

**Proposed Draft** Unicode® Technical Standard #61**UNICODE SET NOTATION**

Version	1 (draft 4)
Editors	Robin Leroy ( <a href="mailto:eggrobin@unicode.org">eggrobin@unicode.org</a> )
Date	2026-03-06
This Version	<a href="https://www.unicode.org/reports/tr61/tr61-1.html">https://www.unicode.org/reports/tr61/tr61-1.html</a>
Previous Version	n/a
Latest Version	<a href="https://www.unicode.org/reports/tr61/">https://www.unicode.org/reports/tr61/</a>
Latest Proposed Update	<a href="https://www.unicode.org/reports/tr61/proposed.html">https://www.unicode.org/reports/tr61/proposed.html</a>
Revision	1

**Summary**

The description of Unicode properties and algorithms frequently requires referring to sets of code points and strings defined using property assignments. This document defines a notation for such sets. The notation is machine-readable and can be used in APIs.

**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This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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 Technical Standard (UTS)** is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS.

Please submit corrigenda and other comments with the online reporting form [[Feedback](#)]. Related information that is useful in understanding this document is found in the [References](#). 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](#)].

**Contents**

- 1 [Introduction](#)
  - 1.1 [Terminology and Notation](#)
- 2 [Lexical Elements](#)
  - 2.1 [Literal Elements](#)
    - 2.1.1 [Semantics](#)
  - 2.2 [Escaped Elements](#)
    - 2.2.1 [Semantics](#)
  - 2.3 [Named Elements](#)
    - 2.3.1 [Semantics](#)
  - 2.4 [Bracketed Elements and Strings](#)
    - 2.4.1 [Semantics](#)
  - 2.5 [Property Queries](#)
    - 2.5.1 [Negations](#)
    - 2.5.2 [Unary Queries](#)
    - 2.5.3 [Binary Queries](#)
      - 2.5.3.1 [Age Queries](#)
      - 2.5.3.2 [Property Comparisons](#)
      - 2.5.3.3 [Identity and Null Queries](#)
      - 2.5.3.4 [Valid Values and Resolved Sets](#)
      - 2.5.3.5 [Property Value Queries](#)
      - 2.5.3.6 [Regular Expression Queries](#)
- 3 [Set Operations](#)
  - 3.1 [Semantics](#)
- 4 [Conformance](#)
- 5 [Use in APIs](#)
- 6 [Use in Higher-Level Syntaxes](#)
- 7 [Best Practices](#)
  - 7.1 [Escaping](#)
  - 7.2 [Bidirectional display](#)
  - 7.3 [Style Guide for Unicode Specifications](#)
- [References](#)
- [Acknowledgements](#)
- [Modifications](#)

## 1 Introduction

Sets of code points can be defined by reference to their properties; for instance:

1. “the characters with the property `XID_Continue`”
2. “the characters whose `Line_Break` property value is `OP` and whose `East_Asian_Width` property value is neither `F`, `W`, nor `H`”
3. “the characters that have the `Other_ID_Start` property, or the `Other_ID_Continue` property, or whose `General_Category` value is one of `Nl`, `Mn`, `Mc`, `Nd`, `Pc`, or one of those in the `L` grouping, but that have neither the `Pattern_Syntax` property nor the `Pattern_White_Space` property.”
4. “the characters whose `General_Category` value is one of `Nl`, `Mn`, `Mc`, `Nd`, `Pc`, or one of those in the `L` grouping, except for the character `U+2E2F VERTICAL TILDE`.”

These kinds of set definitions are used throughout the Unicode Standard, including its annexes, and in the Unicode Technical Standards. They are necessary to the description of Unicode algorithms, such the line breaking algorithm [UAX14] and text segmentation algorithms [UAX29], of relations between properties, as in the derivations in [UAX29], [UAX31] and [UAX44], or of syntaxes as in [UAX31] or [UTS51]. They are also omnipresent in proposals and reports used in the development of these standards.

The use of plain-language definitions of these sets, as above, can become impractical when the definitions are complicated or when the sets are used in higher-level syntaxes, such as grammar rules or regular expressions. A definition that is not machine readable also prevents its direct use in implementations, or its inspection using tooling.

This document defines a formal syntax, *UnicodeSet notation*, for finite sets of code points and strings. In this syntax, the above examples can be expressed as:

1. `\p{XID_Continue}`
2. `[\p{lb=OP}-[\p{ea=F}\p{ea=W}\p{ea=H}]]`
3. `[\p{Other_ID_Start}\p{Other_ID_Continue}\p{L}\p{Nl}\p{Mn}\p{Mc}\p{Nd}\p{Pc}-\p{Pattern_Syntax}-\p{Pattern_White_Space}]`
4. `[\p{L}\p{Nl}\p{Mn}\p{Mc}\p{Nd}\p{Pc}-[\u2E2F]]`

Besides defining sets that are useful in specifications, this notation, if implemented in a tool that displays the contents of the set, can serve as a query language for the Unicode Character Database, allowing maintainers of the standard to answer questions such as:

1. “Which characters have an `Uppercase_Mapping` that differs from their `Simple_Uppercase_Mapping`?”  
`\p{Uppercase_Mapping#@Simple_Uppercase_Mapping@}`.
2. “Which characters changed `Simple_Case_Folding` between Unicode Version 15.0 and Unicode Version 15.1?”  
`\p{U15.1:Simple_Case_Folding#@U15.0:Simple_Case_Folding@}`.
3. “Which CJK characters have the word ‘cat’ in their definition, and which Egyptian hieroglyphs have the word ‘cat’ in their description?”  
`[\p{CJKDefinition=/\bcat\b/} \p{kEH_Desc=/\bcat\b/}]`.
4. “Does `Changes_When_Casefolded` mean the same as ‘different from its `Case_Folding`’? No, the set `[\p{Case_Folding#@code point@}-\p{Changes_When_Casefolded}]` is nonempty.

The document then discusses what subsets of UnicodeSet notation is appropriate for use in APIs, and how it can be incorporated in higher-level syntaxes.

**Review Note:** This syntax, which originates in the API of the ICU class `UnicodeSet`, was previously standardized in [UTS35], see [https://unicode.org/reports/tr35/#Unicode\\_Sets](https://unicode.org/reports/tr35/#Unicode_Sets); however, it is only partially defined there, with reference to [UTS18]:

Unicode property sets are defined as described in UTS #18: Unicode Regular Expressions [UTS18], Level 1 and RL2.5, including the syntax where given. For an example of a concrete implementation of this, see [ICUUnicodeSet].

[UTS18] in turn does not formally define a syntax, but instead presents an example syntax, which differs from UnicodeSet syntax. The UAXes and UTSes that use UnicodeSet syntax currently refer to [UTS35], or sometimes incorrectly refer to [UTS18].

There are five known implementations of UnicodeSet notation maintained by the Unicode Consortium:

1. the ICU4C implementation;
2. the ICU4J implementation;
3. the implementation of the online Unicode tools (referred to as the JSPs), based on ICU4J with extensions and comprehensive property coverage;
4. the implementation used in the invariant tests in the Unicode tools, similar to the preceding one, with slightly different extensions;
5. the ICU4X experimental implementation used in the experimental transliterator module.

In addition, a syntax similar to UnicodeSet is supported by ICU4C regular expressions (but not documented), together with a syntax that uses `&&` and `--` for set operations for compatibility with Java. The Unicode Standard itself (Section A.2.1) defines a notation for sets of code points which is similar to, but different from UnicodeSet syntax. That notation uses `&&` and `--` for set operations. Many technical reports use UnicodeSet syntax instead.

In practice, any usage in CLDR has needed to lie within the common subset supported by ICU4C and ICU4J, regardless of what was written in the LDML specification. As a result, this document mostly follows the ICU4C implementation. Changes with respect to the ICU4C 78 implementation that could be in scope for implementation in ICU are highlighted in yellow in the grammar. Extensions to the ICU4C implementation that are unlikely to be in scope for implementation in ICU are shown with a gray background; these typically originate from the Unicode Tools, and are useful for the development and testing of the Unicode Standard itself, but not for general-purpose internationalization libraries. Divergences in other implementations are described in review notes.

### 1.1 Terminology and Notation

The context-free UnicodeSet syntax is described using a variant of Backus-Naur Form. Production rules are written using the sign ::=, and alternatives are separated by |. Nonterminal symbols, referred to in this document as *syntactic categories*, are written in a serif font, and are links to their definition. A monospace font is used for literal text. The symbol "" is used for the empty string. Some syntactic categories which correspond to character classes, such as [white-space](#), are defined outside of the BNF grammar.

A *construct* is a piece of text that is an instance of a syntactic category. A *constituent* of a construct is the construct itself, or any construct appearing within it. An *immediate* constituent of a construct is one that corresponds to a syntactic category appearing in the right-hand side of the production rule defining the syntactic category of the construct.

Rules shown over a gray background define syntactic categories that are not recommended for support in general-purpose APIs. See [Section 5, Use in APIs](#).

**Example:** The rule

Difference ::= Restriction - UnicodeSet

defines the syntactic category [Difference](#) as consisting of a [Restriction](#), followed by the character U+002D HYPHEN-MINUS which is a [set-operator](#), followed by a [UnicodeSet](#).

