TECHNICAL REPORT

ISO TR 12618

First edition
1994-11-15

Computational aids in terminology — Creation and use of terminological databases and text corpora

Aides informatiques en terminologie — Création et utilisation de bases de données terminologiques et de corpus de textes

Reference number
ISO/TR 12618:1994(E)

Contents

International Organization for Standardization
Case Postale 56 • CH-1211 Genève 20 • Switzerland

Printed in Switzerland

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 main task of ISO technical committees is to prepare International Standards. In exceptional circumstances a technical committee may propose the publication of a Technical Report of one of the following types:

— type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

— type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

— type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

ISO/TR 12618, which is a Technical Report of type 3, was prepared by Technical Committee ISO/TC 37, *Terminology (principles and coordination)*, Sub-Committee SC 3, *Computational aids in terminology*.

Introduction

Because the scope of this Technical Report is limited to computational aids for terminology work, the user is advised to consult ISO 704 and ISO 1087 for questions of basic principles of terminology.

In addition to the advice for creating and using terminological databases given in this Technical Report, an exchange format for terminological and lexicographical data is standardized by ISO 6156 and ISO 12200.

Computers can be employed at various stages in the preparation and use of terminological data collections. The preparation of terminological data collections typically includes the following phases:

a) defining scope;

b) identifying, selecting and recording sources;

c) collecting terms, definitions, explanations, text examples, etc.;

d) elaborating systems of concepts;

e) establishing equivalence relations between concepts in two or more languages;

f) recording terminological information, including information on systems of concepts;

g) updating terminological data.

These phases are presented above in the chronological order of the process, but they often overlap and each phase may have to be repeated subsequently. Depending on the type of project and resources involved, computers may prove useful in many phases, particularly b), c), f), and g).

Computer-aided use of terminological data collections includes retrieval of terminological information stored in a database, and production of printouts and dictionaries.

The emphasis of this Technical Report is on the creation and maintenance of a terminological database (i.e. phases f) and g) above). A short introduction concerning the creation and use of a machine-readable text corpus which may be used in phase c) is given in clause 20, it being borne in mind that the creation of a text corpus precedes the creation of a terminological database. Information on phase b), however, is beyond the scope of this Technical Report.

Computational aids in terminology — Creation and use of terminological databases and text corpora

1 Scope

This Technical Report provides guidance on the basic principles and methods for the application of data processing support in the preparation and use of terminological data collections. This Technical Report is especially applicable to the creation and use of terminological databases and text corpora.

2 References

ISO 704:1987, *Principles and methods of terminology.*

ISO 860:1994, *Terminology work — International harmonization of concepts and terms.*

ISO 1087:1990, *Terminology — Vocabulary.*

ISO 1087-2:—[1], *Terminology work — Vocabulary — Part 2: Computational aids in terminology.*

ISO/IEC 2382-1:1993, *Information processing — Vocabulary — Part 1: Fundamental terms.*

ISO 2382-4:1987, *Information processing — Vocabulary — Part 4: Organization of data.*

ISO 6156:1987, *Magnetic tape exchange format for terminological / lexicographical records (MATER).*

ISO/TR 8393:1985, *Documentation — ISO bibliographic filing rules (International Standard Bibliographic Filing Rules) — Exemplification of bibliographic filing principles in a model set of rules.*

ISO 8777:1993, *Information and documentation — Commands for interactive text searching.*

ISO 8879:1986 [+ Amd 1:1988], *Information processing — Text and office systems — Standard Generalized Markup Language (SGML).*

ISO/IEC 9075:1992, *Information technology — Database Languages — SQL.*

ISO 10241:1992, *International terminology standards — Preparation and layout.*

ISO 12200:—[1], *Computational aids in terminology — Terminological interchange format (TIF) — An SGML application.*

3 Definitions

For the purpose of this Technical Report, the following definitions apply.

NOTE 1 Most of these definitions will be incorporated in ISO 1087-2, and are at present provisional.

3.1 data bank

collection of databases including the organizational framework for managing them

NOTE 2 See also ISO/IEC 2382-1:1993.

3.2 database

collection of data organized according to a conceptual structure

NOTE 3 Adapted from ISO/IEC 2382-1:1993.

1) To be published.

3.3 data category
data element type
instruction for interpreting a given data field

3.4 data element
smallest identifiable unit of content in a given record

3.5 data field
variable or fixed length portion of a record reserved for a particular data element

NOTE 4 Adapted from ISO 6156:1987.

3.6 record
set of data elements treated as a unit
[ISO 2382-4:1987]

3.7 terminography
recording, processing and presentation of terminological data

NOTE 5 Adapted from ISO 1087:1990.

