**ISO/FDIS 12199:2022(E)**

Date: 2022-03-02

ISO/TC 37/SC 2/WG 8

Secretariat: SCC

# Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet

*Mise en ordre alphabétique des données lexicographiques et terminologiques multilingues représentées dans l'alphabet latin*

ISO Copyright Office
CP 401 • CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11
Email: copyright@iso.org

Published in Switzerland.

# Contents

Foreword ... iv

1 Scope ... 6

2 Normative references ... 6

3 Terms and definitions ... 7

4 Preparatory procedures ... 7

5 First ordering level ... 8

5.1 First-ordering-level values ... 8

5.2 First-ordering-level sequence ... 8

5.3 Equivalence between special Latin letters and basic letters ... 9

6 Second ordering level ... 10

6.1 Second-ordering-level values ... 10

6.2 Special Latin letters and letters with diacritical marks ... 10

7 Third ordering level ... 11

7.1 Third-ordering-level values ... 11

7.2 Ordering according to capitalization ... 12

8 Fourth ordering level ... 12

8.1 Fourth-ordering-level values ... 12

8.2 Ordering according to special characters ... 12

Annex A (normative) Word-by-word ordering ... 13

Annex B (informative) Special rules for lexicographical and terminological ordering ... 15

Annex C (informative) Ordering rules for chemical names ... 16

Annex D (informative) Character repertoire of the Latin alphabet ... 18

Annex E (informative) Languages using the Latin alphabet ... 27

Annex F (informative) Alphabetical sequences and character repertoires ... 4

Annex G (normative) Formal description of the rules of the main body of this document ... 24

Bibliography ... 41

# Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO documents 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 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/TC 37, *Language and terminology*, Subcommittee SC 2, *Terminology workflow and language coding*.

This second edition cancels and replaces the first edition (ISO 12199:2000), of which it constitutes a minor revision. The changes ~~compared to the previous edition~~ are as follows:

— the relationship of this document with other International Standards has been updated and transferred from the Foreword to the Introduction;

— in Clause 2 and in the Bibliography, the references have been updated;

— ISO/IEC 14651 is cited informatively and therefore has been moved from Clause 2 to the Bibliography;

— in Annexes D, E and F, the Serbian language has been added among the languages using the Latin alphabet, together with a character set and alphabetical ordering information relating to the Serbian language~~,~~;

— in Annex E, the references to Serbo-Croatian have been deleted~~.~~;

— Annex G is cited informatively and therefore has been changed to "(informative)".

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

In the development of international terminologies, both in printed form and in databases, it is essential to have uniform and internationally recognized rules for the alphabetical ordering of terminological and lexicographical data, to make these terminologies more easily accessible for the users. In addition, it will facilitate the interchange of terminological and lexicographical data.

This document complements other International Standards, such as ISO 10241-1.

# Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet

## 1 Scope

This document specifies the sequence of characters to be used in the alphabetical ordering of multilingual terminological and lexicographical data (terms, term elements, or words) represented in the Latin alphabet. Character sets of languages represented in the Latin alphabet are taken into account insofar as terminological or lexicographical data have been recorded. Character sets used in internationally standardized transliteration into Latin script are also taken into account.

The sequence of alphabetical characters given is intended for multilingual purposes only and is not intended to affect the alphabetical order of any specific language.

The main part of this document specifies letter-by-letter ordering of character strings. ~~Normative annex~~Annex A treats word-by-word ordering, which is a widely used alternative to this system.

~~Informative annex~~Annex B gives two additional rules that ~~may~~can be useful for lexicographical and terminological ordering.

~~Informative annex~~Annex C gives ordering rules for chemical names.

~~Informative annex~~Annex D lists the character repertoire of the Latin alphabet.

~~Informative annex~~Annex E lists languages using the Latin alphabet.

~~Informative annex~~Annex F gives alphabetical sequences derived from the sequence specified in this document for a number of languages that use the Latin alphabet.

~~Normative annex~~Annex G gives a formal description of the rules laid down in the main part of this document conforming with ISO/IEC 14651.

