

Proposed Update Unicode® Technical Standard #18
UNICODE REGULAR EXPRESSIONS

| | |
|------------------------|---|
| Version | 20 |
| Editors | Mark Davis, Andy Heninger |
| Date | 2019-11-08 |
| This Version | http://www.unicode.org/reports/tr18/tr18-20.html |
| Previous Version | http://www.unicode.org/reports/tr18/tr18-19.html |
| Latest Version | http://www.unicode.org/reports/tr18/ |
| Latest Proposed Update | http://www.unicode.org/reports/tr18/proposed.html |
| Revision | 20 |

Summary

This document describes guidelines for how to adapt regular expression engines to use Unicode.

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This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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

- 0 Introduction
 - 0.1 Notation
 - 0.1.1 Character Classes
 - 0.2 Conformance
- 1 Basic Unicode Support: Level 1
 - 1.1 Hex Notation
 - 1.1.1 Hex Notation and Normalization
 - 1.2 Properties
 - 1.2.1 General Category Property
 - 1.2.2 Script Property
 - 1.2.3 Other Properties
 - 1.2.4 Age
 - 1.2.5 Blocks
 - 1.3 Subtraction and Intersection
 - 1.4 Simple Word Boundaries
 - 1.5 Simple Loose Matches
 - 1.6 Line Boundaries
 - 1.7 Code Points
- 2 Extended Unicode Support: Level 2
 - 2.1 Canonical Equivalents
 - 2.2 Extended Grapheme Clusters and Character Classes with Strings
 - 2.2.1 Character Classes with Strings
 - 2.3 Default Word Boundaries
 - 2.4 Default Case Conversion
 - 2.5 Name Properties
 - 2.5.1 Individually Named Characters
 - 2.6 Wildcards in Property Values

[2.7 Full Properties](#)[2.8 Optional Properties](#)[3 Tailored Support: Level 3](#)[3.1 Tailored Punctuation](#)[3.2 Tailored Grapheme Clusters](#)[3.3 Tailored Word Boundaries](#)[3.4 Tailored Loose Matches \(Retracted\)](#)[3.5 Tailored Ranges \(Retracted\)](#)[3.6 Context Matching](#)[3.7 Incremental Matches](#)[3.8 Unicode Set Sharing \(Retracted\)](#)[3.9 Possible Match Sets](#)[3.10 Folded Matching \(Retracted\)](#)[3.11 Submatchers](#)[Annex A: Character Blocks](#)[Annex B: Sample Collation Grapheme Cluster Code](#)[Annex C: Compatibility Properties](#)[Annex D: Data Files](#)[Annex E: Resolving Character Classes with Strings](#)[References](#)[Acknowledgments](#)[Modifications](#)

0 Introduction

Review note: This document is organized into main sections by Levels, not by topics. That means that the topics are split, which makes it harder to follow. An alternative structure would be by topic; within a given topic, there could be multiple conformance clauses and discussions relevant to them. Feedback is welcome as to whether the document should be restructured by topic or not; such restructuring would probably happen in a subsequent release.

The following describes general guidelines for extending regular expression engines (Regex) to handle Unicode. The following issues are involved in such extensions.

- Unicode is a large character set—regular expression engines that are only adapted to handle small character sets will not scale well.
- Unicode encompasses a wide variety of languages which can have very different characteristics than English or other western European text.

There are three fundamental levels of Unicode support that can be offered by regular expression engines:

- Level 1: Basic Unicode Support.** At this level, the regular expression engine provides support for Unicode characters as basic logical units. (This is independent of the actual serialization of Unicode as UTF-8, UTF-16BE, UTF-16LE, UTF-32BE, or UTF-32LE.) This is a minimal level for useful Unicode support. It does not account for end-user expectations for character support, but does satisfy most low-level programmer requirements. The results of regular expression matching at this level are independent of country or language. At this level, the user of the regular expression engine would need to write more complicated regular expressions to do full Unicode processing.
- Level 2: Extended Unicode Support.** At this level, the regular expression engine also accounts for extended grapheme clusters (what the end-user generally thinks of as a character), better detection of word boundaries, and canonical equivalence. This is still a default level—independent of country or language—but provides much better support for end-user expectations than the raw level 1, without the regular-expression writer needing to know about some of the complications of Unicode encoding structure.
- Level 3: Tailored Support.** At this level, the regular expression engine also provides for tailored treatment of characters, including country- or language-specific behavior. For example, the characters *ch* can behave as a single character in Slovak or traditional Spanish. The results of a particular regular expression reflect the end-users' expectations of what constitutes a character in their language, and the order of the characters. However, there is a performance impact to support at this level.
- Review item: remove the above bullet if we remove Level 3.**

In particular:

- Level 1 is the minimally useful level of support for Unicode. All regex implementations dealing with Unicode should be at least at Level 1.
- Level 2 is recommended for implementations that need to handle additional Unicode features. This level is achievable without too much effort. However, some of the subitems in Level 2 are more important than others: see [Level 2](#).
- Level 3 contains information about extensions only useful for specific applications. Features at this level may require further investigation for effective implementation.
- Review item: remove above item if we remove Level 3.**

One of the most important requirements for a regular expression engine is to document clearly what Unicode features are and are not supported. Even if higher-level support is not currently offered, provision should be made for the syntax to be extended in the future to encompass those features.

Note: The Unicode Standard is constantly evolving: new characters will be added in the future. This means that a regular expression that tests for currency symbols, for example, has different results in Unicode 2.0 than in Unicode 2.1, which added the euro sign currency symbol.

At any level, efficiently handling properties or conditions based on a large character set can take a lot of memory. A common mechanism for reducing the memory requirements—while still maintaining performance—is the two-stage table, discussed in Chapter 5 of *The Unicode Standard* [Unicode]. For example, the Unicode character properties required in **RL1.2 Properties** can be stored in memory in a two-stage table with only 7 or 8 Kbytes. Accessing those properties only takes a small amount of bit-twiddling and two array accesses.

Note: For ease of reference, the section ordering for this document is intended to be as stable as possible over successive versions. That may lead, in some cases, to the ordering of the sections being less than optimal.

0.1 Notation

In order to describe regular expression syntax, an extended BNF form is used:

| Syntax | Meaning |
|--------|-------------------------------------|
| x y | the sequence consisting of x then y |
| x* | zero or more occurrences of x |
| x? | zero or one occurrence of x |
| x y | either x or y |
| (x) | for grouping |
| "XYZ" | terminal character(s) |

0.1.1 Character Classes

Review note: The term "character range" has been replaced by Character Class, without highlighting each instance. Character Class is more widely used in implementations (and less misleading).

A Character Class represents a set of characters (and sequences of characters, when following Section 2.2.1 [Character Classes with Strings]). The following syntax for Character Classes is used and extended in successive sections.

```

CHARACTER_CLASS := "[" NEGATION? ITEM ( OPERATOR? ITEM)* "]"
ITEM             := "[" CHARACTER_CLASS "]"
                := CODE_POINT2
                := CODE_POINT2 "-" CODE_POINT2 // range
CODE_POINT2     := ESCAPE CODE_POINT
                := CODE_POINT
NEGATION        := "^"
OPERATOR        := "" // union (no separator): AUB
                := "||" // union (where desired for clarity): AUB
ESCAPE          := "\"
    
```

CODE_POINT refers to any Unicode code point from U+0000 to U+10FFFF. Whitespace is allowed between any elements, but to simplify the presentation the many occurrences of sequences of spaces (" ") are omitted.

Negation affects the entire value in square brackets. That is, [^a...z] = [^a...z].

For the purpose of regular expressions, in this document the terms "character" and "code point" are used interchangeably. Similarly, the terms "string" and "sequence of code points" are used interchangeably. Typically the code points of interest will be those representing characters. A Character Class is also referred to as the set of all characters specified by that Character Class.

In addition, for readability the simple parentheses are used where in practice a non-capturing group would be used. That is, (a|b|c) is written instead of (? :a|b|c).

Code points that are syntax characters or whitespace are typically escaped. For more information see [UAX31]. In examples, the syntax "\s" is sometimes used to indicate whitespace. See also *Annex C: Compatibility Properties*.

Note: This is only a **sample** syntax for the purposes of examples in this document. Regular expression syntax varies widely: the issues discussed here would need to be adapted to the syntax of the particular implementation. However, it is important to have a concrete syntax to correctly illustrate the different issues. In general, the syntax here is similar to that of **Perl Regular Expressions** [Perl]. In some cases, this gives multiple syntactic constructs that provide for the same functionality.

The following table gives examples of Character Classes:

| Character Class | Matches |
|---------------------|---------------------|
| [a-z A-Z 0-9] | ASCII alphanumerics |
| [a-z A-Z 0-9] | |

| | |
|--------------|--------------------------------------|
| [a-zA-Z0-9] | |
| [^a-zA-Z0-9] | anything but ASCII alphanumerics |
| [\ \ \] | the literal characters], -, <space> |

Where string offsets are used in examples, they are from zero to n (the length of the string), and indicate positions *between* characters. Thus in "abcde", the substring from 2 to 4 includes the two characters "cd".

The following additional notation is defined for use here and in other Unicode specifications:

| Syntax | Meaning | Note |
|--------------------|--|--|
| <code>\n</code> | As used within regular expressions, expands to the text matching the <i>n</i> th parenthesized group in the regular expression. (à la Perl) | <i>n</i> is an ASCII digit. Implementations may impose limits on the number of digits. |
| <code>\$n</code> | As used within replacement strings for regular expressions, expands to the text matching the <i>n</i> th parenthesized group in a corresponding regular expression. (à la Perl) | The value of \$0 is the entire expression. |
| <code>\$xyz</code> | As used within regular expressions or replacement strings, expands to an assigned variable value. | The "xyz" is of the form of an identifier. For example, given <code>\$greek_lower = [[:greek:]]&&[:lowercase:];</code> , the regular expression pattern "ab\$greek_lower" is equivalent to "ab[[:greek:]]&&[:lowercase:]". |

Because any character could occur as a literal in a regular expression, when regular expression syntax is embedded within other syntax it can be difficult to determine where the end of the regex expression is. Common practice is to allow the user to choose a delimiter like '/' in `/ab(c)*/`. The user can then simply choose a delimiter that is not in the particular regular expression.

0.2 Conformance

The following section describes the possible ways that an implementation can claim conformance to this [Unicode Technical Standard](#).

All syntax and API presented in this document is *only* for the purpose of illustration; there is absolutely no requirement to follow such syntax or API. Regular expression syntax varies widely: the features discussed here would need to be adapted to the syntax of the particular implementation. In general, the syntax in examples is similar to that of [Perl Regular Expressions](#) [Perl], but it may not be exactly the same. While the API examples generally follow [Java style](#), it is again *only* for illustration.

C0. An implementation claiming conformance to this specification at any Level shall identify the version of this specification and the version of the Unicode Standard.

C1. An implementation claiming conformance to Level 1 of this specification shall meet the requirements described in the following sections:

- RL1.1 Hex Notation
- RL1.2 Properties
- RL1.2a Compatibility Properties
- RL1.3 Subtraction and Intersection
- RL1.4 Simple Word Boundaries
- RL1.5 Simple Loose Matches
- RL1.6 Line Boundaries
- RL1.7 Supplementary Code Points

C2. An implementation claiming conformance to Level 2 of this specification shall satisfy C1, and meet the requirements described in the following sections:

- RL2.1 Canonical Equivalents
- RL2.2 Extended Grapheme Clusters [and Character Classes with Strings](#)
- RL2.3 Default Word Boundaries
- RL2.4 Default Case Conversion
- RL2.5 Name Properties
- RL2.6 Wildcards in Property Values
- RL2.7 Full Properties

C3. An implementation claiming conformance to Level 3 of this specification shall satisfy C1 and C2, and meet the requirements described in the following sections:

RL3.1 Tailored Punctuation
 RL3.2 Tailored Grapheme Clusters
 RL3.3 Tailored Word Boundaries
 RL3.6 Context Matching
 RL3.7 Incremental Matches
 RL3.9 Possible Match Sets
 RL3.11 Submatchers

Review item: remove C3 if we remove Level 3, plus references to C3. and its subclauses.

C4. An implementation claiming partial conformance to this specification shall clearly indicate which levels are completely supported (C1–C3), plus any additional supported features from higher levels.

For example, an implementation may claim conformance to Level 1, plus **Context Matching**, and **Incremental Matches**. Another implementation may claim conformance to Level 1, except for **Subtraction and Intersection**.

A regular expression engine may be operating in the context of a larger system. In that case some of the requirements may be met by the overall system. For example, the requirements of Section 2.1 **Canonical Equivalents** might be best met by making normalization available as a part of the larger system, and requiring users of the system to normalize strings where desired before supplying them to the regular-expression engine. Such usage is conformant, as long as the situation is clearly documented.

A conformance claim may also include capabilities added by an optional add-on, such as an optional library module, as long as this is clearly documented.

For backwards compatibility, some of the functionality may only be available if some special setting is turned on. None of the conformance requirements require the functionality to be available by default.

1 Basic Unicode Support: Level 1

Regular expression syntax usually allows for an expression to denote a set of single characters, such as `[a-z A-Z 0-9]`. Because there are a very large number of characters in the Unicode Standard, simple list expressions do not suffice.

1.1 Hex Notation

The character set used by the regular expression writer may not be Unicode, or may not have the ability to input all Unicode code points from a keyboard.

RL1.1 Hex Notation

To meet this requirement, an implementation shall supply a mechanism for specifying any Unicode code point (from U+0000 to U+10FFFF), using the hexadecimal code point representation.

The syntax must use the code point in its hexadecimal representation. For example, syntax such as `\uD834\uDD1E` or `\xF0x9D\x84\x9E` does not meet this requirement for expressing U+1D11E (𐝞) because "1D11E" does not appear in the syntax. In contrast, syntax such as `\U0001D11E`, `\x{1D11E}` or `\u{1D11E}` does satisfy the requirement for expressing U+1D11E.