In the [Difference](#) [A-Z]-[C], the [Restriction](#) [A-Z], the [set-operator](#) -, and the [UnicodeSet](#) [C] are the immediate constituent constructs of the [Difference](#); the substring [A-Z]-[ is not a construct. Parsing the constituent [Restriction](#) [A-Z] itself, it consists of [set-operators](#) [ and ] and of a [Range](#) A-Z. These are constituent constructs of the [Restriction](#) [A-Z] as well as of the [Difference](#) [A-Z]-[C].

The syntax of UnicodeSet notation is described in two parts: lexical elements, whose grammars are regular and space-sensitive, and the context-free (but not regular) grammar of the ranges and set arithmetic making up the UnicodeSet expression itself, where white space is ignored. Syntactic categories used in the grammars of lexical elements are written in [kebab-case](#); their production rules are space-sensitive. Syntactic categories used in the grammar of [UnicodeSet](#) are written in [CamelCase](#); their production rules implicitly allow for optional [white-space](#) between their constituent lexical elements.

**Example:** [ A-Z ] - [C] is a valid [Difference](#), equivalent to [A-Z]-[C].

This allows for a clear separation between lexical analysis (identifying lexical elements independently from context, which can be done using regular expressions) and syntactic analysis (building up syntactic categories up to [UnicodeSet](#) itself). In particular, this separation makes it easier to perform the insertion of left-to-right marks described in [Section 5.2, Conversion to Plain Text](#), in [Unicode Technical Standard #55, Unicode Source Code Handling \[UTS55\]](#); see also [Section 7.2, Bidirectional Display](#).

Review Note: This approach differs from the one taken in [UTS35], where white space is explicit throughout the grammar, and no distinction is made between the syntactic categories for individual characters in string literals, which should not be directionally isolated, and those for individual characters in sets.

The set of code points is finite; however, since UnicodeSets are finite sets of *strings* rather than just code points, the union of all UnicodeSets is the set of all strings, which is infinite and therefore not a UnicodeSet. In particular, one cannot define a UnicodeSet-valued complement operation  $X \mapsto CX$  on UnicodeSets satisfying  $Y \cap CX = Y \setminus X$  for all UnicodeSets  $X$  and  $Y$ .

The *code point complement*  $[\^X]$  of a UnicodeSet  $X$  is defined as the set of all code points not in  $X$ , that is,  $[\^X] := \mathbb{U} \setminus X$ , where  $\mathbb{U}$  is the set of all code points. For all sets of code points  $X$  and  $Y$ ,  $Y \cap [\^X] = Y \setminus X$ ; however, if  $Y$  contains strings of length other than 1 that are not also in  $X$ , this equality does not hold; instead  $Y \cap [\^X] = (Y \setminus X) \cap \mathbb{U}$ . Likewise, the code point complement is not an involution for sets that contain strings of length other than 1:  $[\^[\^X]] = X \cap \mathbb{U}$ , whereas  $CCX = X$  for the complement in the set of all strings.

## 2 Lexical Elements

An expression in UnicodeSet notation consists of a sequence of separate *lexical elements*. Each lexical element is either a [set-operator](#), a [literal-element](#), an [escaped-element](#), a [named-element](#), a [bracketed-element](#), a [string-literal](#), or a [property-query](#).

In this grammar, [white-space](#) is defined as any character with the Pattern\_White\_Space property. One or more [white-space](#) character is allowed between any two adjacent lexical elements; this is not indicated explicitly in the grammar for [UnicodeSet](#). An [ignorable-format-control](#) is either of the [white-space](#) characters U+200E and U+200F. At least one [white-space](#) character other than an [ignorable-format-control](#) is required between the [set-operator](#) [ and the [literal-element](#) :. If removing any [ignorable-format-control](#) characters between lexical elements changes the sequence of lexical elements, the expression is ill-formed.

**Note:** [white-space](#) is sometimes necessary to separate consecutive lexical elements. For instance, `\00` consists of a single [escaped-element](#), but `\0 0` consists of an [escaped-element](#) followed by a [literal-element](#). In that case, [ignorable-format-control](#) cannot be used to separate the lexical elements. The requirement for a space between [ and : makes it possible to analyse the internal grammar of a [property-query](#) using a lexer with conditional rules; such a lexer can treat [posix-start](#) and [perl-start](#) as tokens, and switch to a mode that expects the parts of a [property-query](#).

Review note: Existing implementations allow an [ignorable-format-control](#) to separate lexical elements. This means `[\x0DF]` (with U+200E between D and F) is the two-element set containing U+000D (carriage return) and the letter F, whereas `[\x0DF]` is the one-element set containing the letter ß. While a similar problem occurs with many more invisible characters, for instance, `[\x0DF]` is the three-element set containing carriage return, VARIATION SELECTOR-256, and the letter F, that can be mitigated by requiring that these characters be escaped; in contrast, [ignorable-format-control](#) characters are expected to be used to ensure that UnicodeSet expressions display properly, and should not be prohibited. For instance, `[0\0]` is only readable if an LRM is inserted between the letter 0 and the `\0`, yielding `[↵0]`: besides the letter 0, that set contains U+0000, not U+0030.

Each lexical element other than a [set-operator](#) represents a set of code point sequences.

A [set-operator](#) is any of &, -, [, ], and ^.

### 2.1 Literal Elements

A **literal-element** is a Unicode scalar value that does not have the `Pattern_White_Space` property, and is neither a set operator nor one of `{`, `}`, `$` or `\`.

### 2.1.1 Semantics

A **literal-element** represents a single code point: itself.

### 2.2 Escaped Elements

An **escaped-element** is defined by the following regular grammar, where:

- **escapable-character** is any Unicode scalar value other than the digits 0 through 7, the letters u, x, U, N, p, P, a, b, t, n, v, f, r, e, c, and the **ignorable-format-control** characters U+200E and U+200F.
- **ascii-printable** is any Unicode scalar value in the range U+0020–U+007E.

**escaped-element** ::=

```
\x up-to-two-hexadecimal-digits
| \u four-hexadecimal-digits
| \U0000 five-hexadecimal-digits
| \U0010 four-hexadecimal-digits
| \x{ hexadecimal-digits }
| \ up-to-three-octal-digits
| \ escapable-character
```

```
\c ascii-printable
```

```
| \a | \b | \e | \t | \n | \v | \f | \r
```

**up-to-three-octal-digits** ::=

```
octal-digit
| octal-digit octal-digit
| octal-digit octal-digit octal-digit
```

**up-to-two-hexadecimal-digits** ::=

```
hexadecimal-digit
| hexadecimal-digit hexadecimal-digit
```

**four-hexadecimal-digits** ::=

```
hexadecimal-digit hexadecimal-digit hexadecimal-digit hexadecimal-digit
```

**five-hexadecimal-digits** ::=

```
hexadecimal-digit hexadecimal-digit hexadecimal-digit hexadecimal-digit hexadecimal-digit
```

**hexadecimal-digits** ::=

```
hexadecimal-digit
| hexadecimal-digits hexadecimal-digit
```

**octal-digit** ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7

**hexadecimal-digit** ::=

```
0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
| A | B | C | D | E | F
| a | b | c | d | e | f
```

**Note:** In this grammar, **hexadecimal-digit** is not equivalent to the set of characters with the property `Hex_Digit`: the fullwidth digits and letters are not allowed in an **escaped-element**.

### 2.2.1 Semantics

An **escaped-element** represents a single code point, as follows.

1. An **escaped-element** consisting of `\` followed by an **escapable-character** represents that **escapable-character**.
2. Any **escaped-element** with constituent **octal-digits** represents the code point whose octal representation is given by its constituent **octal-digits**.
3. Any **escaped-element** with constituent **hexadecimal-digits** represents the code point whose hexadecimal representation is given by its constituent **hexadecimal-digits**.
4. If the constituent **hexadecimal-digits** do not represent a code point, the UnicodeSet expression is ill-formed.
5. An **escaped-element** with the prefix `\c` represents the bitwise AND of the code point of the constituent **ascii-printable** with 0x1F.

**Note:** The `c` stands for “control”. An **escaped-element** starting with `\c` represents one of the characters in U+0000–U+001F, which all have the `General_Category_Control`. This syntax matches a long-standing convention of mapping printable characters to these controls for input and display, especially in terminals. For instance, `Ctrl+H` can be used in many terminals to type U+0008 (BACKSPACE), and U+0008 is displayed by many command-line applications as `^H`. The **escaped-element** `\cH` accordingly represents U+0008.

**Review Note:** ICU allows any Unicode scalar value after `\c`, thus it interprets `\c$` as U+000D. These sequences are ill-formed according to the UnicodeSet syntax defined in this document. This does not prevent ICU from continuing to support them as an extension, but we should not standardize such oddities.

