Annex A
Recommended extended repertoire for user-defined identifiers

The recommended extended repertoire consists of those characters which collectively can be used to generate word-like identifiers for most natural languages of the world. This list comprises the letters (combining or not), syllables, and ideographs from ISO/IEC 10646-1, together with the modifier letters and marks conventionally used as parts of words. The list excludes punctuation and symbols not generally included in words or considered appropriate for use in identifiers. Also excluded are most presentation forms of letters and a number of compatibility characters. The inclusion of combining characters corresponds to those allowed under a level 2 implementation of ISO/IEC 10646-1. These are the minimum required to do a reasonable job of representing word-like identifiers in Hebrew, Arabic, and scripts of South and Southeast Asia, which make general use of combining marks. However, combining marks for level 3 implementations of ISO/IEC 10646-1 are not included in the list, so as to avoid the problem of alternative representations of identifiers.

Attention is drawn to the fact that using the extended repertoire for identifiers may impact source code portability, since the presence of these characters in program text may not be supported on systems that implement less than the full repertoire of ISO/IEC 10646-1.

The character repertoire listed in this annex is based on ISO/IEC 10646-1:2000. It is subject to expansion in the future, to track future amendments to the standard. Characters currently listed in this Annex will not be removed from the recommended extended repertoire in future revisions. However, the use of some characters may be discouraged.

The character repertoire listed in this annex should be conceived of as a recommendation for the minimum extended repertoire for use in user-defined identifiers. Each programming language standard or implementation of the standard can extend the repertoire at the adaptation, in accordance with established practice of identifier usage for the language and any additional user requirements that may be present. For example, the C language should allow U003F LOW LINE in addition to the character repertoire listed below; COBOL should allow U002D HYPHEN-MINUS as well; Java allows a rather large extension to support a level 3 implementation of 10646-1. Some programming language standards may allow half- or full-width compatibility characters from ISO/IEC 10646-1, and some of the standards, e.g. COBOL, may recognize these characters in a width-insensitive manner.

Programming language standards generally have restrictions on what characters may be allowed as the first character of an identifier. For example, digits are often constrained from appearing as the first character of an identifier. To assist in their identification, the decimal digits in ISO/IEC 10646-1 are separately noted in the list below. In addition, combining characters should not appear as the first character of an identifier. To maximize the chances of interoperability between programming languages (as for example, when linking compiled objects between languages), programming language standards and their implementations should follow these restrictions when making use of the extended repertoire for user-defined identifiers.

The characters, recommended for identifiers in programming languages consist of the following character ranges of ISO/IEC 10646-1. Combining characters for scripts are separated out.

The table shows
- the first and the last code point in hexadecimal form for a range of characters
- the property of these characters (see legend below)
- the number of characters in this range between square brackets
- and the names of the first and the last character in the range.

The following table will also be available in electronic form on the ITTF secure web site for downloading. Its file name is TR10176-4-table.txt, the URL is http://xxxxxxxxxxxxxx/TR10176-4-table.txt

Legend:

The following table identifies the property of characters suitable for identifiers, as used in the Unicode Character Database and in the table TR10176-4-table.txt on the ITTF web site.

Abbr.	Description
L&	The symbol "L&" indicates characters of type Lu, Ll, or Lt (see below).
Lu	Letter, Uppercase
Ll	Letter, Lowercase
Lt	Letter, Titlecase
Lm	Letter, Modifier
Lo	Letter, Other
Mn	Mark, Non-Spacing
Mc	Mark, Spacing Combining
Nd	Number, Decimal Digit
Nl	Number, Letter
Pc	Punctuation, Connector