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 Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.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 multi-functional. 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.

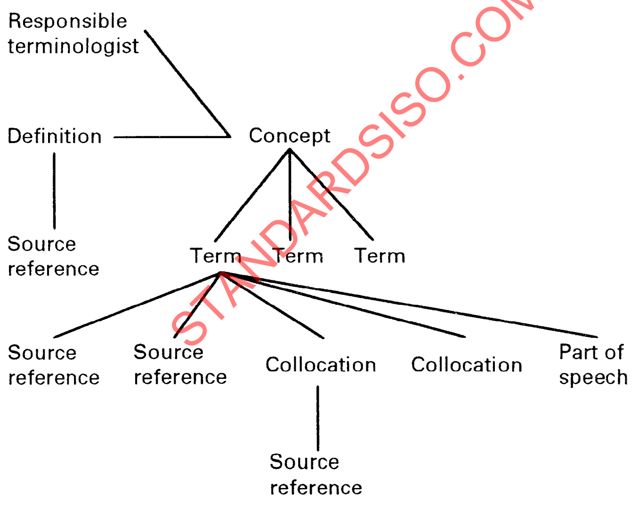


Figure 1 — Data elements in a monolingual entry with three synonymous terms

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.

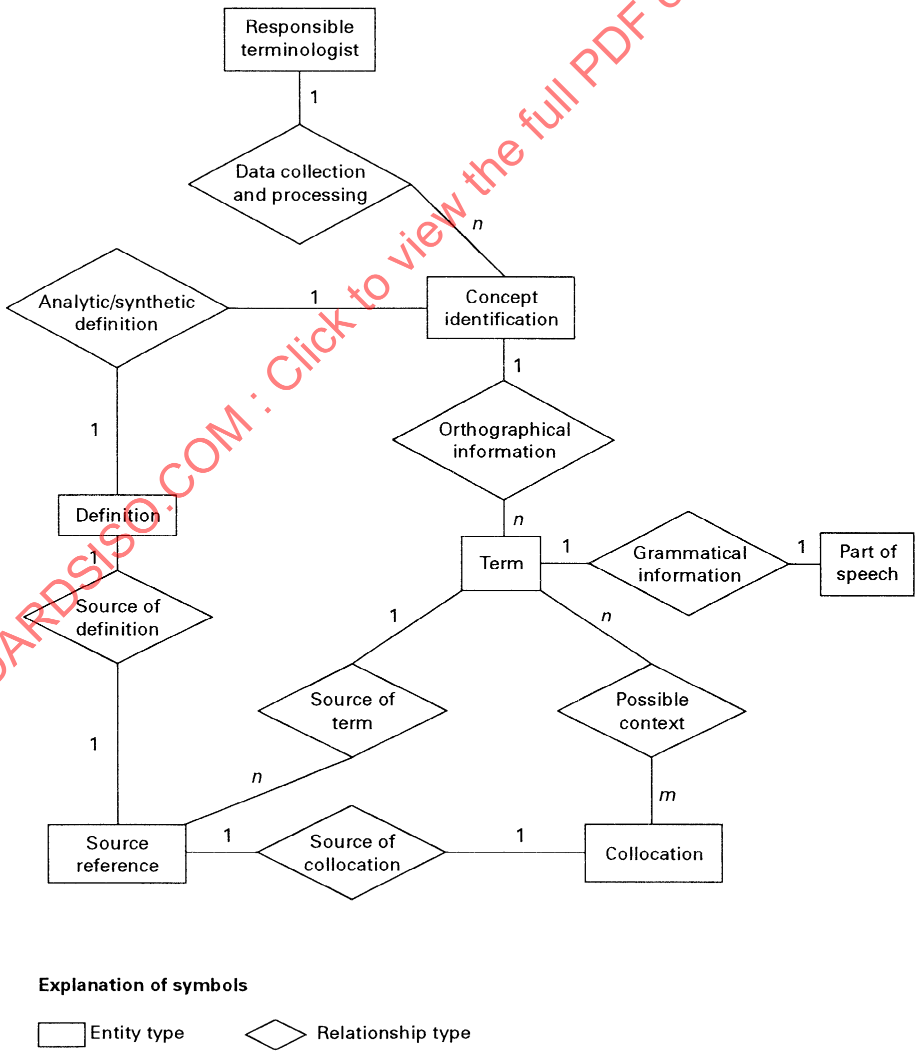


Figure 2 — Example of an entity-relationship diagram for a terminological database

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.

