THE MONTREUX 1989 INTERNATIONAL CHEMICAL INFORMATION CONFERENCE

PROCEEDINGS

Montreux, Switzerland
26–28 September 1989

Harry R. Collier (Ed.)

Proceedings of the Montreux 1989 International Chemical Information Conference

Montreux, Switzerland 26–28 September 1989

Conference Delegate copy. Not for re-sale. The published Proceedings are available throughthe book-trade from Springer-Verlag Berlin, Heidelberg and New York (ISBN 3-540-51804-5)

Easy mapping classification of patent references with microcomputers

Henri Dou, Parina Hassanaly, Luc Quoniam

Centre de Recherche Retrospective de Marseille, Université Aix Marseille III, Centre de St Jérôme, 13397 Marseille cedex 13, France

Abstract

Today, a large quantity of information may be retrieved online, and downloaded into PCs. The most recent hard disks have a storage capacity of more than 300 M/bytes. This means that there are no restrictions to the most powerful help available in this decade in information analysis. The availability of large numbers of references, which can be implemented by your own data, open a new field of research: the PPOS concept (Post Processing of Online Searches). This analysis of downloaded bibliographic outputs leads to a new type of information, used in technological surveys, to follow trends or to map the activities corresponding to critical factor success. This paper will not deal with applications of software commercially available, but will present different ways to analyze offline databases dealing with patents. Some examples will map the activity of large sectors according Chemical Abstracts sections, or Derwent Classification or Manual Codes, others will deal with deeper and precise comparisons of research subjects, productions, and even geographical locations or companies.

I — Material and methods

a — Material

The materials are very simple: downloaded data from various databases, or data obtained from user-created databases. They are in ASCII format and are submitted before the analysis to various tests, to be sure that no references are duplicated and that all the fields to analyze are present.

The PC used needs generally an arithmetic co-processor, and a fast hard disk of at least 20 M/bytes. An EGA graphic facility is usually necessary.

For the presentation of the results, it is often interesting to develop a scenario. This is made by using the IBM Story Teller software.

All the results presented here have been obtained using the databases from *Orbit Information Technologies* which provides one of the better patent clusters; the software has been developed in our laboratory [1].

b — Methods

The methods used vary according to the databases and the goal of the analysis. But, a general principle is as follows: all parts of a bibliographic reference which represent a significant amount of information may be analyzed. This means that all the reference fields are considered as potential sources of indicators. This leads to three different aspects of information processing:

* before working offline with the data, the various information that the host can provide directly must be considered. They are grounded to the software facilities, and the care with which the databases have been loaded. This will include also the use of the GET and related commands.

> e: Various counts, ratios, from direct host data. Examination of the scattering of words, codes, etc. in a reference field for a question.

* the GET commands give rise to files which can be downloaded. They are when worked offline a material of a very high potential. This is the beginning of the PPOS.

> e: Offline treatment of GET results. This is one of the most powerful ways to achieve fast comparisons, and very precise definition of the potential in patents or research for a company, a laboratory, or a town.

* the bibliographic fields will be analyzed according to desired goals. This will necessitate various software, according to the type of groups used to present the results.

> e: Offline analysis of downloaded data. There, various approaches are possible: various sorts, infographic presentations of results, networks, groupings, etc.

We will give now different examples which will present various possible indicators related to the preceding treatments.

II — Various counts and ratios

Very often, simple examination of the data obtained directly are sufficient to provide a quick overview of the subject. This can be made easily it is considered, before going online, how the facilities offere may be used. This will usually require thought as to whether to include in the search a few more questions about the geographical locations of the references, the

production over years, the production of some important competitors, the knowledge of the document type profile, etc.

Question: what is the position of coal chemistry in academic research?

The choice of the database is Chemical Abstracts since CA deals with patents and papers. Then within the database academic interest will be related to the number of theses. The next step will be to create a significant indicator which will show for various parts of the chemistry the academic interest. This indicator will be the ratio of the total number of theses in a research subject divided by the total number of items produced in this field.

The results are indicated in Figure 1. Most of the ratios varied between 1% and 2%, but for coal research the ratio is only 0.1%. This large difference is significant and indicates that academic research is not very concerned with coal chemistry. This is also confirmed by the amount of papers produced in this field.

These type of indicators are very general. The method is to define a ratio

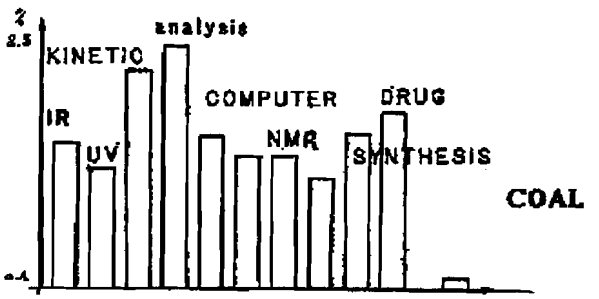

Figure 1

between two (or more) information units, and to repeat it over a sufficient number of queries. Countries, companies, document types, analytical techniques, may be one of the units, the others are provided by one or several keywords according to needs.

Because databases are numerous and well documented, the set of indicators developed may be very large. An example as been provided by I. Dimo to compare technology trends over various countries [2].

III — Offline treatment of GET results

The GET command and its homologues have been studied by different authors [3,4]. But the result of a GET command is only the representation of the scattering of various terms, codes, keywords, contained in a reference field. This is a good assistance to built up questions, or to view the classification of the main companies, countries, codes within a set of answers, but this information is in our opinion not sufficient to be considered as an indicator for exact comparisons, or detection of significant small changes in research trends. This reason prompted us to develop a new way (and software) to process the GET results.[5]

Question: how can one detect small changes in research policy or research productions or orientations? How can one define exactly what is common or different from two queries (e.g. two patent assignees)?

According the Zipf distribution [6], in a GET result the number of individual terms increases when their frequencies decrease. For instance, the number of terms analyzed at frequency 10 will be a lot smaller that the number of terms appearing at frequency 2 or 1. But, we also know that if we want to have meaningful information, we must have it at its very start and this means at low frequencies.

To solve this question, we will combine the power of the statistical GET command, with an offline analysis of its results. Let us summarize the treatment in Figure 2:

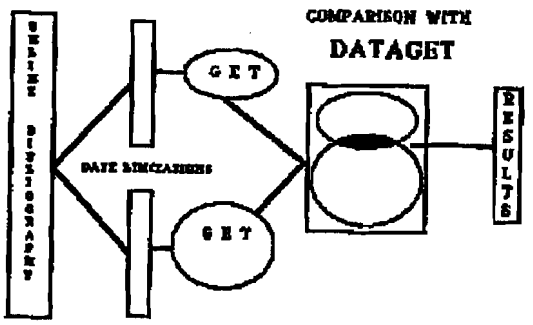

Figure 2

Examples: in 1987 and 1988, what are the chemical compounds used simultaneously by the research teams of the two French towns Marseille and Grenoble?

