

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)

AVA **Documents**
Document Sharing Hub

Contents

	Page
Foreword	iii
Introduction	iv
1 Scope	1
2 References	1
3 Definitions	1
4 Types of terminological data collections	2
5 Criteria for creating a terminological database	2
6 Hardware and software requirements	3
7 Terminological data categories	3
8 Data structure	4
9 Data input	7
10 Character set	8
11 Data retrieval	8
12 Sorting	11
13 Production of printouts and printed vocabularies	12
14 Data protection	12
15 Data transfer	12
16 Feedback from users	12
17 Maintenance and updating	13
18 Portability	13
19 Data communication	13
20 Creation and use of a text corpus	13

© ISO 1994
All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the publisher.

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

Printed in Switzerland

Foreword

ISO (the International Organization for Standardization) is a world-wide 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 recordset 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, print-outs (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 unambiguously 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.