**Proposed Update** Unicode® Standard Annex #42

UNICODE CHARACTER DATABASE IN XML

<table>
<thead>
<tr>
<th>Version</th>
<th>Unicode 15.1.0 (draft 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editors</td>
<td>Eric Muller (<a href="mailto:eric.muller@efele.net">eric.muller@efele.net</a>) Laurențiu Iancu (<a href="mailto:liancu@unicode.org">liancu@unicode.org</a>)</td>
</tr>
<tr>
<td>Date</td>
<td>2023-05-24</td>
</tr>
<tr>
<td>This Version</td>
<td><a href="https://www.unicode.org/reports/tr42/tr42-33.html">https://www.unicode.org/reports/tr42/tr42-33.html</a></td>
</tr>
<tr>
<td>Previous Version</td>
<td><a href="https://www.unicode.org/reports/tr42/tr42-32.html">https://www.unicode.org/reports/tr42/tr42-32.html</a></td>
</tr>
<tr>
<td>Latest Version</td>
<td><a href="https://www.unicode.org/reports/tr42/">https://www.unicode.org/reports/tr42/</a></td>
</tr>
<tr>
<td>Latest Proposed Update</td>
<td><a href="https://www.unicode.org/reports/tr42/proposed.html">https://www.unicode.org/reports/tr42/proposed.html</a></td>
</tr>
<tr>
<td>Schema</td>
<td><a href="https://www.unicode.org/reports/tr42/tr42-33.rnc">https://www.unicode.org/reports/tr42/tr42-33.rnc</a></td>
</tr>
<tr>
<td>Revision</td>
<td>33</td>
</tr>
</tbody>
</table>

**Summary**

This annex describes an XML representation of the Unicode Character Database.

**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This is a draft document which may be updated, replaced, or superseded by other documents at any time.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
2 Overall schema
   2.1 General principles
   2.2 Namespace
   2.3 Datatypes
   2.4 Root Element
   2.5 Common attributes
   2.6 Ordering of elements
3 Description
4 Repertoire
   4.1 Sets of code points
   4.2 Code point types
   4.3 Group
   4.4 Properties
      4.4.1 Age property
      4.4.2 Name properties
      4.4.3 Name Aliases
      4.4.4 Block
      4.4.5 General Category
      4.4.6 Combining properties
      4.4.7 Bidirectionality properties
      4.4.8 Decomposition properties
      4.4.9 Numeric Properties
      4.4.10 Joining properties
      4.4.11 Linebreak properties
      4.4.12 East Asian Width property
      4.4.13 Case properties
      4.4.14 Script properties
      4.4.15 ISO Comment properties
      4.4.16 Hangul properties
      4.4.17 Indic properties
1 Introduction

In working on Unicode implementations, it is often useful to access the full content of the Unicode Character Database (UCD). For example, in establishing mappings from characters to glyphs in fonts, it is convenient to see the character scalar value, the character name, the character East Asian width, along with the shape and metrics of the proposed glyph to map to; looking at all this data simultaneously helps in evaluating the mapping.

Directly accessing the data files that constitute the UCD is sometimes a daunting proposition. The data is dispersed in a number of files of various formats, and there are just enough peculiarities (all justified by the processing power available at the time the UCD representation was designed) to require a fairly intimate knowledge of the data format itself, in addition to the meaning of the data.

Many programming environments (for example, Java or ICU) do give access to the UCD. However, those environments tend to lag behind releases of the standard, or support only some of the UCD content.

Unibook is a wonderful tool to explore the UCD and in many cases is just the ticket; however, it is difficult to use when the task at hand has not been built-in, or when non-UCD data is to be displayed as well.

This annex presents an alternative representation of the UCD, which is meant to overcome these difficulties. We have chosen an XML representation, because parsing becomes a non-issue; there are a number of XML parsers freely available, and using them is often fairly easy. In addition, there are freely available tools that can perform powerful operations on XML data: for example, XPATH and XQUERY engines can be thought of as a “grep” for XML data and XSLT engines can be thought of as “awk” for XML data.

It is important to note that we are interested in exploring the content of the UCD, rather than in using the UCD data to process character streams. Thus, we are not concerned so much by the speed of processing or the size of our representation.

Our representation supports the creation of documents that represent only parts of the UCD, either by not representing all the characters, or by not representing all the properties. This can be useful when only some of the data is needed.

This annex presents only the XML representation format of the UCD. The data itself is part of the Unicode Character Database.

2 Overall schema

2.1 General principles

Our schema can be used to create and validate documents which are intended to represent properties of Unicode code points, blocks, named sequences, normalization corrections, standardized variants, CJK radicals and emoji sources. A document may represent the values actually assigned in a given version of the UCD, or it may represent a draft version of the UCD, or a private agreement on Private Use characters. The validity of a XML document with respect to the schema defined in this annex does not assert anything about the correctness of the values.

Valid documents may provide values for only some of the code points, or some of the Unicode properties. Furthermore, they may also incorporate non-Unicode properties.

Our schema is defined using English. However, a useful subset of the validity constraints can be captured using a schema language, thereby simplifying the task of validating documents. We have chosen Relax NG [ISO 19757], in the compact syntax, as the schema language. It is important to stress that the schema which is defined in English imposes more constraints on the documents than can be validated with the Relax NG schema.

An important characteristic of Relax NG is that its schemas do not modify or augment the infoset of the documents. Therefore, it is possible to process our XML representation without using the schema. Also, the schema is relatively straightforward and can be converted mechanically to other schema languages.

While our XML representation is not intended to be used during processing of characters and strings, it is still a design principle for our schema to support the relatively efficient representation of the UCD. This is achieved by an inheritance mechanism, similar to property inheritance in CSS or in XSL:FO (see section 4.3 Group).

Many invariants impose constraints on the values of the different properties for a given code point. For example, if the value of the Numeric Type property is None, then the value of the Numeric Value property should be the empty string; and if the value of the Other Alphabetic property is true, then the value of the Alphabetic property should be true. Those invariants are not captured in the schema.

2.2 Namespace

The namespace for our elements is “http://www.unicode.org/ns/2003/ucd/1.0”. Our attributes are in the empty namespace.
In all our examples, we assume that this namespace is the default one.

2.3 Datatypes

We use a standard XML Schema datatypes:

```xml
[datatype for code points, 3] =
  single-code-point = xsd:string { pattern = "([1-9A-F][1-9A-F])*" }
  one-or-more-code-points = list { single-code-point + }
  zero-or-more-code-points = list { single-code-point * }
  two-code-points = list { single-code-point, single-code-point }
```

2.4 Root Element

The root element of valid documents is a `ucd`.

```xml
[description, 6] =
  ucd.content &=
    element description { text }?
```

4 Repertoire

The `repertoire` child element of the `ucd` element describes the code points and their properties. As we will see shortly, code points can be described individually or as part of a group:

```xml
[set of code points, 8] =
  set-of-code-points =
    attribute cp { single-code-point }
    | ( attribute first-cp { single-code-point },
      attribute last-cp { single-code-point } )
```

In the `repertoire`, there must be at most one `code-point` element for a given code point.
4.2 Code point types

When thinking about Unicode code points, it is useful to split them into four types:

- those assigned to abstract characters (PUA or not)
- the noncharacters
- the surrogate code points
- the reserved code points

This leads to four elements to describe sets of code points:

\[
\text{[Code points, 9]} =\]
\[
\begin{align*}
\text{code-point} &= \text{element reserved} \\
&\quad \text{set-of-code-points, code-point-attributes} \\
\text{code-point} &= \text{element noncharacter} \\
&\quad \text{set-of-code-points, code-point-attributes} \\
\text{code-point} &= \text{element surrogate} \\
&\quad \text{set-of-code-points, code-point-attributes} \\
\text{code-point} &= \text{element char} \\
&\quad \text{set-of-code-points, code-point-attributes}
\end{align*}
\]

4.3 Group

While we already recognized the situation where a set of code points have exactly the same set of property values, another common situation is that of code points which have almost all the same property values.