Step 1: definition of the corpus: Marseille/CI, Grenoble/CI OR Heres/CI. Limit the corpus to 1987 and 1988: 1 AND 87-87, 2 AND 88-88

Step 2: perform the command GET on IT (index terms for both answers. GET IT RANK TOP xxxx SS y, where xxxx is the number of terms used for the GET (1000 for Marseille and 1500 for Grenoble), and y is the research step. The number of items in the GET where guessed to go approximately till frequency three. (That is to say three papers published during this period of time). The two files are downloaded.

Step 3: clean the GET, to get rid of unnecessary data, or reformat it (spaces, format of frequencies); a GET file may change over a period of time.

Extract from file Marseille 1000 after 'cleaning':

42	GENE AND GENETIC ELEMENT
34	PROTEINS
29	RECEPTORS
25	ANTIGENS
21	AMINO ACIDS
20	CONFORMATION AND CONFORMERS
20	GENE AND GENETIC ELEMENT, ANIMAL
20	PANCREAS
19	CELL MEMBRANE
18	BRAIN
17	KINETICS
17	PROTEIN SEQUENCES
17	7440-21-3
16	7439-92-1
16	7440-50-8
15	AMINO ACIDS, BIOL STUDY
15	GEOLOGICAL SEDIMENTS
15	KINETICS, ENZYMIC
15	LIVER
15	9001-62-1
14	BIOLOGICAL TRANSPORT
14	CRYSTAL STRUCTURE
14	FATTY ACIDS

15 highest index term of file MAIDOC
source C.A.S. 1987-1988

GENE AND GENETIC ELEMENT	42	
PROTEINS	34	
RECEPTORS	29	
ANTIGENS	25	
AMINO ACIDS	21	
CONFORMATION AND CONFORME	20	
GENE AND GENETIC ELEMENT,	20	
PANCREAS	20	
CELL MEMBRANE	19	
BRAIN	18	
KINETICS	17	
PROTEIN SEQUENCES	17	
7440-21-3	17	
7439-98-1	16	
7440-50-8	16	

Orbit Information Technologies, date: 06-21-1989

16 highest AN appearing in Marseille
Lower frequency = 3

40-21-2	337	
39-92-1	307	
40-50-8	267	
40-22-4	267	
-28-5	265	
335-13-2	231	
-85-0	215	
40-70-2	187	
39-89-6	187	
40-02-0	177	
39-55-4	157	
-17-5	156	
01-62-1	159	
04-54-9	147	
40-60-6	157	

Orbit Information Technologies, date: 06-21-1989

Figure 3

Figure 3

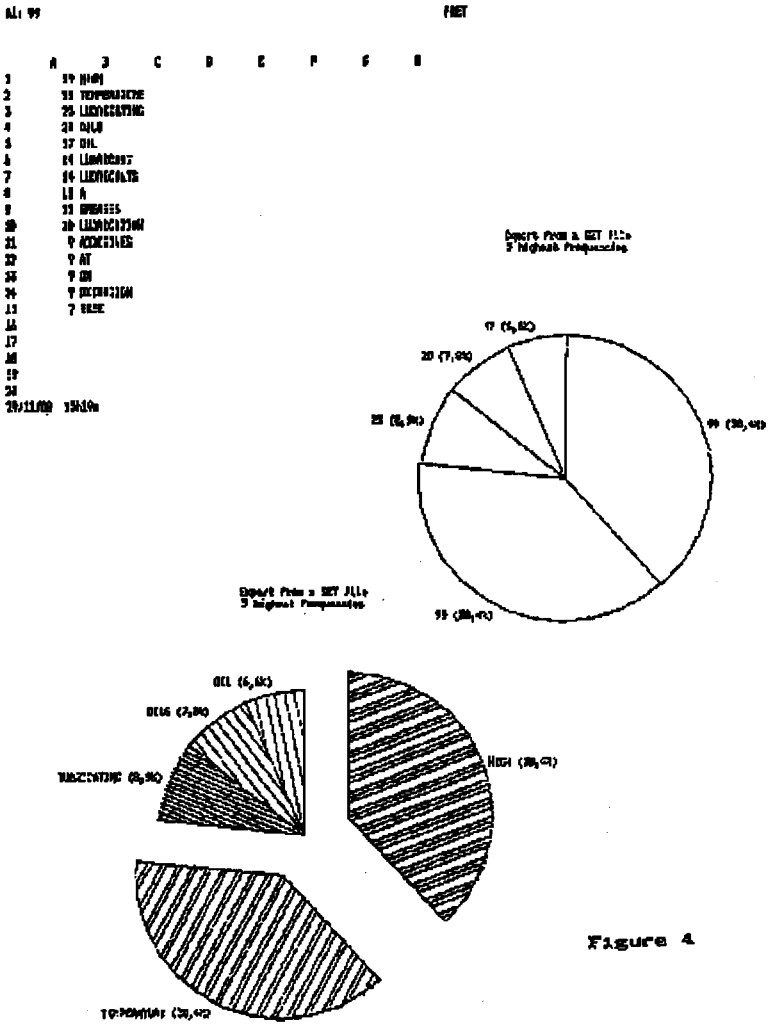

Figure 4

Step 4: determine the GET frequencies profile, to help select the comparison threshold.

Frequency profile of MA1000:

frequency	item number
2	499
3	191
4	106
5	42
6	43
7	18
8	27
9	12
10	6
11	10
12	5
13	2
14	6
15	4
16	1
17	2
18	1
20	2
21	1
29	1
42	1

file selected from frequencies 3 to 42 contains 500 index terms.

Step 5: obtain a graph of the result of the GET (histogram or pie chart)

Step 6: if further calculations are necessary on the GET file transfer it to Lotus 123 (format .PRN)

Step 7: extraction of RN from the two GET files (because IT contains both text and RN). The result will be in a GET format, then further work will be possible.

151 RN are extracted from Marseille and 117 from Grenoble. (Between the upper frequency and the frequency 3)

frequency	RN
33	7440-21-3
30	7439-92-1
26	7440-50-8
26	7440-22-4
26	57-88-5
23	12385-13-6
21	56-86-0
18	7440-70-2
18	7439-89-6
17	7440-02-0
16	7439-95-4
16	64-17-5
15	9001-62-1
15	7704-34-9
15	7440-66-6
15	7440-57-5
15	7440-44-0
14	7429-90-5
14	1333-74-0

When the RN are extracted, an ASCII file is automatically built. This file has a format suitable to send the RN with Xtalk (or other software) to the ORBIT host, and performed a search on the Chemical Dictionaries. (the term prompt must be set to tilda (ALT 126 = ~) in both Xtalk and ORBIT (after login use the TERM PROMPT ON command).

