Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics

Rosary College
River Forest, IL, USA
June 7-10, 1995

Sponsored by The Rosary College
Graduate School of Library and Information Science

FIFTH INTERNATIONAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

PROCEEDINGS—1995

June 7-10, 1995

Sponsored by

ROSARY COLLEGE

GRADUATE SCHOOL OF LIBRARY & INFORMATION SCIENCE

RIVER FOREST, ILLINOIS

USA

Edited by

Michael E.D. Koenig

Rosary College

&

Abraham Bookstein

University of Chicago

Learned Information, Inc.
Medford, NJ

Published for the International Society for Scientometrics and Informetrics by Learned Information, Inc.

Manufactured in the
United States of America

ISBN: 1-57387-010-2
Price: $79.00
Order from: Learned Information
143 Old Marlton Pike
Medford, NJ 08055
USA

Proceedings Manager: Carol Nixon
Cover Design: Jennifer Johansen

The papers published in this volume are in the format submitted by the authors. No editorial intervention was undertaken by the Editors unless absolutely necessary and only if the production schedule allowed. The Editors' task was to solicit contributions and to assist in other editorial matters.

MEETING ORGANIZER

Michael E.D. Koenig
Rosary College

PROGRAM CHAIR

Abraham Bookstein
University of Chicago

REGIONAL PROGRAM CHAIRS

Asia

Ravichandra Rao
Indian Statistical Institute
Bangalore, India

Europe

Leo Egghe
Limburgs Universitair Centrum
Belgium

North America

Abraham Bookstein
University of Chicago
USA

PROGRAM COMMITTEE

Asia

Ravichandra Rao, Indian Statistical Institute, *Chairman*	INDIA
Aparnu Basu, National Institute of Science, Technology, and Development Studies	INDIA
Mari Davis, University of Melbourne	AUSTRALIA
Hajime Eto, University of Tsukuba	JAPAN
B.K. Sen, Indian National Scientific Documentation Centre	INDIA
Yishan Wu, Information Analysis and Research Center	CHINA

Europe

Leo Egghe, Limburgs Univsitair Centrum, *Chairman*	BELGIUM
Manfred Bonitx, M.B.I.	GERMANY
Tibor Braun, Eötvös Lorand University	HUNGARY
Q.L. Burrell, University of Manchester	UK
Wolfgang Glänzel, Library of the Hungarian Academy of Sciences	HUNGARY
Peter Ingwersen, The Royal School of Librarianship	DENMARK
Sylvan Katz, University of Sussex-Brighton	UK
Alexev Korennoy, Ukrainian Academy of Sciences	UKRAINE
Jan Kozlowski, KBN State Committee for Scientific Research	POLAND
Hildrun Kretschmer, Association for Science Studies, e.V.	GERMANY
Thierry Lafouge, Ecole nationale superieure des l'information et des bibliotheques (ENSIB)	FRANCE
Cees le Pair, Stichting voor de Technische Wetenschappen Technology Foundation	NETHERLANDS
Valentina Markusova, VINITI	RUSSIA
Emilio Matricciani, Politenica di Milano	ITALY
Aida Mendez, Institut d'Estudies Avancats de le Illes	SPAIN
H. Moed, University of Leiden, CWTS	NETHERLANDS
Karl Muller, Institute for Advanced Studies	AUSTRIA
Bluma Peritz, The Hebrew University of Jerusalem	ISRAEL
L. Quonian, Centre de Recherche Retrospective de Marseille	FRANCE
A.F.J. Van Raan, University of Leiden, CWTS	NETHERLANDS
Ronald Rousseau, Katholieke Industriele Hogeschool	BELGIUM
William Turner, CERESI/CNRS	FRANCE
Peter Vinkler, Hungarian Academy of Sciences	HUNGARY

PROGRAM COMMITTEE

North America

Abraham Bookstein, University of Chicago, *Chairman*	USA
Terrence Brooks, University of Washington	USA
Susan Cozzens, Rensselaer Polytechnic Institute	USA
Blaise Cronin, Indiana University	USA
Belver Griffith, Drexel University	USA
Kate McCain, Drexel University	USA
Bill McGrath, SUNY - Buffalo	USA
Francis Narin, CHI Research, Inc.	USA
Mike Nelson, University of Western Ontario	CANADA
Miranda Pao, University of Michigan	USA
Jane Russell, Universidad Nacional Autónoma de México	MEXICO
Henry Small, Institute for Scientific Information	USA
Jean Tague, University of Western Ontario	CANADA
Radosvet Todorov, University of Maryland (Visiting)	USA
Howard White, Drexel University	USA