A sample notation for listing hex Unicode characters within strings uses `"\u"` followed by four hex digits or `"\u{"` followed by any number of hex digits and terminated by `"}"`, with multiple characters indicated by separating the hex digits by spaces. This would provide for the following addition:

```
<codepoint> := <character>
<codepoint> := "\u" HEX_CHAR HEX_CHAR HEX_CHAR HEX_CHAR
<codepoint> := "\u{" HEX_CHAR+ "}"
<codepoints> := "\u{" HEX_CHAR+ (SEP HEX_CHAR+)* "}"
SEP := \s+
```

```
U_SHORT_MARK := "\u"
```

The following table gives examples of this hex notation:

| Syntax | Matches |
|---|--|
| <code>[\u{3040}-\u{309F} \u{30FC}]</code> | Hiragana characters, plus prolonged sound sign |
| <code>[\u{B2} \u{2082}]</code> | superscript ² and subscript ₂ |
| <code>[a \u{10450}]</code> | "a" and U+10450 SHAVIAN LETTER PEEP |
| <code>ab\u{63 64}</code> | "abcd" |

More advanced regular expression engines can also offer the ability to use the Unicode character name for readability. See [2.5 Name Properties](#).

For comparison, the following table shows some additional, current examples of escape syntax for Unicode code points:

| Type | Escaped Characters | | | | | Escaped String |
|----------------|--------------------|-----------|--------|--------|--------|--------------------------------------|
| Unescaped | ☹ | € | £ | a | <tab> | ☹€fa<tab> |
| Code Point† | U+1F47D | U+20AC | U+00A3 | U+0061 | U+0009 | U+1F47D U+20AC U+00A3 U+0061 U+0009 |
| CSS† | \1F47D | \20AC | \A3 | \61 | \9 | \1F47D \20AC \A3 \61 \9 |
| UTS18, Ruby | \u{1F47D} | \u{20AC} | \u{A3} | \u{61} | \u{9} | \u{1F47D 20AC A3 61 9} |
| Perl | \x{1F47D} | \x{20AC} | \x{A3} | \x{61} | \x{9} | \x{1F47D}\x{20AC}\x{A3}\u{61} |
| XML/HTML | 👽 | € | £ | a | 	 | 👽€£a	 |
| C++/Python/ICU | \U0001F47D | \u20AC | \u00A3 | \u0061 | \u0009 | \U0001F47D\u20AC\u00A3\u0061\u0009 |
| Java/JS/ICU* | \uD83D\uDC7D | \u20AC | \u00A3 | \u0061 | \u0009 | \uD83D\uDC7D\u20AC\u00A3\u0061\u0009 |
| URL* | %F0%9F%91%BD | %E2%82%AC | %C2%A3 | %61 | %09 | %F0%9F%91%BD%E2%82%AC%C2%A3%61%09 |
| XML/HTML* | 👽 | € | £ | a | 	 | 👽€£a	 |

† Following whitespace is consumed.
 * Does not satisfy [RL1.1](#)

1.1.1 Hex Notation and Normalization

The Unicode Standard treats certain sequences of characters as equivalent, such as the following:

| | |
|-----------|---|
| u + grave | U+0075 (u) LATIN SMALL LETTER U + U+0300 (◌̣) COMBINING GRAVE ACCENT |
| u_grave | U+00F9 (ù) LATIN SMALL LETTER U WITH GRAVE |

Literal text in regular expressions may be normalized (converted to equivalent characters) in transmission, out of the control of the authors of that text. For example, a regular expression may contain a sequence of literal characters 'u' and *grave*, such as the expression [aeiou◌̣◌̇◌̈] (the last three character being U+0300 (◌̣) COMBINING GRAVE ACCENT, U+0301 (◌̇) COMBINING ACUTE ACCENT, and U+0308 (◌̈) COMBINING DIAERESIS. In transmission, the two adjacent characters in Row 1 might be changed to the different expression containing just one character in Row 2, thus changing the meaning of the regular expression. Hex notation can be used to avoid this problem. In the above example, the regular expression should be written as [aeiou\u{300 301 308}] for safety.

A regular expression engine may also enforce a single, uniform interpretation of regular expressions by always normalizing input text to Normalization Form NFC before interpreting that text. For more information, see [UAX #15, Unicode Normalization Forms \[UAX15\]](#).

1.2 Properties

Because Unicode is a large character set **that is regularly extended**, a regular expression engine needs to provide for the recognition of whole categories of characters as well as simply **literal sets of characters and strings**; otherwise the listing of characters becomes impractical, **out of date**, and error-prone. This is done by providing syntax for sets of characters based on the Unicode character properties, **and related properties and functions**. Examples of such syntax are `\p{Script=Greek}` and `[:Script=Greek:]`, which both **stand for the set of characters that have the Script value of Greek**. In addition to the basic syntax, regex engines also need to allow them to be **combined with literal sets of characters and strings**. An example is `\p{Script=Greek}-\p{General_Category=Letter}`, which **stands for the set of characters that have the Script value of Greek and that do not have the General_Category value of Letter**.

There is a large number of **character properties** defined in the Unicode Character Database (UCD), which also provides the official data for mapping Unicode characters (and code points) to property values. See Section 2.7, **Full Properties**; [UAX #44, Unicode Character Database \[UAX44\]](#); and Chapter 4 in [The Unicode Standard \[Unicode\]](#). **For use in regular expressions, properties can also be considered to be defined by Unicode definitions and algorithms, and by data files and definitions associated with other Unicode Technical Standards, such as UTS #51 Unicode Emoji**. For example, this includes the Basic Emoji definition from UTS #51. The full list of recommended properties is in [2.7 Full Properties](#). **The defined Unicode string functions, such as isNFC() and isLowercase(), which also apply to single code points and may be useful to support in regular expressions:**

The values (codomain) of those character properties defined in the Unicode Character Database have the following types: Binary, Enumerated, Numeric, and String. (The UCD Catalog type is the same as Enumerated, while Miscellaneous is generally best treated like String.) Some String properties only ever have values that are single code point strings. Here are some examples.

| Property Type | Property | Code Point | Character | Property Value |
|---------------|--------------------------|------------|-----------|-----------------------------|
| Binary | White_Space | U+0020 | " " | True |
| Enumerated | Script | U+3032 | ⷀ | Common |
| Code point | Simple_Lowercase_Mapping | U+0041 | A | a \u{61} |
| String | Name | U+0020 | " " | SPACE \u{53 50 41 43 45} |

A property value can also be a *set* or *list* with multiple values. For example, the `Script_Extensions` property maps from code points to a set of enumerated Script values, such as the following. Those are categorized as Miscellaneous in the UCD.

| Property Type | Property | Code Point | Character | Property Value |
|--------------------------|-------------------|------------|-----------|----------------|
| Set of Enumerated Values | Script_Extensions | U+3032 | ⷀ | {Hira, Kana} |

Expressions involving properties with multiple values are most often tested for containment, not equality. An expression like `\p{Script_Extensions=Hira}` is interpreted as containment: matching each code point *cp* such that `Script_Extensions(cp) ⊇ {Hira}`. Thus, `\p{Script_Extensions=Hira}` will match both U+3032 (ⷀ) VERTICAL KANA REPEAT WITH VOICED SOUND MARK (with value {Hira Kana}) and U+3041 (ぁ) HIRAGANA LETTER SMALL A (with value {Hira}). That also allows the natural replacement of `\p{Script=Hira}` by `\p{Script_Extensions=Hira}`: the latter just adds characters that may be either Hira or some other script.

For a more detailed example, see [Section 1.2.2 Script and Script Extensions Properties](#).

In addition to properties of characters, there are also properties of strings (sequences of characters). As with properties of characters, properties of strings can have values that are binary, enumerated, code point, or string — or a set of such values. A property of strings is more general than a property of characters. In other words, any property of characters is also a property of strings; its domain is, however, limited to strings consisting of a single character. Data, definitions, and properties defined by the Unicode Standard and other Unicode Technical Standards, which map from strings to values, can thus be specified in this document as defining regular-expression properties. For example:

| Property Type | Property | Code Points | Characters | Property Value |
|---------------|-------------|---------------|------------|----------------|
| Binary | Basic_Emoji | U+231A WATCH | 🕒 | True |
| | | U+23F2 U+FE0F | 🕒 | True |
| | | U+0041 | A | False |
| | | U+0041 U+0042 | AB | False |

Note that such properties can always be “narrowed” to just contain code points. For example, `[\p{Basic_Emoji} && \p{any}]` is the set of characters in `Basic_Emoji`.

Note that negations of properties of strings or Character Classes with strings may not be valid in regular expressions. For more information, see [Annex E: Resolving Character Classes with Strings](#) and [Section 2.2.1 Character Classes with Strings](#).

The recommended names for UCD properties and property values are in [PropertyAliases.txt](#) and [PropertyValueAliases.txt](#). There are both abbreviated names and longer, more descriptive names. It is strongly recommended that both names be recognized, and that loose matching of property names be used, whereby the case distinctions, whitespace, hyphens, and underbar are ignored.

Note: It may be a useful implementation technique to load the Unicode tables that support properties and other features on demand, to avoid unnecessary memory overhead for simple regular expressions that do not use those properties.

Where a regular expression is expressed as much as possible in terms of higher-level semantic constructs such as *Letter*, it makes it practical to work with the different alphabets and languages in Unicode. The following is an example of a syntax addition that permits properties. Following Perl Syntax, the *p* is lowercase to indicate a positive match, and uppercase to indicate a negative match.

```
CHARACTER_CLASS := POSITIVE_SPEC | NEGATIVE_SPEC
ITEM           := POSITIVE_SPEC | NEGATIVE_SPEC
POSITIVE_SPEC  := ("\p{" PROP_SPEC "}") | ("[:" PROP_SPEC ":]")
NEGATIVE_SPEC  := ("\P{" PROP_SPEC "}") | ("[:^" PROP_SPEC ":]")
PROP_SPEC      := <binary_unicode_property>
PROP_SPEC      := <unicode_property> (":" | "=" | "!=" | "!=") PROP_VALUE
```

PROP_SPEC := <script_or_category_property_value> ("|" <script_or_category_property_value>)*
 PROP_VALUE := <unicode_property_value> ("|" <unicode_property_value>)*

The following table shows examples of this extended syntax to match properties:

| Syntax | Matches |
|--|---|
| <code>[\p{L} \p{Nd}]</code> | all letters and decimal digits |
| <code>[\p{letter} \p{decimal number}]</code> | |
| <code>[\p{letter decimal number}]</code> | |
| <code>[\p{L Nd}]</code> | anything that does not have the Greek script |
| <code>\P{script=greek}</code> | |
| <code>\P{script:greek}</code> | |
| <code>\p{script#greek}</code> | |
| <code>[:^script=greek:]</code> | |
| <code>[:^script:greek:]</code> | anything that has the enumerated property value East_Asian_Width = Narrow |
| <code>[\p{East Asian Width:Narrow}]</code> | |
| <code>\p{Whitespace}</code> | anything that has binary property value Whitespace = True |
| <code>\p{scx=Kana}</code> | The match is to all characters whose Script_Extensions property value <i>includes</i> the specified value(s). So this expression matches U+30FC, which has the Script_Extensions value {Hira, Kana} |

Review Note:

The syntax for expressing properties of strings is not settled; feedback is welcome.

The main reason for considering different options is that the match of the negation of set of strings is problematic, and needs to be disallowed in certain regex expressions. For more information, see Annex E: Resolving Character Classes with Strings and Section 2.2.1 Character Classes with Strings.

The options for implementation will be illustrated by two different binary properties:

| Property | Sample code point match | Sample string match | Description |
|-------------|-------------------------|-----------------------------------|--|
| White_Space | U+0020 SPACE | <i>none</i> | Matches a fixed set of code points |
| Basic_Emoji | U+231A WATCH | <U+23F2 U+FE0F> TIMER CLOCK | Matches a fixed set of strings (as defined by <i>ED-20. basic emoji set</i> in <i>UTS #51 Unicode Emoji</i>) — many of the strings being single characters. |

In the following text the term **invalid** means that such syntax should be detected and handled as a syntax error in regular expressions, much as `\p{SomePropertyThatDoesntExist}`.

A. \m notation (for “multicharacter property”)

With this notation, different syntax is used for properties of characters and properties of strings. For properties of strings, there is no single-character (uppercase) negation. Other ways of expressing negation (such as `[^Basic_Emoji:]` or `[^\m{Basic_Emoji}]`) would also be invalid (outside of certain contexts: see Annex E: Resolving Character Classes with Strings).

| | |
|------------------------------|----------------|
| <code>\m{White_Space}</code> | <i>valid?</i> |
| <code>\p{White_Space}</code> | valid |
| <code>\P{White_Space}</code> | valid |
| <code>\m{Basic_Emoji}</code> | valid |
| <code>\p{Basic_Emoji}</code> | <i>invalid</i> |
| <code>\P{Basic_Emoji}</code> | <i>invalid</i> |

| | |
|--|----------------|
| <code>[^\m{Basic_Emoji}]</code> | <i>invalid</i> |
| <code>[\p{Symbol}--\m{Basic_Emoji}]</code> | valid |

The chief reason given in favor of the `\m` notation is that it alerts the reader of the regular expression that a property of strings is being used, and that thus negation may be invalid.

B. `\p` notation

With this notation, `\p` is used both for properties of characters and properties of strings. Any expression that negated a property of strings would be invalid. So `\P{Basic_Emoji}`, `[^\p{Basic_Emoji}]`, and `[:^Basic_Emoji:]` would all be invalid (outside of certain contexts: see Annex E: Resolving Character Classes with Strings).

| | |
|--|---|
| <code>\p{White_Space}</code> | valid |
| <code>\P{White_Space}</code> | valid |
| <code>\p{Basic_Emoji}</code> | valid (same set as <code>\m</code> above) |
| <code>\P{Basic_Emoji}</code> | <i>invalid</i> |
| <code>[^\p{Basic_Emoji}]</code> | <i>invalid</i> |
| <code>[\p{Symbol}--\p{Basic_Emoji}]</code> | valid |

The chief reason given in favor of the `\p` notation is that it is better to have a unified syntax for "properties", and the regex compilation will signal any malformed expression in any event, just as it would for `\p{Bogus=Letter}` or `\p{Script=Bogus}`.

Given A and B, there are two possible options we could take:

1. Use `\p` notation for properties of characters and properties of strings. Add a note that some implementations may wish to use `\m` for properties of strings.
2. Use `\p` notation for properties of characters and `\m` notation for properties of strings. Add a note that some implementations may wish to use `\p` for both.

Feedback on this issue is welcome.

Some properties are binary: they are either true or false for a given code point. In that case, only the property name is required. Others have multiple values, so for uniqueness both the property name and the property value need to be included.

For example, **Alphabetic** is a binary property, but it is also a value of the enumerated `Line_Break` property. So `\p{Alphabetic}` would refer to the binary property, whereas `\p{Line_Break:Alphabetic}` or `\p{Line_Break=Alphabetic}` would refer to the enumerated `Line_Break` property.

