Jointly published by Elsevier Science B. V., Amsterdam and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 33, No. 1 (1995) 41-63

ON THE SIGNIFICANCE OF DATA BASES KEYWORDS FOR A LARGE SCALE BIBLIOMETRIC INVESTIGATION IN FUNDAMENTAL PHYSICS

M. G. SURAUD,* L. QUONIAM,** H. ROSTAING,** H. DOU**

** L. E. R. A. S. S., IUT-A, 115 Route de Narbonne, 31077 Toulouse cedex (France)*
*** C. R. R. M., Centre Saint-Jérome, 13397 Marseille cedex 13 (France)*

(Received November 14, 1994)

We present an automatized bibliometric investigation applied to the field of fundamental research in physics. We briefly describe the scientific context motivating this study and the statistical method used for analyzing the data. We discuss in more detail how we adapted our investigation to the questions motivating this study, namely the identification of relevant groups working in a well defined subfield of physics. We next present the results of our investigation. We particularly focus on an analysis of Index and Free terms, as obtained from the INSPEC data base we used for performing the bibliometric investigation. We discuss the relevance of Index and Free terms by means of a separation between "Noise", "Interesting" and "Trivial" entries. We show that Index and Free terms exhibit somewhat different behaviors when considered as distributions in terms of frequencies of occurrence in the references. We show the particular relevance of Free terms in this analysis. This may be connected to the emerging nature of the subfield of physics under consideration. This shed an interesting light on the respective importance of Index and Free terms, as entries of data bases, in particular in the case of rapidly evolving scientific domains.

Introduction

In recent years a large amount of work has been devoted to the development of specific techniques for helping strategic R & D management.[1-3] The primary goal of these investigations was to provide relevant criteria for decision making in technology-based commercial activity. Very soon it appeared that these techniques, mainly developed for economic purposes, and more specifically for improving industrial competitivity, could be applied to many fields of R & D, also outside the commercial sphere. This is particularly true in the field of public scientific research, where such studies are often connected to the central and recurrent question of quality assessment, very often for funding purposes. The bibliometric tool here appears as a crucial step in the process of analysing a research activity.[4] Indeed,

0138-9130/95/US $ 9.50

scientific information still remains widely spread by means of international journals, in which articles hence appear as unavoidable indicators of a research activity. Note that the term "indicator" does not necessarily reflect any quality assessment. It may as well allow to identify research activities, without hierarchical classification. Altogether scientific "papers" remain one of the main accesses to science, whatever reason one may have to access it. This is actually recognized as well by scientific institutions (of course with variations from one country to another one): funding research projects as well as career evolution still rely at various degrees on publication lists.

One might argue here that with the worldwide dissemination of computer networks "hot" scientific material is spread much more quickly than by means of actual paper journals, not mentioning more media-oriented means used in the last few years. While it is clear that a bunch of information is indeed transmitted through these networks, it remains even clearer that publishing in a widely recognized journal represents an unavoidable step in scientific production and recognition. This may be due to the role of referees (which are for the time beeing absent from computer networks), but also to more psychological dimensions such as the "eternal" status of printed material. Altogether, it is hence clear that paper journals appear as particularly relevant indicators of a research activity, and bibliometric techniques may bring here useful tools of analysis of the bunch of produced material.

Once the importance of the bibliometric tool has been recognized it is worth mentioning some of its characteristics. First let us recall the growing importance of data bases. They allow both the storage of enormous amounts of information *and* a standardized access to this information. Such standards are for example the fixed sequences of data selected to characterize a publication: title, authors, affiliation, abstracts, keywords, ... A high level of standardization in the storage of data makes a computer assisted analysis relatively easy (we shall discuss this aspect in more detail below), which in turn allows to treat large amounts of data and hence to really exploit as much information as possible. This optimistic view should however not hide the underlying difficulties. Data bases and computer assisted analysis of these data have reshaped our access to scientific information but they have in turn generated new sorts of difficulties, precisely connected to the amount of processed data. This is where strategic access to the data becomes relevant. In an industrial context it is motivated by productivity. In a scientific research environment it is somewhat motivated by the access to the most relevant pieces of information. This latter aspect deserves some comments. One may expect that a researcher, well aware of his

research field, should more or less know what relevant researches are indeed performed in his field of expertise. Such an expertise can however hardly become cross-disciplinary. This is often simply due to the lack of access to the proper information, say for example to the finite number of journals a local library may actually buy. As long as a specific research does not require cross-disciplinary sources, the scientific activity may proceed in a quite "linear way", and with little help of large scale bibliometric studies. This is however not always the case.

Object

In the actual example we shall consider below, the original motivation for the bibliometric study has been a cross-disciplinary oriented program developed by a group of french physicists, hence requiring help of a large scale bibliometric survey. It should be noted that the actual starting of this cross-disciplinary program requires an acceptance and a funding by the french central research agency CNRS (Centre National de la Recherche Scientifique). CNRS is a national agency to which most french research laboratories are connected. It is structured in scientific departments and subdepartments, corresponding to the various fields of research. Apart from running and funding laboratories CNRS also supports scientific disciplinary or cross-disciplinary projects like the so-called GdR's (Groupement de Recherche) involving several laboratories all over France.

Before entering the details of the project under concern here let us make a final comment on the bibliometric study presented below. The analysis we have performed concerns a research activity in fundamental physics. But the motivation is little connected to a quality assessment of the scientific activity, which makes the interpretation easier, for obvious reasons. The bibliometric survey aims here at displaying a concise view of a cross-disciplinary field for the group of scientists working on the project. At this stage of the project there is hence no quality assessment underlying the study, even if in a later stage the project may be evaluated for funding. The distinction is important, in particular in a french context where evaluation is often confused with quality assessment. In a sense the test case considered here hence somewhat appears as a textbook case in view of the pure academic motivation.

