Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest

Scientometrics,
Vol. 35, No. 2 (1996) 223-235

COPING WITH THE PROBLEM OF SUBJECT CLASSIFICATION DIVERSITY

ISABEL GÓMEZ,* MARIA BORDONS,* M.T. FERNÁNDEZ,* AIDA MÉNDEZ**

**Centro de Información y Documentación Científica (CINDOC), CSIC, Joaquín Costa 22, E-28002 Madrid (Spain)*
***Institut d'Estudis Avançats de les Illes Balears, CSIC-UIB, E-07071 Palma de Mallorca (Spain)*

(Received June 22, 1995)

The delimitation of a research field in bibliometric studies presents the problem of the diversity of subject classifications used in the sources of input and output data. Classification of documents according to thematic codes or keywords is the most accurate method, mainly used in specialised bibliographic or patent databases. Classification of journals in disciplines presents lower specificity, and some shortcomings as the change over time of both journals and disciplines and the increasing interdisciplinarity of research. Differences in the criteria in which input and output data classifications are based obliges to aggregate data in order to match them. Standardization of subject classifications emerges as an important point in bibliometric studies in order to allow international comparisons, although flexibility is needed to meet the needs of local studies.

Introduction

Standards applied to bibliometric indicators are needed to make bibliometric research reproducible, more transparent for users and to overcome the incompatibilities between indicators produced by different institutions.[1]

Diversity in subject classifications can be pointed out as one more of the obstacles that difficult the comparability among studies. When facing the problem of delimitation of the field in a bibliometric study, a broad spectrum of subject classifications appears in front of us:

- Classification of documents according to thematic codes and/or key-words;
- Subject classification of journals (differences among databases, directories, catalogues);
- Classification of input data, as resources invested or research projects, attending to the UNESCO classification, priority research lines or socio-economic objectives, etc.

0138-9130/96/US $ 15.00

Is it possible to harmonize all these different classifications? Which one reflects the performance in a specific area better? Does any coincidence exist between the classificatory needs at the national and at the international level? For example, a study of the scientific performance of a country in Biomedicine can be done attending to its publications in biomedical journals as covered by multidisciplinary international or specialized databases, but also through the scientific output derived from biomedical research projects or from biomedical research centres. Looking for international comparability, the first approach, based on international databases, could be the best. Research projects and institutes not always follow the same classifications in all the countries, so it hinders comparisons. However, the second approach could be more convenient for internal evaluative purposes.

In this paper, different subfield delimitation approaches are shown, indicating their main advantages and shortcomings and illustrating every case with real examples from our experience in the development of bibliometric indicators.

Different subject delimitation approaches. Advantages and shortcomings

1. Delimitation of fields through database thematic classification of documents

Delimitation of the field through keywords or thematic codes used in databases could be the most appropriate solution when dealing with specific topics. Three different advantages are guaranteed by using a controlled language: flexibility in field delimitation, as the best words or codes to define the area of concern can be chosen; pertinentia of the final set of documents retrieved, as the indexation is made at the article level; and reproducibility of the study, as it can be easily performed again and by others. This possibility is usually offered by specialized bibliographic databases that index at the article level: Chemical Abstracts, INSPEC, BIOSIS, Medline, etc., as well as by the patent databases.

2. Delimitation of fields through journal classification

Some multidisciplinary databases offer a classification of journals into subfields. Delimitation of a field attending to the "journal level" cannot be as accurate as the one performed at the "article level". It is well known that most journals contain articles dealing with a relatively broad range of themes, in spite of their "main subject". Thus, a subject delimitation based on journal classification will probably contain some articles weakly related with the target subject, while some pertinent articles will be missing. The search according to a fixed set of journals will be characterized by a lower degree of flexibility and lower pertinentia of the results. In spite of this, the strategy is easier to be established and it is broadly used in many bibliometric studies. The SCI database is frequently chosen because it selects "main stream" science in all fields, facilitates comparisons through fields and countries, and allows studies on cooperation, as it records all addresses of the authors.

