Jointly published by Kluwer Academic Publishers, Dordrecht
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 50, No. 1 (2001) 33–57

Infometric methods and measures for sharing knowledge over internet

WILLIAM A. TURNER, R. GHERBI, C. JACQUEMIN, M. DE SAINT LEGER

LIMSI – CNRS, Université de Paris XI, Orsay (France)

This paper deals with knowledge sharing over Internet. After defining the concept, we will discuss work aimed at creating a technical system to implement it and at measuring the quality of results obtained. However, the reader will quickly see that the text is organized to address the theme of this special issue of *Scientometrics*. Models, methods and measures characterize scientometric research. What problems arise in attempting to develop them for Internet? In order to answer this question, it is important to distinguish between two schools of practice in the scientometric research field: the first derives from applied statistics and is called bibliometrics; the second derives from cognitive sociology and is called infometrics (*Turner*, 1994).

From bibliometrics to infometrics

Bibliometrics generally assumes that scientific activity is distributed over a very large number of scientific specialties which are studied in their own right, or aggregated using appropriate geographical and/or institutional criteria. Each specialty or aggregated unit is modeled as a productive system which is individually capable of transforming inputs into outputs, but which will have different performances when compared to other, similar systems. The goal of bibliometric research is to fully describe these differences. Three hypotheses justify the methods used. The first is that publications result from a research investment and, consequently, by counting these publications and studying how they are cited, different measures of system output can be developed. The second is that these measures are relevant for evaluating performance because they are built using representative samples of publications produced by the production systems themselves. The third is that we know enough about the social, cognitive and material conditions which affect the propensity to publish in order to apply the qualifying statement "all else being equal" when carrying out comparative studies.

0138–9130/2001/US $ 15.00

Infometrics, on the other hand, integrates science into the general socio-economic process structuring collective activity.* For example, the study presented in this paper concerns genetic resistance to antibiotics. While this is clearly a subject of fundamental importance for biological research, it also has wide socio-economic implications for individuals, the agro-industry, drug firms and public health. Scientific activity is open to ethical, political, economic and social considerations and structured by them; it can't be represented as a mosaic of largely autonomous production systems, each with its own set of resources and a recognized capacity for independently controlling the conditions of research in its area. To the contrary, science takes place in an open, active knowledge space (*Turner*, 1999) and although, at any given moment in time, specific work arrangements structure this space, these work arrangements are not the expression of a formal, socially stable division of labor but are, instead, temporary arrangements reflecting a particular organization of a diversity of interests. The goal of infometrics research is to describe these temporary arrangements as they evolve over time and, in so doing, to help detect the socio-cognitive elements structuring collective action. Obviously this type or research can benefit enormously from the existence of Internet because stakeholders in socio-economic processes are increasingly using this medium to voice their concerns and mobilize support for their various points of view. Because infometrics is concerned with the way in which public involvement in science is influencing both the amount and direction of resources given to fundamental research, Internet is a source of information which can't be ignored.

Infometrics and bibliometrics take very different approaches to modeling, methods and measures because they represent the nature of scientific activity differently. We can illustrate this point by looking critically at the three hypotheses presented above as underlying bibliometric research.

A great deal of work has been done in the past which aims at describing the social, cognitive and material conditions of doing research in different specialties. Bibliometrics has used the results of this work to develop frameworks for comparing the performances of research systems operating in different geographical and/or institutional contexts over time. Explication relies on a distinction between the social system of science on the one hand, and work practices under development in local contexts on the other hand. Three assumptions generally are used to justify the distinction: a) each specialty is a social system which enforces a specific set of socio-cognitive norms through peer review; b) these norms are universal in the sense that they define what

* See http://www.limsi.fr/WkG/PCD2000/indexeng.html for information on the activities of an international network researchers working on distributed collective practices.

constitutes good science; c) this definition is adhered to locally. Taken together, these three assumptions give meaning to publication and citation counts; figures obtained through bibliometric analysis can be used to class performances in terms of their contribution to good science.

Infometrics contests the idea of a universal definition of good science, considering instead that differences between science systems result from specific socio-cognitive arrangements which favor the formation of certain values rather than others. The argument derives from two basic sources: work on the institutional structures underpinning the social systems of science and work on growing public involvement in science. As early as the mid 1980s, a report was produced by the National Science Foundation which aimed at evaluating the peer review system (*NSF*, 1986). The report argued that the system functions well in the United States but that it only provides a technically competent scientific opinion on the value of knowledge claims and that, in fact, voice should be given to wider interests. For example, the growing importance of research for social and economic development implies that decisions on resource allocation have consequences not only for science itself but for industry, education and the society at large. The report concludes that peer review should be included in a larger evaluation process which it calls merit review on the basis that research has its own logic (peer review) and that in order to articulate results produced by the science system with larger social and economic demands, the evaluation process should be open to interested stakeholders from the outset (merit review). For the authors of the NSF report, peer review is the locus of priority setting in science, however, since publication of the report in the 1980s, the research environment has dramatically changed: major public health scares such as the mad cow disease or fears caused by the pursuit of genetic research have made the public suspicious of technical judgements which tend to evacuate ethical and moral issues. Public involvement in science has moved "upstream" in its concerns from scientific output, through scientific practice of designing field trials for genetically modified organisms, for example, to scientific priorities themselves (*Gibbons* et al., 1994). Internet has accompanied this movement by providing voice to stakeholders who are actively seeking to participate in the science arena.

The influence of institutional changes and a modified research environment on doing science and, consequently, on the propensity to publish, is not fully understood. Infometrics research is developing the theoretical concept of an active knowledge space in order to help produce this understanding. The problem addressed is that of representing empirically the mechanisms structuring collective action given the hypotheses outlined above: a) good science is an outcome of interactions between peers, stakeholders and the general public and not a concept which can be defined *a priori*;

b) merit review encompasses peer review and, because of growing public involvement in science, the technical opinions of scientists have to be considered as only one point of view on suitable directions for research, on an equal footing with moral, ethical, economic and social issues defended by other stakeholders in the research process; c) finally, the fact that science develops in an open space that is constrained neither teleologically by a predefined set of values nor practically by a formal set of institutional relationships raises the very difficult empirical question of delimiting the knowledge space underlying collective action. Infometrics requires a robust empirical mechanism in order to fix the limits of an active knowledge space and to obtain, through a document analysis, a meaningful representation of the working arrangements structuring that space.

Infometrics assigns a different role to publications in the research process than the one which is assigned to them in bibliometric research. From an infometrics perspective, publications serve to identify the objects which populate a knowledge space; to position them with respect to one another as elements structuring that space; and, finally, to measure changes in their structuring power over time given variations in the human, material and cognitive resources which they focus and mobilize. The idea underlying bibliometrics that research investments produce publications is redefined in infometrics: science doesn't produce publications; it produces concepts and/or artifacts which are active elements in a socio-economic process structuring collective activity.

