SLOVENSKI STANDARD
oSIST ISO/DIS 5078:2024

01-oktober-2024

Upravljanje terminoloških virov - Ekstrakcija terminologije

Management of terminology resources — Terminology extraction

Gestion des ressources terminologiques — Extraction de terminologie

Ta slovenski standard je istoveten z: ISO/DIS 5078

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

oSIST ISO/DIS 5078:2024 **en,fr**

DRAFT INTERNATIONAL STANDARD
ISO/DIS 5078

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

Secretariat: **DIN**

Voting begins on:
2023-12-05

Voting terminates on:
2024-02-27

Management of terminology resources — Terminology extraction

Gestion des ressources terminologiques — Extraction de terminologie

ICS: 35.240.30; 01.020

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 5078:2023(E)

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 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 on 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 the following URL: 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 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 validated terminology extraction 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 corpora (general principles and types of 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 will provide a reference framework for improving the performance of terminology extraction tools and optimising the use of their output.

Management of terminology resources — Terminology extraction

1 Scope

This document describes methods for extracting candidate terms from corpora and provides guidelines for selecting relevant designations, definitions, concept relations and other terminology-related information.

2 Normative references

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 terminological 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* (3.8) to segments of language data

[SOURCE: ISO 24617-1:2012, 3.2, modified – “information” has been replaced with “metadata”, “or that information itself” has been deleted.]

3.2
bitext
collection of texts in two languages that can be considered translations of each other and that are aligned at the sentence or paragraph level

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.

[SOURCE: [13], modified - “(usually electronic)” has been removed after “collection”.]

3.3
candidate term
term candidate
provisional term
string of characters that has been collected by means of *term extraction* (3.18) but has not yet been selected as a *lexical unit* (3.6) to be considered for inclusion in a terminological data collection

Note 1 to entry: In many languages, a lexical unit consists of a word or a group of words.

[SOURCE: ISO 12616-1:2021, 3.18, modified – “text element” has been replaced with “lexical unit”, “to be documented in the” has been replaced with “to be considered for inclusion in a”.]

3.4
collocation
lexically or pragmatically constrained recurrent cooccurrences of at least two *lexical units* (3.6) which are in a direct syntactic relation with each other

EXAMPLE “commit a crime” instead of “do a crime”

[SOURCE: [18], 980, modified – “and/or” has been replaced with “or", “lexical items” has been replaced with “lexical units”, example has been added.]

3.5
keyness
statistical significance of the frequency of a *lexical unit* (3.6) in a subject-field-specific *text corpus* (3.21), relative to a *reference corpus* (3.14)

3.6
lexical unit
unit of language, belonging to the lexicon of a given language

[SOURCE: ISO 26162-3:2023, 3.8]

3.7
logic-based terminology extraction
method of *terminology extraction* (3.20) that uses logical rules and inference techniques to extract terms or terminological information from a *text corpus* (3.21) or from a knowledge base

3.8
metadata
data that defines and describes other data

[SOURCE: ISO 24531:2013, 4.32]

3.9
n-gram
sequence of n co-occurring words within a given sample of text or speech

Note 1 to entry: Frequently co-occurring words may be an indicator for *termhood* (3.19).

Note 2 to entry: The number of co-occurring words (n) is usually 2, 3 or 4.

3.10
noise
items of non-relevance in search results

3.11
parsing
process of determining the syntactic structure of a language unit by decomposing it into more elementary subunits and establishing the relationships among the subunits

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.2818, modified – term “parse” has been replaced with “parsing”, definition beginning “to determine” has been replaced with “process of determining”, example has been removed.]

3.12
precision
ratio of relevant search results to all search results

Note 1 to entry: “Precision” and “recall” generally have an inverse relationship; when one increases, the other tends to decrease.

[SOURCE: ISO 5127:2017, 3.10.3.13, modified – “of the hits” has been replaced with “of relevant search results”, “all hits” has been replaced with “all search results”, Note 2 to entry has been removed.]

3.13
recall
ratio of the *candidate terms* (3.3) found in a text or *text corpus* (3.21) to all candidate terms that have been or should have been found

Note 1 to entry: “Recall” and “precision” generally have an inverse relationship; when one increases, the other tends to decrease.

[SOURCE: ISO 5127:2017, 3.10.3.11, modified – “relevant hits” has been replaced with “candidate terms”, “found in a text or text corpus” has been added, “all documents which have relevance” has been replaced with “all candidate terms that have been or should have been found”, Note 2 to entry has been removed.]

3.14
reference corpus
text corpus (3.21) to which a given text corpus for *terminology extraction* (3.20) is compared

3.15
rule-based terminology extraction
method of *terminology extraction* (3.20) for extracting terms using a set of predefined rules, which are usually based on text patterns, meta-information (tags) or other relevant aspects specific to the particular subject field

3.16
silence
set of valid *candidate terms* (3.3) that are missing in extraction results

3.17
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 (for example prepositions, articles), brand names and general-language words.

3.18
term extraction
identification and excerption of terms

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

3.19
termhood
degree to which a *lexical unit* (3.6) is recognized as a term

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, appellations, as well as formulas and symbols.

[SOURCE: ISO 26162-3:2023, 3.13, modified – “‘bulk carrier ship’ has stronger termhood than ‘ship’ alone” removed from example, “stronger” added before “termhood” in example.]

3.20
terminology extraction
identification and excerption of terminological data

Note 1 to entry: In addition to designations, terminological data can include definitions, concept relations and contexts.