When comparing both methodologies: classification of journals versus thematic codes of articles for delimiting a field, the results are quite different but complementary. The objectives pursued should be kept in mind when deciding the strategy to be followed.

The delimitation of a field according to the database subject classification of journals is not free of shortcomings. Some of them are shown in the following lines. We are going to illustrate some points with our experience in a specific study devoted to the Biomedical area analyzed through the Science Citation Index database.

2.a The level of aggregation of disciplines in the databases frequently does not fit our own requirements.

In our study, we were interested in the medical area.[2] We chose a multidisciplinary database instead of a medical one because we wanted to study collaboration, only detectable through the SCI database. Delimitation of the medical area in the SCI was done attending to the subject classification of journals into subfields. The delimitation of the field was made in a broad sense, choosing all those subfields with medical interest, either from the SCI or from the SSCI. Medical-related subfields from the SSCI not contemplated in the SCI included: Rehabilitation, Substance Abuse and Health Policy. Some areas were in both databases although differently represented: Public Health (58 journals in SCI and 26 journals in SSCI, 1992 edition), Psychiatry (56 journals in SCI and 71 in SSCI, 1993), Psychology (30 journals in SCI and 285 in SSCI, 1993). Some changes were made to adapt the classification to the project needs: for example, very small subfields were joined to others, in order to obtain significative figures (i.e. Urology, 36 journals, plus Andrology, 4 journals in 1992, change supported by the fact that Spanish andrologists are trained inside the Urology specialization).

Inside the medical area, it is often interesting to delimit two big sub-areas: Clinical Medicine and Biomedical Research. When trying to group the SCI disciplines into these two sub-areas we used the CHI classification of journals into levels (Clinical medicine, levels=1-2; Biomedical research, levels=3-4).[3] We obtained the following for the case of the Spanish output:

- Clinical disciplines with an average level $L<2.5$: Surgery, Internal Medicine, Urology, Gastroenterology, Cardiology, etc.
- Basic Biomedical disciplines with an average level $L>2.5$: Biochemistry, Microbiology, Cytology, Genetics, Physiology, etc.
- A third group of disciplines where a divergence in the indicators used is observed: when considering the production of the Spanish scientists their average basic/clinical level is basic although CHI classification considers them as Clinical Medicine. These disciplines were analyzed separately considering different possible classification criteria. We compared UNESCO classification (code 32 for Medical Sciences and code 24 for Life Sciences), Current Contents, SPRU, and CHI classification with our own experience of the analysis of Spanish biomedical output (average level of the Spanish papers per institution involved). The results are shown in Table 1, where C stands for Clinical and B for Biomedical research. These subfields, in between Clinical Medicine and Biomedicine, were considered as a separate group in our study.

Table 1
Clinical/basic character of research in some disciplines according to different journal classifications

	Av.level Spain	CHI levels	UNESCO	SPRU	Current Contents
Pharmacology	2.8	C	C	B	B
Neurology	3.0	C	B+C	C	B+C
Immunology	2.7	C	B+C	B	B
Endocrinology	3.0	C	C	C	C
Toxicology	2.7	C	C	B	C
Hematology	2.9	C	C	C	C

Making our own decisions about subject delimitation (journals to be included in the medical area, new non-ISI subfields appearing in our study by fusion of small ones, etc.) enables to adapt the fields to our requirements. However, it makes difficult comparisons with other studies.

2.b Fast rate of change of disciplines and journals over time. Up-dating subject classifications of journals is essential to overcome the birth of new journals, and to identify the emergence of new disciplines. For example, some changes detected in the Science Citation Index:

- Biotechnology. Subfield born in 1984 (14 journals), it includes 53 journals in 1993.
- Neurosciences. Since 1991 this subfield, with 156 journals in 1990, splits into 2 different ones: Neurosciences (145 journals in 1991) and Clinical Neurology (21 journals in 1991).
- Critical Care. New subfield since 1991, with 12 journals.
- Materials Science. It shows a remarkable growth over years: the general heading accounts for 34 journals in 1983 vs. 76 journals in 1993. Besides, there were 2 specific subgroups of Materials Science journals in 1983, vs. 7 subgroups in 1993.

