INTERNATIONAL STANDARD

ISO 26162-3

First edition
2023-01

Management of terminology resources — Terminology databases —

Part 3:
Content

Gestion des ressources terminologiques — Bases de données terminologiques —

Partie 3: Contenu

COPYRIGHT PROTECTED DOCUMENT

Contents

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 3, *Management of terminology resources*.

A list of all parts in the ISO 26162 series can be found on the ISO website.

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

Managers, educators and terminology database maintenance authorities conduct both periodic and continuous evaluation of terminology databases containing concept entries for a number of purposes:

— quality assurance-related validation of terminological data collections in business, government and non-governmental organizations;

— formative assessment and summative evaluation and feedback in training and educational environments.

ISO 26162-1 and ISO 26162-2 specify design principles and software considerations for modelling terminology databases (termbases). ISO 26162-1 establishes the general principles of termbase design as outlined in core ISO/TC 37 standards, such as ISO 704, which, among other topics, treats general principles for concept entry content and structure, term identification, basic principles for modelling concept systems and a range of other areas associated with terminology work. ISO 26162-1 also encourages conformity to the terminological metamodel as outlined in ISO 16642. It describes the role that data categories play in modelling terminological data and sets down basic principles for ensuring and evaluating the quality of data stored in termbases, such as data granularity, elementarity and modelling variance. These criteria comprise fundamental benchmarks against which to measure the quality and reliability of terminological data. ISO 26162-2 relates the principles outlined in ISO 26162-1 to the implementation of database design with respect to software and user interface considerations, together with pragmatic workflow implementations in terminology management environments.

This document provides guidance for defining procedures for ensuring high-quality content in terminological data collections designed to meet documentation needs in a range of environments involving, for instance, translation, interpreting and technical communication. Conformity to this document can strengthen processes designed to support a quality management system, such as ISO 9001, and the related auditing procedures in a translation, interpreting or technical communication environment. An error typology is presented in the framework of an overall evaluation model, with generic (non-standardized) options for creating a concept entry evaluation model, depending on the needs of users and of the sponsoring organization.

Annexes A to C provide pragmatic advice on error evaluation practice. Annex A describes the creation of scoring models reflecting the error typology described in Clause 6, allowing for design variations depending on organization needs. For instance, a given scoring model can form the basis for a score card used for students and trainees, which is likely to be different from a score card used for a major enterprise or a national term bank.

Annex B presents a sample term entry. Annex C presents a sample evaluation model that can be adopted or adapted to meet the needs of terminologists, individuals working as freelancers or in companies, governmental organizations and NGOs. The values in this evaluation model can be used to create a scoring method, with the understanding that actual scoring practice is likely to vary according to specifications and user needs.

Management of terminology resources — Terminology databases —

Part 3:
Content

1 Scope

This document specifies content-related aspects of terminology database maintenance. It gives guidance on the content of terminological data collections, with emphasis on data quality evaluation.

This document gives guidance for modellers of concept entries who need to ensure interoperability and high-quality content. It aims to ensure that terminological data collections themselves meet high standards for design conformity with standards such as ISO 12620-1 and ISO 16642, data accuracy and performance. It outlines principles for assuring data quality (see ISO 9001) and evaluating terminological data collections for purposes of continuous improvement. This approach contrasts that of ISO 23185:2009, which focuses on the usability of existing terminology resources.

This document does not apply to the management of text corpora or to term extraction tools.

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

ISO 26162-2:2019, *Management of terminology resources — Terminology databases — Part 2: Software*

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/

3.1
concept entry
entry
terminological entry
part of a *terminological data collection* (3.14) which contains the terminological data related to one concept

Note 1 to entry: A concept entry can contain information treating one or more languages.

Note 2 to entry: In this document, the term *entry* is used as a short form for *concept entry.*

[SOURCE: ISO 30042:2019, 3.5, modified — "entry" made second preferred term, notes to entry added.]

3.2
evaluation model
model for analysing data in a *concept entry* (3.1) according to terminology principles and specified data requirements consistent with the purpose of the *termbase* (3.11)

3.3
core structure
common structure and *data categories* (3.6) that are used in all TermBase eXchange (*TBX*) *dialects* (3.5)

Note 1 to entry: The core structure is compliant with ISO 16642 (TMF).

[SOURCE: ISO 30042:2019, 3.6]

3.4
error typology
systematic list of error types

3.5
TBX dialect
eXtensible Markup Language (XML) that validates according to the *core structure* (3.3) of TermBase eXchange (TBX) and allows exactly those *data categories* (3.6) at those levels specified by a precisely defined configuration of data categories

Note 1 to entry: See ISO 30042 for more detail.

[SOURCE: ISO 30042:2019, 3.12, modified — Simplified for this document, Note 1 to entry replaced.]

3.6
data category
class of data items that are closely related from a formal or semantic point of view

EXAMPLE   /part of speech/, /subject field/, /definition/.

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

Note 2 to entry: In running text, such as in this document, data category names are enclosed in forward slashes, e.g. /part of speech/.

