Procedding :Fourth International Conference on bibliometrics, scientometrics, informetrics,
Berlin, Germany: ISSI, 11-15 sept, 1993

DATAVIEW: BIBLIOMETRIC SOFTWARE FOR ANALYSIS OF DOWNLOADED DATA

H. ROSTAING,* H. DOU,* P. HASSANALY,* C. PAOLI**

* *CRRM, Centre de Recherche Rétrospective de Marseille, Faculté Saint Jérome, 13397 Marseille Cedex 20, (France)*

** *Cedocar, Centre de Documentation de l'Armement, 2 bis rue Lucien Bossoutrot, 00460 ARMEES, (France)*

Most of the studies are not fully automatized, and only two points are generally integrated into a computerized process: collecting data and data analysis. To have a whole automatic bibliometric process specialists need a software which will be the bridge between these two steps. This communication presents a bibliometric software, called Dataview, which has especially been developed to change textual data into numerical data in an automatic way, and to offer various possibilities of data analysis. To have a really efficient software tool to treat scientific and technological information, one cannot restrict to analyse only to one type of bibliographic data dealt with only one analysis method. So, Dataview accept to process various formats of bibliographic data, is able to consider various kind of bibliometric items in the same study and provide numerical data suitable for various statistical techniques. As a matter of fact, this software correspond exactly to the need of companies or national institutes involved in technology watch and competitive intelligence.

Introduction

During most bibliometric treatment, it is noticeable that only a very few number of authors deal with a full automatic processing. Most of the studies are not fully automatized, and only two points are generally integrated into a computerized process (Fig. 1):

(i) for collecting data: bibliographic data are usually obtained by communication software to access to online databases.

(ii) for data analysis: statistical softwares are used to change numerical data into graphic presentations (curves, bar charts, pie charts...) and to interpret the underlying structures of data (network structures, inertia analysis, clusters analysis, multidimensional scaling...)

We note that these data are not the same at the first and second step. Online databases provide textual data and statistical softwares need numerical data. It is obvious that to have a whole automatic bibliometric process specialists need a software which will integrate these two steps.

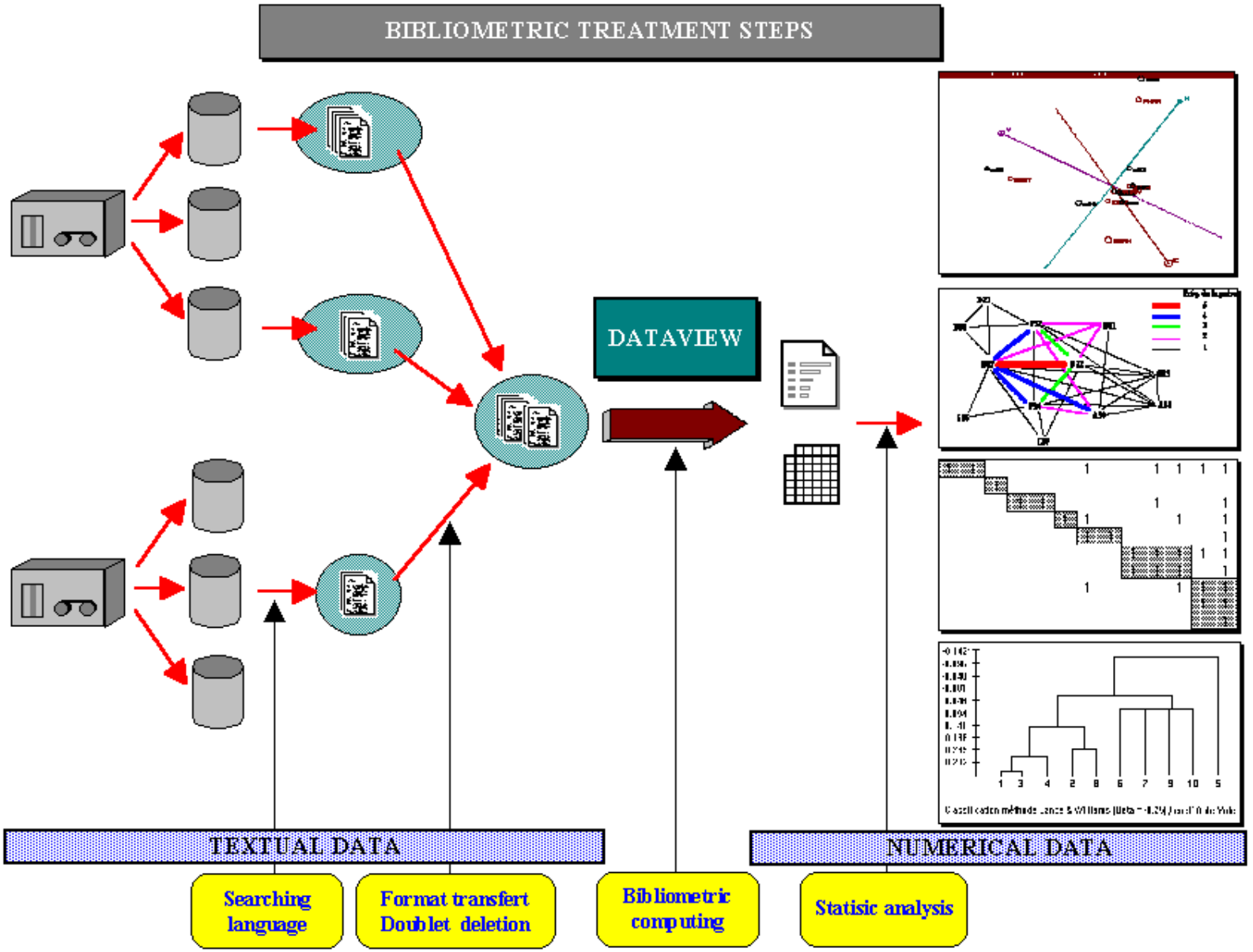
BIBLIOMETRIC TREATMENT STEPS
DATAVIEW
TEXTUAL DATA
NUMERICAL DATA
Searching language
Format transfert Doublet deletion
Bibliometric computing
Statisic analysis

