Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: ISO/IEC JTC1/SC2/WG2/IRG
Title: Proposal to define new Unihan Database property: **kUnihanCore2020**
Source: Ken Lunde (Adobe)
Status: Individual Contribution
Action: For consideration by the UTC & IRG
Date: 2018-10-04 (original proposal was dated 2018-09-03; also see L2/18-066, L2/18-066R & L2/18-066R2)

Per the documents listed directly above in parentheses, two of which are subsequent revisions of the first document, I previously proposed what I considered to be modest changes to the existing *kIICore* property, mainly to address some shortcomings that were identified in a series of five CJK Type Blog articles. Given the reluctance on the part of some national bodies to accept such modest changes, I decided to instead propose a completely new Unihan Database property that releases the set from being hampered by memory constraints that may have been applicable 15 years ago, but which arguably no longer apply to modern environments.

The proposed property name is *kUnihanCore2020*, which includes as part of its name the year in which the first version of Unicode that could include this new property is released, specifically Version 13.0. The attached *unihancore2020-data.txt* data file provides all of the property data as proposed in this document, which covers 20,626 CJK Unified Ideographs and 68 CJK Compatibility Ideographs. Compared to the existing *kIICore* property, the proposed *kUnihanCore2020* property includes 10,906 additional ideographs, and excludes 22 ideographs that have a *kIICore* property value. The following table lists these 22 ideographs, their *kIICore* property value, and the source reference that corresponds to their source tag:

Code Point	kIICore Property Value	Corresponding Source Reference
U+3960 惰	CK	K3-2554
U+4137 秭	CK	K3-2D4F
U+48B5 郘	CG	G5-6F4F
U+48C5 郪	CG	G3-6F29
U+48D3 鄗	CG	G3-7B67
U+49D1 陮	CG	GKX-1352.16
U+4A12 霓	CK	K3-3455
U+4CB3 鳿	CT	T3-5028
U+4D08 鷍	CT	T4-6C52
U+593D 奍	CK	K2-2B54
U+5D44 嵄	CK	K2-2F33
U+5F34 弴	CJ	J13-7436
U+5F45 彅	CJ	J13-743A

Code Point	kIICore Property Value	Corresponding Source Reference
U+66A3 暭	CK	K1-5B6F
U+713F 煿	CT	T3-6552
U+7807 砇	CK	*n/a*
U+7A66 穦	CK	*n/a*
U+974D 靌	CJ	J3-7D68
U+974F 靏	CJ	J13-7D6A
U+9964 饤	CG	G8-2D43
U+997E 饾	CG	G8-2D48
U+9AD9 髙	CJ	*n/a*

Also see the attached *excluded-22.txt* data file.

The seven sections that follow describe the scope of each of the seven supported source tags, which are the same as those used by the existing *kIICore* property.

G—PRC

The proposed scope of the "G" source tag is the union of the GB 2312 (6,763) and TGH-2013/通用规范汉字表/*Tōngyòng Guīfàn Hànzìbiǎo* (8,105—see the kTGH property) standards, which results in 8,230 unique ideographs, all of which are CJK Unified Ideographs. This figure is only 125 ideographs more than TGH-2013 itself.

SPECIAL NOTES: 29 existing *kIICore* ideographs with the "G" source tag are excluded, because they are outside the scope of the two specified standards; only the following six of those 29 ideographs are not covered by one of the other six source tags: U+48B5 邵 (G5), U+48C5 郾 (G3), U+48D3 鄚 (G3), U+49D1 陧 (GKX), U+9964 饤 (G8) & U+997E 饾 (G8). See the attached *excluded-g-29.txt* data file.

H—Hong Kong SAR

The proposed scope of the "H" source tag is the union of the Big Five (13,060—see the kBigFive property) and HKSCS (4,602) standards, which results in 17,662 unique ideographs, 11 of which are CJK Compatibility Ideographs. There is no overlap between these two standards.