3.8 term bank
terminological data bank
data bank containing terminological data

3.9 terminological database
database containing terminological data

3.10 terminological data collection
collection of data containing information on concepts of specific subject fields

3.11 terminological entry
part of a terminological data collection that contains the terminological data related to one concept

NOTE 6 See also ISO 1087:1990, subclause 6.2.2.2.

3.12 text corpus
corpus
systematic collection of machine-readable texts or parts of text prepared, coded and stored according to predefined rules

NOTE 7 A text corpus may be limited according to aspects of subject fields, size or time, e.g. mathematical texts, or certain periodicals from 1986 onwards. It is used as source material for further linguistic analysis or terminology work.

NOTE 8 See also ISO 1087:1990, subclause 6.1.2.2.

4 Types of terminological data collections

The following criteria effect the ways that terminological data collections are manipulated and accessed:

— **size**: number of entries, subject fields, languages, data categories;

— **hardware**: microcomputer, minicomputer, mainframe-computer; hard disk storage, diskette, CD-ROM; standalone system or network system;

— **software**: database management system, information retrieval system, dictionary editing system; off-the-shelf or custom-tailored design;

— **owner/user**: international organization, national institution, company, individual; free access or restricted access;

— **applications**: on-line or off-line retrieval of terms for computer-aided translation, printouts (e.g. containing all entries within a subject field as basic material for a working group), production of printed vocabularies, computer typesetting, use in expert systems or machine translation systems.

Other types of data collection may be integrated with the terminological data collection, e.g.:

— full text databases (see also clause 20);

— graphical databases;

— numerical databases;

— bibliographical databases.

5 Criteria for creating a terminological database

Establishing a terminological database may be useful if one or more of the following criteria are met:

a) There is a need for a harmonized mono-, bi-, or multilingual terminology at international, national or company level.

b) There is a permanent need for updating and revising large volumes of data.

c) There is a need to search within terminological data by means of different criteria or combinations of criteria (e.g. by term in one language, by subject or by source).

d) There is a need for presenting data in different formats according to user specifications (e.g. alphabetically or systematically ordered special vocabularies as subsets for machine translation or computer-aided translation).

e) The number of potential users needing fast access to the data is large enough to justify the investment in hardware, software and human resources (training, programming, maintenance, etc.).

f) The human resources needed are available both for the training of the personnel and for creating and maintaining the database, as well as the financial resources needed for acquiring hardware and software.

6 Hardware and software requirements

The size of the terminological data collection and the number of potential users will determine whether a microcomputer, a minicomputer or a mainframe computer is needed.

Various types of software can be used for recording and using terminological data collections, e.g. word processing systems, dictionary editing systems, database management systems and information retrieval systems. Database management systems — and to some extent information retrieval systems — are the most flexible systems for handling data. This Technical Report, therefore, focuses on the creation and use of a terminological database by means of a database management system or an information retrieval system. In the following text, "database system" is used to cover both database management systems and information retrieval systems.

Many database systems are available for different operating systems, in either single- or multi-user versions. Some systems run on micro-, mini-, and mainframe computers. Ideally it should be possible to upgrade a system (from the micro- to the minicomputer version or from the single-user to the multi-user version) (see clause 18).

Some systems run on microcomputers with very limited internal storage, but it is often advisable to invest in additional RAM capacity. If a database system interacts with other programs and thus forms an integrated part of a more powerful system, even more capacity is needed.

Most database systems require additional storage space for data management purposes (internal markers, indexes etc.) — sometimes up to 10 times the space needed for the "raw" text of the terminological entries. With an average entry size of, e.g. 1 000 characters (bytes), 1 000 entries may occupy up to 10 MB storage capacity, although many systems have facilities for reducing the space occupied by the database. There should be facilities for back-up, e.g. on hard disk or diskettes or by using streaming tape.

A terminological database may be made available on optical media, e.g. CD-ROM (Compact Disk — Read-Only Memory), which can hold large amounts of data. Special equipment is needed to read an optical media database.

7 Terminological data categories

The structure and data categories of terminological entries must be clearly described. These descriptions are necessary for data handling.

Data categories should be defined and delimited independently. Thus each data element can be unambigously assigned to one and only one category. "Subject classification" is an example of a data category. "UDC 621" or "UDC 347" could then be relevant data elements.

Other examples of data categories are:

— term;
— grammatical information;
— definition;
— context;
— collocation;
— relation between concepts;
— source references.