6. The remaining **escaped-elements** are defined by the following table.

escaped-element	Code point (name alias)
<code>\a</code>	U+0007 (ALERT)
<code>\b</code>	U+0008 (BACKSPACE)
<code>\t</code>	U+0009 (HORIZONTAL TABULATION)

<code>\n</code>	U+000A (NEW LINE)
<code>\v</code>	U+000B (VERTICAL TABULATION)
<code>\f</code>	U+000C (FORM FEED)
<code>\r</code>	U+000D (CARRIAGE RETURN)
<code>\e</code>	U+001B (ESCAPE)

**Example:** The `escaped-elements` `\\`, `\134`, `\x5C`, `\u005C`, `\x{05C}`, and `\u0000005C` all represent the code point U+005C. The `escaped-elements` `\a`, `\7`, `\x7`, `\c7`, `\cG`, and `\eG` all represent the code point U+0007. The `escaped-element` `\x{110000}` is ill-formed.

**Review Note:** [UTS35] allows for `\u{2F}` as well as `\x{2F}`, and for wholly-escaped strings with the syntax `\x{2F 2F}` (equivalent to `\x{2F}\x{2F}`). It allows optional `white-space` (including line terminators) inside the braces of a `\x{}` or `\u{}` escape. This is not supported by ICU4C, ICU4J, the JSPs, nor the invariants, but is supported by the ICU4X experimental implementation. [UTS35] does not allow for octal escapes nor for a single hexadecimal digit after `\x`, but since this is supported by ICU4C, ICU4J, and the ICU4J-based Unicode tools, as well as consistent with many programming languages, we include these in the specification.

### 2.3 Named Elements

A `named-element` is defined by the following regular grammar, where a `ucd-identifier-character` is any character in the Basic Latin block whose general category is one of Lu, Ll, Nd, Pc, Pd, or Zs, and where a `named-literal-element` is any Unicode scalar value except `:` and `}`.

```
named-element ::=
  \N{ ucd-identifier }
  | \N{ hexadecimal-digits : ucd-identifier }
  | \N{ hexadecimal-digits : named-literal-element : ucd-identifier }
ucd-identifier ::=
  ucd-identifier-character
  | ucd-identifier ucd-identifier-character
```

**Note:** In UnicodeSet notation, the set of `ucd-identifier-characters` is `[\p{block=Basic_Latin} & [\p{L}\p{Nd}\p{Pc}\p{Pd}\p{Zs}]] = [A-Za-z0-9\N{SPACE}_-]`.

#### 2.3.1 Semantics

A `named-element` represents the single character whose Name or Name Alias matches the constituent `ucd-identifier` according to loose matching rule UAX44-LM2. If there is no such character, the UnicodeSet expression is ill-formed.

If the `named-element` contains `hexadecimal-digits`, these shall be a hexadecimal representation of the code point named by the `ucd-identifier`. If it contains a `named-literal-element`, that `named-literal-element` shall be the named character.

**Examples:** The `named-elements` `\N{SPACE}`, `\N{0020:SPACE}`, and `\N{20: :SPACE}` all represent U+0020 SPACE. The `named-elements` `\N{THIS IS NOT A CHARACTER}`, `\N{0A:LATIN CAPITAL LETTER A}`, and `\N{41:a:LATIN CAPITAL LETTER A}` are ill-formed.

The `named-elements` `\N{PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET}` and `\N{PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET}` both represent U+FE18  $\text{⌈}$ . The `named-element` `\N{Latin small ligature o-e}` represents U+0153  $\text{œ}$  LATIN SMALL LIGATURE OE. The `named-element` `\N{Hangul jungseong O-E}` represents U+1180  $\text{ㅝ}$  HANGUL JUNGSEONG O-E. The `named-element` `\N{Hangul jungseong OE}` represents U+116C  $\text{ㅞ}$  HANGUL JUNGSEONG OE.

**Review Note:** The `\N` escapes with colons are innovations introduced in this document. The need for them has become apparent in the Unicode invariant tests, especially for property comparisons for character additions. They are approximated in the Unicode invariant tests by the use of `\x{code point}\N{name}`, combined in some cases with higher-level checks that the sets have the right size (this is done because earlier iterations of those tests failed to catch incorrect code points or names in draft data when they were testing only one of those). This is however quite brittle (for instance, swapped characters would not be detected).

**Review Note:** [UTS35] allows for arbitrary ignored `white-space` (including line terminators) after the opening curly bracket and before the closing curly bracket, but not within the character name itself (only U+0020 SPACE is allowed within the name). Spaces other than U+0020 within a `\N` escape are not supported by any implementation (ICU4C, ICU4J, JSPs, nor invariants; the ICU4X experimental implementation does not support `\N` at all).

**Review Note:** Neither the Unicodetools implementation nor the ICU implementation consider name aliases.

**Review Note:** `\N` escapes do not allow for the use of named sequences. Should they be allowed?

### 2.4 Bracketed Elements and Strings

The syntactic categories `bracketed-element` and `string-literal` are defined by the following regular grammar, where a `bracketed-literal-element` is any Unicode scalar value except `\` and `}`.

```
bracketed-element ::= { string-element }
string-literal ::=
  {}
  | { string-elements }
string-element ::=
  bracketed-literal-element | escaped-element | named-element | \p | \P | \N
string-elements ::=
  string-element string-element
  | string-elements string-element
```

### 2.4.1 Semantics

A [bracketed-literal-element](#) represents a single code point: itself. A [string-element](#) represents the code point represented by its constituent [bracketed-literal-element](#), [escaped-element](#), or [named-element](#).

A [bracketed-element](#) represents the code point represented by its constituent [string-element](#). A [string-literal](#) represents the sequence of the code points represented by each of its constituent [string-elements](#).

**Review Note:** The ICU4C and ICU4J implementations ignore [white-space](#) in a [bracketed-element](#) or [string-literal](#). The Properties and Algorithms Group and several ICU-TC participants found this to be confusing; it is therefore proposed that string literals be made space-sensitive.

**Review Note:** ICU4C and ICU4J allow [\p](#), [\P](#) and [\N](#) inside a [string-literal](#) or [bracketed-element](#), as if they were [escaped-elements](#), and do not recognize [named-element](#). We propose making the handling of escapes consistent.

**Note:** {} represents the empty string. A [string-literal](#) represents either the empty string or a string consisting of two or more code points.

**Note:** The optional white space has no effect on the semantics of a [string-literal](#) or [bracketed-element](#).

### 2.5 Property Queries

A [property-query](#) is defined by the following regular grammar.

```

property-query ::=
    perl-start query-expression perl-end
  | posix-start query-expression posix-end
perl-start ::= \p{ | \P{
perl-end ::= }
posix-start ::= [: | [ : ^
posix-end ::= :]
query-expression ::=
    unary-query-expression
  | binary-query-expression
unary-query-expression ::= optional-version-qualifier ucd-identifier
binary-query-expression ::= optional-version-qualifier ucd-identifier query-operator property-predicate
optional-version-qualifier ::=
    ""
  | version-qualifier
version-qualifier ::=
    U version-number :
  | U version-suffix :
  | U-1:
version-number ::=
    digits optional-suffix
  | digits . digits optional-suffix
  | digits . digits . digits optional-suffix
optional-suffix ::=
    ""
  | version-suffix
version-suffix ::= α | β | dev
digits ::= digit | digits digit
digit ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
query-operator ::= = | ≠
property-predicate ::=
    property-value
  | regular-expression-match
  | property-comparison
property-value ::= initial-property-value-element | property-value property-value-element
initial-property-value-element ::=
    initial-literal-value-element
  | escaped-element
  | named-element
property-value-element ::=
    literal-value-element
  | escaped-element
  | named-element
literal-value-element ::= initial-literal-value-element | /
where initial-literal-value-element is any Unicode scalar value other than \, :, {, }, =, ≠, or @.
property-comparison ::= @ unary-query-expression @
regular-expression-match ::= / regular-expression /
regular-expression ::=
    ""
  | regular-expression regular-expression-character
regular-expression-character ::= regex-unescape | \ any
where regex-unescape is any Unicode scalar value other than / and \ and any is any Unicode scalar value.

```

**Review Note:** The operator `≠` is not supported by ICU4C and ICU4J, but is specified in [UTS35], and is supported in the JSPs as well as the ICU4X experimental implementation. Experience has shown that the `\P` syntax can lead to confusion, so `\p` with `≠` may be preferable. The double negation resulting from `\P` with `≠` or `[:^` with `≠` should be avoided, and implementations should probably reject it.

**Review Note:** `property-comparison` and `regular-expression-match` are supported only in the JSPs and invariants.

**Review Note:** No implementation supports escapes in property values. This is not a major problem for the ICUs, as they do not support string- or code point-valued properties either, except for `Name`; but it is a problem in the tools. Since the lack of string- or code point-valued properties seems to be serendipitous, rather than fundamental to the scope of general-purpose internationalization libraries, we propose adding support for escapes generally (so they are in yellow, not in gray).

**Review Note:** UTS35 allows for unescaped `:` in Perl-style queries, and for unescaped `}` in POSIX-style queries. However, non-enumerated properties are not supported in any `UnicodeSet` implementation other than those of the Unicode tools (JSPs and invariants), so this poses no real compatibility constraints. Since we are using `:` as a delimiter, it makes sense to require that it be escaped.

### 2.5.1 Negations

A `property-query` is *exteriorly negated* if it starts with the `posix-start` `[:^` or the `perl-start` `\P{`. It is *interiorly negated* if its `query-expression` is a `binary-query-expression` whose `query-operator` is `≠`.

