

Technical Dictionaries Retrieved from a Database

Otto Vollnhals

Volume 27, numéro 2, juin 1982

URI : <https://id.erudit.org/iderudit/004577ar>

DOI : <https://doi.org/10.7202/004577ar>

[Aller au sommaire du numéro](#)

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

1492-1421 (numérique)

[Découvrir la revue](#)

Citer cet article

Vollnhals, O. (1982). Technical Dictionaries Retrieved from a Database. *Meta*, 27(2), 157–166. <https://doi.org/10.7202/004577ar>

Technical Dictionaries Retrieved from a Database

OTTO VOLLNHALS

INTRODUCTION

One side effect of the rapid advance of science and technology is an equally swift development of technical language, marked by the continuous coinage of new terms. This and the increasing trend towards the international exchange of information are two reasons why up-to-date dictionaries, monolingual and interlingual, are more essential today than ever before.

Yet in spite of their importance, most technical dictionaries are still produced in the same way now as they were in the "good old days": Vocabulary is collected, recorded on file cards, and arranged neatly in alphabetical order; it is combined to create entries, rewritten in manuscript form, and otherwise processed by hand until ready for typesetting, after which galley proofs and, finally, page proofs have to be read before printing can begin. Aside from demanding expensive manpower at every stage, this procedure takes years to complete and results in correspondingly long intervals between new editions. In other words, it is wasteful of time and money.

Obviously, dictionaries produced in this manner can never be truly up to date, which means their value as aids to the understanding of contemporary technical and scientific literature is appreciably diminished.

More than a decade ago, members of the Language Services Department of Siemens AG recognized that most of the procedures followed in producing dictionaries, from recording terminology to setting type, could be automated with the help of electronic data processing and computer-controlled typesetting.

Accordingly, at the start of the 70s work was begun on the design of software with this aim in view. The resulting program system was christened TEAM, at once an acronym for Terminology Evaluation and Management, and a symbol of the close cooperation exercised by terminologists, lexicographers, programmers, engineers, and translators in developing the system.

The intellectual tasks entailed in dictionary making, such as researching the meaning of terms and matching terms in different languages, remain the province of man, but even here the machine can do much to speed and facilitate work. Compiling, comparing, and systematizing vocabulary; correcting and supple-

menting entries; building terminology nests¹; producing discussion manuscripts; sorting and inverting entries (e.g., changing French-English into English-French); eliminating double entries; inserting synonym references — all are tasks that can be done by computer.

1. TERMINOLOGY DATABASE

The basic idea of the TEAM Program System is to provide a comprehensive, computer-supported dictionary — a “terminology database” — which contains technical terminology in several languages as well as auxiliary information, such as definitions, synonyms, and labels indicating the source of terms and the subject fields to which they belong. The TEAM System serves first and foremost as an aid to the translating activities of Siemens AG.

All terms in the several languages that go to make up one entry in the computer dictionary, together with all supplemental data pertaining to them, are combined to form a single unit. This unit is then subdivided into information categories, so that each meaningful segment of information — that is, as a rule, the contents of each information category — can be accessed separately.

The computer dictionary can be interrogated directly in the interactive mode via data display or teleprinter terminals, or it can be used as the source of printed word lists and dictionaries. In the latter case, the vocabulary conforming with the specified criteria is selected by the computer from the total inventory of terms, is processed by a qualified person, and is then output by high-speed printer.

For direct interrogation of the dictionary over a data display or teleprinter terminal, a storage with random access is used. In the TEAM System, this is a magnetic disk storage. For other purposes, however, use is generally made of magnetic tape at present. Terms can be selected from the contents of the computer-stored dictionary, or “terminology database in any quantity, any sequence and any combination, and reproduced by means of peripheral equipment.

2. DICTIONARY PROGRAMS

Within the system are other data pools, independent of the internal database used by Siemens for translation purposes, which are also administered with the help of TEAM. These pools contain terminology in one, two, or more languages and are used to produce dictionaries for various publishing houses.

