The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

SIGAPS : a prototype of bibliographic tool for medical research evaluation

P Devos [a] , E Dufresne[b], JM Renard[b], R Beuscart[b]

[a] *Department of Biostatistics, University Hospital, 1 Av O. Lambret, 59800 Lille*
[b] *Department of Medical Informatics, University Hospital, 1 Av O. Lambret, 59800 Lille*

Abstract

Evaluation of research activity is extremely important but remains a complex domain. There's no standardized methods and evaluation is often based on the scientific publications. It is easy to identify, for a researcher, all the publications realized over a given period of time. At the level of an important establishment like an University Hospital, with about 500 researchers, this sort of inventory is very difficult to realize : we have to list the researchers, to list their publications, to determine the quality of articles produced, to store retrieved data and to calculate summary statistics. We have developed a full-Web prototype, using free software which, for a given researchers' list, interrogates the Pubmed server, downloads the found references and stores them in a local database. They are then enriched with local data which allow the realization of more or less complex analyses, the automatic production of reports, or keyword search. This tool is very easy to use, allowing for immediate analysis of publications of a researcher or a research team. This tool will allow to identify those active teams to be maintained or emergent teams to be supported. It will also allow to compare candidate profiles for appointments to research posts.

Keywords:
research evaluation, bibliographic indicators, reports.

1. Problem

Evaluation of research activity is extremely important, since it constitutes a determinant factor for budget allocation, for identifying new areas for research, and for the assessment of research performance. However, it is a complex domain, and there are no methods or standardized tools allowing for perfect evaluation. Scientific publication is one of the elements for quantifying research activities, and is generally used as an indicator of activity in numerous evaluation studies [1,2]. On a practical note, it is necessary to be able to identify and sort, for an individual, a team, or an institution, all the publications realized over a given period of time. While this would be a straightforward exercise on an individual level, it is less so where a research team or an establishment is concerned. At the level of an important establishment like the Regional University Teaching Hospital, this sort of inventory can pose big problems. This is because, each year, more than 500 scientific publications are realized in this establishment, by about 500 researchers, working in about 70 clinical services and belonging to around thirty research teams.

A first approach consists in interrogating the 500 researchers and in asking them for their list of publications. This approach is not reliable for different reasons: poor response rates, doubloon problems, and cost of this type of inquiry. This approach has once been experimented in our establishment and lasted about four months. The second approach is to use bibliographical interrogation software like Reference Manager© or EndNote©, edited by the Institute for Scientific Information **(ISI)** [3]. These tools are well adapted for managing

the publications of one or several researchers, but cannot be automated to allow for working on a much larger scale like that of about 500 researchers.

Worldwide, there are data banks which reference a significant proportion of scientific production. In the medical domain, the most important databases are Medline and Current Contents. Medline, via its Web server Pubmed [4], is certainly the most notorious and has already been used in studies on bibliographical evaluation [5,6]. It is possible, via Pubmed, to obtain the bibliography of a single individual. To carry out the same research at team level, it is necessary to interrogate Pubmed for each author and then summarize the findings. Furthermore, for more detailed analysis, it becomes necessary to type the information again.

To avoid this, we propose an architecture for a semi automated treatment of bibliographical data.

2. Method

We identified the functions required for an application intended to synthesize the scientific production of a given entity. This type of application should be capable of 1) to list the researchers of this entity, 2) to list the publications of these authors 3) to determine the quality of articles produced 4) to store retrieved data 5) to calculate summary statistics.

In responding to these various objectives, the following methodology was adopted :

1) Implementation of a research directory containing information on the researchers (service, research team, thematic, ...).

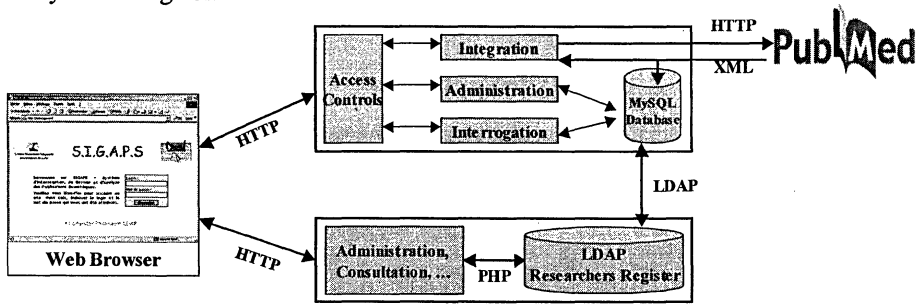
2) Creation of a research module for interrogating the Pubmed server, retrieving the references, decoding them and storing them locally.

