DATACIB : A new automatic tool to link scientific bibliographic references and technical information

Pascal Faucompré, Luc Quoniam, Hervé Rostaing, Henri Dou

Proceedings: 7th International Conference on Scientometrics and Informetrics, Colima, Mexico, July 5-9, 577 (1999)

DATACIB: A NEW AUTOMATIC TOOL TO LINK SCIENTIFIC BIBLIOGRAPHIC REFERENCES AND TECHNICAL INFORMATION

FAUCOMPRÉ P., QUONIAM L., ROSTAING H., DOU H.
C.R.R.M.- UNIVERSITE AIX-MARSEILLE III
FR 13397 MARSEILLE CEDEX 20 - e-mail: crrm@crrm.univ-mrs.fr

ABSTRACT

The aim of this paper is to present a software that will automatically introduce a new bibliographic field in scientific bibliographic references. The used algorithm is, in fact, largely linked to the bibliometrics field, using distributional properties of the suggested links between scientific keywords and International Patent Classification Catchwords. It is a contribution in an attempt to globalize links between different information fields and built generalized relational data on a large scale.

1. DECISION MAKING IS A RELATIONAL PROCESS

The competitiveness of manufactures can be involve with a large number of actions. Therefore, as shown in figure 1, the actions involving information are in a good position in terms of accessibility and cost. The innovative process in R&D and patenting is due to interconnections between facts more than to the discovering novelties act. This comes from the increasing number of new information available and to the increasing complexity of the

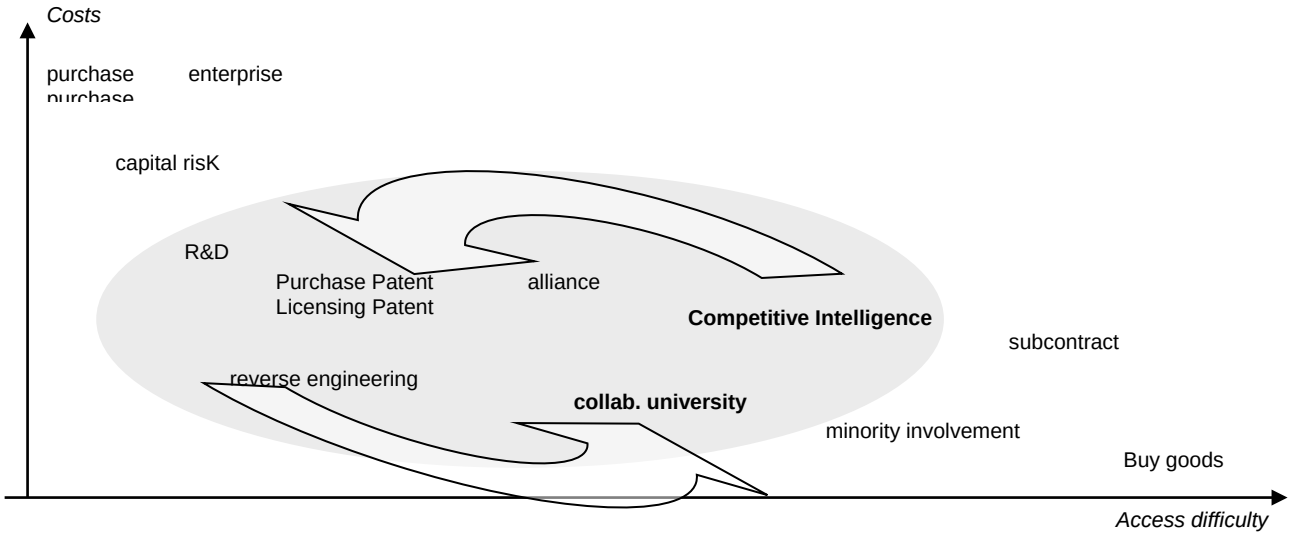

Figure 1 Cost & accessibility of the used activities to increase the enterprises competitiveness

industrial decisional process. Industrial decisions depend as well on environmental, economic, politic, scientific facts and data. Therefore, it is obvious that, even if the decision is the act of a small number of persons, this decision is built over the work of a human network that built a network of knowledge. First, this is due to the brain limitation of each one of us in acquiring such a large number of informations. This is also due in a bad relational organization of the data. Our aim is to tent to built an help to link just two words, scientific data and patent data.