Example of file

ERASEALL (question number is reset to 1)

/RN 7440-21-3 OR 7439-92-1 OR 7440-50-8 OR 7440-22-4 OR 57-88-5

/RN 12385-13-6 OR 56-86-0 OR 7440-70-2 OR 7439-89-6 OR 7440-02- /RN

HIS (retype the research steps and the results)

Step 8: compare the two files in different ways: (*)

i — common RN for Marseille and Grenoble
ii — RN only present in Marseille
iii — RN only present in Grenoble

Example of RN common to Marseille and Grenoble from F max to F=3: The frequency of RN from Marseille appears first.

frequency	RN	
23	12385-13-6	Marseille
6	12385-13-6	Grenoble
7	14265-44-2	Marseille
10	14265-44-2	Grenoble
6	12586-59-3	.....
4	12586-59-3	
5	109064-29-1	
49	109064-29-1	

Example of RN present only in Marseille from F max to F=3:

frequency	RN
32	7439-92-1
26	7440-22-4
26	57-88-5
23	12385-13-6
21	56-86-0
18	7439-95-4
17	7440-02-0
16	64-17-5
15	9001-62-1
15	7704-34-9
15	7440-66-6

Step 9 : transfer these files if necessary to Lotus 123; the new file has a .PRN format.

Step 10: allows a comparison file to be built. These are the target data. This file in a GET format can be used for further comparisons.

Alert index:

1	109064-29-1	All frequencies are
1	64-29-1	set purposely to 1
1	1333-74-0	
1	64-17-5	
1	1810-67-0	
1	1939-14-5	

Example RN common to Marseille and to the index alert file:

frequency	RN	
1	109064-29-1	index
5	109064-29-1	Marseille
1	1333-74-0	index
14	1333-74-0	Marseille
1	64-17-5	index
16	64-17-5	Marseille

This set of comparisons can be performed on various fields (the IT field from Chemical Abstracts is the most complex, since it includes numbers [RN] and text). In some cases, the GET command may give unexpected results e.g. SO, this field can be extracted from downloaded data, and formatted as a GET. Of course, in this case, comparisons will be performed from the beginning to the end of a single line beginning with SO and extracted from the downloaded data.

(*) The comparisons can be made from 1 to 50 characters. This is important since it allows expressions to be more or less precise. (eg. SUKUMITO HYRU and SUKUMITO PRIGHITO can be seen as SUKUMITO, with a frequency of 2 or as two separate items). Notice that even if a GET TOSEL command be used allowing the SELECTion of one or more terms sorted by the GET, the computer will take all the letters from this term and will not allow a selective comparison according a variable key.

Most of the databases supporting the GET command can be used as a starting point. From our point of view, the storage of the appropriate GET files over years is a very powerful means to follow technology and research trends.

Example: What are the changes in technology and research in the field of high temperature lubrica(tion, ting, nt)?

The strategy used is very simple for this example high(W)temperature AND lubrica: on file WPIL, and can be improved if necessary. The field MC (Manual Codes) is submitted to a GET.

Three periods of time have been selected: 1988–1987–1986.

According the period of time the comparison of the GET files will give the answers. Common terms, or terms present only in one file will allow trends in technology to be followed.

For more information on this question, the PN from WPIL have been transferred into the Chemical Abstracts database, and the IT fields of the document type retrieved have been analyzed with a GET command and the GET file store. The RN have been extracted, and compared. This gives a list of the chemicals used. The software allows also to format an ASCII file containing the RN. This file has be transferred under Xtalk to the ORBIT Chemical Dictionaries and the PARENTS analyzed to obtained a core of structural information. Further uses of the RN, and selection of P/DT (Patent documents) will give a comparison with non-patent documents [7].

Partial results are as follows:

→ indicates when unusual behaviour occurs.

Manual codes common to 1988 and 1987:

4	A12-W02A	
6	A12-W02A	→ to be examined closely
2	A12-W02	
2	A12-W02	
2	E05-G08	
1	E05-G08	
2	E10-E04G	
1	E10-E04G	
2	H07-A	
5	H07-A	→ to be examined closely
2	H07-C	
1	H07-C	
1	E05-B01	
1	E05-B01	
1	E05-G09C	
2	E05-G09C	
1	E10-E04K	
1	E10-E04K	
1	E10-G02G	
2	E10-G02G	
1	E10-G02H	
2	E10-G02H	
1	H07-D	
3	H07-D	→ to be examined closely
1	L02-G08	
1	L02-G08	
1	L02-H02B2	
1	L02-H02B2	
1	L02-J01	
1	L02-J01	
1	M21-A06	
1	M21-A06	
1	M22-H03B	
1	M22-H03B	

Manual codes common to 1988 and 1986:

4	A12-W02A	
5	A12-W02A	
2	A12-W02	
1	A12-W02	
2	H07-A	
1	H07-A	
2	H07-C	
4	H07-C	→ to be examined closely
2	H07-X	
1	H07-X	
2	S03-E14F	
1	S03-E14F	
1	E05-A	
1	E05-A	
1	E10-A14	
1	E10-A14	
1	E10-B04A	
1	E10-B04A	
1	E10-D03A	
1	E10-D03A	
1	H07-D	
4	H07-D	→ to be examined closely
1	L02-H04	
1	L02-H04	
1	M21-B03	
2	M21-B03	

Terms common to 1988-1987 and 1986: (terms of both former files are combined):

A12-W02A
A12-W02
H07-A
H07-C
H07-D

This result shows a large dispersion of technology areas, since only a few manual codes are common.

Terms from 1988 differ from 1987 and 1986: (files 1986 and 1987 are concatenated and file 1988 compare with this later file)

2	E10-D03D
2	H07-G
2	H07-G04
1	A04-B05
1	A04-C04
1	A04-E02D

1	A04-F01A
1	A05-E08
1	A05-J11
1	A06-A00E
1	A06-D
1	A10-E04A
1	A10-E08A
1	A12-B01F
1	A12-H03
1	E05-G07
1	E05-G09D
1	E06-D05
1	E06-E01
1	E06-F01
1	E07-D09C
1	E07-D13C
1	E07-E01
1	E07-E04
1	E07-F01
1	E07-F03
1	E07-H
1	E10-A04A
1	E10-A10B
1	E10-A15A
1	E10-A15F
1	E10-A20
1	E10-C03
1	E10-C04K
1	E10-C04L
1	E10-E04B
1	E10-E04M
1	E10-F01
1	E10-F02A3
1	E10-F02C
1	E10-G02F
1	E10-G03
1	E31-N02
1	E34-B02
1	E34-D03
1	G02-A02B
1	G02-A02D2
1	G03-B02E
1	G06-F08A
1	H06-D03
1	H07-G01
1	H07-G03
1	H07-H
1	J04-C02
1	L02-A04
1	L02-H
1	L02-J02B

1	L03-A02B
1	L03-E01B3
1	M22-G03A
1	M22-G03D
1	M22-H02
1	M22-H03A
1	M22-H03G
1	S03-E01
1	S03-E04B1
1	V04-L01B
1	V06-M12
1	X11-J03
1	X16-E01
1	X25-B02B
1	X27-A02A1
1	X27-B

Note the large dispersion of manual codes. This indicates that the technology is moving and that research in many directions is still going on.

On this question, citations of patents or academic papers in US patents may give more information. The post treatment of CT (cited field) of US patents will be discussed by the same authors at the London Online Meeting, December 1989.

IV — Offline analysis of downloaded references with MS/DOS commands

MS/DOS commands [8] can be used very effectively to perform some bibliometric analysis. This shows that if desired, something can always be done with data.