For example, the characters U+1740 BUHID LETTER A .. U+1753 BUHID VOWEL SIGN U all have the age “3.2”, and all have the script “Buhd”. On the one hand, it is convenient to support data files in which those properties are explicitly listed with every code point, at this makes answering questions like “what is the age of U+1749?” easier, because that data is expressed right there. On the other hand, this leads to rather large data files, and it also tends to obscure the differences between similar characters.

Our representation accounts for this situation with the notion of groups. A \text{group} element is simply a container of code points that also holds default values for the properties. If a code point inside a \text{group} does not list explicitly a property but the \text{group} lists it, then the code point inherits that property from its \text{group}. For example, the fragment with explicit properties:

\[
\begin{align*}
&\text{<char cp="1740" age="3.2" na="BUHID LETTER A" gc="Lo" sc="Buhd"/>} \\
&\text{<char cp="1741" age="3.2" na="BUHID LETTER I" gc="Lo" sc="Buhd"/>} \\
&\text{<char cp="1752" age="3.2" na="BUHID VOWEL SIGN I" gc="Mn" sc="Buhd"/>} \\
&\text{<char cp="1820" age="3.0" na="MONGOLIAN LETTER A" gc="Lo" sc="Mong"/>}
\end{align*}
\]

is equivalent to this fragment which uses a \text{group}:

\[
\begin{align*}
&\text{<group age="3.2" gc="Lo" sc="Buhd">} \\
&\text{<char cp="1740"/>} \\
&\text{<char cp="1741"/>} \\
&\text{<char cp="1752"/>} \\
&\text{<char cp="1820"/>}
\end{align*}
\]

The element for U+1740 does not have the \text{age} attribute, and it therefore inherits it from its enclosing \text{group} element, that is “3.2”. On the other hand, the element for U+1820 does have this attribute, so the value is “3.0”.

As this example illustrates, the notion of \text{group} does not necessarily align with the notion of Unicode block. It is entirely defined and limited to our representation. In particular, the value of a property for a code point can always be determined from the XML document alone, assuming that this property and this code point are expressed at all. Of course, one may create an XML representation where the groups happen to coincide with the Unicode blocks.

Groups cannot be nested. The motivation for this limitation is to make the life of consumers easier: either a property is defined by the element for a code point, or it is defined by the immediately enclosing \text{group} element.

\[
\text{[groups, 10]} = \]
\[
\text{group = element group} \\
\quad \text{code-point-attributes, code-point*}
\]

4.4 Properties

Each property, except for the Special_Case_Condition and Name_Alias properties, is represented by an attribute. In an XML data file, the absence of an attribute (may be only on some \text{code-points}) means that the document does not express the value of the corresponding property. Conversely, the presence of an attribute is an expression of the corresponding property value; the implied null value is represented by the empty string.
The Name_Alias property is represented by zero or more `name-alias` child elements. Unlike the situation for properties represented by attributes, it is not possible to determine whether all of the aliases have been represented in a data file by inspecting that data file.

The name of an attribute is the abbreviated name of the property as given in the file `PropertyAliases.txt` in version 6.1.0 of the UCD. For the Unihan properties, the name is that given in the various versions of the Unihan database (some properties are no longer present in version 6.1.0).

For catalog and enumerated properties, the values are those listed in the file `PropertyValueAliases.txt` in version 6.1.0 of the UCD; if there is an abbreviated name, it is used, otherwise the long name is used.

Note that the set of possible values for a property captured in this schema may change from one version to the next.

### 4.4.1 Age property

The `age` attribute captures the version of Unicode in which a code point was assigned to an abstract character, or made a surrogate or non-character.

```
[age, 11]
  code-point-attributes & attribute age {
    "1.1" |
    "2.0" | "2.1" |
    "3.0" | "3.1" | "3.2" |
    "4.0" | "4.1" |
    "5.0" | "5.1" | "5.2" |
    "6.0" | "6.1" | "6.2" | "6.3" |
    "7.0" |
    "8.0" |
    "9.0" |
    "10.0" |
    "11.0" |
    "12.0" | "12.1" |
    "13.0" |
    "14.0" |
    "15.0" |
    "15.1" |
    "unassigned" |
  }
```

### 4.4.2 Name properties

There are two name properties: the name given by the current version of the standard (`na`), and possibly the name this character had in version 1.0 of the standard (`na1`).

```
[name pattern, 12]
  character-name = xsd:string { pattern="([A-Za-z0-9 \#\-\(\)]*)|<control>" }

[name properties, 13]
  code-point-attributes & attribute na { character-name }?
  code-point-attributes & attribute na1 { character-name }?
```

The majority of the characters in Unicode have a name which is of the form CJK UNIFIED IDEOGRAPH-<code point>. It also happens that character names cannot contain the character U+0023 # NUMBER SIGN, so we adopted the following convention: if a code point has the attribute `na` (either directly or by inheritance from an enclosing group), then occurrences of the character # in the name are to be interpreted as the value of the code point. For example:

```xml
<char cp="3400" na="CJK UNIFIED IDEOGRAPH-3400"/>
and
<char cp="3400" na="CJK UNIFIED IDEOGRAPH-#"/>
```

are equivalent. The # can be in any position in the value of the `na` attribute. The convention also applies just as well to a set of multiple code points:

```xml
<char cp="3400" na="CJK UNIFIED IDEOGRAPH-3400"/>
<char cp="3401" na="CJK UNIFIED IDEOGRAPH-3401"/>
```

is equivalent to

```xml
<char cp="3400" na="CJK UNIFIED IDEOGRAPH-#"/>
<char cp="3401" na="CJK UNIFIED IDEOGRAPH-#"/>
```

which in turn is equivalent to:

```xml
<char first-cp="3400" last-cp="3401" na="CJK UNIFIED IDEOGRAPH-#"/>
```

### 4.4.3 Name Aliases

The Name_Alias property is represented by zero or more `name-alias` child elements:

```
[name_alias property, 14]
  code-point-attributes & element name-alias {
    attribute alias { text }?,
    attribute type { "abbreviation" | "alternate" |
      "control" | "correction" |
      "figment" }? } *
```