Most studies focus on production over several years, so it is important to identify classificatory changes in our area of analysis, and decide how to tackle them. What is more convenient, to keep a fixed set of journals for a specific time period, or to evolve following journal birth and death? The second option was selected in our studies, as it enables to follow the evolution of disciplines over time.

2.c Due to the increasing interdisciplinarity of the research, it is becoming more and more difficult to set boundaries between disciplines. The result should be an increasing multi-assignment of journals into subfields and/or a huge group of multidisciplinary journals. Different levels of multidisciplinarity can be found: journals dealing with general topics inside a specific area, i.e. Physics or Medicine, and the commonly called "multidisciplinary journals", i.e. PNAS, Nature or Science, that cover a broad range of fields.

In our studies, we have tried to avoid multidisciplinary sections, as it is difficult to obtain conclusions from such a mixed set. An attempt was made in the biomedical area (Table 2) to assign individual articles to specific subfields, in basis to their titles, key-words, institutional address and/or to their citation profile. This initiative should account for a clearer view of the performance in the biomedical field, but it is a labour intensive "ad-hoc" measure, apart from standard ones, and it difficults the reproducibility of the study.

Table 2
Biomedical documents published in multidisciplinary journals

Subfield	No. documents
Biochemistry & Molecular Biology	47
Neurosciences	28
Physiology	24
Genetics & Heredity	22
Immunology	14
Mierobiology	9
Virology	6
Psychology	5
Cytology & Histology	3
Veterinary Medicine	3
Cancer	2
Pharmacology & Pharmacy	2
Other	7
Total	172

The multi-assignment of journals into subfields is a way to cope with the broad range of topics covered by a journal, even a specialised one; it is quite useful for the study of a specific subfield, but the multiple counting performed has to be kept in mind when analysing results from a global point of view. In our study we have reduced the number of multi-assigned journals, attempting to include all the journals with biomedical interest covered by the SCI & SSCI, but avoiding redundant presences.

3. Inter-database differences in subject classifications

3.a. Inter-database discordance in the classification of journals into subfields makes especially difficult joining information from different databases, even if they both follow the journal classification scheme.

As a sample, we can display the problems we had to face to identify the psychological journals that could be included in the biomedical area. The SCI database considers one subfield called "Psychology", while eight different subgroups of psychological journals are included in the SSCI. Experimental, Clinical and General Psychology were the subfields chosen to be included in our biomedical study, attending to a psychologist advice. To confirm the expert opinion, we consulted the "Catalogue des revues de Psychologie"[4] edited by the French CNRS, that contains a thematic classification of journals into subfields, assigns keywords to define the main topic of each journal, and indicates in which databases they are recorded. If those journals belonging to the Experimental, Clinical and General sections were closer than the rest of the journals to the biomedical area, they should also appear most frequently in a biomedical database. Unfortunately, our hypothesis could not be confirmed. Not only non-selected journals also appeared in databases of biomedical interest, but also discrepancies were found between the CNRS and the SSCI classification of journals into subfields. In summary, it seems a wise advice to avoid trying to combine classifications from different information sources.

3.b. When delimiting a specific topic through thematic classification codes in different databases, comparability is hindered by the different degree of specificity of classifications.

In a study devoted to the analysis of the output of the Environmental Research Programme in Spain, several databases were consulted: the Spanish multidisciplinary database ICYT (UNESCO classification), Chemical Abstracts (CA), BIOSIS, and the Spanish Patent database, among others. We show in Table 3 the difficulties found when trying to select the same concepts through the international patent classification, CA and UNESCO classification in the case of Water pollution. Setting correspondences between different classifications prove especially difficult when descending to more specific concepts inside the classification scheme. Aggregation to general headings was the best solution found.