For dictionary projects, the word material gathered by the author — whether provided in the form of a manuscript or a loose collection of unsorted file cards — is recorded on data media. During the recording process, every entry is broken down into individually addressable categories of information.

2.1 *Input media*

Currently, our primary input medium is OCR-B sheets, which can be produced on a standard electric typewriter equipped with OCR-B-font. To some ex-

1. A nest is part of a dictionary entry made up of compound words and phrases containing the entry keyword; these compound expressions are recorded alphabetically with a tilde inserted in place of the keyword itself.

tent, data is also recorded on-line, using display terminals. In both cases, normal spelling and punctuation are employed, including upper and lower case letters, umlauts, and diacritics.

2.2 *Input format*

The form and sequence in which data and data identification codes are entered in the computer is called the input format.

For the present, an entry can consist of a maximum of 100 items, consecutively numbered from 00 to 99. These numbers are the "addresses" of the various categories of information contained in the entry. They are assigned to clearly defined types of information and are used as needed.

Entered under code 00 is the sequence number, or the "address" of the entry as a whole.

The digits occupying the ten's place in the code identify the language; they are assigned as follows:

10	German	50	Russian
20	English	60	Italian
30	French	70	Portuguese
40	Spanish	80	Dutch

The digits in the unit's place contain the following information for the denoted language:

0	Term	5	Context
1	Part of speech	6	Synonym(s)
2	Source of term	7	Quasi synonym(s)
4	Definition of term	8	Subject field labels

Information that is relevant to all the languages contained in a multilingual entry is shown in categories 01 to 09 at the top of the entry. Information that applies to an entire series of entries can likewise be coded in this way and placed at the beginning of the series; thus it has to be recorded only once.

00	CA2317
03	d
04	1179
05	0201
06	E37
07	ZFE
10	farbmetrische Verzerrung
12	ZFE
14	Die Änderung der Farbvalenz einer Körperfarbe bei Änderung der beleuchtenden Lichtart.
16	Farbverzerrung;f.
20	colorimetric shift
22	ZFE
24	The change in chromaticity and luminance factor of an object colour due to change of the illuminant.
30	distorsion colorimétrique
32	ZFE
34	Changement de la chromaticité et du facteur de luminance d'une couleur de surface dû au changement d'illuminant.
50	kolorimetričeskij sdvig
52	ZFE
54	Izmenenie cvetnosti i koeficienta jarkosti objekta, vyzvanoe izmeneniem spektralnogo sostava izlučeniya.
99fa	

Terms in languages using non-Roman alphabets are transliterated into Roman characters when they are recorded. Since this is done in accordance with the fully reversible transliteration scheme developed by ISO, the terms can be reconverted and output by the computer in their original form, provided the output unit is equipped with the proper character set.

