[*Translating and the computer 11*. Proceedings of a conference … 16-17 November 1989, ed. Catriona Picken (London: Aslib, 1990)]

IFP terminological databases and the compiling of specialised dictionaries

M. Moureau, G. Brace and P. Boisserpe

Institut Français du Pétrole

Paper presented by G. Brace

There is no denying the value of a good scientific and technical dictionary for specialists, scientists, linguists and translators working in the field covered by the dictionary. In 1959, faced with the lack of such a dictionary in the field of petroleum technology, the Institut Français du Pétrole (IFP) set itself the goal of compiling a comprehensive English-French/French-English dictionary for the benefit of the petroleum industry. This work was published by Editions Technip in 1963 and contained more than 40,000 terms and expressions used in petroleum technology. It also contained a great many definitions of specific concepts. This dictionary was the kickoff point for what has become the IFP terminological database, now containing more than 60,000 terms. Since 1963 IFP has had powerful computer facilities (IBM, then CDC, then VAX and Cray). This fact encouraged information specialists to make use from the start of these facilities for compiling terminological tools, dictionaries and thesauri.

Initial achievements

First to take on concrete shape was the English-French version of the dictionary. This took four years to compile. Based on this version, the automatic flip French-English version took just four months to create. This second section was somewhat weak linguistically, because definitions and concepts that existed solely in French were not included. A good example of this is the term *gaz fatal,* which refers to natural gas that is indissociably produced with crude oil, for which the only possible English translation is *associated gas,* a term that has its own French equivalent of *gaz associé.* There is no expression in English that reflects the ineluctability of this association. The creation of this second part was IFP's first experience with computerised documentation.

A second and entirely revamped edition was published in 1979 with over 50,000 terms and expressions. Two of the major changes compared to the first edition were: (1) the addition of genders, the lack of which had been very bothersome to both English and French speaking users of the first edition, and (2) the systematic use of upper and lower case letters to start each term so as to indicate clearly the difference between common and proper nouns. Above all, however, this second edition was reworked in depth with the help of computerised facilities that had not existed for the first edition. The terms and expressions selected for the new version were printed in the form of a KWIC index so that the best alphabetical entry could be chosen for each term.

In both the English and French halves of the dictionary, classification is mainly based on the noun part of the term. For example, the term *relaxation time* was naturally placed under *relaxation* and the term *electric logging* under *logging,* but the term *geostrophic wind* was logically placed under *geostrophic* so as to introduce this concept. Likewise, the term *fossile remanié* naturally led to the creation of a French entry for the term *remanié = reworked* because, if this term had merely been associated with rock or fragment, its absence would have created problems.

However, although the CDC computer IFP was using at that time was a powerful instrument for scientific computing, it could not handle the direct inputting of our terminological data. All it had was a punch-card reader that accepted terms in upper case letters only, meaning that the entire 520 pages of the English - French side of the dictionary would have had to be retyped if we wanted to computerise it.

It was when a VAX network was installed for use by IFP researchers that we obtained the possibility of using a database management system, and this is the software we used to create our terminological database, which we called TERMINO.

INGRES software

INGRES (INteractive Graphics and REtrieval System) is a relational database management system (DBMS) that manages data stored in databases made up of tables. Each table has a given number of columns, or fields, containing information about each entry. The INGRES query language is used to append, retrieve, update or delete the data gathered.

We first started to compile an initial table (DICPET) to enter the new terms we had been gathering since 1979. This table now contains more than 6,000 terms. Our second step was to compile a dictionary of terms specific to the field of seismic prospecting.

Our data input format, or table, was defined with the following columns:

— a *French word* field of 60 characters
— a *French-language definition* or *comment* field of 250 characters
— an *English word* field of 60 characters

- an *English-language definition* field of 250 characters
- a *category* field of 10 characters, in which we entered the specific domain of the word or term (e.g. drilling, economics, etc.).

Two other columns were added, called *bmf* and *bma,* each accepting 60 upper case letters with no accents, no apostrophes, no hyphens, and no cedillas. These columns were used to transform the all-inclusive typography of the terms entered in the French and English *word* fields into an impoverished typography that enabled INGRES to carry out an overall alphabetical classification.

