FINAL DRAFT International Standard

ISO/FDIS 5078

Management of terminology resources — Terminology extraction

Gestion des ressources terminologiques — Extraction de terminologie

ISO/TC **37**/SC **3**

Secretariat: **DIN**

Voting begins on:
2024-11-14

Voting terminates on:
2025-01-09

Reference number
ISO/FDIS 5078:2024(en)

COPYRIGHT PROTECTED DOCUMENT

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

Contents

Page

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.

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.

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, *Computer applications in terminology — 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.

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.

3.13
precision
ratio of relevant search results to all search results

Note 1 to entry: In *terminology extraction* (3.23), "precision" means the ratio of relevant *candidate terms* (3.3) retrieved to the total of candidate terms retrieved.

Note 2 to entry: Precision and *recall* (3.14) generally have an inverse relationship; when one increases, the other tends to decrease.

3.14
recall
ratio of relevant search results to all relevant items in a set that have been or should have been found from a search query

Note 1 to entry: In *terminology extraction* (3.23), "recall" means the relevant *candidate terms* (3.3) in a *text corpus* (3.25).

Note 2 to entry: Recall and *precision* (3.13) generally have an inverse relationship; when one increases, the other tends to decrease.

3.15
reference corpus
text corpus (3.25) to which a given text corpus for *terminology extraction* (3.23) is compared

3.16
relevance
quality of being a successful search result in relation to the search query

3.17
silence
set of relevant search results that have not been found from a search query

Note 1 to entry: In *terminology extraction* (3.23), "silence" means the set of valid *candidate terms* (3.3) that are missing in the extraction results.

3.18
stop word
word that is not taken into account as a *candidate term* (3.3)

Note 1 to entry: Typical stop words are function words (e.g. prepositions, articles), brand names and non-special language words to the specific subject field.

3.19
term
designation that represents a general concept by linguistic means

EXAMPLE "laser printer", "planet", "pacemaker", "chemical compound", "¾ time", "Influenza A virus", "oil painting".

Note 1 to entry: Terms can be partly or wholly verbal.

[SOURCE: ISO 1087:2019, 3.4.2]

3.20
term extraction
identification and excerption of *candidate terms* (3.3)

Note 1 to entry: *Terms* (3.19) can include all types of designations, including appellations, proper names and symbols.

3.21
termhood
degree to which a *lexical unit* (3.8) is recognized as a *term* (3.19)

EXAMPLE "Mouse" has stronger termhood in computer applications and weaker termhood in general language.

Note 1 to entry: Termhood applies to both simple terms (consisting of a single word) and complex terms (consisting of more than one word or lexical unit), and to other designations, such as proper names and appellations, as well as formulas and symbols.

[SOURCE: ISO 26162-3:2023, 3.13, modified — Example revised.]

3.22
terminological data
data related to concepts and their designations

Note 1 to entry: Common terminological data include designations, definitions, contexts, notes to entry, grammatical labels, subject labels, language identifiers, country identifiers, and source identifiers

[SOURCE: ISO 1087:2019, 3.6.1]

3.23
terminology extraction
identification and excerption of *candidate terminological data* (3.4)

3.24
text
content in written form

[SOURCE: ISO 20539:2023, 3.3.1]

3.25
text corpus
collection of natural language data

[SOURCE: ISO 1087:2019, 3.6.4, modified — Admitted term and Note 1 to entry deleted.]

3.26
TF-IDF
term frequency — inverse document frequency
statistical value intended to reflect how important a *lexical unit* (3.8) is to a document in a *text corpus* (3.25)

3.27
token
individual occurrence of a *type* (3.29) in a *text corpus* (3.25)

3.28
tokenization
conversion of *text* (3.24) into *tokens* (3.27)

3.29
type
unique sequence of *characters* (3.5) in a *text corpus* (3.25)

Note 1 to entry: The number of types is different from the number of occurrences (*tokens* (3.27)).

Note 2 to entry: While the number of tokens in a text corpus refers to the total number of occurrences, the number of types refers to the total number of unique occurrences.

3.30
unithood
degree to which a given sequence of words has sufficient collocational strength to form a stable *lexical unit* (3.8)

EXAMPLE “Art deco table” has stronger unithood than “modern table”.

Note 1 to entry: Because unithood derives from the collocational relationship of words making up a given string, it only applies to multi-word *terms* (3.19).

[SOURCE: ISO 26162-3:2023, 3.15]

3.31
validated term
candidate term (3.3) which meets specified criteria

3.32
validated terminological data
candidate terminological data (3.4) which meets specified criteria

3.33
vector
quantity having direction as well as magnitude

[SOURCE: ISO 19123-1:2023, 3.1.51, modified — Note 1 to entry deleted.]

3.34
vector space model
statistical model for representing text information as a *vector* (3.33) of identifiers

Note 1 to entry: Vector space models can be used for information retrieval (IR), natural language processing (NLP) or text mining tasks in order to identify whether *texts* (3.24) are similar in meaning.

[SOURCE: Reference [15], modified — "for Information Retrieval, NLP, Text Mining" moved from the definition to Note 1 to entry; "as a vector of identifiers" added to the definition; Note 1 to entry added.]

4 Principles and methods

4.1 General

Terminology extraction requires a deep understanding of terminology theory and terminology work. In this sense, and to achieve high quality results, the following shall be used:

— established terms and definitions as specified in ISO 1087;

— principles and methods as specified in ISO 704;

— data-modelling criteria as specified in ISO 16642;

— terminology database design principles as specified in ISO 26162-1.

There are various types of text corpora. Selection of corpus type and texts to be included is usually influenced by factors such as project goal, scope and deadlines.

4.2 Text corpora and terminology extraction

Organizations usually produce textual material relating to their industry, activity and the field in which they operate. These kinds of texts include, for example, marketing materials, product documentation, internal memos and bilingual translation memories. Such textual material can contribute to an organization-wide text corpus that forms the basis for terminology extraction.

The usefulness of candidate terminological data extracted from such a text corpus depends on the context and aim of the terminology extraction project as well as on the depth or breadth of the subject-field coverage provided by the text corpus.