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

concept

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

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

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.

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.

9.3 Word-processing and checking utilities

Extensive word-processing capability should be provided for entering, editing, modifying and correcting data.

Data validation is needed in connection with data entry and editorial changes (for example, for checking on field names, field content and field order). Most database systems already offer these capabilities, but if special data entry programs are developed, it is advisable to provide a validation utility.

Examples of data validation:

- Double-entry check: a check to determine that an entry is stored only once.
- Consistency check: a check to determine that interdependent entries comply with predefined conditions.
- Spelling check: a check to determine that all entries comply with predefined spelling rules.
- Picture check: a check that uses a mask to verify the character types used in an input field.
- Completeness check: a check to see that data are present where data are required.
- Format check: a check to determine that data conform to a specified layout.
- Plausibility check: a check to determine that a value conforms to specified criteria.
- Validity check: a check based on known conditions or constraints that apply to a given piece of information or result.

9.4 (Semi)automatic generation or modification of field content

A distinction can be made between semiautomatic and fully automatic generation of field contents. Semiautomatic generation can be used, for example, for producing the full form of a source reference on the basis of an abbreviation entered during either data entry or the editing process.

Both field code and field content may be generated fully automatically. For example, a certain content may be entered automatically when the field code is entered or both field code and content (e.g. information on subject field) may be entered automatically in some or all records.

It should be possible to carry out changes in one record or to change a specified content of one field in all or selected records of the database. For example, it should be possible to change an abbreviation of a given subject field for all concepts belonging to this subject field.

10 Character set

The character set should be as open-ended as possible. Ideally, additions to the character set should be possible at any stage. For most terminological projects, special characters are needed that are not part of the standard set. Special characters should be directly accessible, uniquely stored and appear on-screen and in printouts without special coding.

In a terminological database there is a need not only for accented letters but also phonetic characters, the Greek alphabet, mathematical symbols and upper and lower indices.

It is advisable to structure the data so that font changes in a printed vocabulary can be made automatically and field-specifically. It should be possible to suppress font codes on-screen, and they should not affect data retrieval. It should also be possible to present data in different typefaces on screen and in printouts.

With non-Roman writing systems, automatic conversion (i.e. transcription or transliteration) of the contents of certain data fields may become necessary. Solutions must then be found, for example, for languages with 16-bit coded character sets such as Chinese and Japanese.

11 Data retrieval

Depending on the type of hardware, software and telecommunication capability available, the database may be accessed by one or more users simultaneously.

If possible, searches should not be restricted to a specified number of fields. During the creation and updating of a terminological database, the terminologist may need to search on certain information fields that end-users (e.g. translators) do not need to access.

Although everything should be retrievable in principle, in practice it is an advantage to be able to specify that some fields are not accessible for searching. For example, it should be possible to display and print out internal notes, but it may not be necessary to search for individual words in these notes. The purpose of restricted access is to save storage space and reduce response time.

Restricted access is especially important if the database is available for on-line data retrieval by external users. Differentiated access to data is then essential and can be achieved by means of read/write protection (see clause 14).

11.1 Interactive retrieval

Most database systems include facilities for interactive retrieval of data by means of a special query language. Searches may be command-driven, menu-driven, or form-driven (see 11.9 and 11.10).

Interactive retrieval of data from a terminological database often aims at the retrieval of individual terms and relevant information (equivalents, definitions, contexts, etc.). Searches for individual terms are often combined with searches on other criteria, such as language and subject field. The search term may be a single word or a compound term. Truncation or masking may be used to search for compounds, inflected forms, etc. (see 11.5). It is also possible to search a terminological database for all terms belonging to a certain subject field.

In both cases it is normally possible to specify different user profiles. The retrieved data may be displayed on the screen, printed out or stored in a special file.

11.2 Batch retrieval

Database systems normally feature batch retrieval capability. Batch retrieval may be sensible when so many terms are selected that on-line retrieval would block access for other users. The retrieved data are stored in a file, which may be printed out later.