### 4.4.4 Block
The Block property is represented by the `blk` attribute:

```xml
[block property, 15] =
    code-point-attributes &=
    attribute blk {
        "Adlam"
        "Aegean_Numbers"
        "Ahom"
        "Alchemical"
        "Alphabetic_PF"
        "Anatolian_Hieroglyphs"
        "Ancient_Greek_Music"
        "Ancient_Greek_Numbers"
        "Ancient_Symbols"
        "Arabic"
        "Arabic_Ext_A"
        "Arabic_Ext_B"
        "Arabic_Ext_C"
        "Arabic_Math"
        "Arabic_PF_A"
        "Arabic_PF_B"
        "Arabic_Sup"
        "Armenian"
        "Arrows"
        "ASCII"
        "Avestan"
        "Bali"
        "Bamum"
        "Bamum_Sup"
        "Bassa_Vah"
        "Batak"
        "Bengali"
        "Bhaiksuki"
        "Block_Elements"
        "Bopomofo"
        "Bopomofo_Ext"
        "Box_Drawing"
        "Brahmi"
        "Braille"
        "Buginese"
        "Buhid"
        "Byzantine_Music"
        "Carian"
        "Caucasian_Albamian"
        "Cham"
        "Cherokee"
        "Cherokee_Sup"
        "Chess_Symbols"
        "Chorasmian"
        "CJK"
        "CJK_Compat"
        "CJK_Compat_Forms"
        "CJK_Compat_Idenographics"
        "CJK_Compat_Idenographics_Sup"
        "CJK_Ext_A"
        "CJK_Ext_B"
        "CJK_Ext_C"
        "CJK_Ext_D"
        "CJK_Ext_E"
        "CJK_Ext_F"
        "CJK_Ext_G"
        "CJK_Ext_H"
        "CJK_Radicals_Sup"
        "CJK_Strokes"
        "CJK_Symbols"
        "Compat_Jamo"
        "Control_Pictures"
        "Coptic"
        "Coptic_Epact_Numbers"
        "Counting_Rod"
        "Cuneiform"
        "Cuneiform_Numbers"
        "Currency_Symbols"
        "Cyriic_Syllabary"
        "Cypric_Minoan"
        "Cyriic"
        "Cyriic_Ext_A"
        "Cyriic_Ext_B"
        "Cyriic_Ext_C"
        "Cyriic_Ext_D"
        "Cyriic_Sup"
        "Deseret"
        "Devanagari"
        "Devanagari_Ext"
        "Devanagari_Ext_A"
        "Diacriticals"
        "Diacriticals_For_Symbols"
        "Diacriticals_Sup"
        "Diacriticals_Ext"
        "Dingbats"
        "Dives_Akuru"
        "Dogra"
        "Domino"
        "Duployan"
        "Early_Dynastic_Cuneiform"
        "Egyptian_Hieroglyphs"
    }
```
"Mayan_Numerals"
"Medieval_Druid"
"Mediterranean"
"Mende_Kikakui"
"Meroitic_Cursive"
"Meroitic_Hieroglyphs"
"Miao"
"Misc_Arrows"
"Misc_Math_Symbols_A"
"Misc_Math_Symbols_B"
"Misc_Pictographs"
"Misc_Symbols"
"Misc_Technical"
"Modi"
"Modifier_Letters"
"Modifier_Tone_Letters"
"Mongolian"
"Mongolian_Sup"
"Mro"
"Music"
"Multani"
"Myanmar"
"Myanmar_EXT_A"
"Myanmar_EXT_B"
"Nabataean"
"Nag_Mundari"
"Nandinagari"
"NB"
"New_Tai_Lue"
"Newa"
"NKo"
"Number_Forms"
"Nushu"
"Nyai_keng_Puachue_Hmong"
"OCR"
"Ogham"
"Ol_Chiki"
"Old_Hungarian"
"Old_italic"
"Old_North_Arabian"
"Old_Permic"
"Old_Persian"
"Old_Sogdian"
"Old_South_Arabian"
"Old_Turkic"
"Old_Uyghur"
"Oriya"
"Ornamental_Dingbats"
"Osage"
"Osmanya"
"Ottoman_Siyaq_Numbers"
"Pahawh_Hmong"
"Palmyrene"
"Pau_Cin_Hau"
"Phags_Pa"
"Phaistos"
"Phoenician"
"Phonetic_Ext"
"Phonetic_Ext_Sup"
"Psalter_Pahlavi"
"PUA"
"Punctuation"
"Rejang"
"Rumi"
"Runic"
"Samaritana"
"Saurashtra"
"Sharada"
"Shavian"
"Shorthand_Format_Controls"
"Siddham"
"Sinhala"
"Sinhala_Archaic_Numbers"
"Small_Forms"
"Small_Kana_Ext"
"Sogdian"
"Sora_Sompeng"
"Soyombo"
"Specials"
"Sundanese"
"Sundanese_Sup"
"Sup_Arrows_A"
"Sup_Arrows_B"
"Sup_Arrows_C"
"Sup_Math_Operators"
"Sup_PUA_A"
"Sup_PUA_B"
"Sup_Punctuation"
"Sup_Symbols_And_Pictographs"
"Super_And_Sub"
"Sutton_SignWriting"
"Sylo Nhi"
"Symbols_And_Pictographs_EXT_A"
"Symbols_For_Legacy_Computing"
"Syriac"
"Syriac_Sup"
4.4.5 General Category

The general category is represented by the gc attribute.

```
<gc attribute gc { "Lu" | "Ll" | "Lt" | "Lm" | "Lo" |
  "Mn" | "Mc" | "Me" |
  "Nd" | "Nl" | "No" |
  "Pc" | "Pd" | "Ps" | "Pe" | "Pf" | "Po" |
  "Sm" | "Sc" | "Sk" | "So" |
  "Zs" | "Zl" | "Zp" |
  "Cc" | "Cf" | "Cs" | "Co" | "Cn" }?
```

4.4.6 Combining properties

The combining class is represented by the ccc attribute, which holds the decimal representation of the combining class.

Because the set of values that this property has taken across the various versions of the UCD is rather large, our schema does not restrict the possible values to those actually used.

```
<ccc attribute ccc { xsd:integer { minInclusive="0" maxInclusive="254" }?
```

4.4.7 Bidirectionality properties

The bidirectional class is represented by the bc attribute.

```
<bc attribute bc { "AL" | "AN" |
  "B" | "BN" |
  "CS" |
  "EN" | "ES" | "ET" |
  "FSI" |
  "L" | "LRE" | "LRI" | "LRO" |
  "NKM" |
  "ON" |
  "PDI" |
  "R" | "RLE" | "RLI" | "RLO" |
  "S" |
  "WS" }
```

The mirrored property is represented by the bidi_M attribute, which takes a boolean value.

```
<bidi_M attribute bidi_M { boolean }?
```
The `bmg` attribute is the code point of a character whose glyph is typically a mirrored image of the glyph for the current character.

```xml
[bmg property, 20] =
  code-point-attributes &
  attribute bmg { "" | single-code-point }?
```

Note that we do not express the “Best Fit” element recorded in BidiMirroring.txt. For one thing, it is not meant to be machine readable. More importantly, the idea underlying the mirrored glyph is delicate to use, since it makes assumptions about the design of the fonts, and the best fit goes even farther.

The Bidi_Control property is represented by the `bidi_c` attribute.

```xml
[Bidi_C property, 21] =
  code-point-attributes &
  attribute bidi_c { boolean }?
```

The bidi paired bracket type and bidi paired bracket properties are represented by the `bpt` and `bpb` attributes respectively.

```xml
[bpt and bpb attributes, 22] =
  code-point-attributes &
  attribute bpt { "o" | "c" | "n" }?
  code-point-attributes &
  attribute bpb { "#" | single-code-point }?
```

4.4.8 Decomposition properties

The decomposition type and decomposition mapping properties are represented by the `dt` and `dm` attributes.

Most characters have a decomposition mapping to themselves. This is very similar to the situation we encountered with names, and we adopted a similar convention: if the value of a decomposition mapping is the character itself, we use the attribute value # (U+0023 # NUMBER SIGN) as a shorthand notation; this enables those attributes to be captured in groups.

```xml
[decomposition properties, 23] =
  code-point-attributes &
  attribute dt { "can" | "com" | "enc" | "fin" | "font" | "fra"
             | "init" | "iso" | "med" | "nar" | "nb" | "sml"
             | "sqr" | "sub" | "sup" | "vert" | "wide" | "none" }?
  code-point-attributes &
  attribute dm { "#" | zero-or-more-code-points }?
```

The properties Composition_Exclusion and Full_Composition_Exclusion are represented by the attributes `CE` and `Comp_Ex`:

```xml
[composition properties, 24] =
  code-point-attributes &
  attribute CE { boolean }?
  code-point-attributes &
  attribute Comp_Ex { boolean }?
```

The properties NFC_Quick_Check, NFD_Quick_Check, NFKC_Quick_Check, NFKD_Quick_Check, Expands_On_NFC, Expands_On_NFD, Expands_On_NFKC, Expands_On_NFKD, FC_NFKC_Closure have corresponding attributes.

```xml
[quick check properties, 25] =
  code-point-attributes &
  attribute NFC_QC { "Y" | "N" | "M" }?
  code-point-attributes &
  attribute NFD_QC { "Y" | "N" }?
  code-point-attributes &
  attribute NFKC_QC { "Y" | "N" | "M" }?
  code-point-attributes &
  attribute NFKD_QC { "Y" | "N" | "M" }?
  code-point-attributes &
  attribute XO_NFC { boolean }?
  code-point-attributes &
  attribute XO_NFD { boolean }?
  code-point-attributes &
  attribute XO_NFKC { boolean }?
  code-point-attributes &
  attribute XO_NFKD { boolean }?
  code-point-attributes &
  attribute FC_NFKC { "#" | one-or-more-code-points }?
```

4.4.9 Numeric Properties

The numeric type is represented by the `nt` attribute.