Printout of lists

The main aim was to create alphabetical lists that would give us a constantly - updated state of the terms entered in the database together with their definition. The only real problem in achieving this goal was the interclassification of terms, i.e. to have composite terms come directly after the single terms making up the first word. By this I mean:

—English, e.g.	and not
filter	filter
filter correction	filter correction
filter panel	filtering
filter response curve	filter panel
filter slope	filter response curve
filtering	filter slope

—French, e.g.	and not
écart	écart
écart d’indicatrice	écart d’indicatrice
écart de correction	écart de correction
écart moyen	écartement
écart quadratique	écart moyen
écart type	écart quadratique
écartement	écart type

Searching for a term

A simple term is generally found most quickly on alphabetical paper printouts. However, lefthand truncation using the INGRES query language enables an online search to be made for a word within a compound term:

e.g. request:	→ **result:**
retrieve *raypath	incident raypath
	reflected raypath
	refracted raypath

Final printout

From the data entered it is easy to print out a compiled document after enriching the typography. Figure 1 shows an initial work sheet. Figure 2 shows the final printed version.

coherent noise	* bruit cohérent * bruit organisé
collapse structure	* déformation par glissement * structuration par affaisement
collection	* regroupement (de traces)
color display	* représentation couleur
colored sweep	* balayage non linéaire
column matrix	* matrice-colonne
comb	* peigne
Combisweep	* Combisweep (dénomination commerciale d'une technique d'émission vibrosismique)
comma	* virgule
common	* commun, commune
common bus	* bus commun
common-depth-point gather	* regroupement à point-miroir commun (de traces)
common-depth-point stack(COPS)	* somme à point-miroir commun (de traces)
common-geophone gather	* regroupement à géophone commun (de traces)
common midpoint	* point-milieu commun
common-midpoint gather	* regroupement à point-milieu commun
common-midpoint stack	* somme à point-milieu commun
common mode	* mode commun (à)

Figure 1. Work sheet

This experiment was carried out successfully in 1987 with the publication of our *Dictionary of seismic prospecting.* This database, containing some 4,000 terms, was used to enrich the DICPET database with new terms in the field of seismic prospecting.

Compiling a database from the *Dictionary of petroleum technology*

The problem of computerising all the vocabulary contained in the printed copy of the dictionary was solved in 1988 by renting an Inovatic optical character recognition system (scanner plus RS3 character-recognition software) from 8 March to 15 April, i.e. for just over a month. During this time the OCR system was used to perform the following operations: (1) scanning the English-French part of the *Dictionary of petroleum technology* (520 pages with an average of 4,300 characters per page) and creating a postprocessing program for formatting the data before inputting it into INGRES; (2) scanning the catalogues of all IFP publications for the last 23 years (i.e. more than 1,000 pages). The cost of these operations was mainly: (1) rental of the OCR system for one month (FF 10,900 including taxes), (2) training in how to use the system (FF 2,300 including taxes), (3) two engineer-weeks, and (4) two secretary-weeks.

coherent noise - bruit cohérent, bruit organisé.

collapse structure - déformation par glissement, structuration par affaissement.

collection - regroupement *(de traces)*.

color display - représentation couleur.

colored sweep - balayage non linéaire.

column matrix - matrice-colonne.

comb - peigne.

Combisweep - Combisweep *(dénomination commerciale d'une technique d'émission vibrosismique)*.

comma - virgule.

common - commun, commune.
common bus: bus commun.
common-depth-point gather: regroupement en point-miroir commun *(de traces)*.
common-depth-point stack (CDPS): somme en point-miroir commun *(de traces)*.
common-geophone gather: regroupement en géophone commun *(de traces)*.
common midpoint: point-milieu commun.
common-midpoint gather: regroupement en point-milieu
common-midpoint stack: somme en point-milieu commun.
common-mode: mode commun (à).