BIBLIOMETRIC TOOLS FOR BIBLIOGRAPHIC CODIFICATION DATABASES: TECHNOLOGICAL AND METHODOLOGICAL ASPECTS FOR RELATIONAL USE OF BIBLIOGRAPHIC DATABASES

P. Faucompre, P. Baldit, R. Dos Santos, L. Quoniam, and H. Dou
CRRM—Université d'Aix-Marseille, Marseille, France

Keywords: Relation table; Statistical clustering; Automatic classification; IPC catchwords.

Abstract: We will develop and discuss several techniques for reforming bibliographic classification databases systems, both the specific and generic search properties. We suggest ways to perform bibliometric analysis to provide links between them. Our goal is to enhance the exact one-one relationship between corresponding classification and to produce 1-N non obligatory relations that are in order to provide experts' discussions over classification's correspondence. The conclusion drawn from the analysis suggest the implementation of either a Winhelp (@Microsoft) hypertext file that includes both classifications' meanings and correspondences just for expert discussion about correspondence, or a relation table used for a relational construction of a highly heterogeneous internal database and supported clustering techniques. The aim of this last point is more to provide indicators for existing relations between databases than to induce an exact matching. Exact matching is not the most important factor if we consider the necessity, for the area expert, of a relational help in information analysis for technological transfer, decision processes and competitive intelligence. So this work establish a bridge between different bibliographic databases. An example shows probabilistic links between technical literature and technological patent literature

1. COLLECTION STRATEGY IN VIEW OF INFORMATION INCREASE

In view of the increase in information production and bibliographic database number and coverage, several collection strategies appear. Subject restriction that induces database number reduction and query implement, or a multiplicity of queries over a multiplicity of databases (Ref.[1]). If the first approach seems to be easier, it induces a significant increase of silence during collection, and we will not further discuss it. Various servers allow the second practice with cross-linked files search. The only valuable information to do this with is information that does not depend upon indexing policy (like patent number or cited patents). So, the drawback of the second way is the variety of indexing languages. Each of the databases uses a particular one. The difficulty is in understanding the various indexing languages used in various databases to build specific data collection queries. When a keyword's strategy is required, the problem of translating keywords is inevitable or collection is poor and with many distortions. In this case, it is more relevant to practice a specific strategy for each database.

2. INTERNAL HETEROGENEOUS DATABASE AND DECISION MAKING

In any data collection strategy used, the specific downloaded databases are to be incorporated in an internal database. To do that, many manipulations of the databases need to be done. Those manipulations are physical (subfield's correspondence) and intellectual (unify the indexing language), and are only possible with similar subject databases (for example several patent databases). If no standardization of the indexing language is practiced, the common structure is kept, which generally only includes the minimum description of the article. This leads many manufactures or users to build an internal thesaurus. This practice is very expensive and time consuming.

Unfortunately, information for strategic decision making very often requires a mufti-database investigation, moreover with highly heterogeneous databases. An example of this would be the link between a technological-technical patent database and thematic technical databases through thematic description. This kind of link is useful for building relations between industrial property and technical information. Another example is the link between bibliographic databases with corporate affiliation through thematic description of the articles and of the corporate activity. In that case, unique internal databases are difficult to build because of the differences in indexing (Ref.[2]) and even in the database's structure (for example corporate affiliations and scientific literature). Many works exist about building wide thesauri to build such unique database. Our process opens up a strictly relational way.

3. BIBLIOGRAPHIC CODIFICATION DATABASES

There are now some wild word databases which have existed for a long time. Some of them use classifications and/or thesauri which are of very good quality in their own area and are well implemented.