The numeric value is represented by the `nv` attribute, represented as a fraction.
4.4.10 Joining properties

The joining class of a character is represented by the \texttt{jt} attribute.

The \texttt{jg} attribute is the joining group of the character.

4.4.11 Linebreak properties

The linebreak property is represented by the \texttt{lb} attribute.
### 4.4.12 East Asian Width property

The East Asian width property is represented by the `ea` attribute.

```
[ea property, 30] =
  code-point-attributes &=
    attribute ea { "A" | "F" | "H" | "N" | "Na" | "W" }?
```

### 4.4.13 Case properties

The Uppercase, Lowercase, Other_Uppercase and Other_Lowercase properties are represented by corresponding attributes.

```
[casing properties, 31] =
  code-point-attributes &=
    attribute Upper { boolean }?
  code-point-attributes &=
    attribute Lower { boolean }?
  code-point-attributes &=
    attribute OUpper { boolean }?
  code-point-attributes &=
    attribute OLower { boolean }?
```

Most characters have a case mapping and case folding properties that simply map or fold to themselves. This is very similar to the situation we encountered with names, and we adopted a similar convention: if the value of a case mapping or case folding property is the character itself, we use the attribute value `#` (U+0023 # NUMBER SIGN) as a shorthand notation; this enables those attributes to be captured in groups.

The simple case mappings are recorded in the `suc`, `slc`, `stc` attributes.

```
[casing properties, 32] =
  code-point-attributes &=
    attribute suc { "#" | single-code-point }?
  code-point-attributes &=
    attribute slc { "#" | single-code-point }?
  code-point-attributes &=
    attribute stc { "#" | single-code-point }?
```

The non-simple casing are recorded in the `uc`, `lc` and `tc` attributes.

```
[casing properties, 33] =
  code-point-attributes &=
    attribute uc { "#" | one-or-more-code-points }?
  code-point-attributes &=
    attribute lc { "#" | one-or-more-code-points }?
  code-point-attributes &=
    attribute tc { "#" | one-or-more-code-points }?
```

The Simple_Case_Folding and Case_Folding properties are recorded in the `scf` and `cf` attributes respectively.

```
[casing properties, 34] =
  code-point-attributes &=
    attribute scf { "#" | single-code-point }?
  code-point-attributes &=
    attribute cf { "#" | one-or-more-code-points }?
```

The Case_Ignorable, Cased, Changes_When_Casefolded, Changes_When_Casemapped, Changes_When_Lowercased, Changes_When_NFKC_Casefolded, Changes_When_Titlecased, Changes_When_Uppercased, NFKC_Casefold, and NFKC_Simple_Casefold properties are recorded in these attributes:

```
[casing properties, 35] =
  code-point-attributes &=
    attribute CI { boolean }?
  code-point-attributes &=
    attribute Cased { boolean }?
  code-point-attributes &=
    attribute scf { "#" | single-code-point }?
  code-point-attributes &=
    attribute cf { "#" | one-or-more-code-points }?
```

https://www.unicode.org/reports/tr42/tr42-33.html
attribute CWCF { boolean }?
code-point-attributes &= attribute CWCM { boolean }?
code-point-attributes &= attribute CWL { boolean }?
code-point-attributes &= attribute CWKCF { boolean }?
code-point-attributes &= attribute CWT { boolean }?
code-point-attributes &= attribute CWU { boolean }?
code-point-attributes &= attribute NFKC_CF { "#" | zero-or-more-code-points }?
code-point-attributes &= attribute NFKC_SCF { "#" | zero-or-more-code-points }?

Note that the UCD records more information about case folding than is expressed in the properties, specifically the entries in CaseFolding.txt with status T.

4.4.14 Script properties

The script and script extension properties are represented by the sc and scx attributes respectively.

[script property, 36] =
script = "Adlm" | "Aghb" | "Ahom" | "Arab" | "Arni" | "Armn" | "Avst"
| "Bali" | "Bamu" | "Bass" | "Batk" | "Beng" | "Bhks"
| "Bopo" | "Brhm" | "Bugl" | "Buhd" |
| "Cakm" | "Cans" | "Car2" | "Cham" | "Cher" | "Chs" | "Copt"
| "Cpsr" | "Cpty" | "Cyr1" | "Cyr2" | "Dacr" | "Dnpl"
| "Deva" | "Diak" | "Dnog" | "Ebre" | "Ethi"
| "Eglk" | "Gong" | "Germ" | "Goth" | "Gran" | "Grek"
| "Gujr" | "Guru" | "Hanb" | "Hatr" | "Hebr" | "Hira" | "Hlwn"
| "Hmng" | "Hnnp" | "Hrkt" | "Hung"
| "Ital" | "Java"
| "Kali" | "Kana" | "Kawi" | "Khor" | "Khmr" | "Khoj" | "Kits"
| "Knnd" | "Kthi"
| "Lana" | "Laoo" | "Latn" | "Lepc" | "Limb" | "Lina" | "Linb"
| "Lisu" | "Lyci" | "Lyd1"
| "Mhar" | "Maka" | "Mand" | "Mani" | "Marc"
| "Mdra" | "Mndh" | "Merc" | "Mero" | "Mlym"
| "Mroi" | "Mong" | "Mroo" | "Mtei" | "Mult" | "Nhm1"
| "Ngrm" | "Ngr2" | "Nhat" | "Nlao" | "Nkoo" | "Nshu"
| "Ogsm" | "Olck" | "Orkh" | "Orya" | "Osge" | "Osma" | "Ougr"
| "Pala" | "Paua" | "Pepc" | "Phag" | "Phil" | "Phlp" | "Phnx"
| "Plmr" | "Prt1" | "Qaai"
| "Rogh" | "Rjng" | "Runr"
| "Sarb" | "Saur" | "Sgmw" | "Shaw" | "Shrd" | "Sidd"
| "Sind" | "Sinh" | "Sogd" | "Sogo" | "Sora" | "Soyo" | "Sund"
| "Sylo" | "Sprc" | "Tabr" | "Takr" | "Tamo" | "Tang" | "Tavr"
| "Telu" | "Tfng" | "Tglg" | "Thaa" | "Thai" | "Tibb" | "Tirh"
| "Umio" | "Toto" | "Ugar"
| "Vaii" | "Vith"
| "War2" | "Wcho"
| "Xeo3" | "Xsux"
| "Yezi" | "Yi11" | "Zamb"
| "Zinh" | "Zyyy" | "Zzzz"

code-point-attributes &= attribute sc { script }?
code-point-attributes &= attribute scx { list { script + } }?

4.4.15 ISO Comment properties

The ISO 10646 comment field is represented by the isc attribute.

[isc property, 37] =
code-point-attributes &= attribute isc { text }?

4.4.16 Hangul properties

The property Hangul_Syllable_Type is represented by the hst attribute.

[hst property, 38] =
code-point-attributes &= attribute hst { "L" | "LV" | "LVT" | "T" | "V" | "WH" }?