Figure 2. Final printed version

We found that the scanner used for such an operation must be a high-efficiency piece of equipment, and that its quality is measured mainly by its power of resolution (200 or 300 points per inch) and the number of luminosity and contrast levels it has (only three for low-efficiency equipment). The scanner we rented was a Microtek with 300 points per inch and with 15 contrast and luminosity levels. Tests with less efficient scanners proved unsatisfactory (insufficient power of recognition).

The character-recognition software rented from Inovatic operated by an ‘intelligent analysis’ of the page, including recognition of columns, graphs, tables and underlining. Column recognition resulted in the reading of ‘blocks’ of characters on the page in a sequential order, as is normally done by layout specialists. Graphs and tables were recognised but have not yet been interpreted. We are currently studying how to recognise the composition of a table. Underlining can be recognised and reproduced by special characters for the beginning and end of the underlining. Considerable research is being done on the recognition and reproduction (using the same principle of special characters) of letters in boldface and italics (information that is lost at present because a capital A, whether in lightface, boldface, roman or italics, which are the four possibilities existing in our dictionary, is still recognised only as a capital A).

Some of the lessons learnt from this experience are that the documents to be acquired must: (1) be of very high-quality print, (2) have homogeneous type fonts, and (3) be sufficiently important to justify the time required for the scanner to learn the fonts used (four different types for a dictionary, for example).

Practical experience has shown that:

— A printed document is much more legible than a typewritten document.

— The printed character must stand out clearly to be read accurately. There are times when an ‘m’ becomes ‘rn’ or when ‘ri’ is transformed into ‘n’ or vice versa. There are times when this defect can be corrected by adjusting the luminosity and contrast of the scanner, but in general it is the quality of the original document that is of paramount importance.

— Every new font has to be learnt (about 15 to 20 minutes). A document using several fonts (boldface/lightface, Roman/italics, different size characters, etc.) thus requires an apprenticeship that can sometimes take longer than an hour and a half.

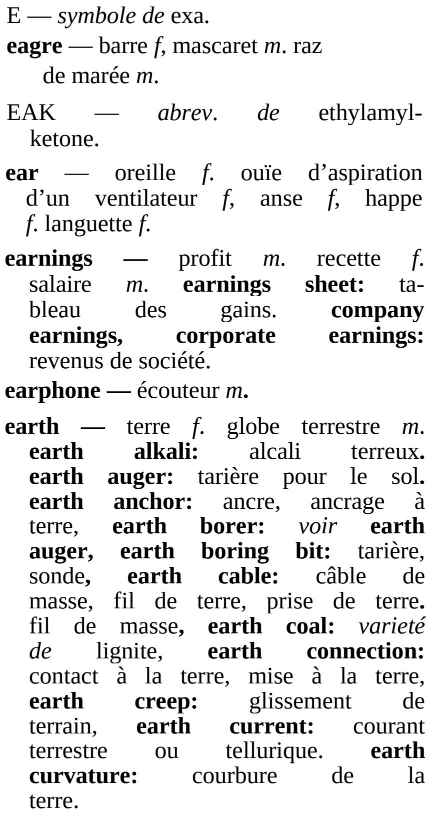
— Finally, the unrecognised characters (3 to 4 per cent) are replaced by asterisks (easily spotted by a text editor), and substitutions are not impossible.

Implementation of the database

The following three figures explain the sequence of operations. Figure 3 shows the start of the E column in the original printed dictionary.

Figure 4 shows the result of scanning after input into the VAX editor. The asterisks indicate characters that were not recognised. The dash after the main entry has become - - -. Text editing consisted in inserting the characters that were not read or were read improperly and that did not come out in the main entries.

Figure 5 is the same as Figure 4 after it was corrected, and after additional entries were made for *eagre, ear, earnings* and *earth.*



E — *symbole de* exa.

eagre — barre *f*, mascaret *m*. raz de marée *m*.