2. NOWADAYS DATA ARE HIGHLY RELATIONAL

Small ([1]), in a previous paper explained truthfully that bibliographic data are highly relational. It is true in a same bibliographic database. It is also true through several databases (i.e.; several bibliographic or patent databases). It is also true from scientific articles to patents (at minimum through the used words). However, it is obvious that those relational aspects are difficult to formalize due to the presentation of the data is completely different. For example, patents use International Patent Classification (I.P.C.) ([2]) to describe the content of the patent. Therefore, the relation to bibliographic data is not a direct link. The way from bibliographic keywords to Patent Classification pass through the catchwords ([3]) (keywords that describe the I.P.C.) and must is intellectually built by the experts who needs a link between scientific and industrial research. This link has a very large impact in economic terms as shown in Figure 2 from a competitive intelligence executive ([4]). There are 72000 I.P.C. Codes that must describe all the patentable technologies due to an international law that oblige each patent in the world to be described by an I.P.C. code. This I.P.C. is represented by hierarchically organized codes. More detailed is the signification of the code, longer is the code.

Our purpose here is to offer this community a new tool that will help their daily work.

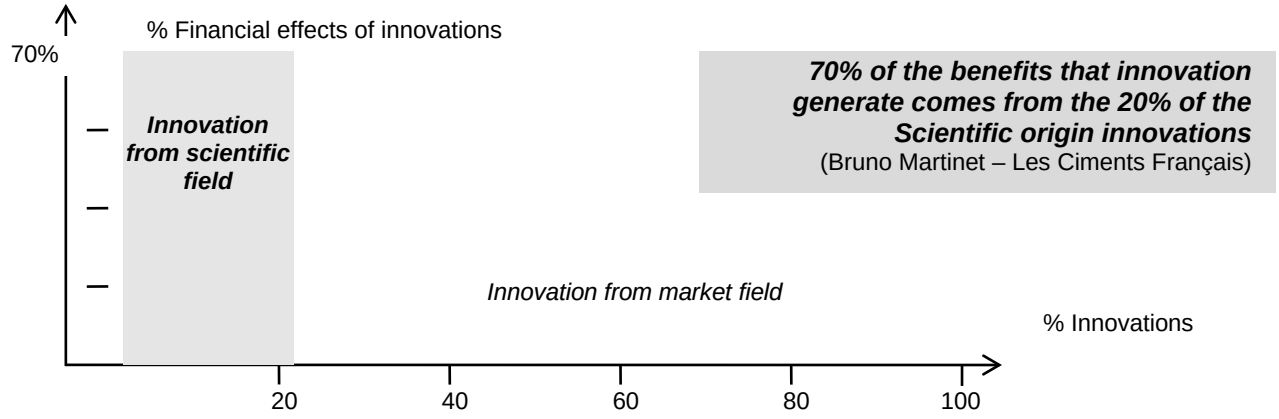

Figure 2 Economic impact of scientific novelties in enterprises

3. HOW IT WORKS?

In previous papers [5,6], we had shown first results of our project to create new relation between scientific information and technical information. In this communication, we present last developments of our study, the DataCIB software. The goal of DataCIB, a program very easy to use, is to generate a new documentary field into scientific bibliographic references, which, generally, never include technical keywords. This new field contains IPC (International Patent Classification) symbols, which have been linked with controlled keywords of these references. These new links are set up with the help of all the official catchwords of the patent classification. Nowadays, DataCIB uses exclusively French and English controlled keywords of the French PASCAL ([7]) international bibliographic database (produced by INIST) ([8]) but it will be soon possible to work with many others vocabularies of indexing. The only basic essential condition is that these controlled terms must to provide a certain formal (verbal) compatibility with the IPC catchword index. In such a condition, links with any other download of any scientific database is possible. Our choice in the Pascal database was conditioned by a double reason. First this database covers any scientific fields and so is a very large test database. Second, the I.P.C. Catchwords are available in both French and English, so a database offering both indexation will lead a stronger link, and this particularity is specific to PASCAL, as far as we know. Out of this limitation, any scientific database can be treated, creating this new field of I.P.C. codes. Virtually is has the same sense as building a relational database system including the several millions of scientific documents with the several millions of patents documents. In that sense, this correspondence is on a very large scale.