[SOURCE: ISO 30042:2019, 3.8]

3.7
data integrity
conformance of data values to a specified set of rules

[SOURCE: ISO/IEC TR 10032:2003, 2.23]

3.8
lexical unit
unit of language, belonging to the lexicon of a given language

[SOURCE: ISO 1951:2007, 3.8, modified — Reference to a dictionary removed.]

3.9
quality assurance
QA
set of planned and systematic activities necessary to provide confidence that a *concept entry* (3.1) satisfies acceptance criteria based on terminology principles and specified data requirements

3.10
specification
document that sets out detailed requirements to be satisfied by a *terminological data collection* (3.14)

Note 1 to entry: Specifications can include procedures for checking conformity to these requirements.

3.11
termbase
terminology database
database comprising a *terminological data collection* (3.14)

[SOURCE: ISO 30042:2019, 3.28, modified — "terminology database" is no longer an admitted term, but a second preferred term.]

3.12
termbase quality evaluator
person who is qualified as a terminologist or subject-matter specialist who conducts a quality evaluation of a *terminological data collection* (3.14)

3.13
termhood
degree to which a *lexical unit* (3.8) is recognized as a term

EXAMPLE   "bulk carrier ship" has stronger termhood than "ship" alone. "Mouse" has termhood in computer applications, whereas it does not 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.

3.14
terminological data collection
resource consisting of *concept entries* (3.1) with associated metadata and documentary information

EXAMPLE   A TBX document instance, ISO 1087.

[SOURCE: ISO 30042:2019, 3.29, modified — Second preferred term "TDC" removed.]

3.15
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.

4   Identifying terms

4.1   Requirements for term selection

Concept entries should meet the needs of their intended audience and purpose as well as organization-specific requirements, including requirements for terms. ISO 704 discusses principles for assigning and analysing designations.

While it is not possible to set universal requirements for term selection and for the content of concept entries, an important goal of a termbase in many commercial environments is prescriptive in nature: directing users away from terms that are problematic for one reason or another, and towards the use of preferred terms. Thus, term selection involves both identifying the organization's preferred terms and documenting the corresponding synonymous terms that are to be avoided. In this context, implementers should document deprecated and obsoleted terms along with preferred terms, as well as non-central terms that nonetheless recur in critical enterprise content.

Common considerations of including a term in the termbase include:

— Human safety: Important terms that, when used incorrectly, could affect the safety of customers or employees. When the wrong terms are used, the product can be used incorrectly, inadvertently damaging the product, or in some cases, posing a risk to the user. The same concerns apply for terms used in documents related to government regulations that affect citizen access to critical healthcare, interaction with the court system, occupational safety and health, and other resources and human services.

   EXAMPLE 1   Terms related to pharmaceutical prescriptions, safety equipment on an oil rig, airplane landing gear assembly and maintenance, or communications between schools and parents.

— Company survival: Terms that protect the company or organization. Using the wrong terms could result in liability suits or loss of intellectual property.

   EXAMPLE 2   Terms related to regulatory requirements in an industry or patents that a company owns.

— Company identity: Terms in this category represent concepts created by a company. Using them correctly protects the brand and trademarks and strengthens market share. More and more global companies today check possible translations for suitability when deciding on brand-related terms and trade names. It is important to make sure that terms will establish the company as an industry leader and set it apart from the competition. These items would be terms that are related to products or services, or terms that are used with a non-standard meaning.

   EXAMPLE 3   Product names and registered company trademarks; components in patented products; names of features or services below the branding level.

— Subject field: Terms used in an industry or domain. These include terms from vertical industries or areas where experts (outside of a given organization) agree on a set of terms.

   EXAMPLE 4   Domains such as accounting, core automotive, biology.

— Public service entities: Terms that will be transparent to a given target audience.

   EXAMPLE 5   Materials written in an under-resourced language sometimes need to use different terminology from the scholarly or industry-standard terms used by experts.

— Pragmatic and locale-related issues: Any terms that benefit from documentation or standardization, including high-frequency terms or difficult terms in cases where documenting and managing these terms will reduce errors or queries. If products are localized or documents are translated into other languages, this category also covers terms that can pose translation challenges, or which can benefit from standardization in one or more languages. Some languages can benefit from a higher level of terminology documentation than others.

4.2   Unithood and termhood

As shown in Figure 1, "termhood" comprises the intersection among unithood, usage in the sponsoring organization's corpus and the purpose of the terminological data collection (see Reference [16]).
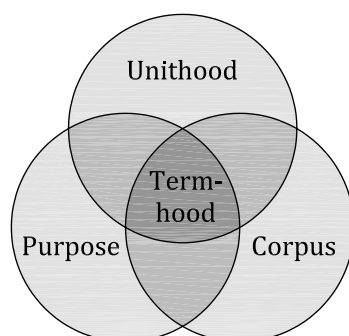

Figure 1 — Termhood

"Unithood" refers to collocational relationships between words and is not applicable to single-word lexical units. Determining whether a given multi-word string comprises a stable, lexical unit that recurs with collocational frequency in the organization's corpus is important for term selection. Because of their relatively stable syntagmatic structures, multi-word units with strong unithood tend to form key communicative structures demanding consistency in commercial environments. While multi-word terms prevail in commercial and professional environments, this does not exclude the documentation of simple terms that have sufficient prominence (frequency and/or dispersion) in the organization's corpus and support the purpose of the terminological data collection.

