Slaying Some Earthly Demons - remove UB 28, 29

Document: n3565

Author: Glenn COATES

Date: 2025-05-10

Undefined Behavior:
(28) A universal character name in an identifier does not designate a character whose encoding falls into one of the specified ranges.
(29) The initial character of an identifier is a universal character name designating a digit.

Analysis:
§6.4.3.1 states: "Each character and universal character name in an identifier shall designate a character whose encoding in ISO/IEC 10646 has the XID_Continue property. The initial character (which can be a universal character name) shall designate a character whose encoding in ISO/IEC 10646 has the XID_Start property. An identifier shall conform to Normalization Form C as specified in ISO/IEC 10646. Annex D provides an overview of the conforming identifiers."

However, this provision appears under the section labeled "Semantics." Meanwhile, §4 of the standard states: "If a 'shall' or 'shall not' requirement that appears outside of a constraint or runtime-constraint is violated, the behavior is undefined."

Recommendation:
The above provision should be relocated under a "Constraints" section.
This change eliminates the following entries from Annex J.2, as the behaviors they describe become diagnosable constraint violations rather than undefined behavior:
(28) A universal character name in an identifier does not designate a character whose encoding falls into one of the specified ranges.
(29) The initial character of an identifier is a universal character name designating a digit.

If this proposal is accepted, the above entries should be removed from Annex J.2.

Suggested Rewording:
…
6.4.3 Identifiers
6.4.3.1 General
Syntax
identifier:

        identifier-start
        identifier identifier-continue

identifier-start:

        nondigit
        XID_Start character
        universal character name of class XID_Start

identifier-continue:

        digit
        nondigit
        XID_Continue character
        universal character name of class XID_Continue

nondigit: one of

        _ a b c d e f g h i j k l m
        n o p q r s t u v w x y z
        A B C D E F G H I J K L M
        N O P Q R S T U V W X Y Z

digit: one of

        0 1 2 3 4 5 6 7 8 9

Constraints

Each character and universal character name in an identifier shall designate a character whose encoding in ISO/IEC 10646 has the XID_Continue property. The initial character (which can be a universal character name) shall designate a character whose encoding in ISO/IEC 10646 has the XID_Start property. An identifier shall conform to Normalization Form C as specified in ISO/IEC 10646. Annex D provides an overview of the conforming identifiers.

Semantics

An XID_Start character is an implementation-defined character whose corresponding code point in ISO/IEC 10646 has the XID_Start property. An XID_Continue character is an implementation-defined character whose corresponding code point in ISO/IEC 10646 has the XID_Continue property. An identifier is a sequence of one identifier start character followed by 0 or more identifier continue characters, which designates one or more entities as described in 6.2.1. It is implementation-defined if a $ (U+0024, DOLLAR SIGN) may be used as a nondigit character. Lowercase and uppercase letters are distinct. There is no specific limit on the maximum length of an identifier.

The character classes XID_Start and XID_Continue are Derived Core Properties as described by UAX #44.[64) Each character and universal character name in an identifier shall designate a character whose encoding in ISO/IEC 10646 has the XID_Continue property. The initial character (which can be a universal character name) shall designate a character whose encoding in ISO/IEC 10646 has the XID_Start property. An identifier shall conform to Normalization Form C as specified in ISO/IEC 10646. Annex D provides an overview of the conforming identifiers.

…