Technical Reports

L2/25-109

Proposed Draft Unicode® Standard Annex #60

DATA FOR NON HAN IDEOGRAPHIC SCRIPTS

Version	Unicode 18.0.0
Editor	Michel Suignard
Date	2025-02-14
This Version	https://www.unicode.org/reports/tr60/tr60-1.html
Previous Version	
Latest Version	https://www.unicode.org/reports/tr60/
Latest Proposed Update	https://www.unicode.org/reports/tr60/proposed.html
Revision	1

Summary

This document describes the Sources and other ancillary data for non Han Ideographic Scripts, including Jurchen, Nüshu, and Tangut.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***A Unicode Standard Annex (UAX)** forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document corresponds to the version of the Unicode Standard of which it forms a part.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "Common References for Unicode Standard Annexes." For the latest version of the Unicode Standard, see [Unicode]. For a list of current Unicode Technical Reports, see [Reports]. For more information about versions of the Unicode Standard, see [Versions]. For any errata which may apply to this annex, see [Errata].

Contents

1 Introduction

This document is a guide to information including sources and other ancillary data related to ideographic scripts other than Han. Historically, a summary and often incomplete version of that information was provided in the data file preambles related to these scripts. This document formalizes these elements in a structure similar to what is done for Han characters in UAX #38 Unihan Han Database information. In common with Han ideographs, elements of these other ideographic scripts are encoded using algorithmic names, including the name of the script and a multi-digit notation indicating the hexadecimal value of the code point, therefore providing little information about the identity of the character. The ancillary data provided by the related data files define additional information such as the various sources for the ideograph identity, and other ancillary information, such as the reading and radical-stroke index. While sources are always provided, the ancillary information varies between scripts. Similar to the Unihan database, this information could grow in the future, such as adding sources or other type of data related to specific code points.

The scripts covered by this document include Jurchen, Nüshu, and Tangut, referred as 'covered scripts' in the rest of this document. Note that while another East Asian encoded script, Khitan Small Script, had properties documented in the various encoding proposals, especially https://www.unicode.org/L2/L2016/16113r-n4725r-khitan-small-script.pdf, they were not surfaced in any Unicode data files. Nothing precludes their addition in the future, as it would improve the knowledge related to that script.

This document is a guide to these data files, one per covered script, describing their mechanics, the nature of their contents, and the status of the various properties. One the main goal of this document is to provide a single point of reference for all property information related to the covered scripts.

2 Mechanics

2.1 Data files Design

The data files consist of a number of fields containing data for each of the covered script's ideographs included in the Unicode Standard. The fields, all of which correspond to properties, have names that consist entirely of ASCII letters and digits with no spaces or other punctuation except for underscore. For historical reasons, they all start with a lowercase `k`.

All data in these data files is stored in UTF-8 using Normalization Form C (NFC). Note, however, that the “Syntax” descriptions below, used for validation of property values, operate on Normalization Form D (NFD), primarily because that makes the regular expressions simpler.

2.2 Data files for Jurchen, Nüshu, and Tangut

Included with the [UCD] are three files called `JurchenSources.txt`, `NushuSources.txt`, and `TangutSources.txt`. These files are single text files, in UTF-8, NFC, and using Unix line endings which contain the values for all properties related to each of the covered scripts. Properties are described by categories in this document but are nevertheless included in a single file per script (unlike, for example the Unihan database which is made of multiple files for the Han script). All properties use a 'k' prefix followed by the four-letter abbreviated version of the script name as described in `PropertyValueAliases.txt`. For example, for the Tangut script, the prefix is 'kTANG', and an example of property value is `kTANG_MergedSrc`.