There are two exceptions to the general rule that expressions involving properties with multiple value should include both the property name and property value. The **Script** and **General_Category** properties commonly have their property name omitted. Thus `\p{Unassigned}` is equivalent to `\p{General_Category = Unassigned}`, and `\p{Greek}` is equivalent to `\p{Script:Greek}`.

RL1.2 Properties

To meet this requirement, an implementation shall provide at least a minimal list of properties, consisting of the following:

- *General_Category*
- *Script and Script_Extensions*
- *Alphabetic*
- *Uppercase*
- *Lowercase*
- *White_Space*
- *Noncharacter_Code_Point*
- *Default_Ignorable_Code_Point*
- *ANY, ASCII, ASSIGNED*

The values for these properties must follow the Unicode definitions, and include the property and property value aliases from the UCD. Matching of Binary, Enumerated, Catalog, and Name values, must follow the

Matching Rules from [UAX44] with one exception: implementations are not required to ignore an initial prefix string of "is" in property values.

RL1.2a Compatibility Properties

To meet this requirement, an implementation shall provide the properties listed in Annex C: Compatibility Properties, with the property values as listed there. Such an implementation shall document whether it is using the Standard Recommendation or POSIX-compatible properties.

In order to meet requirements [RL1.2](#) and [RL1.2a](#), the implementation must satisfy the Unicode definition of the properties for the supported version of The Unicode Standard, rather than other possible definitions. However, the names used by the implementation for these properties may differ from the formal Unicode names for the properties. For example, if a regex engine already has a property called "Alphabetic", for backwards compatibility it may need to use a distinct name, such as "Unicode_Alphabetic", for the corresponding property listed in [RL1.2](#).

Implementers may add aliases beyond those recognized in the UCD. For example, in the case of the the Age property an implementation could match the defined aliases "3.0" and "V3_0", but also match "3", "3.0", "V3.0", and so on. However, implementers must be aware that such additional aliases may cause problems if they collide with future UCD aliases for *different* values.

Ignoring an initial "is" in property values is optional. Loose matching rule [UAX44-LM3](#) in [UAX44] specifies that occurrences of an initial prefix of "is" are ignored, so that, for example, "Greek" and "isGreek" are equivalent as property values. Because existing implementations of regular expressions commonly make distinctions based on the presence or absence of "is", this requirement from [UAX44] is dropped.

For more information on properties, see [UAX #44, Unicode Character Database \[UAX44\]](#).

Of the properties in [RL1.2](#), `General_Category` and `Script` have enumerated property values with more than two values; the other properties are binary. An implementation that does not support non-binary enumerated properties can essentially "flatten" the enumerated type. Thus, for example, instead of `\p{script=latin}` the syntax could be `\p{script_latin}`.

When property_x is defined to have values that are sets of other values, the notation `\p{propertyx=valuey}` represents the set of all code points whose property values contain value_y. For example, the `Script_Extensions` property value for U+30FC (ー) is the set {Hiragana, Katakana}. So U+30FC (ー) is contained in `\p{Script_Extensions=Hiragana}`, and is also contained in `\p{Script_Extensions=Katakana}`.

Review note: the point made in the above paragraph has been moved earlier in this section.

1.2.1 General Category Property

The most basic overall character property is the `General_Category`, which is a basic categorization of Unicode characters into: *Letters, Punctuation, Symbols, Marks, Numbers, Separators, and Other*. These property values each have a single letter abbreviation, which is the uppercase first character except for separators, which use Z. The official data mapping Unicode characters to the `General_Category` value is in [UnicodeData.txt](#).

Each of these categories has different subcategories. For example, the subcategories for *Letter* are *uppercase, lowercase, titlecase, modifier, and other* (in this case, *other* includes uncased letters such as Chinese). By convention, the subcategory is abbreviated by the category letter (in uppercase), followed by the first character of the subcategory in lowercase. For example, *Lu* stands for *Uppercase Letter*.

Note: Because it is recommended that the property syntax be lenient as to spaces, casing, hyphens and underbars, any of the following should be equivalent: `\p{Lu}`, `\p{Lu}`, `\p{uppercase letter}`, `\p{Uppercase Letter}`, `\p{Uppercase_Letter}`, and `\p{uppercaseletter}`

The `General_Category` property values are listed below. For more information on the meaning of these values, see [UAX #44, Unicode Character Database \[UAX44\]](#).

| Abb. | Long form | Abb. | Long form | Abb. | Long form |
|------|------------------|------|-----------------------|------|---------------------|
| L | Letter | S | Symbol | Z | Separator |
| Lu | Uppercase Letter | Sm | Math Symbol | Zs | Space Separator |
| Ll | Lowercase Letter | Sc | Currency Symbol | Zl | Line Separator |
| Lt | Titlecase Letter | Sk | Modifier Symbol | Zp | Paragraph Separator |
| Lm | Modifier Letter | So | Other Symbol | C | Other |
| Lo | Other Letter | P | Punctuation | Cc | Control |
| M | Mark | Pc | Connector Punctuation | Cf | Format |
| Mn | Non-Spacing Mark | Pd | Dash Punctuation | Cs | Surrogate |

| | | | | | |
|----|------------------------|----|---------------------|----|-------------|
| Mc | Spacing Combining Mark | Po | Open Punctuation | Co | Private Use |
| Me | Enclosing Mark | Pe | Close Punctuation | Cn | Unassigned |
| N | Number | Pi | Initial Punctuation | - | Any* |
| Nd | Decimal Digit Number | Pf | Final Punctuation | - | Assigned* |
| Nl | Letter Number | Po | Other Punctuation | - | ASCII* |
| No | Other Number | | | | |

Starred entries in the table are not part of the enumeration of General_Category values. They are explained below.

| Value | Matches | Equivalent to | Notes |
|----------|---|---------------------------------|--|
| Any | all code points | <code>[\u{0}-\u{10FFFF}]</code> | In some regular expression languages, <code>\p{Any}</code> may be expressed by a period (" <code>.</code> "), but that usage may exclude newline characters. |
| Assigned | all assigned characters (for the target version of Unicode) | <code>\p{Cn}</code> | This also includes all private use characters. It is useful for avoiding confusing double negatives. Note that <code>Cn</code> includes noncharacters, so <code>Assigned</code> excludes them. |
| ASCII | all ASCII characters | <code>[\u{0}-\u{7F}]</code> | |

1.2.2 Script and Script Extensions Properties

A regular-expression mechanism may choose to offer the ability to identify characters on the basis of other Unicode properties besides the General Category. In particular, Unicode characters are also divided into scripts as described in UAX #24, *Unicode Script Property* [UAX24] (for the data file, see [Scripts.txt](#)). Using a property such as `\p{sc=Greek}` allows implementations to test whether letters are Greek or not.

Some characters, such as U+30FC (`ー`) KATAKANA-HIRAGANA PROLONGED SOUND MARK, are regularly used with multiple scripts. For such characters the `Script_Extensions` property (abbreviated as `scx`) identifies the set of associated scripts. The following shows some sample characters with their `Script` and `Script_Extensions` property values:

| Code | Char | Name | sc | scx |
|--------|----------------|---|------------------|--------------------------------------|
| U+3042 | <code>あ</code> | HIRAGANA LETTER A | Hira | {Hira} |
| U+30FC | <code>ー</code> | KATAKANA-HIRAGANA PROLONGED SOUND MARK | Zyyy = Common | {Hira, Kana} |
| U+3099 | <code>゛</code> | COMBINING KATAKANA-HIRAGANA VOICED SOUND MARK | Zinh = Inherited | {Hira, Kana} |
| U+30FB | <code>・</code> | KATAKANA MIDDLE DOT | Zyyy = Common | {Bopo, Hang, Hani, Hira, Kana, Yiii} |

The expression `\p{sc=Hira}` includes those characters whose `Script` value is Hira, while the expression `\p{scx=Hira}` includes all the characters whose `Script_Extensions` value contains Hira. The following table shows the difference:

| Expression | Contents of Set |
|---------------------------|--|
| <code>\p{sc=Hira}</code> | [あ-け > - ら じ ろ] |
| <code>\p{scx=Hira}</code> | [、 - " 〃 ‹ -] ≡ - 〰 ~ - \ XX ☒ - ☒ あ-け ズ = ・ - - ^ / - ○ (-) - (至) ⊖ - ⊗ 1月-12月0点-24点 平成- 21日-31日、。- - ' ° じ ろ] |

The expression `\p{scx=Hira}` contains not only the characters in `\p{script=Hira}`, but many other characters such as U+30FC (`ー`), which are either Hiragana or Katakana.

In most cases, script extensions are a superset of the script values (`\p{scx=X} ⊇ \p{sc=X}`). However, in some cases that is not true. For example, the `Script` property value for U+30FC (`ー`) is Common, but the `Script_Extensions` value for U+30FC (`ー`) does not contain the script value Common. In other words, `\p{scx=Common} ⊈ \p{sc=Common}`.

The usage model for the `Script` and `Script_Extensions` properties normally requires that people construct somewhat more complex regular expressions, because a great many characters (Common and Inherited) are shared between scripts. Documentation should point users to the description in [UAX24]. The values for `Script_Extensions` are likely to be extended over time as new information is gathered on the use of characters with different scripts. For more information, see [The Script_Extensions Property](#) in UAX #24, [Unicode Script Property](#) [UAX24].

1.2.3 Other Properties

Review note: this section was moved to Section 2.8 Optional Properties; other sections have not yet been renumbered.

1.2.4 Age

As defined in the Unicode Standard, the Age property (in the `DerivedAge` data file in the UCD) specifies the first version of the standard in which each character was assigned. It does not refer to how long it has been encoded, nor does it indicate the historic status of the character.

In regex expressions, the Age property is used to indicate the characters that were in a particular version of the Unicode Standard. That is, a character has the Age property of that version or less. Thus `\p{age=3.0}` includes the letter `a`, which was included in Unicode 1.0. To get characters that are new in a particular version, subtract off the previous version as described in [1.3 Subtraction and Intersection](#). For example: `[\p{age=3.1} -- \p{age=3.0}]`.

1.2.5 Blocks

Unicode blocks have an associated enumerated property, the Block property. However, there are some very significant caveats to the use of Unicode blocks for the identification of characters: see [Annex A: Character Blocks](#). If blocks are used, some of the names can collide with Script names, so they should be distinguished, with syntax such as `\p{Greek Block}` or `\p{Block=Greek}`.

1.3 Subtraction and Intersection

As discussed earlier, character properties are essential with a large character set. In addition, there needs to be a way to "subtract" characters from what is already in the list. For example, one may want to include all non-ASCII letters without having to list every character in `\p{letter}` that is not one of those 52.

RL1.3 Subtraction and Intersection

To meet this requirement, an implementation shall supply mechanisms for union, intersection and set-difference of sets of characters within regular expression character class expressions.

The following is an example of a syntax extension to handle set operations:

```

ITEM      := "[" ITEM "]" // for grouping
OPERATOR := "" // no separator → union
          := "|" // union: A|B
          := "&&" // intersection: A&B
          := "--" // set difference: A-B
          := "~" // symmetric difference: (A-B)U(B-A) = (A|B)-(A&B)

```

Implementations may also choose to offer other set operations. The [symmetric difference](#) of two sets is [also](#) useful. It is defined as being the union minus the intersection. Thus `[\p{letter}~\p{ascii}]` is equivalent to `[(\p{letter}\p{ascii})--\p{letter}&&\p{ascii}]`.

Review Notes:

- The syntax might be better merged into 0.1.1. Feedback is welcome.
- Symmetric difference is not typically supported, and may just be distracting; feedback is welcome as to whether we should remove it.

There are two ways of adding syntax while maintaining backwards compatibility. One choice is to double the symbols, as in the above notation, which allows the operators to appear adjacent to ranges without ambiguity, such as `[\p{letter}--a-z]`. Alternatively, an engine can have syntax that requires that both sides of every operator be sets or property expressions: `[\p{letter}-[a-z]]`, which is arguably clearer.

For discussions of support by various engines, see:

- <https://www.regular-expressions.info/charclassintersect.html>
- <https://www.regular-expressions.info/charclasssubtract.html>

Binding or precedence may vary by regular expression engine, so [as a user](#) it is safest to always disambiguate using brackets to be sure. In particular, precedence may put all operators on the same level, or may take union as binding more closely. For example, where `A..F` stand for expressions, not characters:

| Expression | Precedence | Interpreted as | Interpreted as |
|--------------|--|----------------------|---|
| [AB--CD&&EF] | Union, intersection, and difference bind at the same level | [[[[[AB]--C]D]&&E]F] | clone(A).add(B) .remove(C).add(D) .retain(E).add(F) |
| | Union binds more closely than difference or intersection | [[[AB]--[CD]]&&[EF]] | clone(A).add(B) .remove(clone(C).add(D)) .retain(clone(E).add(F)) |

Binding at the same level is used in this specification.

The following table shows various examples of set subtraction:

| Expression | Matches |
|---|--|
| [\p{L}--QW] | all letters but Q and W |
| [\p{N}--[\p{Nd}--0-9]] | all non-decimal numbers, plus 0–9 |
| [\u{0}-\u{7F}--[P{letter}]] | all letters in the ASCII range, by subtracting non-letters |
| [\p{Greek}--[\p{GREEK SMALL LETTER ALPHA}]] | Greek letters except alpha |
| [\p{Assigned}--[\p{Decimal Digit Number}--a-fA-F a - f A - F]] | all assigned characters except for hex digits (using a broad definition) |

The boolean expressions can also involve properties of strings or *Character Classes with strings*. The only restriction is that the complete boolean expression, once resolved, cannot be a negated set of strings. Thus the following matches all code points that neither have a Script value of Greek nor are in Basic_Emoji:

```
[^\p{Script=Greek} -- \p{Basic_Emoji}]
```

whereas the following is malformed, and should result in a syntax error:

```
[^\p{Basic_Emoji} -- \p{Script=Greek}]
```

For more information, see *Annex E: Resolving Character Classes with Strings* and *Section 2.2.1 Character Classes with Strings*.

1.4 Simple Word Boundaries

Most regular expression engines allow a test for word boundaries (such as by "b" in Perl). They generally use a very simple mechanism for determining word boundaries: one example of that would be having word boundaries between any pair of characters where one is a <word_character> and the other is not, or at the start and end of a string. This is not adequate for Unicode regular expressions.

