BS ISO/IEC 14651:2019

BSI Standards Publication

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

Documents
Document Sharing Hub

Full document details: BS ISO/IEC 14651:2019

National foreword

This British Standard is the UK implementation of ISO/IEC 14651:2019. It supersedes BS ISO/IEC 14651:2016+A1:2017, which is withdrawn.

The UK participation in its preparation was entrusted to Technical Committee IST/5, Programming languages, their environments and system software interfaces.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2019
Published by BSI Standards Limited 2019

ISBN 978 0 580 99266 7

ICS 35.040.10

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 28 February 2019.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO/IEC 14651:2019

INTERNATIONAL STANDARD

ISO/IEC 14651

Fifth edition
2019-01

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 14651:2019(E)

Documents
Document Sharing Hub

© ISO/IEC 2019

Full document details: BS ISO/IEC 14651:2019

COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2019

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS ISO/IEC 14651:2019

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 ... 8
6.3 Common Template Table: Formation and interpretation ... 9
6.3.1 General ... 9
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** ... **30**
Annex D (informative) **Tutorial on solutions brought by this document to problems of lexical ordering** ... **46**
Annex E (informative) **Searching and fuzzy matches** ... **50**
Bibliography ... **52**

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

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 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 fifth edition cancels and replaces the fourth edition (ISO/IEC 14651:2016), which has been technically revised. It also incorporates the Amendment ISO/IEC 14651:2016/Amd.1:2017.

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

— the conformance clause has been moved from Clause 2 to Clause 5 to align with current ISO/IEC directives;

— ordering data has been added for the new characters standardized in the fifth edition of ISO/IEC 10646 (2017);

— a note has been added to 3.1 (formerly 4.1) concerning the delimitation of a character string;

— the content of 6.2.2 has been substantially revised for more completeness and 6.3.4 and 6.4 have been aligned in consequence;

— the content of D.3 has been modified.

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.

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

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.

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

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 10.0 standard[27], included in ISO/IEC 10646:2017. 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 10.0[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.

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

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 10646:2017, *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 objectNote 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

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

3.7
ordering
collation
process by which, given two strings, it is determined whether the first one is less than, equal to, or greater than the second one

3.8
ordering key
sequence of *subkeys* (3.11) used to determine an order

3.9
preparation
collation preparation
process in which given *character strings* (3.1) are mapped to (other) character strings before the calculation of the *ordering key* (3.8) for each of the strings

3.10
reference comparison method
method for establishing an order between two *ordering keys* (3.8)

Note 1 to entry: See Clause 6.

3.11
subkey
sequence of weights computed for a *character string* (3.1)

3.12
symbol
collating element (3.4)

3.13
weight
collation weight
positive integer value, used in *subkeys* (3.11), reflecting the relative order of *collating elements* (3.4)

3.14
weighting element
list of a given number of weights sequentially ordered by level

4 Symbols and conventions

Following ISO/IEC 10646, characters are referenced as U*X* where *X* stands for a series of one to eight hexadecimal digits (where all the letters in the hexadecimal string are in upper case) and refers to the value of that character in ISO/IEC 10646. This convention is used throughout this document.

In the Common Template Table, arbitrary symbols representing weights are used according to the BNF notation description in 6.3.1.

5 Conformance

A process is conformant to this document if it produces results identical to those that result from the application of the specifications given in 6.2 to 6.5.

A declaration of conformity to this document shall be accompanied by a statement, either directly or by reference, of the following:

— the number of levels that the process supports; this number shall be at least three;

— whether the process supports the forward position processing parameter;

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

— whether the process supports the backward processing parameter and at which level;

— the tailoring *delta* described in 6.4 and how many levels are defined in the delta;

— if a preparation process is used, the method used shall be declared.

It is the responsibility of implementers to show how their delta declaration is related to the table syntax described in 6.3, and how the comparison method they use, if different from the one mentioned in Clause 6, can be considered as giving the same results as those prescribed by the method specified in Clause 6. The use of a preparation process is optional and its details are not specified in this document.