The idea of the cross-disciplinary project (GdR) presently prepared is to gather a community of physicists around the general topic of "Fragmentation". The main goal of the bibliometric study associated to this project is hence to help to identify the

groups working in fragmentation, as fragmentation does not refer to any specific subfield of physics. On the contrary it is well spread over several branches of physics ranging from microscopic (nuclear, particles, cluster physics, ...) to macroscopic (rocks breaking, ...) and even to the large scale structures of galaxies in the universe. This widespread concept in turn suffers from a lack of accurate definition and even sometimes from recognition. This makes a bibliographic approach very difficult because of the non-existence of well defined keywords even in each subfield of physics. The word "fragmentation" for example does not even exists in the PACS[5] (Physics and Astronomy Classification Scheme, edited by the American Physical Society) or alternatively in the INSPEC[6] data base we have used, in spite of the well recognized status of this data base in fundamental physics. This lack of standardization in fact reflects the emerging nature of this topic. In many subfields of physics like nuclear or cluster physics, fragmentation is a young topic. In that respect the PACS or the INSPEC data base have acknowledged this evolution very lately, by introducing the keyword "nuclear multifragmentation" in the 1992 version of the index terms.[a)] But the latter term is relevant only for nuclear physics. On the other hand, subfields as astrophysics have identified fragmentation patterns for long, although it may have only been recognized as such in recent years. One may hence wonder where the emerging nature of the concept of fragmentation is? It lies in the cross-disciplinary character of fragmentation, character which has only been recognized very lately. The project of the physicists is actually to help this recognition. It is hence not so surprising that standard physics data bases have not yet integrated such a recent recognition (or even definition). Performing a bibliometric study in this context hence appears as particularly relevant, although difficult. It is interesting both for physicists who may not have an easy access to some information, and for pure bibliometric aspects as one has to deal with a material at the real boarder of an identified and structured scientific knowledge.

We have divided the complete bibliometric study into two parts. The first part deals with the real cross-linked subject, restricted to France as a geographical area (with 2267 processed references). The results of this "french" investigation have been published in France.[7] In the present article we discuss the second part of this investigation, namely concerning a more restricted scientific area (fragmentation in the subfield of nuclear physics) but with a worldwide coverage. This second study has been performed over 1994 bibliographic references. There are several reasons for the choice of a restricted scientific domain in this second investigation, both practical and theoretical.

First, our computer facilities would not have allowed us to treat fragmentation on a worldwide basis. This actually might have not been very relevant. In order not to ignore relevant activities in fragmentation we were led to perform a relatively open nationwide investigation, hence leading to a large number of publications and in turn to a sizeable noise, noise reduction being as a second step performed by hand by experts. To proceed in the same way worldwide would probably have led to an untractable set of data, hence of little use. This brings us naturally to the idea of specifying a subfield and hence to our second argument. Considering a specific subfield allows a more accurate investigation, while preserving the interest of a worldwide versus a nationwide investigation. It would also allow to assess the french activity in this field as compared to the international one.

The choice of nuclear physics as a subfield is then natural. First, as we already mentioned it, it is the first (and single) subfield of physics in which a PACS or INSPEC entry has been specifically created for fragmentation. Second, the nuclear physics community is a relatively well structured physics community, and hence relatively easy to study. This structuration is due to historic reasons but also to the fact that nuclear experiments devoted to fragmentation require sophisticated and dedicated detection devices and experimental facilities. The groups are hence big and often geographically peaked. Such a structure is reflected by the bibliometric analysis. Usually physicists working in one of these big groups have only *one* research activity, the one of the group. This eliminates spurious connexions to other activities, connexions which may appear when a researcher, working in less heavy structures, is involved in several, possibly different, research activities. The existence of big groups also means that each group is associated to a well defined, homogeneous set of keywords, with little noise from other subfields of physics. The drawback of this structure in big groups is the actual identification of the groups. As we shall see below the bibliometric study often leads to huge, often "artificial", groups, which have to be splitted into several smaller, "relevant" ones. But the latter point may be accommodated by a proper cutoff procedure.

The INSPEC data base for studying fragmentation in nuclear physics

It is well known that one of the difficulties of a bibliographic analysis lies in the way one constitutes the set of data to be processed.[8] This of course does not mean that the automatized processing of the data cannot in turn introduce bias; we shall however discuss this aspect only later on. Let us for the time being focus on the raw

material, namely the set of publications selected for the automatized analysis. As suggested above the constitution of this set is particularly not trivial in the case we consider in this study, mainly because of the emerging nature of the concept of fragmentation. At this stage the major points to be discussed are the choice of the data base to work on, the relevant words used for selecting the publications and the choice of material from the publication (authors, affiliations, keywords, ...) to be processed. Let us briefly discuss these three aspects and evaluate the possible bias we introduced here in our investigation.

Choice of the data base

As already stressed, the topic of our investigation is strongly focused on fundamental research in physics. In this field, ***Physical Abstract*** offers a wide access to the field with more than **4000** journals, including the most famous ones. The term famous should deserve a lot of comments. Without entering the details here, let us mention that the major journals used (and hence recognized) by nuclear physicists working in the field of the investigation, do appear in ***Physical Abstract.*** From the point of view of the "experts" this is hence a reasonable choice. From the bibliometric point of view this is also a relevant choice. The number of journals in the base confirms the all-day experience of physicists. Furthermore, it allows to work on one data base only. This has the important advantage of offering a once-for-ever standardized information for all the publications. This obviously makes the automatic processing much easier. But the interest is not purely technical. The fact that the set of data is standardized should allow a relative homogeneity. One may even hope for instance that the way Index Terms are attributed to papers is to a large extent homogeneous, up to the unavoidable fluctuations due to human factors. Our investigation has been based on the INSPEC data base of Physical Abstract. It is easily accessible from the Central Library of our University in Toulouse, since **1990.** Our investigation has hence been extended over the last **4** years, from **1990** to **1993.** This imposed choice of the period should not introduce too many bias. We outlined above the relatively recent appearance of the field of fragmentation in nuclear physics. In particular, the period between **1990** and **1994** covers the first recognition of "nuclear multifragmentation" by INSPEC, hence possibly allowing, at a more advanced stage of the investigation, a study of the effect of the introduction of this term as an Index Term. Finally, an investigation over these **4** years provided about **2000** publications, which is a quite reasonable number for an automatized processing.