In conclusion, then, we have shown that scientometrics is grounded in an understanding of the role that publications play in the knowledge production process. When publications are presented as the outcome of a research investment, as they are in bibliometrics, the underlying assumption is that information and knowledge can be considered as formally equivalent: a publication is considered to contain knowledge which has been coded onto some support (paper, electronic,…) for communication, storage, and future use. This assumption leads to what we will call in the following section of this paper a market model of scientific communication practices which denies Internet much interest for understanding scientific activity. Infometrics, on the other hand, carefully distinguishes between information and knowledge. Science produces concepts and artifacts but these will only become active elements of a knowledge production process to the extent that multiple investments are made to ensure their use. From an infometrics perspective, publications don't communicate knowledge. They provide information on where the products of a research system are produced, who is taking part in their production and the amount of resources being invested. They focus discussion, raise questions and generate fears which are now increasingly finding voice in the E-publications of Internet. In other words, whereas the market model of scientific communication practices downplays the importance of Internet, infometrics considers, on the contrary, that developing new methods for modeling the flow of E-publications is of fundamental importance to the future of scientometric research. In order to position E-publications within a scientometric framework, we have been working on an active knowledge space model of scientific communication practices as an alternative to the market model used in bibliometrics. This work is justified by what we said before about the changes in the dynamics of the merit review process resulting from growing public involvement in the scientific arena. New confidence building mechanisms are at work in science: the active knowledge space model seeks to identify and explain them.

Sharing knowledge over internet

Sharing knowledge requires confidence. The market model solves this problem by distinguishing three independent spheres of activity – knowledge production, dissemination and use – and then establishing formal relationships between them. In the first sphere, it is generally assumed as we saw above that publications code knowledge and that this knowledge is the outcome of a research investment. In the second sphere, actors of the information industry produce products and services designed to disseminate knowledge. Finally, end-users are located in the third sphere. They have knowledge needs which are inherent to their ways of life and they will consequently use the products and services of the information industry to meet them. Two additional hypothesis create linear relationships between these three spheres: one concerns the role of peer review in organizing the passage of research results into the editorial processes of the information industry; the other assumes that end users will act rationally and that quality, costs and relevance will determine their demand for information products and services.

This three sphere linear market model places information practices downstream from work done in laboratories. The added value produced by the information industry lies in its contribution to the dissemination of research results through a series of specific actions. For example, publishing houses organize the peer review process by setting up editorial committees to review articles submitted for publication. Database producers such as the PASCAL database in France or the *Science Citation Index* in the United States improve current awareness of on-going research by systematically collecting information on new publications, producing standardized bibliographic reference files and publishing a variety of paper, CD-Rom and other on-line bibliographic products.

Finally, brokers, on-line suppliers, libraries and documentation centers structure the information market offering information in a variety of forms (reports, synopses, books, journals), formats (paper, CD-ROMs, Internet) and at various prices.

E-publication practices over Internet have raised questions about the viability of the market model as presented above. One area concerned by these questions is peer review. On one side, the peer review system is defended tooth and nail as a rampart against the dangers inherent in cluttering up knowledge spaces with elements of little cognitive value. According to this argument, experts guard the temple of knowledge creation and dissemination and because they are posted at the temple's gates to filter out the truly useful information for transfer into the public domain, the market value of the selected information is greatly enhanced. End users know that, in principle, they are getting quality for their money. On the other side, however, what was said previously about growing public involvement in the science arena leads to quite a different conclusion. This paper insists upon the need to install merit review over Internet if confidence is to be sustained in the positive contribution of science to collective activity (*Turner* et al., 1996).

The market model uses a rationality hypothesis in its attempt to model the socio-cognitive interactions structuring knowledge production, dissemination and use. The active knowledge space model is grounded in a theory of organizational learning and the role that information access plays in structuring that learning process. The relationship between information access and organizational learning is far from direct. For example, economists insist upon the importance of learning mechanisms for improving performances (*Dosi* et al., 1988). These mechanisms are generally considered essential for establishing the rules, routines, values and codes of behavior that come to dominate a collective practice. However, the flip side of the learning process is that these codes, routines and values become irreversible (*Callon*, 1992). Organizations are initially open to outside influences, but as time goes on, they tend to reproduce tried and tested behavior, respond to environmental solicitations in a routine way, react favorably to certain signals and give preference to specific courses of action but systematically ignore others. In other words, they progressively get locked into a specific way of doing things and of looking at the world which determine their information behavior. Three corollaries derive from this irreversibility hypothesis. First, in terms of our discussion of peer review, we can expect technical evaluations to take on increasing importance over time as rules are forged for collectively working together. Second, this suggests the existence of an inverse relationship between confidence building through organizational learning and openness to the ethical, economic and social issues of merit review. Finally, if this inverse relationship holds, communities which are confident in their socio-cognitive norms should be less open to outside influences than communities whose norms are less well established.

LIMSI has recently initiated a major research program aimed at better understanding distributed collective practices. The work presented in this paper concerns microbiology; another area of study concerns Internet-mediated communication practices in the humanities (*Turner* and *D'Iorio*, 1999); and, finally, a large grant has recently been obtained to examine the use of Internet by civil action movements (*Henry*, 2000). The diversity of these field studies is designed to test for the inverse relationship postulated above. The methodology is the same in each case. An initial corpus is built around a decision to invest resources in a program (the study of genetic resistance to antibiotics, a hypertext for research on Nietzsche (*D'Iorio*, 2000), civil action in favor of North-South cooperation). A set of hypotheses deriving from the active knowledge space model and outlined in the following sections of this paper are then applied to enrich the initial document collection. The hypotheses express a specific point of view on the way to study the relationship between information access and up-date on the one hand, collective learning and doing practices on the other hand. The different field studies are designed to test if this particular point of view can be generalized across a variety of socio-cognitive contexts.

When a decision is taken to launch a collective action, there is no prior experience of organizational learning – no culture of working together – which means that the program is potentially open to outside influences. The movement towards closure of an open information space can be represented as a selection process: a community will elect to invest in a limited number of objects and explore only a limited number of relationships amongst all those suggested by a set of publications. In other words, the fit between results of a document analysis and the elements effectively structuring organizational activity will necessarily have to be improved through an iterative process. Furthermore, the need for such a process will be felt all the more strongly if one admits the principle of multiple viewpoints, the idea, for example, that investment priorities concerning genetic resistance to antibiotics will not be the same when considered from a public health, industrial or a fundamental research perspective. From an infometrics point of view, giving voice to multiple stakeholders raises a general problem which we will address later in the paper, that of corpus enrichment and its influence on how we represent the active elements of an information space.

Here then is a summary of the set of hypotheses we will be developing in the rest of the paper a) scientific activity produces concepts and artifacts; b) lists of words naming these objects can be extracted from document sets; c) the designated objects focus investment strategies which need to be articulated in order for an organization to move up a learning curve; d) statistical techniques can be applied to identify the active elements of an information space structuring this articulation work; e) a recursive strategy for corpus enrichment can be devised using these active elements.