It is strongly recommended that processes present available tailoring options to users.

6 String comparison

6.1 Preparation of character strings prior to comparison

It can be necessary to transform character strings before the reference comparison method is applied to them. Although not part of the Scope of this document, preparation can be an important part of the ordering process. See Annex C for some examples of preparation.

Characters of the input string shall be encoded according to ISO/IEC 10646 (UCS) or a mapping to ISO/IEC 10646 shall be provided if another encoding scheme is used.

Therefore it can be an important part of the preparation phase to map characters from a non-UCS encoding scheme to the UCS for input to the comparison method. This task can, amongst other things, encompass the correct handling of escape sequences in the originating encoding scheme, the mapping of characters without an allocated UCS codepoint to an application-defined codepoint in the private zone area and change the sequence of characters in strings that are not stored in logical order. For example, for visual order Arabic code sets, input strings shall be put into logical order; and for some bibliographic code sets, strings with combining accents stored before their respective base character require that the combining accents be put after their base character. The resulting string sequence may then have to be remapped into its original encoding scheme.

The Common Template Table is designed so that combining sequences and corresponding single characters (precomposed) will have precisely the same ordering. To avoid inadvertently breaking this invariant (and in the process breaking Unicode conformance), tailoring should reorder combining sequences when corresponding precomposed characters are reordered. For example, if Ä is reordered after Z, then the sequence <A>+<combining diaeresis> should also be reordered. To avoid exposing encoding differences that can be invisible to the end-user, it is recommended that strings be normalized according to Unicode normalization form NFD to achieve this equivalence (see [29]).

Escape sequences and control characters constitute very sensitive data to interpret, and it is highly recommended that preparation should filter out or transform these sequences.

NOTE Since the reference method is a logical statement for the mechanism for string comparison, it does not preclude an implementation from using a non-UCS character encoding only, as long as it produces results as if it were using the reference comparison method.

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

6.2 Key building and comparison

6.2.1 Preliminary considerations

6.2.1.1 Assumptions

The collation table is a mapping from collating elements to weighting elements. In each weighting element, four levels are described in the Common Template Table. This number of levels can be extended or reduced, but cannot be less than 3, in tailoring.

NOTE In the Common Template Table, levels generally have the following characteristics, although this purpose is not absolute.

Level 1: This level generally corresponds to the set of common letters of the alphabets for that script, if the script is alphabetic, and to the set of common characters of the script if the script is ideographic or syllabic.

Level 2: This level generally corresponds to diacritical marks affecting each basic character of the script. For some languages, letters with diacritics are always considered an integral part of the basic letters of the alphabet and are not considered at this second level, but rather at the first. For example, in Spanish, N TILDE is considered a basic letter of the Latin script. Therefore, tailoring for Spanish will change the definition of N TILDE from "the weight of an N in the first level and the weight of a TILDE in the second level" to "the weight of an N TILDE (placed after N and before O) in the first level, and indication of the absence of a diacritic in the second level". For some characters, variant letter shapes are also dealt with on level 2. An example of this is ß, the LATIN SMALL LETTER SHARP S, which is treated as equivalent to ss on level 1, but traditionally distinguished from it on level 2.

Level 3: This level generally corresponds to case distinctions (upper and lower case) or to distinctions based on variant letter shapes (like the distinction between Hiragana and Katakana).

Level 4: This level generally corresponds to weighting differences that are less significant than those at the other levels. Often the last level (level 4 in the Common Template Table) is intended to specify additional weighting for "special" characters, i.e. characters normally not part of the spelling of words of a language (such as dingbats, punctuation, etc.), sometimes called "ignorable" characters in the context of computerized ordering.

6.2.1.2 Processing properties

A given tailored table has specific scanning and ordering properties. These properties can have been changed by the tailoring.

A scanning direction (forward or backward) for each level is used to indicate how to process the string. The scanning direction is a global property of each level defined in the tailored table.