3.1 International Patent Classification (IPC)

For example International Patent Classification (Ref.[3]) has existed since 1968. This classification of technological patentable activities is interesting because it exists in two official languages (English and French) but is available on a CD-ROM in three other languages (German, Spanish, Hungarian). Legislation obliges 75 leading patent offices in the world, which practice indexing, to mention it in patents. This leads any patent database to include an IPC field. The Permanent Committee for Patent Information (PCPI) performs the revision of patent classification with a periodicity of five years. This revision period is sufficient to warranty an evolution according to technological development (Ref.[4]). Since 1989 the fifth version of IPC is available. Table one describes complete hierarchical classification numbers. At the most detailed level, more than sixty-four thousand classification codes describe all of the patentable technology.

Table 1. Hierarchical International Patent Classification

IPC	Classes	Sub-classes	Groups	Main-groups	Sub-groups	Σ
A	15	80	7 259	1 042	6 217	7 354
B	34	161	15 560	1 751	13 809	15 755
C	19	91	13 449	1 340	12 109	13 559
D	8	38	3 012	355	2 657	3 058
E	7	30	3 106	323	2 783	3 143
F	17	98	7 976	1 061	6 915	8 091
G	13	75	6 862	641	6 221	6 950
H	5	48	6 799	501	6 298	6 852
Σ	118	621	64 023	7 014	52 345	64 762

We have already used this classification to provide an interface between experts of a studied area (Ref.[5]) and to built up databases' queries (Ref.[6]). Due to difficulties in consulting such classification in a hierarchical form, World Intellectual Property Organisation (OMPI) supplies on the IPC:CLASS CD-ROM (Ref.[7]) a nearly keyword entry for classification. These keywords are extracted from the Official Catchwords Index and Figure 1 shows how they introduce direct access to classification codes.

Figure 1. A direct access to classification codes

Catchword entry
Séchage par **ABSORBANTs** F 26 B 5/16

IPC fulltext
F 26 SECHAGE
F 26 B **SECHAGE** DE MATERIAUX SOLIDES OU D'OBJETS PAR ELIMINATION DU LIQUIDE QUI
Y EST CONTENU ...
5/16 . par contact avec des corps **absorbants** ou adsorbants, p. ex. avec un moule absorbant; par mélange avec des matériaux absorbants ou adsorbants

These keywords simplify the classification process for examiners and additional keywords assigned to groups provide a useful (Ref.[8]) and powerful search tool for codes or subjects search. Figure 2 illustrates how six entries allow to *catch* a classification code (Ref. [9]).

IPC:CLASS CD-ROM contains the English, French and Spanish Official Catchwords Indexes and the bilingual German/English "Stich- und Schlagwörter Verzeichnis". Available since the 4th version of the IPC - 1984, they are not built by the same organism and so they provide different points of view (Ref.[10]) of the representation of the classification through catchwords. In this article, we use French catchwords (about 20,000 entries) because we are looking for a correspondence with another French database classification. First, we verify that catchwords are well connected with IPC:CLASS. The result of this study is abstracted in table two. We also built a hypertext that links (1-1relation) the catchword and the official definition of the IPC fulltext.

Figure 2. Directory of Alphabet Order Patent Descriptors and Index Terms

Stich- und Schlagwörter Verzeichnis - ENG IPC5 StichW (German Patent Office)

belt
  mattress belt A 47 C 31/08
hand-strap
  mattress hand-strap A 47 C 31/08
handle
  mattress handle A 47 C 31/08
mattress belt A 47 C 31/08
mattress hand-strap A 47 C 31/08
mattress handle A 47 C 31/08

ENG IPC5 fulltext A 47 C

31/00 Miscellaneous features in connection with chairs, beds, or the like, e.g. upholstery fasteners, mattress protectors, stretching devices for mattress nets
31/08 . Mattress hand-straps

Table 2. IPC:CLASS and Catchwords recovery

Sections	Classes	Sub-classes	Groups	Main-groups	Sub-groups
A	0.46	0.92	0.32	0.73	0.28
B	0.76	0.95	0.21	0.60	0.17
C	0.68	0.79	0.15	0.48	0.11
D	0.37	0.94	0.18	0.53	0.15
E	0.71	1.00	0.30	0.71	0.28
F	0.94	0.92	0.16	0.46	0.12
G	0.84	0.96	0.26	0.66	0.25
H	0.60	0.93	0.22	0.64	0.28

Then we treat those catchwords to get a database with only keywords and patent classifications. Figures 3a and 3b provide an example of catchwords before and after treatment.

