

Update Annex E based on Unicode 15.0 UAX 31

Steve Downey (sdowney@gmail.com)

Document #: P2653R1
Date: 2022-10-20
Project: Programming Language C++
Audience: SG16, CWG

Abstract

Update Annex E, Conformance with UAX #31, based on the updated guidance from Unicode 15. In particular update the pattern whitespace and syntax section.

Contents

1 What is being proposed	1
2 Wording	1
References	2

1 What is being proposed

The Unicode Consortium in the recent release of the Unicode Standard, version 15, clarified that the Pattern White Space and Pattern Syntax requirements are intended to apply to programming languages. We had interpreted the requirements to apply to facilities like regexp pattern matching. See the Modifications section of [UAX31-15:online] where it says:

Section 4, Pattern Syntax
Clarified that this section is applicable to programming languages.

The diff from the prior version of the Pattern Syntax section [UAX31-DIFF:online]:

Most programming languages have a concept of whitespace as part of their lexical structure, as well as some set of characters that are disallowed in identifiers but have syntactic use, such as arithmetic operators.
~~There are~~ Beyond general programming languages, there are also many circumstances [...]

While it appears possible to achieve conformance with this requirement by defining an appropriate profile showing how C++ classifies white space and the characters it uses for syntax, it is much simpler at this time to continue to disclaim conformance, while updating our reasoning for the disclamation.

2 Wording

The proposed changes are relative to the current working draft [N4917].

◆.4 R3 Pattern_White_Space and Pattern_Syntax characters

[uaxid.pattern]

- ¹ UAX #31 describes how languages that use or interpret patterns of characters, such as regular expressions or number formats, may describe that syntax with Unicode properties.

UAX #31 describes how formal languages such as computer languages should describe and implement their use of whitespace and syntactically significant characters during the processes of lexing and parsing.

- ² C++ does not do this as part of the language, deferring to library components for such usage of patterns. This requirement does not apply to C++.

C++ does not claim conformance with this requirement.

References

- [N4917] Thomas Köppe. N4917: Working draft, standard for programming language c++. <https://wg21.link/n4917>, 9 2022.
- [UAX31-14:online] Mark Davis and Robin Leroy. Unicode identifier and pattern syntax. <https://www.unicode.org/reports/tr31/tr31-35.html>, 08 2021. (Accessed on 10/20/2022).
- [UAX31-15:online] Mark Davis and Robin Leroy. Unicode identifier and pattern syntax. <https://www.unicode.org/reports/tr31/tr31-37.html>, 08 2022. (Accessed on 10/20/2022).
- [UAX31-DIFF:online] Mark Davis and Robin Leroy. Unicode identifier and pattern syntax. <https://www.unicode.org/reports/tr31/tr31-36.html>, 08 2022. (Accessed on 10/20/2022).