**Examples:** The constructs `\P{Cn}`, `[:^Cn:]`, `\P{General_Category=Cn}`, and `[:^General_Category=Cn:]`, and `[:^General_Category≠Cn:]` are exteriorly negated. The constructs `\p{General_Category≠Cn}`, and `[:General_Category≠Cn:]`, and `[:^General_Category≠Cn:]` are interiorly negated.

For a `property-query`, the *corresponding non-negated property-query* is defined by changing any `perl-start` to `\p{`, any `posix-start` to `[:`, and any `query-operator` to `=`.

#### Examples:

property-query	Corresponding non-negated property-query
<code>\P{Cn}</code>	<code>\p{Cn}</code>
<code>\p{General_Category≠Cn}</code>	<code>\p{General_Category=Cn}</code>
<code>\P{General_Category=Cn}</code>	<code>\p{General_Category=Cn}</code>
<code>\p{General_Category=Cn}</code>	<code>\p{General_Category=Cn}</code>
<code>[:^General_Category≠Cn:]</code>	<code>[:General_Category=Cn:]</code>

A `property-query` is *simply negated* if it is either exteriorly negated or interiorly negated, but not both. A simply negated `property-query` represents the code point complement of the set represented by the corresponding non-negated `property-query`.

**Examples:** `\P{Cn}` and `\p{General_Category≠Cn}` are simply negated. They represent the code point complement of `\p{General_Category=Cn}`.

A `property-query` is *doubly negated* if it is both exteriorly negated and interiorly negated. A doubly negated `property-query` represents the same set as the corresponding non-negated `property-query`.

**Note:** While they are well-defined, the use of doubly negated property queries is discouraged. Examples of doubly-negated property-queries: `\P{Decomposition_Type≠compat}` (equal to `\p{Decomposition_Type=compat}`), `[:^Noncharacter_Code_Point≠No:]` (equal to `[:Noncharacter_Code_Point=No:]`).

**Note:** There is no semantic difference between POSIX-style and Perl-style property queries, that is, for any `property-query` `x`, `[:x:]` is equivalent to `\p{x}`, and `[:^x:]` is equivalent to `\P{x}`.

A `property-query` which is neither simply negated nor doubly negated is *non-negated*.

**Note:** For any `property-query`, the corresponding non-negated `property-query` is non-negated.

### 2.5.2 Unary Queries

A non-negated `property-query` whose `query-expression` is a `unary-query-expression` represents a set of code points as follows.

1. If the `ucd-identifier` matches an alias for a binary property under rule UAX44-LM3, the `property-query` represents the set of code points for which the given property is True.
2. If the `ucd-identifier` matches an alias for a Script property value under rule UAX44-LM3, the `property-query` represents the set of code points whose Script property value has that alias.
3. If the `ucd-identifier` matches an alias for a `General_Category` property value under rule UAX44-LM3, then:
  1. if the `ucd-identifier` matches an alias for a grouping of `General_Category` values, the `property-query` represents the set of code points whose `General_Category` property value is in that grouping;
  2. otherwise, the `property-query` represents the set of code points whose `General_Category` property value has the alias matching the `ucd-identifier`.
4. Otherwise, the `UnicodeSet` expression is ill-formed.

**Note:** The invariants of the Unicode character database ensure that only one of these alternatives holds. For example, no Script property value alias matches an alias for a binary property.

No such guarantee is made if unary queries are extended to other properties:

- Properties of other types can match Script or General\_Category aliases; for instance, ISO\_Comment has the alias isc, which matches the alias C for the General\_Category grouping Other.
- Value aliases for properties other than Script and General\_Category can match property aliases for binary properties; for instance, White\_Space is both a Bidi\_Class value and a binary property.
- If  $P$  and  $Q$  are properties and the pair  $\{P, Q\}$  is not  $\{\text{Script}, \text{General\_Category}\}$ , a value alias for  $P$  may match a value alias for  $Q$ . For instance, with  $P=\text{Line\_Break}$  and  $Q=\text{Grapheme\_Cluster\_Break}$ , both properties have a value alias ZWJ. With  $P=\text{Script}$  and  $Q=\text{Block}$ , both properties have a value alias Greek.

**Review Note:** The UnicodeSet implementation of the invariant tests do not implement implicit Script nor implicit General\_Category.

If the [version-qualifier](#) with a [version-number](#) is present, the above set is defined based on the property assignments in the version of the Unicode Character Database given by the [version-number](#). A [version-suffix](#) may be used to refer to unpublished versions of the Unicode Character database.

**Note:** No products or implementations should be released based on the beta, alpha, or earlier draft UCD data files. The use of a version suffix in UnicodeSet expressions should be restricted to documents and tools involved in the development of the Unicode Standard.

**Review Note:** Only the Unicode tools (JSPs and invariants) support [version-qualifiers](#). This is not expected to change: general-purpose internationalization libraries have no reason to ship the entire history of the UCD.

In the absence of a version qualifier, the version of the UCD used depends on context. The [version-qualifier](#) `u-1:` is used to refer to the version of the UCD preceding the one referenced by an absence of version qualifier.

**Review Note:** The [version-qualifier](#) `u-1:` is only supported in the invariant tests, not in the JSPs.

#### Examples:

By default, within the text of the Unicode Standard, a UnicodeSet expression refers to the property assignments in that version of the standard.

In the sentences “the set `\p{Pattern_Syntax}` is immutable” and “the set `\p{XID_Continue}` can only grow over successive versions of the Unicode Standard”, the expression refers to all versions of the UCD.

The encoding stability policy, applicable to Unicode 2.0+, states that

Once a character is encoded, it will not be moved or removed.

This policy implies that `\p{GC=unassigned} ⊆ \p{u-1:GC=unassigned}`, where the implicit version is any version after 2.0.

### 2.5.3 Binary Queries

A non-negated [property-query](#) whose [query-expression](#) is a [binary-query-expression](#) represents a set of code points as follows.

The [ucd-identifier](#) preceding the [query-operator](#) shall match an alias for a property under rule UAX44-LM3. That property is the *queried property*. If the [binary-query-expression](#) starts with a [version-qualifier](#), it defines the *queried version*.

**Note:** The invariants of the Unicode character database ensure that a string matches an alias for at most one property.

If the [property-predicate](#) is a [property-value](#), the *queried value* is defined as the sequence of code points represented by each [initial-property-value-element](#) or [property-value-element](#), where an [initial-literal-value-element](#) or a [literal-value-element](#) represents itself, and an [escaped-element](#) and a [named-element](#) represent a code point as described by their respective semantics.

A [property-value](#) shall consist solely of [literal-value-elements](#) unless the queried property is a string-valued or miscellaneous property.

**Review Note:** The preceding paragraph removes an unnecessary burden on implementers that do not support string properties (they do not need to support `\p{gc=\N{LATIN CAPITAL LETTER L}\N{LATIN SMALL LETTER L}}`), and it establishes some semblance of typing (even though we do not formally have types in this specification).

If the queried version is defined, the property assignments of the queried property used in the definition of the set are those from that version of the Unicode Character Database.

#### 2.5.3.1 Age Queries

If the queried property is the Age property, the [property-predicate](#) shall be a [property-value](#), and the queried value shall match a value alias for the Age property under UAX44-LM3. The [property-query](#) then represents the set of code points whose Age value is less than or equal to the matching Age value.

**Example:** The set `\p{Age=6.0}` contains all characters that were assigned in Unicode Version 6.0, as well as noncharacter code points, surrogate code points, and private use area code points. It is equal to the set `[ \P{U6:Cn} \p{U6:Noncharacter_Code_Point} ]`. The expressions `\p{Age=@U6:Age@}` and `\p{Age=/1/}` are ill-formed.

**Note:** The special handling of the Age property addresses the common use case of matching characters present in some version of Unicode (thus with an age older than or equal to that version of Unicode). This special handling is largely redundant with the more regular [version-qualifier](#) mechanism; specifically for an alias  $x$  of the Age property which satisfies the [version-number](#) grammar, The sets `\p{Ux:gc#Unassigned}` and `[ \p{Age=x} - \p{Noncharacter_Code_Point} ]` are equal. However, the support of [version-qualifier](#) is not recommended for general-purpose APIs, see [Section 5, Use in APIs](#).

**Review Note:** The age property behaves unusually in `UnicodeSet`, in a way that cannot be unified with the other properties. Contrast the `Name` property, which we can make regular by treating formal aliases as value aliases. We therefore do not specify property comparisons nor regular expression matching on the `Age` property.

### 2.5.3.2 Property Comparisons

If the `property-predicate` is a `property-comparison`, the constituent `ucd-identifier` of the `property-comparison` shall either match an alias for a property under rule UAX44-LM3, or it shall match the string `none` or the string `code point` under rule UAX44-LM3. In the first case, that property is the *comparison property*. In the second case, there is no comparison property. If the constituent `unary-query-expression` of the `property-comparison` starts with a `version-qualifier`, it defines the *comparison version*.