Let us consider the various steps of a BIBLIOMETRIC analysis. We will go over these various steps, showing how MS/DOS allows you to solve part of the problem. The starting point is a bibliography performed on WPIL (the subject is not meaningful since only the approach is interesting). This example has been developed here for teaching purpose. Those who are not familiar with post processing will be able to repeat this treatment with their computer.

Example of downloaded data:

Name of the file: EXAMPLE (Number of references 52)

```
-1-
TI - Blood reservoir for extracorporeal blood treatment
       circuit - has storage section with discharge port in
       bottom, and includes partition(s) disposed vertically
       to suppress flow
```

```
DC - P34
PA - (TERU) TERUMO CORP
PN - EP-292395-A 88.11.23 (8847) AU8816459-A 88.11.24 (8903)
PR - 87.05.19 87JP-122243 87.12.18 87JP-320760 87.12.25
       87JP-331480 87.12.29 87JP-333257 87.12.28 87JP-333608
IC - A61M-001/36
NP - 2
-2-
TI - Blood storage appts. - has blood extraction tube with
       removable cover and slide
DC - P31
PA - (SARS-) SARSTEDT W KUNST
PN - US4769025-A 88.09.06 (8838)
PR - 85.11.15 85US-798761 87.05.12 87US-048701
IC - A61B-019/00
NP - 1
```

a — Verification of the numbers of fields

DOS command: FIND /C "PA -" EXAMPLE

FIND /C "TI -" EXAMPLE

results:

EXAMPLE: 52

EXAMPLE: 52

b — Extraction of one field

DOS command: FIND "PA -" <EXAMPLE> EXAMPLE.PA

result: all the lines containing PA — at the beginning are saved in file EXAMPLE.PA

PA — (TERU) TERUMO CORP
PA — (SARS-) SARSTEDT W KUNST
PA — (BAXT) BAXTER TRAVENOL LABS INC
PA — (CHAT/) CHATELAIN N M
PA — (TERU) TERUMO CORP
PA — (KAWA-) KAWASUMI KAGAKU-KOG

c — Cleaning file EXAMPLE.PA (before numbering the lines if necessary):

DOS command: EDIT EXAMPLE.PA

SHIFT F2 kill one line, SHIFT F5 save the file.

(EDIT.COM from MS/DOS Olivetti)

Note that from now, the numbers of PA fields are identical to the numbers of references, the line NUMBER of the PA in increasing order in file EXAMPLE.PA is identical to the reference number in the original bibliography.

d — Numbering the PA line in file EXAMPLE.PA

DOS command: FIND/N "PA -" <EXAMPLE.PA > EXAMPLE.PAN

Result: all the lines of file EXAMPLE.PA are associated with a number identical to the reference number of the downloaded reference. e.g. the first line of the file .PAN is the PA field of the first reference in the bibliography.

[1]PA — (TERU) TERUMO CORP
[2]PA — (SARS-) SARSTEDT W KUNST
[3]PA — (BAXT) BAXTER TRAVENOL LABS INC
[4]PA — (CHAT/) CHATELAIN N M
[5]PA — (TERU) TERUMO CORP
[6]PA — (KAWA-) KAWASUMI KAGAKU-KOG
[7]PA — (TERU) TERUMO CORP

e — Ranking the Patent Assignees

DOS command: SORT EXAMPLE.PA EXAMPLE.SRT

Result: the patent assignees from file EXAMPLE.PA are sorted and stored in file EXAMPLE.SRT

PA — (BAXT) BAXTER TRAVENOL LAB
PA — (BAXT) BAXTER TRAVENOL LAB
PA — (BAXT) BAXTER TRAVENOL LABS INC
PA — (BAXT) BAXTER TRAVENOL LABS INC
PA — (BAXT) BAXTER TRAVENOL LABS INC
PA — (BAXT) BAXTER TRAVENOL LABS INC
PA — (BIOT) BIOTEST SERUM INST GMBH
PA — (BIOT-) BIOTEST PHARMA GMBH
PA — (BUCA/) BUCALO L
PA — (BURT/) BURT R T
PA — (CHAT/) CHATELAIN N M

f — Typing file EXAMPLE.SRT (printer or screen)

DOS command: TYPE EXAMPLE.SRT > PRN (list file on printer)

TYPE EXAMPLE.SRT | MORE (list file on screen)

g — Search for the reference numbers which contain a certain patent assignee:

DOS command: FIND "BAXTER" < EXAMPLE.PAN (listed on screen)

FIND "BAXTER" < EXAMPLE.PAN > RESULT

In this case the results are stored in file .RESULT

Result: the lines containing the reference numbers and the patent assignee will be listed or recorded. Note that this is file EXAMPLE.PAN which is used.

[3]PA — (BAXT) BAXTER TRAVENOL LABS INC
[10]PA — (BAXT) BAXTER TRAVENOL LABS INC
[33]PA — (BAXT) BAXTER TRAVENOL LABS INC
[36]PA — (BAXT) BAXTER TRAVENOL LABS INC
[44]PA — (BAXT) BAXTER TRAVENOL LAB
[46]PA — (BAXT) BAXTER TRAVENOL LAB

h — Verification

```
-3-
TI - Blood contacting material with prolonged anti-coagulative
       activity - comprises porous polymer impregnated with
       collagen bonded to protamine and heparin
DC - A96 B04 D22 P34
PA - (BAXT ) BAXTER TRAVENOL LABS INC
PN - EP-282091-A 88.09.14 (8837) J63290573-A 88.11.28 (8902)
       {JP}
PR - 87.03.13 87US-025670
MC - A12-V03B B04-B04A6 B04-C02E B04-C03 B11-C04 B12-H02
       B12-M10B D09-C01
IC - A61L-033/00
NP - 2
-33-
TI - Addn. of sterol to whole blood on storage - to preserve
       red cell morphology and suppress spontaneous haemolysis
DC - B01 D22 B04 P34
```

```
PA - (BAXT) BAXTER TRAVENOL LABS INC
PN - US4432750-A 84.02.21 (8410)
PR - 81.12.02 81US-326772 MC - B01-D02 B04-B04D B11-C06
       B12-M06 D05-H
IC - A01N-001/00 A61K-035/14 A61M-001/03
NP - 1
```

Note also, that this combination of DOS commands may be made within a batch file, and using the option %, be performed automatically. The only limitations arise from the line length which can suppress some data (most hosts have a maximum line length of 131 characters). This is why DOS commands are more accurate (as far as bibliometry is concerned) with short fields.

V — Offline analysis of downloaded references with appropriate software. Analysis of the Derwent Classification

The goal is to map a set of patents, according the DC codes and to draw the technology (or research) network of the subject. The maps should be easy to read, colourful, and obtained quickly.

a — Mapping the Derwent Codes

From previous work [9], we have develop an infographic representation of research poles on a chess board where various sections (CC from Chemical Abstracts for instance, or DC from WPIL in this case) are represented on each chessboard according an increasing frequency number. Section 1 is at the upper left and the last section at the lower right of the chess board. We have also described how an EGA screen allows a good presentation of about 150 square cases. A lower number is often better. The sections or codes are represented as cylinders, and the height is proportional to their frequency in the database.[10]

To be able to work with the WPIL Derwent Classification, we have divided it into eight areas. Each of these areas being homogeneous as much as possible: eg. general, sx-elect, mechanical ...