RL1.4 Simple Word Boundaries

To meet this requirement, an implementation shall extend the word boundary mechanism so that:

- 1. The class of <word_character> includes all the Alphabetic values from the Unicode character database, from *UnicodeData.txt*, plus the decimals (General_Category=Decimal_Number, or equivalently Numeric_Type=Decimal), and the U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER (Join_Control=True). See also *Annex C: Compatibility Properties*.*
- 2. Nonspacing marks are never divided from their base characters, and otherwise ignored in locating boundaries.*

Level 2 provides more general support for word boundaries between arbitrary Unicode characters which may override this behavior.

1.5 Simple Loose Matches

Most regular expression engines offer caseless matching as the only loose matching. If the engine does offer this, then it needs to account for the large range of cased Unicode characters outside of ASCII.

RL1.5 Simple Loose Matches

To meet this requirement, if an implementation provides for case-insensitive matching, then it shall provide at least the simple, default Unicode case-insensitive matching, and specify which properties are closed and which

are not.

To meet this requirement, if an implementation provides for case conversions, then it shall provide at least the simple, default Unicode case folding.

In addition, because of the vagaries of natural language, there are situations where two different Unicode characters have the same uppercase or lowercase. To meet this requirement, implementations must implement these in accordance with the Unicode Standard. For example, the Greek U+03C3 "σ" *small sigma*, U+03C2 "ς" *small final sigma*, and U+03A3 "Σ" *capital sigma* all match.

Some caseless matches may match one character against two: for example, U+00DF "ß" matches the two characters "SS". And case matching may vary by locale. However, because many implementations are not set up to handle this, at Level 1 only simple case matches are necessary. To correctly implement a caseless match, see *Chapter 3, Conformance* of [Unicode]. The data file supporting caseless matching is [CaseData].

To meet this requirement, where an implementation also offers case conversions, these must also follow *Chapter 3, Conformance* of [Unicode]. The relevant data files are [SpecialCasing] and [UData].

Matching case-insensitively is one example of matching under an equivalence relation:

A regular expression R matches *under an equivalence relation E* whenever for all strings S and T:

If S is equivalent to T under E, then R matches S if and only if R matches T.

In the Unicode Standard, the relevant equivalence relation for case-insensitivity is established according to whether two strings case fold to the same value. The case folding can either be simple (a 1:1 mapping of code points) or full (with some 1:n mappings).

- "ABC" and "Abc" are equivalent under both full and simple case folding.
- "cliff" (with the "ff" ligature) and "CLIFF" are equivalent under full case folding, but not under simple case folding.

In practice, regex APIs are not set up to match parts of characters. For this reason, full case equivalence is difficult to handle with regular expressions. For more information, see *Section 2.1, Canonical Equivalents*.

For case-insensitive matching:

1. Each string literal is matched case-insensitively. That is, it is *logically* expanded into a sequence of OR expressions, where each OR expression lists all of the characters that have a simple case-folding to the same value.
 - For example, /Dåb/ matches as if it were expanded into /(?:d|D)(?:â|À|\u{212B})(?:b|B)/. (The \u{212B} is an angstrom sign, identical in appearance to Å.)
 - Back references are subject to this logical expansion, such as /(?:i)(a.c)\1/, where \1 matches what is in the first grouping.
2. **(optional)** Each character class is closed under case. That is, it is logically expanded into a set of code points, and then closed by adding all simple case equivalents of each of those code points.
 - For example, [\p{Block=Phonetic_Extensions} [A-E]] is a character class that matches 133 code points (under Unicode 6.0). Its case-closure adds 7 more code points: a-e, P, and ð, for a total of 140 code points.

For condition #2, in both property character classes and explicit character classes, closing under simple case-insensitivity means including characters not in the set. For example:

- The case-closure of \p{Block=Phonetic_Extensions} includes two characters not in that set, namely P and ð.
- The case-closure of [A-E] includes five characters not in that set, namely [a-e].

Conformant implementations can choose whether and how to apply condition #2: the only requirement is that they declare what they do. For example, an implementation may:

- A. uniformly apply condition #2 to all property and explicit character classes
- B. uniformly not apply condition #2 to any property or explicit character classes
- C. apply condition #2 only within the scope of a switch
- D. apply condition #2 to just specific properties and/or explicit character classes

1.6 Line Boundaries

Most regular expression engines also allow a test for line boundaries: end-of-line or start-of-line. This presumes that lines of text are separated by line (or paragraph) separators.

RL1.6 Line Boundaries

To meet this requirement, if an implementation provides for line-boundary testing, it shall recognize not only CRLF, LF, CR, but also NEL (U+0085), PARAGRAPH SEPARATOR (U+2029) and LINE SEPARATOR (U+2028).

Formfeed (U+000C) also normally indicates an end-of-line. For more information, see Chapter 3 of [Unicode].

These characters should be uniformly handled in determining logical line numbers, start-of-line, end-of-line, and arbitrary-character implementations. Logical line number is useful for compiler error messages and the like. Regular expressions often allow for SOL

and EOL patterns, which match certain boundaries. Often there is also a "non-line-separator" arbitrary character pattern that excludes line separator characters.

The behavior of these characters may also differ depending on whether one is in a "multiline" mode or not. For more information, see *Anchors and Other "Zero-Width Assertions"* in Chapter 3 of [Friedl].

A newline sequence is defined to be any of the following:

`\u{A} | \u{B} | \u{C} | \u{D} | \u{85} | \u{2028} | \u{2029} | \u{D A}`

1. Logical line number

- The line number is increased by one for each occurrence of a newline sequence.
- Note that different implementations may call the first line either line zero or line one.

2. Logical beginning of line (often "^")

- SOL is at the start of a file or string, and depending on matching options, also immediately following any occurrence of a newline sequence.
- There is no empty line within the sequence `\u{D A}`, that is, between the first and second character.
- Note that there may be a separate pattern for "beginning of text" for a multiline mode, one which matches only at the beginning of the first line. For example, in Perl this is `\A`.

3. Logical end of line (often "\$")

- EOL at the end of a file or string, and depending on matching options, also immediately preceding a final occurrence of a newline sequence.
- There is no empty line within the sequence `\u{D A}`, that is, between the first and second character.
- SOL and EOL are not symmetric because of multiline mode: EOL can be interpreted in at least three different ways:
 - a. EOL matches at the end of the string
 - b. EOL matches before final newline
 - c. EOL matches before any newline

4. Arbitrary character pattern (often ".")

- Where the 'arbitrary character pattern' matches a newline sequence, it must match all of the newline sequences, and `\u{D A}` (CRLF) *should* match as if it were a single character. (The recommendation that CRLF match as a single character is, however, not required for conformance to RL1.6.)
- Note that `^$` (an empty line pattern) should not match the empty string within the sequence `\u{D A}`, but should match the empty string within the reversed sequence `\u{A D}`.

It is strongly recommended that there be a regular expression meta-character, such as `"\R"`, for matching all line ending characters and sequences listed above (for example, in #1). This would correspond to something equivalent to the following expression. That expression is slightly complicated by the need to avoid backup.

`(?:\u{D A}|(?:!\u{D A})[\u{A}-\u{D}\u{85}\u{2028}\u{2029}])`

Note: For some implementations, there may be a performance impact in recognizing CRLF as a single entity, such as with an arbitrary pattern character ("."). To account for that, an implementation may also satisfy R1.6 if there is a mechanism available for converting the sequence CRLF to a single line boundary character before regex processing.

For more information on line breaking, see [UAX14].

1.7 Code Points

A fundamental requirement is that Unicode text be interpreted semantically by code point, not code units.

RL1.7 Supplementary Code Points

To meet this requirement, an implementation shall handle the full range of Unicode code points, including values from $U+FFFF$ to $U+10FFFF$. In particular, where UTF-16 is used, a sequence consisting of a leading surrogate followed by a trailing surrogate shall be handled as a single code point in matching.

UTF-16 uses pairs of 16-bit code units to express code points above $FFFF_{16}$, while UTF-8 uses from two to four 8-bit code units to represent code points above $7F_{16}$. Surrogate pairs (or their equivalents in other encoding forms) are to be handled internally as single code point values. In particular, `[\u{0}-\u{10000}]` will match all the following sequence of code units:

| Code Point | UTF-8 Code Units | UTF-16 Code Units | UTF-32 Code Units |
|------------|------------------|-------------------|-------------------|
| 7F | 7F | 007F | 0000007F |
| 80 | C2 80 | 0080 | 00000080 |
| 7FF | DF BF | 07FF | 000007FF |
| 800 | E0 A0 80 | 0800 | 00000800 |
| FFFF | EF BF BF | FFFF | 0000FFFF |

| | | | |
|-------|-------------|-----------|----------|
| 10000 | F0 90 80 80 | D800 DC00 | 00010000 |
|-------|-------------|-----------|----------|

For backwards compatibility, some regex engines allow for switches to reset matching to be by code unit instead of code point. Such usage is discouraged. For example, in order to match € it is better to write `\u{1F44E}` rather than `\uD83D\uDC4E` (using UTF-16) or `\xF0\x9F\x91\x8E` (using UTF-8).

Note: It is permissible, but not required, to match an isolated surrogate code point (such as `\u{D800}`), which may occur in Unicode **16-bit** Strings. See [Unicode String](#) in the Unicode [\[Glossary\]](#).

2 Extended Unicode Support: Level 2

Level 1 support works well in many circumstances. However, it does not handle more complex languages or extensions to the Unicode Standard very well. Particularly important cases are canonical equivalence, word boundaries, extended grapheme cluster boundaries, and loose matches. (For more information about boundary conditions, see UAX #29, *Unicode Text Segmentation* [UAX29].)

Level 2 support matches much more what user expectations are for sequences of Unicode characters. It is still locale-independent and easily implementable. However, for compatibility with Level 1, it is useful to have some sort of syntax that will turn Level 2 support on and off.

The features comprising Level 2 are not in order of importance. In particular, the most useful and highest priority features in practice are:

- [RL2.3 Default Word Boundaries](#)
- [RL2.5 Name Properties](#)
- [RL2.6 Wildcards in Property Values](#)
- [RL2.7 Full Properties](#)

2.1 Canonical Equivalents

The equivalence relation for canonical equivalence is established by whether two strings are identical when normalized to NFD.

For most full-featured regular expression engines, it is quite difficult to match under canonical equivalence, which may involve reordering, splitting, or merging of characters. For example, all of the following sequences are canonically equivalent:

- A. o + horn + dot_below
1. U+006F (o) LATIN SMALL LETTER O
 2. U+031B (◌◌◌) COMBINING HORN
 3. U+0323 (◌◌◌) COMBINING DOT BELOW
- B. o + dot_below + horn
1. U+006F (o) LATIN SMALL LETTER O
 2. U+0323 (◌◌◌) COMBINING DOT BELOW
 3. U+031B (◌◌◌) COMBINING HORN
- C. o-horn + dot_below
1. U+01A1 (◌◌◌) LATIN SMALL LETTER O WITH HORN
 2. U+0323 (◌◌◌) COMBINING DOT BELOW
- D. o-dot_below + horn
1. U+1ECD (◌◌◌) LATIN SMALL LETTER O WITH DOT BELOW
 2. U+031B (◌◌◌) COMBINING HORN
- E. o-horn-dot_below
1. U+1EE3 (◌◌◌) LATIN SMALL LETTER O WITH HORN AND DOT BELOW

The regular expression pattern `/o\u{31B}/` matches the first two characters of A, the first and third characters of B, the first character of C, part of the first character together with the third character of D, and part of the character in E.

In practice, regex APIs are not set up to match parts of characters or handle discontinuous selections. There are many other edge cases: a combining mark may come from some part of the pattern far removed from where the base character was, or may not explicitly be in the pattern at all. It is also unclear what `./` should match and how back references should work.

It is feasible, however, to construct patterns that will match against NFD (or NFKD) text. That can be done by:

1. Putting the text to be matched into a defined normalization form (NFD or NFKD).
2. Having the user design the regular expression pattern to match against that defined normalization form. For example, the pattern should contain no characters that would not occur in that normalization form, nor sequences that would not occur.
3. Applying the matching algorithm on a code point by code point basis, as usual.

2.2 Extended Grapheme Clusters **and Character Classes with Strings**

One or more Unicode characters may make up what the user thinks of as a character. To avoid ambiguity with the computer use of the term *character*, this is called a *grapheme cluster*. For example, "G" + *acute-accent* is a grapheme cluster: it is thought of as a single character by users, yet is actually represented by two Unicode characters. The Unicode Standard defines *extended grapheme*

clusters that treat certain sequences as units, including Hangul syllables and base characters with combining marks. The precise definition is in UAX #29, Unicode Text Segmentation [UAX29]. These extended grapheme clusters are not the same as tailored grapheme clusters, which are covered in Section 3.2, Tailored Grapheme Clusters. However, the boundary definitions in CLDR are strongly recommended: they are more comprehensive than what is defined in UAX #29: Unicode Text Segmentation and include Indic extended grapheme clusters such as ksha.

RL2.2 Extended Grapheme Clusters and Character Classes with Strings

To meet this requirement, an implementation shall provide a mechanism for matching against an arbitrary extended grapheme cluster, a literal cluster, Character Classes with strings, and extended grapheme cluster boundaries.

For example, an implementation could interpret \x as matching any extended grapheme cluster, while interpreting "." as matching any single code point. It could interpret \b{g} as a zero-width match against any extended grapheme cluster boundary, and \B{g} as the negation of that.

More generally, it is useful to have zero width boundary detections for each of the different kinds of segment boundaries defined by Unicode ([UAX29] and [UAX14]). For example:

| Syntax | Zero-width Match at |
|--------|--|
| \b{g} | a Unicode extended grapheme cluster boundary |
| \b{w} | a Unicode word boundary. Note that this is different than \b alone, which corresponds to \w and \W. See Annex C: Compatibility Properties. |
| \b{1} | a Unicode line break boundary |
| \b{s} | a Unicode sentence boundary |

Thus \x is equivalent to .+\b{g}; proceed the minimal number of characters (but at least one) to get to the next extended grapheme cluster boundary.

2.2.1 Character Classes with Strings

Regular expression engines should also provide some mechanism for easily matching against Character Classes with strings, because they are more likely to match user expectations for many languages. One mechanism for doing that is to have explicit syntax for strings in Character Classes, as in the following addition to the syntax of Section 0.1.1 Character Classes:

```

ITEM := "\q{" [CODE_POINT (SP_CODE_POINT)* } "}"
SP := \u{20}

```

This syntax can also be used for tailored grapheme clusters (see Tailored Grapheme Clusters).

Review item: remove the above line if we remove Level 3.

