INTERNATIONAL STANDARD

ISO/IEC 14651

Sixth edition
2020-12

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

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

Reference number
ISO/IEC 14651:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **2**
3 Terms and definitions ... **2**
4 Symbols and conventions ... **3**
5 Conformance ... **3**
6 String comparison ... **4**
6.1 Preparation of character strings prior to comparison ... 4
6.2 Key building and comparison ... 5
6.2.1 Preliminary considerations ... 5
6.2.2 Reference ordering key formation ... 6
6.2.3 Reference comparison method for ordering character strings ... 8
6.2.4 Key ordering definition ... 9
6.3 Common Template Table: Formation and interpretation ... 10
6.3.1 General ... 10
6.3.2 BNF syntax rules for the Common Template Table in Annex A ... 10
6.3.3 Well-formedness conditions ... 12
6.3.4 Interpretation of tailored tables ... 13
6.3.5 Evaluation of weight tables ... 15
6.3.6 Conditions for considering specific table equivalences ... 15
6.3.7 Conditions for results to be considered equivalent ... 15
6.4 Declaration of a delta ... 15
6.5 Name of the Common Template Table and name declaration ... 17
Annex A (normative) **Common Template Table** ... **18**
Annex B (informative) **Example tailoring deltas** ... **20**
Annex C (informative) **Preparation** ... **29**
Annex D (informative) **Tutorial on solutions brought by this document to problems of lexical ordering** ... **45**
Annex E (informative) **Searching and fuzzy matches** ... **49**
Bibliography ... **51**

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.

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) or the IEC list of patent declarations received (see patents.iec.ch).

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

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded character sets*.

This sixth edition cancels and replaces the fifth edition (ISO/IEC 14651:2019), which has been technically revised.

The main changes compared to the previous edition are as follows:

— ordering data has been added for the new characters standardized in the sixth edition of ISO/IEC 10646 (2020);

— the content of 6.2.2 has been revised for more completeness;

— the weights of character U+A9B5 (JAVANESE VOWEL SIGN TOLONG) have been changed, as the latter is considered a variant of character U+A9B4 (JAVANESE VOWEL SIGN TARUNG). This needed to be fixed.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document provides a method, applicable around the world, for ordering text data, and provides a Common Template Table which, when tailored, can meet a given language's ordering requirements while retaining reasonable ordering for other scripts.

The Common Template Table requires some tailoring in different local environments. Conformance to this document requires that all deviations from the template, called "deltas", be declared to document resultant discrepancies.

This document describes a method to order text data independently of context.

ISO/IEC TR 30112 has specifications for ordering that informatively complement the specifications in this document and indicates where additional information can be sought on ordering keywords defined in this document.

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

1 Scope

This document defines the following.

— A reference comparison method. This method is applicable to two character strings to determine their collating order in a sorted list. The method can be applied to strings containing characters from the full repertoire of ISO/IEC 10646. This method is also applicable to subsets of that repertoire, such as those of the different ISO/IEC 8-bit standard character sets, or any other character set, standardized or not, to produce ordering results valid (after tailoring) for a given set of languages for each script. This method uses collation tables derived either from the Common Template Table defined in this document or from one of its tailorings. This method provides a reference format. The format is described using the Backus-Naur Form (BNF). This format is used to describe the Common Template Table. The format is used normatively *within* this document.

— 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 all characters encoded in the Unicode 13.0 standard,[27] included in ISO/IEC 10646:2020. 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).

NOTE 1 This Common Template Table is to be modified to suit the needs of a local environment. The main worldwide benefit is that, for other scripts, often no modification is required and the order will remain as consistent as possible and predictable from an international point of view.

NOTE 2 The character repertoire used in this document is equivalent to that of the Unicode Standard version 13.0[27].

— A reference name. The reference name refers to this particular version of the Common Template Table, for use as a reference when tailoring. In particular, this name implies that the table is linked to a particular stage of development of the ISO/IEC 10646 Universal coded character set.

— Requirements for a declaration of the differences (delta) between the collation table and the Common Template Table.

This document does *not* mandate the following.

— A specific comparison method; any equivalent method giving the same results is acceptable.

— A specific format for describing or tailoring tables in a given implementation.

— Specific symbols to be used by implementations, except for the name of the Common Template Table.

— Any specific user interface for choosing options.

— Any specific internal format for intermediate keys used when comparing, nor for the table used. The use of numeric keys is not mandated either.

— A context-dependent ordering.

— Any particular preparation of character strings prior to comparison.

NOTE 1 It is normally necessary to do preparation of character strings prior to comparison even if it is not prescribed by this document (see Annex C).

NOTE 2 Annex D describes problems that gave way to this International Standard with their anticipated solutions.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 10646:2020, *Information technology — Universal Coded Character Set (UCS)*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at https://www.iso.org/obp

3.1
character string
sequence of characters considered as a single object

Note 1 to entry: Note to entry: A character string to be ordered does not normally include the characters that delimit it, as for example an “end of line” control character in a text file to be sorted.

3.2
collating symbol
symbol (3.12) used to specify weights assigned to a *collating element* (3.4)

3.3
collation table
weighting table
mapping from *collating elements* (3.4) to *weighting elements* (3.14)

3.4
collating element
sequence of one or more characters that are considered a single entity for *ordering* (3.7)

3.5
delta
list of the differences between a given *collation table* (3.3) and another one

Note 1 to entry: The given collation table, together with a given delta, forms a new collation table.

Note 2 to entry: Unless otherwise specified in this document, the term “delta” always refers to differences from the Common Template Table as defined in this document.

3.6
level
collation level
sequence number for a *subkey* (3.11) in the series of subkeys forming a key