A much larger amount of data would have generated computational difficulties, a much smaller one would have strongly enhanced statistical bias and hence lead to questionable conclusions.

Keywords for characterizing nuclear fragmentation

Before presenting explicitely the way we constituted our set of data let us comment on the possible keywords to be included for defining it. As already mentioned (multi)fragmentation is a young topic in nuclear physics. It is well known that formulating and using specific words is a crucial step in the process of building knowledge. But this may take several years. Normalization comes later, once the topic has been properly characterized (and recognized) by a growing community. At the bibliographic level this process is reflected by the evolution of the keywords entries of data bases. For data collection in an area where the keywords are not yet well established, it is very important to have several discussions with experts in the studied area.[b] Then, one must control the opportunity of the keywords with data base experiments to ensure that no noise appear. Our data collection strategy is partly related in Table 1. We shall discuss some of its properties later on.

Table 1. a
Example of the information kept for characterizing a typical bibliographical reference after the first automatized treatment of the data

-1594-	
AU	- Bozek; Ploszajczak
OS	- GANIL, Caen, France
IT	- Heavy ion-nucleus reactions; Nuclear fragmentation; Nuclear reaction and scattering theory
FT	- Particle production; Nuclear multifragmentation; Relativistic heavy-ion collisions; Projectile dependence; Source-size dependence; Spatio-temporal intermittency

Table 1. b
Constitution of a typical group. This group contains three authors. The frequency of appearance of each author (namely the number of references in which it belongs to the author list (entry AU of Table 1.a) is indicated in the first row

Freq.	Author
6	Ploszajczak
6	Bozek
2	Tucholski

Table 1. c

IT's associated to the example group of Table 1.b. In the first row the frequency of occurrence of a given IT is given, for each author. For this group 4 different IT's were retained, which characterize the activity of this group

Freq.	Index Term (IT)	Author
3	Statistical theory of nuclear reactions and scattering	Bozek
5	Heavy ion-nucleus reactions	Bozek
2	Fluctuations	Ploszajczak
2	Nuclear fragmentation	Ploszajczak
2	Statistical theory of nuclear reactions and scattering	Ploszajczak
4	Heavy ion-nucleus reactions	Ploszajczak
2	Heavy ion-nucleus reactions	Tucholski
2	Statistical theory of nuclear reactions and scattering	Tucholski

Table 1. d

FT's associated to the example group of Table 1.b. In the first row the frequency of occurrence of a given FT is given, for each author. For this group 3 different FT's were retained

Freq.	Free Term (FT)	Author
3	Nuclear multifragmentation	Bozek
4	Scaled factorial moments	Bozek
2	Fluctuations	Ploszajczak
4	Nuclear multifragmentation	Ploszajczak
4	Scaled factorial moments	Ploszajczak
2	Nuclear multifragmentation	Tucholski
2	Scaled factorial moments	Tucholski

We are now in a better position to discuss the way to constitute the set of data to be processed. In order to make our strategy clearer it is worth reminding the content of the INSPEC data base. The INSPEC data base offers several items on a given publication: title (TI), authors (AU), affiliation of the first author (OS), journal (CO), abstract (AB), index terms (IT), classification codes (CC) and uncontrolled or free terms (FT). It is obvious that IT and CC entries offer a relatively rigid (but highly standardized) access to the document. On the other hand TI, AB and FT entries are much more evolutionary and might be more adapted to an emerging concept. The drawback is of course the larger diversity in the terms used, diversity which hinders standardized analysis and which might altogether generate a sizeable

noise. Note however that in the case we consider here this latter suspicion might not be well founded as we shall see below.

In our access to the raw data we processed in a mixed way, considering on the same footing the appearance of relevant terms in any of the entries (AB, TI, IT and FT) and also filtering the data with CC's. As a final precaution we proceeded in a multistep way. We started with a set of a priori relevant keywords. In a first exploration of the data base we selected a few "characteristic" references which we examined together with the experts. This allowed to modify our set of keywords. We iterated this procedure up to convergence, namely up to a set of keywords generating a stable set of data. This analysis also confirmed our suspicion according to which controlled indexation (IT's, in particular) often remains very general and little specific, in a rapidly evolving situation such as the one we were considering. Relevant information in fact came from uncontrolled entries (TI, AB and FT's). Once the set of keywords had been defined we finally filtered the data to be processed by the relevant CC's, which allowed to focus the investigation on a small subfield of nuclear physics. With this procedure we constituted a set of 1994 references to be processed automatically.

Automatized treatment of data

First level treatment of raw data

As already stressed above the fact that we used only one data base allows the constitution of a highly homogeneous set of data, at least from the point of view of an automatized treatment, which is a welcome feature. INSPEC items are numerous, but well presented and always in the same sequence. As a first step we selected some of the entries and created a new set of "reduced" data in which we also took care of double counting. In these reduced data we kept only AU, OS, IT and FT entries, in a properly normalized way (see Table 1.a). This first step of the analysis is very important. It leads to a perfectly homogeneous set of data, which in turn can safely be treated at an automatized level.