A multi-word lexical unit which meets the criteria for unithood functions as a term if it designates an identifiable concept in the textual and operational context in question. Each concept shall be documented as a separate concept entry in a termbase, adhering to the principle of concept orientation.

EXAMPLE   Sometimes one word in one language requires multiple concept entries. For instance, the English word (which in some contexts can function as a term) "river" denotes two concepts as illustrated by the French "fleuve" (a river which flows into a sea or ocean) and "rivière" (a river that flows into another river or lake). The German term "Abstandsbolzen" has two conceptual references as evidenced in English: "distance rivet" and "spacer bolt". Both of these cases would require separate concept entries in a termbase.

4.3   Corpora and term extraction

In commercial environments, identifying terms is usually a process informed by the textual context in which they occur. It is recommended to use corpora to identify term candidates. Corpus analysis informs the depth or breadth of subject-field coverage without overloading the termbase with unnecessary entries. Corpora may consist of any kind of written materials produced by the sponsoring organization, including marketing materials, product documentation, internal memos, transcripts, bilingual translation memories or other collections of textual materials. The organization-wide corpus provides evidence of the frequency of occurrence for a given term candidate, as well as its dispersion across various types of textual materials. Both frequency and dispersion are criteria for determining whether a candidate meets the criteria for termhood.

EXAMPLE 1   Certain low-frequency lexical units can be important terms for an organization because they appear across multiple types of content (marketing, sales, training, online content, user guides, etc.), or they can nonetheless represent key concepts.

A range of tools and techniques are available to assist in the extraction of lexical units for the purpose of identifying terms (see ISO 26162-2 as well as ISO 12616-1 for additional information on term extraction tools).

NOTE   ISO 5078[3] on terminology extraction is also being developed.

Two key factors to consider for selecting relevant concepts and terms are:

— marketing: key concepts involving brand recognition that distinguish an organization from its competitors;

   EXAMPLE 2   Concepts and related terms associated with patented and trademarked products or processes.

— customer satisfaction: areas where terms and concepts have historically caused problems and can reflect past quality assurance (QA) problems or other serious risk criteria.

   EXAMPLE 3   Particular topics that have generated a higher than usual number of support calls or customer complaints.

5 Collecting terminological data

5.1 Data requirements

5.1.1 Evaluation procedure

The evaluation procedure proposed in this document is designed to measure the quality of a termbase based on the evaluation of the concept entries it contains and the terms assigned to those concepts, as well as the correctness of related data and its fitness for use. Correctness and fitness for use are judged with respect to adherence to the termbase data model and related specifications.

Quality data meets the requirements for its intended purpose (see ISO 8000-1). The purpose shall be clearly articulated and shall comprise the basis for specifying data requirements. Data requirements shall be aligned with the needs of the sponsoring organization and shall be made available to users and contributors to the termbase. Data correctness and fitness for use can be measured as a function of conformity to specifications.

5.1.2 Quality data

Data shall adhere to best practices outlined in ISO 26162-1 and ISO 26162-2, and shall:

— meet specified requirements for the terminological data collection;

— reference a stable data model;

— use explicitly defined data categories consistent with the data model;

— be portable as specified in ISO 26162-2.

5.1.3 Purpose of the termbase

The purpose of the terminological data collection shall be specified relative to the needs of the organization. A termbase used primarily to support authoring or translation processes requires different entries from those selected for a termbase designed for search optimization or ontological modelling, although these purposes need not be mutually exclusive. For instance:

— Discourse-oriented termbases often provide a range of concept entries reflective of a wide spectrum of subject fields. For instance, a manufacturing company can also document frequently used concepts for human resources management, accounting or facilities maintenance, in addition to core product and process-related concepts.

— Product and process-related entries can be closely linked to a central product catalogue or to computer systems designed for logistics and process control.

— Information management and ontology-oriented systems can be designed for open data reference across the enterprise or even with open accessibility to the Semantic Web.

5.1.4 Data correctness

5.1.4.1 Data correctness is a function of data validation as defined in ISO 1087:2019, 3.6.6. It can be related to:

— the content of individual data fields (see 5.1.4.2);

— the relationship among several data fields within one concept entry (see 5.1.4.3); or

— relationships among multiple concept entries (see 5.1.4.4 and "consistency check" in ISO 1087:2019, 3.6.8).

5.1.4.2   The data provided in a concept entry:

— shall meet the specifications for each data field as set by the organization according to the needs and purposes of the termbase; this criterion is closely reflective of "format" and "completeness" checks cited in ISO 1087:2019, 3.6.11 and 3.6.10;

> EXAMPLE   A given data model can have the following rules: "a definition must be written according to guidelines in ISO 704"; "the content of a term note must follow organization-specific guidelines"; "a context is required according to the guidelines of the TBX-Basic specification". (See Reference [15].)

— shall follow the conventions appropriate for the respective language, e.g. gender, spelling reforms, script choices, register, formality.