4. USED ALGORITHMS

Previously we tested various algorithms to build this automatic link. It seems that the most efficient is the simplest algorithm that consider the distribution of the linked I.P.C.. The actual algorithm tent to link a scientific reference to all the matched I.P.C. codes using the keywords from this reference. What occurs is that for each reference there is a pool of I.P.C. codes that are possible. Some of those codes are redundant e.g. there are codes that are linked to the reference with a higher probability. A selection process then occurs which is done over this frequency distribution of the linked codes. Therefore, this process is not simple due to the hierarchical presentation of I.P.C. codes. We must obtain the most detailed codes (longer ones) that occurs more for each reference

Obtained results can be analyzed following two different ways. In the relation science-technology, we can observe codes contained in a particular bibliographic reference or, through a more bibliometric and/or scientometric approach, study the whole of codes which

are associated with this downloaded corpus [9]. We can also get patents, which are classified with the same symbols, i.e. published in the same technical domains. In the opposite sense, in this case the relation technology-science, the initial step can be a particular IPC symbol or all extracted codes from a set of patent documents and then, the second step, the search of fundamental references linked with these selected codes.

5. AUTOMATIC ANALYSIS WITH A CONTROL FROM THE EXPERT

Main steps of the reindexing process of bibliographic references are illustrated by the four following pictures. The Figure 3 shows the global result of reindexing of 40 references, which present publications of laboratories dealing with the biological and industrial pollution problem. By the reindexing, 77,5% of references contain now a supplementary field, which

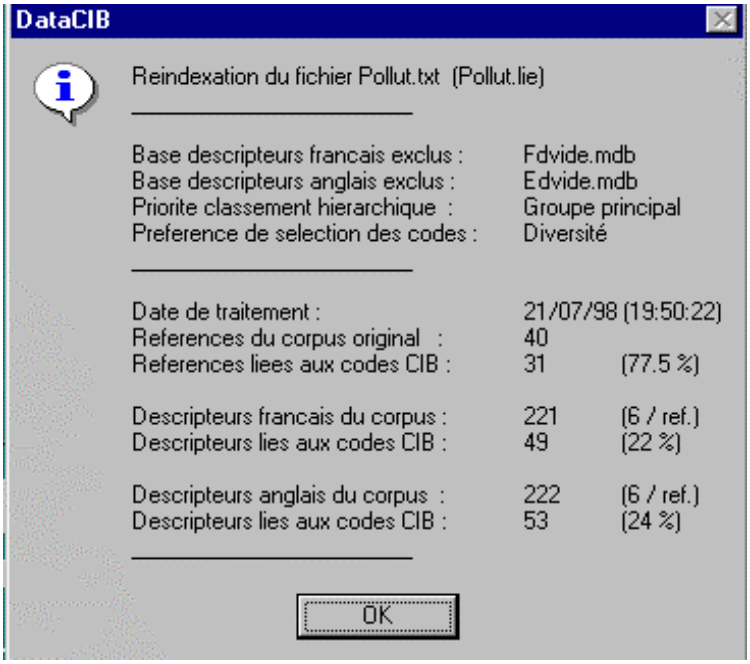

Figure 3 Reindexing scientific reference with IPC codes

includes patent classifications codes. "No linked references" (22,5%) have not obvious industrial interest or they have a too fundamental problematic: their keywords can not be matched with any terms or syntagms of the correspondence table. We must not forget not that all the scientific field may be patentable. For each reindexed reference, the definition of associated classification symbols is given.

Picture 4 presents the global distribution of new inserted codes into references, their associated keywords with codes, and their links together. The idea of the software is to provide a tool for the industrial experts, but not to substitute their activity by an automatic treatment, so this step will be included in a validation step of the accuracy of the link. In this sample, main problematic concern damages from water pollution, heavy metals, and their

consequences for the human health. Any one could understand that only the highest frequencies have a good probability to represent the technological aspects of the scientific article. The first idea is to keep only those high frequency codes, but considering a truncation at a higher level in the hierarchy could inverse the distribution of codes. Another reason to provide an analysis off the hierarchical structure of the linked codes is to help the customer