The following table shows examples of use of the \q syntax:

| Expression | Matches |
|------------------------------|--|
| [a-z\q{x\u{323}}] | The characters a–z, and the string x with an under-dot (used in American languages) |
| [a-z\q{aa}] | The characters a–z, and the string aa (treated as a single character in Danish) |
| [a-z ñ \q{ch} \q{ll} \q{rr}] | Some lowercase characters in traditional Spanish |
| [a-z \q{🍌} \q{FR}] | Characters a–z and two emoji. Note that this is equivalent to [a-z 🍌 \q{FR}] because the first emoji is a single code point, while the second is two codepoints and thus requires the \q syntax. |

In implementing Character Classes with strings, the expression /[a-m \q{ch} \q{chh} \q{rr} ß-ç]/ should behave as the alternation [(ch | ch | rr) [a-m ß-ç]]. Note that such an alternation must have the strings ordered as longest-first to work correctly in arbitrary regex engines, because some regex engines try the leftmost matching alternative first. So it does not work to (logically) have shorter strings first (except, of course, shorter strings that are not initial substrings of longer strings).

If the implementation supports empty alternations, such as (ab|[ac-m]), then it can also handle empty strings: [\q{ab}[ac-m]\q{}].

Of course, such alternations can be optimized internally for speed and/or memory, such as (ab|[ac-m]) → ((ab?)[c-m]).

Like properties of strings, negated Character Classes with strings need to be handled specially: see Annex E: Resolving Character Classes with Strings.

Matching a *complemented* set containing strings like `\q{ch}` may behave differently in the two different modes: the normal mode where code points are the unit of matching, or the mode where extended grapheme clusters are the unit of matching. That is, the expression `{^ a m \q{ch} \q{rr}}` should behave in the following way:

| Mode | Behavior | Description |
|------------------|--|---|
| normal | <code>{?! ch rr [a-m]} {\u{0} \u{10FFFF}}</code> | failing with strings starting with a-m, ch, or rr, and otherwise advancing by one code point |
| grapheme cluster | <code>{?! ch rr [a-m]} \X</code> | failing with strings starting with a-m, ch, or rr, and otherwise advancing by one extended grapheme cluster |

An *extended* character set containing strings like `\q{ch}` plus embedded complement operations is interpreted as if the complement were pushed up to the top of the expression, using the following rewrites recursively:

| Original | Rewrite |
|-------------------------------|--------------------------------|
| <code>^^x</code> | <code>x</code> |
| <code>^x ^y</code> | <code>^(x && y)</code> |
| <code>^x y</code> | <code>^(x --- y)</code> |
| <code>x ^y</code> | <code>^(y --- x)</code> |
| <code>^x && ^y</code> | <code>^(x y)</code> |
| <code>^x --- y</code> | |
| <code>^x && y</code> | <code>y --- x</code> |
| <code>^x --- ^y</code> | |
| <code>x && ^y</code> | <code>x --- y</code> |
| <code>x --- ^y</code> | <code>x && y</code> |
| <code>^x --- ^y</code> | <code>x --- y</code> |
| <code>^x --- y</code> | <code>^(x --- y)</code> |
| <code>x --- ^y</code> | |

Applying these rewrites results in a simplification of the regex expression. Either the complement operations will be completely eliminated, or a single remaining complement operation will remain at the top level of the expression. Logically, then, the rest of the expression consists of a flat list of characters and/or multi-character strings; matching strings can then be handled as described above.

2.2.1 Grapheme Cluster Mode

A grapheme cluster mode behaves more like users' expectations for character boundaries, and is especially useful for handling canonically equivalent matching. In a grapheme cluster mode, matches are guaranteed to be on extended grapheme cluster boundaries. Each atomic literal of the pattern matches complete extended grapheme clusters, and thus behaves as if followed by `\b{g}`. Atomic literals include: a dot, a character class (like `[a-m]`), a sequence of characters (perhaps with some being escaped) that matches as a unit, or syntax that is equivalent to these. Note that in `/abc?/`, the "abc" is not matching as a unit; the `?` modifier is only affecting the last character, and thus the `ab` and the `ε` are separate atomic literals. To summarize:

| Syntax | Behaves Like | Description |
|------------------------|-----------------------------|--|
| <code>τ</code> | <code>\X</code> | matches a full extended grapheme cluster going forward |
| <code>{abc{gh}}</code> | <code>{abc{gh}}\b{g}</code> | matches only if the end point of the match is at a grapheme cluster boundary |
| <code>abcd</code> | <code>abcd\b{g}</code> | matches only if the end point of the match is at a grapheme cluster boundary |

Note that subdivisions can modify the behavior in this mode. Normally `/x(y)/` is equivalent to `/x(y)/` in terms of matching (where `x` and `y` are arbitrary literal character strings); that is, only the grouping is different. That is not true in grapheme cluster mode, where each atomic literal acts as if it is followed by `\b{g}`:

For example, `/x\u{308}/` is not the same as `/x(\u{308})/` in matching. The former behaves like `/x\u{308}\b{g}/` while the latter behaves like `/x\b{g}(\u{308}\b{g})/`. The latter will never match in grapheme cluster mode, since it would only match if there were a grapheme cluster boundary after the `x` and if `x` is followed by `\u{308}`, but that can never happen simultaneously.

2.3 Default Word Boundaries

RL2.3 Default Word Boundaries

To meet this requirement, an implementation shall provide a mechanism for matching Unicode default word boundaries.

The simple Level 1 support using simple `<word_character>` classes is only a very rough approximation of user word boundaries. A much better method takes into account more context than just a single pair of letters. A general algorithm can take care of character and word boundaries for most of the world's languages. For more information, see UAX #29, *Unicode Text Segmentation* [UAX29].

Note: Word boundaries and "soft" line-break boundaries (where one could break in line wrapping) are not generally the same; line breaking has a much more complex set of requirements to meet the typographic requirements of different languages. See UAX #14, *Line Breaking Properties* [UAX14] for more information. However, soft line breaks are not generally relevant to general regular expression engines.

A fine-grained approach to languages such as Chinese or Thai—languages that do not use spaces—requires information that is beyond the bounds of what a Level 2 algorithm can provide.

2.4 Default Case Conversion

RL2.4 Default Case Conversion

To meet this requirement, if an implementation provides for case conversions, then it shall provide at least the full, default Unicode case folding.

Previous versions of RL2.4 included full default Unicode case-insensitive matching. For most full-featured regular expression engines, it is quite difficult to match under code point equivalences that are not 1:1. For more discussion of this, see 1.5 *Simple Loose Matches* and 2.1 *Canonical Equivalents*. Thus that part of RL2.4 has been retracted.

Instead, it is recommended that implementations provide for full, default Unicode case conversion, allowing users to provide both patterns and target text that has been fully case folded. That allows for matches such as between U+00DF "ß" and the two characters "SS". Some implementations may choose to have a mixed solution, where they do full case matching on literals such as "Strauß", but simple case folding on character classes such as [ß].

To correctly implement case conversions, see [Case]. For ease of implementation, a complete case folding file is supplied at [CaseData]. Full case mappings use the data files [SpecialCasing] and [UData].

2.5 Name Properties

RL2.5 Name Properties

To meet this requirement, an implementation shall support individually named characters.

When using names in regular expressions, the data is supplied in both the **Name (na)** and **Name_Alias** properties in the UCD, as described in UAX #44, *Unicode Character Database* [UAX44], or computed as in the case of CJK Ideographs or Hangul Syllables. Name matching rules follow *Matching Rules* from [UAX44#UAX44-LM2].

The following provides examples of usage:

| Syntax | Set | Note |
|---|--------------------------|--|
| <code>\p{name=ZERO WIDTH NO-BREAK SPACE}</code> | <code>[\u{FEFF}]</code> | using the Name property |
| <code>\p{name=zerowidthno breakspace}</code> | <code>[\u{FEFF}]</code> | using the Name property, and <i>Matching Rules</i> [UAX44] |
| <code>\p{name=BYTE ORDER MARK}</code> | <code>[\u{FEFF}]</code> | using the Name_Alias property |
| <code>\p{name=BOM}</code> | <code>[\u{FEFF}]</code> | using the Name_Alias property (a second value) |
| <code>\p{name=HANGUL SYLLABLE GAG}</code> | <code>[\u{AC01}]</code> | with a computed name |
| <code>\p{name=BEL}</code> | <code>[\u{7}]</code> | the control character |
| <code>\p{name=BELL}</code> | <code>[\u{1F514}]</code> | the graphic symbol 🔔 |

Certain code points are not assigned names or name aliases in the standard. With the exception of "reserved", these should be given names based on *Code Point Label Tags* table in [UAX44], as shown in the following examples:

| Syntax | Set | Note |
|--|-----|------|
| <code>\p{name=private-use-E000}</code> | | |

| | | |
|----------------------------|------------|--|
| | [\u{E000}] | |
| \p{name=surrogate-D800} | [\u{D800}] | would only apply to isolated surrogate code points |
| \p{name=noncharacter-FDD0} | [\u{FDD0}] | |
| \p{name=control-0007} | [\u{7}] | |

Characters with the <reserved> tag in the [Code Point Label Tags](#) table of [UAX44] are *excluded*: the syntax \p{reserved-058F} would mean that the code point U+058F is unassigned. While this code point was unassigned in Unicode 6.0, it *is* assigned in Unicode 6.1 and thus no longer "reserved".

Implementers may add aliases beyond those recognized in the UCD. They must be aware that such additional aliases may cause problems if they collide with future character names or aliases. For example, implementations that used the name "BELL" for U+0007 broke when the new character U+1F514 (🔔) BELL was introduced.

Previous versions of this specification recommended supporting ISO control names from the Unicode 1.0 name field. These names are now covered by the name aliases (see [NameAliases.txt](#)). In four cases, the name field included both the ISO control name as well as an abbreviation in parentheses.

U+000A LINE FEED (LF)
 U+000C FORM FEED (FF)
 U+000D CARRIAGE RETURN (CR)
 U+0085 NEXT LINE (NEL)

These abbreviations were intended as alternate aliases, not as part of the name, but the documentation did not make this sufficiently clear. As a result, some implementations supported the entire field as a name. Those implementations might benefit from continuing to support them for compatibility. Beyond that, their use is not recommended.

The \p{name=...} syntax can be used meaningfully with wildcards (see [Section 2.6 Wildcards in Property Values](#)). For example, in Unicode 6.1, \p{name=/ALIEN/} would include a set of two characters:

- U+1F47D (👽) EXTRATERRESTRIAL ALIEN,
- U+1F47E (👾) ALIEN MONSTER

The namespace for the \p{name=...} syntax is the namespace for character names plus name aliases.

2.5.1 Individually Named Characters

The following provides syntax for specifying a code point by supplying the precise name. This syntax specifies a single code point, which can thus be used in [wherever \u{...} can be used](#).

```
<codepoint> := "\N{<character_name>}"
```

The \N syntax is related to the syntax \p{name=...}, but there are [three](#) important distinctions:

1. \N matches a single character [of a sequence](#), while \p matches a set of characters [\(when using wildcards\)](#).
2. The \p{name=<character_name>} may silently fail, if no character exists with that name. The \N syntax should instead cause a syntax error for an undefined name.

The namespace for the \N{name=...} syntax is the namespace for character names plus name aliases. Name matching rules follow [Matching Rules](#) from [UAX44#UAX44-LM2].

As with other property values, names should use a loose match, disregarding case, spaces and hyphen (the underbar character " _ " cannot occur in Unicode character names). An implementation may also choose to allow namespaces, where some prefix like "LATIN LETTER" is set globally and used if there is no match otherwise.

There are, however, three instances that require special casing with loose matching, where an extra test shall be made for the presence or absence of a hyphen:

- U+0F68 TIBETAN LETTER A and U+0F60 TIBETAN LETTER -A
- U+0FB8 TIBETAN SUBJOINED LETTER A and U+0FB0 TIBETAN SUBJOINED LETTER -A
- U+116C HANGUL JUNGSEONG OE and U+1180 HANGUL JUNGSEONG O-E

The following table gives examples of the \N syntax:

| Expression | Equivalent to |
|------------------------|---------------|
| \N{WHITE SMILING FACE} | \u{263A} |
| \N{whitesmilingface} | |

| | |
|---|--------------------------------|
| <code>\N{GREEK SMALL LETTER ALPHA}</code> | <code>\u{3B1}</code> |
| <code>\N{FORM FEED}</code> | <code>\u{C}</code> |
| <code>\N{SHAVIAN LETTER PEEP}</code> | <code>\u{10450}</code> |
| <code>[\N{GREEK SMALL LETTER ALPHA}-\N{GREEK SMALL LETTER BETA}]</code> | <code>[\u{3B1}-\u{3B2}]</code> |

2.6 Wildcards in Property Values

RL2.6 Wildcards in Property Values

To meet this requirement, an implementation shall support wildcards in Unicode property values.

Instead of a single property value, this feature allows the use of a regular expression to pick out a set of characters based on whether the property values match the regular expression. The regular expression must support at least wildcards; other regular expressions features are recommended but optional.

```
PROP_VALUE := <value>
            | "/" <regex expression> "/"
            | "@" <unicode_property> "@"
```

Note: Where regular expressions are used in matching, the case, spaces, hyphen, and underbar are significant; it is presumed that users will make use of regular-expression features to ignore these if desired.

The @...@ syntax is used to compare property values, and is primarily intended for string properties. It allows for expressions such as `[:^toNFKC_Casefold=@toNFKC@:]`, which expresses the set of all and only those code points **CP** such that `toNFKC_Casefold(CP) = toNFKC(CP)`. The value *identity* can be used in this context. For example, `\p{toLowerCase#=@identity@}` expresses the set of all characters that are changed by the `toLowerCase` mapping.