## 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 1087, *Terminology work and terminology science — Vocabulary*~~.~~

ISO/IEC 10646-1~~:1993,~~[1], *Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane*~~.~~

~~NOTE In this Minor Revision of ISO 12199:2000 reference continues to be made to ISO/IEC 10646-1:1993. ISO/IEC 10646-1 and ISO/IEC 10646-2 have since been merged into ISO/IEC 10646.~~

~~ISO/IEC 14651, *Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering*.~~

[1] ~~Cancelled and replaced by ISO/IEC 10646:2020.~~ In this minor revision of ISO 12199:2000, reference continues to be made to ISO/IEC 10646-1:1993. ISO/IEC 10646-1 and ISO/IEC 10646-2 have since been merged into ISO/IEC 10646:2020.

## ~~5~~3 Terms and definitions

For the ~~purpose~~purposes of this document, the terms and definitions given in ISO 1087 ~~(for terminological concepts )~~ and the following apply.

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

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

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

**3.1**
**character**
member of a set of elements used for the organization, control or representation of data

**3.2**
**letter**
*character* (3.1) used for writing natural language, often representing a sound in the language

**3.3**
**digit**
*character* (3.1) used to represent the numeric value, or part thereof, of a number

**3.4**
**special character**
*character* (3.1) that is not a *letter* (3.2) nor a *digit* (3.3)

EXAMPLE The space character is a special character.

**3.5**
**ligature**
*character* (3.1) resulting from the joining of two or more *letters* (3.2)

~~NOTE —~~Note 1 to entry: The resulting character is, in some cases, considered a separate letter.

**3.6**
**polygraph**
two or more consecutive *letters* (3.2) that are regarded as one letter for some purpose

~~NOTE —~~Note 1 to entry: A polygraph consisting of two or three letters may be referred to as a digraph or a trigraph, respectively.

**3.7**
**diacritical mark**
*character* (3.1) that is not a *letter* (3.2) and is placed over, under, or through a letter or a combination of letters

**3.8**
**ordering**
act of bringing strings of *characters* (3.1) into a well-defined sequence according to a string comparison specification

## ~~6~~4 Preparatory procedures

In the process of alphabetical ordering, character strings are compared according to a set of rules. This document specifies the set of rules to be used for the ordering, but does not address the means of selection of relevant character strings, nor any modification of the strings that ~~may~~can be needed for a given purpose.

Consequently, certain preparatory procedures ~~may~~can be needed before applying the ordering rules. Depending on the needs in each individual case, it is possible that:

— the relevant character strings ~~may~~ have to be selected, e.g. relevant terms ~~may~~ have to be extracted from a corpus~~,~~;

— the character strings ~~may~~ have to be modified, e.g. sentence-initial uppercase letters ~~may~~ have to be changed to lowercase letters, plural form of words ~~may~~ have to be changed to singular form~~, or~~;

— leading zeroes or spaces ~~may~~can be added, e.g. in lists containing numerals.

Polygraphs are treated as sequences of separate letters.

An application may arrange information into several ordering fields, and determine ranking order with several separate and independent comparisons. This document only defines a single comparison for one such field, where the field is a character-string field.

Only the characters that appear in the string and their arrangement are taken into account. Apart from the ordering rules and passes, no other knowledge about the words in the character string is used. For example, dictionary information or rules about language syntax, phonetics and semantics are not used.

# ~~7~~5 First ordering level

## ~~7.1~~5.1 First-ordering-level values

When comparing strings to be ordered, the first-ordering-level values of the strings shall be considered first. The subsequent ordering-level values need to be considered only if two or more strings have identical first-ordering-level values.

For multilingual ordering, the following rules shall be applied (~~see annex A~~Annex A shall be applied for word-by-word ordering)~~:~~.

## ~~7.2~~5.2 First-ordering-level sequence

Digits and letters have the following ordering values:

a) **Digits:**

`0 1 2 3 4 5 6 7 8 9`

NOTE 1 Sequences of digits ~~will be~~are ordered from left to right as written, thus generating the following order, ~~e.g.:~~for example: `1 10 100 11 110 111 12 19 190 2 21 3`.