**Example:** In both `\p{scf=@1c@}` and `\p{U15.1:scf=@U15.1:1c@}`, the queried property is `Simple_Case_Folding` and the comparison property is `Lowercase_Mapping`. In `\p{U15.0:Line_Break#@U15.1:Line_Break@}`, the queried version is 15.0, and the comparison version is 15.1. In `\p{kIRG_GSource=@none@}` and `\p{case folding=@code point@}`, there is no comparison property. The expressions `\p{kIRG_GSource=@U16:none@}` and `\p{case folding=@U16:code point@}` are ill-formed.

If there is no comparison property, the constituent `unary-query-expression` of the `property-comparison` shall not start with a `version-qualifier`.

If the comparison version is defined, the property assignments used of the comparison property used in the definition of the set are those from that version of the Unicode Character Database. For both properties, if the version is absent, it depends on context. If both version qualifiers are absent, the same context-dependent version is used.

**Example:** The statement “the set `\p{scf=@1c@}` shrank between Unicode 15.0 and Unicode 15.1” is a statement about the sets `\p{U15.1:scf=@U15.1:1c@}` and `\p{U15.0:scf=@U15.0:1c@}`

If there is a comparison property, its type shall be compatible with that of the queried property, that is, one of the following shall hold:

1. Both are binary properties.
2. Both are (possibly multivalued) string-valued properties.
3. Both are (possibly multivalued) numeric properties.
4. Both are (possibly multivalued) enumerated or catalog properties with the same underlying enumeration.
5. They are the same property.

The `query-expression` then represents the set of code points that have the same value for the queried property and comparison property. For unordered multivalued properties, the sets of values are compared. For ordered multivalued properties, the sequences of values are compared.

**Examples:** The expression `\p{Decomposition_Mapping=@Ideographic@}` is ill-formed, as the string-valued `Decomposition_Mapping` property and the binary `Ideographic` property have incompatible types. The following are well-formed expressions from each of the three categories above:

1. The set `\p{Uppercase#@Changes_When_Lowercased@}` is the set of characters whose `Uppercase` value differs from their `Changes_When_Lowercased` value. It is equal to `[[\p{Uppercase}\p{Changes_When_Lowercased}]-[\p{Uppercase}&\p{Changes_When_Lowercased}]]`, that is, the set of characters that are either `Uppercase` or `Changes_When_Lowercased`, but not both.
2. The set `\p{scf#@cf@}` is the set of characters whose `Simple_Case_Folding` differs from their (full) `Case_Folding`.
3. The set `\p{Numeric_Value=@kPrimaryNumeric@}` is the set of characters that either have a single `kPrimaryNumeric` value, or have neither `kPrimaryNumeric` nor `Numeric_Value` (both are NaN).
4. The set `\p{U15.0:Line_Break#@U15.1:Line_Break@}` is the set of code points whose `Line_Break` assignment changed between Unicode Version 15.0 and Unicode Version 15.1.

The set `\p{U16.0:kPrimaryNumeric#@U17.0:kPrimaryNumeric@}` contains U+5146, as the values are ordered and the order changed in Unicode Version 17.0. The set `\p{Script_Extensions=@Script@}` is the set of characters whose `Script_Extensions` value is a single value equal to their `Script` value. These are the characters not listed in `ScriptExtensions.txt`, to which the line `@missing: 0000..10FFFF; <script>` applies.

**Review Note:** We allow only sensible `property-comparisons`. The `UnicodeTools` allow `\p{Decomposition_Mapping=@Ideographic@}`, which is equal to `[No]` (via the value `No`), and we don't want to specify this sort of silliness.

### 2.5.3.3 Identity and Null Queries

If the `property-predicate` is a `property-comparison` and there is no comparison property:

1. If the `ucd-identifier` matches `code point`, the property shall be a string-valued property. The `query-expression` represents the set of code points that are mapped to themselves by the queried property.
2. If the `ucd-identifier` matches `none`, the property shall be a string-valued property or a miscellaneous property. The `query-expression` represents the set of code points for which no value is defined for the queried property.

**Examples:** The set `\p{scf=@code point@}` is equal to the set of code points which map to themselves under simple case folding. The set `[:*kIRG_GSource=@none@:]` is the set of CJK ideographs that have a “G” source mapping. The sets `\p{Bidi_Paired_Bracket=@none@}` and `\p{Bidi_Paired_Bracket_Type=None}` are equal.

**Review Note:** The only known implementation to support identity and null queries is the one used by the invariant tests. UTS #18 suggests `@identity@` instead of `@code point@` and does not have `@none@`. The use of `@code point@` and `@none@` is consistent with the use of `<code point>` and `<none>` in UCD `@missing` lines in a shared namespace with property names, with `<script>`.

### 2.5.3.4 Valid Values and Resolved Sets

A string `s` is a *valid value* for a property `p` if one of the following holds:

1.  $p$  is the Name property and  $s$  matches a value of the Name property or a value of the Name\_Alias property under matching rule UAX44-LM2.
2.  $p$  is the Name\_Alias property and  $s$  matches one the values of the Name\_Alias property under matching rule UAX44-LM2.
3.  $p$  is a property for which property value aliases are defined, and  $s$  matches a value alias under matching rule UAX44-LM3.
4.  $p$  is some other string-valued or miscellaneous property.
5.  $p$  is a numeric property, and:
  1.  $s$  matches the string NaN under matching rule UAX44-LM3,
  2.  $s$  matches the regular expression  $[+-]?[0-9]+(/[0-9]*[1-9][0-9]*)?$ , or
  3.  $s$  matches the regular expression  $[+-]?[0-9]+\.[0-9]+$ .

The *resolved set* of  $p$  for  $s$  is then respectively:

1. The set whose sole element is the character whose name or name alias matches  $s$ .
2. The set whose sole element is the character whose name alias matches  $s$ .
3. If  $p$  is the General\_Category property and  $s$  is an alias for a grouping of General\_Category values, the set of characters whose General\_Category is one of the values in that grouping. Otherwise, the set of characters for which one of the values of  $p$  has an alias matching  $s$ .
4. The set of characters for which the value of  $p$  is the string  $s$  itself.
5. The set of characters for which the value  $x$  of  $p$  is such that, respectively:
  1.  $x$  is NaN,
  2.  $x$  is the rational number expressed by  $s$ ,
  3. the [IEEE754] binary64 floating-point number nearest to  $x$  is equal to the binary64 closest to the decimal number  $s$

**Note:** This implements matching rule UAX44-LM1.

#### 2.5.3.5 Property Value Queries

If the [property-predicate](#) is a [property-value](#), the queried value shall be a valid value for the queried property.

The [query-expression](#) represents the resolved set of the queried property for the [property-predicate](#).

**Examples:** The set  $\{p\{\text{Uppercase}=\text{True}\}$  is equal to the set  $\{p\{\text{Uppercase}\}$ . The set  $\{p\{\text{Uppercase}=\text{NO}\}$  is equal to the set  $\{p\{\text{Uppercase}\}$ . The set  $\{p\{\text{Script\_Extensions}=\text{Latin}\}$  is the set of characters that have Latin as one of their Script\_Extensions values. The sets  $\{p\{\text{nv}=\text{2}/\text{12}\}$  and  $\{p\{\text{Numeric\_Value}=\text{1}/\text{6}\}$  are equal. For all formal name aliases  $x$ ,  $\{p\{\text{Name\_Alias}=\text{x}\}$  and  $\{p\{\text{Name}=\text{x}\}$  are equal.

#### 2.5.3.6 Regular Expression Queries

If the [property-predicate](#) is a [regular-expression-match](#), the queried property shall not be a numeric property. The text of the [regular-expression](#) is interpreted as a regular expression. Where ambiguous, the specific regular expression syntax and options used should be described.

**Review Note:** Defining regular expression matching on numeric values would require us to define a finite set of preferred string representations of the numeric values, filling the same role as the exact spellings of name aliases. This would be a nontrivial exercise, and likely a pointless one, as matching numbers with regular expressions is inconvenient.

If the queried property is the Name property, the [query-expression](#) represents the set of code points whose character name matches the regular expression, or that have a formal name alias matching the regular expression. Otherwise the [query-expression](#) represents the set of code points for which one of the aliases of one of the values of the queried property matches the regular expression.

**Examples:** The set  $\{p\{\text{Name}=\text{CAPITAL LETTER}\}$  is the set of all characters whose name contains “CAPITAL LETTER”. The set  $\{p\{\text{Block}=\text{/Cyrillic}\}$  is the set of all code points in a block whose name starts with “Cyrillic”. The set  $\{p\{\text{scx}=\text{/Gondi}\}$  contains all code points that have either Gunjala\_Gondi or Masaram\_Gondi among their Script\_Extensions values. The set  $\{p\{\text{gc}=\text{/P}\}$  contains punctuation characters (whose short aliases match), as well as private use characters and U+2029 PARAGRAPH SEPARATOR (whose long aliases match).