3.c. A multidatabase approach to the study of a specific field may require combining not only different classifications but also different levels of analysis attending to the database. Thus, a specific topic may be delimited by the journals of the field in those databases with subject classification of journals, while keywords or thematic codes assigned to an article could lead to a more specific selection in other specialized databases.

We studied the Spanish output in Neuroscience using database BIOSIS and selecting the documents classified as "nervous system", concept code 2050.[5] When analysing the journals used for publication of the documents retrieved, according to the SCI journal subject categories, we found the following:

- More than 80% of the documents were covered by SCI;
- 31% were published in Neuroscience journals;
- 14% in Clinical Medicine, 6% in Biochemistry, 5.8% in Pharmacology, and lower percentages in Physiology, Endocrinology, Anatomy and other 50 disciplines.

Table 3

Delimiting the topic "Water pollution" through different classifications: International Patent Classification, UNESCO and Chemical Abstracts

	C02F+ (not 5+)	B63J4+	D21C11+	E02B15+	C09K3/ 32	E03F5/ 14,16	E03F5/ 18
330800	X			X		X	X
330802	X						
330804				X		X	X
330806	X	X	X				
330810	X	X	X	X	X	X	X
330811	X						
61-2	X	X	X	X	X		
61-3							
61-5	X	X	X	X	X		
47						X	X

CIBEPAT

C02F+(not 5+)	Wastewater treatment
B63J4+	Ship wastewater treatment
D21C11+	Recycling of white waters
E02B15+	Cleaning of water surface
C09K3/32	Substances for liquid polluants treatment
E03F5/14,16	Devices to separate liquids and solids from sewage waters
E03F5/18	Chambers for desinfection, neutralization and cooling of sewage waters

ICYT

330800	Environmental technology and engineering
330802	Industrial wastes
330804	Pollution engineering
330806	Reclamation of water
330810	Sewage water technology
330811	Water pollution control

C.A.

61-2	Water pollution
61-3	Analysis
61-5	Purification and chemicals in water treatment
47	Apparatus and Plant Equipment

The delimitation through SCI Neuroscience journals included articles not selected through the BIOSIS strategy due to the broad scope of many journals.

This mixed approach enables to analyse coverage of SCI on Neuroscience, and to study the multidisciplinary character of this subfield through the main related subfields. A different picture of a field can be obtained depending on the subject delimitation approach used. The type of study and its goals are key points in deciding which strategy has to be followed in each case.

4. Customized studies

Sometimes, the criteria for field delimitation are not of an international scope, but answer national needs. This is the case in studies carried out under request, in which specific "ad hoc" methods are to be developed. The difficulties for international comparisons are obviously highest in these cases. Some examples are shown in the following lines.

4.a. Facing the analysis of the Spanish pharmacological scientific production, the "subject" definition problem emerged. What can be considered as pharmacological research? Research published in "Pharmacology & Pharmacy" journals or that published by pharmacologists and pharmaceutical scientists, independently of the publication journal? Both approaches were used.[6,7] Main features of the Spanish research system were evident in both studies: irregular geographic distribution of scientific production, the university as the main type of productive institution, low contribution of the pharmaceutical industry etc. From a general point of view, differences found between both studies were not very large. However, the study based on pharmacologists was better at showing the multidisciplinary character of the field: 40% of the authors production was published in non-pharmacological journals, spread over 46 different subfields. Some of these articles could be done in collaboration with non-pharmacologists -in fact, collaboration rate was higher than in the pharmacological journal publications- so it could be one reason for being in extra-pharmacological journals. However, it is only one cause. The truth is that setting frontiers among disciplines in base to subject classification of journals is becoming more and more difficult due to the increasing multidisciplinary character of research.