3) Characterization of articles based on an indicator derived from the Impact Factor (**IF**) [7].

4) Implementation of a data base containing the data extracted from Pubmed, information concerning the authors, and scientific evaluation indicators.

5) Development of an interrogation module for obtaining standard statistics and graphs concerning an author or a team.

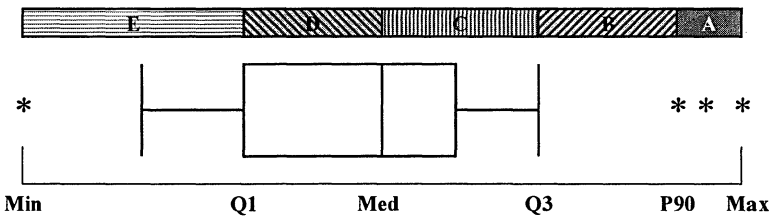
Next, we have developed SIGAPS (**S**ystème d'**I**nterrogation, de **G**estion et d'**A**nalyse des **P**ublications **S**cientifiques), a full-Web application. This application, based on a LDAP directory [8], a data base MySQL [9], an Apache Web server and a set of modules written in PHP [10] allows for automatic integration of data, consultation or creation of data boards from any web navigator.



Graph 1

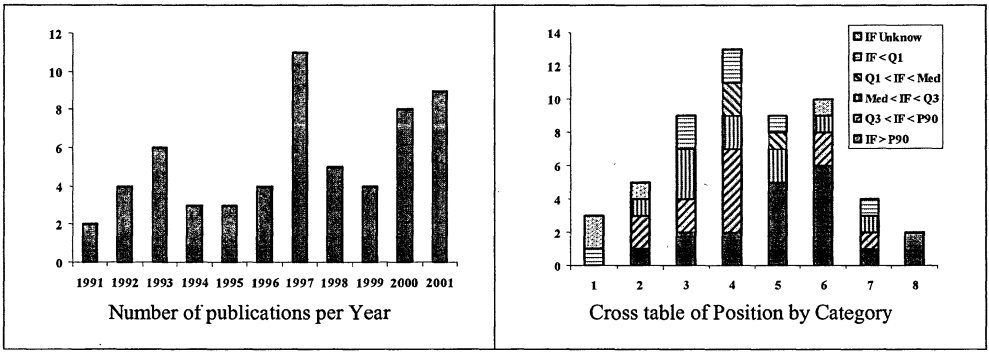
Authors' characteristics (postings, research themes, ...) are contained in a LDAP directory (OpenLDAP). The software allows the creation of a specific LDAP directory but can also operate using an existing LDAP directory. Publication integration is done using a queries file : it is enough to create a text file containing the Pubmed requests (like Smith J [Author]) which one wishes to make. If, for an author, the query is either too complex, or even impossible (numerous homonyms), the PMID which is a unique Pubmed identification number, can be used to retrieve publications. The software then interrogates the Pubmed server and gets back the corresponding publications. The software verifies the existence of each publication in the base, decodes the XML and loads the data into the base. It also interrogates the LDAP directory to find information concerning the authors.

To evaluate the "quality" of a publication, we propose a quality indicator based on the IF, which is the notoriety index of a review calculated by the ISI. This indicator, much in use but also very controversial [11] varies from one discipline to the other. For example, if one considers the IF for the year 2001, it is observed that for the discipline "Medical Informatics", the IF varies from 0.431 to 2.091 with a median equal to 0.808. For the discipline, "Cell Biology" it varies from 0.25 to 29.2 with a median at 2.188. We did a statistical classification of reviews, by discipline, into 5 categories (A to E) corresponding to quartiles and percentiles of the distribution (Graph 2). If an article is published into a journal not registered by the ISI (proceedings of congress, for example), then this article is affected to a 6th category (NC : Not Classified)