If the last level is greater than three, there is an optional property of this level of comparison called “position” option: when active, a comparison on the numeric position of each “ignorable” character in the two strings is effected, before comparing their weights. In other words, for two strings equivalent at all levels except the last one, the string having an ignorable character in the lowest position comes before the other one. In case corresponding ignorable characters are at the same position, then their weights are considered, until a difference is found. *Support* for this kind of processing is *optional* and is not necessary to claim conformance to this document.

NOTE The scanning direction (forward or backward) is not normally related to the natural writing direction of scripts. The scanning direction applies to the logical sequence of the coded character string.

According to ISO/IEC 10646, for scripts written right to left, such as Arabic, the first characters in the logical sequence correspond to the rightmost characters in their natural presentation sequence. Conversely, for the Latin script, written left to right, the first characters in the logical sequence correspond to the leftmost characters in their natural presentation sequence.

Scanning forward starts with the lowest position in the logical sequence, while scanning backward starts from the highest position, independently of the presentation sequence. The scanning direction for ordering purposes is a global property of each level described in the table.

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

In ISO/IEC 10646, the Arabic script is artificially separated into two pseudo-scripts: 1) the logical, intrinsic Arabic, coded independently of contextual shapes, and 2) the Arabic presentation forms. Both allow the complete coding of Arabic, but intrinsic Arabic is normally preferred for better processing, while presentation-form Arabic is preferred by some presentation-oriented applications. ISO/IEC 10646 does not prescribe that the presentation forms be stored in any specific order, and in some implementations, the storage order for the latter is the reverse of the storage order used for intrinsic Arabic. It is therefore advisable that the preparation phase be used to make sure that Arabic presentation forms and other Arabic characters be fed to the comparison method in logical order.

A tailored sort table can be separated into sections for ease of tailoring. Each section is then assigned a name consistent with the specification in 6.3.1. One of the tailoring possibilities is to assign a given order to each section and to change the relative order of an entire section relative to other sections.

6.2.2 Reference ordering key formation

6.2.2.1 General

When two strings are to be compared to determine their relative order, the two strings are first parsed into a sequence of collating elements taking into account the multi-character "collating-element" statements declared and used in a tailored table (*if* the syntax of 6.3.2 is used). For the syntax used for expressing the Common Template Table, the name of a collating element consisting of a single character, is formed by the UCS value of the character, expressed as a hexadecimal string, prefixed with "U". For multi-character collating elements, the name and association to characters can be found via the collating elements declarations.

NOTE Collating elements with more characters have preference over shorter ones. As an example, if a multicharacter collating element is defined for "abc" and another one is defined for "ab" or for "bc", then if "abc" is encountered, the collating element for "abc" will apply and not the one for "ab" or "bc".

Then, a sequence of *m* intermediary subkeys is formed out of a character string, where *m* is the number of levels described in a tailored collation weighting table.

Each ordering key is a sequence of subkeys. Each subkey is a list of numeric weights. A subkey is formed by successively appending the list of the weights assigned, at the level of the subkey, to each collating element of the string. The keyword "`IGNORE`" in the Common Template Table at the place of a sequence of collating symbols at a level indicates that the sequence of weights at that level for that collating element is an empty sequence of weights.

6.2.2.2 Weighting elements to be ignored

When forming a sort key, collating elements ignored at the first level or at the two first levels and that follow a collation element ignored at all levels but the last one, do not keep their weights as defined in the common template table (or a tailored table); each of these weights shall be zeroed (this means that "IGNORE" shall be assigned to each non-nil weight).

6.2.2.3 Implicit weights computing

If there is no entry in the tailored table for a character of the input string, then the character's weights are undefined. In this case, one shall compute an "implicit" primary weight consisting of a pair of 16-bit words — call them "aaaa" and "bbbb" — and assume that lines like the following were added to the weighting table:

<UXXXX> "<R{aaaa$_{16}$}><T{bbbb$_{16}$}>";<BASE>;<MIN>;<SFFFF>

NOTE <SFFFF> (at the last level) is the largest level 1 weight in the Common Template Table.