Data processing was performed at CRRM (Centre de Recherches Rétrospectives de Marseille). For this purpose we used the softwares developed by the group of H. Dou in Marseille. More precisely this first step of the analysis was performed with INFOTRANS.[9]

Method of analysis: group identification

Let us now discuss our method of analysis of the data. Remember that the primary goal of the physicists was to have an identification of the various groups working in the field of fragmentation, mostly on a nationwide but cross-disciplinary basis. This will be of importance for the choice of the method of analysis. In the investigation of multifragmentation in nuclear physics the goal remains essentially the same, namely to identify the groups working in fragmentation, but in the subfield of nuclear physics. We have chosen to identify groups by linking people who effectively worked together. We hence worked on the author entry (AU) of the INSPEC data base in order to constitute the groups. All the group constitution we discuss here was performed automatically with the DATAVIEW software.[10] This choice of constituting groups by linking authors obviously maps the preoccupation of physicists, but is also quite interesting from the bibliometric point of view. Let us be slightly more specific.

The idea is to constitute groups of authors having published jointly. Groups are in turn formed by successive links. The key quantity is obviously the number of joint publications above which two persons are considered to be linked. In the nationwide inquiry we fixed this threshold at 2, which allowed a nice identification of most of the groups working in fragmentation ... except in nuclear physics. Most of nuclear physicists were gathered inside a non significant huge group of more than 400 persons. The reason for the occurrence of this oversized group is however relatively simple to figure out. As we outlined above, nuclear physics is a highly structured community, where people are gathered in often large experimental groups of typically 10 to 20 persons. Furthermore, experiments require national or international facilities where physicists come to perform their experiments. Occasional collaborations stem from contacts between the physicists belonging to the facility as staff members and visiting scientists coming to perform their own experiment. This results in the constitution of a huge community around the facility, overlooking true, long-living collaborations between individuals. One has hence to fix a much higher threshold in this case. In the nuclear physics part of the nationwide inquiry we processed by hand the oversized group and fixed a threshold at 6. This means that we defined groups of persons having published at least 6 times (by pairs) over the last 4 years. In the worldwide examination under study here, we again fixed the threshold at 6, which allowed to identify several groups.

One should however remain cautious with the interpretation of the results in view of the crucial role played by the threshold in the actual identification of the groups. The argument we just developed presumably essentially holds for physics in a stationary regime. If we consider a well established activity one may expect that people will publish several (at least more than 1) papers by year. If a collaboration is indeed well established between several people one may expect that this will be all the more true. However, such a stationary situation is not the single possible one to occur. Remember that we consider an emerging topic. One may hence expect new people to enter the field and to need some time before effectively producing some material to be published. This is true for theoreticians but even more for experimentalists. We mentioned the huge detectors needed to perform experiments on nuclear multifragmentation. The building of such detectors may take several years during which people involved will effectively publish much less than expected. Altogether, one should hence consider with caution the details of the analysis in groups we present below. With these restrictions in mind, let us now discuss in more detail the information gathered around identified groups.

Characterization of group activity

To each identified group of authors one can associate the corresponding entries of INSPEC kept in our reduced data set, namely IT's and FT's. This allows to characterize the activity of each group by a set of keywords. A typical group is presented in Table 1 where we give an example of the pieces of information used for characterizing a reference (Table 1.a) and the group itself (Table 1.b, c, d). This group is constituted by 3 physicists linked by the abovedescribed procedure. It is associated to 6 bibliographical references out of which 4 IT's and 3 FT's have been retained for characterizing the activity of the group. The retained IT's or FT's are the ones which appear at least twice in the IT's and FT's of all the bibliographical references associated to this group. This threshold allows to exhibit the dominant activity of the group.

This procedure of associating keywords to a group is also interesting from the methodological point of view. The linking procedure used for constituting groups is completely independant of IT's and FT's entries. The underlying assumption for the constitution of a group is only to consider that 2 authors having a common interest and a recognized collaboration should have published together a certain number of times during the last four years. This by no means implies anything on the topic or

the characteristics of their joined work. But once the group has been constituted the analysis of the associated IT's or FT's simply defines the field in which the group is working. An a contrario proof of the reliability of such an approach is the case of homonymy, which might lead to the constitution of a fake group.[c] Index and Free Terms may then be used as indicators for the true existence of this group. In general it is easy for experts to identify the inhomogeneous set of IT's and FT's generated in such cases. In the study we consider here, note that the fact that neither IT's nor FT's were used to constitute the groups allow them to be used as relevant criteria of the reliability of the classification in groups. It should finally be noted that our experience, in the course of this investigation, is that such polluted groups were effectively very few, which in turn validated our method for constituting the groups.

It is nearly impossible, for automatic bibliometric treatment out of downloaded data, to perform a perfect analysis. Automatic bibliometric treatments are reading guides for experts in a studied area, but should never be considered as a substitution of the experts, aiming at an "objective truth".

Analysis of results: IT's and FT's

On identified groups

A good indicator of the quality of our computerized treatment is precisely the homogeneity (once depolluted) of the IT's associated to each of the identified groups. This high degree of homogeneity helps a lot in the further stages of the analysis. It allows to identify very easily the particular subfield of physics the group is working on. In turn this makes the work of selection of the relevant groups very simple. Remember that we constituted our primary set of data in a relatively open way, in order not to eliminate possible interesting groups, at the boarder of the field. The latter groups might actually play a particularly important role in the case of an emerging topic, as the one we are working on. As a consequence of this non restricted selection of data we expected a relatively large number of groups not relevant for (multi)fragmentation. This is indeed what we observed. The large fraction of irrelevant groups however does not only reflect our data collection strategy. We mentioned the fact that the word fragmentation and other peripheral terms are used in nuclear physics in contexts which are not relevant for our purpose [see Notes and Comments b)]. The identification of irrelevant groups is hence to a large extent unavoidable and these groups have indeed to be eliminated by hand by the experts. The IT + FT's entries allowed here, as we explained, a quick and safe selection of relevant groups.