4.b. Another example that can be shown is a study of the output of Materials Science in Spain.[8] Delimiting this field was not an easy task, as it includes a complex and multidisciplinary research in Physics, Chemistry, Metallurgy, Ceramics, etc. The strategy defining the field through journals or thematic classification of the publications proved to be either too broad or too narrow. Finally we took into account the goal of the study, which was to detect the changes in the Spanish research involved in New Materials originated by the policy measures of establishing this area as a national priority. We therefore decided to delimit the community of Spanish scientists working in the area as those who had obtained financial support for research projects through the National Programme of New Materials or who worked in the recently created Institutes of Materials Science.

The production and topics of interest of these authors were compared before and after the Programme being launched. The results obtained were the following: a great diversity in the background of the scientists and in the disciplines -most of them of basic character- involved in the Programme; an important increase in scientific output over time; but no change in the multidisciplinary profile of the research along the programme accomplishment was observed. Thus, it seems that the National Programme failed in one of its objectives: leading research towards more technological fields. It is worth mentioning that our bibliometric results were confirmed by a quite different study devoted to the analysis of this National Programme in Spain, performed by questionnaires and sociological methodologies.[9]

4.c. Multidisciplinarity of research institutes. The analysis of the scientific output of research institutes may also be undertaken from different viewpoints.

The Spanish Scientific Research Council (CSIC) is an autonomous body of the Ministry of Education and Science and the largest multidisciplinary research institution in Spain, with 90 research Institutes all over the country. This institution is equivalent to the French CNRS or the Italian CR. The research activity of the CSIC is organized in 8 research areas, and the institutes are distributed along these areas.

Facing the study of the scientific output of the CSIC, we had to decide whether the "subject" is defined by the research area or by the Institute, independently of the topic of the journal used for publishing, or the "subject" is defined by the publication journals, without taking into consideration where the research was performed. In the first case, for instance, "Chemistry" comprises all publications produced by scientists included in the research area "Chemistry: Science and Technology" or by the publications of scientists working in the "Institute of Organic Chemistry" and in the

"Institute of Inorganic Chemistry" etc., whatever the journals used for publication could be. According to the second choice, Chemistry is determined by the articles published in journals classified in Chemistry sub-fields, whatever the area or the Institute in which the research was developed. As both definitions could provide quite interesting perspectives, we decided to analyze the publications of the Research Council under both viewpoints.[10]

As a result, we found that the subjects are so scattered among institutes, as the publications of the institutes are scattered among subjects (Table 4). Thus, multidisciplinarity is more than guarantee. These data from our parallel analyses can be of interest for science policy management of the Research Council as a whole.

Table 4
Scattering of CSIC publications among CSIC research areas and SCI subfields

		No. Institutes	No. Documents	No. SCI subfields
CSIC research area	Materials Science & Technology	11	1388	48
SCI subfield	Materials Science	27	250	
CSIC research area	Chemistry Science & Technology	14	1425	81
SCI subfield	Chemistry	43	307	
	Analytical Chemistry	34	190	
	Applied Chemistry	25	89	
	Medical Chemistry	5	20	
	Organic Chemistry	28	446	
	Inorganic Chemistry	21	251	

5. Subject classification problems in input-output analysis

Any sort of input-output analysis in a research system, has to face the difficulty of harmonizing the different classification schemes used for input and output data. Thus, thematic distribution of input data -as investment and manpower- can hardly be confronted with thematic distribution of patents or publications, the latter done attending to any of the methods mentioned above. As an example, and concerning the Spanish case, the Ministry of Education and therefore our financing agencies classify research projects following the UNESCO International Classification of Science and Technology (with 22 scientific fields, disciplines and subdisciplines) and the socio-economic objectives scheme. On the other hand, the statistical data of budget and manpower from the Spanish Instituto Nacional de Estadística follow the UNESCO classification for only 5 major areas; the National Classification of Economic Activities for productive sectors; the socio-economic objectives classification (Frascati Manual) or product groups (in the industrial sector). Output data are usually obtained from bibliographic databases (each with their own classification criteria) or from patent databases (International Patent Classification), whose differences have already been discussed.