The paper presents work to install a merit review process over Internet. It should be apparent from what was said above that the role of infometrics in this process lies at three levels: the application of natural language processing techniques to describe the open information space of a scientific activity; the empirical identification of the active elements structuring knowledge exchanges in that space; the adoption of corpus enrichment strategies which will help organizations move up the learning curve and ensure widespread stakeholder participation in the merit review process. Once again, our goal in the following discussion is to illustrate these themes as they relate to the subject of this special issue of *Scientometrics.* Research presented here started six months ago and for the moment our concern has been essentially to identify and experiment appropriate technologies for the task at hand. That said, the discussion which follows should help to understand the nature of an infometric approach to knowledge sharing over Internet.

Methods and measures

Describing information spaces using natural language processing techniques

In the subject area of this paper – genetic resistance to antibiotics – DNA and protein sequence data are of the utmost importance. These data are gathered and classed in two distinct categories: the core data and the annotation. For each sequence entry into a factual database, the core data consists of the sequence data, the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein) while the annotation consists of the description of its function, its structure and its biological action. Annotation information is obtained through publications and consultation of scientific experts. SWISSPROT is an annotated protein sequence database used in this study. It provides cross-references in the form of pointers to information related to SWISSPROT entries and found in data collections other than SWISSPROT. One such collection is GENBANK a DNA sequence database produced by the National Center for Biotechnology Information (NCBI) in the United States.

One of NCBI's missions is to build an integrated information space where researchers can move easily between annotated core data and the results of on-going biological research which are published in research journals. The NCBI supports and distributes GENBANK and a variety of other factual databases and is one of the prime movers in the United States of the PubMed journal project. PubMed is being promoted by the National Institute of Health as a Web search interface that provides access to 9 million journal citations in MEDLINE and contains links to full-text articles at participating publishers' Web sites and to core data in factual databases such as GENBANK or SWISSPROT. Projects like PubMed are under discussion in Europe and elsewhere. From a socio-cognitive point of view they raise the following question: how will people carve out active knowledge areas for themselves in these huge information spaces? As we explained before, this question is formally equivalent to the one which lies at the heart of infometrics research: namely that of better understanding how collective practices move up a learning curve; how, in other words, a set of values and norms is adopted which admits specific information processing rules and routines but not others.

The closure mechanisms at work in an information space are focused by debate over the choice of objects that should be collectively constructed which means that infometric research can't be understood in technical terms only, as a set of methods serving to feed information into this selection process. Evaluation techniques exist which address the technical performances of natural language processing techniques* but what we are seeking to do, is to say why certain topics extracted from the literature should be considered more important than others. Two types of narratives exist (schematically) for this purpose: the first is conservative while the second stresses the need for change. We have explained above why we favor the latter: transforming information into active elements of knowledge requires advancing arguments and developing these arguments require, at a more fundamental level of collective practice, adopting a critical attitude with respect to positions of authority. For example, no stakeholder should be able to impose a point of view based on status acquired through political force, economic might or scientific expertise and yet, all these attributes of power come to play when debate is engaged about how to organize collective investments in future forms of joint action. Its for this reason, then, that infometrics research seeks to develop a critical stance with respect to arguments of authority.

* LIMSI is actively involved in developing these techniques for both spoken and written language processing (see, for example, *Mariani* and *Paroubek*, 1999)

A simple method for identifying the objects structuring a research activity is to use a controlled vocabulary. For example, the names contained in the *United Medical Language System* (UMLS) of MEDLINE describing genes, species and antibiotics were used to obtain a first list of elements structuring research in our subject area. However, this list of names is an authoritative list and, as we just said, infometrics requires a test of relevancy. To what extent is a controlled vocabulary, which is defined *a priori*, representative of science in action and to what extent does it give voice to the range of subjects that mobilize the various interests of concerned stakeholders? In order to answer this type of question, we adopted two approaches to term extraction: the first is a top-down pattern matching procedure exploiting the UMLS word list; the second is a bottom-up term acquisition procedure that doesn't use a controlled vocabulary. Experiments were run on two document sets, one containing 557 bibliographical notices from MEDLINE and the other containing information from 450 SWISSPROT notices. An example of the content of these two data sets follows:

```
<TI>OVEREXPRESSION OF THE D-ALANINE RACEMASE GENE CONFERS RESISTANCE TO D-CYCLOSERINE IN
        MYCOBACTERIUM SMEGMATIS.</TI>
<AU>CACERES-NE; HARRIS-NB; WELLEHAN-JF; FENG-Z; KAPUR-V; BARLETTA-RG</AU>
<AD>DEPARTMENT OF VETERINARY AND BIOMEDICAL SCIENCES, UNIVERSITY OF NEBRASKA, LINCOLN 68583-0905,
        USA.</AD>
<SO>J-BACTERIOL. 1997 AUG; 179(16): 5046-55</SO>
<JN>JOURNAL-OF-BACTERIOLOGY</JN><ISSN>0021-9193</ISSN><PY>1997</PY><LA>ENGLISH</LA><CP>UNITED-
        STATES</CP>
<AB>D-CYCLOSERINE IS AN EFFECTIVE SECOND-LINE DRUG AGAINST MYCOBACTERIUM AVIUM AND MYCOBACTERIUM
        TUBERCULOSIS. TO ANALYZE THE GENETIC DETERMINANTS OF D-CYCLOSERINE RESISTANCE IN
        MYCOBACTERIA, A LIBRARY OF A RESISTANT MYCOBACTERIUM SMEGMATIS MUTANT WAS CONSTRUCTED. A
        RESISTANT CLONE HARBORING A RECOMBINANT PLASMID WITH A 3.1-KB INSERT THAT CONTAINED THE
        GLUTAMATE DECARBOXYLASE (GADA) AND D-ALANINE RACEMASE (ALRA) GENES WAS IDENTIFIED.
        SUBCLONING EXPERIMENTS DEMONSTRATED THAT ALRA WAS NECESSARY AND SUFFICIENT TO CONFER A D-
        CYCLOSERINE RESISTANCE PHENOTYPE. THE D-ALANINE RACEMASE ACTIVITIES OF WILD-TYPE AND
        RECOMBINANT M. SMEGMATIS STRAINS WERE INHIBITED BY D-CYCLOSERINE IN A CONCENTRATION-
        DEPENDENT MANNER. THE D-CYCLOSERINE RESISTANCE PHENOTYPE IN THE RECOMBINANT CLONE WAS
        DUE TO THE OVEREXPRESSION OF THE WILD-TYPE ALRA GENE IN A MULTICOPY VECTOR. ANALYSIS OF A
        SPONTANEOUS RESISTANT MUTANT ALSO DEMONSTRATED OVERPRODUCTION OF WILD-TYPE ALRA ENZYME.
        NUCLEOTIDE SEQUENCE ANALYSIS OF THE OVERPRODUCING MUTANT REVEALED A SINGLE TRANSVERSION
        (G--&GT; T) AT THE ALRA PROMOTER, WHICH RESULTED IN ELEVATED BETA-GALACTOSIDASE REPORTER
        GENE EXPRESSION. FURTHERMORE, TRANSFORMANTS OF MYCOBACTERIUM INTRACELLULARE AND
        MYCOBACTERIUM BOVIS BCG CARRYING THE M. SMEGMATIS WILD-TYPE ALRA GENE IN A MULTICOPY
        VECTOR WERE RESISTANT TO D-CYCLOSERINE, SUGGESTING THAT ALRA OVERPRODUCTION IS A POTENTIAL
        MECHANISM OF D-CYCLOSERINE RESISTANCE IN CLINICAL ISOLATES OF M. TUBERCULOSIS AND OTHER
        PATHOGENIC MYCOBACTERIA. IN CONCLUSION, THESE RESULTS SHOW THAT ONE OF THE MECHANISMS OF D-
        CYCLOSERINE RESISTANCE IN M. SMEGMATIS INVOLVES THE OVEREXPRESSION OF THE ALRA GENE DUE TO
        A PROMOTER-UP MUTATION.</AB>
<MJME>*ALANINE-RACEMASE-GENETICS; *ANTIBIOTICS,-ANTITUBERCULAR-PHARMACOLOGY; *CYCLOSERINE-
        PHARMACOLOGY; *MYCOBACTERIUM-GENETICS</MJME>
<MIME>ALANINE-RACEMASE-BIOSYNTHESIS; ALANINE-RACEMASE-METABOLISM; AMINO-ACID-SEQUENCE; CLONING,-
        MOLECULAR; DRUG-RESISTANCE,-MICROBIAL-GENETICS; GENE-EXPRESSION-REGULATION,-BACTERIAL;
        GENOMIC-LIBRARY; MOLECULAR-SEQUENCE-DATA; MUTATION-; MYCOBACTERIUM-DRUG-EFFECTS;
        MYCOBACTERIUM-ENZYMOLOGY; MYCOBACTERIUM-AVIUM-COMPLEX-GENETICS; MYCOBACTERIUM-BOVIS-
        GENETICS; OPEN-READING-FRAMES; PROMOTER-REGIONS-GENETICS; SEQUENCE-ALIGNMENT</MIME>
<TG>SUPPORT,-NON-U.S.-GOV'T; SUPPORT,-U.S.-GOV'T,-NON-P.H.S.; SUPPORT,-U.S.-GOV'T,-P.H.S.</TG>
<SH>BIOSYNTHESIS; GENETICS; METABOLISM; PHARMACOLOGY; DRUG-EFFECTS; ENZYMOLOGY</SH>
<PT>JOURNAL-ARTICLE</PT>
<SI>GENBANK/U70872; GENBANK/U00020; GENBANK/Z77165; GENBANK/M19142; GENBANK/M16207; GENBANK/U00020;
        GENBANK/Z77165; GENBANK/L02948; GENBANK/K02119; GENBANK/M12847; GENBANK/U00006;
        GENBANK/L46206</SI>
<RN>EC 5.1.1.1; 0; 68-41-7</RN>
<NM>ALANINE-RACEMASE; ANTIBIOTICS,-ANTITUBERCULAR; CYCLOSERINE</NM>
<CN>AI40365AINIAID</CN>
<JC>BACTERIOLOGY</JC>
<AN>1997405901</AN><DA>970905</DA><UD>199711</UD><SC>0021-9193(199708)179:16L.5046:OARG;1-L</SC>
```