Seitenzähler	84
Seitenzähler <i>m</i> COBOL / счётчик страниц / page counter Sektor <i>m</i> Mpl. Sw / сектор <i>n</i> / sector <i>n</i> Sektoradresse <i>f</i> ProzRe. Sw / адрес сектора / sector address Sekundärspeicher <i>m</i> DigDV / память второго уровня, внешняя память / secondary storage selbstdefinierender Wert Sw / самоопределяющаяся величина / self-defining value, self-defining term selbstkorrigierender Code Sich. DÜ / корректирующий код / error-correcting code, error-correction code, self-correcting code selbstladend <i>adj</i> Sw / самозагружающийся / self-loading <i>adj</i> selbstladender Speicherauszug Sw / самозагружающаяся распечатка памяти / self-loading memory print (SLMP) Selbstorganisation <i>f</i> Kyb / самоорганизация <i>f</i> / self-organization <i>n</i> selbstprüfender Code Sich. DÜ / код, обнаруживающий ошибки / error-detecting code, self-checking code Selektion <i>f</i> Lk / выбор <i>m</i>, селекция <i>f</i> / selection <i>n</i> Selektor <i>m</i> Lk / селектор <i>m</i> / selector <i>n</i> Selektorkanal <i>m</i> ZE / селекторный канал / selector channel Semantik <i>f</i> MaÜbers / семантика <i>f</i> / semantics <i>pl</i> semantische Information MaÜbers, Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information semantischer Gehalt eines Zeichens Kyb / значение <i>n</i> / meaning <i>n</i>: - Informationsbetrag MaÜbers, Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information; - Informationsgehalt MaÜbers, Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information Semiotik <i>f</i> MaÜbers, Kyb / семиотика <i>f</i> / semiotics <i>pl</i> sequentiell, indizierte -e	Zugriffsmethode Sw / метод индексно-последовательного доступа / indexed sequential access method (ISAM) sequentieller Zugriff / последовательное обращение, последовательный доступ / sequential access seriell <i>adj</i> Sw / последовательный <i>adj</i>, серийный <i>adj</i> / serial <i>adj</i> Serienbetrieb <i>m</i> DigDV, DÜ / последовательный режим работы / serial operation Serien-Parallelbetrieb <i>m</i> DÜ / последовательно-параллельный режим / series parallel operation Serien-Parallel-Digitalrechner <i>m</i> DigDV / цифровая вычислительная машина параллельно-последовательного действия / parallel series computer Serien-Parallelsystem <i>n</i> DÜ / последовательно-параллельная система / series parallel system Serien-Parallelumsetzung <i>f</i> DÜ / последовательно-параллельное преобразование / series parallel conversion Serienrechner <i>m</i> DigDV / последовательная ЭВМ, цифровая вычислительная машина последовательного действия, машина последовательного действия / serial computer Seriensystem <i>n</i> DÜ / последовательная система / serial operation Servomechanismus <i>m</i> DigDV / серводвигатель <i>m</i> / servomechanism <i>n</i>, servo <i>n</i> Servosystem <i>n</i> ProzRe, NumSt / сервосистема <i>f</i> / servo system Setzen Programmaske Befehl / ввод программной маски / Set Program Mask; - Speicherschlüssel Befehl / ввод ключа защиты памяти / Set Storage Key Setzmaschine <i>f</i> Satz, Info / наборная машина / typesetting machine Sheffer-Funktion <i>f</i> Boole / штрих Шеффера, функция Шеффера / Sheffer's stroke, Sheffer function, Sheffer stroke function Sheffersche Funktion Boole / штрих Шеффера, функция Шеффера / Sheffer's stroke, Sheffer function,

After data has been recorded, the input data log is proof-read. The system is designed to allow corrections at any time. If the typist is aware of having made a mistake while keying in data, she can either correct it in the entry itself, or she can make a separate correction entry and key it into the system at random. Corrections and additions which proofreading of the input log has shown to be necessary, or which are made at a later date, can likewise be entered in the system at any spot, since the computer automatically inserts them in the proper place.

Whenever a correction is made, only the affected category of information is altered, and, usually, only a portion of that. The remaining data is left strictly alone. Normally this prevents the commission of new errors while correcting existing ones.

3. SUPPLEMENTAL PROGRAMS

An entire series of programs is available for processing data further.

3.1 *Synonym entries*

With the help of a supplemental program, the synonyms recorded for stored terms can be used to automatically generate new entries. This greatly reduces the workload at the data input end.

3.2 *Transposed entries*

Using another program, it is possible to reduce compound expressions to their component parts and rearrange the parts in any order desired. For example, if so instructed the program will change an adjective-noun sequence to noun-adjective. Thus "hybrid computer" becomes "computer, hybrid". The term can then appear in the dictionary under either "hybrid" or "computer" or both — an advantage of inestimable value for the automatic generation of indexes. Longer phrases, too, can be treated in this manner.