Distributions of IT's and FT's

The analysis of the distribution of IT's and FT's is very interesting from the bibliometric point of view.[11, 12] From our 1994 references we identified 266 IT's and 2345 FT's. The distribution of IT's and FT's are presented in Figs 1 and 2 as a function of the number of occurrences (frequencies). We shall discuss in detail below the thresholds we chose for a relevant interpretation of these two curves. Both curves present the standard shape of Zipf law,[13] that one expects. Nevertheless it is interesting to look at them in some more detail.

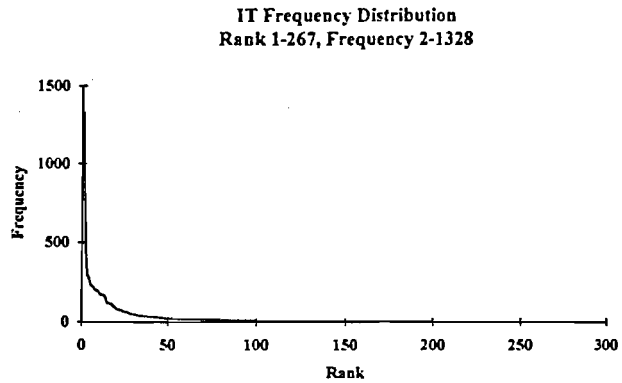

Fig. 1. Frequency curve of Index Terms (IT's). The IT's have been sorted by decreasing frequency. The resulting curve has a standart Zipf shape

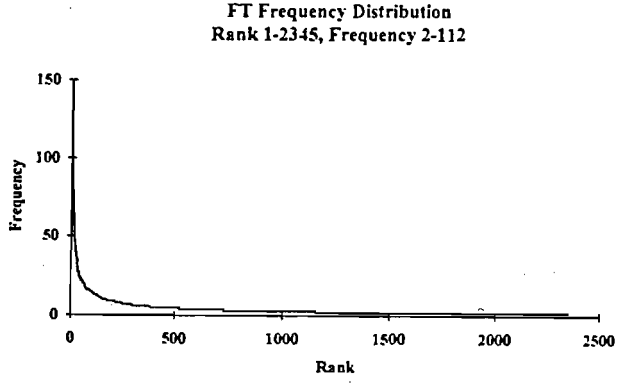

Fig. 2. Same as Fig. 1 for Free Terms (FT's)

Let us as a first step focus on the first 50 entries and magnify this part of the curves (Fig. 3). We see here that the IT curve is clearly "diverging" at origin. The first entry appeared 1328 times in the 1994 references we worked on, while the second one occurs only 463 times. On the contrary the FT curve is not truly "diverging" at origin (112 occurrences for entry number 1, against 93 for entry number 2).[d] These differences may look obvious at first sight. Indeed, due to the lack of IT entries specific to nuclear multifragmentation, IT entries only reflect very general characteristics of the work, whence the divergence. These trivial entries are used so many times that, to some extend, they lose sense, at least at the level of identification of a so to say "subsubfield" of nuclear physics, we are interested in. On the other hand FT's are much more numerous and well spread over the references. One may hence expect, as observed, that the peak value of FT's is much smaller that the one of IT's. It is indeed more than an order of magnitude smaller than the IT peak value. This is however not the most interesting aspect of this comparison. The relevant point comes from the comparison of the shapes of the curves, which are very different. The FT curve is much smoother than the IT one. This a priori indicates that there exists a sizeable set of FT's which are used with comparable frequencies, for characterizing this field of physics. This is probably where we can learn something on the relevant terms used in nuclear multifragmentation.

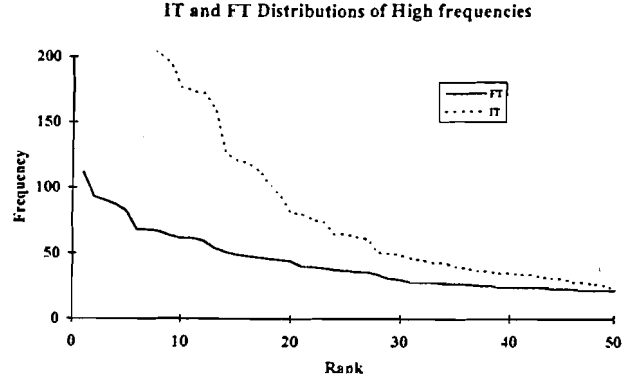

Fig. 3. Comparison of the high frequency part of IT's and FT's curve. One can note from this magnification the different shapes of the curves

In order to investigate this possibility in some more detail let us consider explicitely the first FT's of the curve. The first 20 FT's with highest frequencies

(frequencies between 112 and 44) have been reported in Table 2.a. The striking feature of Table 2.a is the fact that to a large extend these first FT's are highly significant. Let us mention for example the term "Multifragmentation" or "Intermediate Mass Fragments" which occur respectively with frequencies 67 and 76, to be compared to the peak value of 112. With such frequencies they respectively take the 9th and 6th positions. The occurrence of these terms with such levels of frequencies call for several comments. First, these occurrences confirm the relevance of FT's in such a study. This is probably due to a large extend to the fact that Index Terms do not yet (except for 1993) include so specific terms. Second, it is interesting to note that the term "Intermediate Mass Fragments", which characterizes the fragments obtained in the course of multifragmentation, appears more frequently than "Multifragmentation" itself. This reflects the fact that the recognition of multifragmentation, even inside the nuclear physics community, is recent, while its manifestations were identified before its actual recognition.