Example: Chemdoc.

01=section number, Name and contain, E11=DC

01 General organic. Containing E11
02 General organic. Organometallics E12
03 General organic. Heterocyclics E13
04 General organic. Aromatics E14
05 General organic. Alicyclics E15
06 General organic. Aliphatics E16

07 General organic. Other aliphatics E17
08 General organic. General hydrocarbon mixtures E18
09 General organic. Other organic compounds general E19
10 Dyestuffs. Azo E21
11 Dyestuffs. Anthracene E22
12 Dyestuffs. Heterocyclic E23
13 Dyestuffs. Other dyes, all precursors E24
14 General inorganic. Compounds of metals of groups Va E31
15 General inorganic. Compounds of metals of groups IVa E32
16 General inorganic. Compounds of metals of groups IIa E33
17 General inorganic. Compounds of group Ia metals E34
18 General inorganic. Ammonia E35
19 General inorganic. Non-metallic elements E36
20 General inorganic. Mixtures of many components E37
21 Separation, evaporation, crystallisation, chromato, extract. J01
22 Mixing, crushing, spraying, dispersing, atomaising (other than paint) J02
23 Electrochemical processes and electrophoresis (ozone, water, .) J03
24 Chemical/Physical processes/ apparatus catalysis catalysts colloid-chem. J04
25 Boiling and boiling apparatus, steam generation unless poxer plant J05
26 Storing or distribution gas or liquids pipes (excluding oil gas and HC) J06
27 Refrigeration, ice, gas liquefaction, solidification, separation J07
28 Heat transfer and drying, include steam condensers, exchangers J08
29 Furnaces, Kilns, ovens, retorts, furnaces J09