Conclusions

As shown above, the concept of "subject" can be analyzed from different points of view. Field delimitation in bibliometric studies can be done attending to document thematic codes or keywords, publication journals, institutional addresses, disciplinary areas of a centre, professional training of the authors...and so on. Each of these methods possesses advantages and shortcomings. The final objective of a study is the key point in deciding which method has to be used for delimiting a field. However, it is desirable to use a systematic approach to whatever the method, to guarantee reproducible studies.

Undoubtedly, the use of subject classifications widely spread all over the world has to be pursued whenever possible and for the sake of international comparisons, although national objectives cannot be forgotten. Bibliometric analyses performed by request of governmental agencies are a common practice in most countries. They can be defined as "customized" studies, since they are performed according to the customer (institution, government, etc.) needs -including field delimitation- and are often far away from international standards. The role of the bibliometrician would be to satisfy the customers demands trying to follow international standards as much as possible. Standards should be flexible enough to both enable international comparisons and meet the needs of local studies.

However, both national and international classifications have to be periodically updated. Concerning journal subject classifications, one of the most frequently used approaches to field delimitation, it accounts for reproducibility and international comparability of studies, although a new question is emerging now: is its validity decreasing as the multidisciplinary character of research and the rate of change of disciplines both grow over time?

Finally, disagreement among subject classifications of input and output data appears as an important obstacle in studies with evaluative purposes. The need of subject standardization emerges strongly when trying to confront bibliometric results with other kind of input and output measures. Working close to the national bodies should be convenient in these cases, in order to obtain data of economic and personal resources in the desirable level of aggregation, data usually not publicly available.

References

1. H.F. Moed, R.E. de Bruin, A.J. Nederhof, A.F.J. van Raan, R.J.W. Tijssen, State of the Art of Bibliometric Macro-Indicators, Commission of the European Communities. Luxembourg, 1992.
2. I. Gómez, J. Camí, *La producción científica española en Biomedicina y Salud. Un estudio a través del Science Citation Index (1986-89),* Report under contract FIS 90/4001. Madrid, 1992.
3. E. Noma, *Subject Classification and Influence Weights for 3000 Journals,* Research report under CHI and NIH contracts. Computer Horizons Inc. Research, New Jersey, 1986.
4. Catalogue des revues de Psychologie. CNRS, 1988.
5. I. Gómez, E. Sanz, A. Méndez, Utility of bibliometric analysis for research policy: a case study of Spanish research in neuroscience, *Research Policy*, 19 (1990) 457-466.
6. M. Bordons, F. García-Jover, S. Barrigón, Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-89). Part 1. Contribution to the "Pharmacology & Pharmacy subfield", *Scientometrics*, 24 (1992) 163-177.
7. M. Bordons, S. Barrigón, Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-89). Part II. Contribution to subfields other than "Pharmacology & Pharmacy", *Scientometrics,* 25 (1992) 425-446.
8. A. Méndez, M.A. Insúa, I. Gómez, G. López, C. Refolio, *Dinámica de la investigación multidisciplinar sobre nuevos materiales en España,* CINDOC, Madrid, 1993.
9. J. Espinosa de los Monteros, F. Martínez, M.A. Toribio, E. Muñoz, *El Programa Nacional de Nuevos Materiales en el período 1988-92.* Su evaluación mediante una metodología dual. Instituto de Estudios Sociales Avanzados, Madrid, 1994.
10. A. Pestaña, I. Gómez, M.T. Fernández, M.A. Zulueta, A. Méndez, Scientometric evaluation of R&D activities in medium-size institutions: a case study based on the Spanish Scientific Research Council (CSIC). *Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics,* Medford, Learned Information Inc, 1995. p. 425-434.