[*Translating and the Computer 9*. Proceedings of a conference… 12-13 November 1987, ed. Catriona Picken (London: Aslib, 1988)]

The Termium termbank: today and tomorrow

Alain Landry

Assistant Under Secretary of State, Official Languages and Translation, Ottawa, Ontario, Canada

CONTEXT

In order to place the Canadian government's termbank in its historical and operational context, I would like to mention that the Translation and Official Languages Sector of the Department of the Secretary of State, which includes the Translation Bureau, is responsible for providing translation, interpretation and terminological services for all federal government departments and agencies in Canada. This responsibility, which dates from the early days of Confederation, was reinforced with the adoption in 1969 of the Official Languages Act under which all documents intended for the public and issued under the authority of the federal administration must be published in both English and French. Since then, the demand for linguistic services has more than doubled.

As a result, the size of the Translation Bureau has increased from 500 employees in 1967 to a current strength of 1,500, 800 of whom are translators and terminologists. The annual translation workload represents more than one million pages. Owing to factors such as the increasing use of native languages, international trade, inter-governmental relations and information technology, translation demand in Canada is increasing by 10 per cent every year. Increasing human productivity with computer aids is thus vital to the nation's economy. The Translation Bureau has drawn up short-, medium- and long-term plans to tackle this problem.

MANDATE

The Canadian Government Linguistic Data Bank was established in 1974, after Cabinet made the Translation Bureau responsible for 'verifying and standardising English and French terminology used throughout the federal public service and in all government agencies reporting to the Parliament of Canada'. Fulfilment of this mandate required, among other things, the organisation and promotion of terminology research projects, the establishment of a termbank for the purpose of increasing the efficiency of translation services in all fields, and the development of cooperative ties with language research and standardisation centres across Canada and abroad.

The bank was to serve three main purposes:

(1) to provide an efficient research tool for government translators, interpreters and writers;
(2) to help standardise terminology; and
(3) to facilitate the implementation of the Canadian government's official languages policy.

Over the years, the Termium bank has outgrown its original mandate of service to translators and other public servants in the federal administration. It is now considered to be a key element in bilingualisation programmes throughout Canada. Recognised internationally as a major achievement in the area of language technology, Termium provides concrete evidence of the success of Canadian efforts to support the use of the country's two official languages.

BACKGROUND

Where did Termium come from? The basic system was developed in the early 1970s by a research team assembled by the University of Montreal which gave it the name Termium (Termi for terminology and urn for University of Montreal). This database was acquired by the Department of the Secretary of State at the end of 1975. It had been created primarily as a translation tool and was therefore well suited to the Translation Bureau's needs. It is interesting to note that the 150,000 University of Montreal records which served as a base for the initial Termium file were also included in Eurodicautom, the terminology databank of the European Communities.

In January 1977, once the necessary adaptation of the software had been made (resulting in the second generation of Termium), the system was transferred to a central computer in Ottawa. The hundreds of thousands of terminology records accumulated by Bureau translators over the years were then added to the initial database. At this stage, the services of terminologists were enlisted: duplicate records were eliminated, inaccurate information was screened out and records pertaining to the same concept were consolidated. By the end of 1980, the bank contained 600,000 records, all accessible online.

It soon became evident that, because of the limited technical capabilities of

this second generation system and the expanding network of Termium users both in Canada and abroad, a new system would be required. Consequently, from 1982 to 1985, the Department of the Secretary of State developed its own third generation system using the BASIS software package in conjunction with the Telidon technology which permits the use of menus and colour graphics. Moreover, two computers were operated to serve the bank's needs and to facilitate user access. Operational since the spring of 1985, Termium III offers, as we shall see later, a diversified approach to client needs.

At present, the bank's approximately 2.5 million terms can be accessed by the 2,500 users who make up the Termium network. Terminals providing access to Termium serve not only the Canadian government's translators, but also the entire federal public service, several provincial governments, universities, various research institutes, translation companies and foreign organisations such as the Commissariat Général de la Langue Française in Paris, the headquarters of the United Nations in New York, the International Monetary Fund in Washington and the International Maritime Organisation in London (see Figure 1).

THE CONTENT OF TERMIUM

The Termium bank now consists of four linguistic data collections and one documentary database (see Figure 2).