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

Distinctions are needed among characters with no entry in the weighting table, and implicit primary weights are thus computed as defined here:

a) Unified Han ideographs:

For a given Han character at code point *cp*:

base_weight = `0xFB40` for original URO and the unified Han characters from the CJK compatibility area

base_weight = `0xFB80` for Extension A through Extension F Han characters

aaaa = $[\text{base_weight} + (\text{cp} >> 15)]_{16}$

bbbb = $[(\text{cp} \,\&\, 0\text{x}7\text{FFF}) \mid 0\text{x}8000]_{16}$

b) Tangut ideographic and component characters:

For a given Tangut character at code point *cp*:

base_weight = `0xFB00`

aaaa = $[\text{base_weight}]_{16}$

bbbb = $[(\text{cp} - 0\text{x}17000) \mid 0\text{x}8000]_{16}$

c) Nüshu ideographic characters:

For a given Nüshu character at code point *cp*:

base_weight = `0xFB01`

aaaa = $[\text{base_weight}]_{16}$

bbbb = $[(\text{cp} - 0\text{x}1\text{B}170) \mid 0\text{x}8000]_{16}$

d) Any other code point not mentioned in the table (non-characters, surrogates, and any characters that are not assigned in the current repertoire):

For a given character at code point *cp*:

base_weight = `0xFBC0`

aaaa = $[\text{base_weight} + (\text{cp} >> 15)]_{16}$

bbbb = $[(\text{cp} \,\&\, 0\text{x}7\text{FFF}) \mid 0\text{x}8000]_{16}$

Thus, the value of the code point is used to calculate the desired weights, by operating on individual bits. The two numbers, converted to four-character hexadecimal values, will then take the following form: "<Raaaa><Tbbbb>".

Decomposable characters are excluded from these computings, as they have an entry in the Common Template Table (with dual primary weights).

If a string contains ill-formed code unit sequences, two approaches show up: the sequence can be handled as if it were U+FFFD (REPLACEMENT CHARACTER), or each subsequence can be ignored. The same approaches can be adopted for any out-of-range values (cp > 10FFFF_{16}).

NOTE 1 Areas in which Han, Tangut and Nüshu characters are found are defined in the comment lines added to the end of the Common Template Table.

NOTE 2 When computed weights are used, characters without explicit mappings are sorted in code point order within each set they belong to (Han, Tangut, Nüshu, others) and they sort properly with respect to the rest of the characters.

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

6.2.2.4 Subkeys formation

There are three ways of forming subkeys: subkeys formed using the "`forward`" processing parameter, subkeys formed using the "`backward`" processing parameter, and subkeys formed using the "`forward,position`" processing parameter. Subkeys that use the "`position`" option can only occur at the last level, and only if that level is greater than three. Support of the "`position`" option is not required for conformance. If the processing parameter "`forward,position`" is not supported, "`forward,position`" shall be interpreted as if the processing parameter had been "`forward`".

6.2.2.5 Formation of subkeys for the first three levels

With the "`forward`" or "`forward,position`" processing parameter, during (forward) scanning of each collating element of the input character string, one or more weights are obtained. These weights are obtained by matching the collating element in the given tailored collation weighting table, obtaining the list of weights assigned to the collating element at the particular level. The obtained weight list is appended to the end of the subkey.

Subkeys, at a particular level, formed with the "`backward`" level processing parameter are built by forming a subkey as with the "`forward`" parameter, then reversing that subkey weight by weight.

6.2.2.6 Formation of the level 4 subkey

At the last level, the subkey is built as described in the preceding paragraph. However, once the ordering key is completely formed (when the end of the string is reached), there are two possible approaches:

a) With the "`forward`" processing parameter: *all* the <SFFFF> symbols are backed out of the subkey;

b) With the "`forward,position`" processing parameter: any *trailing* sequence of <SFFFF> symbol(s) shall be removed from the subkey (leaving intact the remaining part of the subkey).

6.2.3 Reference comparison method for ordering character strings

The reference comparison method for ordering two given character strings (*after* collation preparation, which is not part of the reference comparison method itself) is to compare ordering keys generated by the reference key formation method described in 6.2.2.