11.3 Specifying a "stop-word" list

In a terminological database it may not be necessary to search for words like "of", "for" or "on". Therefore, it may be desirable to draw up a list of "trivial" words not subject to search. In this way, storage space can be saved and response times reduced. Ideally, it should be possible to define "stop-word" lists for particular fields, e.g. for all fields containing information in one language. This would obviate the problem of "homography" between languages. Defining the word "and" as a "stop-word" throughout a multilingual database would, for example, make it impossible to search for Danish "and" (duck).

11.4 Browsing

It should be possible to browse through the database without searching for anything in particular. It should be possible to define different browsing orders for the terminological entries, e.g. alphabetical or systematic orders. Browsing in a systematic order offers the user an overview of a certain subject field or parts of it.

11.5 Specified searching

For specified searches, it should be possible to define search profiles, i.e. specify that the search is to be restricted to certain fields. A term can be stored in various fields of many entries, e.g. in the definition or text fields describing related terms. One way to reduce the number of responses is, for example, to restrict the search to terms only.

A free text search capability is very useful, e.g. for searching on one or more words in a text field or for searching on part of compound terms. This kind of search is immediately possible in free text information retrieval systems, while some database management systems only offer this capability by virtue of truncation.

When searching for a compound term like "separate system" one may also want terms like "separate sewer system", where the words specified for the search are separated in the text. In a search for "information retrieval", responses may also be required where the words are both separated and in reverse order, e.g. "retrieval of information". Such searches are possible by means of logical operators (AND, OR) or positional operators.

In addition it should be possible to specify how far the words may be separated from one another, to avoid "noise" in the form of irrelevant responses. Ideally, a system should offer this facility as an option. Some systems allow this type of search by means of positional operators.

It should be possible to carry out searches for the first, last or a randomly positioned element in a word (right/left truncation or masking), e.g. searches for first elements in compounds, endings or particular derivational forms.

It should be possible to retrieve all entries containing one or more given data elements; entries containing definitions, internal editorial notes, etc.

Searches with several criteria that employ logical operators (AND, OR, NOT) may be used for the retrieval of a specific term belonging to a particular subject classification. In the production of a printed dictionary, this capability may be used to exclude all terms that are not genuine terms, but rather proposed translations.

11.6 Ignoring uppercase and lowercase, hyphens and diacritics in searches

Ignoring uppercase and lowercase, hyphens and diacritics may be necessary when searching in texts. Otherwise it is impossible to search for identical words that begin with upper- or lowercase or contain hyphens or accents. On the other hand, the opposite capability, the ability to distinguish between, for example, uppercase and lowercase, should also be available.

11.7 Information on the number of responses to a search

The number of responses to a search is valuable statistical information to a user searching for a specific term and its equivalent in a given language. For example, a search with "charge" or "absorption" may produce many responses because the words are common to many subject fields and exist in several languages. It may not be of interest to study all the responses. Instead, a more specific search can be reformulated and limited to a particular subject classification, e.g. with "absorption" as headword and "nuclear energy" as subject classification. It should be possible to issue a new query using the results of the first query.

11.8 Presentation

Presentation includes on-screen display, hardcopy and various printouts on paper, e.g. dictionary printout.

In some systems, search and presentation commands are combined. In other systems, it is possible first to search and then to choose presentation of all or some of the retrieved entries on the basis of the number of hits. The separation of search and presentation is an advantage when searching in a terminological database.

It should be possible to define presentation profiles involving one or more fields (selective presentation).

It should be possible to configure for presentation with or without displayed field names and additional text. Some systems will show the data in the typographical form of a printed glossary.

The hierarchical structure of terminological entries should be represented in the system to enable presentation which reflects the hierarchical structure. The links between terms, contexts and sources should be clearly marked on screen displays and in printouts of terminological entries. It should also be possible to specify that given data elements always appear together with other specified elements.