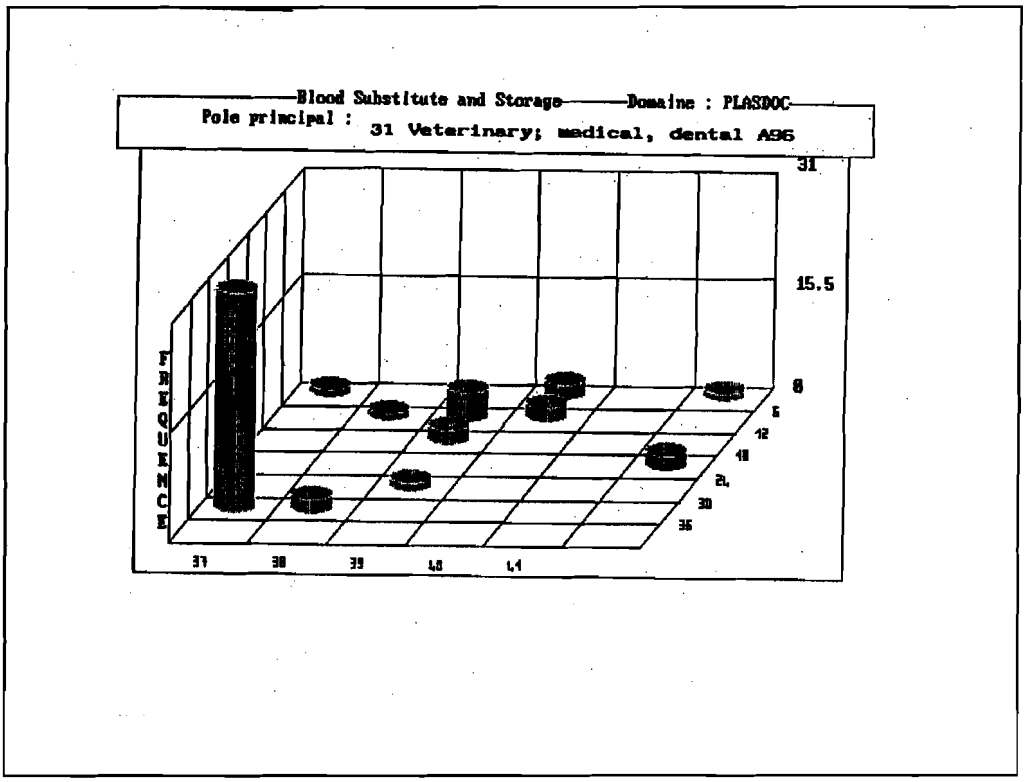

Figure 5

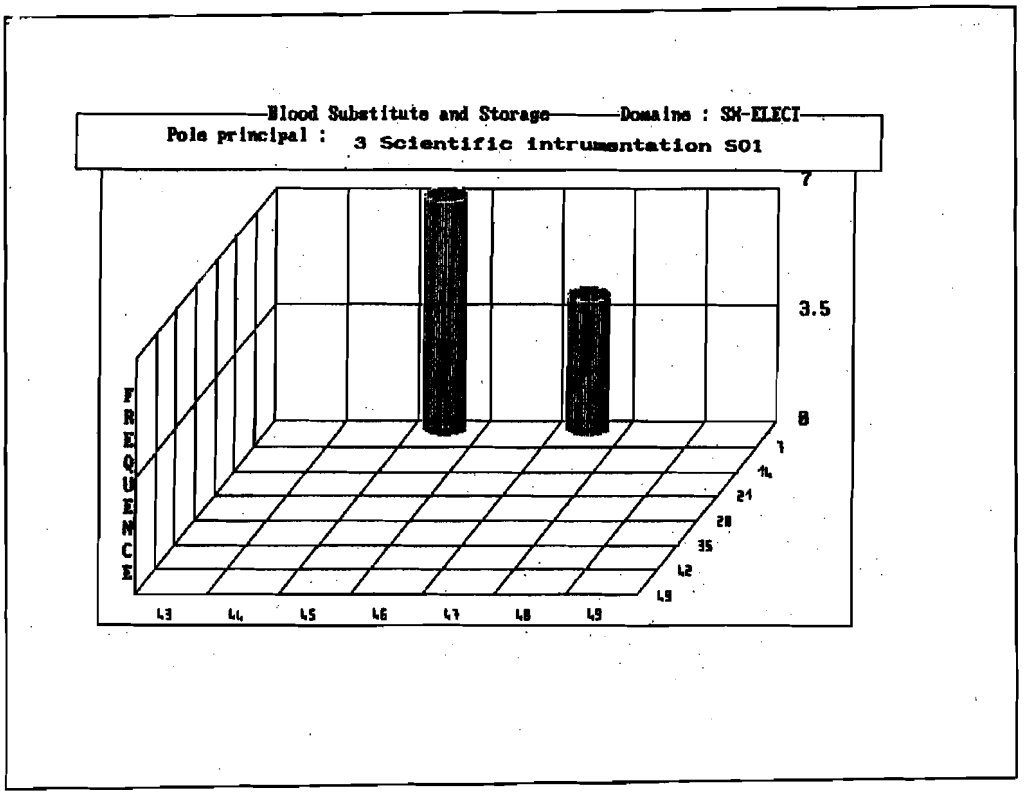

Figure 6

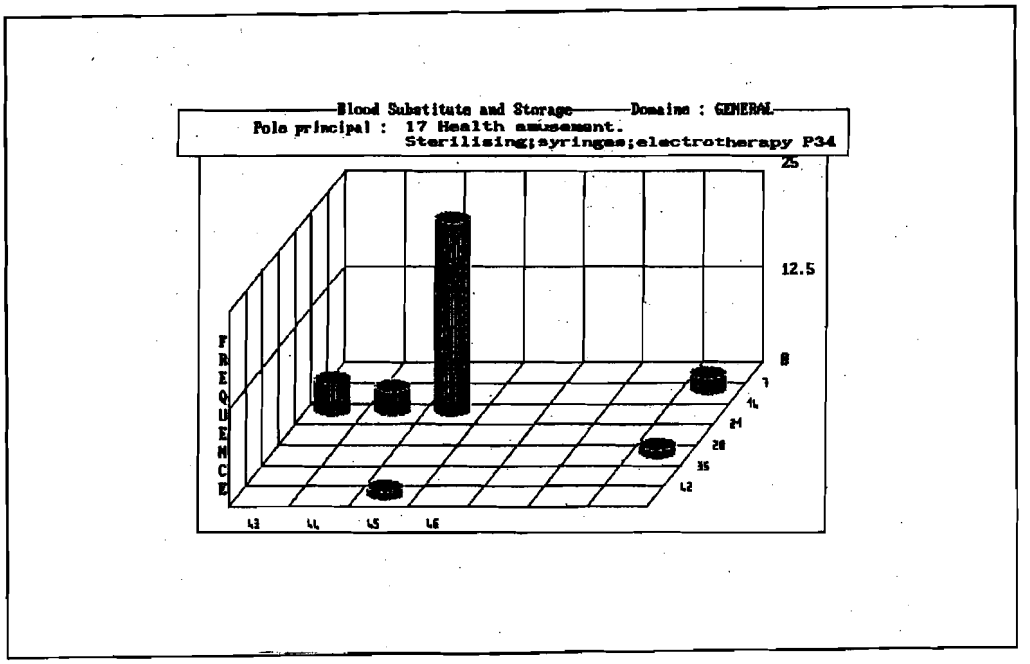

Figure 7

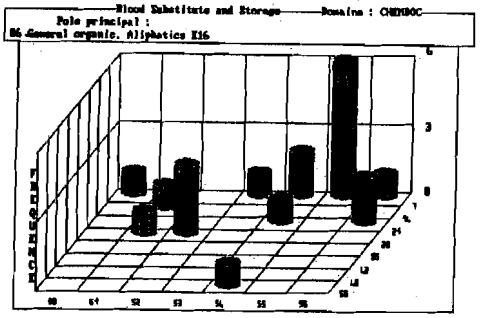

Figure 8

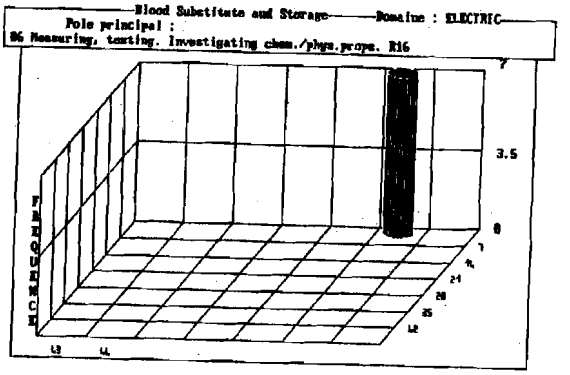

Figure 9

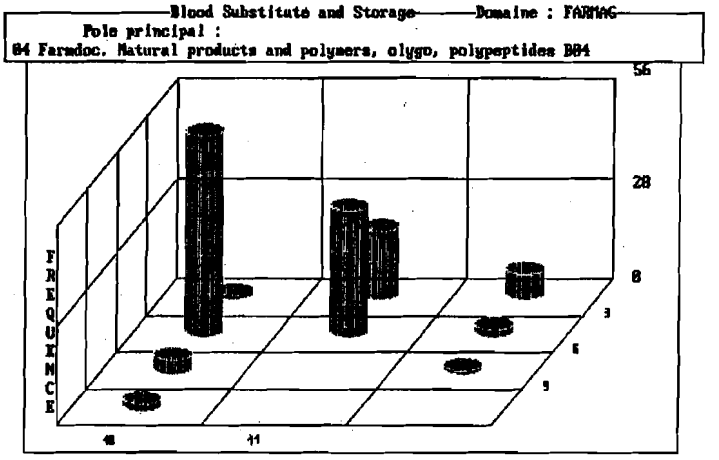

Figure 10

30 Fire Fighting, fire extinguishing, excld. fire engines, clothing K01
31 Protextion again chemical warfare, fuses, blasting exclud. missiles K02
32 Explosives, matches, includ. detonators, lighter, chem.lasers, smokes K04
33 Nuclear reactors and simulators. Exclud. power plants K05
34 Nuclear power plant — Includ. reprocessing of nuclear fuel K06
35 Health physic, includ. radiation protect. decontamination, wastes K07
36 Nucleonics, X-rays, neutrons, electrons beams, plasma welding K08
37 Glass, hemical comp. furnaces, flat-glass, bottles, containers L01
38 Refractories, ceramics, cements, soil roads, magnesia, abrasives L02
39 Electro-(in)organic, conductors, resistors, magnets, batteries,semicond. L03
40 Inorganic pigments — and non fibrous fillers G01
41 Inks, paints, polishes G02
42 Adhesives — excluding dispensers therefore G03
43 Miscellaneous compositions — incl. luminescent and tenebrescent matter. G04
44 Printing materials and processes G05
45 Photosensitive composition and bases; photographic processes G06
46 Photo-mechanical production of printing surfaces G07
47 Electrography, electrophotography and magnetography G08
48 Obtaining crude oil and natural gas — incl. explor., drilling, producing H01
49 Unit operations — incl. distill., sorption and solvent extraction H02
50 Transportation and storage — only large scale systems are included H03
51 Petroleum processing — incl. treating, cracking, reforming and catalysis H04
52 Refinery engineering H05
53 Gaseous and liquid fuels — incl. pollution control H06
54 Lubricants and lubrication — excl. self lubricating surfaces H07
55 Petroleum products other than fuels and lubricants H08
56 Fuel product not of petroleum origin H09