Different types of data collection require different combinations of data categories. A database used for producing printed dictionaries contains data categories different from those in a database used for the retrieval of individual terms. Different user groups (e.g. students, translators, subject field experts) need different types of information.

Very often a terminological database is multifunctional. It is not, however, always possible to foresee all future needs during the planning stage. Therefore it is advisable to define a database structure to be as flexible as possible so that it allows for the addition of new data categories at any later stage.

8 Data structure

8.1 Terminological data structure

To be able to describe the structure of the entries in a terminological data collection, information is needed on the relations between the data elements.

Each data element is related to the concept as a whole or to any other data element, typically a term. Data elements may be optional or mandatory, repeatable or not.

The internal data format normally differs from the external format, i. e. the format presented to the user.

Data elements such as definition, responsible terminologist, etc. are customarily related to the concept. Part of speech, collocations, etc. are related to terms. Source reference may be related to definitions, terms, collocations, etc. These relationships are illustrated in figure 1.

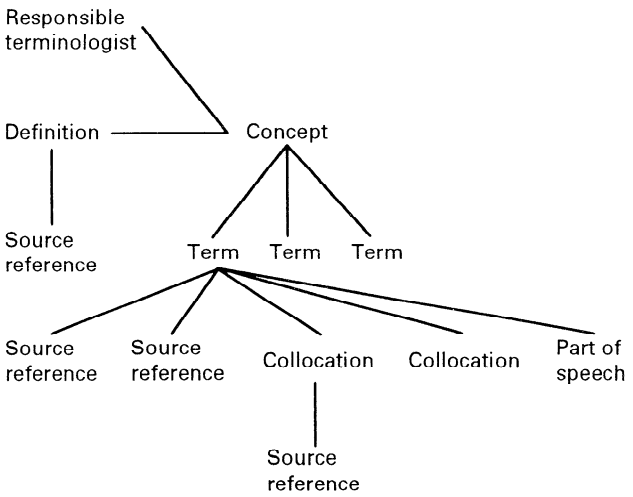

Figure 1 — Data elements in a monolingual entry with three synonymous terms

The terms in figure 1 are synonymous terms that may be accompanied by information on stylistic or regional usage restrictions. It is often necessary to append a number of items to each term (e.g. source references, collocations, notes, etc.) The terms illustrated in figure 1 could also be arranged as one preferred term (in the TERM field) and two admitted terms (in a SYNONYM field). This procedure would apply in a terminological database for standardized terminology.

Each terminological entry contains the information on one concept in one or more languages. It may, however, sometimes be sensible to include partially equivalent concepts in two languages in the same entry if, despite the differences, it is reasonable to use the terms for the two languages as translations of each other. In such cases, however, the differences between the two concepts should be clearly indicated in a note on equivalence.

Within subject fields where there are no significant equivalence differences — which is often the case in technical subject fields — one entry can contain information on more than two languages. In subject fields like law, social sciences, education, etc. the equivalence differences often make a multilingual approach impossible. In such cases it is better that the information in each entry refer to a single language pair. Ideally, it should be possible to store mono-, bi- and multilingual terminology in the same database.

A database may consist of language pairs, where one language is always the same, e.g. English in an English term bank. In case searches between two other languages, e.g. French and German, should be permitted, it is advisable to build in automatically generated restrictions or warnings that point out that the equivalence relationship is only established between a given language and English. If, for example, an English concept is related to both a German and a French concept, and partial equivalence has been shown in both language pairs (English—German and English—French), the user should be informed that the equivalence relationship between German and French has not been verified.

In ideal cases the terminology of a subject field is worked out in parallel in two or more languages. Concept systems are established, and concepts are defined independently for all the languages. Equivalence relationships between the languages are then established by comparing the definitions and the systems of concepts (see ISO 860). All information categories, definitions, contexts, sources, etc., may be supplied for each concept in

all the languages. Consequently no language is considered as the source or the target language in the database.

When the database is used for interactive retrieval and production of a printed vocabulary, each language may be chosen as either the source or the target language, and it is possible to select different subsets of information according to the user group and purpose of the dictionary (see clause 11).

When a concept in one language does not have an equivalent in another language a translation may be suggested, but this proposed translation should never appear as a source language term in a printed dictionary. Therefore, such proposed translations need to be marked as such in the database.

8.2 Database implementation

To establish a terminological database, a formalized description of the data structure is needed. For this purpose, various types of diagrams may be used. One type of diagram is the entity-relationship diagram for the description of the data structure in a hierarchical, relational or network database. The simplified tree structure of a terminological entry in figure 1 may be represented in an Entity-Relationship diagram as shown in figure 2, where the following types of relationship occur:

— one-to-one (1:1) relationships, e.g. between term and part of speech;
— one-to-many (1:*n*) relationships, e.g. between concept and term;
— many-to-many (*n*:*m*) relationships, e.g. between term and collocation.