Figure 1. Example of a MEDLINE bibliographic notice

Swissprot ID	Protein names and synonyms	Genes	Source Species
AAC1_PSEAE	GENTAMICIN 3'-ACETYLTRANSFERASE (EC 2.3.1.60) (GENTAMICINACETYLTRANSFERASE I) (AMINOGLYCOSIDE N3'-ACETYLTRANSFERASE I)(AAC(3)-I).	AACC1	Pseudomonas aeruginosa
AAC2_ACIBA	AMINOGLYCOSIDE N3'-ACETYLTRANSFERASE III (EC 2.3.1.81) (GENTAMICIN-(3)-N-ACETYL-RANSFERASE) (AAC(3)II).	AACC2	Acinetobacter baumannii
.... PAT_STRHY	.... PHOSPHINOTHRICIN ACETYLTRANSFERASE (EC 2.3.1.-).	.... BAR	.... Streptomyces hygroscopicus

Figure 2. Example of information found in SWISSPROT

A top-down pattern matching procedure exploits a controlled vocabulary, each word being composed of a chain of characters (pattern) that then has to be identified as appearing in a data set. Using the UMLS word list defined above, this procedure was applied to index the MEDLINE data set and 334 subjects were identified. The notion of "top-down" applies to the fact that, by construction, all the elements in a controlled vocabulary are relevant to describing the information space of a research effort. That said, the technique enforces an authoritative definition of scientific practice and for this reason a second method was designed to implement an empirical approach to defining the active elements of the subject area's information space.

A bottom-up term acquisition strategy was used in order to identify strings of characters (words) which were simultaneously present in both the SWISSPROT and MEDLINE data. We interpreted the co-presence of words in the two data sets as indicating objects of investment where researchers are making an effort to enrich existing factual information by on-going research; as a measure, in other words, of the potential usefulness of research publications cited in MEDLINE for augmenting the information in SWISSPROT on the structure, function and action of genes explaining their resistance to antibiotics. This interpretation is coherent with the goal of infometric research: active elements of knowledge are those which have not yet been stabilized to the point where new investments are considered as being a waste of collective resources. They focus an underlying tension between the stabilized core of a scientific practice on the one hand and the constant need to redefine this core in the light of new research on the other hand. Building new objects implies modifying the borderline between what is taken for granted and what should be revised. Merit review was the name given above to the process of collectively fixing borderlines. It is during merit review that a research community moves up the learning curve: arguments are advanced, combated, accepted or rejected and its through these interactions that the norms and routines of a collective practice are forged. For all these reasons, then, we used the co-presence of words in MEDLINE and SWISSPROT data to empirically identify the active elements of a merit review process.

We will not describe the technical work carried out at LIMSI to implement the bottom-up term acquisition strategy.* It produced a list of 1057 terms which is approximately three times bigger than the list of 334 terms produced by the top-down technique for describing exactly the same information space. Here, then, is a measure of

* Patrick Paroubek who works in the Spoken Language Processing Group at LIMSI developed the method used. It is described in an interim report to the Life Science Department of the CNRS intitled, "Bilan d'une expérience de veille scientifique" "Microbiologie – résistance aux antibiotiques", CNRS/SdV, Inist/Uri, CNRS/Limsi, Octobre 2000.

the potential bias in using authoritative, controlled vocabularies for indexing on-going research. However, this interpretation is obviously subject to caution. The difference in the size of the two lists might be more simply explained by an automatic procedure that has little cognitive relevance. This problem of giving a socio-cognitive interpretation to the results of an infometric analysis has to be addressed if infometricians are to actively take part in debates fixing the borderlines of collective activity.