EAK — *abrev. de* ethylamylketone.

ear — oreille *f*. ouïe d'aspiration d'un ventilateur *f*, anse *f*, happe *f*. languette *f*.

earnings — profit *m*. recette *f*. salaire *m*. **earnings sheet:** tableau des gains. **company earnings, corporate earnings:** revenus de société.

earphone — écouteur *m***.**

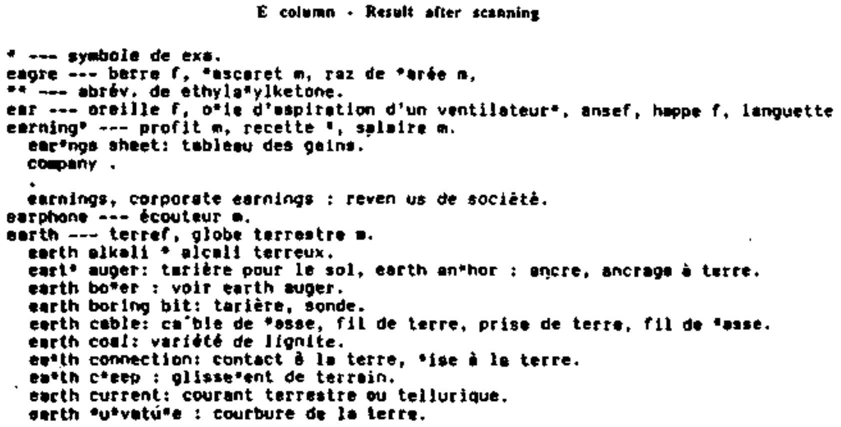
earth — terre *f*. globe terrestre *m*. **earth alkali:** alcali terreux**.** **earth auger:** tarière pour le sol**.** **earth anchor:** ancre, ancrage à terre, **earth borer:** *voir* **earth auger, earth boring bit:** tarière, sonde**, earth cable:** câble de masse, fil de terre, prise de terre**.** fil de masse**, earth coal:** *varieté de* lignite, **earth connection:** contact à la terre, mise à la terre, **earth creep:** glissement de terrain, **earth current:** courant terrestre ou tellurique. **earth curvature:** courbure de la terre.

Figure 3. Example from the original printed dictionary

Each letter was processed as a separate table. Before a table was copied into INGRES, a check was made of the length of the fields. The following rules were adopted:

- — the first entry = an *English word* or composite term with the space ending by - - - or:
- — the second entry = a *French word* followed by a full stop or a comma. Any text coming after the comma was considered to be a *definition.* If the second entry began by **abrév., voir,** or **dénomination,** it was inserted into the *French comment* field rather than the *French word* field.

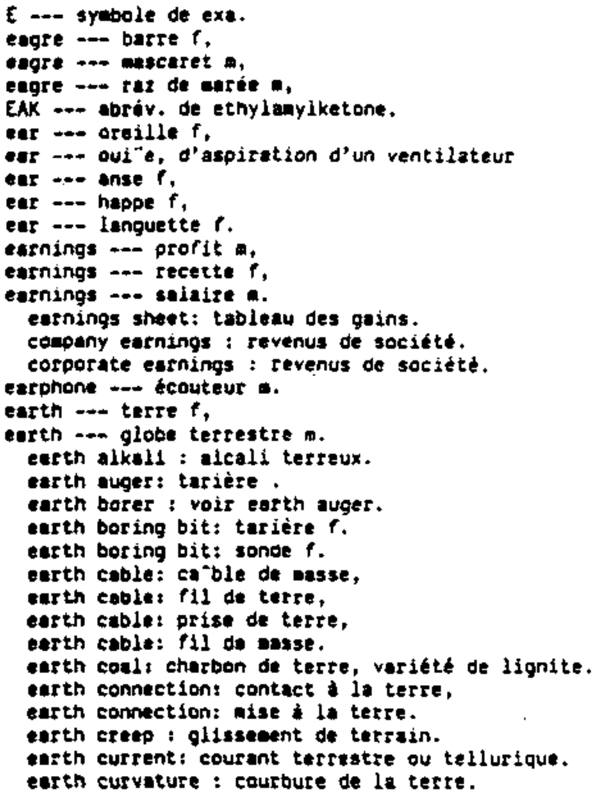
An error recognition program indicated the lines or rules not respected. In general the error had to do with the separator between the term and the comment or definition (Figure 6).