3.3 *Merging double entries*

When data volumes are large it can happen that the same term is recorded twice, or that two nearly identical terms are recorded. If certain criteria are met — ideally, if there is complete agreement between the terms in the source and the target language — these so-called doublet entries are combined by the program to form a single entry. In the process, all other information of interest, such as data on source and subject field, is retained in full.

4. SELECTION OPTIONS

On the basis of numerous criteria, terms can be selected from the data bank in any quantity and order, and with practically any combination of elements. For instance, selection can be made by requesting:

- a specific combination of languages
- a specific subject field or fields
- a particular source
- a certain part of speech
- a particular entry date (as when terminology is being checked for currency)

- a particular class of usage (preferred, accepted, unacceptable)
- terms with definitions and/or context samples
- a particular technical system or type of equipment
- concordances.

Selection criteria can be applied singly or in combination. All entries that meet the specified criteria are selected from the total inventory, and either output or stored for further processing. All unwanted portions of information contained in the selected entries can be eliminated by a different program. Selection programs are very useful as well for producing condensed versions of larger dictionaries.

5. ERROR SEARCH

Selection and concordance programs are also excellent tools for searching out errors, since they offer countless possibilities for combining elements. In this application they far outmatch the performance of human beings, not only by their enormous speed, but also by their uncompromising precision.

For example, for the new edition of a large technical dictionary with 150 000 entries it was necessary to change the spelling of certain words in the German portion to conform with new rules. Thus "oxyd" had to be made into "oxid"; "jod" into "iod," etc. This was accomplished by using the programs to locate all such character strings and alter them accordingly. Of course, revisions like this cannot be left solely to the computer to perform, but while subsequent human checking is imperative, the time involved is incalculably less than would otherwise be required.

A further example: Using a selection program, frequency statistics were compiled for the keywords of a large bilingual dictionary with French as the source language. This was done shortly before the start of typesetting to make sure that entries beginning with words like "électricité" were recorded with properly placed diacritics throughout the entire work. On the basis of frequency statistics, corrections could be made very quickly, for although a word may occur a thousand or more times in correct form, any wrong version is likely to crop up only two or three times at most, and is precisely pinpointed by the program. Error elimination of this sort is vital, because any variation in the way a keyword is written causes the typesetting program to interpret it as a new word and to therefore cease its "nest building" and go on instead to form a new entry.

6. CHANGES AND NEW ENTRIES

Right up to the starting gun for typesetting, so to speak, the author can modify entries stored in the computer. Certain senses, subject fields, examples of use, etc., can be altered, added, or deleted as he pleases. He is also free to introduce any number of new and exceptionally up-to-date terms without upsetting either the alphabetical order or the printing format. In a matter of minutes the computer integrates all changes and additions into the work and suitably rearranges the surrounding material. The danger of perpetrating new errors is slight, since nothing is touched but the affected passages.

7. MANIFOLD OUTPUT POSSIBILITIES

The TEAM System is also extremely flexible at the output end. Using available programs and program variants, it can work with output formats and equipment of many descriptions. The high-speed printer is especially suitable for producing working and discussion manuscripts and short-lived word lists like those prepared for specific translation projects. A number of programs can be used for this purpose, among them, one designed for outputting lists with the high-speed library printer, a unit specially developed by Siemens for producing documentation and equipped with diacritics as well as upper and lower case font. More recently, it has also become possible to produce lists using the modern laser printer, which can be programmed for any alphabet or set of characters. All programs allow the suppression of any information in the entries whose output is not wanted.

8. PHOTOCOMPOSITION

The selected entries, stored in the desired order on magnetic tape, are prepared for typesetting and then re-transcribed to tape. In this way a data carrier is created that includes, besides the sheer contents of the dictionary, all the control characters required for electronic typesetting.