The property Jamo_Short_Name is represented by the jsn attribute:
4.4.17 Indic properties

The property Indic_Syllabic_Category is represented by the InSC attribute.

```xml
<jamo property, 39> =
code-point-attributes &
attribute InSC { xml:IDNS { pattern="[A-Z][0-3]" } }?
```

4.4.18 Identifier and Pattern and programming language properties

The properties ID_Start, Other_ID_Start, XID_Start, ID_Continue, Other_ID_Continue, and XID_Continue are represented by corresponding attributes:
The properties Pattern_Syntax and Pattern_White_Space are represented by corresponding attributes:

```
[pattern properties, 44] =
  code-point-attributes &
  attribute Pat_Syn { boolean }?
  code-point-attributes &
  attribute Pat_WS { boolean }?
```

4.4.19 Properties related to function and graphic characteristics

The properties Dash, Hyphen, Quotation_Mark, Terminal_Punctuation, Sentence_Terminal, Diacritic, Extender, Soft_Dotted, Alphabetic, Other_Alphabetic, Math, Other_Math, Hex_Digit, ASCII_Hex_Digit, Default_Ignorable_Code_Point, Other_Default_Ignorable_Code_Point, Logical_Order_Exception, Prepended_Concatenation_Mark, White_Space, Vertical_Orientation and Regional_Indicator describe the function or graphic characteristic of a character, and have each a corresponding attribute.

```
[properties related to function and graphic characteristics, 45] =
  code-point-attributes &
  attribute Dash { boolean }?
  code-point-attributes &
  attribute Hyphen { boolean }?
  code-point-attributes &
  attribute QMark { boolean }?
  code-point-attributes &
  attribute Term { boolean }?
  code-point-attributes &
  attribute STerm { boolean }?
  code-point-attributes &
  attribute Dia { boolean }?
  code-point-attributes &
  attribute Ext { boolean }?
  code-point-attributes &
  attribute PCM { boolean }?
  code-point-attributes &
  attribute SD { boolean }?
  code-point-attributes &
  attribute Alpha { boolean }?
  code-point-attributes &
  attribute OAlpha { boolean }?
  code-point-attributes &
  attribute Math { boolean }?
  code-point-attributes &
  attribute OMath { boolean }?
  code-point-attributes &
  attribute Hex { boolean }?
  code-point-attributes &
  attribute AHex { boolean }?
  code-point-attributes &
  attribute DI { boolean }?
  code-point-attributes &
  attribute OD1 { boolean }?
  code-point-attributes &
  attribute LOE { boolean }?
```
attribute WSpace { boolean }?

code-point-attributes &=
    attribute vo { "U" | "K" | "Tu" | "Tt" }?

code-point-attributes &=
    attribute RI { boolean }?

4.4.20 Properties related to boundaries

The properties Grapheme_Base, Grapheme_Extend, Other_Grapheme_Extend, Grapheme_Link, Grapheme_Cluster_Break, Word_Break and Sentence_Break each have a corresponding attribute:

[properties related to boundaries, 46] =

code-point-attributes &=
    attribute Gr_Base { boolean }?

code-point-attributes &=
    attribute Gr_Ext { boolean }?

code-point-attributes &=
    attribute Gr_Link { boolean }?

code-point-attributes &=
    attribute GCB { "CN" | "CR" | "EB" | "EBG" | "EM" | "EX" | "GAZ" | "L" | "LF" | "LV" | "LVT" | "PP" | "RI" | "SM" | "T" | "V" | "XX" | "ZwJ" }?

code-point-attributes &=
    attribute WB { "CR" | "DQ" | "EB" | "EBG" | "EM" | "EX" | "Extend" | "FO" | "GAZ" | "HL" | "KA" | "LE" | "LF" | "MB" | "ML" | "MN" | "NL" | "NU" | "RI" | "SQ" | "WSegSpace" | "XX" | "ZHO" }?

code-point-attributes &=
    attribute SB { "AT" | "CB" | "CL" | "CR" | "EX" | "FO" | "GAZ" | "HL" | "KA" | "LE" | "LF" | "MB" | "ML" | "MN" | "NL" | "NU" | "RI" | "SC" | "SE" | "SP" | "ST" | "UP" | "XX" }?

4.4.21 Properties related to ideographs

The properties Ideographic, Unified_Ideograph, Equivalent_Unified_Ideograph, IDS_Binary_Operator, IDS_Trinary_Operator and Radical have corresponding attributes:

[properties related to ideographs, 47] =

code-point-attributes &=
    attribute Ideo { boolean }?

code-point-attributes &=
    attribute UIdeo { boolean }?

code-point-attributes &=
    attribute EqUIdeo { single-code-point }?

code-point-attributes &=
    attribute IDSB { boolean }?

code-point-attributes &=
    attribute IDST { boolean }?

code-point-attributes &=
    attribute IDSU { boolean }?

code-point-attributes &=
    attribute Radical { boolean }?
4.4.22 Miscellaneous properties

The properties Deprecated, Variation_Selector, and Noncharacter_Code_Point have corresponding attributes:

\[
\text{[miscellaneous properties, 48]} = \\
\begin{align*}
\text{code-point-attributes } &= \text{attribute Dep } \{ \text{boolean } \}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute VS } \{ \text{boolean } \}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute NChar } \{ \text{boolean } \}\?
\end{align*}
\]

4.4.23 Unihan properties

The Unihan properties (from the Unihan database) are represented as attributes.

\[
\text{[Unihan properties, 49]} = \\
\begin{align*}
\text{code-point-attributes } &= \text{attribute kAccountingNumeric } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9]+"}\}\}? \\
\text{code-point-attributes } &= \text{attribute kAlternateHanYu } \\
& \quad \{ \text{text }\}\? \#old \\
\text{code-point-attributes } &= \text{attribute kAlternateJEF } \\
& \quad \{ \text{text }\}\? \#old \\
\text{code-point-attributes } &= \text{attribute kAlternateKangXi } \\
& \quad \{ \text{text }\}\? \\
\text{code-point-attributes } &= \text{attribute kAlternateMorohashi } \\
& \quad \{ \text{text }\}\? \\
\text{code-point-attributes } &= \text{attribute kAlternateTotalStrokes } \\
& \quad \{ \text{"-" | list } \{ \text{xsd:string } \{ \text{pattern="[0-9]-[BHJKMPSUV]+"}\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kBigFive } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCCCCII } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9A-F]{6}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCNS1986 } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[12E]-[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCNS1992 } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[123]-[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCangjie } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[A-Z]+"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCheungBauer } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[a-z][1-6]"}\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kCheungBauerIndex } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[0-9]-[0-9][1-6]"}\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kCihaiY } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[1-9][0-9][0-9][0-9][0-9][0-9][0-9]\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kCompatibilityVariant } \\
& \quad \{ \" | \text{xsd:string } \{ \text{pattern="U+2?[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kCowles } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[-][0-9][1,2][\.[0-9]]?\} +}\}? \\
\text{code-point-attributes } &= \text{attribute kDaeJaweon } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9][4,6,8][0-9][0-158]\} +}\}
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kDefinition } \\
& \quad \{ \text{text }\}\? \\
\text{code-point-attributes } &= \text{attribute kEACC } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kFenn } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[0-9]+[A-KP]+\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kFennIndex } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[0-9]-[0-9][0-9][0-9][0-9]\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kFourCornerCode } \\
& \quad \{ \text{list } \{ \text{xsd:string } \{ \text{pattern="[0-9][4,6,8][\.[0-9]]?\} +}\}\? \\
\text{code-point-attributes } &= \text{attribute kFrequency } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[1-5]\} +}\}\?
\end{align*}
\]

\[
\begin{align*}
\text{code-point-attributes } &= \text{attribute kBGB0 } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9A-F]{4}"\}}\}? \\
\text{code-point-attributes } &= \text{attribute kBGB1 } \\
& \quad \{ \text{xsd:string } \{ \text{pattern="[0-9A-F]{4}"\}}\}?
\end{align*}
\]
code-point-attributes &= attribute kGB3
  ( xsd:string {pattern="[0-9A-F]{4}\}\]})

code-point-attributes &= attribute kGB5
  ( xsd:string {pattern="[0-9A-F]{4}\}\]})

code-point-attributes &= attribute kGB7
  ( xsd:string {pattern="[0-9A-F]{4}\}\]})

code-point-attributes &= attribute kGB8
  ( xsd:string {pattern="[0-9]{4}\}\]})

code-point-attributes &= attribute kGradeLevel
  ( xsd:string {pattern="[1-6]\]})

code-point-attributes &= attribute kGSR
  ( list ( xsd:string {pattern="[0-9]{4}\]\]}) )

code-point-attributes &= attribute kHangul
  ( text )

code-point-attributes &= attribute kHanYu
  ( list ( xsd:string {pattern="[a-z\-\^\’\""]\]\]}) )