E column - Result after scanning

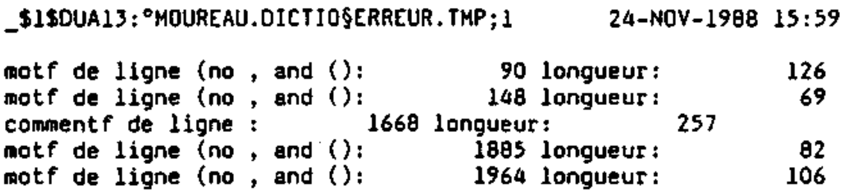
```
* --- symbole de exa.
eagre --- barre f, *ascaret m, raz de *arée m,
** --- abrév. de ethyla*ylketone.
ear --- oreille f, o*ie d'aspiration d'un ventilateur*, ansef, happe f, languette
earning* --- profit m, recette *, salaire m.
  ear*ngs sheet: tableau des gains.
  company .
  .
  earnings, corporate earnings : reven us de société.
earphone --- écouteur m.
earth --- terref, globe terrestre m.
  earth alkali * alcali terreux.
  eart* auger: tarière pour le sol, earth an*hor : ancre, ancrage à terre.
  earth bo*er : voir earth auger.
  earth boring bit: tarière, sonde.
  earth cable: ca^ble de *asse, fil de terre, prise de terre, fil de *asse.
  earth coal: variété de lignite.
  ea*th connection: contact à la terre, *ise à la terre.
  ea*th c*eep : glisse*ent de terrain.
  earth current: courant terrestre ou tellurique.
  earth *u*vatú*e : courbure de la terre.
```

Figure 4. Result after scanning



```
£ --- symbole de exa.
eagre --- barre f,
eagre --- mascaret m,
eagre --- raz de marée m,
EAK --- abrév. de ethylamylketone.
ear --- oreille f,
ear --- oui¨e, d'aspiration d'un ventilateur
ear --- anse f,
ear --- happe f,
ear --- languette f.
earnings --- profit m,
earnings --- recette f,
earnings --- salaire m.
  earnings sheet: tableau des gains.
  company earnings : revenus de société.
  corporate earnings : revenus de société.
earphone --- écouteur m.
earth --- terre f,
earth --- globe terrestre m.
  earth alkali : alcali terreux.
  earth auger: tarière .
  earth borer : voir earth auger.
  earth boring bit: tarière f.
  earth boring bit: sonde f.
  earth cable: ca^ble de masse,
  earth cable: fil de terre,
  earth cable: prise de terre,
  earth cable: fil de masse.
  earth coal: charbon de terre, variété de lignite.
  earth connection: contact à la terre,
  earth connection: mise à la terre.
  earth creep : glissement de terrain.
  earth current: courant terrestre ou tellurique.
  earth curvature : courbure de la terre.
```

Figure 5. Result after correction



```
_$1$DUA13:°MOUREAU.DICTIO§ERREUR.TMP;1       24-NOV-1988 15:59

motf de ligne (no , and ():          90 longueur:          126
motf de ligne (no , and ():         148 longueur:           69
commentf de ligne :        1668 longueur:          257
motf de ligne (no , and ():        1885 longueur:           82
motf de ligne (no , and ():        1964 longueur:          106
```

Figure 6. Error recognition program

After having entered the English words, French translations and French comments, we inserted the English comments and definitions (about 5,000 of them) one by one, which gave us a first opportunity to review and update the basic database. Then we reread and updated the entire database a second time.

This has been our schedule so far:

— April to November 1988: correction of the scanned version.

— September 1988 to May 1989: insertion of the English definitions and creation of three different tables, one for *drilling,* one for *refining and petrochemicals* and one for *energy economics.* To date, each letter, or almost each one, still forms a separate table so as to be easier to manipulate than an overall integrated table. Table 1 shows the number of entries for each letter.

A	2352	H	1993	Q-R	2863
B	3050	IJK	2338	S	6460
C	6332	L	2161	T	2845
D	3442	M	2148	U	674
E	1919	N	854	V	1113
F	3002	O	1225	W	1549
G	1836	P	3950	XYZ	298

Table 1. Number of terms in each letter table

At the same time, in cooperation with drilling engineers in different French companies, the *Dictionary of drilling and boreholes* has been compiled, and the manuscript is ready for printing. We plan to present our publishing subsidiary with a hard-copy printout of the computer file as well as a magnetic tape with all the entries in alphabetical order.