NOTE 2 Leading zeroes ~~may~~can be inserted as a preparatory procedure, e.g. to generate the following order: `0001 0002 0003 0010 0011 0012 0019 0021 0100 0110 0111 0190`.

b) **Basic letters of the Latin alphabet:**

`a A b B c C d D e E f F g G h H i I j J k K l L m M n N`

`o O p P q Q r R s S t T u U v V w W x X y Y z Z þ Þ`

NOTE ~~1~~ 3 This order has been established for use in multilingual environments so as to conflict with as few individual languages as possible. See ~~informative annex~~ Annex F for examples of deviations from this sequence in some languages.

Uppercase and lowercase letters shall be treated as equivalent (see ~~clause~~Clause 7). Letters of the Latin alphabet with diacritical marks shall be treated as equivalent to the corresponding basic Latin letters (see ~~clause~~Clause 6). Special letters of the Latin alphabet shall be treated as equivalent to basic Latin letters according to Table 1 in 5.3 (see ~~clause~~Clause 6).

The Turkish language distinguishes ı/I from i/İ, while other languages have the pair i/I only. To order multilingual data including Turkish text, the i/I pair shall be expanded as follows:

1: ı/I U0131/U0049 LATIN LETTER DOTLESS I (Turkish)

2: i/I U0131/U0049 LATIN LETTER I (non-Turkish)

3: i/İ U0069/U0130 LATIN LETTER I WITH DOT ABOVE (Turkish)

It should also be noted that, for example, í (U00ED LATIN SMALL LETTER I WITH ACUTE) in normal print is represented as LATIN SMALL LETTER DOTLESS I WITH ACUTE. For the purpose of ordering, however, it shall be treated as equivalent to i (U0069 LATIN SMALL LETTER I) on the first ordering level.

NOTE ~~2~~ 4 Throughout this document, characters are referenced as UXXXX, where X is any hexadecimal digit and refers to the position of the character in ISO/IEC 10646-1. Character names are given as in ISO/IEC 10646-1. Most names of Latin letters start with "LATIN SMALL LETTER …" and "LATIN CAPITAL LETTER …". When referring to both lowercase and uppercase letter, the name "LATIN LETTER …" is used. When there is no danger of misinterpretation, the words "LATIN LETTER" are sometimes omitted.

c) **Letters of other alphabets:**

Letters of other alphabets follow in the sequences established for each alphabet. The order of non-Latin alphabets shall be: the Greek alphabet, the Cyrillic alphabet, other alphabets.

NOTE 5 It is outside the scope of this document to establish the sequences for alphabets other than the Latin alphabet. The Greek alphabet has the following sequence of letters:

α Α β Β γ Γ δ Δ ε Ε ζ Ζ η Η θ Θ ι Ι κ Κ λ Λ μ Μ ν Ν ξ Ξ

ο Ο π Π ρ Ρ σ Σ τ Τ υ Υ φ Φ χ Χ ψ Ψ ω Ω~~.~~

All other characters, e.g. punctuation marks, shall be ignored. See ~~clause~~Clause 8.

### ~~7.3~~5.3 Equivalence between special Latin letters and basic letters

Special Latin letters shall be treated as equivalent to basic letters of the Latin alphabet according to Table 1. Uppercase and lowercase letters shall be treated as equivalent.

**Table 1 — Equivalence between special Latin letters and basic letters**

| Position | Character name in ISO/IEC 10646-1 | Character position for lowercase/uppercase in ISO/IEC 10646-1 | | Equivalent to |
|---|---|---|---|---|
| 01 | LATIN LETTER AE | U00E6 | U00C6 | ae |
| 02 | LATIN LETTER B WITH HOOK | U0253 | U0181 | b |
| 03 | LATIN LETTER C WITH HOOK | U0188 | U0187 | c |
| 04 | LATIN LETTER D WITH STROKE | U0111 | U0110 | d |
| 05 | LATIN LETTER D WITH HOOK | U0257 | U018A | d |
| 06 | LATIN LETTER ETH | U00F0 | U00D0 | d |
| 07 | LATIN LETTER G WITH HOOK | U0260 | U0193 | g |
| 08 | LATIN LETTER H WITH STROKE | U0127 | U0126 | h |
| 09 | LATIN LETTER K WITH HOOK | U0199 | U0198 | k |
| 10 | LATIN SMALL LETTER KRA | U0138 | [a] | k |
| 11 | LATIN LETTER L WITH STROKE | U0142 | U0141 | l |