**Note:** Neither loose matching rule LM2 nor LM3 is applied in regular expression queries. The set  $\{p\{\text{Name}=\text{NO BREAK SPACE}\}$  is empty, whereas the set  $\{p\{\text{Name}=\text{/NO-BREAK SPACE}\}$  contains NO-BREAK SPACE, NARROW NO-BREAK SPACE, and ZERO WIDTH NO-BREAK SPACE. The set  $\{p\{\text{Script}=\text{/Gondi}\}$  is empty, whereas the set  $\{p\{\text{Script}=\text{/Gondi}\}$  contains Gunjala Gondi and Masaram Gondi characters. General\_Category groupings are not taken into account in regular expression queries: the set  $\{p\{\text{gc}=\text{/Cased\_Letter}\}$  is empty. If  $x$  is the exact spelling of a value alias for property  $p$ , or if  $P$  is Name and  $x$  is either the exact spelling of a name or a name alias, the sets  $\{p\{p=\text{x}\}$  and  $\{p\{p=\text{/^x}\}$  are equal.

**Review Note:** Neither the JSPs nor the invariant tests take Name\_Alias into account for regular expression queries on the Name property. We want to take Name\_Alias into account for value queries for compatibility with ICU (which follows the recommendations in UTS18), see the review note above. We also want to be consistent between regular expression queries and value queries (specifically, we want the property stated at the end of the note above). We therefore need to consider name aliases as aliases of the Name property here too.

### 3 Set Operations

UnicodeSet expressions are defined by the syntactic category [UnicodeSet](#) in the following context-free space-insensitive grammar, whose terminals are the lexical elements defined in Section 2, Lexical Elements.

```
UnicodeSet ::=
  [ Union ]
  | Complement
  | property-query
```

```

named-element
Complement ::= [ ^ Union ]
Union ::=
    Terms
  | UnescapedHyphenMinus Terms
  | Terms UnescapedHyphenMinus
  | UnescapedHyphenMinus Terms UnescapedHyphenMinus
UnescapedHyphenMinus ::= -
Terms ::=
    ""
  | Terms Term
Term ::=
    Elements
  | Restriction
Restriction ::=
    UnicodeSet
  | Intersection
  | Difference
Intersection ::= Restriction & UnicodeSet
Difference ::= Restriction - UnicodeSet
Elements ::= Element | Range
Range ::= RangeElement - RangeElement
RangeElement ::=
    literal-element
  | escaped-element
  | named-element
  | bracketed-element
Element ::= RangeElement | string-literal | bracketed-element

```

**Note:** The above grammar is LR(2) rather than LR(1). After `[a`, if the next lexical element is the [set-operator](#) `-`, there is an ambiguity between a [Range](#) and an [Element](#) followed by an [UnescapedHyphenMinus](#) (a shift-reduce conflict). This ambiguity is resolved by looking ahead one more lexical element: the `-` is an [UnescapedHyphenMinus](#) only if it is followed by the [set-operator](#) `]`. The grammar can be rewritten to be LR(1), see [Knuth1965]. However, such a transformation obscures the definition of the syntax, as it requires introducing syntactic categories for constructs such as `a-` that could either be the beginning of a range or an element followed by an unescaped hyphen, and those such as `[a-z]-` that could turn out to be either the beginning of a difference or a restriction followed by an unescaped hyphen.

The grammar can also be straightforwardly rewritten to be LL(2), so that it lends itself to top-down predictive parsing. [Restriction](#) must then be analysed with right rather than left recursion, as [UnicodeSet RightHandSides](#), where [RightHandSides](#) ::= "" | & [UnicodeSet RightHandSides](#) | - [UnicodeSet RightHandSides](#). The tree resulting from this right-recursive grammar is not an expression tree, as set difference is not an associative operation, and the operators `-` and `&` are left-associative in UnicodeSet syntax: a construct whose syntactic category is [RightHandSides](#) does not represent a set. Instead a top-down UnicodeSet parser must shrink the set corresponding to the [Restriction](#) as it encounters additional operators `&` and `-`. Left factoring of `[ ^ Union ]` and `[ Union ]` can be used to parse those constructs with only one lexical element of lookahead, but as in the LR case, it is most practical to handle [UnescapedHyphenMinus](#) by looking ahead two lexical elements.

Review Note: ICU puts [named-element](#) as an alternative in [UnicodeSet](#) rather than [Element](#), making `\N{SPACE}` equivalent to `[x{20}]` rather than `\x{20}`; see ICU-22851.

This is misleading, as the expression `\N{LATIN SMALL LETTER A}-\N{LATIN SMALL LETTER Z}` is then valid, but is the singleton `[a]` rather than the set of 26 letters `[a-z]`. This has led to bugs in practice.

The proposal to move it to [Element](#) fixes that.

This means that expressions of the form `UnicodeSet - named-element`, e.g., `[p{Changes_When_Casefolded}-\N{COMBINING GREEK YPOGEGRAMMENI}]`, which would make sense and work in earlier versions of ICU, become invalid. Likewise, an unbracketed `\N{SPACE}` is currently a valid and unproblematic UnicodeSet, and would become invalid. In both cases, brackets need to be added to restore the old semantics, thus `[p{Changes_When_Casefolded}-\N{COMBINING GREEK YPOGEGRAMMENI}]` and `[\N{SPACE}]` respectively.

The ICU-TC approved this backward-incompatible change to its implementation. An earlier draft of this grammar included affordances for backward compatibility, allowing a [named-element](#) to stand as a set on the right-hand-side of a [Restriction](#) or as an entire UnicodeSet expression. The ICU-TC considers that the backward compatibility was outweighed by the added complexity in the grammar and by the discrepancy in behaviour between a C++ `\N` escape and a `\\N` (representing a [named-element](#)) in a string literal containing a UnicodeSet expression.

Review Note: ICU4J allows string ranges such as `{aa}-{zz}` (all 2-letter lowercase ASCII strings). ICU4C disallows string ranges, but also disallows [bracketed-element](#) in ranges, thus disallowing `{a}-{z}`. UTS35 used to allow string ranges, but they were retracted, leaving only the single-character `{a}-{z}`. ICU4X follows UTS35 and allows for ranges of [bracketed-element](#), but not string ranges.

Experience in CLDR has shown that the systematic usage of brackets is useful in avoiding surprises with combining marks:

```

[\p{Latn} - \p{Changes_When_NFKC_Casefolded} & [a-ä]] is a set of 31 Latin letters equal to [a-z áâãäå], whereas
[\p{Latn} - \p{Changes_When_NFKC_Casefolded} & [a-ä]] is equal to [a-q], because [a-q] is [a-q \N{COMBINING DIAERESIS}]. If brackets are
used, [\p{Latn} - \p{Changes_When_NFKC_Casefolded} & {a-ä}] remains valid, but
[\p{Latn} - \p{Changes_When_NFKC_Casefolded} & {a}-{ä}] is a syntax error, exposing the issue.

```

As a result, we are proposing to allow bracketed-clement as a RangeElement, while disallowing string ranges.

### 3.1 Semantics

A **RangeElement** represents the single code point represented by its constituent lexical element.

A **Range** represents the set of code points that are both greater than or equal to the code point represented by the initial **RangeElement** and less than or equal to the final **RangeElement**. If the code point represented by the initial **RangeElement** is greater than the code point represented by the final **RangeElement**, the **UnicodeSet** expression is ill-formed.

**Examples:** The **Range** a-z represents a set of 26 elements. The **Range** z-a is not the empty set; it is ill-formed.

An **UnescapedHyphenMinus** represents the set whose sole element is U+002D - HYPHEN-MINUS.

**Review Note:** ICU4C and ICU4J also support a final \$ in a **Union**, which represents U+FFFF. However, this is better understood as a conformant extension designed for an environment where U+FFFF signals string boundaries, in particular for use in higher-level syntaxes such as transliterator rules. This is therefore discussed in the sections on conformance and higher-level syntaxes. [TODO: Which I have not yet written.]

A **Complement** represents the code point complement of the set represented by its constituent **Union**, that is, the set of code points not in the set represented by the **Union**.

An **Intersection** represents the intersection of the sets represented by the **Restriction** and **UnicodeSet** either side of the &.

A **Difference** represents the set of elements of set represented by the **Restriction** that are not elements of the set represented by the **UnicodeSet**.

For all other syntactic categories defined in the **UnicodeSet** grammar, the construct represent the union of the sets represented by their immediate constituent constructs.

**Examples:** The **UnicodeSet** [ac-z] contains twenty-five elements; it is the union of the sets represented by the **Element** a and the **Range** c-z.

**Note:** The empty **Terms** represents the empty set, and the **UnicodeSet** [] is therefore the empty set.

**Note:** The operators & (intersection) and - (set difference) have equal precedence and are left-associative: [ [a-z] - [c] & [d] ] is equal to [d], whereas [ [a-z] - [[c] & [d]] ] is the empty set. Set union, denoted by juxtaposition, has a lower precedence: [ [a-z] - [c] [d] ] is equal to [a-b d-z], whereas [ [a-z] - [[c] [d]] ] is equal to [a-b e-z].

## 4 Conformance

An implementation of **UnicodeSet** syntax is *consistent* if, for every valid **UnicodeSet** expression defined by this specification, the implementation either rejects the expression or evaluates it according to this specification.

### Examples:

1. An implementation that rejects any input string is consistent.
2. An implementation is consistent if it rejects any **UnicodeSet** expression that makes use of the syntactic categories whose definition has a gray background in the grammar, but accepts and correctly interprets all other **UnicodeSet** expressions.
3. An implementation which interprets [a] and [b] as the same set is not consistent.
4. An implementation which interprets [ \d ] as \p{Nd} is not consistent.