Table 2. a

List of the first 20 FT's extracted from the 1994 references treated in our study. The "Trivial/Interesting" threshold has been fixed here between frequencies 54 and 60, which corresponds to ranks 12 and 13

Rank	Freq.	Free Terms
1	112	Fission fragments
2	93	Heavy-ion collisions
3	90	Fragmentation
4	87	Multiplicity
5	82	Angular distributions
6	76	Intermediate Mass Fragments
7	68	Fragments
8	68	Central collisions
9	67	Multifragmentation
10	64	Nuclear matter
11	62	Fluctuations
12	60	Multiplicity distributions
13	54	Cross sections
14	51	Nucleus-nucleus collisions
15	49	Intermittency
16	49	Heavy-ion collisions
17	48	Excitation energy
18	46	Multiplicities
19	45	(P,X)
20	44	Fission

Table 2. b
List the first 20 IT's extracted from the 1994 references treated in our study. The "Trivial/Interesting" threshold has been fixed here between frequencies 94 and 102, which corresponds to ranks 18 and 19

Rank	Freq.	Index Terms
1	1328	Heavy-ion nucleus reactions
2	463	Stat. theory of nuclear reactions and scattering
3	300	Nuclear reactions and scattering theory
4	277	Nuclei with mass number 190 to 219
5	231	Nuclei with mass number 220 or higher
6	226	Nuclear matter
7	207	Nuclei with mass number 20 to 38
8	203	Fission
9	195	Nuclei with mass number 6 to 19
10	177	Nuclei with mass number 90 to 149
11	174	Proton-nucleus reactions
12	172	Nuclear fragmentation
13	161	Direct nuclear reactions and scattering
14	127	Nuclei with mass number 59 to 89
15	121	Fission products
16	119	Nuclei with mass number 39 to 58
17	113	Nuclear fusion
18	102	Nuclei with mass number 150 to 189
19	94	Fission of uranium
20	81	Fluctuations

Of course a sizeable fraction of the first FT entries bring little information, as for example "Fission fragments" (rank 1, which concerns a somewhat different field of physics and is hence not truly related to multifragmentation), or as "Fragments" (rank 7) or "Multiplicity" (rank 4) because of their lack of specificity. But altogether these entries represent a much less trivial set of words as compared to IT's. For comparison we have reported in Table 2.b the 20 first IT's with highest frequencies (frequencies between 1328 and 81). It is striking, from expert's remarks, that these high frequency IT's bear very little specific significance for the field of physics under study. This can be illustrated by several general terms such as "Nuclear Matter" (rank 6) or "Fission" (rank 8). A remarkable feature is also the recurrent appearance of the IT's "Nuclei with mass number... to...". There exist 10 such IT's which cover the whole set of nuclei studied by nuclear physicists. Out of these 10 entries 8 appear in the first 20 IT's of our study, which confirms the relatively poor significance of these high

frequency IT's. The single exception to this general trend is the entry number 12, "Nuclear fragmentation", which could appear as quite significant in our inquiry. However, one should keep in mind the difficulties connected with the use of this term in nuclear physics (see the above discussion in footnote b).

The general poor significance of high frequency terms is not surprising. It is actually to some extend a side product of the characteristic shape of the IT curve. The important point here is hence not the behavior of the IT curve, which is quite "normal" but rather the fact that the IT curve strongly differs from the FT one where high frequency entries may be highly significant. The two FT terms "Intermediate Mass Fragment" and "Multifragmentation" discussed above are actually recognized by experts as the most significant terms for characterising this field of physics.

Noise-Interesting-Trivial (N-I-T) analysis of IT's and FT's frequency distributions

In order to complete this discussion we have further analyzed our IT and FT curves by means of a "Noise-Interesting-Trivial" (N-I-T) splitting of the curve. The high frequency entries of each of the FT or IT curves are considered as trivial (hence bringing little significance) while the very low frequency entries are considered as generating noise (hence not relevant). In fact noisy keywords may be interesting in terms of innovation, as they may reflect an emerging field. The drawback is that they cannot be considered without a proper selection by the experts of the studied area. This leads us to leave them for a further automatic analysis.

Thresholds between "Noise" and "Interesting" and between "Interesting" and "Trivial" have to be fixed arbitrarily by considering the shape of the frequency curves and also with the help of the experts of the studied area, who considered the significance of the keywords explicitely. The thresholds have been assigned to significant decreases in the curve, from one item to the following. For the IT curve the thresholds are respectively between frequencies 94–102 (which corresponds to ranks 18–19) for the "Trivial" to "Interesting" transition and between frequencies 9–10 (which corresponds to ranks 100–101) for the "Interesting" to "Noise" transition. In the case of FT's the thresholds have been chosen between frequencies 54–60 (which corresponds to ranks 12–13) for the "Trivial" to "Interesting" transition and between frequencies 5–6 (which corresponds to ranks 374–375) for the "Interesting" to "Noise" transition.

Before discussing further the results of this N–I–T analysis a word of caution is necessary. Remember that we noted significant differences between the most frequent IT's and FT's. In particular we identified 2 highly significant FT's among the first 10 most frequent ones. In the present N–I–T separation these 2 terms will be considered as Trivial which may be questioned from a non statistical point of view. It will be important to keep in mind this restriction in the interpretation of the following discussions.

The N–I–T cloud

An interesting outcome of this further gross classification of IT's or FT's by means of the N–I–T filter, is that one can attribute to each IT or FT a simple characteristic that we shall abbreviate by N for Noise, I for Interesting and T for Trivial. One may then characterize each reference, in which appear a given number of IT's and FT's according to its "coordinates" along N, I, and T axis. Normalization of the coordinates is performed according to the *maximum* total number of IT's (or FT's) appearing in any of the references. For example the maximum number of IT is 18; if a reference contains 6 IT's out of which 3 are "Trivial", 2 "Interesting" and 1 "Noise" the "coordinates" of this reference are respectively 3/18 along T axis, 2/18 along I axis and 1/18 along N axis. Such a coordinate representation of the ensemble of references allows a 3–D representation of this ensemble according to a reference frame constituted by the N, I and T axis, for Index and Free terms. In these IT and FT spaces each point represents one bibliographic reference. The position and the shape of the distribution of references shed a new light on the significance of IT's and FT's. However, because of the finite number of IT's and FT's, N, I and T coordinates also take a finite (and restricted) number of values so that a 3-dimensional representation of the N–I–T cloud brings little information. In other words this representation is not able to reflect the weights of the points of the cloud. In order to recover the latter feature it appears more relevant to consider histograms of the coordinates.