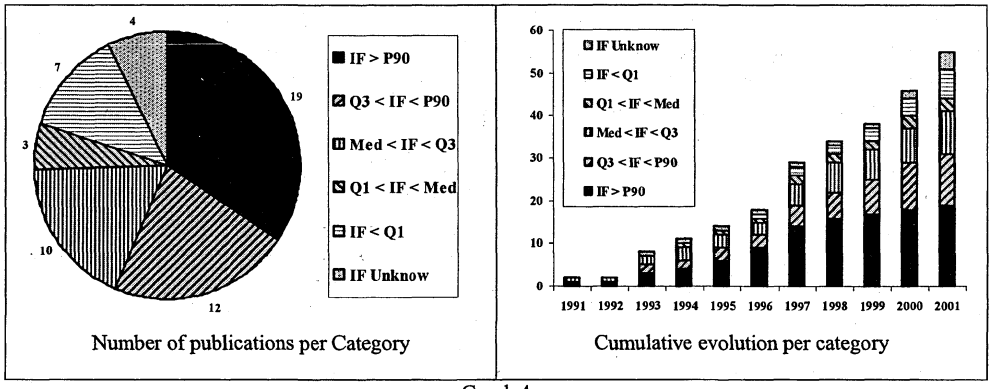


Graph 2

Via any browser, a summary table can be generated for a given author, one or several services, one or several research teams. The software then generates a report in HTML containing the number of publications a year, the position by category, the evolution by year and by category, and the main collaborations (co-authors). This information is also available in graphic form (Graph 3,4).



Graph 3



Graph 4

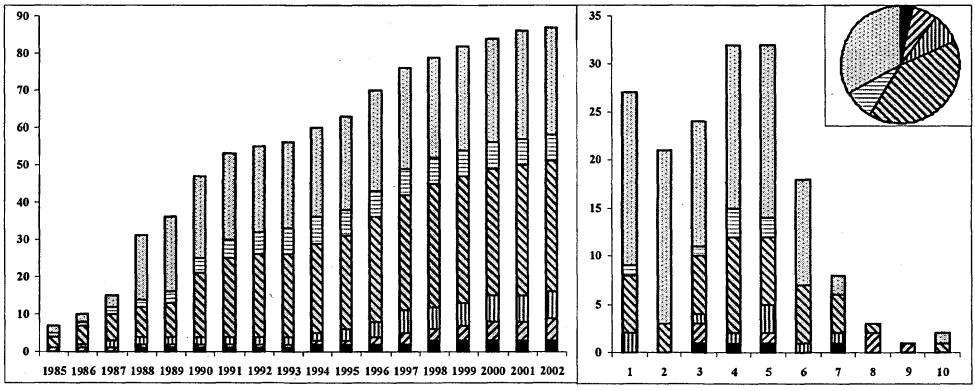
The application manages two profile types: an Administrator profile, for accessing all the functions and the User profile, for interrogation only. A log system allows for tracing of loading operations or data consultation.

Finally, by using an ODBC connection at Intranet level, it is possible to connect on the base and use the data in a standard office environment for statistical analyses, scheduling requests or generating bibliography automatically.

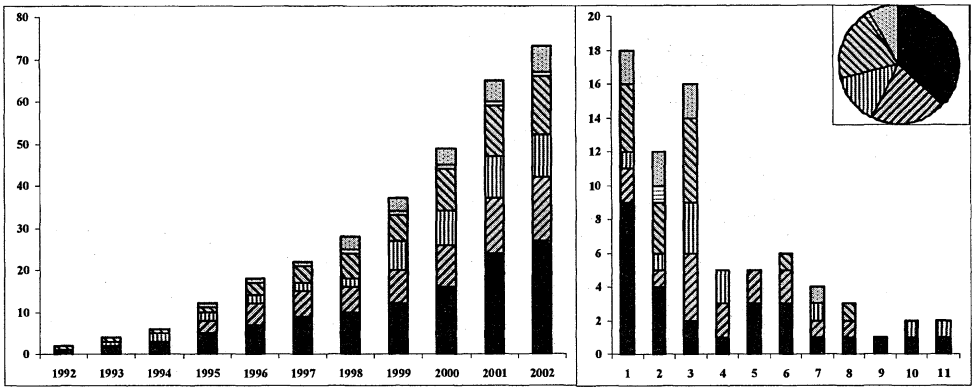
3. Results

Preliminary queries and studies show some very interesting results. First, we can see a high number of articles classified in category A, corresponding to articles published in famous journals.

Secondly, we observe a lot of different profiles of publication. We present here two profiles of 2 real researchers. The researcher A (Graph5) has a lot of publications referenced in the Pubmed database, but mainly in proceedings of congress (NC category) or in journals with low impact factors (categories D or E), in the forth or fifth position in the list of authors.



Graph 5



Graph 6

The researcher B (Graph6) has also a lot of publications referenced in the Pubmed database. However, one third of its publications is classified into category A, another third in B or C. The proportion of NC is very low. Finally, we can observe this researcher is essentially cited among the first three authors.