**Note:** Consistency is not required of conformant implementation, as it prevents the use of notations that are common in regular expressions, such as \d for digits, or the use of identifiers without sigils, as in [UAX14]. However, since they lead to interoperability issues when reusing an expression in another implementation, the inconsistencies must be declared.

An implementation that interprets expressions that are not valid **UnicodeSet** expressions according to this specification implements a *pure extension*.

**Note:** **UnicodeSet** syntax does not have many reserved characters: most characters are valid **literal-elements**. In particular, **Pattern\_Syntax** characters other than \$ are not reserved, and cannot be given a syntactic meaning as a pure extension. However, some character sequences cannot occur in well-formed **UnicodeSet** expressions, and could thus be used to define pure extensions:

1. The sequence of lexical elements -x-, where x is a **literal-element**, can only occur in a well-formed **UnicodeSet** expression if it is at the beginning or the end of a **Union**; the sequence -- can only occur if it is the entirety of a **Union**. These sequences can therefore be used as infix operators as a pure extension.
2. The sequences of lexical elements && and &x, where x is a **literal-element**, are always ill-formed, and can therefore be used in pure extensions.
3. A lexical element cannot start with \x followed by a character other than a hexadecimal digit or {. \x can therefore be used as part of additional lexical elements in pure extensions.
4. The character \$ is reserved, and can be used to define pure extensions.

**Note:** Any pure extension may be assigned a meaning in a future version of this specification; while using pure extensions to implement new features avoids changing the interpretation of currently standardized **UnicodeSet** expressions, it does not guarantee that expressions using the extensions are forward compatible.

**Examples:** The following are pure extensions:

- Accepting a final \$ in a [Union](#) and interpreting it as representing the character U+FFFF.
- Interpreting a non-negated [property-query](#) whose [ucd-identifier](#) is `exemplar` as the set of all characters that are CLDR exemplars for the language whose language code is given by the [property-predicate](#).
- Accepting the operators `--` as set difference and `&&` as set intersection, in addition to `-` and `&`.
- Adding `UnicodeSet -u UnicodeSet` as an alternative in [Union](#) with the semantic of a disjoint union (the union of both constituent [UnicodeSets](#), ill-formed if they intersect).
- Defining `Transform ::= &transform ( identifier , UnicodeSet )` and adding it as an alternative in [UnicodeSet](#).
- Defining a lexical element variable  `::= $ identifier`, with `identifier ::= XID_Start | identifier XID_Continue`, and adding it as an alternative in [UnicodeSet](#); see [Section 6, Higher-Level Syntaxes](#).
- Adding `UnicodeSet \xor UnicodeSet` as an alternative in [Union](#) with the semantic of a symmetric difference.

**Note:** The International Components for Unicode interpret a final \$ in a [Union](#) as U+FFFF. This is related to the behavior of out-of-range indexing in ICU, which returns U+FFFF as a sentinel value. A character class containing U+FFFF can therefore be used to match the end of a string.

An implementation of `UnicodeSet` syntax is *syntactically complete* if, for some subset of lexical elements which contains at least all [set-operators](#), it supports all productions of the `UnicodeSet` grammar and interprets them according to this document.

**Examples:** As the syntactic categories whose definitions have a gray background in the grammar are part of the grammar of lexical elements, an implementation is syntactically complete if does not support these, but accepts and correctly interprets all other `UnicodeSet` expressions.

An implementation is not syntactically complete if it supports the entirety of the [property-query](#) grammar, but does not support the [Complement](#) syntax.

A syntactically complete implementation interprets `[]` as the empty set and `[^]` as the set of all code points.

**Note:** A syntactically complete implementation need not be consistent. For instance, such an implementation can remove `\d` from the set of [escaped-elements](#), give it the meaning of `\p{Nd}`, and add it as an alternative in `UnicodeSet`. It would therefore give `[\d]` a different meaning than that given by this specification.

A syntactically complete implementation is *minimally consistent* if, for any lexical element in the following list, the implementation either rejects the lexical element, or interprets it according to this specification:

- Any [escaped-element](#) with constituent [hexadecimal-digits](#).
- Any [named-element](#).
- Any [property-query](#).

**Note:** The definition of syntactic completeness requires that a minimally consistent implementation interpret all [set-operators](#) according to this specification.

**Example:** An implementation can be minimally consistent even if it interprets `\d` as the set `\p{Nd}` rather than as an [escaped-element](#). An implementation that interprets `\p{IsGreek}` as the set of code points in the Greek and Coptic block, instead of the set of characters with `Script=Greek`, is not minimally consistent.

**UTS61-C1** A conformant implementation of `UnicodeSet` syntax shall be syntactically complete and minimally consistent.

**Example:** An implementation that interprets `\p{IsGreek}` as the set of code points in the Greek and Coptic block is not a conformant `UnicodeSet` implementation.

**UTS61-C2** A conformant implementation of `UnicodeSet` syntax shall declare any restrictions to the set of lexical elements defined by this syntax.

**Note:** A lack of support for the syntactic categories defined with a gray background can be described as “supporting only property queries that are recommended for general-purpose APIs”. Support for a subset of UCD properties in property queries is easiest to describe by enumerating the supported properties.

**UTS61-C3** A conformant implementation of `UnicodeSet` syntax that is not consistent shall declare itself as a tailoring of `UnicodeSet` syntax. It shall declare the expressions that are interpreted differently from this specification.

**Example:** A syntactically complete and minimally consistent implementation that excludes `XID_Continue` characters from [literal-element](#), adds default identifiers to the `UnicodeSet` production, and interprets `x` as `\p{!b=x}` for any default identifier `x`, is not consistent, since it interprets `[qu]` as a different set from `[{Q} {U}]`. It is a conformant tailoring of `UnicodeSet` syntax.

## 5 Use in APIs

The support of [version-qualifier](#) require carrying a long-obsolete versions of the Unicode Character Database; this represents a large amount of data, and a burden on implementers to support variations in format over the years. It is therefore not recommended for general-purpose APIs.

Similarly, the support of [property-comparison](#) and [regular-expression-match](#) in a [property-query](#) requires a significant amount of bespoke logic from implementers, and are primarily useful for exploratory queries on the Unicode Character Database, rather than to define character classes used in practical application. It is not recommended for general-purpose APIs.

General-purpose APIs should not expose the properties that are contributory, obsolete, deprecated, or otherwise not recommended for support in public property APIs. See [Section 5.1, Property Index](#), in [UAX44].

**Note:** `UnicodeSet` expressions using such properties are well-defined, and it is useful for them to be supported in tools used in the development of the Unicode Standard. For instance, the stability policy statement that decomposition mappings are limited to a single

value or a pair can be checked by verifying that the sets [ `\p{Decomposition_Type=Canonical} & \p{Decomposition_Mapping=}` ] and [ `\p{Decomposition_Type=Canonical} & \p{Decomposition_Mapping=/.../}` ] are empty, even though `Decomposition_Type` is not appropriate for general-purpose APIs.

## 6 Use in Higher-Level Syntaxes

UnicodeSet syntax can be used within higher-level syntaxes. In particular, as it defines a syntax for character classes, it can be used for the character classes in a regular expression syntax.

In many cases, it can be useful to include variables in a higher-level syntax based on UnicodeSet. A syntax allowing variables in UnicodeSet syntax should incorporate the identifiers into the grammar. Textual replacement prior to parsing the UnicodeSet syntax is not advisable, as it results in misleading behaviour: [ `$x $y $z` ] would be the range [ `a-z` ] for `$x=a`, `$y=-`, `$z=z`, but the three-element set [ `az-` ] for `$x=a`, `$y=z`, `$z=-`.

The UnicodeSet syntax disallows an unescaped U+0024 \$ DOLLAR SIGN, so identifiers starting with \$ can be made a lexical element as a pure extension of the syntax. Alternatively, default identifiers as defined in [UAX31] may be used. If default identifiers are used, characters with the `XID_Start` property must be removed from the syntactic category `literal-element`.

**Example:** In [UAX14], short aliases of `Line_Break` property values stand for the set of code points with that property; for instance, `QU` stands for `\p{1b=QU}`. If the algorithm were to special-case the letter Q in one of its regular expressions, it would need to refer to it using an `escaped-element` such as `\x51`, a `named-element` such as `\N{LATIN CAPITAL LETTER Q}`, or a `bracketed-element` such as `{Q}`.

In addition to defining a lexical element identifier, a syntax using UnicodeSet with identifiers must incorporate this lexical element in the `UnicodeSet` grammar. If the variables can only represent sets, identifier can be added as an alternative in the `UnicodeSet` production without further complication: [ `$a-$b` ] is then always a set difference. If the variables are also allowed to represent single code points for use in ranges, the category variable can be added as an alternative in the `RangeElement` production. This makes the grammar ambiguous (that is, it has a reduce-reduce conflict), so that the types of the variables must be known to parse it correctly: [ `$a-$b` ] may be a range, a set difference, or erroneous depending on the types of `$a` and `$b`.