Review Note: There is ongoing work to clarify the status of these data sets in term of Unicode properties. These data files currently contain data which is only in scope for the script they are

addressing. In that aspect they are different from typical Unicode properties which encompass the whole Unicode repertoire. As such, they were not subject to the typical constraints of Unicode properties, such stability, consistency, etc. By moving these defintion in a UAX, the use of their status as Normative or Informative creates a stability requirement that may not be desired. As this moment, only properties that are essential to the identities are qualified as 'normative', all others are qualified as 'provisional'. this will also requires an update in UAX #44 concerning the description of these properties.

In this file, blank lines may be ignored; lines beginning with # are comment lines used to provide the header and footer. Each of the remaining lines is one entry, with three, tab-separated fields: the Unicode Scalar Value, the property name, and the value for the property for the given Unicode Scalar Value. For most of the properties, if multiple values are possible, the values are separated by spaces. No ideograph may have more than one instance of a given property associated with it, and no empty properties are included in these data files.

There is no formal limit on the lengths of any of the property values. Any Unicode character may be used in the property values except for control characters (especially tab, newline, and carriage return).

The data lines are sorted by Unicode Scalar Value and property-type as primary and secondary keys, respectively.

The file's header includes a summary of the properties each of these data files contains.

3 Property Types

The data in these data files serves a multitude of purposes, and the properties are grouped into categories according to the purpose they fulfill. A general discussion of the various categories is provided here, followed by a detailed description of the individual properties, alphabetically arranged. Among these categories, because the source information is essential in determining identity for characters which have algorithmically constructed names, the status of source related properties is 'normative'; all other properties have a 'provisional status.

3.1 Sources

Sources are among the normative parts of these data files and refer to ideograph collections which identifies encoded characters. These sources are defined as `kNSHU_DubenSrc` for Nüshu, `kTANG_MergedSrc` for Tangut, and `kJURC_Src` for Jurchen. These sources were typically documented in the encoding proposals for these scripts. Detailed descriptions of the syntax used for these sources are to be found in Section 4, *Script Properties*, below.

3.2 Radical-Stroke Counts

Two of the scripts include radical-stroke counts: Jurchen with `kJURC_RSUnicode` and Tangut with `kTANG_RSUnicode`. All the radical-stroke properties used here are loosely derived from the radical system introduced by the 18th-century *Kangxi Dictionary* and used in the Unihan database for the Han ideographs. Each Tangut ideograph is assigned one of the 883 Tangut components, and each Jurchen ideographs is assigned one of the 51 Jurchen radicals. In all these cases, unlike Han, the component or radical assignment is never a semantic signifier; it is solely based on the ideograph's structure and is mainly meant to facilitate lookup of a specific ideograph in these large lists. The same two scripts also include a stroke count, and unlike the Han equivalent, the count includes the component or radical. It should be noted that the Nüshu repertoire is ordered by stroke counts (one to sixteen) but this is not reflected in any property.

3.3 Readings

Two of the scripts include a reading property: Jurchen with with `kJURC_NCReading` and Nüshu with `kNSHU_Reading`. Any attempt at providing a reading or set of readings for an ideograph is bound to be fraught with difficulty, because the readings will vary over time and from place to place, even within a language. However, because these readings have been documented in the encoding proposals and related to well-known sources, these are provided when available in these data files.

3.4 Other Data

This category includes properties that are typically specific to a given script. Currently only one property is defined in this category: `kJURC_Numeric` used for Jurchen.

4 Scripts Properties

Below is a listing of all properties for each of the covered scripts. Each of these lists is ordered alphabetically, with information on the property contents and syntax.

For each property we give the following information in the alphabetical listing: its *Property* tag, its Unicode *Status*, its *Category* as defined above, the Unicode version in which it was *Introduced*, its *Delimiter*, its *Syntax*, and its *Description*.

The *Property* name is the tag used in the data files to mark instances of this property.

The Unicode *Status* is either *Normative*, *Informative*, or *Provisional*, depending on whether it is a normative part of the standard, an informative part of the standard, or neither. We may also include *Deprecated* as a Unicode Status if the property is no longer to be used.