The following table shows examples of the use of wildcards:

| Expression | Matched Set |
|---|--|
| Characters whose NFD form contains a "b" (U+0062) in the value: | |
| <code>\p{toNfd=/b/}</code> | U+0062 (b) LATIN SMALL LETTER B U+1E03 (ȃ) LATIN SMALL LETTER B WITH DOT ABOVE U+1E05 (ȅ) LATIN SMALL LETTER B WITH DOT BELOW U+1E07 (ȇ) LATIN SMALL LETTER B WITH LINE BELOW |
| Characters with names starting with "LATIN LETTER" and ending with "P": | |
| <code>\p{name=/^LATIN LETTER.*P\$/}</code> | U+01AA (Ɔ) LATIN LETTER REVERSED ESH LOOP U+0294 (ʔ) LATIN LETTER GLOTTAL STOP U+0296 (ʓ) LATIN LETTER INVERTED GLOTTAL STOP U+1D18 (Ɔ) LATIN LETTER SMALL CAPITAL P |
| Characters with names containing "VARIATION" or "VARIANT": | |
| <code>\p{name=/VARIA(TION NT)/}</code> | U+180B () MONGOLIAN FREE VARIATION SELECTOR ONE ... U+180D () MONGOLIAN FREE VARIATION SELECTOR THREE U+299C (ㄣ) RIGHT ANGLE VARIANT WITH SQUARE U+303E (ㄩ) IDEOGRAPHIC VARIATION INDICATOR U+FE00 () VARIATION SELECTOR-1 ... U+FE0F () VARIATION SELECTOR-16 U+121AE (𐎠) CUNEIFORM SIGN KU4 VARIANT FORM U+12425 (𐎠𐎫) CUNEIFORM NUMERIC SIGN THREE SHAR2 VARIANT FORM U+1242F (𐎠𐎫𐎪) CUNEIFORM NUMERIC SIGN THREE SHARU VARIANT FORM U+12437 (𐎠𐎫𐎩) CUNEIFORM NUMERIC SIGN THREE BURU VARIANT FORM U+1243A (𐎠𐎫𐎪𐎩) CUNEIFORM NUMERIC SIGN THREE VARIANT FORM ESH16 ... U+12449 (𐎠𐎫𐎪𐎩𐎫) CUNEIFORM NUMERIC SIGN NINE VARIANT FORM ILIMMU A U+12453 (𐎠𐎫𐎪𐎩𐎫𐎪) CUNEIFORM NUMERIC SIGN FOUR BAN2 VARIANT FORM U+12455 (𐎠𐎫𐎪𐎩𐎫𐎪𐎩) CUNEIFORM NUMERIC SIGN FIVE BAN2 VARIANT FORM U+1245D (𐎠𐎫𐎪𐎩𐎫𐎪𐎩𐎫) CUNEIFORM NUMERIC SIGN ONE THIRD VARIANT FORM A U+1245E (𐎠𐎫𐎪𐎩𐎫𐎪𐎩𐎫𐎪) CUNEIFORM NUMERIC SIGN TWO THIRDS VARIANT FORM A U+E0100 () VARIATION SELECTOR-17 ... U+E01EF () VARIATION SELECTOR-256 |

Characters in the Letterlike symbol block with different toLowercase values:

| | |
|--|-------------------------------|
| <code>[\p{toLowercase≠@cp@}</code> | U+2126 (Ω) OHM SIGN |
| <code>& \p{Block=Letterlike Symbols}]</code> | U+212A (K) KELVIN SIGN |
| | U+212B (Å) ANGSTROM SIGN |
| | U+2132 (∩) TURNED CAPITAL F |

The lists in the examples above were extracted on the basis of Unicode 5.0; different Unicode versions may produce different results.

The following table some additional samples, illustrating various sets. A click on the link will use the online Unicode utilities on the Unicode website to show the contents of the sets. Note that these online utilities currently use single-letter operations.

| Expression | Description |
|---|--|
| <code>[[:name=/CJK/:]-[:ideographic:]]</code> | The set of all characters with names that contain CJK that are not Ideographic |
| <code>[[:name=/\bDOT\$/:]]</code> | The set of all characters with names that end with the word DOT |
| <code>[[:block=/(?i)arab/:]]</code> | The set of all characters in blocks that contain the sequence of letters "arab" (case-insensitive) |
| <code>[[:toNFKC=/\./:]]</code> | the set of all characters with toNFKC values that contain a literal period |

2.7 Full Properties

RL2.7 Full Properties

To meet this requirement, an implementation shall support all of the properties listed below that are in the supported version of the Unicode Standard (or Unicode Technical Standard, respectively), with values that match the Unicode definitions for that version.

To meet requirement RL2.7, the implementation must satisfy the Unicode definition of the properties for the supported version of Unicode (or Unicode Technical Standard, respectively), rather than other possible definitions. However, the names used by the implementation for these properties may differ from the formal Unicode names for the properties. For example, if a regex engine already has a property called "Alphabetic", for backwards compatibility it may need to use a distinct name, such as "Unicode_Alphabetic", for the corresponding property listed in RL1.2.

The list excludes provisional, contributory, obsolete, and deprecated properties. It also excludes specific properties: Unicode_1_Name, Unicode_Radical_Stroke, and the Unihan properties. The properties shown in the table with a gray background are covered by RL1.2 Properties. For more information on properties, see UAX #44, Unicode Character Database [UAX44]. Properties marked with * are properties of strings, not just single code points.

| General | Case | Shaping and Rendering |
|------------------------------|--------------------------|------------------------------|
| Name (Name_Alias) | Uppercase | Join_Control |
| Block | Lowercase | Joining_Group |
| Age | Lowercase_Mapping | Joining_Type |
| General_Category | Titlecase_Mapping | Vertical_Orientation |
| Script (Script_Extensions) | Uppercase_Mapping | Line_Break |
| White_Space | Case_Folding | Grapheme_Cluster_Break |
| Alphabetic | Simple_Lowercase_Mapping | Sentence_Break |
| Hangul_Syllable_Type | Simple_Titlecase_Mapping | Word_Break |
| Noncharacter_Code_Point | Simple_Uppercase_Mapping | East_Asian_Width |
| Default_Ignorable_Code_Point | Simple_Case_Folding | Prepended_Concatenation_Mark |
| Deprecated | Soft_Dotted | |
| Logical_Order_Exception | Cased | Bidirectional |
| Variation_Selector | Case_Ignorable | Bidi_Class |
| | Changes_When_Lowercased | Bidi_Control |

| | | |
|------------------------------|------------------------------|---------------------------|
| Numeric | Changes_When_Uppercased | Bidi_Mirrored |
| Numeric_Value | Changes_When_Titlecased | Bidi_Mirroring_Glyph |
| Numeric_Type | Changes_When_Casefolded | Bidi_Paired_Bracket |
| Hex_Digit | Changes_When_Casemapped | Bidi_Paired_Bracket_Type |
| ASCII_Hex_Digit | | |
| | Normalization | Miscellaneous |
| Identifiers | Canonical_Combining_Class | Math |
| ID_Continue | Decomposition_Type | Quotation_Mark |
| ID_Start | NFC_Quick_Check | Dash |
| XID_Continue | NFKC_Quick_Check | Sentence_Terminal |
| XID_Start | NFD_Quick_Check | Terminal_Punctuation |
| Pattern_Syntax | NFKD_Quick_Check | Diacritic |
| Pattern_White_Space | NFKC_Casefold | Extender |
| Identifier_Status | Changes_When_NFKC_Casefolded | Grapheme_Base |
| Identifier_Type | | Grapheme_Extend |
| | Emoji | Regional_Indicator |
| CJK | Emoji | Indic_Positional_Category |
| Ideographic | Emoji_Presentation | Indic_Syllabic_Category |
| Unified_Ideograph | Emoji_Modifier | |
| Radical | Emoji_Modifier_Base | |
| IDS_Binary_Operator | Emoji_Component | |
| IDS_Tertiary_Operator | Extended_Pictographic | |
| Equivalent_Unified_Ideograph | Basic_Emoji* | |
| | RGI_Emoji_Modifier_Sequence* | |
| | RGI_Emoji_Tag_Sequence* | |
| | RGI_Emoji_ZWJ_Sequence* | |
| | RGI_Emoji* | |

Property metadata for properties not listed in the UCD can be found via [Annex D: Data Files](#).

Review Notes:

- Some of the links under Emoji will not work yet.
- There were additional possible candidates for this list. These have been moved into 2.8 Optional Properties. Feedback is requested as to whether any of those would be useful to move here.
- The properties in SpecialCasing.txt such as Lowercase_Mapping are conditional on context. Their support would lead implementations to ignore the context, which may give incorrect results. They have thus been retracted.

The **Name** and **Name_Alias** properties are used in $\backslash p\{name=..}$ and $\backslash N\{..}$. The data in NamedSequences.txt is also used in $\backslash N\{..}$. For more information see [Section 2.5, Name Properties](#). The **Script** and **Script_Extensions** properties are used in $\backslash p\{scx=..}$. For more information, see [Section 1.2.2, Script_Property](#).

See also [2.5 Name Properties](#) and [2.6 Wildcards in Property Values](#).

2.8 Optional Properties

Review Note: This section was moved from [Section 1.2.3 Other Properties](#) and retitled to better reflects the optional nature of these properties.

| | |
|-------------------------------|---|
| <code>{:isUppercase:}</code> | The set of all characters X such that $\text{toUppercase}(X) = X$ |
| <code>{:isTitlecase:}</code> | The set of all characters X such that $\text{toTitlecase}(X) = X$ |
| <code>{:isCaseFolded:}</code> | The set of all characters X such that $\text{toCasefold}(X) = X$ |
| <code>{:isCased:}</code> | The set of all cased characters. |

3 Tailored Support: Level 3

This section has been retracted. It last appeared in [version 19](#).

Review Note: Level 3 was originally included to consider some features that were attempted in certain implementations (such as including characters in ranges based on locale-specific ordering), and some more speculative possible features. However, at this point it does not appear that there is any value to retaining this section. Feedback is welcome as to whether it should be retained or removed. If removed, then other changes will need to be made where Level 3 is referenced, such as in the Introduction and Conformance sections.

All of the above deals with a default specification for a regular expression. However, a regular expression engine also may want to support tailored specifications, typically tailored for a particular language or locale. This may be important when the regular expression engine is being used by end users instead of programmers, such as in a word processor allowing some level of regular expressions in searching.

For example, the order of Unicode characters may differ substantially from the order expected by users of a particular language. The regular expression engine has to decide, for example, whether the list `[a-ä]` means:

- the Unicode characters in binary order between 0061_{16} and 00E5_{16} (including 'z', 'z', 'l', and 'x'), or
- the letters in that order in the users' locale (which *does not* include 'z' in English, but *does* include it in Swedish).

If both tailored and default regular expressions are supported, then a number of different mechanisms are affected. There are two main alternatives for control of tailored support:

- coarse-grained support:* the whole regular expression (or the whole script in which the regular expression occurs) can be marked as being tailored.
- fine-grained support:* any part of the regular expression can be marked in some way as being tailored.

For example, fine-grained support could use some syntax such as the following to indicate tailoring to a locale within a certain range:

```
\T{<locale_id>}\E
```

Locale (or language) IDs should use the syntax from locale identifier definition in [\[UTS35\]](#), Section 3. *Identifiers*. Note that the locale id of "root" or "und" indicates the root locale, such as in the CLDR root collation.

There must be some sort of syntax that will allow Level 3 support to be turned on and off, for two reasons. Level 3 support may be considerably slower than Level 2, and most regular expressions may require Level 1 or Level 2 matches to work properly. The syntax should also specify the particular locale or other tailoring customization that the pattern was designed for, because tailored regular expression patterns are usually quite specific to the locale, and will generally not work across different locales.

Sections 3.6 and following describe some additional capabilities of regular expression engines that are very useful in a Unicode environment, especially in dealing with the complexities of the large number of writing systems and languages expressible in Unicode.

3.1 Tailored Punctuation

The Unicode character properties for punctuation may vary from language to language or from country to country. In most cases, the effects of such changes will be apparent in other operations, such as a determination of word breaks. But there are other circumstances where the effects should be apparent in the general APIs, such as when testing whether a curly quotation mark is opening or closing punctuation.

RL3.1 Tailored Punctuation

*To meet this requirement, an implementation shall allow for punctuation properties to be tailored according to locale, using the locale identifier definition in [\[UTS35\]](#), Section 3. *Identifiers*.*

As just described, there must be the capability of turning this support on or off.

3.2 Tailored Grapheme Clusters

RL3.2 Tailored Grapheme Clusters

To meet this requirement, an implementation shall provide for collation grapheme clusters matches based on a

locale's collation order.

Tailored grapheme clusters may be somewhat different than the extended grapheme clusters discussed in Level 2. They are coordinated with the collation ordering for a given language in the following way. A collation ordering determines a *collation grapheme cluster*, which is a sequence of characters that is treated as a unit by the ordering. For example, *ch* is a collation grapheme cluster for a traditional Spanish ordering.

The tailored grapheme clusters for a particular locale are the collation grapheme clusters for the collation ordering for that locale. The determination of tailored grapheme clusters requires the regular expression engine to either draw upon the platform's collation data, or incorporate its own tailored data for each supported locale.

For example, an implementation could interpret `\u(es-u-co-trad)` as matching a collation grapheme cluster for a traditional Spanish ordering, or use a switch to change the meaning of `\X` during some span of the regular expression.

See [Section 6.9, Handling Collation Graphemes](#) in UTS #10, *Unicode Collation Algorithm* [UTS10] for the definition of collation grapheme clusters, and [Annex B: Sample Collation Grapheme Cluster Code](#) for sample code.

3.3 Tailored Word Boundaries

RL3.3 Tailored Word Boundaries

To meet this requirement, an implementation shall allow for the ability to have word boundaries to be tailored according to locale.

For example, an implementation could interpret `\b(x+)` as matching the word break positions according to the locale information in CLDR [UTS35] (which are tailorings of word break positions in [UAX29]). Thus it could interpret expressions as show here:

| Expression | Matches |
|------------------------|--------------------------|
| <code>\b(w+und)</code> | a <i>root</i> word break |
| <code>\b(w)</code> | a Japanese word break |
| <code>\b(w+ja)</code> | a Japanese line break |

Alternatively, it could use a switch to change the meaning of `\b` and `\B` during some span of the regular expression.

Semantic analysis may be required for correct word boundary detection in languages that do not require spaces, such as Thai. This can require fairly sophisticated support if Level 3 word boundary detection is required, and usually requires drawing on platform OS services.

3.4 Tailored Loose Matches (Retracted)

RL3.4 Tailored Loose Matches (Retracted)

Previous versions of RL3.4 described loose matches based on collation order. However, for most full featured regular expression engines, it is quite difficult to match under code point equivalences that are not 1:1. For more discussion of this, see [1.5 Simple Loose Matches](#) and [2.1 Canonical Equivalents](#). Thus RL3.4 has been retracted.

3.5 Tailored Ranges (Retracted)

RL3.5 Tailored Ranges (Retracted)

Previous versions of RL3.5 described ranges based on collation order. However, tailored ranges can be quite difficult to implement properly, and can have very unexpected results in practice. For example, languages may also vary whether they consider lowercase below uppercase or the reverse. This can have some surprising results: `[a-z]` may not match anything if `Z < a` in that locale. Thus RL3.5 has been retracted.

