L2/22-187

Title: Proposal for new provisional Unihan Database property: **kMojiJoho** (was L2/20-146)
Author: Ken Lunde
Date: 2022-08-19

This proposal was first submitted as L2/20-146, but 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, 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 proposed property to be added to the Unihan database.

Japan has recently promulgated the Moji Jōhō Kiban database, which is reflected in the normative kIRG_JSource Unihan database property, introduced in ISO/IEC 10646:2017 (Annex A.5.10) as Collection 390, *MOJI-JOHO-KIBAN IDEOGRAPHS-2016*, and introduced in ISO/IEC 10646:2020 (Annex A.5.11) as Collection 391, *MOJI-JOHO-KIBAN IDEOGRAPHS-2018*. Moji Jōhō Kiban database serial numbers map to CJK Unified Ideographs, SVSes (*Standardized Variation Sequences*), and registered *Moji_Joho* IVSes (*Ideographic Variation Sequences*).

The proposed provisional Unihan database property, **kMojiJoho**, maps CJK Unified Ideographs to Moji Jōhō Kiban database serial numbers, to include those that correspond to SVSes and registered *Moji_Joho* IVSes. Such a property is useful, because there are upwards of 60,000 entries in the Moji Jōhō Kiban database. In contrast, the *kIRG_JSource* property covers only 16,226 CJK Unified Ideographs as of Unicode Version 15.0, and ISO/IEC 10646 Collections 390 and 391 are merely listings of CJK Unified Ideographs, SVSes, and IVSes with no mapping information.

This proposed property provides mappings from CJK Unified Ideographs, along with SVSes and registered *Moji_Joho* IVSes that use the CJK Unified Ideograph as a BC (*Base Character*), to Moji Jōhō Kiban database serial numbers. The proposed property is based on Version 006.01 of the Moji Jōhō Kiban database, which is the latest version as of this writing.

In terms of syntax, if a colon (“:”) and VS (*Variation Selector*) follow a Moji Jōhō Kiban database serial number, the sequence of the CJK Unified Ideograph, serving as a BC, followed by a VS, corresponds to the Moji Jōhō Kiban database serial number. Such sequences are either SVSes or registered *Moji_Joho* IVSes. Also, if a Moji Jōhō Kiban database serial number appears both by itself and followed by a colon and VS, the registered IVS that corresponds to the latter is considered the default (aka encoded) form.

The following table provides the fields for this property as reflected in UAX #38:

Field	Text
Property	`kMojiJoho`
Status	`Provisional`
Category	`Dictionary-like Data`
Introduced	`TBD`
Delimiter	`space`
Syntax	`MJ\d{6}(:(FE0[01]|E01[01][0-9A-F]))?`

Field	Text
Description	This property provides mappings from CJK Unified Ideographs, along with SVSes (Standardized Variation Sequences) and registered Moji_Joho IVSes (Ideographic Variation Sequences) that use the CJK Unified Ideograph as a BC (Base Character), to Moji Jōhō Kiban database (文字情報基盤データベース) serial numbers. The property is based on Version 006.01 of the Moji Jōhō Kiban database. See: https://moji.or.jp/mojikiban/mjlist/ If a colon (“:”) and VS (Variation Selector) follow a Moji Jōhō Kiban database serial number, the sequence of the CJK Unified Ideograph, serving as a BC, followed by the VS, corresponds to the Moji Jōhō Kiban database serial number. Such sequences are SVSes or Moji_Joho IVSes. If a Moji Jōhō Kiban database serial 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. 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.

Examples

A total of 52,515 CJK Unified Ideographs are assigned this proposed property. Among them, 5,057 also serve as BCs for SVSes (89) and registered *Moji_Joho* IVSes (11,384, which is all of them). The following table indicates—on a per-block basis—the number of CJK Unified Ideographs that are assigned this proposed property, along with the percentage that this number represents within each block as of Unicode Version 15.0:

	URO	Compatibility	Extension A	Extension B	Extension C	Extension D	Extension E	Extension F
Characters	18,253	12	5,855	25,744	387	117	502	1,645
Percent	87%	3%	89%	60%	9%	53%	9%	22%

The following are three prototypical examples of the proposed property that use U+4E00 一, U+5304 匄, and U+5606 嘆 as their BCs:

```
U+4E00   kMojiJoho   MJ006294
U+5304   kMojiJoho   MJ007755 MJ007755:E0100 MJ007756:E0101
U+5606   kMojiJoho   MJ008578 MJ030251:FE00 MJ008578:E0102 MJ030251:E0103
```

The Moji Jōhō Kiban database serial numbers that are shown in red specify which registered *Moji_Joho* IVS corresponds to the default form. There are 5,057 such instances in this proposed property.

The Moji Jōhō Kiban database serial numbers that are shown in cyan demonstrate that SVSes always have a corresponding registered *Moji_Joho* IVS. There are 89 such instances in this proposed property.