4. Discussion - Conclusion

We have developed a prototype completely based on Internet technology, and using free software which, for a given researchers' list, interrogates the Pubmed server, downloads the found references and stores them in a local data base. They are then enriched with local data which allow the realization of more or less complex analyses, the automatic production of reports, or keyword search. We chose the Pubmed server because it is the reference in the medical domain and it allows the export of data in XML format, automatically interpretable by a computer program. There are however numerous other bibliographical servers on Internet. Regrettably, a large number of them do not allow XML export or are only accessible via the Z39.50 protocol. This protocol, dedicated to bibliographical research, is less adapted to automatic data processing than the XML export. There are however initiatives being taken to make these two solutions converge. It would be interesting afterwards to widen the search to the other servers, notably as regards disciplines outside the medical domain, as biostatistics, medical informatics, or biophysics for example. For these disciplines, an important part of their works are published in more fundamental reviews which are not systematically referenced in Medline. The major problem remains the decoding of the information retrieved, generally supplied in free text format.

The implementation of a researchers' directory allows for a representation of research by services and by team, as well as by research theme. Synchronizing the base with the directory allows to have transversal statistics by team or by theme. However, the major problem of directories is the absence of historical perspective. The directory only contains information on current appointment and affects all the publications of an author to his current service. We are presently working about a mechanism which would make it possible to affect a publication according to the successive appointments of an author.

Another problem concerns the misspellings or identical patronymic names. In fact, an author can have several publication names (concerning women, for example, the marital name or maiden name) and several persons can have the same name and surname's initial. In this

cases, a manual validation by the different authors is essential in order to attribute to each person its own articles. The first problem is easily solvable by the directory, not the second one.

Review categorization into 5 classes is presently a statistical classification based on the Impact Factor. The proposed categorization, although partially correct, can be improved. The major problem of the Impact Factor results from the fact that it is based on the number of quotations. Therefore, there are some evident flaws, for example, the ascendancy of the American reviews, or the high value of the Impact Factor of reviews which have a widely spread electronic diffusion. We suggest crossing this classification with a classification realized by an experts' panel on every discipline, to correct these flaws.

From a strategic view point, we now have a very easy to use tool, allowing for immediate analysis of the whole publications of a researcher or a research team. This tool can be useful for different levels: at the research team level, it will allow automatic generation of bibliography, and will allow an auto-evaluation. At the institutional level, it will allow to identify those active teams to be maintained or emergent teams to be supported. It will also allow to compare candidate profiles for appointments to research posts.

The distribution of this software to other hospitals will allow them to have a common tool for comparison. For every establishment, it will be possible to analyze research activity and to determine the disciplines having consequential research activities.

We are presently also working on the implementation of a framework for the constitution of a central database and a national directory. The analysis of this national database will allow to establish a synthetic representation of hospital research teams. It will also contribute to the reflection on the implementation of networks for federating research teams. Finally, it will facilitate information availability and promote faster and more efficient reactivity to national or European, public or private invitations to tender.

References

[1] E.C.M Noyons, M. Luwel, H.F Moed, Assessment of Flemish R&D in the field of information technology. A bibliometric evaluation based on publication and patent data, combined with OECD research input statistics, *Research Policy*, 27 (1998) 285-300.

[2] W. Glänzel, U. Schoeplin, A bibliometric study of reference literature in the sciences and social sciences, *Information Processing and Management*, 35 (1999) 31-44.

[3] ISI Homepage - http://www.isinet.com.

[4] Entrez Pubmed - http://www.ncbi.nlm.nih.gov/Pubmed/

[5] RE Gagnon, AJ Macnab, FA Gagnon, A quantitative ranking of Canada's research output of original human studies for the decade 1989 to 1998, *Canadian Medical Association Journal*, 162 (2000) 37-40.

[6] K. Dickersin et al, Problems with indexing and citation of articles with group authorship, *Journal of the American Medical Association*, 287(2002) 2772-2774.

[7] E. Garfield, Citation analysis as a tool in journal evaluation, *Science*, 178(1972) 471-479.

[8] The OpenLDAP Project - http://www.openldap.org

[9] The World's Most Popular Open Source Database - http://www.mysql.org

[10] PHP: Hypertext Preprocessor - http://www.php.net

[11] P.O Seglen, Why the impact factor of journals should not be used for evaluating research, *British Medical Journal*, 314(1997) 498-502.

Adress for correspondance : Patrick Devos, Department of Biostatistics, University Hospital, 1,Av O. Lambret 59800 Lille, France
Tel. +33 3 20 62 68 09 E-mail : pdevos@univ-lille2.fr