3.6 Context Matching

RL3.6 Context Matching

To meet this requirement, an implementation shall provide for a restrictive match against input text, allowing for context before and after the match.

For parallel, filtered transformations, such as those involved in script transliteration, it is important to restrict the matching of a regular expression to a substring of a given string, and yet allow for context before and after the affected area. Here is a sample API that implements such functionality, where `m` is an extension of a `Regex.Matcher`.

```
if (m.matches(text, contextStart, targetStart, targetLimit, contextLimit)) {
    int end = p.getMatchEnd();
}
```

The range of characters between contextStart and targetStart define a *precontext*; the characters between targetStart and targetLimit define a *target*, and the offsets between targetLimit and contextLimit define a *postcontext*. Thus contextStart ≤ targetStart ≤ targetLimit ≤ contextLimit. The meaning of this function is that:

- a match is attempted beginning at targetStart.
- the match will only succeed with an endpoint at or less than targetLimit.
- any zero-width look-arounds (look-aheads or look-behinds) can match characters inside or outside of the target, but cannot match characters outside of the context.

Examples are shown in the following table. In these examples, the text in the pre- and postcontext is italicized and the target is underlined. In the output column, the text shown with a gray background is the matched portion. The pattern syntax "(← x)" means a backwards match for x (without moving the cursor). This would be (?←x) in Perl. The pattern "(→x)" means a forwards match for x (without moving the cursor). This would be (?→x) in Perl.

| Pattern | Input | Output | Comment |
|---------------|----------------------------------|----------------------------------|---|
| /←a(bc)*←d/ | <i>1</i> abc bc <i>d2</i> | <i>1</i> abc bc <i>d2</i> | matching with context |
| /←a(bc)*←bcd/ | <i>1</i> abc bc <i>d2</i> | <i>1</i> abc bc <i>d2</i> | stops early, because otherwise 'd' would not match |
| /(bc)*d/ | <i>1</i> abc bc <i>d2</i> | <i>no match</i> | 'd' cannot be matched in the target, only in the postcontext |
| /←a(bc)*←d/ | <i>1</i> abc bc <i>d2</i> | <i>no match</i> | 'a' cannot be matched, because it is before the precontext (which is zero-length, in this case) |

While it would be possible to simulate this API call with other regular expression calls, it would require subdividing the string and making multiple regular expression engine calls, significantly affecting performance.

There should also be pattern syntax for matches (like ^ and \$) for the contextStart and contextLimit positions.

Internally, this can be implemented by modifying the regular expression engine so that all matches are limited to characters between contextStart and contextLimit, and so that all matches that are not zero-width look-arounds are limited to the characters between targetStart and targetLimit.

3.7 Incremental Matches

RL3.7 Incremental Matches

To meet this requirement, an implementation shall provide for incremental matching.

For buffered matching, one needs to be able to return whether there is a partial match; that is, whether there *would be* a match if additional characters were added after the targetLimit. This can be done with a separate method having an enumerated return value: *match*, *no_match*, or *partial_match*.

```
if (m.incrementalMatches(text, cs, ts, tl, cl) == Matcher.MATCH) {
    ...
}
```

Thus performing an incremental match of /bebec←d/ against "1abc**bc**d2" would return a *partial_match* because the addition of an e to the end of the target would allow it to match. Note that /(bc)*←d/ would *also* return a partial match, because if bc were added at the end of the target, it would match.

The following table shows the same patterns as shown above in Section 3.6, *Context Matching*, but with the results for when an incremental match method is called:

| Pattern | Input | Output | Comment |
|---------------|----------------------------------|----------------------|--|
| /←a(bc)*←d/ | <i>1</i> abc bc <i>d2</i> | <i>partial match</i> | 'bc' could be inserted |
| /←a(bc)*←bcd/ | <i>1</i> abc bc <i>d2</i> | <i>partial match</i> | 'bc' could be inserted |
| /(bc)*d/ | <i>1</i> abc bc <i>d2</i> | <i>partial match</i> | 'd' could be inserted |
| /←a(bc)*←d/ | <i>1</i> abc bc <i>d2</i> | <i>no match</i> | as with the matches function; the backwards search for 'a' fails |

The typical usage of incremental matching is to make a series of incremental match calls, marching through a buffer with each successful match. At the end, if there is a partial match, one loads another buffer (or waits for other input). When the process terminates (no more buffers or input are available), then a regular match call is made.

Internally, incremental matching can be implemented in the regular expression engine by detecting whether the matching process ever fails when the current position is at or after targetLimit, and setting a flag if so. If the overall match fails, and this flag is set, then the return value is set to *partial_match*. Otherwise, either *match* or *no_match* is returned, as appropriate.

The return value `partial_match` indicates that there was a partial match: if further characters were added there could be a match to the resulting string. It may be useful to divide this return value into two, instead:

- *extendable_match*: in addition to there being a partial match, there was also a match somewhere in the string. For example, when matching `/(ab)*/` against "aba", there is a match, *and* if other characters were added ("a", "aba",...) there could also be another match.
- *only_partial_match*: there was no other match in the string. For example, when matching `/abed/` against "abc", there is only a partial match; there would be no match unless additional characters were added.

3.8 Unicode Set Sharing (Retracted)

Previous versions described a technique to reduce memory consumption by sharing the underlying implementation data structures for character classes. That technique has been retracted because it assumed a very specific implementation environment and did not specify any Unicode related pattern or matching features.

3.9 Possible Match Sets

RL3.9 Possible Match Sets

To meet this requirement, an implementation shall provide for the generation of possible match sets from any regular expression pattern.

There are a number of circumstances where additional functions on regular expression patterns can be useful for performance or analysis of those patterns. These are functions that return information about the sets of characters that a regular expression can match.

When applying a list of regular expressions (with replacements) against a given piece of text, one can do that either serially or in parallel. With a serial application, each regular expression is applied the text, repeatedly from start to end. With parallel application, each position in the text is checked against the entire list, with the first match winning. After the replacement, the next position in the text is checked, and so on.

For such a parallel process to be efficient, one needs to be able to winnow out the regular expressions that simply could not match text starting with a given code point. For that, it is very useful to have a function on a regular expression pattern that returns a set of all the code points that the pattern would partially or fully match.

```
myFirstMatchingSet ← pattern.getFirstMatchSet(Regex.POSSIBLE_FIRST_CODEPOINT);
```

For example, the pattern `/[[\u{0} \u{FF}] && [:\Latin:]] * [0-9]/` would return the set {0..9, A..Z, a..z}. Logically, this is the set of all code points that would be at least partial matches (if considered in isolation).

Note: An additional useful function would be one that returned the set of all code points that could be matched at any point. Thus a code point outside of this set cannot be in any part of a matching range.

The second useful case is the set of all code points that could be matched in any particular group, that is, that could be set in the standard `$0`, `$1`, `$2`, ... variables.

```
myAllMatchingSet ← pattern.getAllMatchSet(Regex.POSSIBLE_IN$0);
```

Internally, this can be implemented by analysing the regular expression (or parts of it) recursively to determine which characters match. For example, the first match set of an alternation `(a|b)` is the union of the first match sets of the terms `a` and `b`.

The set that is returned is only guaranteed to *include* all possible first characters; if an expression gets too complicated it could be a proper superset of all the possible characters.

3.10 Folded Matching (Retracted)

RL3.10 Folded Matching

Previous versions of RL3.10 described tailored folding. However, for most full featured regular expression engines, it is quite difficult to match under folding equivalences that are not 1:1. For more discussion of this, see 1.5 [Simple Loose Matches](#) and 2.1 [Canonical Equivalents](#). Thus RL3.10 has been retracted.

3.11 Submatchers

RL3.11 Submatchers

To meet this requirement, an implementation shall provide for general registration of matching functions for providing matching for general linguistic features.

There are over 70 properties in the Unicode character database, yet there are many other sequences of characters that users may want to match, many of them specific to given languages. For example, characters that are used as vowels may vary by language. This goes beyond single character properties, because certain sequences of characters may need to be matched; such sequences may not be easy themselves to express using regular expressions. Extending the regular expression syntax to provide for registration of arbitrary properties of characters allows these requirements to be handled.

The following provides an example of this. The actual function is just for illustration.

```
class MultipleMatcher implements RegExSubmatcher {
// from RegExFolder, must be overridden in subclasses
/**
 * Returns 1 if there is no match; otherwise returns the endpoint;
 * an offset indicating how far the match got.
 * The endpoint is always between targetStart and targetLimit, inclusive.
 * Note that there may be zero width matches.
 */
int match(String text, int contextStart, int targetStart, int targetLimit, int contextLimit) {
// code for matching numbers according to numeric value.
}

// from RegExFolder, may be overridden for efficiency
/**
 * The parameter is a number. The match will match any numeric value that is a multiple.
 * Example: for "2.3", it will match "0002.3000", "4.6", "11.5", and any non Western
 * script variants, like Indic numbers.
 */
RegExSubmatcher clone(String parameter, Locale locale) {...}
}

...

RegExSubmatcher.registerMatcher("multiple", new MultipleMatcher());

...

p = Pattern.compile("xxx\\M[multiple-2.3]xxx");
```

In this example, the match function can be written to parse numbers according to the conventions of different locales, based on OS functions available for such parsing. If there are mechanisms for setting a locale for a portion of a regular expression, then that locale would be used; otherwise the default locale would be used.

Note: It might be advantageous to make the Submatcher API identical to the Matcher API; that is, only have one base class "Matcher", and have user extensions derive from the base class. The base class itself can allow for nested matchers.

Annex A: Character Blocks

The Block property from the Unicode Character Database can be a useful property for quickly describing a set of Unicode characters. It assigns a name to segments of the Unicode codepoint space; for example, `[\u{370}-\u{3FF}]` is the Greek block.

However, block names need to be used with discretion; they are very easy to misuse because they only supply a very coarse view of the Unicode character allocation. For example:

- **Blocks are not at all exclusive.** There are many mathematical operators that are not in the Mathematical Operators block; there are many currency symbols not in Currency Symbols, and so on.
- **Blocks may include characters not assigned in the current version of Unicode.** This can be both an advantage and disadvantage. Like the General Property, this allows an implementation to handle characters correctly that are not defined at the time the implementation is released. However, it also means that depending on the current properties of assigned characters in a block may fail. For example, all characters in a block may currently be letters, but this may not be true in the future.
- **Writing systems may use characters from multiple blocks:** English uses characters from Basic Latin and General Punctuation, Syriac uses characters from both the Syriac and Arabic blocks, various languages use Cyrillic plus a few letters from Latin, and so on.
- **Characters from a single writing system may be split across multiple blocks.** See the following table on Writing Systems versus Blocks. Moreover, presentation forms for a number of different scripts may be collected in blocks like Alphabetic Presentation Forms or Halfwidth and Fullwidth Forms.

The following table illustrates the mismatch between writing systems and blocks. These are only examples; this table is not a complete analysis. It also does not include common punctuation used with all of these writing systems.

Writing Systems Versus Blocks

| Writing System | Associated Blocks |
|----------------|---|
| Latin | Basic Latin, Latin-1 Supplement, Latin Extended-A, Latin Extended-B, Latin Extended-C, Latin Extended-D, Latin Extended-E, Latin Extended Additional, Combining Diacritical Marks |
| Greek | Greek, Greek Extended, Combining Diacritical Marks |
| Arabic | Arabic, Arabic Supplement, Arabic Extended-A, Arabic Presentation Forms-A, Arabic Presentation Forms-B |
| Korean | Hangul Jamo, Hangul Jamo Extended-A, Hangul Jamo Extended-B, Hangul Compatibility Jamo, Hangul Syllables, CJK Unified Ideographs, CJK Unified Ideographs Extension A, CJK Compatibility Ideographs, CJK Compatibility Forms, Enclosed CJK Letters and Months, Small Form Variants |

| | |
|---------|--|
| Yi | Yi Syllables, Yi Radicals |
| Chinese | CJK Unified Ideographs, CJK Unified Ideographs Extension A, CJK Unified Ideographs Extension B, CJK Unified Ideographs Extension C, CJK Unified Ideographs Extension D, CJK Unified Ideographs Extension E, CJK Compatibility Ideographs, CJK Compatibility Ideographs Supplement, CJK Compatibility Forms, Kangxi Radicals, CJK Radicals Supplement, Enclosed CJK Letters and Months, Small Form Variants, Bopomofo, Bopomofo Extended, ... |

For the above reasons, Script values are generally preferred to Block values. Even there, they should be used in accordance with the guidelines in UAX #24, *Unicode Script Property* [UAX24].

Annex B: Sample Collation Grapheme Cluster Code

This annex was retracted with Level 3.

Review Note: The only purpose of this section was for Level 3.

The following provides sample code for doing Level 3 collation grapheme cluster detection. This code is meant to be illustrative, and has not been optimized. Although written in Java, it could be easily expressed in any programming language that allows access to the Unicode Collation Algorithm mappings.

```

/**
 * Return the end of a collation grapheme cluster.
 * @param s the source string
 * @param start the position in the string to search
 *         forward from
 * @param collator the collator used to produce collation elements.
 * This can either be a custom built one, or produced from
 * the factory method Collator.getInstance(someLocale).
 * @return the end position of the collation grapheme cluster
 */
static int getLocaleCharacterEnd(String s,
    int start, RuleBasedCollator collator) {
    int lastPosition = start;
    CollationElementIterator it
        = collator.getCollationElementIterator(
        s.substring(start, s.length()));
    it.next(); // discard first collation element
    int primary;
    // accumulate characters until we get to a non-zero primary
    do {
        lastPosition = it.getOffset();
        int ce = it.next();
        if (ce == CollationElementIterator.NULLORDER) break;
        primary = CollationElementIterator.primaryOrder(ce);
    } while (primary == 0);
    return lastPosition;
}

```

Annex C: Compatibility Properties

The following table shows recommended assignments for compatibility property names, for use in Regular Expressions. The standard recommendation is shown in the column labeled "Standard"; applications should use this definition wherever possible. If populated with a different value, the column labeled "POSIX Compatible" shows modifications to the standard recommendation required to meet the formal requirements of [POSIX], and also to maintain (as much as possible) compatibility with the POSIX usage in practice. That modification involves some compromises, because POSIX does not have as fine-grained a set of character properties as in the Unicode Standard, and also has some additional constraints. So, for example, POSIX does not allow more than 20 characters to be categorized as digits, whereas there are many more than 20 digit characters in Unicode.