The histogram for IT's and FT's coordinates along the 3 N, I and T axis are presented in Figs 4 and 5, respectively. Once again these two figures exhibit striking differences between IT's and FT's. This observation may be further quantified by evaluating the average values of the distribution along each axis. The results of these calculations are reported in captions of Figs 4 and 5 respectively for IT's and FT's. The IT cloud exhibits a strong trivial component as expected from the previous

analysis of the IT curve. This is reflected by a high average for T (0.1308) as compared to the ones for I (0.0638) and N (0.0244). On the contrary the FT cloud is dominated by Noise (average 0.1818), which is not surprising as many FT's have a marginal significance for the topic. The average along I (0.0547) in turn is larger than the one along T (0.0115) which agains reflects the importance of FT's. Altogether these results nicely complement the analysis of the IT and FT curves. They once again reflect the relevance of FT's in this investigation.

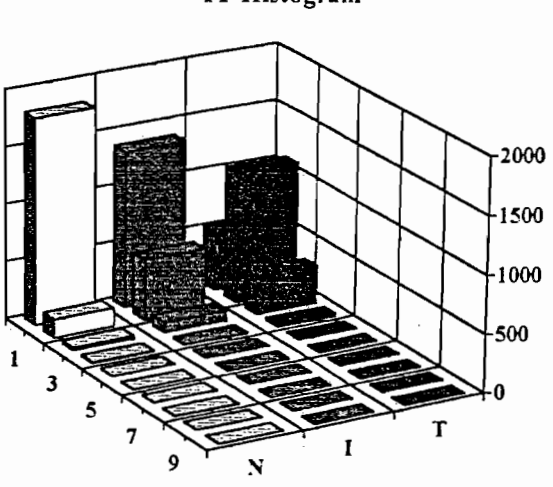

Fig. 4. Histograms of Index Terms (IT) coordinates along the three axis defined in the text "Noise" axis (N) associated to frequencies between 1 and 9. "Interesting" axis (I) associated to frequencies between 10 and 94 and "Trivial" axis (T) associated to frequencies between 102 and 1328. The coordinates are between 0 and 1 by definition (see text) and are binned in bins of 0.1 length. First entry hence corresponds to bin 0.0 to 0.1 (bin labelled "1") second bin to 0.1 to 0.2, a. s. o. ... The average values of the histograms along the three axis are respectively 0.0244 for Noise, 0.0638 for Interesting and 0.1308 for Trivial. The numbers of exactly zero values are respectively 1503 for Trivial, 746 for Interesting and 105 for Trivial

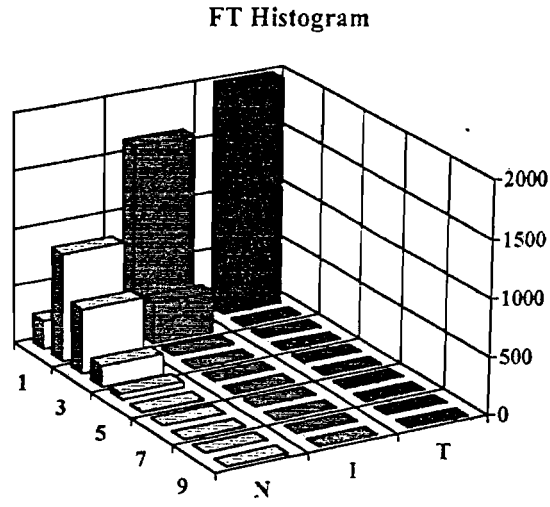

Fig. 5. Histograms of Free Terms (FT) coordinates along the three axis defined in the text "Noise" axis (N) associated to frequencies between 1 and 5, "Interesting" axis (I) associated to frequencies between 6 and 54 and "Trivial" axis (T) associated to frequencies between 60 and 112. The coordinates are between 0 and 1 by definition (see text) and are binned as in Figure 4. The average values of the histograms along the three axis are respectively 0.1818 for Noise, 0.0547 for Interesting and 0.0115 for Trivial. The numbers of exactly zero values are respectively 6 for Trivial, 293 for Interesting and 1247 for Trivial

It should finally be noted that this kind of N–I–T representations are useful in order to classify the bibliographic references among the indexing policy of each field (this remarks concerns controlled (IT's) as well as uncontrolled (FT's) keywords). But they are also very important to give a first reading guide to the experts of the studied area: marginal articles presumably are essentially "noisy"; review-like papers will presumably contain mainly "trivial" terms; more "mixed" cases might in turn be considered as subfield indicators.

Conclusions

In this text we have presented the results of a large scale bibliometric survey applied to fundamental research in physics. The investigation was performed in

connection with the onset of a cross-disciplinary project in physics, whose aim is to gather a wide community of physicists around the topic of fragmentation. This work constituted a strategic help to this project by allowing an identification of the relevant groups in the field. From the bibliometric point of view a major interest of the investigation lies in the fact that the topic under study is only emerging as a recognized field in physics. This means that standardization, unavoidable in the process of constituting data bases, is still in an evolving stage. We hence had to rely strongly on low standardized entries, in particular on the numerous Free Terms associated to the references we worked on.