(1) The *terminology data collection* containing English-French term equivalents consists of 775,000 records. It is subdivided into three files according to the quality status of the data. Records are also classified according to a classification scheme reflecting semantic links with given scientific, technical or socio-economic fields of knowledge. The built-in hierarchy allows for the retrieval of very limited subsets of data (eye surgery, for example) or very large ones (such as the field of medicine).
(2) *The proper names collection* contains 75,000 titles, including trade names, brands, toponyms, official titles of programmes, Acts and Regulations, official names of various organisations, titles of documents, etc. This data collection is considered by translators to be among the most useful and it has seen substantial growth in recent years.
(3) The *multilingual data collection* consists of 33,000 records. Each record establishes a relationship between two languages, one of them being either English or French and the other a foreign language. At present, there are only records with German or Spanish equivalents.
(4) The *translation problems collection* contains roughly 10,000 records dealing with common translation difficulties (for example, proverbs and idioms, phraseology, syntactic difficulties).

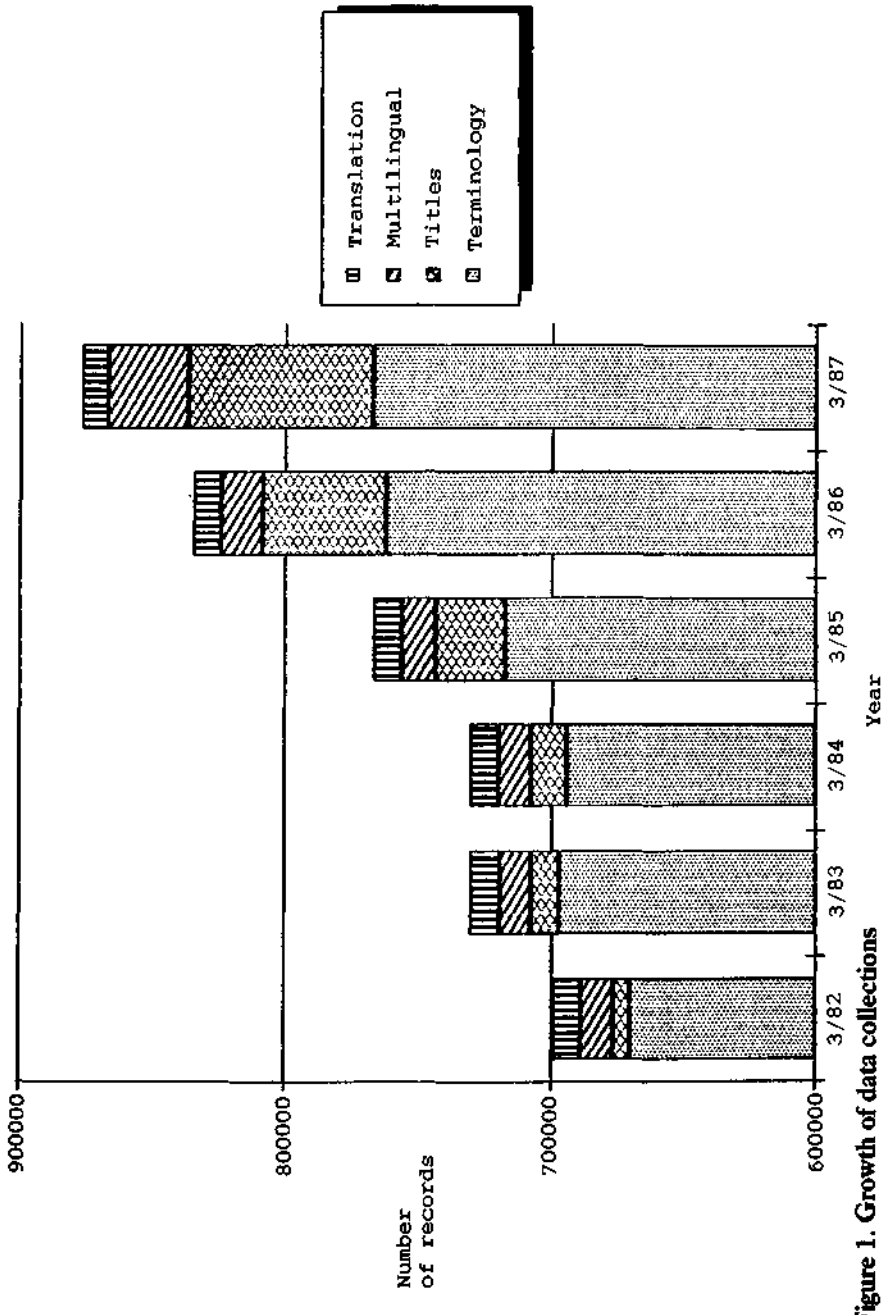

Figure 1. Growth of data collections

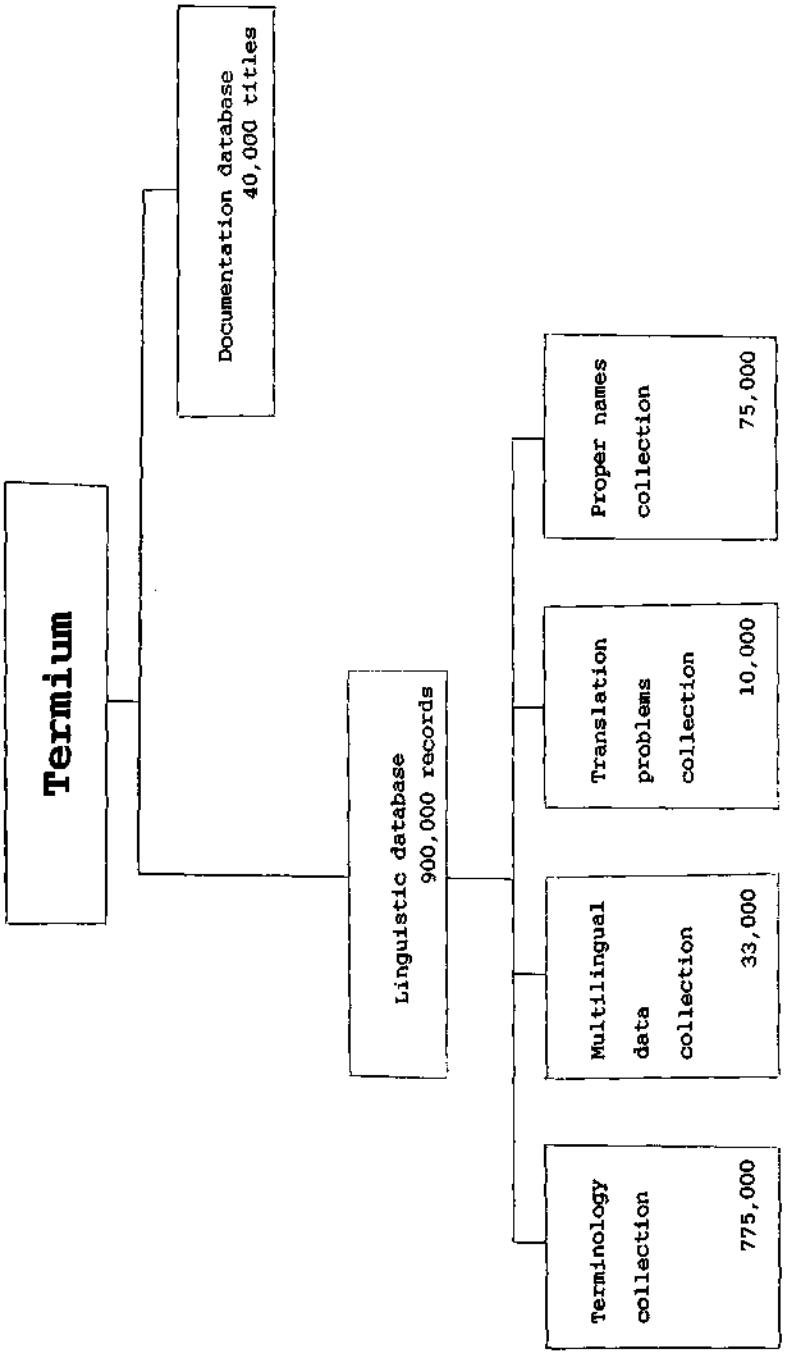

Figure 2. The content of Termium

As mentioned earlier, Termium also contains a *documentation database* used for online decoding of source codes on terminology records and provides clients with full bibliographic references and document locations for the approximately 40,000 documents used as source material by the terminologists.

One final statistic – if we were to print out all the terminology records in the termbank on 210x297 millimetre sheets, we would have a pile of paper 22 metres high – so please do not ask us for printouts!