— Begin by building an ordering key, using a given tailored collation weighting table, for each of the two given character strings being compared.

— Then compare the resulting keys according to the key ordering definition presented in 6.2.4. Keys can be compared either up to a given level, or up to the last level of the given tailored collation weighting table.

NOTE The comparison can be made *while* generating the ordering keys for two strings to be compared, stopping the key generation when the order of the strings can be determined. Such a technique is sometimes termed lazy evaluation, and some systems support it by default. This avoids generating the full ordering key when an ordering difference is found early in the keys. When a bigger set of strings are to be ordered, one can generate the ordering keys, for example, and store each key or an initial segment of each, before comparing the keys.

6.2.4 Key ordering definition

Weights for different levels should not be compared, which implies that subkeys at different levels should not be compared, nor should keys generated from different tailored tables be compared.

NOTE 1 This allows implementations to assign weightings at each level independently of the other levels, and independently of other tailorings.

m is the maximal level of a given tailored table. Recall that a key is a list, of length m, of subkeys; a subkey is a list of weights; and a weight is a positive integer. Other notations used below are the following:

— L_z is the length of the subkey z, i.e., the number of weights in the subkey;

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

— $z_{\mathrm{wt}(a)}$, where $1 \le a \le L_z$, is the weight at index position a (an integer > 0) of the subkey z;

— $u_{sk(b)}$, where $1 \le b \le m$, is the subkey at level b (an integer > 0) of the key u.

The orderings of weights, subkeys, and ordering keys (up to a given level, or up to the last level) are total order relations, defined for a given tailored collation table as follows.

a) Weights are positive integers (in the reference method) and are compared as such for the purposes of collation.

b) A subkey v is *less than* a subkey w (written $v < w$) **if and only if** there exists an integer, i, where $1 \le i \le L_v + 1$ and $i \le L_w$, such that

   — $i = 1$ and $v_{\mathbf{wt}(i)} < w_{\mathbf{wt}(i)}$, or

   — for all integers, j, where $1 \le j < i$, the equality $v_{\mathbf{wt}(j)} = w_{\mathbf{wt}(j)}$ holds, and either

      — $i \le L_v$ and $v_{\mathbf{wt}(i)} < w_{\mathbf{wt}(i)}$, or

      — $i = L_v+1$ and $0 < w_{\mathbf{wt}(i)}$.

   A subkey v is *greater than* a subkey w (written $v > w$) **if and only if** w is less than v. A subkey v is *equal to* a subkey w (written $v = w$) **if and only if** neither v is less than w, nor w is less than v.

c) An ordering key x is *less than* an ordering key y at level s (written $x <_s y$) **if and only if** there exists an integer i, where $1 \le i \le s$ and $i \le m$, such that

   — $i = 1$ and $x_{\mathbf{sk}(i)} < y_{\mathbf{sk}(i)}$, or

   — for all integers j, where $1 \le j < i$, the equality $x_{\mathbf{sk}(j)} = y_{\mathbf{sk}(j)}$ holds, and $x_{\mathbf{sk}(i)} < y_{\mathbf{sk}(i)}$.

   An ordering key x is *greater than* an ordering key y at level s (written $x >_s y$) **if and only if** y is less than x at level s. An ordering key x is *equal to* an ordering key y at level s (written $x =_s y$) **if and only if** neither x is less than y at level s, nor y is less than x at level s.

d) For ordering keys, <, >, and = are defined as $<_m$, $>_m$, and $=_m$ respectively.

NOTE 2 For ordering keys, $x <_t y$ implies $x <_{t+1} y$, $x >_t y$ implies $x >_{t+1} y$, $x =_t y$ implies $x =_{t-1} y$, $x <_0 y$ is false, $x >_0 y$ is false, and $x =_0 y$ is true. Above level m, for a given tailored table, there are no further ordering distinctions. Note that this definition implies that if two ordering keys are in the 'less than' relationship at level 1, they will also be in the 'less than' relationship at levels 2, 3, 4, etc. In general, whenever two ordering keys are less than at a given level, they will also automatically be less than at all subsequent, higher levels. Conversely, if two ordering keys are equal at a given level, they will also automatically be equal at all preceding, lower levels.