Figure 1

Only a few research progress have been made in this way:
(i) Brookes and his *Bibliometrics toolbox* software[1]
(ii) Derwent LTD and its *PatStat*+ software[2]
(iii) Battelle Europe and its *Patent Trend Analysis* software[3]
(iv) CSI - CDST collaboration and its *Leximappe* software[4]
(v) and some authors who developped their own computer programmes for limited treatments like *Todorov and Winterhager*[5]...

Each of these computer tools have always been developped in a specific way: the *Brookes' Toolbox* allows to fit data with bibliometric laws, *PatStat*+ and *Patent Trend Analysis* softwares allow to analyse patents references coming from WPIL or US Patents databases, *Leximappe* is made to treat only one bibliographic field in the same study with only one possibility of analysis (co-word analysis with AHC techniques).

The CRRM* laboratory has been working on automatic data processing and bibliometric softwares for a long time. This communication presents a bibliometric software which is the result of several previous software experiments (*Datacode*[6], *Datalink*[7], *Datrans*[8]). This software, called *Dataview*, has especialy been developped to change textual data into numerical data in an automatic way, and to offer various possibilities of data analysis. This software fill the gap existing in most of the bibliometric processes[9].

Dataview: a versatile bibliometric tool

The purpose of the software conception has not been to create of a new bibliometric method. On the contrary, we wanted to provide a bridge between the various information funds and the various data analysis methods. *Dataview* is a software tool for experts of scientific and technological information processing. These experts will use this tool to build their own analysing techniques according to the most suitable statistic methods. To reach this purpose, we focused our research works on three points:
(i) *Dataview* should accept the various formats of information funds
(ii) *Dataview* should be able to consider various kind of bibliometric items in the same study
(iii) *Dataview* should provide numerical data fittable to various statistical techniques

* The CRRM is located at the University of Aix-Marseille in France, and is a research and teaching unit, especialized in Technology Watch and Competitive Intelligence.