The socio-cognitive signification of natural language processing results

When the same corpus is used to generate word lists, the results obtained through application of different algorithms are often compared by assuming that the corpus delimits an information space and that each word list will consequently be a more or less precise inventory of the objects which one would expect to find in that space. The hypothesis is then controlled by expert evaluation; scientists are called upon to say if they are satisfied with the word list produced automatically. An alternative form of authoritative control is to use a controlled vocabulary as a reference for judging the performance of automatic techniques. However, these approaches are anchored in the idea of peer review while the goal of infometrics is to defend merit review for the reasons explained above.

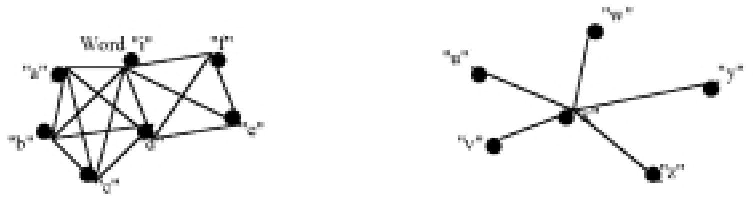
Merit review is a process through which the diversity of stakeholder interests structuring collective activity are progressively articulated. Articulation work generates conflicts of interpretation, debate and tension but these are generally surmounted: when irreversibility sets in, the political, social and economic dimensions of collective activity fade into the background and the onus is put on the technical skills needed for competent social participation. There is no need to defend peer review; collective practices inevitably produce situations where social interaction skills are judged on the basis of technical criteria. The formation of standardized vocabularies codes the bureaucratization of collective practices.* We are attempting to use natural language and co-word processing techniques to better understand the socio-cognitive dynamics explaining the formation of these standardized vocabularies.

Co-word analysis techniques were initially developed in France through cooperation between the CNRS and the Paris School of Mines in the 1970s and the 1980s (*Callon* et al., 1986; *Turner* et al., 1988). These techniques use several statistics to describe the structure of word association patterns which characterizes a document set. These statistics are derived from frequency counts (C_i, C_j); co-occurrence counts (C_{ij}) and the

* This point is excellently made in a recent book by *Bowker* and *Star* (1999).

number of different documents in which a word appears, given the total number of documents in the set (N). Co-word techniques can thus be used to exploit the results obtained after applying the top-down and bottom-up term extraction procedures described earlier. Obviously, the statistics will differ: N is the same for both applications (557 MEDLINE notices) but frequency counts and co-occurrence counts will vary considerably given that the automatic extraction process produced 3 times more words than the controlled vocabulary approach (1057 for the former, 334 for the latter). Equally obvious, then, is that the word association patterns produced by the application of co-word techniques will not be the same when the initial statistical data differs. These structural differences in word association graphs are studied in order to empirically implement the perspectives set out in the preceding paragraph.

Many different statistical techniques can be applied to visualizing word association patterns depending upon the properties of word use that one is seeking to study. We consider word lists as being an inventory of objects structuring an information space; in visualizing their association patterns, we want to clarify the role of these objects in structuring this information space. New co-word analysis measures were recently developed for this purpose (*de Saint Leger*, 1997). The structure of a word association graph presents two types of local configurations: a tightly reticulated pattern is formed by totally connected components; star-like patterns link the highly reticulate regions of a graph together.



A) Highly reticulated totally connected components

B) Star-like configuration patterns

Figure 3. Different word association patterns generated by a co-word analysis

The new software classes words according to the power they exert on the structure of local patterns found in word association graphs. For example, in the configuration shown on Figure 3B, only the central node of the star (word "x") exerts a structuring influence on the word association graph, whereas in figure 3A, we see that word "i" contributes to reticulating a complex containing 9 totally connected components:

"i", "a", "b";
"i", "b", "c"
"i", "c", "d"
"i", "a", "c"
"i", "b", "d"
"i", "a", "d"
"i", "d", "e"
"i", "e", "f"
"i", "d", "f"

A comparative study of word lists generated through a natural language analysis consists, first, in locating where the words on these lists fit into the association structures visualized on Figures 3A and 3B above; second, in automatically generating document retrieval strategies on the basis of these word association patterns in order to identify stakeholders investing in a specific region of the information space; and, third, in using this information on stakeholder investment strategies to take part in the merit review process. With respect to the second point, the link between word association patterns and investment strategies is established through the boolean operator "or" to connect the three words defining totally connected components (Figure 3A) and through the operator "and" to connect words in star-like configurations (Figure 3B : "x" and "u"; "x" and "v", etc.). This new evaluation perspective will be better understood if we consider a concrete example.

The gene *vanA* appears on both lists of our study. It was identified in the lexicon of 334 terms and in the lexicon of 1057 terms as being an element structuring the microbiological research field under study. When these lexicons were used to build word association graphs, despite variations in the basic statistics serving for the co-word analysis (C_i, C_j, C_{ij}, N), *vanA* appeared in local, totally connected configurations on both graphs (the parameters used to visualize the word association patterns were the same in both cases). However, the level of reticulation was not the same: *vanA* appeared in a configuration containing 5 words and only three totally connected components in the graph of 334 terms, while in the graph of 1057 terms, its association environment was defined by 7 words and 6 totally connected components. The boolean operator "or" was used to combine the elements of these two word sets, producing two distinct document retrieval strategies. We discovered that 3 out of the 5 words in the first set and 5 out of the 7 words in the second set were specific to each graph, which means that only two words were common to both retrieval strategies.

This observation points to the existence of a naming problem: when natural language is used to code knowledge often the same word will designate different objects and the same object will be designated by different words. Controlled vocabularies offer a solution to this problem, however, we have already criticized this line of reasoning: what is called a naming problem isn't in fact a problem at all; its a crucial feature of research, the expression of a collective effort to move up the learning curve. A standardized vocabulary is the result of social interactions, but it doesn't condition them. Our assumptions are consequently different to those which derive from concern for the naming problem: a) when word lists are automatically extracted from articles presenting on-going research they provide more accurate information on objects structuring science in action than a controlled vocabulary; b) using standardized vocabularies to launch studies on investment strategies privileges dominant investment strategies, not those which are seeking to move collective practices into new areas; c) naming objects must be subordinated to the general problem of identifying stakeholders investing in structuring an information space.

It is this third point which we address in the following example. When we applied the two document retrieval strategies defined above, a much broader perspective on the information horizon structuring *vanA* research was obtained using automatic term extraction techniques.