5.1.4.3   Regarding appropriateness of the content of data fields within one entry with respect to each other:

— all terms entered in the /term/ fields of an entry shall represent the concept in question;

— all terms that are synonyms or equivalents representing the concept shall be entered in their own /term/ fields and not hidden in other fields where they are inaccessible to filters or search routines;

— definitions, if they exist in multiple languages, shall define the same concept;

— values in term, language and concept-related data fields shall be correct for the term, language and concept, respectively; for instance, /part of speech/ = "noun" is correct only if assigned to a noun.

— synonyms and equivalents associated with a given concept shall be included in a single entry; duplicate entries shall be merged or removed (see "double-entry check", ISO 1087:2019, 3.6.7).

— homophones representing different concepts shall be treated in separate concept entries; if the same lexical unit represents both a noun and a verb, the terms shall be represented in separate concept entries.

5.1.4.4   Regarding consistency of data distributed among multiple concept entries:

— explicit relationships among concept entries shall be correctly identified (see "consistency check", ISO 1087:2019, 3.6.8);

— content of individual data fields shall be consistent across all entries;

— links shall be maintained intact, i.e. if an entry is removed, existing links shall also be removed as specified in ISO 26162-2:2019, 4.7.

5.1.5   Fitness for use

To be fit for use, terminological data shall be entered in the correct data fields (e.g. definitions are in /definition/ fields, notes in /note/ fields).

Depending on the terminology workflow, the current life cycle phase should be indicated for each entry or parts of entries in accordance with system specifications (e.g. unprocessed, provisionally processed, finalized).

Data should allow for classification of subsections of the terminology database, e.g. according to:

— categorization of subject fields within the discipline in question;

— customers or projects within the organizational workflow;

— security requirements;

> EXAMPLE 1   Restricting distribution for security reasons by classifying concepts as public or confidential.

— intellectual property requirements.

EXAMPLE 2   Restriction placed on a term for the purpose of protecting the right of an organization to the exclusive use of the term (e.g. trademark, service mark).

5.2   Data model

The data model shall meet the needs of the organization and reflect the purpose of the termbase.

Each concept entry shall make available the minimum data fields to be fit for use, which can entail including data fields that are only used in certain cases.

EXAMPLE 1   A data field can be included for /transfer comment/, which would only be used in cases where equivalence relationships require explication.

EXAMPLE 2   A term in a concept entry is not labelled as /full form/ if that concept is not also represented by an /abbreviated form/.

If a terminological data collection is identified as compliant with a known data model, such as a TBX dialect, the structure shall comply with that model.

5.3   Data categories

Concept entries shall use explicitly defined data categories consistent with the declared data model. Requirements for individual data categories are set by the organization according to the needs and purposes of the termbase.

Data category names shall comply with specified data category names in the declared data model or clear guidelines shall be provided for mapping local names and values to values used in the declared model. For example, if a terminological data collection specifies a TBX dialect as its data model, data categories shall be compliant with or mapped to that dialect (see ISO 30042).

NOTE   Data category definitions can be found at Reference [14].

5.4   Data portability

The terminological data collection should be portable and accessible by other applications, as specified in ISO 26162-2:2019, Clause 4.

6   Validating concept entry quality

6.1   General validation criteria

The quality of concept entries is validated by evaluating their content with respect to system requirements.

6.2   Error typology and system design

For purposes of consistency and inter-rater reliability, it is recommended to use error typologies that are consistent with the termbase data model. This document outlines detailed criteria for assessing concept entries and provides an illustrative example of one such error typology, as presented in Annex C for the TBX-Basic dialect. Any concept-entry-related error typology shall reflect the specifications for the termbase in question. For instance, in TBX-Basic, either a definition or a context shall be present, but in a given instantiation, the creators of the termbase can have determined that definitions are mandatory, and contexts are optional (see ISO 30042). TBX-compliant dialects should account for expressing concept relations and subject fields as per ISO/TS 24634:2021.

Termbase design and content can vary according to the needs of an organization, as well as cultural, linguistic and regional requirements. As a consequence, actual evaluation (scoring) schemes used with these error typologies are likely to vary as well.

6.3 Error types

6.3.1 Termbase specification and maintenance

The collection of concept entries in the terminology database shall meet the specifications for the data model in question and shall meet the criteria for termhood described in Clause 4.

The termbase shall be updated on an ongoing basis to remain relevant in existing subject fields, to support new subject fields and to address known problem areas. The evaluation of multilingual termbases should be distributed among termbase quality evaluators who are familiar with the respective languages and the needs of translators, technical writers and the subject field.

The error typology in Annex C categorizes error types for:

a) errors involving the data integrity of the entire concept entry;

b) errors at the concept level of the entry;

c) errors at the language level;

d) errors at the term level.

6.3.2 Error types associated with the inclusion of concept entries

Nonconformity with termbase specifications can involve the entire concept entry, as evidenced by the entry data integrity issues listed in the sample error typology shown in Annex C. A systematic response to such errors can include:

— merging entries (where multiple entries exist for the same concept);

— splitting entries (where more than one concept is treated in a single entry);

— removing an entry that does not serve the purpose of the termbase.

6.3.3 Error types associated with automatically generated content

Generally, the following group of data fields is automatically generated by the terminology management system:

— Identifier;

— Creation date;

— Created by;

— Last modified date;

— Last modified by.