11.9 Reuse of search command and presentation profiles in a query language

Searches in a term bank can be facilitated by using predefined search and presentation profiles. The macro capability provided by many database systems offers the option of storing frequently used and complex commands as simple commands. A presentation macro can contain a list of field names, e.g.

```
FR_DEF, FR_TERM, FR_REF, FR_TEXT,
EN_DEF, EN_TERM, EN_REF, EN_TEXT
```

saving the user from typing all these field names on each selective presentation.

It is also possible to define macros containing search commands, where the object of the search can be respecified each time while the search profile remains the same. Macros can also prompt the user, e.g. "Enter subject area" and "Enter search term".

In some systems it is also possible to combine search and presentation commands in one multiple-command macro.

11.10 Customized menus and special search forms

With many database systems today the user can search by choosing from a menu. A menu system provides a number of search and presentation options. However, the system should allow for both "menu-driven" and "command-driven" searches and combinations thereof.

A menu system for a terminological database could be designed so that the following options are available:

- subject;
- source language;
- target language;
- search profile and presentation profile;
- search with truncation, masking, or format, etc.

Another possibility is to set up a special form containing headings or labels corresponding to the fields of the database. The form, which may consist of several pages, is presented on the screen and the user simply fills in the search criteria, e.g. language, subject field, search term; and the result of a search is presented in the same form. This type of "form-driven" search may also be combined with a "menu-driven" search.

11.11 Statistical functions

It should be possible to obtain information on the frequency of words for the database as a whole, or for parts of the database or for specific text segments.

EXAMPLE

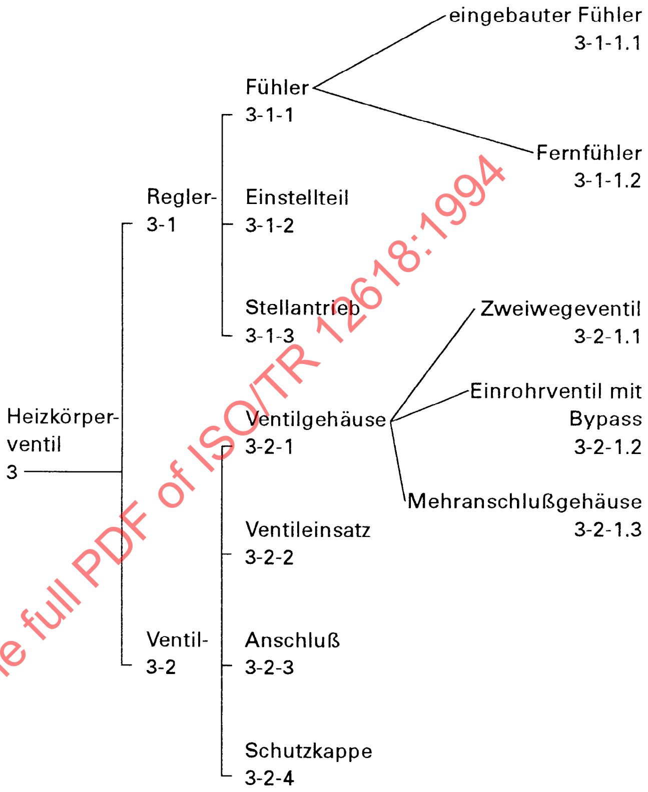
> In order to define a "stop-word" list, it may be desirable to have a frequency or rank list for words in the whole database or certain fields in the database. Alphabetical frequency lists may also be useful for identifying spelling errors.

12 Sorting

In principle, it should be possible to sort the responses in a given order and produce sorted printouts from them. Further, it should be possible to sort on the content of all information types. In practice, however, it is rarely relevant to sort items such as internal editorial notes.

For instance, it should be possible to produce either alphabetic or systematic lists. Systematic lists can be sorted according to the position of the concept in a system of concepts.

Figure 4 shows an example of a partial system of concepts for radiator thermostats and a systematic list produced on the basis of positional information.



3	Heizkörperventil
3-1	Regler
3-1-1	Fühler
3-1-1.1	eingebauter Fühler
3-1-1.2	Fernfühler
3-1-2	Einstellteil
3-1-3	Stellantrieb
3-2	Ventil
3-2-1	Ventilgehäuse
3-2-1.1	Zweiwegeventil
3-2-1.2	Einrohrventil mit Bypass
3-2-1.3	Mehranschlußgehäuse
3-2-2	Ventileinsatz
3-2-3	Anschluß
3-2-4	Schutzkappe

Figure 4 — Partial system of concepts and corresponding systematic list

The user will understand the structure of the subject field and the meaning of the terms better if definitions, notes and information on equivalence are printed out in addition to terms.

