L2/13-147

ISO/IEC JTC 1/SC 2/WG2 *N4436*
Date: 2013-06-02

Title: Proposal to change data format for CJK sources
Source: Project Editor
Distribution: WG2 members and liaison organizations

The following document describes a proposed change to the data source format used to describe the CJK Ideographs source references in order to simplify synchronization between ISO/IEC 10646 and the Unicode Standard.

Background:

Today ISO/IEC 10646 uses the following data file to describe the source references for CJK Ideographs:

- CJKU_SR.txt describes sources for the CJK Unified Ideographs. It contains 11 fields (UCS code point, Radical/Stroke index, 9 sources: G, H, M, T, J, K, KP, V, and U).
- CJKC_SR.txt describes sources for the CJK Compatibility Ideographs. It contains 9 fields (UCS code point, corresponding CJK Unified Ideograph code point, Radical/Stroke, 6 sources: T, H, J, K, U, and U). Three CJK sources have no compatibility fields: G, M, and V.

In addition, another data file: IICORE.txt describes the International Core subset of the CJK Unified repertoire. It contains 9 fields (UCS code point, 7 source related identifiers: G, T, J, H, K, M, and KP, and General category).

On the Unicode side, the same information is conveyed through the following three files:

- Unihan_IRGSources.txt contains the IICore info (kIICore which is a simple version number: 2.1) and the 9 sources references (kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_USource, kIRG_VSource, and kIRG_MSource). The file covers both CJK Unified and Compatibility ideographs.
- Unihan_Variants.txt contains (among other fields) kCompatibilityVariant which is the CJK Unified Ideograph when appropriate).
- Unihan_RadicalStrokeCounts.txt contains (among other fields) kRSUnicode which is an augmented version of the Radical/Stroke count available with ISO/IEC 10646.

Comparisons/issues with the current data field from both standards.

ISO/IEC 10646 presents the information in a tabular fashion. Example of entries for 4E00 and 4E01 in CJKU_SR.txt:
04E00;1.0;G0-523B;T1-4421;J0-306C;K0-6C69;V1-4A21;HB1-A440;KP0-FCD6;;
04E01;1.1;G0-3621;T1-4423;J0-437A;K0-6F4B;V1-4A22;HB1-A442;KP0-E8B9;;
Unicode (Unihan) presents the same information in a sequential fashion. Examples for the same entries in Unihan_IRGSources.txt:

U+04E00	kIRG_GSource	G0-523B
U+04E00	kIRG_HSource	HB1-A440
U+04E00	kIRG_JSource	J0-306C
U+04E00	kIRG_KPSource	KP0-FCD6
U+04E00	kIRG_KSource	K0-6C69

U+04E00 kIRG_TSource T1-4421
U+04E00 kIRG_VSource V1-4A21
U+04E00 kRSUnicode 1.0
U+04E01 kIRG_GSource G0-3621
U+04E01 kIRG_HSource HB1-A442
U+04E01 kIRG_JSource J0-437A
U+04E01 kIRG_KPSource KP0-E8B9
U+04E01 kIRG_KSource K0-6F4B
U+04E01 kIRG_TSource T1-4423
U+04E01 kIRG_VSource V1-4A22
U+04E01 kRSUnicode 1.1

Unicode merge the source information for both CJK Unified and CJK Compatibility ideographs in a single file which makes sense because they are issued from the same source sets.

The field format for source references is the same between Unicode and ISO/IEC 10646 (see example above).

The Unicode Radical/Stroke counts kRSUnicode contains one or two counts. The second optional count document cases where the main count is considered not optimal. The count included in the ISO/IEC 10646 always corresponds to the first of these two counts and use the same syntax.

The Unicode kIICore value contains the value '2.1' corresponding to the IICORE version when it was standardized as collection 370 by ISO/IEC 10646. The IICORE information in the ISO/IEC 10646 is much more complex, because it contains information about the source, its priority within the source and its overall priority. While more complete than the Unicode version it has its own issue which is redundancy with the source reference (and thus possible lack of synchronization).