Figure 3a. Example of catchwords before treatment

VEHICULE
  (1°) VEHICULEs en général
    Caractéristiques, aspects ou aménagements communs à plusieurs genres de VEHICULEs ou propres à certains VEHICULEs terrestres B 60
    (a) Détails de structure, aménagements divers, accessoires
      amortisseurs de chocs et ceinture de sécurité pour passagers (Dispositifs) B 60 R 21/00
      A 62 B 35/00
      anti-vol (Equipement) B 60 R 25/00

Figure 3b. Example of catchwords after treatment

AN CATCHWORD_V.5_19287
EP VEHICULES
ES
ET DETAILS STRUCTURES AMENAGEMENT DIVERS ACCESSOIRES
EQ AMORTISSEURS CHOCS CEINTURES SECURITES PASSAGERS DISPOSITIFS
IC B 60 R 21/00 ; A 62 B 35/00
SO VEHICULE // (1°) VEHICULEs en général // (a) Détails de structure, aménagements divers, accessoires // amortisseurs de chocs et ceinture de sécurité pour passagers (Dispositifs)
.... (EP=main entry, ES= secondary entry, ...)

3.2 CETIM Thesaurus

The French Centre d'Etude Technique des Industries Mécaniques (CETIM) produces a database which is available on the ESA server. For this database indexing, CETIM use an internal thesaurus ([11]). With this thesaurus they describe the CETIM subscriber's activity, as well as the CETIM activity or the abstracted articles in the database (Ref.[12]). There are three thousand two hundred entries in this thesaurus. This thesaurus describes technical activities, but the databases do not include patents. For strategic reasons, links with patents may be interesting. Direct links between those two indexing systems are not relevant because some individual words may match but not the multi-term entries of the thesaurus. Before a matching test, vocabulary homogeneity is performed with IPC:CLASS. Figure 4 shows some examples of single and multi-term keyword matches between two classification databases. In this first study, we only consider the terms and multi-terms as isolated graphical forms without their semantic meaning (Ref.[13]).

Figure 4. Multi-terms matching between CETIM Thesaurus and IPC:CLASS.

AN CETIM_5
EP ABRASIONS RESISTANCES
SA BT_ABRASION RT_ABRASION (ESSAIS)
RT_ALLIAGES RESISTANTS A L'USURE
RT_CARBURES (REVETEMENTS PAR)
RT_USURE (RESISTANCE A L')
SO ABRASION (RESISTANCE A L')

AN CATCHWORD_V.5_29
EP ABRASIONS
SA ABRASIF(IVE)
ES ACIERS ALLIES DOTES RESISTANCES SPECIALES
IC C 22 C 38/00
SO ABRASION // Acier allié doté d'une résistance à l'ABRASION spéciale

AN CATCHWORD_V.5_35
EP ABRASIONS
SA ABRASIF(IVE)
ES DETERMINATIONS RESISTANCES
IC G 01 N 3/56
SO ABRASION // Détermination de la résistance à l'ABRASION

4. MATRIX CONSTRUCTION AND CLUSTERING STRATEGY

Finding correspondence with a database system seems difficult. First, if a query system is needed, because of the number of queries one would need to produce. Relational database should be as difficult because of the nonunique correspondence between keywords. This is the reason we use a clustering technique that will allow nonexclusive links. We then automatically construct a binary matrix containing as columns all of the common vocabulary of the CETIM Thesaurus and IPC:CLASS. Both classifications' entries, which have common words, correspond to rows. This absence (0)/presence (1) of words within classifications must be clustered. The clustering technique here must classify both rows (classification entries) and columns (terms). This is why we use a block seriation technique. First developed in IBM CEMAP ([14]), the algorithm is now implemented in our laboratory ([15]). This clustering technique, unemployed in IR field which use numerous automatic classification, itself determines the number of clusters. The only parameter is a required density of clusters which induces a number of clusters. High density clusters provide

numerous small groups, and low density clusters induce fewer larger groups. Relations will be different too. With small groups, the relation between classifications within the group will be strong with many common words. Figure 5 provides an example of block seriation with such a matrix. It is possible to notice that the correspondence between classification entries is built with nonunique matching and with global optimization criteria.