| Position | Character name in ISO/IEC 10646-1 | Character position for lowercase/uppercase in ISO/IEC 10646-1 | | Equivalent to |
|---|---|---|---|---|
| 12 | LATIN LETTER ENG | U014B | U014A | n |
| 13 | LATIN LETTER O WITH STROKE | U00F8 | U00D8 | o |
| 14 | LATIN LIGATURE OE | U0153 | U0152 | oe |
| 15 | LATIN SMALL LETTER SHARP S | U00DF | [a] | ss |
| 16 | LATIN LETTER T WITH STROKE | U0167 | U0166 | t |

[a] No corresponding uppercase letter.

# ~~8~~6 Second ordering level

## ~~8.1~~6.1 Second-ordering-level values

If the comparison of two strings results in identical first-ordering-level values, second-ordering-level values shall be applied according to 6.2.

The rule shall be applied from left to right.

## ~~8.2~~6.2 Special Latin letters and letters with diacritical marks

Special Latin letters, that have been treated as equivalent to basic Latin letters according to Table 1, shall be ordered according to the order in Table 1.

Diacritical marks shall be ordered according to Table 2.

NOTE This order has been established for multilingual environments so as to be in conflict with as few individual languages as possible. See ~~informative annex~~ Annex F for examples of deviations from this sequence in some languages.

**Table 2 — Ordering of diacritical marks**

| Position | Name | Position for combining diacritical mark in ISO/IEC 10646-1 |
|---|---|---|
| 0000 | none | — |
| 0100 | ACUTE ACCENT | U0301 |
| 0200 | GRAVE ACCENT | U0300 |
| 0300 | BREVE | U0306 |
| 0301 | BREVE AND ACUTE | — |
| 0302 | BREVE AND GRAVE | — |
| 0310 | BREVE AND HOOK ABOVE | — |
| 0311 | BREVE AND TILDE | — |
| 0313 | BREVE AND DOT BELOW | — |
| 0315 | BREVE AND COMMA BELOW | — |
| 0400 | CIRCUMFLEX ACCENT | U0302 |
| 0401 | CIRCUMFLEX AND ACUTE | — |
| 0402 | CIRCUMFLEX AND GRAVE | — |
| 0410 | CIRCUMFLEX AND HOOK ABOVE | — |

| Position | Name | Position for combining diacritical mark in ISO/IEC 10646-1 |
|---|---|---|
| 0411 | CIRCUMFLEX AND TILDE | — |
| 0413 | CIRCUMFLEX AND DOT BELOW | — |
| 0500 | CIRCUMFLEX ACCENT BELOW | U032D |
| 0600 | CARON | U030C |
| 0614 | CARON AND CEDILLA | — |
| 0700 | RING ABOVE | U030A |
| 0701 | RING ABOVE AND ACUTE | — |
| 0800 | DIAERESIS | U0308 |
| 0813 | DIAERESIS AND DOT BELOW | — |
| 0817 | DIAERESIS AND MACRON | — |
| 0900 | DOUBLE ACUTE ACCENT | U030B |
| 1000 | HOOK ABOVE | U0309 |
| 1100 | TILDE | U0303 |
| 1200 | DOT ABOVE | U0307 |
| 1300 | DOT BELOW | U0323 |
| 1400 | CEDILLA | U0327 |
| 1500 | COMMA ABOVE/BELOW | U0313 and U0326[a] |
| 1600 | OGONEK | U0328 |
| 1700 | MACRON | U0304 |
| 1713 | MACRON AND DOT BELOW | — |
| 1800 | MACRON BELOW | U0331 |
| 1900 | PRECEDED BY APOSTROPHE | — |
| 2000 | FOLLOWED BY APOSTROPHE | — |
| 2100 | HORN | U031B |
| 2101 | HORN AND ACUTE | — |
| 2102 | HORN AND GRAVE | — |
| 2110 | HORN AND HOOK ABOVE | — |
| 2111 | HORN AND TILDE | — |
| 2113 | HORN AND DOT BELOW | — |

[a] The position of combining comma above and below the base character.