When a dictionary is produced on order for a publishing house, the control characters are parameterized to accord with the publisher's wishes. The computer program assigns the control characters to the dictionary text, and transcribes them together with the text on the input tape for the photocomposition unit; control characters for vertical spacing and page option are then added. Running heads and page numbers are likewise generated by the program, which also has a special segment for performing tasks such as converting transliterated Russian texts back to the Cyrillic alphabet.

The printing format can be varied in any way designated by the author or publisher. By altering the parameters, the sequence and configuration of the terms and supplementary information in the entries are correspondingly changed.

The program then automatically constructs lines, columns, and pages, carries out pagination, and generates the running heads and initials at the top of the alphabet sections.

Systematized multilingual dictionaries often feature indexes listing the pages on which terms can be found; these indexes, too, can be generated automatically by the program.

With electronic typesetting, in contrast to conventional methods, typographical errors are totally ruled out. No changes of spelling, punctuation, or alphabetical order can occur, and nothing can be "forgotten." The text, automatically reproduced on film in page makeup, thus requires no further proofreading. It is advisable, however, to glance through the film to be sure no damage has been caused by faulty exposure.

9. MULTIPLE UTILIZATION OF DATA

Unlike other so-called computer typesetting systems, TEAM was designed from the outset for making multiple use of stored data. Multilingual TEAM en-

tries are stored in data pools in a form so neutral that any language can serve as either source or target language. Moreover, a given language can be combined with any number of target languages. This has two outstanding advantages :

a) **Inversion.** If the companion volume of a bilingual dictionary is needed — say, the English-French counterpart of a French-English work — the computer can quickly invert all vocabulary in the existing volume and arrange it alphabetically in the new source language. Even though extensive human processing is still required, the author is completely spared a particularly tedious and time-consuming task.

b) **Creation of new language combinations.** If a publisher wishes to expand a series of dictionaries — e.g. by adding a French-English volume, after publishing German-English and German-French versions — this too can be done by automatically calling up the relevant terminology in the desired configuration from the data pool².

In each case, not only is a huge amount of time saved (also adding to the dictionary's currency), but, what is possibly more important to publishers, relatively lower costs are incurred, since the material for the new work already exists in machine-readable form and needs only the application of certain criteria to bring it into conformity with new requirements. The investment of time and money is therefore comparatively small in this respect.

10. ADVANTAGES OF "COMPUTERIZED PUBLISHING"

1. The TEAM Program System performs all tasks that are capable of being automated — from data recording, to sorting, to phototypesetting — so that it can in a true sense be called an "automated publishing system."

2. The system is conveniently equipped with a large character set for normal spelling and punctuation, with facilities for on-line data acquisition and interrogation in the interactive mode via display terminal, and with a wide variety of programs.

3. The terminology inventory can be modified and expanded with a minimal investment of time and money.

4. Errors can be systematically, swiftly, and reliably located with the help of selection programs.

5. The danger of committing new errors during error correction is extremely small.

6. No proofreading is required after page makeup.

7. Apart from dictionary projects, the TEAM System is also suitable for all non-numeric EDP applications where large volumes of data have to be frequently and quickly updated (as for city directories, telephone books, catalogs, etc.) as well as for the fully automatic compilation of indexes for encyclopedias and similar works (that is, provided the texts lend themselves to the automatic extraction of keywords).

2. This method is currently being used in the technical *Dictionary of Engineering and Technology/ Dictionnaire de la Technique industrielle* by Richard Ernst. The dictionary, which will be published early in 1982, comprises some 160 000 entries.

8. It is not necessary to know at the time of data acquisition the configuration, order, typeface, or format in which the data for unilingual and multilingual dictionaries is to be set.

9. The substantial savings in time and money offered by computer-aided lexicography are increased many times over by the multiple use of stored data to produce new editions or inverted versions of existing works, dictionaries in new combinations of languages, and similar publications.

Computer-aided lexicography of the type provided by the TEAM Program System represents an ideal collaboration between man and machine.