5. HYPERTEXT BUILDING ACCORDING TO CLUSTER ANALYSIS

If the result of the cluster analysis is presented with @Microsoft Excel, the legibility of group constitution is poor. This is the reason we use a hypertext representation of the clustered classifications. Adding some specific fields, it is possible to give some navigational rules in a hypertext according to the statistical criteria of words' cooccurrences in various classification entries. We use a standard hypertext which allows a great portability of the product. The aim of this hypertext is now to provide some tracks of correspondence to the experts of the studied area. Such matching techniques cannot provide infallible links but are relevant to submit a variety of high probability links within several thousands of possible links. The @Microsoft Winhelp provides some powerful tools for expert validation (cut and paste with external applications, printing, history, personal notes and bookmarkers). Figure 6 provides a view of a hypertext entry.

Figure 5. Clustered binary matrix with block seriation technique

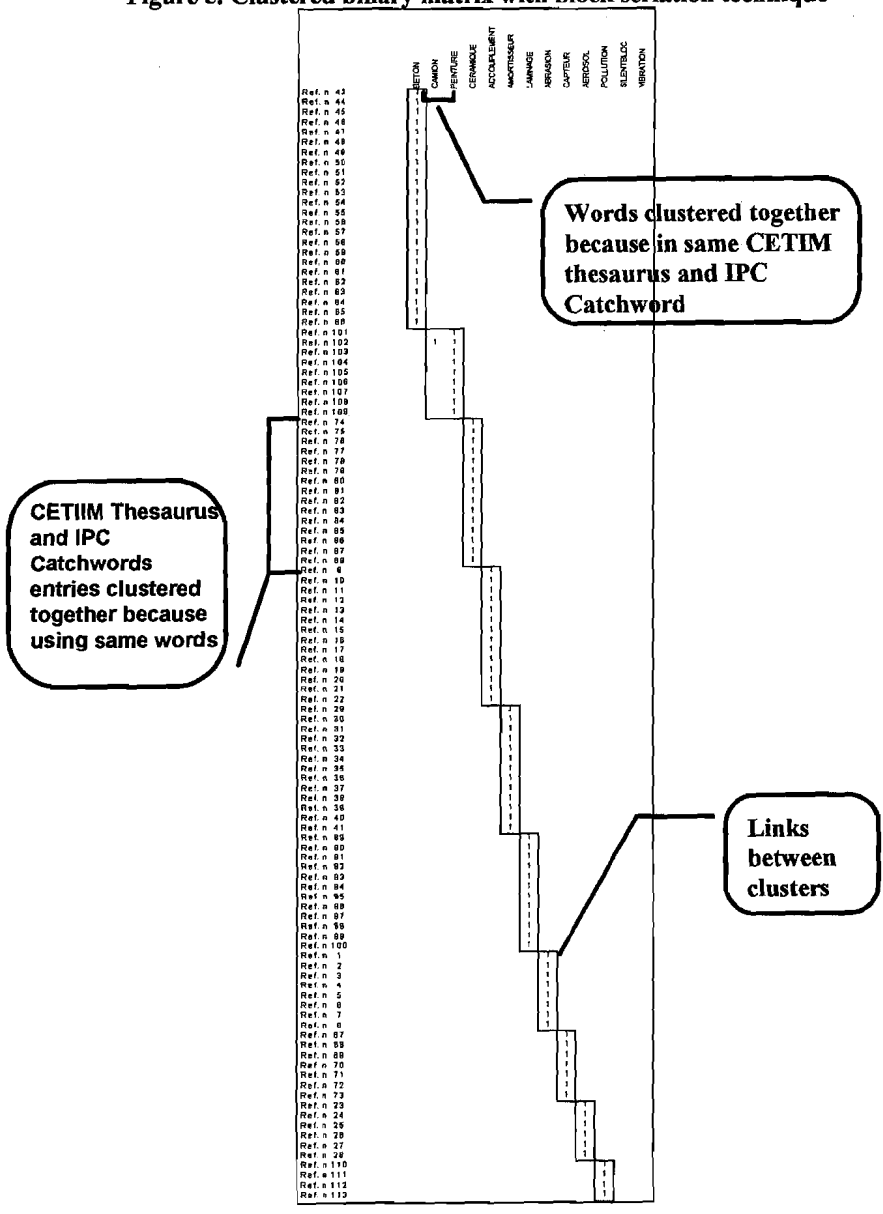
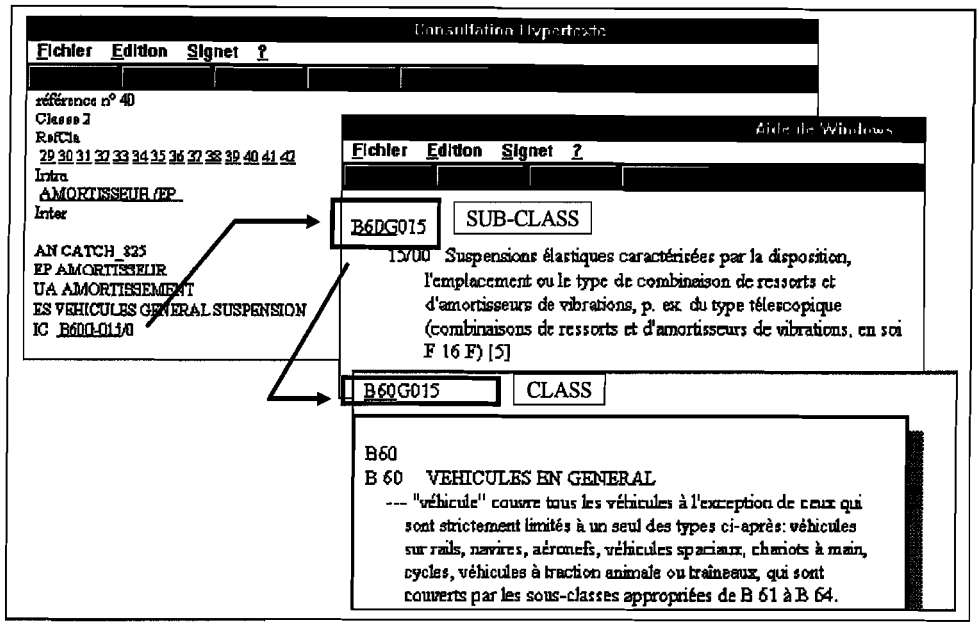


