INTERNATIONAL STANDARD

ISO
12199

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

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 12199:2000
https://standards.iteh.ai/catalog/standards/sist/e045c3cb-2cf5-4b63-adc8-
6bb2e9abd774/iso-12199-2000

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 12199:2000
https://standards.iteh.ai/catalog/standards/sist/e045c3cb-2cf5-4b63-adc8-
6bb2e9abd774/iso-12199-2000

Contents

Page

iii

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this International Standard may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

International Standard ISO 12199 was prepared by Technical Committee ISO/TC 37, *Terminology (principles and coordination),* Subcommittee SC 2, *Layout of vocabularies.*

It complements other International Standards prepared by ISO/TC 37, such as ISO 10241:1992, *International terminology standards — Preparation and layout* and ISO 12200:1999, *Computer applications in terminology — Machine-readable terminology interchange format (MARTIF) — Negotiated interchange.*

Annexes A and G form a normative part of this International Standard. Annexes B to F are for information only.

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.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 12199:2000
https://standards.iteh.ai/catalog/standards/sist/e045c3cb-2cf5-4b63-adc8-
6bb2e9abd774/iso-12199-2000

ISO 12199:2000
https://standards.iteh.ai/catalog/standards/sist/e045c3cb-2cf5-4b63-adc8-
6bb2e9abd774/iso-12199-2000

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

1 Scope

This International Standard 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 International Standard specifies letter-by-letter ordering of character strings. Normative annex A treats word-by-word ordering, which is a widely used alternative to this system.

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

Informative annex C gives ordering rules for chemical names.

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

Informative annex E lists languages using the Latin alphabet.

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

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

2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this International Standard. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO 1087:1990, *Terminology — Vocabulary.*

ISO 1087-1:—[1], *Terminology work — Vocabulary — Part 1: Theory and application.*

ISO 1087-2:2000, *Terminology work — Vocabulary — Part 2: Computer applications.*

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

ISO/IEC 14651:—[1], *Information technology — International string ordering — Method for comparing character strings and description of a default tailorable ordering.*

_____

1) To be published.

3   Terms and definitions

For definitions of terminological concepts, see ISO 1087, ISO 1087-1 and ISO 1087-2.

For the purpose of this International Standard, the following terms and definitions apply.

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

3.2
letter
character used for writing natural language, often representing a sound in the language

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

3.4
special character
character that is not a letter nor a digit

EXAMPLE   The space character is a special character.

3.5
ligature
character resulting from the joining of two or more letters

NOTE   The resulting character is, in some cases, considered a separate letter.

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

NOTE   A polygraph consisting of two or three letters may be referred to as a digraph or a trigraph respectively.

3.7
diacritical mark
character that is not a letter and is placed over, under, or through a letter or a combination of letters

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

4   Preparatory procedures

In the process of alphabetical ordering, character strings are compared according to a set of rules. This International Standard 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 be needed for a given purpose. Consequently, certain preparatory procedures may be needed before applying the ordering rules. Depending on the needs in each individual case

—   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 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 International Standard 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.

5   First ordering level

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 for word-by-word ordering):

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 ordered from left to right as written, thus generating the following order, e.g.: 1 10 100 11 110 111 12 19 190 2 21 3.

NOTE 2   Leading zeroes may 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   This order has been established for use in multilingual environments so as to conflict with as few individual languages as possible. See informative annex F for examples of deviations from this sequence in some languages.

Uppercase and lowercase letters shall be treated as equivalent (see clause 7). Letters of the Latin alphabet with diacritical marks shall be treated as equivalent to the corresponding basic Latin letters (see 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 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   U0069/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   Throughout this International Standard, 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   It is outside the scope of this International Standard to establish the sequences for alphabets other than the Latin alphabet. The Greek alphabet has the following sequence of letters:

α A   β B   γ Γ   δ Δ   ε E   ζ Z   η H   θ Θ   ι I   κ K   λ Λ   μ M   ν N   ξ Ξ
o O   π Π   ρ P   σ Σ   τ T   υ Y   φ Φ   χ X   ψ Ψ   ω Ω.

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

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
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.				

6   Second ordering level

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.

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 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	–
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.

7 Third ordering level

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.

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   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.

8 Fourth ordering level

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.

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 A), the space character and possibly other special characters may 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, this International Standard 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 International Standard. 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 be the country names, the second key may be the delegates' last names, and the third key may 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 International Standard are invoked for one key at a time.

NOTE 1   In addition to the space characters, some or all punctuation marks may be defined as key separators. It may 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 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.