.dd)`, `(GIDC-ddd)`, `(GJZ-ddddd)`, `(GKX-dddd.dd)`, `(GXC-dddd.dd)`, `(GZFY-ddddd)`, `(GZH-dddd.dd)`, or `(GZJW-ddddd)`
- 4th field: Hanzi T sources `T1-hhhh)`, `(T2-hhhh)`, `(T3-hhhh)`, `(T4-hhhh)`, `(T5-hhhh)`, `(T6-hhhh)`, `(T7-hhhh)`, `(TB-hhhh)`, `(TC-hhhh)`, `(TD-hhhh)`, `(TE-hhhh)`, or `(TF-hhhh)`
- 5h field: Kanji J sources `(J0-hhhh)`, `(J1-hhhh)`, `(J3-hhhh)`, `(J3A-hhhh)`, `(J4-hhhh)`, `(JA-hhhh)`, `(JH-xxxxxx)`, `(JK-ddddd)`, or `(JARIB-hhhh)`
- 6th field: Hanja K sources `(K0-hhhh)`, `(K1-hhhh)`, `(K2-hhhh)`, `(K3-hhhh)`, `(K4-hhhh)`, or `(K5-hhhh)`
- 7th field: ChuNom V sources `(V0-hhhh)`, `(V1-hhhh)`, `(V2-hhhh)`, `(V3-hhhh)`, or `(V4-hhhh)`
- 8th field: Hanzi H source `(H-hhhh)`
- 9th field: Hanja KP sources `(KP0-hhhh)` or `(KP1-hhhh)`
- 10th field: Unicode U source `(UTC-ddddd)`

- 11th field: Hanzi M source `(MAC-ddddd)`

The format definition uses `'d'` as a decimal unit, `'h'` as a hexadecimal unit, and `'x'` as an alphanumerical unit (0 to 9 and A to Z). Uppercase characters, digits and all other symbols between parentheses appear as shown.