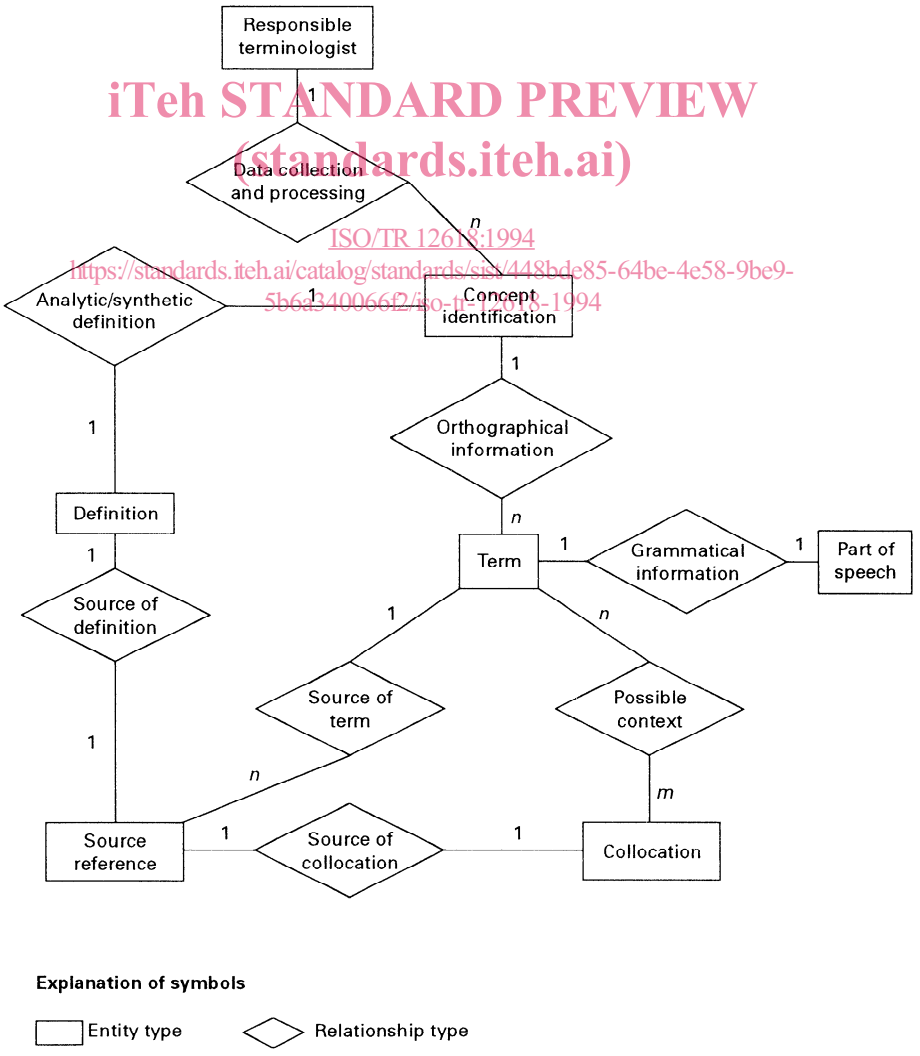

Figure 2 — Example of an entity-relationship diagram for a terminological database

The structure of the terminological entry is implemented in different ways in various types of system. Figure 3 illustrates one possible implementation in a relational database system of the data structure shown in figure 2.

concept

ID	LANG	RESP
60201	da	HPL
60201	fr	HPL
60201	es	MMJ
60204	da	HPL
60204	fr	HPL
60204	es	MMJ

def

ID	LANG	DEF
60201	da	Juridisk er forsikring en aftale, hvor den ene part, forsikringsgiveren, forpligter sig til at udbetale en erstatning til den anden part, forsikringstageren, såfremt en af aftelen omfattet begivenhed indtræder. Som modydelse betaler forsikringstageren en præmie.
60201	fr	Une opération par laquelle une partie, l'assuré, se fait promettre, moyennant une rémunération, la prime, pour lui ou pour un tiers, en cas de réalisation d'un risque, une prestation par une autre partie, l'assureur, qui, prenant en charge un ensemble de risques, les compenses conformément aux lois de la statistique.
60204	da	Ved livsforsikring forstås dels en forsikring, hvor forsikringssummen udbetales ved eller en bestemt tid efter en persons død, og dels en forsikring, hvor summen udbetales i levende live, fx. ved opnåelse af en bestemt alder eller ved indgåelse af ægteskab.
60204	fr	Les assurances sur la vie sont déstinées à garantir, soit le risque de mort de la personne assurée (assurance en cas de décès), soit le risque de sa survie à une époque déterminée (assurance en cas de vie).
60204	es	El seguro sobre la vida comprenderá todas cas combinación que pueden hacerse, pactando entregas de primas o entrega de capital a cambio de disfrute de renta vitalicia o hasta cierta edad, o percibo de capitales al fallecimiento de persona cierta.

