Proposal for C23
WG14 N 3111

Title:	Six versus eight-digit short identifiers v2
Author, affiliation:	Robert C. Seacord, Woven Planet
Date:	2023-2-13
Proposal category:	Defect
Target audience:	Implementers
Abstract:	Six versus eight-digit short identifiers for universal character names
Prior art:	C

Six versus eight-digit short identifiers v2

Reply-to: Robert C. Seacord (rcseacord@gmail.com)
Document No: **N 3111**
Reference Document: **N 3019**
Date: 2023-9-2

Change Log

2023-2-13:

- Initial version

2023-2-19:

- Changed “hexadecimal number” to “hexadecimal value.”
- Changed “0000FFFF” to “00FFFF” and change “00010000” to “010000.”

1.0 Introduction and Rationale

NB comment GB-012 from [n3019] identifies the issue that the 2011 edition of ISO/IEC 10646 removed eight-digit short identifiers that were present in the 2003 edition (and this removal still applies as of the 2020 edition) but the current C23 draft supports eight-digit short identifiers but not six-digit short identifiers.

SC 22 N 5777, Subclause 6.4.3, “Universal character names” paragraph 4 states that:

> *The universal character name \Unnnnnnnn designates the character whose eight-digit short identifier (as specified by ISO/IEC 10646) is nnnnnnnn.80) Similarly, the universal character name \unnnn designates the character whose four-digit short identifier is nnnn (and whose eight-digit short identifier is 0000nnnn).*

Ideally, the C standard would only use short identifiers with no more than six digits. However, this would break backwards compatibility.

2.0 Proposed Solution

This solution is largely editorial. It does not change any syntax or semantics, but simply eliminates or corrects any incorrect references to short identifiers.

3.0 Wording

Replace **Subclause** 6.4.3, “Universal character names”, paragraph 4:

The universal character name \Unnnnnnnn designates the character whose eight-digit short identifier (as specified by ISO/IEC 10646) is nnnnnnnn. Similarly, the universal character name \unnnn designates the character whose four-digit short identifier is nnnn (and whose eight-digit short identifier is 0000nnnn).

with

A universal character name designates the character in ISO/IEC 10646 whose code point is the hexadecimal value represented by the sequence of hexadecimal digits in the universal character

name.

[Editor's note: Remove footnote 80]

In Subclause 5.2.4.1 paragraph 1, change "0000FFFF" to "00FFFF" and change "00010000" to "010000".

— 31 significant initial characters in an external identifier (each universal character name specifying a short identifier of 0000FFFF or less is considered 6 characters, each universal character name specifying a short identifier of 00010000 or more is considered 10 characters, and each extended source character is considered the same number of characters as the corresponding universal character name, if any)[18)]

4.0 Acknowledgements

I would like to recognize the following people for their help with this work: Corentin Jabot, Aaron Ballman, Steve Downey, Peter Bindels, Jens Gustedt, and Joseph Myers.

5.0 References

[n2785] Corentin Jabot, Aaron Ballman. Delimited escapes sequences. https://www.open-std.org/jtc1/sc22/WG14/www/docs/n2785.pdf

[n3019] Keaton, David. CD1 9899 ballot comments with progress from first week of ballot resolution.

[P2071R0] Tom Honermann and Peter Bindels. P2071R0: Named universal character escapes. https://wg21.link/p2071r0, 1 2020.

[P2290R3] Corentin Jabot. P2290R3: Delimited escape sequences. https://wg21.link/p2290r1, 6 2021.