Various formats of information funds

Usually, there are three possible funds for bibliometric studies. It is either home databases or online or CDROM databases. In the three cases information is presented in the same way. The main structure is the bibliographic reference and the sub-strucutres are the bibliographic fields (Fig. 2). *Dataview* needs three tags in the text to determine this bibliographic structure:

(i) division tag between two references
(ii) field beginning tag
(iii) field finishing tag

As *Dataview* has more especially been developped to deal with data coming from online or CDROM databases, tags are defined below:

(i) an empty line or a repetition of empy lines indicates a separation between two references
(ii) the beginning of a field is determined by its name. The field name is set at the beginning of a line and the field content is brought to alignment with this name.
(iii) a field is finished when there is a new name at the beginning of a line

Dataview allows to process any computer ASCII files if the file format fit with the previous definition. Therefore, references downloaded from online hosts like Questel, Cedocar, Orbit, Dialog, STN... are directly interpreted by *Dataview*.

Various kind of bibliometric items

Bibliometric methods does not always deal with the same kind of bibliographic information. For instance, co-citation analysis considers information coming from citation fields of the Science Citation Index database[11], co-word analysis considers information coming from keyword fields[4], co-heading analysis[5] or co-subfield analysis[11] considers information coming from code fields...(a review of different bibliometric methods was attemped[9, 12])

Dataview should allow to study of a references set using any kind of bibliometric methods. *Dataview*'s user specifies which information items he wants to process. In this way, two indications are asked to him:

(i) which fields contain informations items
(ii) which separation tags are set between information items in these fields