As of now, we are engaged in a rereading of the French-English version of each letter, because the errors are much more visible in this direction. Reading in this version is less monotonous because the vocabulary is more dispersed.

It is simple to see what remains to be done for the dictionary as a whole. A systematic revision has to be made of the vocabulary in all the fields not already making up the subject of an individual table. Some fields that are already covered must also be revised. The fields already covered include:

- geology
- geochemistry
- geophysics
- remote sensing
- drilling
- well logging
- safety

The fields that still remain to be covered are:

- production and reservoir engineering
- transportation and storage
- use of petroleum products
- environment
- refining
- energy economics
- petrochemicals

To enter all this new vocabulary, we have opened a supplementary table called PETROLE, which already contains more than 1,500 terms. By the end of December 1989 we hope to be able to integrate all the separate alphabetical tables, including the new vocabulary contained in DICPET, to form one giant table. Not all the terms contained in the specialised dictionaries will be included per se in the giant table. This macrotable will be the basic support for the publication of the new paper dictionary (Figure 7), which we hope to be able to complete by the end of 1990. But it will still be maintained in the form of a database so that it can be corrected and updated.

The new dictionary will be very different from the old one because the computer sets a very strict alphabetical order. Since the adjective generally (but not always) comes after the noun in French, the French-English side will mainly have the nouns in alphabetical order. But since adjectives generally precede nouns in English, the English-French side will, more often than not, be alphabetically indexed by the adjective for composite terms, as opposed to the manly noun-based indexing strategy of the old dictionary.

Other specialised dictionaries may be published, such as one on *Geology*. Likewise, consideration can be given to the possibility of proposing the new dictionary as an online service or in the form of a CD-ROM. The construction of this new tool is now well advanced, and it might even be the starting point for a whole line of new products.

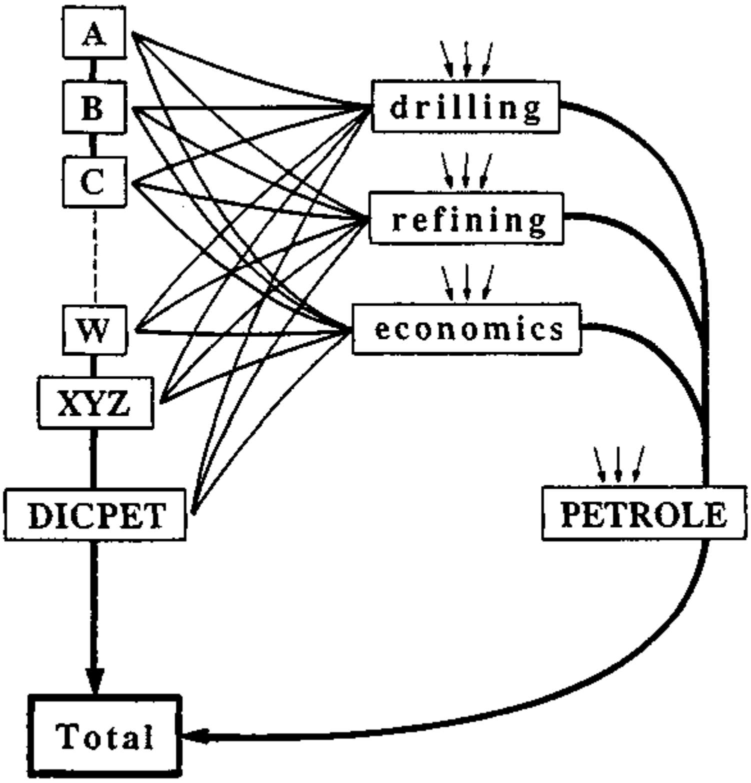


Figure 7. Formation of new macrotable

AUTHORS

M. Moureau, G. Brace and P. Boisserpe, Institut Français du Pétrole, B.P. 311, 92506 Rueil-Malmaison, France.