Proposal

Create a new single file containing the following fields available in sequential mode (each line contains first the UCS code point followed by the field name and then the value, each separated by a TAB). These lines are repeated as many times as there are fields defined for a given UCS code point. The possible fields and syntax are described in the following table:

Name	Type	Regex Syntax
kIRG_GSource	Hanzi G source	G(4K|BK|CH|CY|FZ|HC|HZ|((BK|CH|GH|HC|XC|ZH)-[0-9]{4}\.[0-9]{2})|HZ-[0-9]{5}\.[0-9]{2}|(KX-[01][0-9]{3}\.[0-9]{2})|((CYY|FZ|JZ|ZFY|ZJW)-[0-9]{5})|([0135789ES]-[0-9A-F]{4})|(IDC-[0-9]{3})|(K-[0-9A-F]{4})|(H-\d{4}))
kIRG_HSource	Hanzi H source	H((3)|(B[012]))?-[0-9A-F]{4}
kIRG_MSource	Hanzi M source	MAC-[0-9]{5}
kIRG_TSource	Hanzi T source	T[1-7B-F]-[0-9A-F]{4}
kIRG_JSource	Kanji K source	J((([0134AK]|3A|ARIB)-[0-9A-F]{4,5})|(H-(((IB|JT|[0-9]{2})[0-9A-F]{4}S?))))
kIRG_KSource	Hanja K source	K[0-57]-[0-9A-F]{4}
kIRG_KPSource	Hanja KP source	KP[01]-[0-9A-F]{4}
kIRG_VSource	ChuNom V source	V[0-4]-[0-9A-F]{4}

kIRG_USource	Unicode U source	U(TC|CI)-[0-9]{5}
kIICore	IICORE info	[ABC]{1}[GTJHKMP]{1-7}
kCompatibilityVariant	Compatibility info	U\+2?[0-9A-F]{4}
kRSUnicode	Radical-Stroke counts	[1-9][0-9]{0,2}\'?\.[0-9]{1,2}

All fields except kIICore are using existing syntax and are already documented in the Unicode Unihan database (http://www.unicode.org/reports/tr38/). The kIICore would preserve the overall priority and letters referring one of the 7 possible sources: G,H,J,K,M,P, and T (P representing the KP source).

Following are examples for 3687, 4E00, 4E07, and F928. Note that the three first definitions correspond to CJK Unified Ideographs, with 4E00 and 4E07 part of the IICore set, while the fourth definition corresponds to a CJK Compatibility Ideograph.

```
U+03687      kIRG_GSource             G3-3A36
U+03687      kIRG_KPSource            KP1-3C87
U+03687      kIRG_KSource             K3-2339
U+03687      kIRG_TSource             T4-2861
U+3687       kRSUnicode               35.6 66.6

U+04E00      kIRG_GSource             G0-523B
U+04E00      kIRG_HSource             HB1-A440
U+04E00      kIRG_JSource             J0-306C
U+04E00      kIRG_KPSource            KP0-FCD6
U+04E00      kIRG_KSource             K0-6C69
U+04E00      kIRG_TSource             T1-4421
U+04E00      kIRG_VSource             V1-4A21
U+04E00      kRSUnicode               1.0
U+04E00      kIICore                  AGTJHKMP

U+04E07      kIRG_GSource             G0-4D72
U+04E07      kIRG_HSource             HB2-C945
U+04E07      kIRG_JSource             J0-4B7C
U+04E07      kIRG_KPSource            KP0-DAB9
U+04E07      kIRG_KSource             K0-5832
U+04E07      kIRG_TSource             T2-2126
U+04E07      kIRG_VSource             V1-4A24
U+04E07      kRSUnicode               1.2
U+04E07      kIICore                  AGJKP

U+0F928      kIRG_JSource             J3-742E
U+0F928      kIRG_KSource             K0-5227
U+0F928      kRSUnicode               53.9
U+0F928      kCompatibilityVariant    U+5ECA
```

Additional Notes

This proposal would require a rewrite of clause 23 ‘Source References for CJK Ideographs’ merging the definition for CJK Unified and Compatibility Ideographs. It may also be desirable to use the formal regex expression instead of the ad hoc definition currently used. A formal regex reference would be required if done that way. Sub-clause A.4.1 ‘370 IICORE’ would also need to be slightly modified to describe the new format and new file.

Otherwise, conversion to the new format is a simple matter because it is very close to an internal file already generated to create the CJK multicolumn charts for both ISO/IEC 10646 and Unicode.