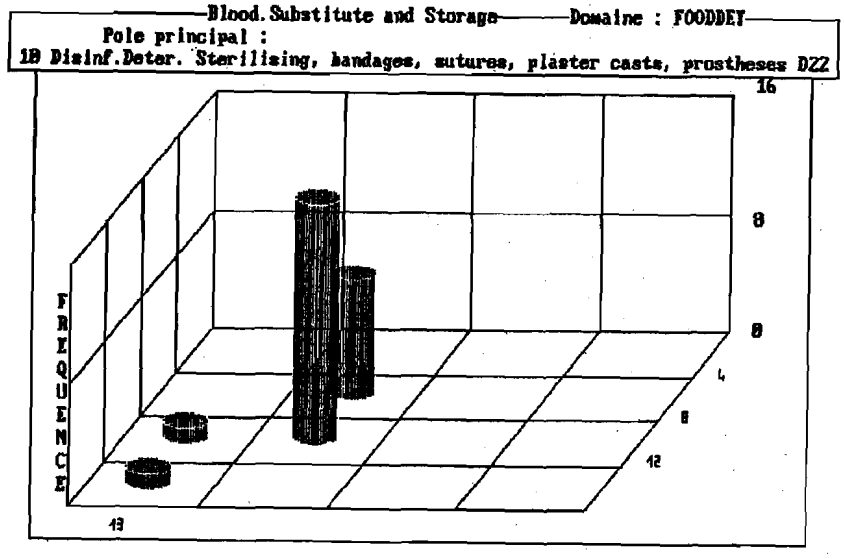

Figure 11

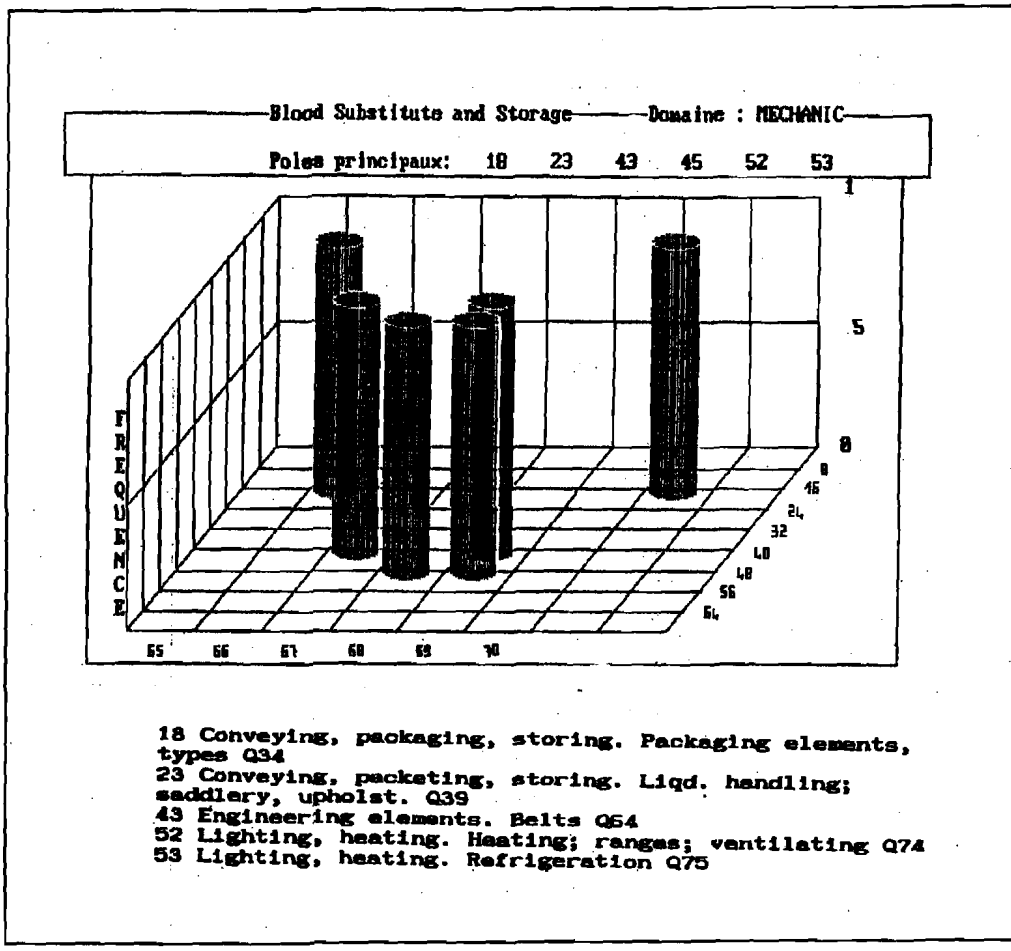

Figure 12

According to the list, the program will test, extract, sort, count and draw the maps of the eight areas. All frequencies under 25% of the maximum frequency appearing on a map will have a different colour. The work will be performed in sequences, the only information to be provided to the program is the name of the file to use. To obtain a print screen of the map, a commercial software must be used with the EGA graphic format. To capture the screen on hard disk or floppies, IBM Story Teller is used. The program can also be used with CGA configuration but the quality of the drawing is inferior.

In the following example, data from the search for blood storage and blood substitutes are mapped. This is presented in Figures 5 to 12.

The subject was selected because we wished to emphasize the differences in strategies between Japan, the Western World and USSR. Because several daily papers indicated that Japan was buying human blood on a large scale,