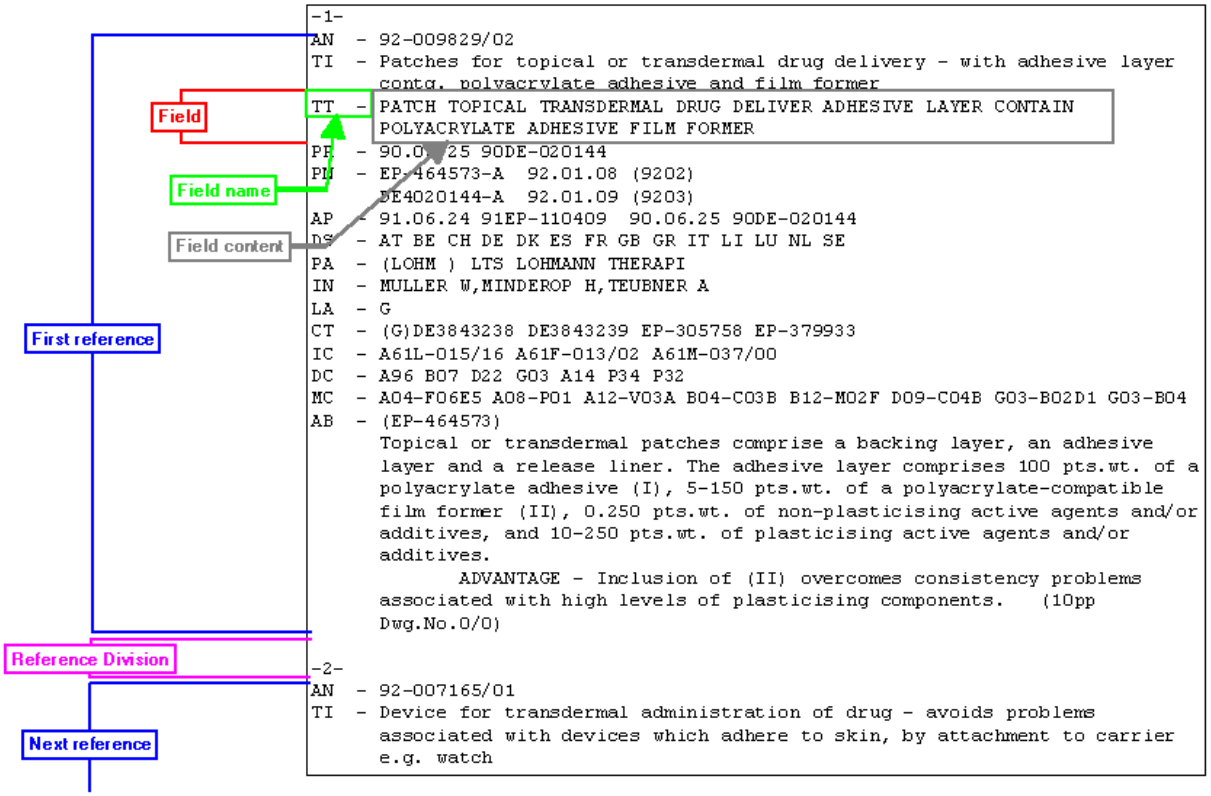

figure 2

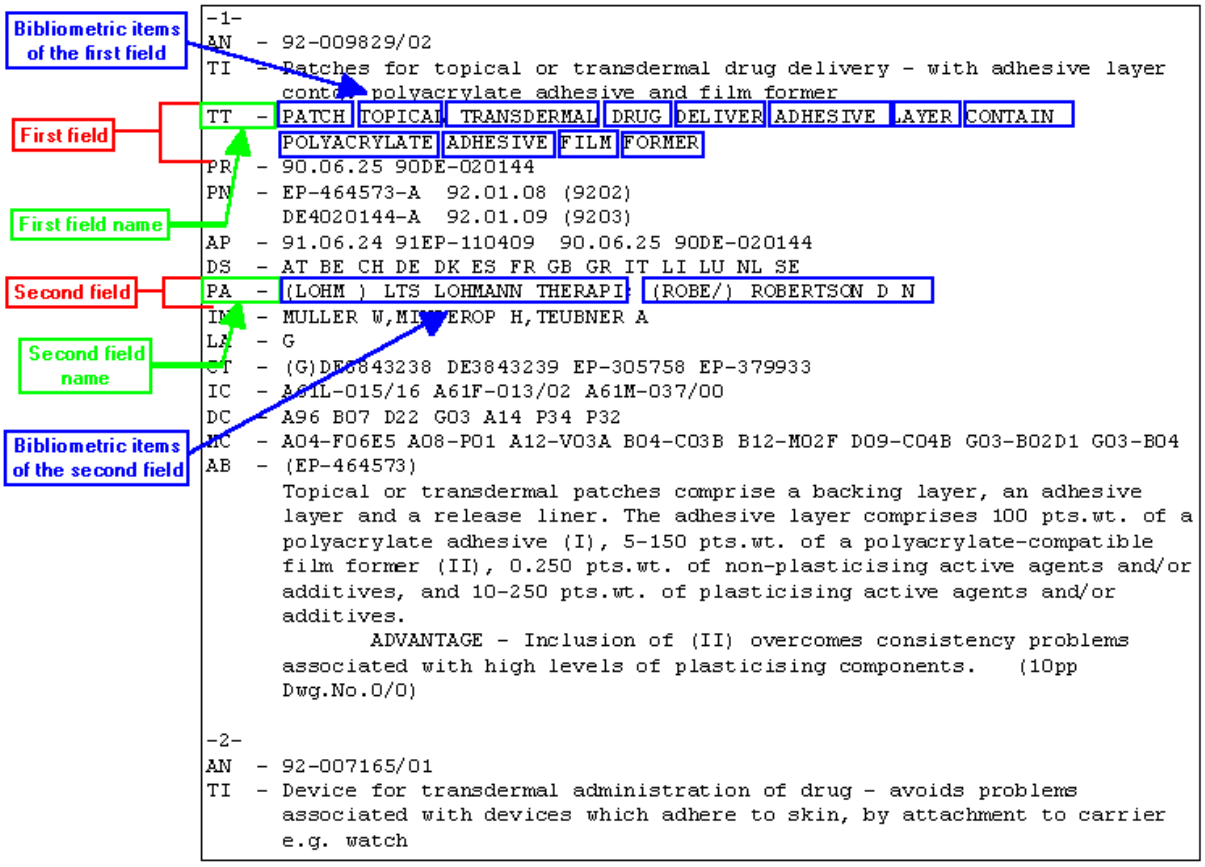

figure 3

These two data acquisition realised, *Dataview* will process all strings of characters belonging to the specified fields and enclosed by the specified separation tags. As these strings of characters can represent different kind of bibliographic items we have chosen to call them "forms" ("formes" in French). This name is the abreviation of "graphic forms" ("formes graphiques" in French) which had been already used by *Lebart et Salem*[13]. During all the process, *Dataview* considers these forms as the bibliometric items to analyse.

