L2/22-188

Title: Proposal to remove/improve provisional Unihan database properties (was L2/21-032)
Author: Ken Lunde
Date: 2022-09-01

This proposal was first submitted as L2/21-032, and was explicitly marked as a draft, because the CC BY-SA 2.1 JP license that is associated with the Moji Jōhō Kiban database (文字情報基盤データベース), the source of the property data, would have prevented it from being accepted by the UTC. The Unicode Consortium has since licensed the Moji Jōhō Kiban database from CITPC (*Character Information Technology Promotion Council* 文字情報技術促進協議会), the organization that currently owns the database, which allows the provisional *kMorohashi* Unihan database property to be improved.

This document consists of the following two Unihan database proposals:

- Remove the provisional *kIRGDaiKanwaZiten* property
- Improve the provisional *kMorohashi* property

Proposal #1: Remove the provisional *kIRGDaiKanwaZiten* property

The provisional *kIRGDaiKanwaZiten* property is a pure subset of the provisional *kMorohashi* property: all 17,865 records of the former property are present in the latter property with identical property values.

The provisional *kMorohashi* property additionally includes property values for U+43EE 腋 (Extension A) and U+657B 夐.

The *kIRGDaiKanwaZiten* property no longer serves any meaningful purpose, and offers no data that is not already included in the *kMorohashi* property, so I propose that it be removed from the Unihan database in the next version of the Unicode Standard.

Proposal #2: Improve the provisional *kMorohashi* property

The provisional Unihan database property, *kMorohashi*, maps CJK Unified Ideographs to *Dai Kanwa Jiten* (大漢和辞典) dictionary index numbers. In the Unicode Version 15.0 Unihan database, this property covers 17,867 CJK Unified Ideographs. The dictionary itself, which consists of 15 volumes, includes a total of 51,284 entries.

I propose that the following three improvements be made to the provisional *kMorohashi* property for the next version of the Unicode Standard based on the attached proposed improved property data:

- **Expand** the coverage of this property to include a significantly greater number of CJK Unified Ideographs
- **Enhance** the property values by including mappings to SVSes (*Standardized Variation Sequences*) and *Moji_Joho* IVSes (*Ideographic Variation Sequences*) where appropriate
- **Correct** existing property values

The proposed improved property data is based on Version 006.01 of the *Moji Jōhō Kiban database* (文字情報基盤データベース), which is the latest version as of this writing.

Expanding the provisional *kMorohashi* property

The proposed improved property data covers 49,071 CJK Unified Ideographs, 13,807 of which have identical property data when compared to the existing provisional *kMorohashi* property. The table below breaks down the 51,284 dictionary entries, and indicates how many are covered by the proposed improved property data:

		Volumes 1–12 + 索引			補巻
Mappings	Total	Standard	Prime	Double Prime	Supplemental
Dictionary	51,284	49,964	513	3	804
Proposed Data	49,486	48,669	508	3	306
Missing	1,798	1,295	5	0	498

Among the 1,798 *Dai Kanwa Jiten* index numbers that are not associated with a CJK Unified Ideograph, the following two are included in the *Moji Jōhō Kiban Database*, but have been explicitly excluded from the proposed improved property data, because they do not map to CJK Unified Ideographs:

00092 U+303B 〻 VERTICAL IDEOGRAPHIC ITERATION MARK

00097 U+3005 々 IDEOGRAPHIC ITERATION MARK

Some may claim that it is possible to associate *Dai Kanwa Jiten* index number 00097 with U+206A4 𠚪, but until the semantics of that CJK Unified Ideograph are known, I feel that it is risky to do so.

Enhancing the provisional *kMorohashi* property

The proposed improved property data, like the existing provisional *kMorohashi* property data, provides mappings from CJK Unified Ideographs to *Dai Kanwa Jiten* index numbers, but additionally provides mappings from SVSes and registered *Moji_Joho* IVSes.