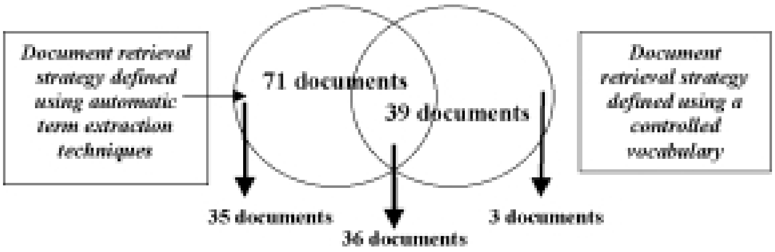

Document retrieval strategy defined using automatic term extraction techniques
71 documents
39 documents
Document retrieval strategy defined using a controlled vocabulary
35 documents
36 documents
3 documents

Figure 4. Retrieval strategies and the definition of an information space

Figure 4 shows that the document set retrieved using the automatic term extraction technique was 56% bigger than the one retrieved using the controlled vocabulary: 71 documents were identified in the first case; 39 in the second, and 36 documents were common to both. Only three documents were identified using the controlled vocabulary which weren't identified using the automatic techniques. However, these latter techniques produced 35 more documents than the controlled approach. The overlap between the two approaches gives confidence in the fact that a core set of documents exists defining the information space structuring *vanA* research. The authors (laboratories and countries) of the 36 documents in this core set are considered as the "principle stakeholders" in the field. The authors of the 35 documents identified through the automatic procedure are considered as "outsiders".

Outsider participation in the merit review process

The merit review process is punctuated by moments of collective discussion and debate which are essential for focussing efforts to collectively move up a learning curve. Organizing these discussions over Internet requires appropriate human-machine interfaces. The following (schematic) example uses the results presented in Figure 4 to show how we envisage the computer's role in the merit review process. We consider it as an assistant for visualizing investment strategies: the horizontal axis on the following diagram corresponds to a word list extracted from the 36 documents produced by the principle "stakeholders" in the field; the vertical axis corresponds to a second word list extracted from the "outsider" literature (35 documents). We have put these two words in quotation marks in order to insist upon the fact that the definition of "stakeholders" and "outsiders" derives from the classification algorithms used for document analysis. Its important to clearly understand that the two categories don't define a socio-cognitive reality; their usefulness resides in the fact that they can be used to debate the question of what that reality actually is. Infometricians need a definition of what is important and what isn't in order to take part in this debate and the computer helps to produce it by identifying and classing the objects structuring "outsider" and "stakeholder" investment strategies. The elements on each list are ordered by decreasing importance according to the power they exert on the structure of a co-word graph (see Figure 3). Five situations needing interpretation are defined on the following diagram:

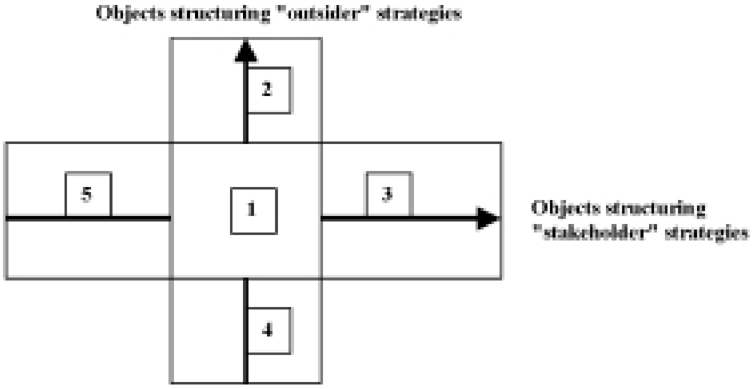


Figure 5. Making sense out of infometric methods and measures in order to sustain the merit review process over Internet

Designing a computer environment for merit review implies a clear definition of the sense-making process underlying the review. Figure 5 shows two lines of interpretation. We can expect that the subjects arrayed along the horizontal axis will be active elements of the process leading to closure of the information space, the reason being that these subjects are extracted from documents that have been indexed using a controlled vocabulary. We have suggested that the formation of a standardized vocabulary is a good indicator of growing irreversibility; by ordering this vocabulary along the X axis of Figure 5 we can identify the objects that are, according to the results of our statistical analysis, likely to be the most active elements structuring the information space (group 3) and the least active (group 5). The vertical axis orders the subjects which are (by definition) of peripheral interest to the principle stakeholders in the research field. The subjects listed in group 2 are likely to be useful in an attempt to keep the merit review process open to outside influences, more useful in any case than the subjects in group 4, which have a lower power of attraction. The interpretation of group 1 words is ambiguous because, as Figure 5 shows, they delimit a region of the information space which is shared in common by both "outsiders" and "stakeholders". Does this mean that the objects structuring exchanges in this area are more likely to focus tensions and conflicts than those which are located in the segregated regions of the graph (identified by groups 2 and 4 for "outsiders" and groups 3 and 5 for "stakeholders")?

Words are the active elements underlying the socio-cognitive dynamics of computer-mediated interactions. By grouping these words in different classes and by assigning to each class a function in structuring collective practice, we have distinguished three categories: words designating "common interests" (group 1); words designating "stakeholder" interests (groups 3 and 5); and words designating "outsider interests" (groups 2 and 4). The next phase of our research will consist in using these categories to define corpus enrichment strategies which are coherent with the merit review process. This means concretely that when words from one of these three categories are used to formulate a document retrieval strategy, only the documents which are indexed by a set of words which are similar to the ones serving to define the category will be retained. Natural language processing techniques will be used together with similarity measures to do this filtering. A number of questions are being addressed such as overcoming the naming problem mentioned earlier and developing appropriate statistical thresholds for determining similarity.

It might be asked why the naming problem needs to be treated in connection with corpus enrichment, but should be left aside when the goal is to generate the word groupings shown in Figure 5. In attempting to answer this question, we will present our solutions for managing the dynamics of the recursive merit review process. As we explained above, the formation of standardized vocabularies is considered as an indication of progression towards the use of increasing technical criteria for evaluating interaction skills. How then do we track this movement? Given that standardized vocabularies are forged through debate and intense social interaction, the merit review process can be represented as a succession of moments "t", "t+1",... when these debates formally take place. Figure 5 represents the synchronic dimension of the process: word groupings at "t+1" need to be compared with those at "t" in order to evaluate progress in the formation of a standardized vocabulary. Corpus enrichment concerns the diachronic dimension of this evaluation: what makes it legitimate to compare two moments in time? Temporal comparisons take on meaning in infometrics to the extent that the similarity between the document sets used to construct word groupings at "t" and "t+1" is statistically significant. The coherence of corpus enrichment strategies has to be statistically tested in order to build a useful human-machine interface for organizing the merit review process over Internet.