THE TERMIUM APPROACH

From the beginning, the management and expansion of the Termium bank have been based on a concept-oriented approach and a sharing of responsibility with its users for the maintenance and upgrading of its content.

The first principle (concept-oriented approach) concerns the entry into the computer of concept-based records, that is, records that contrary to previous models proceed from a concept to its designation and not from the designation of a concept in one language to its designation in another. This approach makes it possible to enter on a single record the different designations used for a given concept (synonyms, abbreviations, spelling variants, etc.) and to confine the addition of records solely to those that deal with a new concept (see Figure 3).

In our view, Termium is a co-operative endeavour. Although the Termium bank remains the property of the Government of Canada, other organisations contribute to its operation. Its clients are not only users of a service, but also suppliers of data, contributing their knowledge and experience to the build-up of the central database, either by adding new records or by upgrading existing data. This collaboration by our clients is most important since, once a termbank has been established, keeping it up to date is an enormous – almost impossible – task. For example, this year we expect to add new records and to modify or cancel existing ones at the rate of about 4,000 a week. We have come a long way from the 150,000 records making up the University of Montreal's original database! But there is also a long way to go before we have a bank that is fully up-to-date – given the tens of thousands of new terms entering the language every year – and that contains nothing but correct, verified data.

That is why – despite the varying degrees of reliability and completeness of the records loaded – it was decided that all terminology records, no matter how incomplete, contained potentially valuable information. They were therefore worth loading into the bank, provided users were made aware of any records that might contain incomplete data and provided they agreed to help improve less complete records. In other words, the bank had to be upgraded through use.

OPERATION OF THE BANK

The operation of Termium is dependent on a network of terminologists and specialists connected with the bank by a series of terminals, and on data entry

```
FICHE 1

                ** Terminologie - Fichier uninotionnel **

DOMAINE(S)          : *Equipement (imprimerie et arts graphiques)
                      *Périphériques (Informatique)

EN                                          FR
laser printer*b*CORRECT,CORRECT;UNIF,OF     imprimante à laser*a;b*CORRECT,CORRECT
   FIC                                         ;FEM,FEM;UNIF,OFFIC

laser beam printer*c*CORRECT,CORRECT        imprimante au laser*d*CORRECT,CORRECT;F
                                               EM,FEM


DEF*A nonimpact printer that operates       CONT*Dans l'imprimante à laser, un
   at well over 10,000 lines per               rayon laser excite une couche de
   minute, using a low-power laser to          sélénium sur un tambour pour que,
   produce image-forming charges a line        par son électrisation, il attire
   at a time on the photoconductive            les encres spéciales sur les zones
   surface of a drum; dry powder that          touchées, ces encres étant ensuite
   adheres only to charged areas is            transférées par contact sur le
   applied to the drum, transferred to         papier.*b
   plain paper, and fused by heat.*b        OBS*Bien que certains auteurs emploient
OBS*Although the terms "laser printer"         parfois les termes "imprimante à
   and "electrographic printer" are            laser" et "imprimante
   sometimes used interchangeably, they        électrographique" de façon
   are not perfectly synonymous.*e             interchangeable, ces termes ne sont
OBS*laser printer: OFFIC-GESC*f                pas parfaitement synonymes.*e
                                            OBS*imprimante à laser : UNIF-ONGI*f


CLE : electrographic printer                CLE : imprimante électrographique

SOURCES
-------
   a*ZWRV*1984
   b*CBT-146*1986***168
   c*COMPN*1983*7*12*31
   d*INBUR*1983*4*8*40
   e*3JCD*1986
   f*3TNH*1987
```

Figure 3. Example of a concept-based record

and querying methods adapted to the type of terminological information processed.

Network of researchers

The research network includes about a hundred Bureau terminologists whose work methods are based on universally accepted principles of terminological research. Other participants include translation units of the Bureau and terminologists, translators and editors employed by outside users of the bank. Terminologists regularly receive research requests from clients throughout the network. These requests are then forwarded to those specialising in the subject field concerned and research findings are stored in Termium. In addition, data received from clients at various points in the network are loaded into the bank: this enables everyone to benefit, even though some findings may be only tentative.