In terms of syntax, if a colon (":") and VS (*Variation Selector*) follow a *Dai Kanwa Jiten* index number, the sequence of the CJK Unified Ideograph, serving as a BC (*Base Character*), followed by a VS, corresponds to the *Dai Kanwa Jiten* index number. Such sequences are either SVSes or registered *Moji_Joho* IVSes. Also, if a *Dai Kanwa Jiten* index number appears both by itself and followed by a colon and VS, the registered *Moji_Joho* IVS that corresponds to the latter is considered the default (aka encoded) form of the CJK Unified Ideograph.

The following table provides revised Syntax and Description for this property as reflected in UAX #38:

Field	Text
Syntax	`\d{5}[\'\"]?|H\d{3}(:(FE0[01]|E010[0-9A-F]))?`

Field	Text
Description	The index of the ideograph in the Dai Kanwa Jiten (大漢和辞典) Japanese kanji dictionary (1984–1986, 大修館書店)—often referred to as Morohashi (諸橋), the family name of its chief editor—or in the Dai Kanwa Jiten Hokan (大漢和辞典補巻) supplemental volume (2000, 大修館書店). Index numbers are five zero-padded integer values with an optional single apostrophe (') or quotation mark (") suffix that correspond to the appearance of a prime or double prime. Index numbers that appear in the supplemental volume (補巻) are prefixed with “H” and consist of three zero-padded integer values. If a colon (:) and VS (Variation Selector) follow an index number, the sequence of the CJK Unified Ideograph, serving as a BC (Base Character), followed by the VS, corresponds to the index number. Such sequences are SVSes or Moji_Joho IVSes. If an index number appears both by itself and followed by a colon and VS, the registered Moji_Joho IVS that corresponds to the latter is considered the default (aka encoded) form of the CJK Unified Ideograph. The Moji Jōhō Kiban database and its mappings are owned by CITPC (Character Information Technology Promotion Council 文字情報技術促進協議会), and are used under license.

Correcting the provisional *kMorohashi* property

When the existing property data was compared to the proposed improved property data, I determined that 72 mappings were corrected as a result. I made one additional correction. All 73 corrections were confirmed in the dictionary itself, and are provided in the table below:

Unicode		Current Property Data	Proposed Improved Property Data
U+5185	内	01512	00366' 00366':E0101
U+53C4	叅	03089	03099
U+5433	吳	03365'	03365
U+5676	噶	04426	04421:E0102
U+58F7	壷	05662	05657 05657:E0101
U+58FA	壺	05657	05662
U+594A	奐	05897	05909
U+594B	奋	05901	05900
U+594C	奌	05909	05908
U+5DDF	巟	08679	49065:E0100
U+5E5A	幫	09057	09097
U+626E	扮	11829	11830 11830:E0101
U+6287	抃	11870	11871

Unicode		Current Property Data	Proposed Improved Property Data
U+6718	朘	14365	29531
U+67F9	柹	14582	14596
U+69D8	様	15352	15352'
U+6A28	樨	15483	15472
U+6BAA	殪	16629	16578
U+6C77	汐	17170	17124
U+6D16	洖	17456	49211
U+7123	煦	19115	19116
U+71C5	燄	19399	19398
U+736A	獪	20929	20729
U+7611	瘍	22240	22340
U+77C9	矉	23778	23777
U+77CA	矊	23779	23778
U+7985	禅	24787	24787' 24787':E0101 24754:E0102
U+79BC	离	24889	24891
U+7D4B	絋	27289	H478
U+7DFB	緻	27700	H484
U+7F6F	罯	28313	28312
U+7F84	羂	38394	28394
U+7F95	羕	28449	28488:E0102
U+80FC	胼	29586	29453
U+8127	朘	29531	14365
U+8192	膃	29812	29809
U+8400	莵	31243	31244
U+8401	莾	31244	31248:E0102
U+8608	蘋	42429	32429
U+862D	蘭	32477'	32477" 32477":E0102 32519:E0103
U+8641	虆	32601	05747 05747:E0101
U+865C	虜	32710'	32720' 32710:FE00 32720':E0102 32710:E0103
U+86CA	蚩	32972	32974
U+8744	蜩	33217	H543 H543:E0101