J—Japan

The proposed scope of the "J" source tag is the union of the JIS X 0208 (6,356), 常用漢字/*Jōyō Kanji* (2,136—see the kJoyoKanji property), 人名用漢字/*Jinmei-yō Kanji* (863—see the kJinmeiyoKanji property), and 表外漢字/*Hyōgai Kanji* (1,022) standards, which results in 6,485 unique ideographs, 58 of which are CJK Compatibility Ideographs. This figure is only 129 more ideographs than JIS X 0208 itself.

SPECIAL NOTES: Two existing *kIICore* ideographs with the "J" source tag are excluded, because they are outside the scope of the specified standards; only the following ideograph is not covered by one of the six other source tags: U+9AD9 髙 (*no kIRG_JSource*). See the attached *excluded-j-2.txt* data file.

K—ROK

The proposed scope of the "K" source tag is the union of the KS X 1001 (4,620) and 한문 교육용 기초 한자/漢文教育用基礎漢字/*Hanmun Gyoyug-yong Gicho Hanja* (1,800—see the kKoreanEducationHanja property) stan-

dards, which results in 4,632 unique ideographs, all of which are CJK Unified Ideographs. This figure is only 12 more ideographs than KS X 1001 itself.

SPECIAL NOTES: 134 existing *kIICore* ideographs with the "K" source tag are excluded, because they are outside the scope of the two specified standards; only the following eight of those 134 ideographs are not covered by one of the other six source tags: U+3960 憎 (K3), U+4137 秅 (K3), U+4A12 霠 (K3), U+593D 奄 (K2), U+5D44 嵄 (K2), U+66A3 暵 (K1), U+7807 砇 (*no kIRG_KSource*) & U+7A66 穻 (*no kIRG_KSource*). See the attached *excluded-k-134.txt* data file.

M—Macao SAR

The proposed scope of the "M" source tag is the union of the Big Five standard (13,060—see the kBigFive property) and the existing *kIICore* ideographs that have the "M" source tag (4,954), which results in 13,119 unique ideographs, all of which are CJK Unified Ideographs. This figure is only 59 more ideographs than Big Five itself.

SPECIAL NOTES: Only one existing *kIICore* ideograph with the "M" source tag, U+5F66 彦, is excluded for reasons explained in the 2018-02-15 CJK Type Blog article, but is covered by four of the other six source tags (G, J, K & P): *Only one ideograph, U+5F66 彦, stands out as odd in that its source references do not suggest Macao SAR use. Its related ideograph, U+5F65 彥, is also tagged "M" in* kIICore *(ATHM), and its source references, particularly T1-507D, more strongly suggest Macao SAR use.* See the attached *excluded-m-1.txt* data file.

P—DPRK

The proposed scope of the "P" source tag is the KPS 9566 (4,653) standard, which means that this is unchanged from *kIICore*.

T—ROC

The proposed scope of the "T" source tag is the union of the CNS 11643 Levels 1 & 2 (13,063) and Big Five (13,060—see the kBigFive property) standards, which results in 13,064 unique ideographs, all of which are CJK Unified Ideographs.

SPECIAL NOTES: 93 existing *kIICore* ideographs with the "T" source tag are excluded, because they are outside the scope of the two specified standards; only the following three of those 93 ideographs are not covered by one of the other six source tags: U+4CB3 鵁 (T3), U+4D08 鶊 (T4) & U+713F 煿 (T3). See the attached *excluded-t-93.txt* data file.

No Priority Tags

Because the notion of priority is largely source-specific, the proposed *kUnihanCore2020* property does not have a provision to specify priority tags. The author of this proposal feels that they are not necessary, and that the source tags are sufficient.

CJK Compatibility Ideographs

Although the proposed *kUnihanCore2020* property specifies source tags for 68 CJK Compatibility Ideographs—11 with the "H" source tag, and 57 with the "J" source tag—it is expected that their corresponding SVSes (*Standardized Variation Sequences*) be used in actual implementations. In addition, the CJK Compatibility Ideographs that correspond to the Big Five (2) and KS X 1001 (268) standards are intentionally excluded, because they represent genuine duplicate ideographs. See the attached *svs-68.txt* data file that provides a correspondence between the SVSes and CJK Compatibility Ideographs.

That is all.