The study of these FT's and a comparison with the IT's is very instructive. We have seen that the frequency curves of IT's and FT's exhibit different patterns. Furthermore, while high frequency IT entries bring little significance, some high frequency FT's may be very relevant for the investigation. This presumably reflects the fact that the emerging concept we have been working on is currently being recognized. A few relevant terms are associated to it and appear frequently as FT's before a recognition as IT's as it was the case for one of them in 1992. At the level of our investigation we however did not identify any sizeable effect due to the recognition of this relevant keyword as an IT. This may be connected to the fact that only the 1993 references may benefit from this recognition, so that the statistical weight would at most be of order 25% as compared to the whole set of references.

We have further analysed our set of references according to a separation of IT's or FT's between Noise, Interesting and Trivial. One has however to be cautious here with the true meaning to be given to this classification. It depends on the chosen thresholds and presumably has a different meaning for IT's and FT's as discussed in the text. With these restrictions in mind one can nevertheless represent each reference as a point in a three dimensional N–I–T space and study the distribution of coordinates of the references in this space. This provides a new representation of the set of references we worked on. This analysis complements the study of frequency curves and again reflects the importance of non standardized entries in the investigation of an emerging concept. This kind of N–I–T study may also be performed with a selection by experts of the field under study. But this induces a less innovative point of view, which reflects the point of view of experts and hence may be sometimes dangerous, especially when one considers an emerging topic.

Notes and Comments

a) It is clear that a data base can only slow down the actual recognition of an emerging concept. For structural reasons a data base cannot accommodate any "new" keyword as soon as it appears: the number of entries of the data base would almost immediately diverge! But above all, increasing the number of entries in an uncontrolled way simply amounts to annihilate any classification scheme, while classification schemes represent the core of any data base. We come back here to the literature we mentioned in the introduction of this text. Papers disseminated by computer networks or even in marginal journals without referees may, in the context of an emerging concept, contain much more relevant information than more "standardized" literature, in part also for reasons connected to the publication delay. A large scale bibliometric investigation based on such a non standardized material is hardly conceivable. Only qualitative investigations might make sense in this case.

b) In order to avoid confusion at the level of relevant terms describing fragmentation in nuclear physics one should actually be a little bit more specific. For several years the term "fragmentation" has been used in nuclear physics in a context relatively different from the one under study here, which has hence to be properly identified by experts. The recognition of the term "multifragmentation" however, does not solve all the problems. First, remember that it was introduced in 1992, only. It is hence missing for characterizing earlier publications. Second, in order not to restrict too much the field of investigation through too specific and too few keywords, and again because the concept we are studying is still emerging, we have included other words which may characterize the process of interest. Let us cite for example the term "fragment". While this term is not recognized as an entry in INSPEC index terms it is obviously relevant for our investigation, and widely used by authors, either in the abstracts of their papers or as free term characterizing a paper. But it is widely used by other communities of nuclear physicists working in disconnected areas. This example is of course not isolated. We have chosen to keep these "peripheral" terms in our investigation, mainly because of the emerging nature of the concept of multifragmentation in nuclear physics. In turn we hence produce a sizeable noise, which had to be reduced "manually" by the experts.

c) Let us consider the (relatively rare, in fact) case of two persons with the same "common" last name, say for example Mr. Smith. In our reduced set of data such an occurrence has a non vanishing probability as we only keep last names. Note by the way that keeping first names would only reduce this probability, not cancel it, and might also raise problems in the opposite direction, in the sense that initials are not always abbreviated in the same way, for a given person. Anyway, it may occur that *two* Mr. Smith are working in *two* different fields and hence have nothing to do with each other from the point of view of our investigation. They will of course be identified as only *one* Mr. Smith and the groups they are associated to will be merged into one big group.

d) If one compares the relative occurrences of entries number 2 as compared to entries number 1 for IT's and FT's the difference is even more striking. For IT's the ratio is of order 35% while for FT's it is of order 83%.

References

1. D. J. De Solla Price, D. Beaver, *American Psychologist*, 21 (1966) 1011–1018.
2. H. Haon, C. Paoli, H. Rostaing, Perception d'un programme de R & D à travers l'analyse bibliométrique des banques de données d'origine japonaise, *Actes du Congrès IDT93*, Juin 1993.
3. H. Dou, H. Hassanaly, L. Quoniam, Informations stratégiques en chimie. Analyse topologique automatique de la base Chemical Abstract, *Revue Française de Bibliométrie*, 7 (1990).
4. H. P. J. Peters, A. F. J. Van Raan, Structuring scientific activities by co-author analysis. An exercise on a university faculty level, *Scientometrics*, 20 (1991) 235–255.

5. PACS, Physics and Astronomy Classification Scheme, published yearly by The APS (American Physical Society) in *Physical Review Letter*, December 1993.
6. INSPEC, Institution of Electrical Engineers (IEE), M. Faraday House, Six Hills way, Steven age, HERTS SGI 2AY, United Kingdom.
7. M. G. SURAUD, L. QUONIAM, H. ROSTAING, H. DOU, Analyse bibliométrique appliquée au cas de l'émergence d'un concept en physique fondamentale, *Revue Française de Bibliométrie*, (1994).
8. C. DUTHEUIL, Du corpus documentaire à l'interprétation des résultats de l'analyse de données, *Revue Française de Bibliométrie*, 6 (1990).
9. I+K, Information und Kommunication, I+K France, 9 Avenue Ville Preux, 78340 Clayes sous Bois.
10. H. ROSTAING, *Veille technologique et bibliométrie: concepts, outils, applications*, Ph. D thesis, Université Aix-Marseille, January 1993.
11. G. WHEELER, Maintaining a controlled vocabulary for a large on line data base, *Computers in Libraries International 1991, Proceedings of the fifth Annual Conference on Computers in Libraries*, London, February 1991.
12. Y. EBINUMA, S. TAKAHASHI, S. HABARA, H. YOKOO, Promotion of keyword assignment to scientific literature by contributors, *International Forum on Information and Documentation*, (1983).
13. G. K. ZIPF, *Human Behaviour and the Principle of Least Effort*, Addison Wesley, 1949.