Figure 6. Hypertext presentation



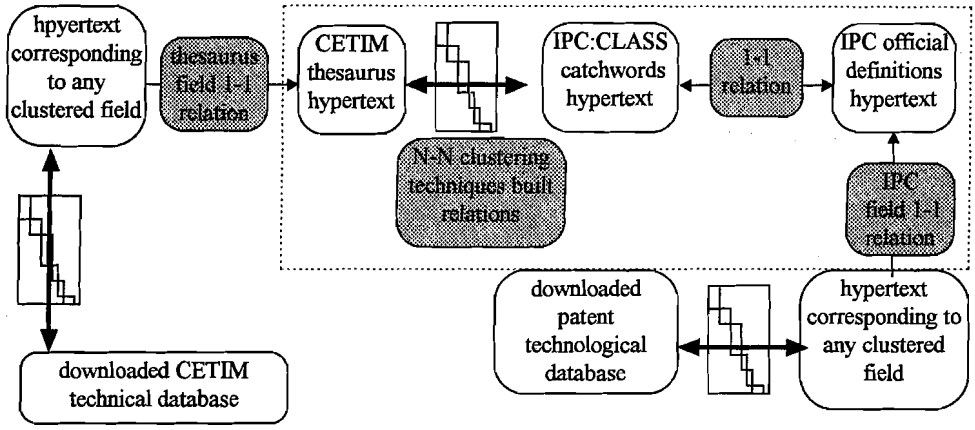
6. LINKS BETWEEN DOWNLOADED DATA

This hypertext system may be the center of a relational system between databases. It is possible to value each statistical link and after this stage build a correspondence table which will allow an automatic reindexing. It is possible also to keep this system of hypertext like a margin between two databases. For instance when working with a CETIM downloaded database, it is possible, after some treatment (i.e., block seriation of the authors) to link (in this stage 1-1 relation) with the previous hypertext. The end user of the system may have a look over two mixed hypertext system. One with the clusters of authors, the other with connection to IPC:CLASS to validate some links with technological patentable activities. Figure 7 represents the complete scheme of the treatment, dotted line limits our field application in this present work;