Identifiers can appear at concept entry level (entry identifier) or term level (term identifier). The other items can occur anywhere in the entry where the designers wish to document responsibility and date information. Although these items are included in the error typology for the sake of completeness and conformity to the commonly used data models, it is unlikely that any error will occur during their generation. In the event that there is a problem, the solution probably lies at the system level and is not necessarily correctable from a simple data maintenance perspective.

6.3.4 Error types associated with open data categories

The /term/, /definition/ and /context/ data fields are used to record the most critical terminological information in a concept entry (see ISO 704:2022, Clauses 6 and 7, and ISO 26162-1:2019, 4.3). These are free text fields with no formal restrictions. As a consequence, they are subject to a wider range of potential errors than most other fields. In addition, errors occurring in these fields can in some cases generate misunderstanding on the part of the user or incorrect assignment of terms in original texts or equivalents in translated texts.

Errors associated with open data categories should be weighted accordingly in any associated evaluation model. The /note/ field is also open and can contain free text, but it does not usually impact the overall quality of an entry to the same degree as /term/, /definition/ and /context/. Consequently, /note/ is not necessarily subject to the same level of weighting.

Defining specifications for the information allowed in each of these data fields helps prevent errors. Specifications shall indicate how terms are entered into the termbase (capitalization, canonical form, use of parentheses), the format for definitions, and conventions for information recorded in usage notes.

In addition to ensuring that information is free from spelling mistakes, estimating the accuracy of text fields is extremely useful. It is beneficial to be able to search open data fields and identify:

— double spaces;

— embedded hard line breaks;

— hidden codes;

— capitalization of terms;

— consistency between part-of-speech and definition;

— definitions, contexts or examples with extremely short or long strings;

— personal data, such as phone numbers;

— duplicate values.

6.3.5 Error types associated with closed data categories

The remaining data categories are usually configured as picklist values. Although it is possible that the wrong value is selected during creation or modification of a concept entry, picklist values are subject to less potential variance and hence to lower risk from a risk analysis viewpoint. Critical data categories in this group include /subject field/ and /part of speech/. The latter is essential, especially in cases where the form of verbs, nouns and adjectives can be identical. Correct subject field identification can be critical for filtering purposes, ensuring the applicability of definitions, and assigning equivalents in multiple languages.

Annex A
(informative)

Evaluation models

A.1 Scoring models

Scoring models and score cards based on them can be configured to meet the needs and expectations of a particular local, regional or international audience. The rigor and coverage of the scoring model can vary depending on the type and size of a termbase. The implementation of the error typology can also vary depending on the design of the termbase and the mission of the organization in question.

A.2 Reviewing and editing concept entries

In many organizations, concept entries are reviewed and edited periodically by a terminologist or by the appropriate subject-matter specialist. The evaluation of multilingual termbases will usually require a team of termbase quality evaluators. In many cases, these revisions can simply be implemented without significant statistical analysis or detailed documentation. Reviewers can provide feedback resulting in discussion and eventual correction of entries, or in some cases, reviewers simply implement edits. Even if there is no statistical analysis, reviewers can benefit from using the error types identified in Annex C as a guide in evaluating the quality and completeness of concept entries.

A.3 Weighted scoring scales

In other environments, the analysis of termbase concept entries factors into overall quality assurance with respect to technical communication, translation and localization. Ongoing ratings for these language resources can count among key performance indicators for documenting the effectiveness of information management in the enterprise.

One possible approach to structuring an error evaluation model assigns weights to specific parts of the concept entry as more important than others. For instance, a rating model can:

a)   assign a 10-point scale to each language section in a concept entry;

b)   distribute major points to the important data fields, e.g. "term" = 3, "definition" = 2 and "context" = 2;

c)   assign fractional points to other less risk-entailed data fields, e.g. divide the remaining three points up among /source/, /note/ and any other fields, e.g. grammar or usage-related fields;

d)   record scores reflecting the possible points for the category in question subtracted by any error points assigned to the category in question;

e)   calculate total scores for each language section, then tally for entire entries and for the entire termbase or the filtered section of the termbase under consideration.

This suggested approach is analytically simple, and when termbase quality evaluators assign error points, they can account for whatever error types they find appropriate as reflected in the list in Annex C. Implementers can configure evaluation scorecards as needed to meet their needs. Table A.1 provides such a sample score card.

Table A.1 — Sample score card for language section

Error level	Graded score
Language section as a whole (10 points total)	
Critical open data categories (7 points), of which:	
Definition (2 points)	
Term (3 points)	
Context (2 points)	
Low-risk data categories (3 points distributed fractionally among them)	
Total	

Some errors, such as the documentation of an inappropriate concept in one language section of an entry, can invalidate an entire language section for that entry, resulting in an error point value of 10. In the event that the entire entry is inappropriate to the termbase, then the score would be (–10) multiplied by the number of language sections in the entry.

The resulting raw scores can be difficult to quantify unless they are converted into something like a 100-point scale. If, for instance, a termbase for a particular project or subject field contains 367 entries with 5 language sections each, and they are absolutely perfect, the resulting score would be 367 multiplied by 5 multiplied by 10 (points per language section) or 18 350.

