BS ISO 5078:2025

BSI Standards Publication

Management of terminology resources — Terminology extraction

Documents
Document Sharing Hub

Full document details: BS ISO 5078:2025

National foreword

This British Standard is the UK implementation of ISO 5078:2025.

The UK participation in its preparation was entrusted to Technical Committee TS/1, Terminology.

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

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

© The British Standards Institution 2025
Published by BSI Standards Limited 2025

ISBN 978 0 539 16705 4

ICS 01.020; 35.240.30

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 2025.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO 5078:2025

International Standard

ISO 5078

First edition
2025-02

Management of terminology resources — Terminology extraction

Gestion des ressources terminologiques — Extraction de terminologie

Reference number
ISO 5078:2025(en)

© ISO 2025

Documents
Document Sharing Hub

Full document details: BS ISO 5078:2025

COPYRIGHT PROTECTED DOCUMENT

© ISO 2025

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

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Principles and methods **5**
4.1 General 5
4.2 Text corpora and terminology extraction 5
4.3 Compilation of text corpora 6
4.3.1 Text corpora used for terminology extraction 6
4.3.2 Criteria for selecting texts for a text corpus 6
4.3.3 Considerations for text corpus creation 7
4.4 Terminology extraction approaches and methods 8
4.4.1 Classification of terminology extraction approaches 8
4.4.2 Extraction method according to the number of languages 10
4.4.3 Extraction method according to the process 11
4.4.4 Extraction method according to the underlying technique 11
4.4.5 Extraction method according to the underlying technology 14
4.4.6 Extraction method according to the extracted items 16
4.5 Term extraction output 17
4.5.1 Filtering candidate term lists 17
4.5.2 Assessing term eligibility 18
4.6 Uses for terminology extraction output 19
5 Implementation of terminology extraction **19**
5.1 General 19
5.2 Initial considerations for terminology extraction 19
5.3 Terminology extraction workflow 20
5.3.1 Overview 20
5.3.2 Starting the terminology extraction workflow 20
5.3.3 Building or selecting a text corpus 20
5.3.4 Preprocessing the text corpus 20
5.3.5 Identifying candidate terms 21
5.3.6 Selecting relevant terms 21
5.3.7 Allocating terms to concepts 22
5.3.8 Identifying concept relations and building concept systems 22
5.3.9 Completing terminological entries 22
Bibliography **23**

© ISO 2025 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025

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 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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 3, *Management of terminology resources*.

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 2025 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025

Introduction

Over the past decades, extracting relevant designations, mostly terms (i.e. linguistic designations), from text corpora has become an increasingly important task carried out in a wide variety of different fields. Terminology extraction, which goes beyond mere extraction of terms, is undertaken by a range of specialists including language professionals in general, and terminologists in particular, as well as ontology engineers, and both information and data scientists. Terminology extraction also serves several purposes that go beyond the compilation of glossaries or the population of terminology databases, including the identification of concepts and of concept relations for building ontologies.

The widespread use of terminology extraction tools in terminology management, as well as in other fields such as information retrieval, stands in stark contrast to the rarity of individual documents that provide definitions, requirements or best practices.

However, although terminology extraction tools save time, money and effort in terminology management, their output becomes even more relevant when it is assessed and validated, using both qualitative and quantitative approaches and criteria for selecting entities such as relevant terms, definitions and concept relations. This extracted and then validated terminological data supports the building of high-quality terminology resources and, thus, terminology management.

This document covers the following aspects that form the core of terminology extraction methods and practices in general:

— compilation of text corpora (general principles and types of text corpora);

— methods and criteria employed by mainstream terminology extraction tools (statistical, linguistic, hybrid and neural);

— criteria for selecting terms (filtering candidate term lists and assessment of term eligibility);

— tool characteristics.

By objectively specifying these aspects, this document provides a reference framework for improving the performance of terminology extraction tools and optimizing the use of their output.

© ISO 2025 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025

Documents
Document Sharing Hub

Management of terminology resources — Terminology extraction

1 Scope

This document specifies methods for extracting candidate terms from text corpora and gives guidance on selecting relevant designations, definitions, concept relations and other terminology-related information.

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 704, *Terminology work — Principles and methods*

ISO 1087, *Terminology work and terminology science — Vocabulary*

ISO 16642, *Management of terminology resources — Terminological markup framework*

ISO 26162-1, *Management of terminology resources — Terminology databases — Part 1: Design*

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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
annotation
process of adding *metadata* (3.10) to segments of language data

[SOURCE: ISO 24617-1:2012, 3.2, modified — “information” replaced by “metadata”; “or that information itself” deleted.]

3.2
bitext
collection of *texts* (3.24) in two languages that can be considered translations of each other and that are segmented and aligned

Note 1 to entry: Bitexts play a key role in training, evaluating and improving localization technologies, such as translation memories, terminology management tools or machine translation engines.

© ISO 2025 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025

3.3
candidate term
term candidate
provisional term
string of *characters* (3.5) that has been collected by means of *term extraction* (3.20) but has not yet been selected as a relevant *term* (3.19) to be considered for inclusion in a *terminological data* (3.22) collection

[SOURCE: ISO 12616-1:2021, 3.18, modified — "text element to be documented in the" replaced by "term to be considered for inclusion in a".]

3.4
candidate terminological data
string of *characters* (3.5) that has been collected by means of *terminology extraction* (3.23) but has not yet been selected as relevant *terminological data* (3.22)

3.5
character
unit of textual information represented by one or more bytes

EXAMPLE Single letter, numeral, punctuation mark, diacritic, symbol, ideograph, space.

[SOURCE: ISO/IEC 14840:1996, 4.10, modified — "textual" added to the definition; example added.]

3.6
collocation
lexically or pragmatically constrained recurrent cooccurrence of at least two *lexical units* (3.8) which are in a direct syntactic relation with each other

EXAMPLE "Commit a crime" instead of "do a crime".

3.7
keyness
quantity proportional to the frequency of a *lexical unit* (3.8) in a subject-field-specific *text corpus* (3.25), relative to a *reference corpus* (3.15)

3.8
lexical unit
meaningful element in the *lexicon* (3.9) of a language

3.9
lexicon
complete set of meaningful elements in a language

3.10
metadata
data that defines and describes other data

[SOURCE: ISO 24531:2013, 4.32]

3.11
n-gram
sequence of *n* adjacent *tokens* (3.27)

Note 1 to entry: Frequently adjacent tokens can be an indicator for *termhood* (3.21).

Note 2 to entry: The number of adjacent tokens (*n*) is usually 2, 3 or 4.

3.12
noise
non-relevant search results

Note 1 to entry: In *terminology extraction* (3.23), "noise" means non-relevant data in the extraction output.

© ISO 2025 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 5078:2025