code-point-attributes &= attribute kHangyuPinlu
  ( list ( xsd:string {pattern="[0-9]{4}\]\]}) )

code-point-attributes &= attribute kHanyuPinlinu
  ( list ( xsd:string {pattern="[0-9]{4}\]\]}) )

code-point-attributes &= attribute kHanyuPinyin
  ( list ( xsd:string {pattern="[0-9]{5}\]\]}) )

code-point-attributes &= attribute kHDZRadBreak
  ( xsd:string {pattern="[U\+2?\[0-9A-F]{4}\]\]})

code-point-attributes &= attribute kHKGlyph
  ( list ( xsd:string {pattern="[0-9]{4}\]\]}) )

code-point-attributes &= attribute kHKSCS
  ( xsd:string {pattern="[0-9A-F]{4}\]\]})

code-point-attributes &= attribute kIICore
  ( xsd:string {pattern="[1-9]\]\]})

code-point-attributes &= attribute kIRGDaeJaweon
  ( xsd:string {pattern="[0-9]{4}\]\]})

code-point-attributes &= attribute kIRGDaiKanwaZiten
  ( xsd:string {pattern="[0-9]{5}\]\]})

code-point-attributes &= attribute kIRGKangXi
  ( xsd:string {pattern="[01]\[0-9]{3}\]\]})

code-point-attributes &= attribute kIRG_GSource
  ( list ( xsd:string {pattern="G0-[0-9A-F]{4}\]\]}) )

...
code-point-attributes &= attribute kIRG_USOURCE
{ "" | xsd:string {pattern="U+2\[0-9A-F\]{4}\(UTC\[0-9\]{5}\)"
| xsd:string {pattern="USAT\[0-9\]{5}\)"
|} }

code-point-attributes &= attribute kIRG_UKSOURCE
{ "" | xsd:string {pattern="UK\[0-9\]{5}\)"
|} }

code-point-attributes &= attribute kIRG_SOURCE
{ "" | xsd:string {pattern="U\+2\[0-9A-F\]{4}\(UTC\[0-9\]{5}\)"
| xsd:string {pattern="USAT\[0-9\]{5}\)"
|} }

code-point-attributes &= attribute kJa
{ xsd:string {pattern="[0-9A-F]{4}S?"}
|} }

code-point-attributes &= attribute kJapanese
{ list { xsd:string {pattern="[A-Z]+"} +}
|} }

code-point-attributes &= attribute kJHJ
{ text
|} }

code-point-attributes &= attribute kJoyoKanji
{ xsd:string {pattern="(20[0-9]{2})|(U\+2\[0-9A-F\]{4})?"
|} }

code-point-attributes &= attribute kKangXi
{ list { xsd:string {pattern="\.[0-9]{2}[01]"}
|} }

code-point-attributes &= attribute kKPS0
{ xsd:string {pattern="[0-9A-F]{4}"
|} }

code-point-attributes &= attribute kKPS1
{ xsd:string {pattern="[0-9A-F]{4}"
|} }

code-point-attributes &= attribute kKSC0
{ xsd:string {pattern="[0-9]{4}"
|} }

code-point-attributes &= attribute kKSC1
{ xsd:string {pattern="[0-9]{4}"
|} }

code-point-attributes &= attribute kMainlandTelegraph
{ xsd:string {pattern="[0-9]{4}"
|} }

https://www.unicode.org/reports/tr42/tr42-33.html
```
code-point-attributes &= attribute kMandarin
  { list { xsd:string {pattern="[A-ZÜ]+[1-5]"}
       | xsd:string {pattern="[a-زل]+"}} }?

code-point-attributes &= attribute kMatthews
  { xsd:string {pattern="[0-9]{1,4}(a|\.)?"}}?

code-point-attributes &= attribute kMeyerWempe
  { list { xsd:string {pattern="[1-9][0-9,\(][a-٦]\)*"}} }?

code-point-attributes &= attribute kMojioho
  { list { xsd:string {pattern="MJ[0-9]{6},(FE0[01]|E0)[010][0-9-A-F]"}+} }?

code-point-attributes &= attribute kMorohashi
  { list { xsd:string {pattern="(\[0-9]{5}'{0,2}|H[0-9]{3})(:(FE0[01]|E0\[0-9A-F])?"}} +} }?

code-point-attributes &= attribute kNelson
  { list { xsd:string {pattern="[0-9]{4}"}} +}?

code-point-attributes &= attribute kOtherNumeric
  { list { xsd:string {pattern="[0-9]+"}} +}?

code-point-attributes &= attribute kPhonetic
  { list { xsd:string {pattern="[1-9]\[0-9\]{0,3}\[A-Dx\]^?"}} +}?

code-point-attributes &= attribute kPrimaryNumeric
  { list { xsd:string {pattern="[0-9]{4}"}} +}?

```
4.4.24 Tangut data

The Tangut data are represented as attributes.

[Tangut data, 50] =

```xml
<target>
  code-point-attributes &= attribute kRSTUnicode
  ( xsd:string {pattern="[0-9]+.[0-9]+"} )
</target>
```

4.4.25 Nushu data

The Nushu data are represented as attributes.

[Nushu data, 51] =

```xml
<target>
  code-point-attributes &= attribute kSrc_NushuDuben
  ( xsd:string {pattern="[0-9]+.[0-9]+"} )
</target>
```

4.4.26 Emoji properties

The Emoji properties are represented as attributes.

[Emoji properties, 52] =

```xml
<target>
  code-point-attributes &= attribute Emoji { boolean }
  code-point-attributes &= attribute EPres { boolean }
  code-point-attributes &= attribute EMod { boolean }
  code-point-attributes &= attribute EBase { boolean }
  code-point-attributes &= attribute EComp { boolean }
  code-point-attributes &= attribute ExtPict { boolean }
</target>
```

5 Blocks

The `blocks` child of the `ucd` describes the blocks. It has one child `block` element per block, with attributes to describe the extent and name of the block.

[blocks, 53] =

```xml
<target>
  ucd.content &=
  element blocks {
    element block {
      attribute first-cp { single-code-point },
      attribute last-cp { single-code-point },
      attribute name { text } ) +
    }
  }
</target>
```

6 Named Sequences

The `named-sequences` child of the `ucd` describes the named sequences. It has one child `named-sequence` element per named sequence, with attributes to describe the name and sequence.

[named sequences, 54] =

```xml
<target>
  ucd.content &=
</target>
```
element named-sequences {
  element named-sequence {
    attribute cps { one-or-more-code-points },
    attribute name { text }) + }?
}

ucd.content &=
  element provisional-named-sequences {
  element named-sequence {
    attribute cps { one-or-more-code-points },
    attribute name { text }) + }?
}

7 Normalization Corrections

The normalization-corrections child of the ucd describes the normalization corrections. It has one child normalization-correction element per correction, with attributes to describe the code point affected, its old normalization, its new normalization and the version of Unicode in which the correction was made.

[normalization corrections, 55] =
ucd.content &=
  element normalization-corrections {
    element normalization-correction {
      attribute cp { single-code-point },
      attribute old { one-or-more-code-points },
      attribute new { one-or-more-code-points },
      attribute version { text }) + }
}

8 Standardized Variants

The standardized-variants child of the ucd describes the standardized variant. It has one child element standardized-variant per variant. The attributes on that last element capture the variation sequence, the description of the desired appearance, and the shaping environment under which the appearance is different.

[standardized variants, 56] =
ucd.content &=
  element standardized-variants {
    element standardized-variant {
      attribute cps { two-code-points },
      attribute desc { text },
      attribute when { text }) + }
}

9 CJK Radicals

The cjk-radicals child of the ucd describes the CJK radicals. It has one child element cjk-radical per radical. The attributes on that last element capture the radical number, the corresponding CJK radical character, and the corresponding CJK unified ideograph.

