INTERNATIONAL STANDARD

ISO/IEC 14651

Third edition
2011-08-15

AMENDMENT 2
2015-04-01

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 2

Technologies de l'information — Classement international et comparaison de chaînes de caractères — Méthode de comparaison de chaînes de caractères et description du modèle commun et adaptable d'ordre de classement

AMENDEMENT 2

Reference number
ISO/IEC 14651:2011/Amd.2:2015(E)

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

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT), see the following URL: Foreword — Supplementary information.

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, SC 2, *Coded character sets.*

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 2

Page 1, Clause 1

Replace the second bullet with the following:

A Common Template Table. A given tailoring of the Common Template Table is used by the reference comparison method. The Common Template Table describes an order for the characters corresponding to the Unicode 7.0 repertoire, a substantial subset of characters encoded in ISO/IEC 10646:2014. It allows for a specification of a fully deterministic ordering. This table enables the specification of a string ordering adapted to local ordering rules, without requiring an implementer to have knowledge of all the different scripts already encoded in the Universal Coded Character Set (UCS).

Page 2, Clause 3

Replace the normative reference with the following:

ISO/IEC 10646:2014, *Information technology — Universal Coded Character Set (UCS)*

Page 16, 6.5

Replace 6.5 with the following:

6.5 Name of the Common Template Table and name declaration

The name ISO14651_2014_TABLE1 shall be used whenever the Common Template Table is referred to externally as a base point in a given context, whether in a process, contract, or procurement requirement. If another name is used due to practical constraints, a declaration of conformance shall indicate the correspondence between this other name and the name ISO14651_2014_TABLE1.

The use of a defined name is necessary to manage the different stages of development of this table. This follows from the nature of the reference character repertoire, for which development will be ongoing for a number of years or even decades.

Page 17, Annex A

Replace Annex A with the following.

Annex A
(normative)
Common Template Table

In order to minimize formatting problems and the risk of errors in reproduction, the common template table is provided separately in a machine-readable file as a normative component of this International Standard. The file name for this language version is different from the normative reference name specified in 6.5 of this International Standard due to the existence of file versions commented in other natural languages. The file for this language version can also be retrieved on the ITTF web site at the following URL:

http://www.iso.org/ittf/ISO14651_2014_TABLE1_en.txt

There is an official French version of the file which only differs in its comments (its technical content is identical), and its name is: ISO14651_2014_TABLE1_fr.txt

NOTE 1 This International Standard deprecates, but does not preclude specific reference to, the previous tables, which contained and still contain ordering information applicable to the repertoires of previous versions of ISO/IEC 10646 and their amendments. The previous tables can be found at the following URLs:

[ordering information on the repertoire of characters as defined in ISO/IEC 10646-1:1993 including Amendments 1-9] http://www.iso.org/ittf/ISO14651_2000_TABLE1.htm

[ordering information on the combined repertoire of characters of ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001] http://www.iso.org/ittf/ISO14651_2002_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2003] http://www.iso.org/ittf/ISO14651_2003_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2003 including Amendments 1-2] http://www.iso.org/ittf/ISO14651_2006_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2003 including Amendments 1-4] http://www.iso.org/ittf/ISO14651_2008_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2011] http://www.iso.org/ittf/ISO14651_2010_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2012] http://www.iso.org/ittf/ISO14651_2012_TABLE1_en.txt

The current Common Template Table reflects the repertoire of characters of Unicode 7.0, included in ISO/IEC 10646:2014.

NOTE 2 The repertoire targeted by this International Standard is equivalent to the repertoire of *The Unicode Standard Version 7.0, published by The Unicode Consortium*. This repertoire is a substantial subset of ISO/IEC 10646:2014, and not its full set, due to synchronization issues with the two documents.

When ordering data applicable to other amendments of ISO/IEC 10646 becomes available, this International Standard and specifically its Common Template Table will be amended accordingly to cover the ordering of the additional characters and scripts. To meet cultural requirements of specific communities, delta declarations will have to be applied to the amended table as defined in this International Standard.

ISO_14651_2014_TABLE1 is the name that is used for referring to this table in this version of this International Standard.

Page 44, after Annex D

Insert a new annex, Annex E.

Annex E
(informative)
Searching and fuzzy matches

At the fundamental level of comparison of a string with a target corpus, simple textual information retrieval widely uses, implicitly or not, the same concepts and the same conventions as string collation.

It is indeed often useful to be able to search without regard to case, diacritical marks or even special characters ("ignorable" characters). Thus, as a case in point using French, a search for keyword "contremaitre" could equally retrieve as legitimate targets the words "contremaître" or "contre-maître" (the latter being considered an old spelling of the same word), or "contremaitre" (without circumflex accent on the "i", a new spelling now admitted in this particular unambiguous case). It could as well retrieve "CONTREMAITRE", as some uses of the language do not put accents on capitals and thus may very well be found in a text corpus.

To ensure adequacy of a match – fuzzy or exact – with user expectations, it is appropriate to consider that the comparison functions used be based on the same comparison method as the one used for sorting. However, the concepts of “fuzziness” used in information retrieval extend far beyond the simple concepts of case-, accent- or special-character-independence. There are particular search functions on prefixes, taking into account word morphology (word search based on their roots), phonetics, or even equivalences in a foreign language. Complex fuzzy search – a requirement that goes far beyond the scope of this International Standard – could, for example, regard as equivalent two strings according to their broad phonetic value, such as “cinq sens” and “Saint-Saëns” in French (pronounced exactly the same). Similarly, some could want to consider as equivalent inflected grammatical forms such as “œil” and “yeux” in French (the latter simply being the plural of the former word). An example in English would be the words “man” and “men”.

Given the wide variety of fuzziness concepts, we will only discuss here the simple fuzziness cases based on the different levels of weighting used in this International Standard.

If one follows the scope of this International Standard, the search operation should use the same comparison method, separating base characters, diacritical marks, case and other characters (special characters, typographical marks and symbols). However, some levels may be omitted in the comparison, depending on the level of precision necessary to consider a match.

This means that, as cases in point:

— A search that is independent of diacritical marks should consider as equivalent (marked “**<>**” in examples) all strings that only differ at level 2 in the comparison process

Example : contremaître **<>** contremaitre

— A search that is independent of case should consider as equivalent all strings that only differ at level 3 in the comparison process

Example : contremaître **<>** CONTREMAÎTRE

— A search that is independent of special (“ignorable”) characters should consider as equivalent all strings that only differ at level 4 in the comparison process.

Example : contremaître **<>** contre-maître

— A search that intends to retrieve all loosely-similar variants should consider as equivalent all words which only match at level 1.

Example :

contremaître **<>** contre-maitre **<>** contremaitre **<>** CONTREMAÎTRE **<>** Contre-maître

[etc.]

The latter case is to be considered typical in simple searches and should be the preferred way in absence of specialized needs.

It is recommended that the user be made aware of the search parameters using appropriate user interface elements.

NOTE For discussions about this subject, see:

—Unicode Technical Standard #10 Unicode Collation Algorithm, subclause 8.2 (publication in English only);

—Standard sur le tri alphabétique et la recherche de chaînes de caractères, in particular Clauses 4 and 6 (publication in French only).

ICS 35.040
Price based on 3 pages