Network of terminals

Expansion plans for Termium call for installation of the maximum number of terminals possible, in order to provide improved access to the database for all Translation Bureau translators, federal employees and various categories of outside clients. We now have about 2,500 clients linked to our system (see Figure 4).

o

Querying

The Termium bank was designed primarily to meet translation needs, that is to provide immediate answers to queries about individual terms (as opposed to entire vocabularies). For this reason a direct communication mode was provided which allows users to key in a term or group of terms and immediately obtain one or more records on a video display screen or printout. Data may also be accessed via an offline print service which, though slower, allows the user to obtain, in the form of computer listings, bilingual glossaries or large numbers of records, retrieved by subject field, originator or a combination of these.

In the latest version of Termium, a number of features ensure personalised access to the system. These include individual user names and passwords, user profiles which enable the client to choose the language of consultation, the display format, the type of data provided at each step in the query, and so on. Thus a user may choose to have access only to certain parts of the record, for example equivalents and definitions or equivalents and sources.

A series of menus displayed by the system enables the user to obtain an answer promptly and easily. If the term requested is a complex unit not available in the bank, a search may be carried out for a part of the term requested, or the computerised index may be consulted to locate a related term or one with a different spelling. Such mechanisms, by giving access to related terms, increase the response rate and provide translators with at least partial answers.