Compatibility Property Names

| Property | Standard | POSIX Compatible | Comments |
|--------------|-----------------------------|------------------|--|
| alpha | <code>\p{Alphabetic}</code> | | Alphabetic includes more than <code>gc = Letter</code> . Note that combining marks (Me, Mn, Mc) are required for words of many languages. While they could be applied to non-alphabets, their principal use is on alphabets. See DerivedCoreProperties for Alphabetic. See also DerivedGeneralCategory . Alphabetic should <i>not</i> be used as an approximation for word boundaries: see word below. |

| | | | |
|-------------------------------------|--|--|---|
| lower | <code>\p{Lowercase}</code> | | Lowercase includes more than <code>gc = Lowercase_Letter (Ll)</code> . See DerivedCoreProperties . |
| upper | <code>\p{Uppercase}</code> | | Uppercase includes more than <code>gc = Uppercase_Letter (Lu)</code> . |
| punct | <code>\p{gc=Punctuation}</code> | <code>\p{gc=Punctuation}</code> <code>\p{gc=Symbol}</code> <code>-- \p{alpha}</code> | POSIX adds symbols. Not recommended generally, due to the confusion of having <i>punct</i> include non-punctuation marks. |
| digit (<code>\d</code>) | <code>\p{gc=Decimal_Number}</code> | <code>[0..9]</code> | Non-decimal numbers (like Roman numerals) are normally excluded. In U4.0+, the recommended column is the same as <code>gc = Decimal_Number (Nd)</code> . See DerivedNumericType . |
| xdigit | <code>\p{gc=Decimal_Number}</code> <code>\p{Hex_Digit}</code> | <code>[0-9 A-F a-f]</code> | <code>Hex_Digit</code> contains 0–9 A–F, fullwidth and halfwidth, upper and lowercase. |
| alnum | <code>\p{alpha}</code> <code>\p{digit}</code> | | Simple combination of other properties |
| space (<code>\s</code>) | <code>\p{Whitespace}</code> | | See PropList for the definition of Whitespace. |
| blank | <code>\p{gc=Space_Separator}</code> <code>\N{CHARACTER TABULATION}</code> | | "horizontal" whitespace: space separators plus U+0009 <i>tab</i> . Engines implementing older versions of the Unicode Standard may need to use the longer formulation: <code>\p{Whitespace} --</code> <code>[\N{LF} \N{VT} \N{FF} \N{CR} \N{NEL} \p{gc=Line_Separator}</code> <code>\p{gc=Paragraph_Separator}]</code> |
| cntrl | <code>\p{gc=Control}</code> | | The characters in <code>\p{gc=Format}</code> share some, but not all aspects of control characters. Many format characters are required in the representation of plain text. |
| graph | <code>[^</code> <code>\p{space}</code> <code>\p{gc=Control}</code> <code>\p{gc=Surrogate}</code> <code>\p{gc=Unassigned}]</code> | | <i>Warning:</i> the set shown here is defined by <i>excluding</i> space, controls, and so on with <code>^</code> . |
| print | <code>\p{graph}</code> <code>\p{blank}</code> <code>-- \p{cntrl}</code> | | Includes graph and space-like characters. |
| word (<code>\w</code>) | <code>\p{alpha}</code> <code>\p{gc=Mark}</code> <code>\p{digit}</code> <code>\p{gc=Connector_Punctuation}</code> <code>\p{Join_Control}</code> | n/a | This is only an approximation to Word Boundaries (see b below). The Connector Punctuation is added in for programming language identifiers, thus adding "_" and similar characters. |
| <code>\X</code> | Extended Grapheme Clusters | n/a | See [UAX29] . Other functions are used for programming language identifier boundaries. |
| <code>\b</code> | Default Word Boundaries | n/a | If there is a requirement that <code>\b</code> align with <code>\w</code> , then it would use the approximation above instead. See [UAX29] . Note that different functions are used for programming language identifier boundaries. See also [UAX31] . |

Annex D: Data Files

There is one [draft](#) data file associated with this specification: `RegexPropertyAliases.txt`. It provides property metadata that can be used to support property syntax. That information is used to match and validate properties and property values for syntax such as `\p{pname=pvalue}`, so that they can be used in the same way as UCD properties.

The format of the file and meaning of the fields is described in the header.

Review Note: The name, location, format, and contents of the file are subject to review/feedback/change.

The draft data file is available for review at www.unicode.org/review/pr404/RegexPropertyAliases.txt.

Annex E: Resolving Character Classes with Strings

Matching the negation of set of strings is problematic, and needs to be disallowed in regex expressions [see [why-ban-the-use-of-these-properties-within-character-classes](#)]. However, the negation of a Character Class with strings or of a property of strings is allowed *internal* to a boolean expression as long as the fully resolved version of that expression is not a negation Character Class with strings.

For example, suppose that C is a Character Class without strings or property of characters, and S is a Character Class with strings or property of strings.

- `[^[S]]` is allowable
- `[C--S]` is allowable
- `[C&&[S]]` resolves to `[C--S]` and is thus allowable — it does not contain any strings.
- `[^C--S]` is not valid

A narrowed set of single characters can always be represented by intersecting with the set of single characters, such as `[\p{Basic_Emoji}-\p{any}]`.

The following describes how a boolean expression can be resolved to a Character Class with characters, a Character Class with strings, or a negated Character Class with strings.

When incrementally parsing and building a resolved boolean expression, the process can be analyzed in terms of a series of core operations. In parsing Character Classes, the intermediate objects are logically *enhanced sets* of strings, such as A and B. The enhancement is the addition of a boolean to indicate whether the set is negated or not. The operations here are expressed as operations on A and B, where A_p is the flag: true for the normal case, and false if negative, and A_s is the normal set of strings. The operations on the strings in the set depend on the two flags A_p and B_p .

Handle each operation as follows:

| Operation | A_p | B_p | Action | Comments |
|------------------------------|-------|-------|--|---|
| <code>[[^]A]</code> | — | n/a | $A_p = !A_p$ | |
| <code>\p{...}</code> | — | n/a | $A_s = R$ $A_p = true$ | R = set from resolved expression |
| <code>\P{...}</code> | — | n/a | $A_s = R$ $A_p = false$ | |
| <code>A B</code> | T | T | $A_s = A_s \cup B_s$ | Standard operation, both being positive sets. |
| | F | F | $A_s = A_s \cap B_s$ | |
| | T | F | $A_s = B_s \setminus A_s$ $A_p = false$ | |
| | F | T | $A_s = A_s \setminus B_s$ | |
| <code>A && B</code> | T | T | $A_s = A_s \cap B_s$ | Standard operation, both being positive sets. |
| | F | F | $A_s = A_s \cup B_s$ | |
| | T | F | $A_s = A_s \setminus B_s$ | |
| | F | T | $A_s = B_s \setminus A_s$ $A_p = true$ | |
| <code>A -- B</code> | T | T | $A_s = A_s \setminus B_s$ | Standard operation, both being positive sets. |
| | F | F | $A_s = B_s \setminus A_s$ $A_p = true$ | |
| | T | F | $A_s = A_s \cap B_s$ | |
| | F | T | $A_s = A_s \cup B_s$ | |

Where the result of recursively resolving the expression is an enhanced set A, and $A_p == false$, the expression is invalid.

As usual, this is a logical expression of the process; implementations can optimize as long as they get the same results.

Review Note: The above needs careful review. Comments welcome.

References

- [Case] Section 3.13, *Default Case Algorithms* in [Unicode]
- [CaseData] <http://www.unicode.org/Public/UCD/latest/ucd/CaseFolding.txt>

- [Friedl] Jeffrey Friedl, "Mastering Regular Expressions", 2nd Edition 2002, O'Reilly and Associates, ISBN 0-596-00289-0
- [Glossary] Unicode Glossary
<http://www.unicode.org/glossary/>
For explanations of terminology used in this and other documents.
- [Perl] <http://perldoc.perl.org/>
See especially:
<http://perldoc.perl.org/charnames.html>
<http://perldoc.perl.org/perlre.html>
<http://perldoc.perl.org/perluniintro.html>
<http://perldoc.perl.org/perlunicode.html>
- [POSIX] The Open Group Base Specifications Issue 6, IEEE Std 1003.1, 2004 Edition, "Locale" chapter
http://www.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap07.html
- [Prop] <http://www.unicode.org/Public/UCD/latest/ucd/PropertyAliases.txt>
- [PropValue] <http://www.unicode.org/Public/UCD/latest/ucd/PropertyValueAliases.txt>
- [ScriptData] <http://www.unicode.org/Public/UCD/latest/ucd/Scripts.txt>
- [SpecialCasing] <http://www.unicode.org/Public/UCD/latest/ucd/SpecialCasing.txt>
- [UAX14] UAX #14, *Unicode Line Breaking Algorithm*
<http://www.unicode.org/reports/tr14/>
- [UAX15] UAX #15, *Unicode Normalization Forms*
<http://www.unicode.org/reports/tr15/>
- [UAX24] UAX #24, *Unicode Script Property*
<http://www.unicode.org/reports/tr24/>
- [UAX29] UAX #29, *Unicode Text Segmentation*
<http://www.unicode.org/reports/tr29/>
- [UAX31] UAX #31, *Unicode Identifier and Pattern Syntax*
<http://www.unicode.org/reports/tr31/>
- [UAX38] UAX #38, *Unicode Han Database (Unihan)*
<http://www.unicode.org/reports/tr38/>
- [UAX44] UAX #44, *Unicode Character Database*
<http://www.unicode.org/reports/tr44/>
- [UData] <http://www.unicode.org/Public/UCD/latest/ucd/UnicodeData.txt>
- [Unicode] The Unicode Standard
For the latest version, see:
<http://www.unicode.org/versions/latest/>
- [UTR50] UTR #50, Unicode Vertical Text Layout
<http://www.unicode.org/reports/tr50/>
- [UTR51] UTR #51, Unicode Emoji
<http://www.unicode.org/reports/tr51/>
- [UTS10] UTS #10, *Unicode Collation Algorithm (UCA)*
<http://www.unicode.org/reports/tr10/>
- [UTS35] UTS #35, *Unicode Locale Data Markup Language (LDML)*
<http://www.unicode.org/reports/tr35/>
- [UTS39] UTS #39, Unicode Security Mechanisms
<http://www.unicode.org/reports/tr39/>
- [UTS46] UTS #46, Unicode IDNA Compatibility Processing
<http://www.unicode.org/reports/tr46/>

Acknowledgments

Mark Davis created the initial version of this annex and maintains the text, with significant contributions from Andy Heneringer.

Thanks to Julie Allen, [Mathias Bynens](#), Tom Christiansen, [David Corbett](#), Michael D'Errico, Asmus Freytag, Jeffrey Friedl, Norbert Lindenberg, Peter Linsley, Alan Liu, Kent Karlsson, Jarkko Hietaniemi, Gurusamy Sarathy, [Markus Scherer](#), Xueming Shen, Henry Spencer, Kento Tamura, Philippe Verdy, Tom Watson, Ken Whistler, Karl Williamson, and [Richard Wordingham](#) for their feedback on the document.

Modifications

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

Revision 20

Summary:

Broadened the scope of properties for regex to allow for properties of strings as well as properties of characters. In doing so, cleaned up the discussion of Character Classes with strings. Added a new annex to clarify exactly when negations of properties of strings and/or Character Classes with strings would invalidate a regular expression.

Updated the full property list to include newer UCD properties plus Emoji properties and UTS #39 properties, and added a data file with property metadata for supporting non-UCD properties. Removed some sections that are no longer useful, and added a number of clarifications.

- General
 - The term "character range" has been replaced by Character Class, without highlighting each instance. Character Class is more widely used (and less misleading).
- Section 1.2 **Properties**
 - Broadened the scope of properties to allow for properties of strings.
 - Moved in material that used to be split across different sections below about properties of strings and property values that are sets. For example, the behavior of `Script_Extensions` was moved up.
 - Added pointer to **Annex E: Resolving Character Classes with Strings** for handling negated properties of strings.
 - Added review note requesting feedback on the syntax issue: `lp` vs `vm`.
- Section 1.2.3 **Other Properties**
 - Moved down to 2.8 (shouldn't have been discussed at Level 1), and revised to make the optional nature clear.
- Section 1.3 **Subtraction and Intersection**
 - Described how subtraction and intersection behave with properties of strings.
- Section 1.7 **Code Points**
 - Added clarifications about code points vs code units.
- Section 2.2 **Extended Grapheme Clusters and Character Classes with Strings**
 - For clarity
 - Renamed section to reflect contents
 - Reworded *Character Classes containing literal clusters* to *Character Classes with strings*.
 - Added new subsection 2.2.1 **Character Classes with Strings**.
 - Noted that the boundary definitions in CLDR are more comprehensive.
- Section 2.2.1 **Grapheme Cluster Mode**
 - Removed subsection that should have been in Level 3.
 - Fixed lack of whitespace in the syntax for ITEM.
 - Also allowed empty strings `\q{}` for implementations that allow empty alternatives.
 - Moved the section on validation to **Annex E: Resolving Character Classes with Strings**, and revised almost completely.
- Section 2.5.1 **Individually Named Characters**
 - This section implied that `\N` was a property of strings. Modified that to clearly be a property of code points, and moved discussion of `Named_Sequences` to Section 2.8 **Optional Properties**.
- Section 2.7 **Full Properties**
 - Updated for newer versions of Unicode.
 - Added Emoji properties
 - Added review note requesting feedback on whether any properties from Section 2.8 **Optional Properties** should be moved here.
- Section 2.8 **Optional Properties**
 - Moved from Section 1.2.3 **Other Properties**
 - Made it clear that these are entirely optional.
 - Removed some properties now covered in Section 2.7 **Full Properties**
 - Changed certain cases be properties of strings.
 - Removed examples of `toNFC_CaseFold`, `CJK`, and other properties
 - Added `Named sequences` (was previously included as part of Section 2.5.1 **Individually Named Characters**).
 - Clarified Exemplar properties and others via examples.

- Section 3 **Tailored Support: Level 3**
 - Entire section removed, with a review note requesting feedback on whether we should remove or retain.
 - The most prominent text elsewhere where changes would need to be made is also marked with review notes.
- Removed **Annex B: Sample Collation Grapheme Cluster Code** along with Level 3.
- Added **Annex D: Data Files**
- Added **Annex E: Resolving Character Classes with Strings**

Modifications for previous versions are listed in those respective versions.

Copyright © 2019 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.