It should be possible to sort according to several criteria, e.g. firstly on the basis of subject and secondly alphabetically.

When working with multilingual data, different sorting orders are needed, e.g. for the German and Swedish letter ä (see ISO/TR 8393):

German	**Swedish**
. . .	. . .
. . .	. . .
. . .	. . .
Ahne	hydrofor
ähneln	hårdhet
ahnen	hävert
. . .	högzon
. . .	. . .

13 Production of printouts and printed vocabularies

All data or subsets of data in a terminological database may be used for the production of printouts and printed vocabularies. Many database systems offer special report writing facilities with sophisticated layout options.

Unless SGML tags (see ISO 8879) are used, typographical codes should be inserted into the retrieved data for computer-typesetting. Such codes can normally be inserted automatically according to the data fields (see ISO 12200). It is possible to use the same database for the production of dictionaries with a different number of languages, subject fields, presentation profile, order of entries, layout and typography.

14 Data protection

It is desirable to define differentiated access profiles for different users. Ideally, it should be possible to establish user-based read/write protection at the

— database access level;
— record level;
— field level;
— character level.

EXAMPLES

Database access level: Only users who know the password for the database can gain access.

Record level: Certain users (e.g. external users) should not have access to records still in the process of revision; some users should be allowed to use, but not alter, the data.

Field level: External users should not have access to fields with internal editorial notes.

Character level: Typographical codes should be suppressed if the database is used for on-line retrieval by external users.

It may be advisable to prevent unauthorized downloading.

15 Data transfer

Authorized personnel should be able to copy the entire database or parts of it to external files so that the data can be transferred to another database.

It should be possible to import external data into the database in machine-readable, structured form. These data can be extracted from another database or data stored in files with data elements marked by means of a special data entry system or using an ordinary word-processing system (see 9.2). Ideally it should be possible to import both whole records and parts of records.

Importing data from external files becomes necessary during data exchange or when hardware or software used for a terminological database is replaced by another type of hardware or software.

For data exchange between terminological databases, it is necessary to classify and describe the content and to describe the structure of the terminological entries of the databases (see clauses 7 and 8).

Data exchange requires a description of the storage medium used (tape, diskette, etc.), of file and record structure and of the character set or sets (see ISO 6156 and ISO 12200).

16 Feedback from users

Well-established user-contacts are very important. A mailbox facility allows the users to send any questions or suggestions to the data bank supplier. This mailbox facility may be a part of the data bank, or it may be a general mailbox system. For standalone microcomputers, a feedback procedure should be implemented.

17 Maintenance and updating

Running a terminological database involves two separate processes: system maintenance and data updating. It is crucial that the responsiblities for maintenance and updating be clearly defined. Only authorized persons should be allowed to modify the data.

17.1 System maintenance

Database software is likely to undergo further developments. New versions may improve performance or reliability, make the system more user-friendly or include new features.

Most database systems include some options that may be specified by the user or data supplier. These options may include terminal screen layout, printer configuration, indexing, contents control, etc. New data or new uses may dictate changes in these specifications. Depending on the nature of the modifications and the database system, it may be necessary to reinstall the entire database or parts of it. Whenever such modifications take place, it is necessary to modify all copies of the database system.

17.2 Data updating

Terminological data have to be updated to reflect technical and scientific development and changing usage. This may be done interactively or in batch depending on the database system and volume of changes. Software tools exist and are being developed that detect possible errors in the database (see 9.3).

18 Portability

In addition to importing or exporting data, it is also important that the system used (e.g. information retrieval or database management system) can be transported from one machine to another. The system may have to be moved to a smaller or bigger version of the same computer make or from one make to another.

Many commercially available systems run on different versions and makes of computers. Sometimes it is possible just to upgrade the system when changing to a bigger version of the same computer. In most cases, however, it is necessary to buy a new licence. This is also the case if one changes over to another type of computer.