[cjk radicals, 57] =
ucd.content &=
  element cjk-radicals {
    element cjk-radical {
      attribute number { xsd:string {pattern="[0-9]{1,3}'{0,2}"}},
      attribute radical { single-code-point? },
      attribute ideograph { single-code-point }) + }
}

10 Emoji sources

The emoji-sources child of the ucd describes the emoji sources.

[datatype for code points, 58] =
  jis-code-point = xsd:string { pattern = "[0-9A-F]{4}" }

[emoji sources, 59] =
ucd.content &=
  element emoji-sources {
    attribute unicode { one-or-more-code-points },
    attribute docomo { jis-code-point? },
    attribute kddi { jis-code-point? },
    attribute softbank { jis-code-point? }) + }

11 The full schema

Our schema is just the accumulation of the pieces we have described so far:

[UCD RelaxNG schema, 60] =
  [namespace declaration: 1]
  [datatypes: 2, 3, 12, 58]
  [schema start: 4]
  [boolean type: 5]
  [description: 6]
  [repertoire: 7, 8, 9, 10]
  [attributes: 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]
  [Tangut data: 50]
  [Nushu data: 51]
  [blocks: 53]
  [named sequences: 54]
  [normalization corrections: 55]
  [standardized variants: 56]
An expanded version is linked from the top of this document.