In order to carry out this statistical testing we need to resolve the naming problem. It derives from morphological, syntactic and semantic variations of word use in natural language. A system, FASTR, has been developed at LIMSI to detect these variations and will be applied in the next phase of our work (*Jacquemin*, 2000). An example of results produced by FASTR concerns the different forms that the concept "resistance strains" takes in the texts of MEDLINE citations (see Figure 1). More than 30 different forms were automatically identified:

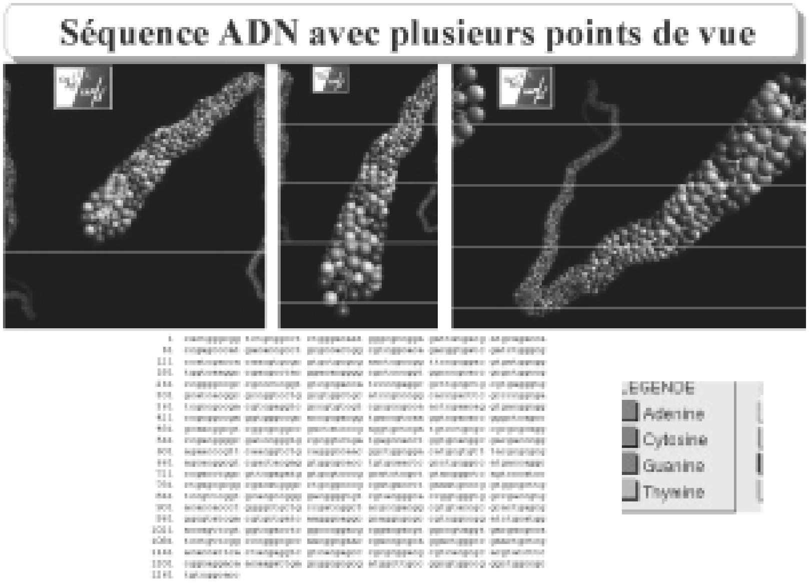
```
vancomycin-resistant strains
vancomycin-resistant enterococci.strain
tetracycline-resistant strains
streptomycin-resistant strains
streptomycin-resistant and streptomycin-dependent strains
rifampin-resistant and rifampin-susceptible strains
rifampicin-resistant strains
rifampicin-resistant pseudomonas fluorescens strain
resistant streptococcus agalactiae strain
resistance of n. asteroides strain
resistance of a wild-type p. aeruginosa strain
resistance of a nalb strain
resistance in prsp strains
pza-resistant strains
piperacillin-resistant strains
oxazolidinone-resistant strain
ofloxacin-resistant strains
mupirocin-resistant strains
mls-resistant strains
methicillin-resistant staphylococcus aureus strains
methicillin-resistant s. aureus strains
methicillin-resistant and methicillin-susceptible strains
kanamycin-resistant strains
irovanoxacin-resistant strains
inh-resistant m. tuberculosis strain
imipenem-resistant strains
gyrase-resistant ) strains
erythromycin-resistant clinical strain
emb-resistant strains
clarithromycin-resistant strain
ceftazidime-resistant strains
cefotaxime-resistant salmonella typhimurium strains
beta-lactam-resistant strains
aztreonam-resistant k. oxytoca strain
```

A final point concerns the rapid evolution of computer-mediated interactions through the implementation of new multi-modal applications. Multi-modal research aims at integrating different forms of interaction (speech, vision, gestures, text-based exchanges) through interfaces that are as "natural" as possible: social interactions are needed to transform information into active knowledge for collective activity; these interactions will be made easier with interfaces integrating different modalities. PubMed provides evidence that this same hypothesis is seriously being entertained by the database industry. The reading environments which are being offered through on-line Web sites by publishers taking part in the PubMed project include 3D displays of DNA and protein sequences. For the moment, these 3D images can't be manipulated, nor can vocal or textual comments be attached to the display. That said, LIMSI is a fundamental research laboratory of the CNRS doing multi-modal research and it is now starting to develop interface applications for use in the field of E-publishing. One attempt in this direction has already been made.

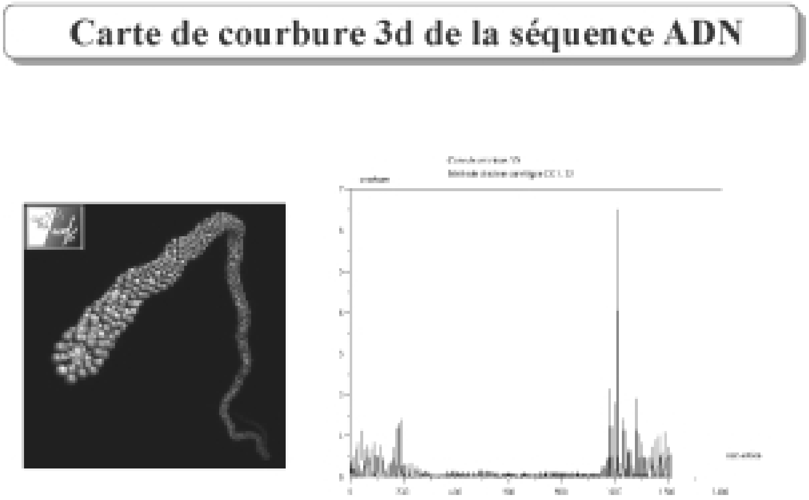
LIMSI is currently developing a 3D viewer for displaying and manipulating DNA sequences (*Gherbi* and *Hérisson*, 2000). The MEDLINE notice appearing in Figure 1 contains several cross-references to GENBANK. In other words, the article cited is classed as being a contribution to the core information on genetic structures available through GENBANK. The sequence which is the object of the empirical research in the article is described as follows in GENBANK:

```
   1  cactgggcgg tctgcggcct ctgggacaat gggcgccgga gattatgacg   atgcagacca
  61  ccgagcccat gacaccgcct gcgccactgg cgtcggcaca gacggtgatc gatctgggcg
 121  ccatcgacca caacgtgcgc gtgctgcgcg aactcgccgg ttccgcggac gtgatggcgg
 181  tggtcaaggc cgacgcctac ggacacgggg cgctcccggt ggcccgcacc gcgctggccg
 241  ccggggccgc cgccctcggt gtcgcgacca tccccgaggc gcttgcgctg cgtgagggtg
 301  gcatcacggc gccggtcctg gcgtggctgc atccgcccgg caccgacttc gccccggcga
 361  tcgccgccga cgtcgaggtc gccgtgtcgt cgcgccgcca actcgaacag gtgacggcgg
 421  ccgcggccga ggtgggccgc accgcgacgg tgaccgtcaa ggtcgacacc gggctcagcc
 481  gcaacggcgt cggcgcggcc gactatcccg aggtgctcga tgtcctgcgc cgcgcgcagg
 541  ccgacggggc gatccgggtg cgcggtctga tgagccacct ggtgcacggc gacgacccgg
 601  agaacccgtt caacggtctg cagggccaac ggctggcgga catgcgtgtt tacgcgcgcg
 661  agcacggcgt cgactacgag gtggcgcacc tgtgcaactc gcctgcggcc atgaccaggc
 721  ccgacctggc gttcgagatg gtgcgtcccg gcatctcgct gtacgggctc agtcccatcc
 781  ctgagcgcgg cgacatgggc ctgcggcccg cgatgacctt gaaatgcccg gtggcgcttg
 841  tccgttcggt gcacgccggg gacggggtgt cgtacgggca ccggtgggtg gccgaccgtg
 901  acaccaccct ggggctgctg ccgatcggct acgccgacgg cgtgtaccgc gcactgagcg
 961  ggcgtatcga cgtgctgatc aagggcaggc gcaggcgcgc cgtcggccgg atctgcatgg
1021  accagttcgt ggtcgacctc ggcccggacg cggacgacgt ggccgtaggt gacgacgcca
1081  tcctgttcgg cccgggcgcc aacggcgaac cgaccgcgca ggactgggcc gaactgctcg
1141  acaccattca ctacgaggtc gtcacgagcc cgcgcggacg cgtcacgcgc acgtatcttc
1201  cggcaggaca acaagattga   gcggcgcgcg atggcttgcc ggcgtggccg ggctggccgc
1261  tgtcggcacc
```

The grey part of the above table was used directly in the DNA Viewer of LIMSI to produce the following displays:



Séquence ADN avec plusieurs points de vue



Carte de courbure 3d de la séquence ADN

The environment for accessing the information spaces of science is rapidly changing with the evolution of new reading interfaces for on-line journals. In its effort to develop new multi-modal interfaces, LIMSI is experimenting the conditions of providing new computer-mediated access to scientific information. In the near future, we will be able to produce information spaces that contain publications, the list of objects extracted from those publications, 3D visual displays of their factual content and annotations added either vocally or textually to each object. This research is being undertaken in order to better understand the socio-cognitive conditions of using a computer to help share collectively produced knowledge.