## ~~9~~7 Third ordering level

### ~~9.1~~7.1 Third-ordering-level values

If the comparison of two strings results in identical first- and second-ordering-level values, third-ordering-level values shall be applied according to 7.2.

The rule shall be applied from left to right.

### ~~9.2~~7.2 Ordering according to capitalization

A lowercase letter shall be ordered before the corresponding uppercase letter. [See 5.2, item b), first paragraph after ~~note 1~~NOTE 3.]

NOTE The terms “lowercase letter” and “uppercase letter” are used for members of the sets “a b c ...” and “A B C ...”, respectively. In character names, the naming conventions of ISO/IEC 10646-1 are used. ISO/IEC 10646-1 uses “LATIN SMALL LETTER” and “LATIN CAPITAL LETTER”, respectively.

## ~~10~~8 Fourth ordering level

### ~~10.1~~8.1 Fourth-ordering-level values

If the comparison of two strings results in identical first-, second- and third-ordering-level values, fourth-ordering-level values shall be applied according to 8.2.

The rule shall be applied from left to right.

### ~~10.2~~8.2 Ordering according to special characters

Special characters are ordered according to the sequence of the default template of ISO/IEC 14651. For most special characters, this is the order in which they are listed in ISO/IEC 10646-1.

NOTE In word-by-word ordering (see ~~normative annex~~Annex A), the space character and possibly other special characters ~~may~~can have special functions as key separators.

# Annex A
(normative)

# Word-by-word ordering

## A.1 Principles of word-by-word ordering

As noted in the ~~scope~~Scope, this document specifies the letter-by-letter ordering of character strings. Word-by-word ordering is a widely used alternative to this system. Table A.1 illustrates the difference between letter-by-letter ordering and word-by-word ordering.

**Table A.1 — Letter-by-letter and word-by-word ordering**

| Letter-by-letter ordering | Word-by-word ordering |
|---|---|
| ad | ad |
| adhesive | ad hoc |
| ad hoc | ad infinitum |
| adieu | adhesive |
| ad infinitum | adieu |
| adipose | adipose |

## A.2 Multiple-key ordering

Single-key ordering is described in the main body of this document. In multiple-key ordering, all the ordering rules are applied to one key before they are applied to the next, until all the keys have been considered or a unique sequence has been established.

NOTE One typical example of multiple-key ordering is a list of delegates to a meeting, where the first key ~~may~~can be the country names, the second key ~~may~~can be the delegates’ last names, and the third key ~~may~~can be the delegates’ first names. In this example, if a country has one delegate only, the second key (last names) will not be considered.

## A.3 Word-by-word ordering as multiple-key ordering

In word-by-word ordering, space characters, and possibly also by definition other characters, are key separators. The key-separator characters function as key separators only, and they have no position in the ordering sequence.

When the character string has been divided into a sequence of keys, the ordering rules of the main body of this document are invoked for one key at a time.

NOTE 1 In addition to the space characters, some or all punctuation marks ~~may~~can be defined as key separators. It ~~may~~can also be useful to define some space characters as key separators, while other space characters remain special characters within a key. The choices ~~will~~ depend on the language(s) and type of strings to be ordered.

NOTE 2 If space characters and hyphens are defined as key separators, the title of this clause would be split into the following keys: *<A.3> <Word> <by> <word> <ordering> <as> <multiple> <key> <ordering>*, where each key is contained within < and >, and the spaces are added for increased readability.

## A.4 Simple word-by-word ordering

If the text to be ordered using word-by-word ordering contains very few special Latin letters and diacritical marks, the following extension to the rules in the main body of this document will produce the same or nearly the same output as the rules described in ~~clause~~Clause A.3.

On the first ordering level (see 5.2), the **space character** is added as the first item. Items 1, 2, and 3 in 5.2 then become items 2, 3, and 4. The space character is not treated as a special character on the fourth ordering level (~~clause~~see Clause 8).

NOTE Depending on the language(s) and type of strings to be ordered, it ~~may~~can be useful to treat even other special characters (e.g. hyphens) in the same way as the space character.