12 Examples

Here is a fragment of the UCD for a few representative characters (only some of the properties are represented):

```xml
<ucd xmlns="http://www.unicode.org/ns/2003/ucd/1.0">
  <repertoire>
    <char cp="001F" age="1.1" na="&lt;control&gt;" gc="Cc" bc="S" lb="CM"/>
    <char cp="0020" age="1.1" na="SPACE" gc="Zs" bc="WS" ea="Na" lb="SP"/>
    <char cp="0026" age="1.1" na="AMPERSAND" gc="Po" bc="ON" ea="Na"/>
    <char cp="0028" age="1.1" na="LEFT PARENTHESIS" na1="OPENING PARENTHESIS" gc="Ps" bc="ON" Bidi_M="y" bmg="0029" ea="Na" lb="OP"/>
    <char cp="0041" age="1.1" na="LATIN CAPITAL LETTER A" gc="Lu" ssc="0061" ea="Na" sc="Latn"/>
    <char cp="AC00" age="2.0" na="HANGUL SYLLABLE GA" gc="Lo" dt="can" dm="1100 1161" ea="W" lb="ID" sc="Hang"/>
    <char cp="20094" age="3.1" na="CJK UNIFIED IDEOGRAPH-20094" gc="Lo" ea="W" lb="ID" sc="Hani" kIRG_GSource="KX" kIRGHanyuDaZidian="10036.060" kIRG_TSource="5-214E" kRSUnicode="4.3" kIRGKangXi="0082.090"/>
    <group age="3.2" gc="Lo" sc="Buhd">
      <char cp="1740" na="BUHID LETTER A"/>
      <char cp="1741" na="BUHID LETTER I"/>
      <char cp="1752" na="BUHID VOWEL SIGN I" gc="Mn"/>
      <char cp="1820" age="3.0" na="MONGOLIAN LETTER A" sc="Mong"/>
    </group>
  </repertoire>
</ucd>

Acknowledgments

Thanks to Markus Scherer and Mark Davis for their help developing this XML representation. Thanks to the reviewers: Julie Allen, Ernest van den Boogaard, Daniel Bünzli, John Cowan, Asmus Freytag, Felix Sasaki, Andrew West.

Modifications

This section indicates the changes introduced by each revision.

Revision 33

- New value for the age attribute: 15.1.
- New value for the blk attribute: CJK_Ext_I.
- New values for the lb attribute: AK, AP, AS, VF, VI.
- Modified values for the number, radical attributes of the cjk-radical element.
- Changed single value into list for the nv code point attribute.
- New code point attributes: ID_Compat_Math_Continue, ID_Compat_Math_Start, IDSU, NFKC_SCF.
- Modified patterns for the kBigFive, kIRG_GSource, kMorohashi, kRSUnicode attributes.
- Changed single values into lists for the kMorohashi, kPrimaryNumeric Unihan attributes.

Revision 32

- New value for the age attribute: 15.8.
- New values for the script attribute: Kawi, Nagm.
- New Unihan attribute: kAlternateTotalStrokes.
- Modified patterns for the `kIRG_GSource, kIRG_HSource, kIRG_TSource, kSemanticVariant, kSpecializedSemanticVariant, kZVariant attributes.

Revision 31 being a proposed update, only changes between revisions 30 and 32 are noted here.

Revision 30
- New value for the `age` attribute: 14.8.
- New values for the `script` attribute: Cpmn, Ougr, Tnsa, Toto, Vith.
- New values for the `jg` attribute: Thin_Yeh, Vertical_Tail.
- New Unihan attribute: kStrange.
- Modified patterns for the `kIRG_GSource, kIRG_HSource, kIRG_VSource, kPhonetic, kSpoofingVariant attributes.
- Removal of the `kWubi` attribute, which has never been present in released versions of the UCD.

Revision 29 being a proposed update, only changes between revisions 28 and 30 are noted here.

Revision 28
- New value for the `age` attribute: 13.8.
- New values for the `blk` attribute: Chorasmian, CJK_Ext_G, Dives_Akuru, Khitan_Small_Script, Lisu_Sup, Symbols_For_Legacy_Computing, Tangut_Sup, Yezidi.
- New values for the `script` attribute: Chrs, Diak, Kits, Yezi.
- New value for the `InPC` attribute: Top_And_Bottom_And_Left.
- New Unihan attributes `kSpoofingVariant, kUnihanCore2020, kIRG_SSsource, kIRG_UKsource, KTGH2013`.
- Modified patterns for the `kIRG_GSource, kIRG_HSource, kIRG_KPsource, kIRG_KSource, kIRG_TSource, kKangXi, kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTraditionalVariant attributes.`

Revision 27 being a proposed update, only changes between revisions 26 and 28 are noted here.

Revision 26
- New value for the `age` attribute: 12.1.

Revision 25
- New value for the `age` attribute: 12.0.
- New values for the `script` attribute: Elym, Hemp, Nand, Wcho.
- New values for the `blk` attribute: Egyptian_Hieroglyph_Format_Controls, Elymaic, Nandinagari, Nyiakeng_Puachue_Hmong, Ottoman_Siyaq_Numbers, Small_Kana_Ext, Symbols_And_Pictographs_Ext_A, Tamil_Sup, Vanch.
- New Unihan attributes `kSemanticVariant, kIRG_KSource, kIRG_TSource, kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTraditionalVariant attributes.`

Revision 24 being a proposed update, only changes between revisions 23 and 25 are noted here.

Revision 23
- New value for the `age` attribute: 11.8.
- New values for the `blk` attribute: Chess_Symbols, Dogra, Georgian_Ext, Gunjala_Gondi, Hanifi_Rohingya, Indic_Siyaq_Numbers, Makasar, Mayan_Numerals, Medefaidrin, Old_Sogdian, Sogdian.
- New values for the `script` attribute: Dogr, Gong, Maka, Medf, Rohg, Sogd, Sogo.
- New values for the `jg` attribute: Hanifi_Rohingya_Kinna_Ya, Hanifi_Rohingya_Pa.
- New value for the `we` attribute: kSegSpace.
- New values for the `InSC` attribute: Consonant_Initial_Postfixed.
- New attributes: EqUIdeo, kJinmeiyoKanji, kJoyoKanji, kKoreanEducationKanja, kKoreanName, kTGH.
- Modified patterns for the `kTGT_MergedSrc` attribute.
- Modified patterns for the `kIRG_GSource, kIRG_HSource and kIRG_VSource attributes.`

Revision 22 being a proposed update, only changes between revisions 21 and 23 are noted here.

Revision 21
- New value for the `age` attribute: 10.8.
- New values for the `blk` attribute: CJK_Ext_F, Kana_Ext_A, Masaram_Gondi, Nushu, Syriac_Sup, Zanabazar_Square.
- New values for the `sc` attribute: Gonn, Nshu, Soyo, Zanb.
- New values for the `jg` attribute: Malayalam_Nga, Malayalam_Ja, Malayalam_Nya, Malayalam_Tta, Malayalam_Nna, Malayalam_Nnna, Malayalam_Rha, Malayalam_Lla, Malayalam_Llla, Malayalam_Ssa.
- New value for the `InPC` attribute: Bottom_And_Left.
- Modified patterns for the `kIRG_GSource`, `kIRG_JSource`, `kIRG_KSource` attributes.
- New code point attributes: vo, RI
- New code point attributes for Nushu data: kSrc_NushuDuben and kReading.

Revision 20 being a proposed update, only changes between revisions 19 and 21 are noted here.

Revision 19
- New value for the `age` attribute: 9.0.
- New values for the `sc` attribute: Adlm, Bhks, Marc, Newa, Osge, Tang.
- New values for the `blk` attribute: Adlam, Bhaiksuki, Cyrillic_Ext_C, Ideographic_Symbols, Marchen, Mongolian_Sup, Newa, Osage, Tangut, Tangut_Components.
- New values for the `gcb` attribute: EB, EBG, EM, GAZ, ZWJ.
- New values for the `wb` attribute: EB, EBG, EM, GAZ, ZWJ.
- New values for the `lb` attribute: EB, EM, ZWJ.
- New code point attributes: PCM, kRSUnicode and kTGT_MergedSrc.
- Modified patterns for the `kRSUnicode`, `kRSKangXi`, `kMandarin`, `kIRG_JSource`, `kIRG_USource` and `kFennIndex` attributes.

Revision 18 being a proposed update, only changes between revisions 17 and 19 are noted here.

Revision 17
- New value for the `age` attribute: 8.0.
- New values for the `sc` attribute: Ahom, Hatr, Hluw, Hung, Mult, Sgnw.
- New values for the `blk` attribute: Ahom, Anatolian_Hieroglyphs, Cherokee_Sup, CJK_Ext_E, Early_Dynastic_Cuneiform, Hatran, Multani, Old_Hungarian, Sup_Symbols_And_Pictographs, Sutton_SignWriting.
- New values for the `InSC` attribute: Consonant_Killer, Consonant_Prefixed, Consonant_With_Stacker, Syllable_Modifier.
- New code point attributes: InPC, kJa.
- New patterns for the `kIRG_GSource` attribute: GFC-, GGFZ-.

Revision 16 being a proposed update, only changes between revisions 15 and 17 are noted here.

Revision 15
- New value for the `age` attribute: 7.0.
- New values for the `jg` attribute.
- New values for the `sc` attribute.
- New values for the `blk` attribute.
- New values for the `InSC` attribute.
- New values for the `kIICore` attribute.
- New values for the `kIRG_GSource` attribute.

Revision 14 being a proposed update, only changes between revisions 13 and 15 are noted here.

Revision 13
- New value for the `age` attribute: 6.3.
- New values DQ, HL, SQ for the `wb` attribute (for Unicode 6.3).
- New code point attributes bpt and bpb (for Unicode 6.3).
- New values for the `bc` attribute: LRI, LRI, FSI, PDI (for Unicode 6.3).
- Updated the patterns for kHanyuPinlu and kTotalStrokes (for Unicode 6.3).
- Updated the patterns for kIRG_HSource and kIRG_HSource (for Unicode 6.2).
- Clarified that the child elements list-like elements are in no particular order.

Revision 12 being a proposed update, only changes between revisions 11 and 13 are noted here.

Revision 11
- New value for the age attribute: 6.2.
- New value for the gb, gb and lb attributes: RI (for Unicode 6.2).
- Updated the patterns for kIRG_GSource and kIRG_HSource (for Unicode 6.2).

Revision 10 being a proposed update, only changes between revisions 9 and 11 are noted here.

Revision 9
- Clarified the default values.
- Indicate that property values may change from one release to the next.
- Introduced the blk attributes, for the Block property.
- Introduced the scx attribute, for the ScriptExtensions property.
- Introduced the name-alias element, for the Name_Alias property.
- New value for the age attribute: 6.1.
- New values for the script attribute: Cakm, Merc, Mero, Plrd, Shrd, Sora, Takr.
- New values for the lb attribute: HL and CI.
- New value for the jg attribute: Rohingya_Yeh.
- The value of the fc_nfkc attribute must now be either # or one-or-more-code-points.
- For the nv attribute, the absence of a numeric value is now represented by NaN rather than by the empty string.
- The values of the ccc are now restricted to 0..254, instead of 0..255.
- Updated the patterns for kSemanticVariant, kSpecializedSemanticVariant, kIRG_USource and kMandarin.

Revision 8 being a proposed update, only changes between revisions 7 and 9 are noted here.

Revision 7
- New value for the age attribute: 6.8.
- New value for the jg attribute: Teh_Marbuta_Goal
- New values for the script attribute: Batk, Brah, Mand.
- Updated the patterns for kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KSource, kIRG_MSource, kIRG_TSource, kIRG_VSource.
- Added the InSC and InMC elements.
- Added the emoji-sources element.

Revision 6 being a proposed update, only changes between revisions 5 and 7 are noted here.

Revision 5
- Changed the type of block/@first-cp, block/@last-cp and normalization-corrections/@cp from text to single-code-point
- Changed the type of named-sequence/@cps, provisional-named-sequences/@cps, normalization-correction/@old and normalization-correction/@new from text to one-or-more-code-points.
- Changed the type of standardized-variants/@cps from text to two-code-points.
- New values for the jg attribute: Farsi_Yeh and Nya.
- New value for the age attribute: 5.2.
- New values for the sc attribute: Lana, Tavt, Avst, Egyp, Samr, Lisu, Banu, Java, Mtei, Armi, Sarh, Prti, Phli, Orkh, Kthi.
- New value for the lb attribute: CP.
- New value for the sc attribute: Zinh.
- New code point attributes CI, Cased, Cwcf, Cwcm, Cwcf, Cwt, Cwu, NFkc_CF.
- New attributes kHanyuPinlu and kIRG_HSource.
New element cjk-radicals

Updated the patterns for kIRG_GSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_VSource, kHanyuPinlu, kMandarin, kSemanticVariant, kSpecializedSemanticVariant, kVietnamese, kZVariant.

Point out that Relax NG schemas do not modify or augment the infoset, and that it is possible to convert mechanically our schema to other schema languages.

Revision 4 being a proposed update, only changes between revisions 3 and 5 are noted here.

Revision 3

First approved version, for Unicode 5.1.0.

For optional elements which acts as collections, such as repertoire and named-sequences, impose that there be at least one element in the collection.

Remove the constraint that the value jg is limited when jt has certain values; similarly for bmg/Bidi_M and for nv/nt.

Value NL added to the wa attribute (for Unicode 5.1).

Value PP added to the gcb attribute (for Unicode 5.1).

Corrected the vai script value to Vaii.

Removed the discussion of elements or attributes in different namespace.

Removed the code-point element.

Revision 2

Promoted to Draft UAX.

Changed the title from "An XML representation of the UCD"

Value 5.1 added to the age attribute (for Unicode 5.1).

Value SM added to the gcb attribute (for Unicode 5.1).

Values CR, Extend, LF, MB added to the wa attribute (for Unicode 5.1).

Values CR, EX, LF, SC added to the sa attribute (for Unicode 5.1).

Value Burushaski_Yeh_Barree added to the jg attribute (for Unicode 5.1).

Value Alef_Maqsurah added to the jg attribute (for Unicode 2.x).

Values Cari, Cham, Kali, Lepc, Lyci, Lydi, Olck, Rjng, Saur, Sund and vai added to the sc attribute (for Unicode 5.0).

jamo attribute renamed to JSN

sfc attribute renamed to scf

Attribute kXHC1983 added (for Unicode 5.1.0).

Pattern for attribute kIRG_USource extended (for Unicode 5.1.0).

Element provisional-named-sequences added (for Unicode 5.0)

Revision 1

First working draft.