For instance, to study WPIL references from Orbit (Fig. 3) with co-word analysis on keyword fields, one must specified "TT - " as the name of the field and " " as the separation tag within this field. To know which research fields companies are working on, two bibliographic fields "PA - ", "TT - " are specified and their seperation tags are respectively "**;**" and " ".

Entrance to various statistic techniques

When the user introduced into the software the necessary descriptive parameters, it executes the "encoding" process* ("codage" in French). This step makes an inventory of the whole forms existing in the references set. This encoding process also draws up the bibliometric data for these forms:

(i) their locations within the references
(ii) their occurrence frequencies
(iii) their inter-relationship strengths according to co-occurrence frequencies and according to statistical association measures (see Appendix for listing of available statistical measures with *Dataview*).

Then, the *Dataview*'s user can exploit this "bibliometric database" to build his own edition of issues. *Dataview* provides the main necessary issues and the main necessary edition formats used for bibliometric analysis (Fig. 4):

(i) bibliometric distributions
  (1) size-frequency distribution for forms (data used for Lotka and Bradford laws)
  (2) size-frequency distribution for pairs of forms
  (3) distribution of forms number per field
  (4) frequency-rank distribution for forms (data used for Zipf law, logistic curve...)
  (5) frequency-rank distribution (or statistical measure-rank distribution) for pairs of forms

* This process was called "encoding" because it applies a H-coding technique to code strings of characters (the forms) to numerical values. this technique speed up the process[7].

(ii) bibliometric matrices
  (1) occurrence matrices
    (a) presence/absence matrix
    (b) total disjonctif matrix
  (2) co-occurrence matrices
    (a) symetric matrix
    (b) asymetric matrix (transaction matrix)
    (c) cross-multifields matrix (Burt matrix)
  (3) statistical association measures matrix
    (a) symetric matrix (symilarity, asymilarity or distance matrix)
    (b) asymetric matrix
    (c) cross-multifields matrix

The user can select the set of forms which will be concerned during the distributions edition. This set is chosen according to rank frequency intervals, according to field belonging, and according to mask retrieval.

In the same way, column and raw headers are chosen by the user. Therefore, he can allocate the forms, which seems to him to contain relevant interactions, to the two dimensions of the matrix. This facility to select and allocate forms allows the user to built up as well classical bibliometric matrices as his customized bibliometric matrices.

For an efficient exploitation, outcoming bibliometric data can be exported to statistic softwares. At the present time, these outcoming can be exported to softwares used for our studies:
(i) *Excel*
(ii) *StatItcf* (ITCF's statistic data analysis software)
(iii) *Clustan* (Wishart's cluster techniques software)
(iv) *Arcade* (CEMAP-IBM's relational analysis software)
(v) *Tetralogie* (IRTI's 4D representation software)

An export feature to the *Statistica* (StatSoft's data analysis software) will be developped in few months.