The advantage of using the same system in another computer is that no restructuring of data is needed. Restructuring is always necessary when the system is changed. A change of system often means that the user (terminologist, translator, etc.) has to learn a new query language. There are standardization efforts within this field, e.g. CCL (Common Command Language — see ISO 8777), which is used in connection with information retrieval systems, and SQL (Structured Query Language — see ISO/IEC 9075), which is used in many database management systems.

19 Data communication

The data communications available play an important role at all stages of preparation and use of terminological data collections.

Data communication comprises direct connections (terminals or microcomputers connected to a central computer), local area networks (LAN), dial-up connections via telephone and modem to another central computer or host, or communication between two computers (hosts) via national or international networks.

20 Creation and use of a text corpus

20.1 Creating a machine-readable text corpus

20.1.1 Introduction

Terminology work may necessitate the creation and use of a text corpus. In terminology work a corpus may be used for the following purposes:

— to identify candidate terms;

— to examine the use of terms in context;

— to retrieve implicit definitions.

With appropriate software, such as text analysis and retrieval packages, it is possible to obtain word counts, indexes, concordances and statistical information from any text in the corpus.

20.1.2 Building up a text corpus

There are several ways of building up a text corpus. The most common ways include

- using an optical scanner (OCR — Optical Character Recognition);
- keyboarding the text in the traditional way; or
- using texts already in machine-readable form (e.g. printers' textbases).

In each case, the advantages and disadvantages of the method must be carefully evaluated in order to identify the most cost-effective method for the specific application.

In most cases it is necessary to proof-read the final version of the text, regardless of the method used.

20.1.3 Corpus size and structure

The size and content of the text corpus have to be defined separately for each terminological project (see also ISO 10241). In addition to the text itself, each corpus should include at least bibliographical information, classification of text types and subject fields.

Ideally, a corpus can be used for several different applications. Therefore, during the encoding process, layout features of the original text should be preserved wherever possible. In addition to linguistic information, it is useful to record layout information and typographical information in the database in such a way that any combination of data elements can be searched for. All recorded information should be described in a data exchange format to allow exchange of data and further software development. Corpus maintenance forms an integral part of corpus management. A strict procedure should therefore be worked out for alterations and exchanges of corpus material. All operations should be fully documented.

20.1.4 Pre- and post-editing

Even when texts are available in machine-readable form, they normally need a great deal of editing before they are ready to be included in the text corpus. They have to be pre-edited to conform to a uniform standard (preferably a common standard). They also have to be typographically disambiguated, e.g. by using well-defined delimiters for such structural elements as words and sentences.

A text corpus can also be post-edited to give additional systematized information, e.g. on word-classes, grammatical function, etc. This process (often called tagging) consists of adding one or more identifiers to the relevant data elements, e.g. words or phrases.

Elements can be tagged manually, automatically or, in most cases, semi-automatically. Tagging is a complex process and needs thorough planning. The documentation on the tagging conventions used can serve both as a supplementary guide to the user of the tagged corpus and later also as a basis for revisions.

20.2 Examples of the use of a text corpus

20.2.1 Production of lists

This section introduces a number of lists produced from a text corpus.

The lists in 20.2.1.1 to 20.2.1.3 give different types of information about the lexical material in the corpus:

a) **an index** to all occurrences (tokens) in the corpus, in alphabetical order and supplied with a text reference;

b) **concordances**, listing the words in the corpus in alphabetical order:

   1) in the context of a computer text line (KWIC-index);

   2) in sentence context;

c) lists supplying the user with **statistics** about the corpus.

NOTE 9 The examples come from an existing machine-readable text corpus consisting of text extracts from the *Rules of the Supreme Court* (34 760 tokens and 3 170 types).

20.2.1.1 Index

A corpus index contains all the types in the corpus in alphabetical order and with a text reference. Figure 5 gives an extract from an index referring to the number of the record in which the token occurs.

claim	302002 : 302002 : 302002 : 302002 : 302002 302002 : 302002 : 302007 : 302007 : 302007 302007 : 302007 : 302015 : 302020 : 302020
[. . .]	
claimant	302115 : 302115 : 302115
claimants	302099 : 302099 : 302115
claimed	302002 : 302002 : 302004 : 302020 : 302021 302022 : 302024 : 302033 : 302035 : 302051
[. . .]	
class	302113 : 302139
classes	302065 : 302066
clear	302021 : 302093 : 302095 : 302095 : 302112
clearly	302024

Figure 5 — Extract from index of tokens referring to record numbers

20.2.1.2 Concordances

1) Figure 6 shows an extract from a KWIC-index of the token "claim" sorted alphabetically to the right of the token.