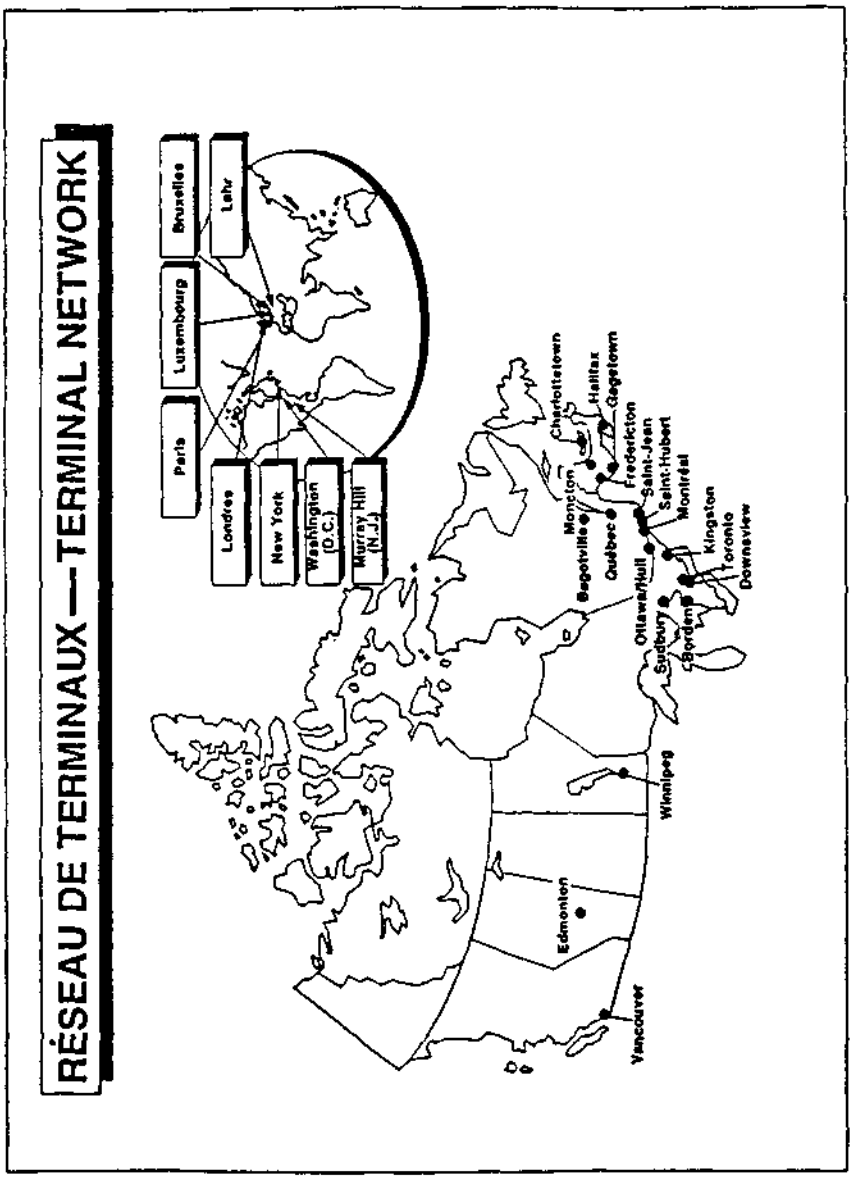

Figure 4. Terminal network

```
                                INDEX

TEN=asychronous communications device
0 records for TEN=ASYCHRONOUS COMMUNICATIONS DEVICE


                              No-Hit Request

* 1  Query Another Term                4  Execute a Stem Search
  2  Search by Key Term                5  List of Terms not in the Bank
  3  Consult the Index                 6  Return to Main Menu

?3

    5/ LOOK (20)TEN=ASYNCHRONOUS COMMUNICATIONS DEVICE

   ITEMS      TERMS
A       1  TEN=ASYNCHRONIZATION LIGHT
B       2  TEN=ASYNCHRONOUS
C       1  TEN=ASYNCHRONOUS ADMITTANCE
D       1  TEN=ASYNCHRONOUS ATTACK
E       1  TEN=ASYNCHRONOUS CHANNEL GROUP
F       1  TEN=ASYNCHRONOUS CHANNEL MULTIPLEXER
G       1  TEN=ASYNCHRONOUS COMMUNICATION
H       1  TEN=ASYNCHRONOUS COMMUNICATION BASE
I       1  TEN=ASYNCHRONOUS COMMUNICATION SYSTEM
J       1  TEN=ASYNCHRONOUS COMMUNICATIONS ADAPTER
***** YOUR TERM *****
K       1  TEN=ASYNCHRONOUS COMMUNICATIONS INTERFACE ADAPTER
L       1  TEN=ASYNCHRONOUS COMPUTER
M       1  TEN=ASYNCHRONOUS CONDENSER
N       1  TEN=ASYNCHRONOUS DATA SET
O       1  TEN=ASYNCHRONOUS DATA TRANSMISSION
P       1  TEN=ASYNCHRONOUS GENERATOR
Q       1  TEN=ASYNCHRONOUS HANDSHAKE
R       2  TEN=ASYNCHRONOUS IMPEDANCE
S       1  TEN=ASYNCHRONOUS IMPEDANCE SYNCHRONOUS MACHINE
T       1  TEN=ASYNCHRONOUS LINE MULTIPLEXOR

    5/ /X LISTOFF NOLIST

  1  Select Entry from Index          3  Query Another Term
* 2  Return to No-Hit Request Menu    4  Return to Main Menu


?1

Enter the letter(s) corresponding to the index entry or entries you
wish to access; separate letters with commas.  Note that you will
receive only one breakdown for the entire group of documents you have
indicated.

LETTERS=g-k,t
```

Figure 5. Special file containing terms not found in the bank

Terms not found in the bank when it is queried online can be placed in a special file whose content is regularly reviewed to help terminologists identify users' problems and guide terminological research (see Figure 5).

Termium III also provides a complete electronic mail service, including a 'mailbox' for each user, which records messages received and sent. This system of communication among users of the bank enables them to contact specialists in a given field directly and to obtain terminological information from the source.

Our statistical data indicate that translators initiate 64 per cent of the total number of term queries on the system. Over the last six months, an average of 100,000 terms were queried by our clients each month. The total number of queries has increased dramatically over the past year, now that some technical problems related to the system's processing capacity have been ironed out. To give you some idea of the total demand for terminology, we expect that this year online querying of the termbank will exceed one million queries.

The CD-ROM

In order to satisfy this huge and growing demand, the Department of the Secretary of State has been looking at new technologies to provide better access to the database. Two of the main constraints we were faced with were the technical limits of the present system (we cannot keep adding larger and larger computers!) and the need to have access to Termium twenty-four hours a day (in Canada alone, we have six different time zones). That is why the Translation Bureau is currently developing a CD-ROM (Compact Disk Read-Only Memory) application for Termium. As you may know, a CD-ROM disk, using the same laser-read optical data-storage system that can hold more than an hour of high-quality sound, can store enormous amounts of information. In our trial run, the entire terminology data collection was stored on a single compact disk. Moreover, the disk's capacity, combined with its method of production, makes it extremely cost-efficient. A single feather-light CD-ROM platter can store 200,000 pages of text (about the same as sixty personal computer hard disks) at a cost of a few cents per million characters. In fact, the CD-ROM technology has the potential to bring mainframe performance to a desktop IBM-compatible computer. A microcomputer, a disk reader and a printer are all that is needed. Imagine having at your fingertips the two million terms stored in Termium, just like the content of an ordinary reference book!