Properties which allow multiple property values have a *Delimiter* defined as “space” (`U+0020` SPACE). Properties which do not have multiple property values have this defined as “N/A.” Some properties do not currently have multiple values in the data but may do so in the future.

For most properties with multiple values, the order of the values is arbitrary and has no particular significance. The most common order in such cases is alphabetical or numerical.

Validation is done as follows: The entry is split into subentries using the *Delimiter* (if defined), and each subentry converted to Normalization Form D (NFD). The value is valid if and only if each normalized subentry matches the property’s *Syntax* regular expression. Note that any given property’s *Syntax* is not guaranteed to be stable and may change in the future.

Finally, the *Description* contains not only a description of what the property contains, but also source information, known limitations, methodology used in deriving the data, and so on.

4.1 Jurchen

The properties covered in the table are: kJURC_Numeric, kJURC_NCReading, kJURC_Src, and kJURC_RSUnicode.

Property	**kJURC_Numeric**
Status	Provisional
Category	Other Data
Introduced	18.0
Delimiter	N/A
Syntax	[1-9]\d{0,4}
Default	N
Description	Numeric value of the Jurchen character. It only applies to a few characters.

Property	**kJURC_NCReading**
Status	Provisional
Category	Readings
Introduced	18.0
Delimiter	N/A

Syntax	[^\t"]+
Default	N/A
Description	Reading given in Nǔzhēnwén Cídiǎn (Jīn), although it can be expressed in any Unicode character the value is typically a single string of Latin characters with optional parenthesis

Property	**kJURC_Src**
Status	Normative
Category	Description
Introduced	18.0
Delimiter	N/A
Syntax	NC:\d{3}\.\d{2}(,\d{3}\.\d{2})? | SJ-B\:\d{3}[A-Z]\.\d | JJ\:\d{3} | N5131\:X-\d{4}
Default	N/A
Description	The Jurchen sources are made of the following categories: NC Jīn Qǐzōng 金啓孮, Nǔzhēnwén Cídiǎn 女真文辞典 (Beijing: Wenwu chubanshe, 1984). The first number is the page number in Nǔzhēnwén Cídiǎn, the second number is the order of the entry on that page. There are multiple entries for some characters in the NC source, but this document only references a single entry for each character. SJ-B Berlin copy of the Sino-Jurchen Vocabulary. The first number is the folio, the second number is the position in # the folio. JJ Jin Guangping 金光平 and Jīn Qǐzōng 金啓孮, "Nüzhen Yuyan Wenzi Yanjiu" 真语言文字研究 (Beijing: Wenwu chubanshe, 1980). The number indicates the page reference. N5131-X Sun Bojun, Nie Hongyin, Jing Yongshi, "A Supplementary Proposal to Encode the Jurchen Characters in UCS" (WG2 N5131), The sequence number is defined in WG2 N5131.

Property	**kJURC_RSUnicode**
Status	Provisional
Category	Sources
Introduced	18.0
Delimiter	N/A
Syntax	[1-9]\d{0,1}\.[1-9]\d{0,1}
Default	N/A
Description	The first number is the radical number, and the second number is the total stroke count.

4.2 Nüshu

The properties covered in the table are: kNSHU_DubenSrc and kNSHU_Reading.

Property	**kNSHU_DubenSrc**
Status	Normative
Category	Sources
Introduced	10.0
Delimiter	N/A
Syntax	[1-9]\d\.\d{2}

Default	N/A
Description	The only source documented in the file is Nǚshū Dúběn (NSDB) 女书读本 'Nüshu Reader': the first number is the page number in the NDSB, the second number is the order of the item on that page. While other sources have been mentioned in discussion about the proposal such as Nüshu Yongzi Bijiaon[NSYZBJ] 女书用字比较 "A Comparison of characters used for writing Women's Script", they are not documented in the data file.