RSC302020	(a) an action which includes a **claim** by the plaintiff
RSC302020	(b) an action which includes a **claim** by the plaintiff
RSC302056	(b) a **claim** by the plaintiff based on an allegation of
RSC302020	a statement of **claim** has been served on a defendant and
RSC302026	may direct that the **claim** in question and any other claim
CCR302161	(2) If the plaintiff's **claim** is for unliquidated
RSC302038	after the statement of **claim** is served on him, whichever

Figure 6 — KWIC-Index of a token

2) Figure 7 shows an extract of a sentence concordance using "statement of claim" as keyword, the same reference location as in figures 5 and 6 and in addition the number of the sentence in that particular record.

(2)

RSC302020	1 . - (1) Where in an action to which this rule applies
RSC302020	a STATEMENT OF CLAIM has been served on a defendant and
RSC302020	that defendant has given notice of intention to defend
RSC302020	the action, the plaintiff may, on the ground that that
RSC302020	defendant has no defence to a claim included in the writ
RSC302020	or to a particular part of such a claim, or has no defence
RSC302020	to such a claim or part except as to the amount of any
RSC302020	damages claimed, apply to the Court for judgement against
RSC302020	that defendant.

(2)

RSC302038	2 . - (1) Subject to paragraph (2) a defendant who gives
RSC302038	notice of intention to defend an action must, unless the
RSC302038	Court gives leave to the contrary, serve a defence on
RSC302038	the plaintiff before the expiration of 14 days after the
RSC302038	time limited for acknowledging service of the writ or
RSC302038	after the STATEMENT OF CLAIM is served on him, whichever
RSC302038	is the later.

Figure 7 — Sentence concordance

Figure 8 shows an extract from an alphabetical list indicating the frequency of occurrence of each type and the rate of occurrence in % in the text corpus.

city	4	0,0103
civil	10	0,0257
claim	119	0,3054
claimant	3	0,0077
claimants	3	0,0077

Figure 8 — Index of frequency of occurrence

20.2.1.3 Statistics

It is possible to produce a list summing up the different explicit and implicit information as in figure 9. The seven columns represent

a) the number of word types counted;

b) the rank of the specific occurrence, giving the most frequent occurrence rank number 1, the next most frequent rank number 2, and so on, giving words with the same number of occurrences the same rank number;

c) the word (the type);

d) the number of occurrences (the tokens);

e) the percentage value for the tokens proportional to the total amount of tokens in the corpus;

f) the percentage value summed up when adding the types in the list;

g) the number of occurrences summed up when adding the occurrences in the list.

a)	b)	c)	d)	e)	f)	g)
1	1	the	2 947	7,563 0	7,563 0	2 947
2	2	of	1 706	4,378 2	11,941 2	4 653
3	3	to	1 173	3,010 3	14,951 5	5 826
4	4	or	1 144	2,935 9	17,887 4	6 970
5	5	a	969	2,486 8	20,374 2	7 939
6	6	in	878	2,253 2	22,627 4	8 817
7	7	be	634	1,627 1	24,254 5	9 451
8	8	and	554	1,421 8	25,676 3	10 005
		[. . .]				
32	32	where	183	0,469 6	44,751 8	17 438
33	33	defendant	177	0,454 2	45,206 0	17 615
34	33	it	177	0,454 2	45,660 2	17 792
35	34	application	169	0,433 7	46,0939	17 961
36	35	such	168	0,431 1	46,525 0	18 129
37	36	notice	166	0,426 0	46,951 0	18 295
38	37	service	157	0,402 9	47,353 9	18 452
39	38	this	150	0,385 0	47,738 9	18 602
40	39	other	148	0,379 8	48,118 7	18 750

Figure 9 — Evaluation chart