If, however, an evaluation using the weighted values shown above yields a score of 16 000, a theoretical percentage value can be achieved by dividing 16 000 (the actual raw quality value) by 18 350 (the ideal value), yielding a weighted score of 87 %. This kind of value provides an intuitively meaningful benchmark for stakeholders creating, maintaining or using the termbase, and improvements in this value over time can be used to measure continuous improvements in the system.
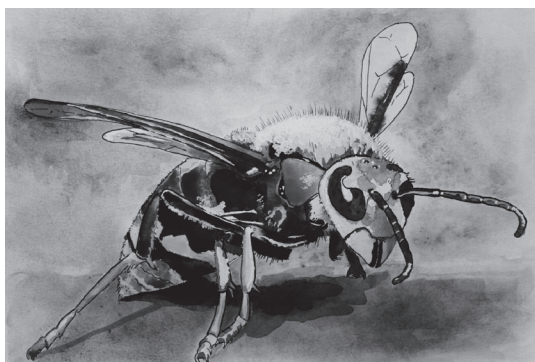
Annex B
(informative)

Sample concept entry[1]

Subject field: entomology

Entry identifier: 7

Security status: public

Figure:

Source of figure: Northern Giant Hornet, ©, reproduced by permission of LTAC Global

English

Definition: insect (order Hymenoptera, of the family Vespidae), which is among the largest eusocial wasps, is found throughout Korea, Japan, Taiwan, and China, as well as the Indian subcontinent, and is notorious for destroying apiaries

Source of definition: Based on: TERMIUM Plus®[2], Retrieved 2020-09-09, https://www.btb .termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=murder+hornet& codom2nd_wet=1#resultrecs[3] and Retrieved 2022-04-17, https://entnemdept.ufl.edu/creatures/ MISC/BEES/Vespa_mandarinia.html.

Term: northern giant hornet

Part of speech: noun

Usage status: preferred

Context: The Entomological Society of America (ESA) announced "northern giant hornet" as the common name for *Vespa mandarinia* (Asian giant hornet) in their list of "Common Names of Insects and Related Organisms" on 2022-07-25.

Source of context: https://www.aphis.usda.gov/aphis/ourfocus/planthealth/plant-pest-and -disease-programs/honey-bees/agh

---

1)   The layout in this sample entry features a bold/italic font for data category names and normal Roman (upright) font for content. There is no intention of replicating any familiar termbase display format.

2)   *TERMIUM Plus*® is an example of a suitable product available for users on the web. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product.

3)   The form of the /source/ entry represented here comprises a single citation as prescribed by APA style[13] and should not be construed as a violation of the principles of granularity and term elementarity.

13

Term: Asian giant hornet

> *Part of speech*: noun

> *Usage status*: deprecated

> *Context*: With queens that can grow to two inches long, Asian giant hornets can use mandibles shaped like spiked shark fins to wipe out a honeybee hive in a matter of hours, decapitating the bees and flying away with the thoraxes to feed their young.

> *Source of context:* Baker, Mike. 'Murder Hornets' in the U.S.: The Rush to Stop the Asian Giant Hornet. *The New York Times*. 2020-05-02.

Term: *Vespa mandarinia* Smith

> *Part of speech*: noun

> *Term type:* international scientific term

> *Usage status*: admitted

> *Context*: *Vespa mandarinia* Smith holds the title for being the world's largest hornet.

> *Source of context*: Department of Entomology and Nematology, University of Florida, Featured Creatures, Retrieved 2020-09-09, https://entnemdept.ufl.edu/creatures/MISC/BEES/Vespa_mandarinia.html

Term: murder hornet

> *Part of speech*: noun

> *Register*: slang

> *Usage status*: admitted

> *Geographical usage*: en-US

> *Context*: Junichi Takahashi, a researcher at Kyoto Sangyo University in Japan, said the species had earned the "murder hornet" nickname there because its aggressive group attacks can expose victims to doses of toxic venom equivalent to that of a venomous snake; a series of stings can be fatal.

> *Source of context*: Department of Entomology and Nematology, University of Florida, *Featured Creatures*, Retrieved 2020-09-09, https://entnemdept.ufl.edu/creatures/MISC/BEES/Vespa_mandarinia.html

French

> *Definition:* insecte (ordre des hyménoptères, de la famille des Vespidae), qui est un redoutable tueur d'abeilles, dont les piqûres pour l'homme sont dangereuses

> *Source of definition*: Based in part on: TERMIUM Plus®[2]), Retrieved 2020-09-09, https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=murder+hornet&codom2nd_wet=1#resultrecs

Term: frelon géant d'Asie

> *Part of speech*: noun

> *Grammatical gender*: masculine

> *Usage status*: preferred

Context: [Le] frelon géant d'Asie, Vespa mandarinia, [une espèce] dont les ouvrières [peuvent mesurer] entre 3,5 et 3,9 cm, et les reines jusqu'à 5,5 cm, [est] un féroce prédateur de nombreux insectes et araignées, mais surtout d'Hyménoptères sociaux (abeilles, guêpes, frelons).