Property	**kNSHU_Reading**
Status	Provisional
Category	Readings
Introduced	10.0
Delimiter	N/A
Syntax	[a-z]+[1-9]\d{0,1}
Default	N/A
Description	Reading based on Nüshu Duben [NDSB], the numeric value after ascii text indicates the tones in five-degree contour tone marks

4.3 Tangut

The properties covered in the table are: kTANG_MergedSrc and kTANG_RSUnicode.

Property	**kTANG_MergedSrc**
Status	Normative
Category	Description
Introduced	9.0
Delimiter	N/A
Syntax	H2004-[AB]-\d{4} | H2021-\d{6} | L(19(86|97) | 20(06|12))-\d{4} | L2008-\d{4}([AB]|-\d{4})? | N1966-\d{3}-\d{2}[0-9A-Z]{1,2} | N5217-\d{2} | S1968-\d{4} | UTN42-\d{3}
Default	N/A
Description	The Tangut sources are made of the following categories: H2004 = Hán Xiǎománg (韓小忙), 西夏文正字研究 (Xīxiàwén Zhèngzì Yánjiū) [Research into the Correct Forms of Tangut Characters]. 2004. H2021 = Hán Xiǎománg (韓小忙), 西夏文词典: 世俗文献部分 (Xīxiàwén Cídiǎn: Shìsú Wénxiàn Bùfēn) [Tangut Word Dictionary: Secular Literature Part, 9 vols.]. 2021. WG2 N5286 2024-10-14. L1986 = Lǐ Fànwén (李範文), 同音研究 (Tóngyīn Yánjiū) [Study of the Homophones]. Yinchuan. 1986. L1997 = Lǐ Fànwén (李範文), 夏漢字典 (Xià-Hàn Zìdiàn) [Tangut-Chinese Dictionary]. Beijing. 1997. L2006 = Lǐ Fànwén (李範文), 《五音切韵》与《文海宝韵》比较研究 (Wǔyīn Qiēyùn yǔ Wénhǎi Bǎoyùn bǐjiào yánjiū), In 西夏研究 (Xīxià Yánjiū) [Western Xia Studies] no.2. Beijing. 2006

	L2008 = Lǐ Fànwén (李範文). 夏漢字典 (Xià-Hàn Zìdiàn) [Tangut-Chinese Dictionary]. Beijing, 2008. L2012 = Lǐ Fànwén, 2012 abridged edition, 2008 Tangut-Chinese Dictionary, cited in WG2 N 4724, page 2, 2014-04-21. N1966 = Nishida Tatsuo (西田龍雄), 西夏文小字典 (Seikabun Shōjiten) [Little Dictionary of Tangut], In 西夏語の研 究 (Seikago no kenkyū) [A Study of the Hsi-Hsia Language] (1964-1966) vol.2. Tokyo, 1966. N5217 = Andrew West, Proposal to encode 2 Tangut components and 28 Tangut ideographs, WG2 N5217 = L2/23-149. 2023-10-02. S1968 = Sofronov M. V. (М. В. Софронов), Грамматика тангутского языка (Grammatika tangutskogo jazyka) [Grammar of the Tangut Language]. Moscow, 1968. UTN42 = Andrew West and Viacheslav Zaytsev, Tangut Character Additions and Glyph Corrections, Unicode Technical Note #42. 2019-12-21.

Property	**kTANG_RSUnicode**
Status	Provisional
Category	Sources
Introduced	9.0
Delimiter	N/A
Syntax	[1-9]\d{0,2}\.[1-9]\d{0,1}
Default	N/A
Description	The first number is the component number, and the second number is the total stroke count.

5 History

The information presented in this document used to be partially located in preambles attached to each of the data file. It was augmented by details found in original encoding proposals for the covered scripts.

References

For references for this annex, see Unicode Standard Annex #41, “Common References for Unicode Standard Annexes.”

Acknowledgements

To be added.

Modifications

Revision 1

- **Proposed draft** of the first version of UAX#60 for Unicode 18.0.0.

Previous revisions will be accessed with the “Previous Version” link in the header when appropriate.