INTERNATIONAL STANDARD

ISO 12199

Second edition
2022-06

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

Reference number
ISO 12199:2022(E)

© ISO 2022



COPYRIGHT PROTECTED DOCUMENT

© ISO 2022

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
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO 2022 – All rights reserved

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Preparatory procedures ... **2**
5 First ordering level ... **3**
5.1 First-ordering-level values ... 3
5.2 First-ordering-level sequence ... 3
5.3 Equivalence between special Latin letters and basic letters ... 4
6 Second ordering level ... **4**
6.1 Second-ordering-level values ... 4
6.2 Special Latin letters and letters with diacritical marks ... 5
7 Third ordering level ... **6**
7.1 Third-ordering-level values ... 6
7.2 Ordering according to capitalization ... 6
8 Fourth ordering level ... **6**
8.1 Fourth-ordering-level values ... 6
8.2 Ordering according to special characters ... 6
Annex A (normative) **Word-by-word ordering** ... **7**
Annex B (informative) **Special rules for lexicographical and terminological ordering** ... **9**
Annex C (informative) **Ordering rules for chemical names** ... **10**
Annex D (informative) **Character repertoire of the Latin alphabet** ... **12**
Annex E (informative) **Languages using the Latin alphabet** ... **20**
Annex F (informative) **Alphabetical sequences and character repertoires** ... **27**
Annex G (informative) **Formal description of the rules of the main body of this document** ... **40**
Bibliography ... **50**

© ISO 2022 – All rights reserved

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 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;

— in Annexes E and F, the entries related to Moldovan have been corrected in line with ISO 639-1 and ISO 639-2;

— 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.

© ISO 2022 – All rights reserved

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.

© ISO 2022 – All rights reserved

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. Annex A treats word-by-word ordering, which is a widely used alternative to this system.

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

Annex C gives ordering rules for chemical names.

Annex D lists the character repertoire of the Latin alphabet.

Annex E lists languages using the Latin alphabet.

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

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[1)], *Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 1087 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/

1) 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.

© ISO 2022 – All rights reserved

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 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 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

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

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

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

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

Polygraphs are treated as sequences of separate letters.

© ISO 2022 – All rights reserved