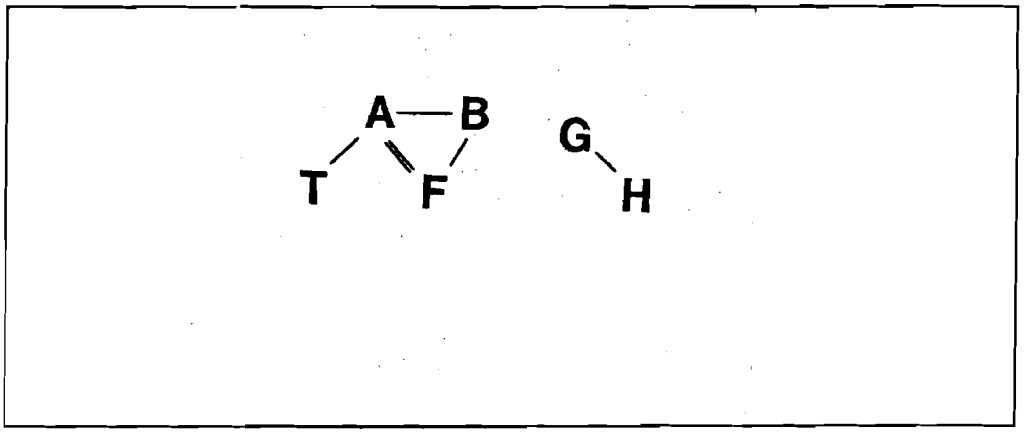

Figure 13

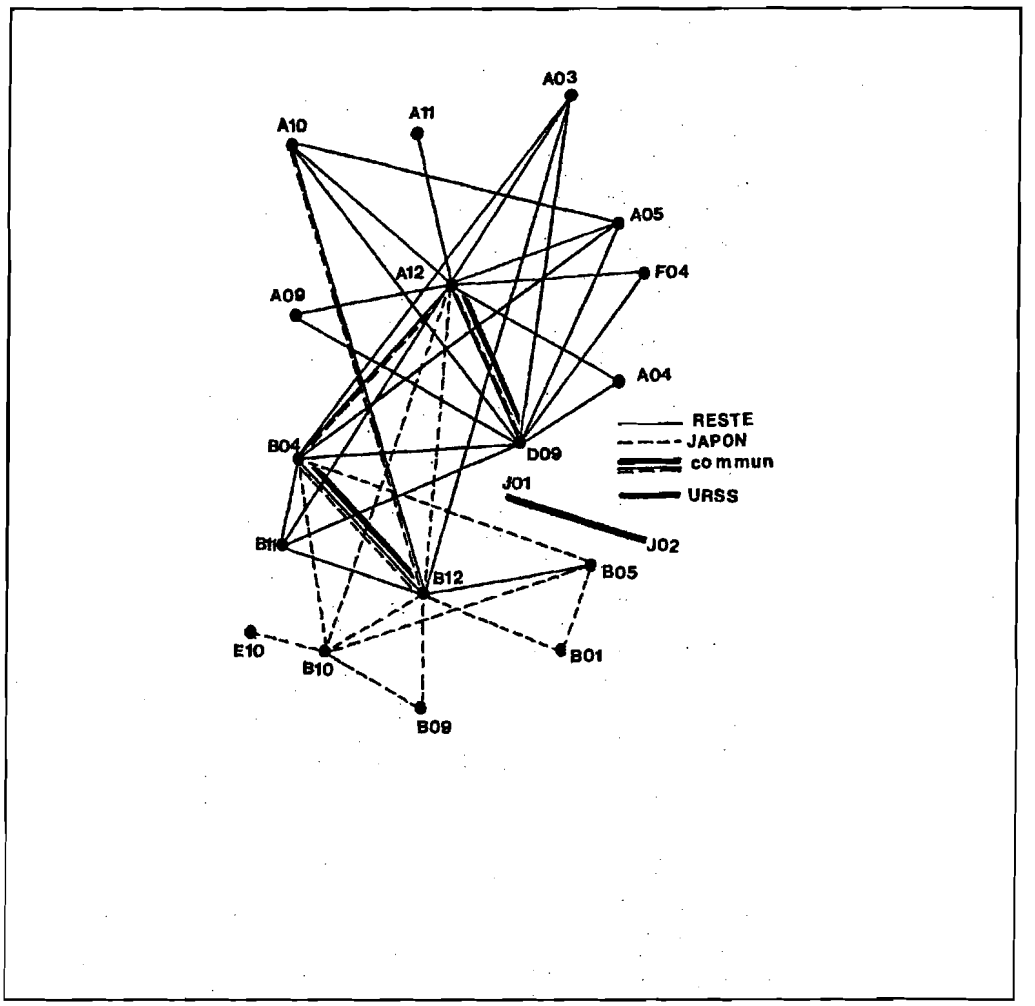

Figure 14

two concepts have been used to described the corpus of references selected in WPIL: blood storage and blood substitute.

Because of the strategic importance of the results the presentation makes in this paper, if it does not alter the general trends, has been voluntary modified on certain points.

b — Drawing the network of research using the manual codes

The manual codes (or the Derwent codes) may be used to draw the research network. To reach this goal, we consider the following:

MC — A B F
MC — F T A
MC — G H

If a set of WPIL references shows the former MC fields, we say that when several manual codes are simultaneously present in a reference, this mean that a potential connection exists between all the manual codes present. For instance if A B F are present there are three connections A-B, A-F, B-F. This view is made legitimate by the fact that the patents have been indexed by specialists, according to the same rules. Then, the three MC fields above give rise to the following network:

In Figure 14, we can see the network obtained from the manual codes. This network shows the trends in Japan, the rest of the Western World and the USSR in this area of research. The various strategies are clearly identified, and they can be followed in different periods of time.

It is also obvious that the same treatment can be applied to other subjects, to patent assignee(s), inventors, institutions, even countries.

VI — Conclusion

There is more than one way to perform a statistical analysis of downloaded information. In this paper, we have not described Factorial Analysis, Dendograms, and other clustering methods that we used in our laboratory. This is because the post processing of downloaded data may follow various routes.

These methods are selected according the goal of the analysis, the number of references — if necessary, various fields (PA, DC, MC, IN) — to analyse separately or simultaneously, and also the facilities provided offline for the analysis.

We believe that within a few years some of these treatments will be available online. The coming period must be taken as experimental and used to test various methodologies and to train people. The sets of data stored by downloading have, according to the subject and its volume, two levels of

information: one of which is the bibliography level (the contents of one reference), the second the organization of data and their correlations (hidden order). This is this level of information which is much more powerful than the former and which is the core of modern information science; it can only be reached by post processing.

We should not forget, however, that the final goal is to provide indicators, to master the information necessary to have the best possible knowledge of the critical factor success used in technological surveys [11]. Performing treatment without this goal in mind is useless.

This paper which deals with various approaches, all easy to perform even by end users, shows that it is not necessary to have large and powerful facilities to achieve good performance. It is often the capacity to analyze a problem and to think of new solutions, which will be the key to the development of good indicators. Nevertheless there are also various more complicated analyses, such as factorial analysis, groupings with automatic networking such as Clustan [12]. In these cases, only trained people should use them, because even though it may always be possible to obtain a graph, it is more complicated to interpret the graph and to devine its real meaning.

Bibliography

1. Système intractif d'aide à la décision SIAD: analyse statistique dynamique des banques de données. Albert La Téla, Thése Sciences Marseille, 1987

2. Indicateurs de développement technologique. Iona Dimo, Colloque sur l'Information Elaborée. SFBA BP 1507, 75327 Paris cedex 07, Ile Rousse p. 123, 1989

3. GET, MAP, MEM, ZOOM et les autres. Jean Pierre Lardy, *Revue Française de Bibliometrie,* 1, 23, 1987

4. The GET command: a powerful new patent searching tool from Pergamon Infoline. John Terragno, *World Patent Information,* **6**, p.69–73, 1984

5. Post Processing of GET results. GET software, available from CRRM (13397 Marseille cedex 13, France). MS DOS compatibles.

6. Bradford's law and related statistical patterns. Eugene Garfield, *Current Contents,* May 12th, 1980

7. French Orbit User Day — Paris June 20th 1989 Post processing of online searches (PPOS) Henri Dou,

8. Teaching Bibliometric analysis and MS/DOS commands. Henri Dou, Luc Quoniam, Parina Hassanaly, *Education for Information,* **6,** p.411–423, 1988

9. Mapping the Scientific network of patent and non-patent documents from chemical abstracts for a fast scientometric analysis. Henri Dou, Parina Hassanaly, *World Patent Information* 1988, 2

10. Clustering Multidisciplinary Chemical Papers To Provide New Tools for research Management and Trends. Application to Coal and Orgnaic Matter Oxidation. Henri Dou, Parina Hassanaly, Luc Quoniam, Jacky Kister, *J. Chem. Inf. Comput. Sci.* 1989, **29**, 45–51

11. 'Maîtriser l'information critique', François Jakobiak, Les Editions d'Organisation, ISBN 2708108743

12. Clustan Ltd

Acknowledgement: We will like to thank very much Orbit Information Technologies, part of the Maxwell Online Group, for its help during this work. Orbit Information Technologies offers one of the best clusters of patents databases to perform comparative analysis and to develop trend indicators.