Source of context: TERMIUM Plus®[2]), Retrieved 2020-09-09, https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=murder+hornet&codom2nd_wet=1#resultrecs

Term: *Vespa mandarinia*

Part of speech: noun

Term type: international scientific term

Usage status: admitted

German

Definition: Insekt, Insektenordnung der Hymenoptera (Hautflügler) der Familie Vespidae (Wespen), das bis zu zehn Zentimeter groß wird, ganze Bienenvölker vernichten und bei Menschen tödliche allergische Reaktionen hervorrufen kann.

Source of definition: Based on TERMIUM Plus®[2]), Retrieved 2020-09-09, https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=murder+hornet&codom2nd_wet=1#resultrecs

Term: Asiatische Riesenhornisse

Part of speech: noun

Grammatical gender: feminine

Usage status: preferred

Context: Die Asiatische Riesenhornisse (lateinischer Name: *Vespa mandarinia*) schwirrt vorwiegend in den ländlichen Gegenden Ost- und Südostasiens umher.

Source of context: Galileo, Retrieved 2020-09-09, https://www.galileo.tv/natur/die-groesste-hornisse-der-welt/

Term: Honigbienenkillerin

Part of speech: noun

Grammatical gender: feminine

Usage status: admitted

Register: slang

Note: Translation from American English for illustrative purposes; not commonly used in German.

Context: In diesem Jahr wurde in den USA eine rund fünf Zentimeter große Asiatische Hornisse gefunden, [deren] Stachel … für manche Menschen tödlich sein [soll] und [die] daher … in den Medien auch als „Mörder-Hornisse" oder „Honigbienenkillerin" betitelt [wurde].

Source of context: Allgäuer Zeitung, Retrieved 2020-09-13, https://www.allgaeuer-zeitung.de/bayern/asiatische-hornisse-der-neue-feind-unserer-honigbiene-n%C3%A4hert-sich-bayern_arid-229827

Term: *Vespa mandarinia*

> *Part of speech*: noun
>
> *Term type*: international scientific term
>
> *Usage status*: admitted

Annex C
(informative)

Sample error typology

The list of error types catalogued in Table C.1 provides an exhaustive sample that is based on a typical TBX-compliant data model (see ISO 30042). The different errors can be weighted as "minor – major – critical" or assigned a numeric weighting scheme, depending on user needs. Some terminology management systems do not use all these criteria, and others can feature additional ones. Not all criteria apply to all languages or all environments.

Table C.1 — Error types

Data levels and categories	Error types
Entry data integrity	(6.3.1 a), 6.3.2)[a]
	Multiple entries for same concept
	More than one concept treated in single entry
	Entry does not pertain to the subject field or purpose of the termbase
	Concept/term assignment treated in concept entry does not meet the criteria for termhood
	Terms treated in concept are trivial and are not terms as per data quality specifications
	Processing errors, such as assigning terms or term and concept-related fields to the wrong entries during import or filtering
Concept level	(6.3.1 b))
	[The first four data fields in the following list can be repeated at various levels in the concept entry]
Creation date (6.3.3)	Missing (usually automated)
Created by (6.3.3)	Missing (usually automated)
Last modified date (6.3.3)	Missing (usually automated)
Last modified by (6.3.3)	Missing (usually automated)
Subject field	Subject field classification incorrect
	Subject field classification not indicated where required
	Concept orientation is violated by assigning inappropriate subject fields in a single entry in cases where this introduces conceptual ambiguity
Security status (subset)	Missing value (Confidential/Public)
	Incorrect value
Entry identifier (6.3.3)	(Usually automated; not likely to be in error)
Figure	(Not usually mandatory)
	Figure or image inappropriate to the entry
	Poor quality figure, poor representation of the concept
	Copyright citation
	Accessibility issues (with online resources)
Source	Source missing for existing image or figure
	Source inaccurate or inactive URL
	Misuse of /Source identifier/ mechanism
[a] All numbers in parenthesis refer to the relevant subclauses in this document.	

Table C.1 *(continued)*

Data levels and categories	Error types
Note (6.3.4)	Spelling error
	Grammar error
	Irrelevant information
	Definitional information in note
Definition (6.3.4)	Definition missing (where required)
	Spelling error
	Grammar error
	Definition incorrect or false
	Definition inconsistent with specified capitalization and punctuation rules (either starts with capital or not, or ends with period or not, or uses article or not)
	Inappropriate definition for concept, e.g. from the wrong subject field
	Improper form as per data quality specifications, for instance improper form as per ISO 704, if ISO 704 style is specified; examples can include:
	— inclusion of term in definition
	— use of verb "is" [or equivalent] in definition (if omission is the norm)
	— missing essential characteristics; imprecise
	— exceeds single sentence fragment
	— contains supplemental material appropriate for an additional note
	Two definitions in same "definition" field, especially definitions for more than one concept in the same entry
	Definition translated literally from another language rather than representing authentic usage in the language in question
Source of definition	Source missing
	Source inaccurate or inactive URL
	Source unauthoritative or potentially flawed
Cross-reference/Reference	Broken link
	Incorrect link
Language level	(6.3.1 c))
	Language identifier missing
	Elements (terms, definitions, etc.) entered in wrong language section
Definition	See "Definition" above
Source of definition	Source missing
	Source inaccurate or inactive URL
	Source unauthoritative or potentially flawed
Note (6.3.4)	Spelling error
	Grammar error
	Irrelevant information
	Definitional information in note
^a All numbers in parenthesis refer to the relevant subclauses in this document.	

