
Information technology — Programming languages, their environments and system software interfaces — Extensions for the programming language C to support new character data types

Technologies de l'information — Langages de programmation, leurs environnements et interfaces de logiciel système — Extensions pour que le langage de programmation C supporte des types de données de caractères nouveaux

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

Withdrawn

© ISO/IEC 2004

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword.....	iv
Introduction.....	v
1 Scope.....	1
2 Normative references.....	1
3 The new typedefs	2
4 Encoding	3
5 String literals and character constants.....	4
5.1 String literals and character constants notations.....	4
5.2 The string concatenation.....	4
6 Library functions.....	5
6.1 The mbrtoc16 function.....	5
6.2 The c16rtomb function	6
6.3 The mbrtoc32 function.....	7
6.4 The c32rtomb function	8
7 ANNEX A Unicode encoding forms: UTF-16, UTF-32	9

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, the joint technical committee may propose the publication of a Technical Report of one of the following types:

- type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;
- type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;
- type 3, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 19769, which is a Technical Report of type 2, was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 22, *Programming languages, their environments and system software interfaces*.

Introduction

The C language has evolved over the last decades, various code pages and multibyte libraries have been introduced, and extended character set support has been introduced; however, the support for extended character data types in the C language is still limited. Today, the introduction and the success of the Unicode/ISO10646 standard and of its implementation in modern computer languages create ever increasing demands on the C language to give Unicode/ISO10646 better support. This paper addresses the introduction of new extended character data types in the C language in order to support future character encoding forms, including Unicode/ISO10646.

The Unicode standard supports 3 encoding forms:

- UTF-8
- UTF-16
- UTF-32

Each encoding form has advantages and disadvantages, so the choice of the encoding form should be left to the application. Currently, some C applications implement UTF-8 using `char` type, UTF-16 using `unsigned short` or `wchar_t`, and UTF-32 using `unsigned long` or `wchar_t`. The current situation, however, faces the following major problems:

- The size of `wchar_t` is implementation defined. While `wchar_t` offers a form of platform portability for C applications, Unicode offers the possibility to write platform independent applications using a platform independent data format.
- There is no string literal for 16- or 32-bit based integer types, but the Unicode encoding forms require string literals.

It is sensible to give all the Unicode encoding forms appropriate data type support. UTF-8 is normally considered as the preferred multibyte encoding, for sequences of one or more elements of type `char`. This paper suggests the implementation of 16 and 32 bit character data types: `char16_t` and `char32_t`. The new data types guarantee program portability through clearly defined character widths. The encoding of the new data types should be as generic as possible in order to support not only Unicode but also other character encodings.

It is generally desirable that C applications process entire strings at once rather than process individual characters in isolation. This paper does not specify the detail of library functions for the new data types, except one set of character conversion functions.

Information technology — Programming languages, their environments and system software interfaces — Extensions for the programming language C to support new character data types

1 Scope

This Technical Report specifies two extended character data types as an extension to the programming language C, specified by the international standard ISO/IEC 9899:1999.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 9899:1999, *Programming Languages – C*

ISO/IEC 10646-1:2000, *Information technology – Universal multiple-octet coded character set (UCS) – Part 1: Architecture and Basic Multilingual Plane*