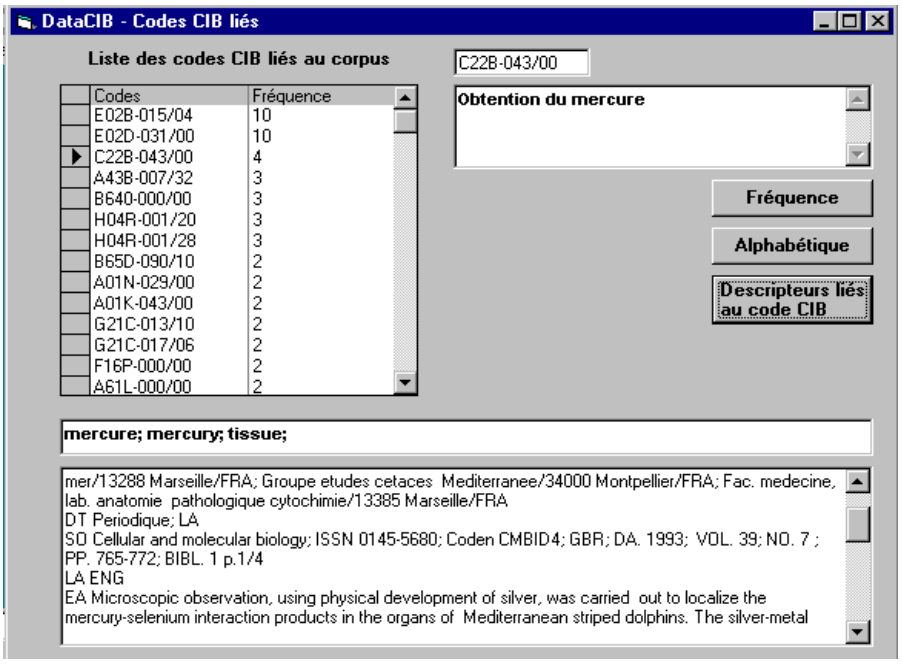

Figure 4 Linked classification codes and associated keywords

for further search of patents documents through databases (WPI, EPAT, ...). The Figure 5 shows how the software let a consultation of the four level hierarchical levels corresponding to a main group. The expert validation is possible too at this step.

Then the reindexation process is performed. The output is available as a text file that can be reused as any common downloaded corpuses. In particular, they show how the simple projection of patent codes about scientific bibliographic references can bring a new industrial representation of more fundamental works. So, they provide a potential interesting articulation between science and technology and are tools for a bijective link from the scientific world to the industrial world using a common language. An example of a reference including the new field is given in Figure 6.