7. FUTURE DEVELOPMENTS

There are other big classifications' databases available which could provide interesting links for decision making. For instance, we can use all classifications about corporate activity like CITI for UNO, NACE for EEC, NAF for France which depend upon one another. Those classifications linked with IPC:CLASS could provide a link with patents even for companies that have no patent activity. The same relation could be provided through the corporate products classifications like CPC from UNO, CPA from EEC, CPF from France. Other classification databases are available in large bibliographic databases such as INSPEC, CAS, PASCAL. Similar analysis performed with such databases could allow links between fundamental works and applied technological works. This type of link is useful as well for corporations that need some scientific help as well as for scientists who look for corporate applications. The main work in such a project is to build the central relations between different classifications. A connection with bibliometric analysis of downloaded peripheral databases is easier to build.

Figure 7. Complete treatment



CONCLUSION

Access to and downloading of bibliographic databases are getting easier every day. Bibliometric mapping of sets of answers is now a classic technique which is used in corporations. There are new possible applications for bibliometric analysis. The one we present here can be a new development. Bibliographic databases are now extensive to allow some interfacing techniques between these databases. This application will also be useful for other fields of bibliometric analysis: Competitive Intelligence and decision making.

BIBLIOGRAPHY

1 CHADWICK T. B. "Searching across files and systems: multiple search techniques." *Online/CD-ROM'91 conference*, Chicago, 11-13 Nov. 1991.

2 YANOSKO CHAMIS, A. " Selection of online databases using switching vocabularies." *J. of the American Soc. for Inf. Science*, Vol. 39, No 3, 217-218, 1988.

3 OMPI. "Informations générales sur la cinquième édition de la CIB." in : *Manuel sur l'information et la documentation en matière de propriété industrielle*, 1990.

4 VRIES, S. de. "Points of interest concernig the new IPC[5] ." *World patent information*, Vol. 11, No 3, 115-120, 1989.

5 DOS SANTOS R., BARETTA A., DOU H. "L'information sur les surfactants biodégradables et non toxiques." *Journées Chevreul et Colloque GERLI.* 19-20 Nov., Marseille, 1994.

6 MACEDO DOS SANTOS R., RODRIGUES GREGOLIN J., VARGAS L., QUONIAM L. "IC&T: stratègia de exploração da informação para a tomada de decisão." *18ième simposio de gestão da inovação Tecnológica,* São Paulo 24-26 Octobre, 1994.

7 HANSSON,B. "The IPC:CLASS CD-ROM : a new search system in the patent information field." *World patent information,* , Vol. 14, No 4, 227-230,1992.

8 VIJVERS, W. G. "The International Patent Classification as a search tool." *World patent information*, Vol. 12, No 1, p. 36-30,1990.

9 ZAKARIA S. "Automated patent classification for German patent documents" Saarbrücken : Universität des Saarlandes, dissertation, 1989.

10 WORLD INTELLECTUAL PROPERTY ORGANIZATION. "IPC:CLASS : International Patent Classification, Classification and lingusitic advanced support system.".

11 CENTRE TECHNIQUE DES INDUSTRIES MECANIQUES "Thésaurus de la mécanique - 3rd ed." Senlis (FR) : CETIM, 1993.

12 DUMAS S. "Développement d'un système de veille stratégique dans un centre technique." Marseille : Univ. Aix-Marseille 3, France :Thèse de doctorat, 335 p., 1994.

13 LEBART L., SALEM A. "Analyse statistique des données textuelles." Paris : Dunod, 1988

14 MARCOTORCHINO F. "La classification automatique aujourd'hui." *Publication scientifique et technique d'IBM France*. No 2, 35-95, 1991.

15 QUONIAM L., HASSANALY P., BALDIT P., ROSTAING H., DOU H. "Bibliometric analysis of patent documents for R&D management." *Research Evaluation*, Vol. 3, No 1, 13-18, April, 1993.