NOTE 3 The decomposition of character string into ordering keys of different levels of comparison from the most significant to the least significant also facilitates fuzzy searches and searches for equivalences. Annex E documents this kind of comparison.

6.3 Common Template Table: Formation and interpretation

6.3.1 General

This subclause specifies the following:

— the syntax used to form the Common Template Table in Annex A or a tailored table based upon the Common Template Table as expressed in Annex A;

— conditions of well-formedness of a table using this syntax;

— interpretation of tailoring statements in deltas for tables formed using this syntax;

— evaluation from symbols to weights of tailored tables formed using this syntax;

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

— conditions for considering two tables as equivalent;

— conditions for considering comparison results as equivalent.

6.3.2 BNF syntax rules for the Common Template Table in Annex A

6.3.2.1 General

Definitions between <angle brackets> make use of terms not defined in this BNF syntax and assume general English usage.

Other conventions:

* indicates 0 or more repetitions of a token or a group of tokens;

\+ indicates 1 or more repetitions of a token or a group of tokens;

? indicates optional occurrence of a token or a group of tokens (0 or 1 occurrences);

parentheses are used to group tokens;

production rules are terminated by a semicolon.

Define collation tables as sequences of lines:

```
weight_table = common_template_table | tailored_table ;
common_template_table =
                 simple_line+ ;
tailored_table = table_line+ ;
```

Define the line types:

```
simple_line =      (symbol_definition | collating_element |
                   weight_assignment | order_end)?
                   line_completion ;
table_line =       simple_line | tailoring_line ;
tailoring_line =   (reorder_after | order_start | reorder_end |
                   section_definition | reorder_section_after)
                   line_completion ;
```

Define the basic syntax for collation weighting:

```
symbol_definition =
                 'collating-symbol' space+ symbol_element ;
symbol_element = symbol | symbol_range ;
symbol_range =   symbol '..' symbol ;
symbol =         simple_symbol | ucs_symbol ;
ucs_symbol =     ('<U'  one_to_eight_digit_hex_string  '>') |
                 ('<U-' one_to_eight_digit_hex_string '>') ;
simple_symbol =  '<' identifier '>' ;
collating_element =
                 'collating-element' space+ symbol space+
                 'from' space+ quoted_symbol_sequence ;
quoted_symbol_sequence =
                 '"' simple_weight+ '"' ;
weight_assignment =
                 simple_weight | symbol_weight ;
simple_weight =  symbol_element | 'UNDEFINED' ;
symbol_weight =  symbol_element space+ weight_list ;
weight_list =    level_token (semicolon level_token)* ;
level_token =    symbol_group | 'IGNORE' ;
symbol_group =   symbol_element | quoted_symbol_sequence ;
order_end =      'order_end' ;
```


© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

Define the tailoring syntax:

```
reorder_after =  'reorder-after' space+ target_symbol ;
target_symbol =  symbol ;
order_start =    'order_start' space+ multiple_level_direction ;
multiple_level_direction =
                 (direction semicolon)* direction (',position')? ;
direction =      'forward' | 'backward' ;
reorder_end =    'reorder-end' ;
section_definition =
                 section_definition_simple |
                 section_definition_list ;
section_definition_simple =
                 'section' space+ section_identifier ;
section_identifier =
                 identifier ;
section_definition_list =
                 'section' space+ section_identifier space+
                 symbol_list ;
symbol_list =    symbol_element (semicolon symbol_element)* ;
reorder_section_after =
                 'reorder-section-after' space+ section_identifier
                 space+ target_symbol ;
```