Excluded Moji Jōhō Kiban database serial numbers

The following three (3) characters are assigned Moji Jōhō Kiban database serial numbers, but were explicitly excluded from this property, because they are not CJK Unified Ideographs: U+3005 々 IDEOGRAPHIC ITERATION MARK (MJ000001), U+3006 〆 IDEOGRAPHIC CLOSING

MARK (MJ000002), and U+303B 〻 VERTICAL IDEOGRAPHIC ITERATION MARK (MJ000003). U+3005 々 looks like U+206A4 𠚪 (Extension B), which is not referenced in the Moji Jōhō Kiban database, U+3006 〆 is related to U+4E44 乄 (URO), which has its own Moji Jōhō Kiban database serial numbers (MJ006376 and MJ006377), and U+303B 〻 has no equivalent CJK Unified ideograph.

The following 19 Moji Jōhō Kiban database serial numbers have been necessarily excluded from this property, because they have no mappings, and the preferred Moji Jōhō Kiban database serial number, which has a mapping and is included in this proposed property, is provided in parentheses (those marked in red lack a URL with any meaningful content): MJ003719 (MJ003718), MJ006065 (MJ006064), MJ014004 (MJ014005), MJ029825 (MJ029826), MJ029893 (MJ029894), MJ033216 (MJ033215), MJ035887 (MJ035886), MJ037229 (MJ068077), MJ037904 (MJ037903), MJ037910 (MJ037909), MJ040282 (MJ040281), MJ040579 (MJ040580), MJ041326 (MJ041325), MJ041470 (MJ041469), MJ042077 (MJ068085), MJ043219 (MJ043218), MJ053026 (MJ053025), MJ055353 (MJ055352), and MJ059043 (MJ059042).

Data file

The *kMojiJoho-data.txt* data file, which is a PDF attachment, provides everything that is necessary for adding this property to the Unihan Database, and for adding a full description of the property to UAX #38, *Unicode Han Database (Unihan)*.

Unihan Database Lookup

Because the full records of the Moji Jōhō Kiban database are online, and use URLs that terminate with their serial numbers, the online Unihan Database Lookup tool can be updated to add links to the Moji Jōhō Kiban database. The following is the URL for the Moji Jōhō Kiban database record for U+5263 剣 whose serial number is MJ007553:

`https://moji.or.jp/mojikibansearch/info?MJ%E6%96%87%E5%AD%97%E5%9B%B3%E5%BD%A2%E5%90%8D=`**`MJ007553`**

Relationship with the normative kIRG_JSource property

The normative *kIRG_JSource* property includes the “JMJ-” prefix (*Moji Joho Kiban Project* (文字情報基盤整備事業)). As of Unicode Version 15.0, only 1,647 CJK Unified Ideographs have a *kIRG_JSource* property value with the “JMJ-” prefix. 1,645 of them are in Extension F (Unicode Version 10.0), and the remaining two were appended to the URO (Unicode Version 11.0).

The following indented paragraphs capture the complete text of a CJK Type Blog article that was published on 2018-01-27 in which I briefly examined the Moji Jōhō Kiban database. This article provided two data files that could be used to prepare a massive horizontal extension, and to completely replace existing *kIRG_JSource* property values that use the “JA-” (*Unified Japanese IT Vendors Contemporary Ideographs, 1993*) and “JH-” (*Hanyo-Denshi Program* (汎用電子情報交換環境整備プログラム)*, 2002-2009*) prefixes with “JMJ-” ones.

> As evidenced by the very last paragraph of IRG N1964 (aka L2/13-192), which was discussed during IRG #41 that took place in Tōkyō, Japan at the end of 2013, I have been curious as to why many ideographs that are commonly used in Japan lack a UAX #38 kIRG_JSource property value. As suggested by this recent tweet, I have been thinking about this again…

I first checked these three ideographs—U+9592 閒, U+9AD9 髙 & U+20BB7 𠮷—against the official search page for the Moji Jōhō Kiban Project (文字情報基盤整備事業) and found that they could use JMJ-027430, JMJ-028902, and JMJ-032129, respectively, as their kIRG_JSource property values, but then figured this may be the tip of the proverbial iceberg. After locating and working with the latest data files, I was correct.

The Unihan Database currently includes 16,224 ideographs that have a kIRG_JSource property value. I found that there are 36,427 ideographs in the Moji Jōhō Kiban Project that lack a kIRG_JSource property value. 36,416 of these are CJK Unified Ideographs, and the remaining 11 are CJK Compatibility Ideographs. This would represent a rather massive **horizontal extension**, which entails adding new source references to existing ideographs (see Sections 2.2.1.e and 2.2.1.f of IRG N2275, which is Version 10 of the IRG's Principles & Procedures). Please see the data file that I prepared.

This suggested horizontal extension also represents a good opportunity for Japan to get rid of the kIRG_JSource property's "JA" (*Unified Japanese IT Vendors Contemporary Ideographs*, 1993) source prefix altogether, because the 575 remaining ones have corresponding Moji Jōhō Kiban Project source references. Like what was done when JIS X 0213 source references replaced "JA" ones ("JA3" and "JA4" were used), a new kIRG_JSource source prefix, such as "JAMJ," should be used to indicate that they formerly had "JA" source references, and the 575 "JA" source references should then be moved to the "kJa" property. The same treatment can be applied to all 107 ideographs that use the "JH" (*Hanyo-Denshi Program*, 2002-2009) source prefix, though a new UAX #38 property, such as "kJh," would need to be defined in order to preserve the 107 original source references. Please see the data file that I prepared.

That is all.