Unicode		Current Property Data	Proposed Improved Property Data
U+8746	蝆	33084	33218
U+882A	袈	33833	33832 33832:E0100
U+8873	袳	34125	34131
U+88A3	裭	34211	34212 34212:E0101
U+88BB	裻	34299	34249
U+88C0	裀	34297	34257
U+8A29	詶	35284	35282
U+8AB0	誰	35686	35586
U+8AB4	諒	35595	35593
U+8AE3	諤	35714	35712
U+8B64	譥	35006	36006
U+8C50	豐	36296	36304
U+8C51	豲	36304	36318
U+8CAE	貳	36703	H575
U+8E6B	躇	36852	37852
U+8EFF	軿	38396	38286
U+8FAC	辯	38667	13483:E0100
U+9039	達	30991'	H607:E0104
U+9089	邉	39214	H611 H611:E010F
U+90F9	郹	39501	39500
U+9138	鄤	39697	39696
U+928F	鈏	40373	40324
U+96C3	雃	41971	42015
U+985E	類	43608	43608' 43636:FE00 43608':E0102 43636:E0103
U+9D0B	鳩	46765	46715
U+9D36	鵵	45857	46857
U+9D37	鵷	45859	46859
U+9EEC	黦	47149	48149
U+9EF1	鼱	47180	48180

CJK Unified Ideograph coverage

The proposed improved property data covers a total of 49,071 CJK Unified Ideographs. Among them, 4,825 also serve as BCs for SVSes (73) and registered *Moji_Joho* IVSes (5,356, which is less than half of them). The following table indicates—on a per-block basis—the number of CJK Unified Ideographs that are assigned the proposed improved *kMorohashi* property data, along with the percentage that this number represents within each block as of Unicode Version 13.0:

	URO	Compatibility	Extension A	Extension B	Extension C	Extension D	Extension E	Extension F
Characters	18,100	4	5,711	25,115	25	3	30	83
Percent	86%	1%	87%	59%	1%	1%	1%	1%

Examples

The following are three prototypical examples from the proposed improved *kMorohashi* property data that use U+4E00 一, U+4E26 並, and U+4FAE 侮 as their BCs:

```
U+4E00   kMorohashi   00001
U+4E26   kMorohashi   00054 00054:E0102 00053:E0103
U+4FAE   kMorohashi   00629' 00630:FE00 00629':E0102 00630:E0103
```

The *Dai Kanwa Jiten* index numbers that are shown in red specify which registered *Moji_Joho* IVS corresponds to the default form of the CJK Unified Ideograph. There are 3,011 such instances in the proposed improved property data.

The *Dai Kanwa Jiten* index numbers that are shown in cyan demonstrate that SVSes always have a corresponding registered *Moji_Joho* IVS. There are 73 such instances in the proposed improved property data.

Playing the Devil's advocate

One question that may be raised by the UTC, which I would like to short-circuit if at all possible, is why not flip the proposals? In other words, why not remove the provisional *kMorohashi* property and instead improve the provisional *k**IRG**DaiKanwaZiten* property?

The IRG had nothing to do with this proposal to significantly improve the provisional *kMorohashi* property, so it would make less sense to improve the provisional property that includes "IRG" as part of its name. In addition, the IRG no longer uses this dictionary for its ongoing work. Perhaps the most compelling argument is that the *kMorihashi* property is referenced in a small number of *kSemanticVariant* and *kZVariant* property values.

Clearly, one of the provisional properties should be removed, because both refer to the same dictionary.

Data file

The *kMorohashi-data.txt* data file, which is a PDF attachment, provides everything that is necessary for improving this provisional property in the Unihan database, and for modifying its Syntax and Description in UAX #38, *Unicode Han Database (Unihan)*.

That is all.