Define low-level tokens used by the rest of the syntax:

```
identifier =   (letter | digit) id_part* ;
id_part =      letter | digit | '-' | '_' ;
line_completion =
               space* comment? EOL ;
comment =      comment_char character* ;
one_to_eight_digit_hex_string =
               hex_upper |   hex_upper hex_upper |
               hex_upper hex_upper hex_upper |
               hex_upper hex_upper hex_upper hex_upper |
               hex_upper hex_upper hex_upper hex_upper hex_upper |
               hex_upper hex_upper hex_upper
               hex_upper hex_upper hex_upper |
               hex_upper hex_upper hex_upper
               hex_upper hex_upper hex_upper hex_upper |hex_upper
               hex_upper hex_upper hex_upper
               hex_upper hex_upper hex_upper hex_upper ;
hex_numeric_string =
               hex_upper+ ;
space =        ' ' | <TAB> ;
semicolon =    ';' ;
comment_char = '%' ;
digit =        '0' | '1' | '2' | '3' | '4' |
               '5' | '6' | '7' | '8' | '9' ;
hex_upper =    digit | 'A' | 'B' | 'C' | 'D' | 'E' | 'F' ;
letter =       'a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' |
               'h' | 'i' | 'j' | 'k' | 'l' | 'm' | 'n' |
               'o' | 'p' | 'q' | 'r' | 's' | 't' | 'u' |
               'v' | 'w' | 'x' | 'y' | 'z' |
               'A' | 'B' | 'C' | 'D' | 'E' | 'F' | 'G' |
               'H' | 'I' | 'J' | 'K' | 'L' | 'M' | 'N' |
               'O' | 'P' | 'Q' | 'R' | 'S' | 'T' | 'U' |
               'V' | 'W' | 'X' | 'Y' | 'Z' ;
EOL =          <end-of-line in the text conventions in use> ;
character =    <any member of the repertoire of the encoded
                character set in use, not including any
                characters used to delimit the end of lines> ;
```

© ISO/IEC 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019

6.3.2.2 Keyword usage

The usage of the following locale syntax keywords is as follows.

collating-symbol Defines a collating symbol representing a weight.

collating-element Defines a collating element symbol representing a multi-character collating element.

order_start Defines collation rules. This statement is (after reordering is done) followed by one or more collation order statements, assigning multi-level collation weightings to collating elements. A collating element is either a character or a defined substring.

order_end Specifies the end of the collating order statements.

reorder-after Redefines collating rules. Specifies after which collating symbol weighting the lines between the “reorder-after” and the next (or “reorder-end”) are to be moved. This statement is followed by one or more collation table lines. The result is to reassign collating symbol’s values or collating element’s weightings.

reorder-end Specifies the end of the "reorder-after" collating order statements.

section Defines a section of the table. A section can be moved as a whole by “reorder-section-after”.

reorder-section-after Redefines the order of sections. This statement is followed by a section symbol and a target symbol. The result is to reassign collating symbol’s values or collating element’s weightings.

6.3.3 Well-formedness conditions

WF 1. Any *simple_symbol* occurring in a *weight_list* shall also occur in the initial *symbol_element* of a *symbol_weight*, or *symbol_definition* that occurs no later in the sequence of lines that constitute a *weight_table*.

All *simple_symbol*s shall be “defined” before they are “used”.

WF 2. No *symbol* that occurs in a *symbol_definition* in a *weight_table* that contains no *tailoring_line*s may occur in another *symbol_definition* in the same *weight_table*.

Duplication of collation weighting symbols is prohibited. This is true for the Common Template Table itself. It shall remain true for a *tailored_table* after all reordering of lines has been applied.

WF 3. All *weight_list*s in a *tailored_table* shall contain the same number of *level_token*s. An empty *level_token* shall be interpreted as the *collating_element* itself.

A tailorable table shall be consistent in its use of levels throughout.

WF 4. A tailored_table shall contain exactly one order_start statement followed later by exactly one order_end statement. The order_start statement shall appear after the symbol_definition entries and before the symbol_weight entries after all reordering of lines has been applied.

WF 5. A multiple_level_direction in a tailored_table shall contain the same number of directions as the number of level_tokens of any weight_list in that tailored_table.

Any *order_start* shall have the same number of levels as is generally used in the table.

© ISO/IEC 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO/IEC 14651:2019