A pilot project giving access to Termium on CD-ROM to a limited number of users from the Translation Bureau and the private sector was initiated this summer. A preliminary analysis of the results indicates that there has been an enthusiastic response. In general, we noted an increased use of the termbank by translators who found the CD-ROM admirably suited to their day-to-day needs. In our experiment, Termium was only accessible through a microcomputer devoted to that use. Maximum efficiency would be obtained if the CD-ROM was linked directly to the translator's workstation.

The translator's workstation

The transfer of termbanks to compact disks will greatly facilitate access for translators working with microcomputers. We are pushing ahead with research into a translator's workstation. This study is being carried out at the Canadian Workplace Automation Research Centre, in cooperation with the Department of Communications. The purpose of the workstation project is to provide

translators with a set of tools capable of improving their working conditions and productivity. It has been found that translators spend a large proportion of their time on activities other than actual translation (for example, research). Considerable savings could therefore be achieved with the help of relatively simple equipment and software. We will soon be entering the technical specifications phase and we hope to have a prototype ready for field testing within a year. What will this prototype consist of? It will be a network of workstations linked to one another, equipped with high-resolution, full-page screens and sharing several draft- and letter-quality printers. Each translator will have the following: a word-processing software package to meet his or her needs as regards layout and special characters, a spell checker, a section-wide terminology management program, a file comparison program and an index of previous translations.

We expect that the workstation will allow the translator to consult dictionaries on a split screen and automatically paste the selected word on to the current translation. The workstation would also be interfaced with a communications network in order to receive and transmit original and translated texts. In this way, costly retyping and delays would be avoided, since many clients prefer to receive the translation online through a communication device.

Ideally, each translator should be equipped with a workstation comprising a word-processing facility and a video display terminal. The display unit, which should have multiple windows, would allow the translator to write the text while simultaneously having direct access to a terminology bank and various computer-aided translation systems. The new laser disk or CD-ROM, now in the testing phase as a vehicle for Termium, could be incorporated, together with other similar databases, among the panoply of source materials. One of the stations would be reserved for administrative support functions, such as word counting, communications outside the network, indexing, text tracking, management information and so on.

We are also using, on an experimental basis, a computer-aided translation system. For the past nine months, a translation unit in Montreal has been working in close cooperation with LOGOS, a private-sector company specialising in software development. This unit is using the software provided by LOGOS while at the same time improving it to meet the Canadian government's needs. As research and development continues we find that the use of the LOGOS software is exceeding our translators' expectations.

No decision has yet been reached on the network architecture. The choice will depend on both the cost and the potential for expansion. This is because the workstation must be conceived in modular fashion: features could be selected 'à la carte' in accordance with individual needs and, as technical progress is made, new components could be added (contextual spelling checker, one or more computer-assisted translation software packages, and later – who knows? – voice recognition).

Decentralised working files

Other research is currently being done on the feasibility of having decentralised databases updated directly by the translators. These working files would allow translators linked to a local area network (LAN) to share instantly the results of their research. These local databases would replace the individual files kept by the translator on cards. At present, if translators want their colleagues to benefit from their research, they must write the information on a card and put it in the unit's index card file or send it to the terminology service. By directly updating a decentralised database, the translators would enable everyone linked to the same local area network to have instant access to the information. Once these 'minibanks' are established, it is planned to link them with Termium so that their content can be used to update the bank's content. The potential and usefulness of various communication and database softwares on the market is currently being tested. Preliminary results indicate that this link-up is something that could be implemented in the near future within a local area network.

What we have in mind is a completely integrated system where both translators and terminologists will benefit from computer-assisted aids. The terminologists, although working for a computerised bank, still do a large part of their work manually. The possibility of updating a working file directly during the course of research is being studied. Rapid access to all the available information is the key element. To keep up to date, the termbank must receive without delay all data collected by either translators or terminologists.