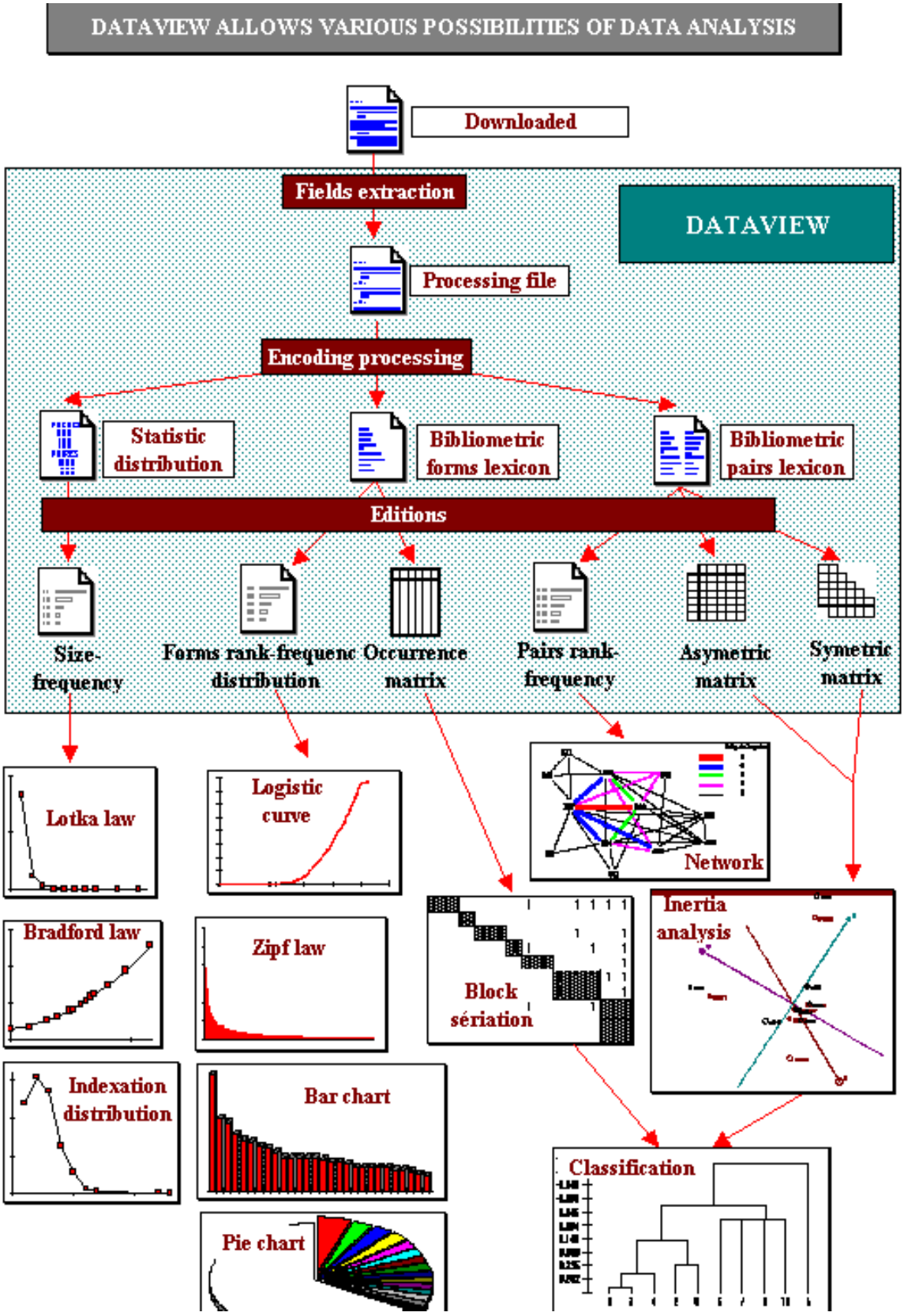
DATAVIEW ALLOWS VARIOUS POSSIBILITIES OF DATA ANALYSIS
Downloaded
Fields extraction
DATAVIEW
Processing file
Encoding processing
Statistic distribution
Bibliometric forms lexicon
Bibliometric pairs lexicon
Editions
Size-frequency
Forms rank-frequenc distribution
Occurrence matrix
Pairs rank-frequency
Asymetric matrix
Symetric matrix
Lotka law
Logistic curve
Network
Bradford law
Zipf law
Block sériation
Inertia analysis
Indexation distribution
Bar chart
Classification
Pie chart

Figure 4

Dataview software specifications

Dataview has been developped to run on DOS environment. It has been programmed using the professional Basic language (PDS 7). As a matter of fact, we actally are transfering the *Dataview* application to Visual Basic language and in a few times *Dataview* will run on Windows graphical environment.