Table C.1 *(continued)*

Data levels and categories	Error types
Term level	([6.3.1](#) d))
	Term missing (where required)
	Term does not meet criteria for inclusion as per termbase specifications
	Important term "hidden" in another concept entry or in another text field, hence unsearchable
	Term in the wrong language section
	Term is a false calque from another language and incorrectly flagged as preferred or admitted
	Term is a mistranslation from another language and incorrectly flagged as preferred or admitted
	Term does not conform to a requirement that prefers native forms to loan words and is incorrectly flagged as preferred
	Term inappropriate as synonym or equivalent to other terms
	Term inappropriate to the definition
	Term designates a superordinate or subordinate concept
	Form error in term representation:
	— term misspelled
	— case error (incorrect upper or lower case)
	— diacritics or other special font features
	— other language-specific representation errors
	Introduction of noun by article
	Introduction of verb by "to," "zu," or other particle
	Incorrect number or case (plural, dative, etc.)
	Elementarity issue: more than one term per /term/ field, e.g. United Nations (UN) in same field instead of recording the abbreviation in its own /term/ field
	Conflation of forms, e.g. hiragana and katakana in same /term/ field
	Conflation of forms, e.g. variant spellings in same /term/ field
	Inclusion of punctuation (e.g. parentheses, periods)
Source of term	Source missing (if required)
	Source inaccurate or inactive URL
	Broken link
Term identifier ([6.3.3](#))	(Usually automated; not likely to be in error)
Part of speech ([6.3.5](#))	Missing
	Incorrect
Grammatical gender	Missing (if required)
	Incorrect
Grammatical number	Failure to identify mass nouns and occasional documented plurals
Usage status	Missing (if required or appropriate)
Term type	Missing (if required or appropriate)
	Indication of "full form" where there is no parallel "abbreviated form" (failure to adhere to data category definitions)
Geographical usage	Missing (if required or appropriate)
ᵃ All numbers in parenthesis refer to the relevant subclauses in this document.	

Table C.1 *(continued)*

Data levels and categories	Error types
Context (6.3.4)	Missing (if required)
	Spelling error
	Grammar error
	Relevant term not present in context sample
	Context does not reflect the concept or usage
	Uninformative context
	Translation from another language (does not represent authentic usage in the language in question)
	Embedded term has wrong part of speech (e.g. the concept is a noun, but the term in the context is an adjective or verb)
	Presence of extended or extraneous material
	Instead of following capitalization and punctuation from the original context, rules for definition style are inappropriately applied to text in the /context/ field
Source of context	Source missing
	Source inaccurate or inactive URL
	Source unauthoritative or potentially flawed
Note (6.3.4)	Spelling error
	Grammar error
	Irrelevant information
	Definitional information in note
Usage note	Critical collocations and co-occurrences missing or misreported, where appropriate
Cross-reference	Broken link
	Incorrect link
Term location	Missing (if required)
	Inaccurate
Customer	Missing or inaccurate (if required)
Project	Missing or inaccurate (if required)
External cross-reference	Broken link
	Incorrect link
Concept system	Incorrect relations
	Relations expressed in the termbase (definitions, contexts) are incongruent with the relations modelled in the concept system
	Inconsistent modelling representation (UML, graphic representation, etc.)
^a All numbers in parenthesis refer to the relevant subclauses in this document.	

Bibliography

[1] ISO 704:2022, *Terminology work — Principles and methods*

[2] ISO 1951:2007, *Presentation/representation of entries in dictionaries — Requirements, recommendations and information*

[3] ISO 5078,[4)]*Management of terminology resources — Terminology extraction*

[4] ISO 8000-1, *Data quality — Part 1: Overview*

[5] ISO 9001, *Quality management systems — Requirements*

[6] ISO 12616-1, *Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography*

[7] ISO 12620-1, *Management of terminology resources — Data categories — Part 1: Specifications*

[8] ISO 16642, *Computer applications in terminology — Terminological markup framework*

[9] ISO 23185:2009, *Assessment and benchmarking of terminological resources — General concepts, principles and requirements*

[10] ISO/TS 24634:2021, *Management of terminology resources — TBX-compliant representation of concept relations and subject fields*

[11] ISO 30042:2019, *Management of terminology resources — TermBase eXchange (TBX)*

[12] ISO/IEC TR 10032:2003, *Information technology — Reference Model of Data Management*

[13] APA Style, https://apastyle.apa.org, accessed 2022-03-16

[14] datcatinfo website, http://datcatinfo.net, accessed 2020-05-13

[15] TBX info website, https://www.tbxinfo.net, accessed 2019-12-12

[16] Warburton, Kara. The Corporate Terminologist. In: *Terminology and Lexicography Research and Practice*. Volume 21. L'Homme, M.-C. and Kageura, K. (Ed.). John Benjamins B.V., Amsterdam/ Philadelphia, 2021

---

4) Under preparation. Stage at the time of publication: ISO/WD 5078:2021.

ICS 35.240.30; 01.020

Price based on 21 pages