Conclusions

We used the PubMed project which is being promoted by the National Institute of Health in the United States as an example of a project which has motivated the infometrics research described in this paper. Web search interfaces are being produced which will allow for extensive navigation between journal databases and factual data, but how will people use these huge information spaces in structuring their collective activity? In trying to answer this question we have insisted upon the need to look more closely at the dynamics of social interaction and, in particular, at the conditions for collectively moving up a learning curve. We have argued that doing things together requires articulating stakeholder interests and, as time goes on, irreversibility sets in: certain values, routines and practices come to dominate others and those that don't respect them are considered as "outsiders". Raising the question of *Scientometrics and Internet* requires taking a position with respect to the role we would like to see technology play in organizing collective practices. Will computer-mediated interactions reinforce a tendency to exclude outsiders from taking part in collective learning processes?

This question is broader than the one normally addressed in scientometric research, which is generally limited to that of defining suitable indicators for describing and explaining the operation of the science system. However, as we have said before, science is an area of growing public involvement in policy debates and we are witnessing, because of this involvement, an increasing focalization on the distinction between experts and non-experts. By definition, non-experts – the adverted public, educators, students, public opinion leaders – are outsiders. We feel that *a priori* definitions of this type should be avoided and that the onus should be placed instead on using Internet to encourage wider public participation in the science arena. A better understanding of how information is transformed into active knowledge for collective action is required to organize this participation, together with the design of new multi-modal interfaces for collectively sharing knowledge.

*

Work presented here is part of project organized by the Life Science Department of the CNRS. LIMSI's specific contribution to this bio-informatics project lies in the design of computer-mediated interaction interfaces exploiting speech, text and 3D visualization technologies currently under development in the laboratory. Other CNRS units participating in the project are the Bibliometrics Unit of the Life Science Department (project coordinator); the Research and Innovation Unit of the Technical and Scientific Information Institute (INIST), the Service for Molecular and Genome Bioinformatics and the Science Policy Indicator Unit of the Direction for Strategy and Programs.

The authors would like to acknowledge the help received from Patrick *Paroubek*, Marie-Thérèse *Maunoury*, Philippe *Dessen*, Bernard *Arrio* and Claude *Henry* during the learning and doing experience that we have gone through together.

Bibliography

BOWKER, G.C., STAR, S. L. (1999), *Sorting things out : Classification and its consequences.* Cambridge, MA : MIT Press.

CALLON, M. (1992), Variety and irreversibility in networks of technical conception and adoption, In: FORAY, D., FREEMAN, C. (Eds), *Technology and the Wealth of Nations,* London : Francis Pinter

CALLON, M., LAW, J., RIP, A. (1986*), Mapping the Dynamics of Science and Technology*, London : MacMillan.

DE SAINT LEGER, M. (1997), Modélisation des flux d'information scientifique et technique par le bruit : vers un suivi de l'évolution des domaines de connaissances, *Thèse d'Université*, CERESI/CNRS-CNAM.

D'IORIO, P. (2000), *HyperNietzsche*, Paris: Presses Universitaires de France, Collection Ecritures Electroniques

DOSI, G., FREEMAN, C., NELSON, R., SILVERBERG, SOETE, L. (Eds) (1988), *Technical Change and Economic Theory,* London : Francis Pinter.

GHERBI R., HÉRISSON, J. (2000), ADN_Viewer: a software framework toward 3d modeling and stereoscopic visualization of the genome, *10th International Conference on Computer Graphics and Computer Vision,* Moscow, Russia, August.

GIBBONS M., SCOTT, P., LIMOGES, C., SCHWARTZMAN, S., TROW, M., NOWOTNY, H. (1994), *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*, London: Sage Publications.

HENRY, C. (2000), *Outiller les Alliances*, Projet labellisé par le Réseau national de recherche en télécommunications (RNRT), France.

JACQUEMIN, C. (2000), *Spotting and Discovering Terms through NLP*, MIT Press, Cambridge MA, to appear.

MARIANI, J., PAROUBEK, P. (1999), Human Language Technologies Evaluation in the European Framework,, *Proceedings of the DARPA Broadcast News Workshop*, Washington, February 1999, Morgan Kaufman Publishers, ISBN-1-55860-638-6, pp. 237–242.

NSF (1986), *Final Report of the NSF Advisory Committee on Merit Review*, Washington, D.C.: National Science Foundation.

TURNER, W. A. (1999), Interaction skills of information professionals: Studies for Collaboratory Design, *Workshop on Digital Collaboration Technologies, the Organization of Scientific Work and Economics of Knowledge Access*, Laxemburg, Austria : organized by the National Science Foundation, International Institute for Applied Systems Analysis, European Science Foundation, December

TURNER, W. A., D'IORIO, P. (1999), *Nietzsche sur Internet : L'observation des collaborations médiatisées par ordinateurs dans les sciences de l'érudition*, CNRS : Micro Bulletin Thématique : L'information scientifique et technique et l'outil Internet, Paris : Editions CNRS.

TURNER, W. A. (1994), What's in a R: info*R*metrics or *info*metrics? *Scientometrics*, 30 (2-3) 471–480.

TURNER, W. A., CHARTRON, G., LAVILLE, F., MICHELET, B. (1988), Packaging information for peer review : new co-word analysis techniques, In: VAN RAAN, A. F. J. (Ed) *Handbook of Quantitative Studies of Science and Technology*, Amsterdam : Elsevier.

TURNER W. A., DE GUCHTENEIRE, P., VAN METER K. (1996), Evaluation of scientific merit in collaboratories: A new look at an old question, *Accountability in Research*, 5 : 73–93.

WAYNES, C. L. (1998), Topic Detection and Tracking: A Case Study in Corpus Creation and Evaluation Methodologies, *1st International Conference on Language Resources and Evaluation* (LREC98), Granada, Spain, May, pp. 111–115.

Received November 10, 2000.

Address for correspondence:
WILLIAM A. TURNER
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)
National Center for Scientific Research (CNRS)
Bat. 508, Université de Paris XI,
BP. 133, 91403 Orsay Cedex (France)
E-mail: turner@limsi.fr