During the development of *Dataview*, we tried expand its limits to:

(i) encoding process limits
  (1) 32 000 references
  (2) 15 000 000 digits per references
  (3) 32 000 forms (i.e. a vocabulary of 32 000 items)
  (4) 2 000 000 pairs of forms (i.e. *Dataview* indexes until two billions of relations between forms)

(ii) working limits
  (1) 9 fields can be simultaneously studied
    10 separation tags can be specified for each field
  (2) no limit for distribution editings
  (3) matrix limits: 32 000 cases x 32 000 variables
    (a) for an occurrence matrix building it represents the locating of 32 000 forms within 32 000 references
    (b) for an co-occurrence matrix building it represents more than 1 000 000 000 relations between two sets of forms

Conclusion

Nowadays, industrial world and national decision makers of science policy are involved in competitive intelligence or technology assessment. Our laboratory has been working to develop efficient software tools to facilitate in these activities. The emphasis of our research works is to provide handy tools which will allow to draw up valuable analysis.

Centres of interest for a decision maker depends widely on scientific and technical fields interesting his firm or his institute. So, necessary informations for competitive intelligence can not be found in only one database. Moreover, according to the questions of the decision makers, the bibliographic items to treat will not always be the same. And to have the best representation of relations and structures within these items, the statistic method to use can not always be the same. To have an really efficient competitive intelligence, firms can not restrict to analyse only one information fund with only one analysis method.

This paper has presented how *Dataview* software can performe in this way. The use of *Dataview* in three international companies and two national institutes (Cetim, Cedocar) give stronghly that there is not only an interest for this kind of software tool but also *Dataview* corresponds exactly to the need.

Unfortunatly, for confidentiality reasons it is not possible to publish results of these exploitations. All the same, few publications[14, 15, 16] described the methodology and part of *Dataview*[*] .

References

1. T.A. BROOKES, Bibliometrics Toolbox, Software and documentation available from City Bibliometrics 15825 6th Ave. NE. WA 98155, 1987
2. DERWENT PUBLICATIONS LTD, Rochdale House, Theodablds, London WC1 X3RP, GB
3. BATTELLE EUROPE, 7 route de Drize, CH-1227 Carrouge-Genève, Suisse
4. B. MICHELET, L'analyse des associations, Thesis: Université Paris VII, 1988
5. R. TODOROV, M. WINTERHAGER, Mapping autralian graphics: a co-heading analysis, Scientometrics, 20 (1991) No. 1, 163-172
6. H. DOU, P. HASSANALY, L. QUONIAM, Infographic analytical tools for decisions makers, Scientometrics, 17 (1989) No. 1-2, 61-70
7. A. LA TELA, Système intéractif d'aide à la décision (SIAD), Thesis: Université Aix-Marseille III, 1987
8. L. QUONIAM, Bibliométrie informatisée et information stratégique. Système automatique d'analyse des fichiers télédéchargés sur micro-ordinateur, Thesis: Université Aix-Marseille III, 1988
9. H. ROSTAING, Veille technologique et bibliométrie: concepts, outils, applications, Thesis: Université Aix-Marseille III, 1993
10. H.G. SMALL, Co-citation in the scientific literature: a new measure of relationship between two documents, Journal of American society for information science, 24 (1973) No. 4, 265-269
11. A.F.J. VAN RAAN, H.P.F. Peters, Dynamic of scientific field analysed by co-subfield structures, Scientometrics, 15 () No. 5-6, 607-620
12. A.D. WHITE, K.W. MC CAIN, Bibliometrics, Annual review of information science and technology (ARIST), 24 (1989)
13. L. LEBART , A. SALEM, Analyse statistique des données textuelles, Dunod, Paris, 1988

[*] For more informations on Dataview, contact Rostaing Hervé at CRRM, 13397 Marseille Cedex 20

14. W. NIVOL, Système de surveillance systématique pour le management stratégique de l'entreprise. Le traitement automatique de l'information brevet, Thesis: Université Aix-Marseille III, 1993

15. H. ROSTAING, W. NIVOL, L. QUONIAM, A. LA TELA, Le logiciel bibliométrique Dataview et son application comme outil d'aide à la l'évaluation de la concurrence, Proceedings of "Les systèmes d'information élaborée", Ile Rousse, France, 9-11 juin 1993

16. H. ROSTAING, W. NIVOL, L. QUONIAM, C. BEDECARRAX, C. HUOT, Exploitation systématique des bases de données. Des analyses stratégiques pour l'entreprise, Proceedings of "Journées de l'ADEST", Paris, France, 1-2 juin 1992

Appendix

Statistic association measures computed with DATAVIEW:

We have

		Form X	
		Presence	Absence
Form Y	Presence	N_A	N_B
	Absence	N_C	N_D

And $N_A + N_B + N_C + N_D = M$

N_A = number of references containing forms X and Y
N_B = number of references containing only the form X
N_C = number of references containing only the form Y
N_D = number of references containing neither the form X or the form Y
M = number of references

Name	Formula	Type
Bray & Curtis	$(N_B + N_C) / (2 * N_A + (N_B + N_C))$	dissimilarity [0, 1]
Simple matching	$(N_A + N_D) / M$	similarity [0, 1] (Sokal & Mich. 85)
Pearson correlation	$((N_A*N_D)-(N_B*N_C))/\sqrt{((N_A+N_B)*(N_A+N_C)*(N_B+N_D)*(N_C+N_D))}$	similarity [-1, 1]
Czekanowski-Dice	$(2 * N_A) / (2 * N_A + N_B + N_C)$	similarity [0, 1]
Binary shape difference	$(M * (N_B + N_C) - (N_B - N_C)^2) / M^2$	dissimilarity [0, 1]
Binary pattern difference	$N_B * N_C / M^2$	dissimilarity [0, 1]
Binary size difference	$(N_B + N_C)^2 / M^2$	dissimilarity [0, 1]
Dispersion	$((N_A * N_D) - (N_B * N_C)) / M^2$	similarity [-1, 1]
Binary Euclidean	$(N_B + N_C) / M$	dissimilarity [0, 1]
Equivalence	$N_A^2 / (N_A + N_B) * (N_A + N_C)$	similarity [0, 1]
Faith	$(N_A + N_D / 2) / M$	(Faith 83)
Hamman	$((N_A + N_D) - (N_B + N_C)) / M$	similarity [-1, 1]
Inclusion	$N_A / \min \{(N_A + N_B), (N_A + N_C)\}$	
Jaccard	$N_A / (N_A + N_B + N_C)$	similarity [0, 1] (Jaccard 1900)
Kulczynski 1	$N_A / (N_B + N_C)$	similarity [0, ∞] (Kulczynski 28)
Kulczynski 2	$(N_A / (N_A + N_B) + N_A / (N_A + N_C))^2$	similarity [0, 1] (Sokal Sneath 63)
Average squared	$(N_B + N_C) / M$	dissimilarity [0, 1]
Ochiai 1	$N_A / \sqrt{((N_A + N_B) * (N_A + N_C))}$	similarity [0, 1] (Ochiai 1957)
Ochiai 2	$N_A / \sqrt{((N_A+N_B)*(N_C+N_D)*(N_A+N_C)*(N_B+N_D))}$	
Q de Yule	$((N_A*N_D)-(N_B*N_C)) / ((N_A * N_D)+(N_B*N_C))$	similarity [-1, 1]
Rogers & Tanimoto	$(N_A+N_D) / ((N_A+N_D) + 2*(N_B+N_C))$	similarity [0, 1] (Rogers Tanim. 60)
Russel & Rao	N_A / M	similarity [0, 1] (Russel Rao 40)
Shannon	$2 * (N_B+N_C) * \text{Log}(2)$	dissimilarity [0, ∞]
Sokal & Sneath 1	$2 * (N_A+N_D) / (2 * (N_A+N_D) + (N_B+N_C))$	similarity [0, 1] (Sokal Sneath 63)
Sokal et Sneath 2	$N_A / (N_A + 2 * (N_B+N_C))$	similarity [0, 1] (Sokal Sneath 63)
Sokal et Sneath 3	$(N_A + N_D) / (N_B + N_C)$	similarity [0, ∞]
Sokal et Sneath 4	$(N_A/(N_A+N_B) + N_A/(N_A+N_C) + N_D/(N_B+N_D) + N_D/(N_C+N_D))/4$	similarity [0, 1]
Sokal et Sneath 5	$N_A * N_D / \sqrt{((N_A+N_B) * (N_A+N_C) * (N_B+N_D) * (N_C+N_D))}$	similarity [0, 1]
Binary variance	$(N_B + N_C) / (4 * M)$	similarity [0, 1]