Desktop publishing

Since the end result of terminologists' research is the loading of the information collected into Termium, the production of a publication can sometimes make them deviate from the original objective. This should not be so. With the integration of electronic publishing systems now available on the market into the workstation, the terminologist should be able both to format the material gathered during research into a document ready for printing and to send the data for loading into the bank. The presentation of glossaries prepared directly from the bank could be improved with the help of desktop publishing software packages. The glossary file could be transferred to a microcomputer where tasks such as modification of type style, addition of titles and subtitles of various point sizes, and incorporation of graphics or diagrams could be done through a desktop publishing program.

Quality management program

In its search for excellence, the Translation Bureau has also carried out research to improve the tools used to manage the bank's content. By systematically evaluating samples taken from the bank, we are now able to determine the

quality of the data in Termium. Recently, a program was launched to extract from the bank a statistically valid sample of either the whole database or a particular subset (a field for example) and to analyse it, thus dispensing with the need to scan the entire database for the information needed. This mechanism will enable us to document the evolution of the database and to assess the impact of management decisions on the content of Termium.

An initial computerised analysis indicates that roughly two-thirds of the records in Termium have at least one supporting text, whether a definition, a context or an observation. Moreover, 50 per cent of those records are complete records, that is they have supporting texts in both English and French. (See Figure 6). This analysis also showed that only 17 per cent of our records have synonyms, which means that more than 80 per cent of the records have only one English term and one French term! Sources, contexts, definitions and quality ratings are among the data to be analysed further. By periodically reviewing the content of the bank, we expect to be able to improve the efficiency

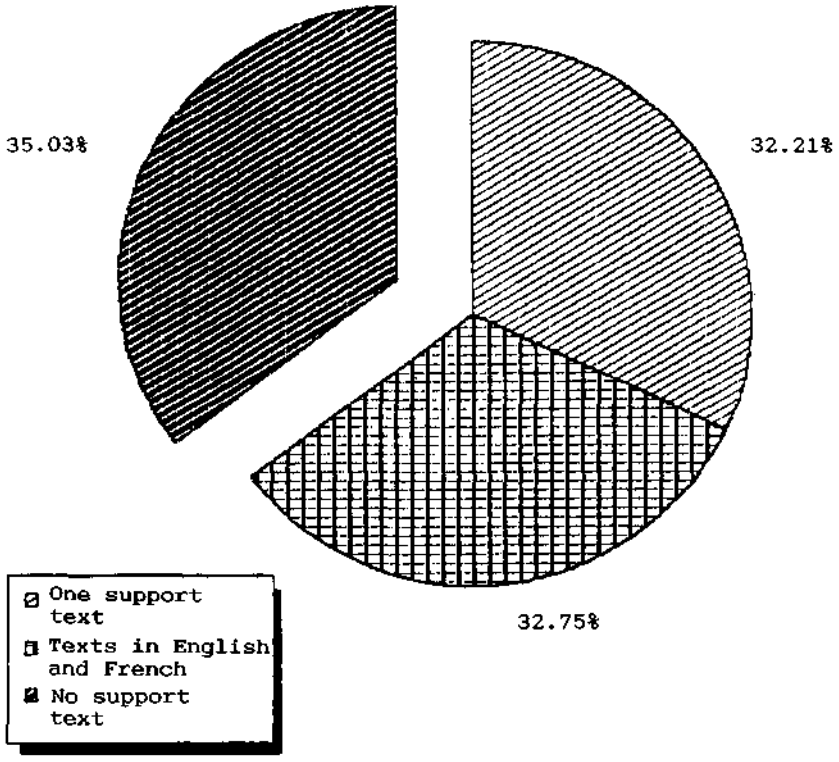

Figure 6. Quality management program - supporting texts

of the research programs designed to update its content. As one can see, computerised tools are opening new avenues for all professionals including managers. So much the better, since the search for excellence is a never-ending one.

CONCLUSION

Excellence, productivity, rapidity, accessibility and cost-effectiveness are the key words that govern what we do in terminology and that help us decide how to do it. Termium has already proved itself with respect to these keys. We are justifiably proud of what we have accomplished so far, but, at the same time, we are aware of the need to keep up with the increasing and varied demands placed upon us. To do this, we must adapt our methods and technology to make full use of all innovations in a constantly evolving field.

Terminology is a young and growing discipline. The tools that terminologists use have gone, in only a few years, from the pre-industrial to the post-industrial mode. In fifteen years, we have moved from index card files to laser optical disks. Collectively, the terminology community has adapted remarkably to this new environment. I am confident that, collectively, we will continue to move ahead in the years to come.