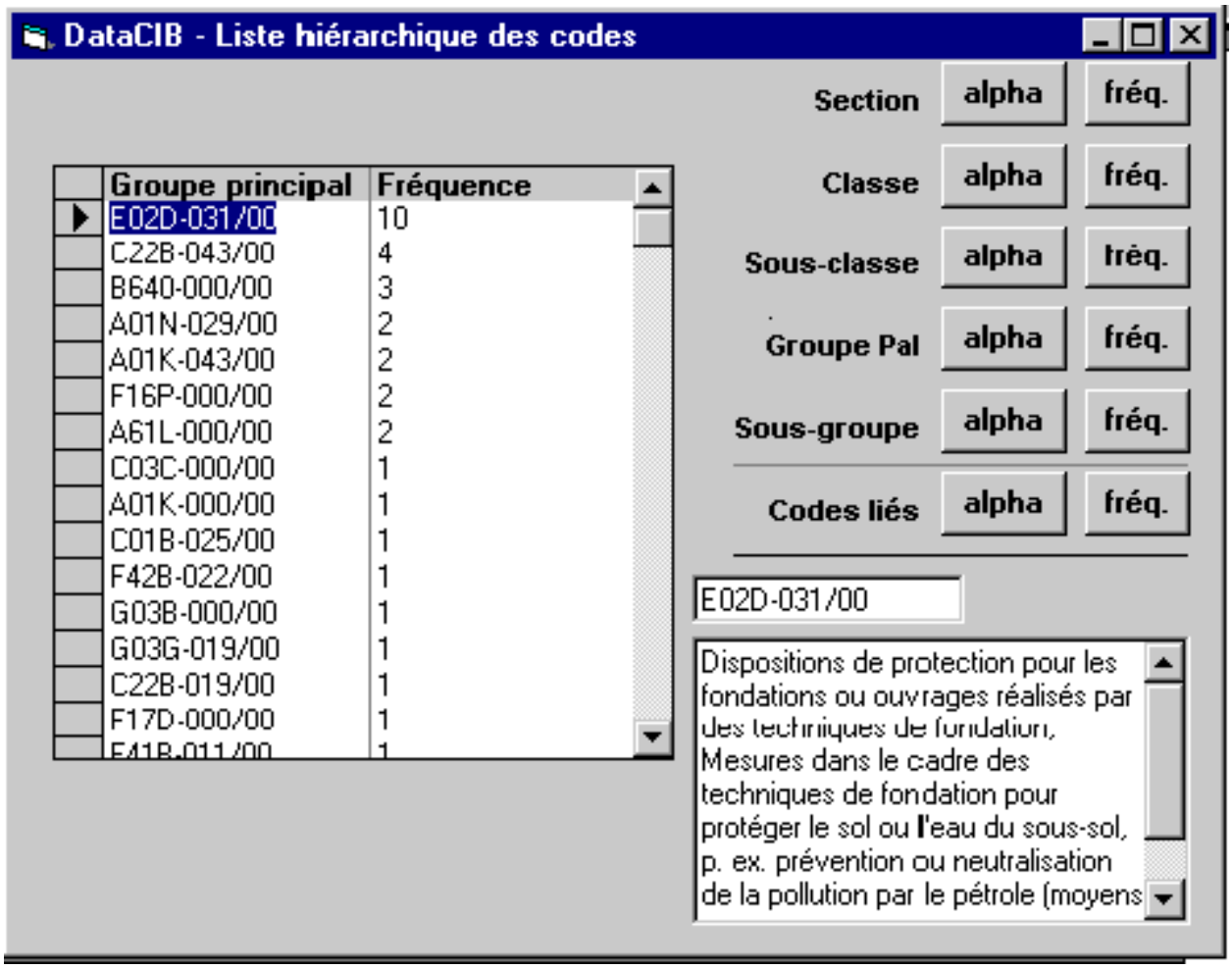

Figure 5 Main groups hierarchical level

AN PASCAL 93-0494817 INIST
ET **Hydrocarbon biodegradation and hydrocarbonoclastic bacterial communities composition grown in seawater as a function of sodium chloride concentration**
AU BERTRAND J C; BIANCHI M; AL MALLAH M; ACQUAVIVA M; MILLE G
AF Cent. oceanologie Marseille/13288 Marseille/FRA; CNRS Fac. sci. Luminy, lab. microbiologie marine/Marseille/FRA; CNRS Fac. sci. tech. Saint Jerome, lab. chimie analytique environnement/Marseille/FRA
DT Periodique; LA
SO Journal of experimental marine biology and ecology; ISSN 0022-0981; Coden JEMBAM; NLD; DA. 1993; VOL. 168; NO. 1 ; PP. 125-138; BIBL. 22 ref.
LA ENG
CC 002A14C02
ED Environmental factor; Salinity; Pollutant; Hydrocarbon; Bacteria; Biodegradation; NaCl structure; Concentration factor; France
IC C12N-001/20

Figure 6: Exemple of a reference with the extra IC Field

6. CONCLUSIONS

Obviously, the new established links can not be considered as a strong and reliable relation. This tool, here presented is in its first version, it contains today defects and some developments will be necessary to improve the basic algorithms. We also know, on a strict documentary plane, that our method only based on a statistical approach generates inevitable noise. However, and it is the most important fact, these relations provide presumption of links which are sufficient to create the necessary condition of emergence of something of new and, therefor, to bring new relevant matter to the technological watch or to the competitive intelligence. A more complete description of this methodology and this automatic tool will be given in our final communication.

7. BIBLIOGRAPHY

[1] Small H. relational bibliometrics *5th Int. Conf. on Scientometrics and Informetrics, Chicago, June 7-10, 1995* p.525-532

[2] World International Property Organisation *Classifications Internationales - Locarno, Nice* 1998 [Online] URL address http://www.wipo.org/fre/main.htm

[3] World International Property Organisation *CIB 6 — Index Officiel des mots clés* 1998 [Online] URL address http://www.wipo.int/eng/clssfctn/ipc/ipc6fr/nfcatch/index.htm

[4] MARTINET B., MARTI Y.-M. *L'intelligence économique : les yeux et les oreilles des entreprises*. Paris : Ed. d'Organisation, 1996

[5] Faucompré P., Baldit P., Dos Santos R., Quoniam L., Dou H. Bibliometric tools for bibliographic codification databases : technological and methodological aspects for relational use of bibliographic databases. *5th Int. Conf. on Scientometrics and Informetrics, Chicago, June 7-10, 1995*

[6] Faucompré P., Quoniam L., Dou H. An effective link between science and technology. *6th Int. Conf. on Scientometrics and Informetrics, Jerusalem, June 16-19, 1997*

[7] Spécial IST : base de données PASCAL. *La lettre des utilisateurs des produits et services de l'INIST*, 1995, oct., 3 p.

[8] L'INIST *INIST-CNRS : Accueil*. 1998 [Online] URL address http://www.inist.fr/

[9] Faucompré P., Quoniam L., Dou H. The function-application relation through a double link between indexing-classifying and science-technology. *World patent information, 1997, vol. 19, n° 3, p. 167-174*