defref

ID	LANG	REF
60201	da	Bac p 9
60201	fr	BES p 2
60204	da	PL p 383
60204	fr	Bes p 32
60204	es	CC 416

term

ID	LANG	TNO	TERM
60201	da	1	forsikring
60201	fr	1	assurance
60201	es	1	seguro
60204	da	1	livsforsikring
60204	fr	1	assurance sur la vie
60204	fr	2	assurance-vie
60204	es	1	seguro sobre la vida
60204	es	2	seguro de vida

termref

ID	LANG	TNO	REF
60201	da	1	Bac p 9
60201	fr	1	Bes p 2
60201	es	1	MMJ
60204	da	1	PL p 383
60204	fr	1	Bes p 32
60204	fr	2	Vey II p 180
60204	es	1	CCE p 416
60204	es	2	MMJ

coll

ID	LANG	TNO	CNO	COLL
60201	da	1	1	tegne forsikring
60201	da	1	2	forsikringen dækker tab
60201	fr	1	1	contracter une assurance
60201	fr	1	2	conclure une assurance

collref

ID	LANG	TNO	CNO	REF
60201	da	1	1	PL p 126
60201	da	1	2	Bac p 11
60201	fr	1	1	Bes p 75
60201	fr	1	2	Bes p 75

Figure 3 — Tables in a relational database with sample data

In a relational database the terminological entry is split up into several records in various interconnected tables. As an example, the relationship between a concept and one or more synonymous terms is given by means of IDentification number and LANGuage. During retrieval, data elements connected to one terminological entry are linked together and presented as a unit.

In other systems, e.g. an information retrieval system, all data elements of one terminological entry are stored in one record. Regardless of the type of system used, it is very important for interactive retrieval, production of vocabularies and data exchange that each data element and its connections to other data elements can be identified separately. If this requirement is not met, it is not possible, for example, to specify user-group-specific search and presentation profiles (see 11.9).

8.3 Modifying the data structure

Although a prescribed entry structure is needed before a term bank can be set up, there should be facilities for making changes in the data structure at any time.

For instance, it should be possible to

- add a field;
- reorganize hierarchical structures;
- change the order of fields;
- subdivide or merge fields;
- change the data types of fields (e.g. integer, character, date).

EXAMPLE

In the first version of a terminological database, synonymous terms are classified as one term (TERM field) and one or more synonyms (SYNONYMS field). At a later stage it is decided to classify all synonymous terms as terms and delete the SYNONYM field (see 8.1).

Ideally, names of the fields should be mnemonic abbreviations such as TERM, DEF, REF, etc.

8.4 Quantitative requirements

Some database systems have quantitative restrictions which are unacceptable for terminology work, where the following conditions need to be satisfied:

- no limit to the number of terminological entries included in the database (in practice, a maximum of about one million will often be sufficient);
- no limit to the number of fields (data elements) per entry;
- no limit to the number of characters per field (field length).

A database system that allowed only fixed-length data fields would be inadequate, because terminological data often are of variable length and may include optional data elements.

9 Data input

Data may be entered by interactive or batch data input or by combinations of these.

9.1 Interactive data input

Most database systems have the capability for direct data input and allow updating and corrections in interactive mode. This normally means that corrected data are immediately available for retrieval purposes. This form of data input is practical only when limited volumes of data are to be entered. Otherwise, updating takes place at regular intervals.

9.2 Batch data input

Database systems will normally have a batch input capability to transfer externally created data to the database proper. Data are entered using an external data entry utility and transferred to the database as a batch process when a suitable amount of data has been entered. The external entry utility may, for instance, be a word-processing program.

Terminological data in machine-readable form and data that may be made machine-readable, e.g. by optical scanning, can normally be transferred to a terminological database. Such data usually have to be restructured in terms of record format and character set. The nature and extent of the restructuring will depend on the source data. If source data are printed dictionary data, the typeface and the punctuation have to be analysed to determine the corresponding categories. In some cases, very sophisticated and specialized parsing programs need to be developed.