**Review Note:** The Unicode invariant tests, the implementation of segmentation rules in the Unicode tools, and ICU transliterators all support variables in UnicodeSets, all using variables with `$sigils`.

The invariant tests and segmentation rules use textual replacement, but check that the values of the variables are valid UnicodeSet expressions; except for special handling of `\N` with the grammar as amended here, this is equivalent to having identifier as an alternative in `UnicodeSet`.

The ICU4C and ICU4J transliterators use textual replacement, but do not check that the variables are valid UnicodeSet expressions. The variables are used in ranges in practice by some transliterators in CLDR.

The ICU4X implementation of transliterators incorporates variables into its UnicodeSet grammar, using the types to disambiguate, but disallowing a variable from turning into a set operator.

As part of a higher-level syntax that allows comments, it can be useful to allow comments within multiline UnicodeSet expressions. In that case, the comment initiator character must be removed from the `literal-element` category. The character U+0023 # NUMBER SIGN is a common choice, being compatible with the comment syntax of many space-insensitive regular expression syntaxes.

**Review Note:** The Unicode invariant tests allow comments in multiline UnicodeSet expressions.

## 7 Best Practices

### 7.1 Escaping

The use of an `escaped-element` with a constituent `escapable-character` is not recommended when that `escapable-character` is neither a space (U+0020) nor a `Pattern_Syntax` character; such unnecessary escaping is especially ill-advised for letters in the Basic Latin block. Indeed, escape sequences consisting of a Basic Latin letter frequently have a different meaning in higher level syntaxes. This is in particular the case in regular expressions, where, for instance, `\d` typically stands for digits (`\p{Nd}`) or [0-9] depending on the implementation), rather than the letter U+0064 d LATIN SMALL LETTER D.

Conversely, it is recommended to escape the character U+0023 # NUMBER SIGN, as it may be a comment initiator in higher-level syntaxes.

### 7.2 Bidirectional display

**TODO** Describe the atoms for the purpose of <https://www.unicode.org/reports/tr55/#Conversion-To-Plain-Text>.

### 7.3 Style Guide for Unicode Specifications

Many aspects of UnicodeSet syntax exist for compatibility with existing practice in regular expression and other pattern syntaxes. Prominent examples are the profusion of escape syntaxes, including octal, and the dual POSIX-style [ `...:` ] and `\p{...}` options. The specification includes these options to ensure that standard UnicodeSet expressions are interoperable with commonly-used UnicodeSet implementations, and that commonly-used UnicodeSet expressions are well-defined.

However, actually using multiple redundant options is detrimental to the clarity of specifications. As a result, a limited subset of UnicodeSet syntax is used in the text of the Unicode Standard and associated Unicode Technical Reports. The rules in this section define this limited subset.

Besides making a choice between redundant alternatives, the subset of UnicodeSet syntax used in Unicode specifications also excludes some of the advanced features that function as a query language on the UCD. While it is valuable in the preparation of the standard to have a well-defined notation for discussing the relation between properties, or historical values of properties, the actual standard should not rely on these constructs. If a set defined by a relation between properties is useful to an algorithm, it should be turned into a derived binary property, instead of requiring users of the standard to derive it themselves.

UTS61-SG1 Do not use POSIX-style property queries.

UTS61-SG2 Use only the [posix-start](#) `\p`, not `\P`. Use a [binary-query-expression](#) with `=No` or `≠` instead of negating a [unary-query-expression](#) with `\P`.

UTS61-SG3 Prefer changing an intersection to a difference, or vice-versa, to using a negated property query as its right-hand side.

UTS61-SG4 Only use the following [escaped-elements](#):

- `\u` [four-hexadecimal-digits](#)
- `\x{` [hexadecimal-digits](#) `}`

UTS61-SG5 Do not use [regular-expression-match](#), [property-comparison](#), or [version-qualifier](#).

**Table 1. Style Guide Examples**

Rule	Do not use	Use instead
UTS61-SG1	<code>[ :Lowercase_Letter: ]</code>	<code>\p{Lowercase_Letter}</code>
UTS61-SG2	<code>\P{Unassigned}</code>	<code>\p{General_Category≠Unassigned}</code>
	<code>\P{Deprecated}</code>	<code>\p{Deprecated=No}</code>
UTS61-SG3	<code>[ [ \u0000-\uFFFF ] &amp; \p{General_Category≠Unassigned} ]</code>	<code>[ [ \u0000-\uFFFF ] - \p{Unassigned} ]</code>
UTS61-SG4	<code>\0</code>	<code>\u0000</code>
	<code>\U00010FFFF</code>	<code>\x{10FFFF}</code>
UTS61-SG5	<code>\p{Uppercase#@Changes_When_Lowercased@}</code>	<code>[ [ \p{Uppercase}\p{Changes_When_Lowercased} ] - [ \p{Uppercase}&amp;\p{Changes_When_Lc</code>
	<code>\p{Bidi_Paired_Bracket=@none@}</code>	<code>\p{Bidi_Paired_Bracket_Type=None}</code>
	<code>\p{scf#@cf@}</code>	(If this set is useful in an algorithm, a property should be defined for it.)

**Review Note:** Many more rules will be added in subsequent drafts.

## References

**Review Note:** The list of references will be updated in a future draft of this document.

- [IEEE754] *IEEE Standard for Floating-Point Arithmetic*  
IEEE 754-2019:  
<https://standards.ieee.org/ieee/754/6210/>
- [Unicode] *The Unicode Standard*  
Latest version:  
<https://www.unicode.org/versions/latest/>
- [UAX14] *Unicode Standard Annex #14: Unicode Line Breaking Algorithm*  
Latest version:  
<https://www.unicode.org/reports/tr14/>
- [UAX29] *Unicode Standard Annex #29: Unicode Text Segmentation*  
Latest version:  
<https://www.unicode.org/reports/tr29/>
- [UAX31] *Unicode Standard Annex #31: Unicode Identifiers and Syntax*  
Latest version:  
<https://www.unicode.org/reports/tr31/>
- [UTS18] *Unicode Technical Standard #18: Unicode Regular Expressions*  
Latest version:  
<https://www.unicode.org/reports/tr18/>

## Acknowledgements

Robin Leroy authored the bulk of the text, under direction from the Unicode Technical Committee.

Thanks also to the following people for their feedback or contributions to this document: Mark Davis, Asmus Freytag,

## Modifications

The following summarizes modifications from the previous revision of this document.

### Revision 1

- Initial version of the Proposed Draft based on [L2/25-127](#), authorized by decision [183-C26](#).
- Draft 2: Made [string-literal](#) space-sensitive (it is space-insensitive in ICU), removed the optional-white-space production.
- Draft 2: Split `[^` into two lexical elements (`[`, already a [set-operator](#) in draft 1, and `^`). This means spaces are allowed between `[` and `^` in a [Complement](#).
- Draft 2: Corrected the change markers in the [Element](#) production to correctly reflect the ICU4C behaviour prior to the proposed changes: [bracketed-element](#) is an [Element](#) in ICU4C. No change to the grammar resulting from the highlighted changes, [bracketed-element](#) becomes a [RangeElement](#).
- Draft 2: Expanded the note on parsing considerations to consider top-down parsing.
- Draft 3: Corrected nonsensical productions for [version-number](#) and [property-value](#). Changed [property-value](#) to permit non-initial `/` which was used in examples.

- Draft 3: Prohibited `[]`: unless it forms a [property-query](#), matching the existing behaviour of implementations and simplifying some implementation strategies.
- Draft 3: Added a definition of [ignorable-format-control](#) characters and prohibited these from separating lexical elements. This is a change with respect to the behaviour of existing implementations.
- Draft 3: [2.5.3.4, Valid Values and Resolved Sets](#): added support for a decimal mark and matching based on binary64 floating-point, to match existing implementations.
- Draft 3: [1, Terminology and Notation](#): added a definition of the code point complement and a discussion of its properties.
- Draft 3: Changed the proposed `\xcN` to `\xIN` in [named-element](#), since `\xcN` is currently parsed as `\x0C N`, whereas `\xIN` is currently a lexical error.
- Draft 4: Simplified the main [UnicodeSet](#) grammar removing backward compatibility measures for [named-element](#) as a set, based on feedback from ICU-TC.
- Draft 4: Added highlighting to the [string-element](#) production to reflect the lack of support for [named-element](#) in ICU 78; added struck-out `\P`, `\p`, and `\N`.
- Draft 4: Changed the proposed `\xIN` and `\xN` to `\N` in [named-element](#) (using the same prefix for `{hex:literal:name}`, `{hex:name}`, and `{name}`) based on feedback from ICU-TC.
- Draft 4: [4, Conformance](#): Added more hypothetical options for pure extensions, and a discussion of compatibility considerations.
- Draft 4: Added `\e` and `\c` escapes to [escaped-element](#) to match ICU behaviour.

© 2025 Unicode, Inc. All Rights Reserved. The Unicode Consortium makes no expressed or implied warranty of any kind, and assumes no liability for errors or omissions. No liability is assumed for incidental and consequential damages in connection with or arising out of the use of the information or programs contained or accompanying this technical report. The Unicode [Terms of Use](#) apply.

Unicode and the Unicode logo are trademarks of Unicode, Inc., and are registered in some jurisdictions.