Field Delineation Using Medical Subject Headings (MeSH) – An Alternative Way to Aggregate Data in the Web of Science

Håkan Carlsson*[1], Ed C.M. Noyons[2]

[1] *Hakan.Carlsson@ub.gu.se*
Gothenburg University Library, University of Gothenburg, P.O. Box 222, SE-405 30 Gothenburg (Sweden)

[2] *Noyons@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX Leiden (the Netherlands)

Abstract

Field delineation is important in bibliometrics. Particularly when measuring scientific performance, the demarcation of the area, to which a science actor belongs, is of the greatest importance. This paper present a method based on searches on Boolean combinations of medical subject headings (MeSH-terms) in a combined Web of Science (WoS) – Medline database. The construction of the MeSH area definitions is described and recall is discussed and compared to other methods.

Introduction

The use of automated methods for field-based delineation is important in bibliometrics as the need of large datasets for statistical purposes and the interest in comparisons to others (benchmarks) makes it impossible to select publications manually.

MeSH-terms are added to all Medline-records and based on a well-defined thesaurus of subject terms. In this study we show them to be suitable as building blocks in defining medical research areas.

In other studies, primarily based on the Web of Science database (WoS), fields generated through the WoS journal categories have often been used. This has shown to be convenient, but not problem free, as it leaves out large numbers of publications. (Aknes *et al*, 2000) In a recent attempt, the original categories were extended by adding publications in other areas with at least 10% of their references to the original category journals. (López-Illescas *et al*, 2009) This extended method was used as a reference in this work.

Materials and Method

A number of MeSH-based definitions were constructed in collaboration with medical experts at Lund University. The refined workflow of construction started by a simple search on common topic names related to the area in PubMed Entrez. The resulting starter sets of bibliographic records were analyzed for frequency of the contained MeSH terms. These were determined by generating a background reading of a whole year of PubMed data and then comparing the frequency of the individual MeSH terms of the generated starter sets to the background. To simiplify this procedure and to take the MeSH tree structure into account, the web-based tool, Meva, was used. (Tenner *et al*, 2003)

The resulting definitions were improved by an iterative process where experts from each individual subfield were confronted with the raw definitions and publication test data from PubMed. This finalized the definitions.

The MeSH-areas were in our study used to collect publications in Web of Science for citation benchmark purposes. This was done by merging the Medline and Web of Science datasets in a matching procedure and then searching on the MeSH-terms in the combined database.

To verify the method, a study was made comparing the data collected for the areas of neuroscience and cancer research with the related WoS subject area and corresponding reference-extended areas. (López-Illescas *et al*, 2009) A number of indicators were calculated to illustrate the nature of the collected data sets.

Results

The final MeSH-area definition (Tab 1) and the calculated indicators in the verification study (Tab. 2) are tabulated for the area of neuroscience. The results for oncology/cancer research show very similar trends and will be presented in the poster.

Table 1. MeSH area definition of the field of Neuroscience. All MeSH tree subareas should be included in each search term

Neurosciences OR (Nervous System Diseases NOT (Cerebrovascular Accident OR Pain OR Sensation

Disorders OR Muscular Diseases OR Muscular Disorders, Atrophic OR Carotid Artery Diseases OR Sleep Apnea, Obstructive OR Pituitary Diseases)) OR Nervous System OR (Neurosurgical Procedures NOT Hypophysectomy) OR Brain Mapping OR Brain Chemistry OR (Diagnostic Techniques, Neurological NOT Pain Measurement) OR Headache OR Cerebrovascular Accident

Table 2. Indicators for the area of Neuroscience using the MeSH definition, classic and extended WoS categories. P (number of publications), C (number of citations until 2008 without self-citations) and CPP (citations per publication) were collected for publications 2003-2006.

	P	C	CPP
MeSH Area	211,431	961,219	4.08
WoS Category	109,102	476,720	4.55
Extended WoS Category	243,701	994,650	4.37

Discussion

The construction of a MeSH area definition was a very tedious process, but the product showed to be a versatile tool in the delineation of the field in question. The MeSH area for both cancer research and neuroscience collects about twice the number of publications compared to the corresponding Web of Science journal category areas. On the other hand, the recall is comparable to that of the extended areas. This indicates the limited scope of the original WoS categories, which exclude many publications in both the outskirts of the areas, as well as a large number of publications categorized to e.g. the multidisciplinary category, where journals such as Nature and Science are found.

The citations per paper is slightly elevated for the MeSH area, which can be attributed to the lack of about 10% of lowly-cited WoS category journals in Medline and hence in the combined database.

Conclusions

The creation of MeSH areas allows access to the added granularity of a publication level delineation method and conforms well to other publication level methods. The MeSH-term-based method gives larger freedom of how to form the fields compared to the recent reference-based WoS category extension method, which demands a similar original area. Based on recall, the methods give similar results. The collaborative work of Christer Larsson of the Lund medical faculty is kindly acknowledged.

References

Aksnes, D.W. & Olsen, T.B. & Seglen, P.O. (2000). *Scientometrics* 49 : 7-22.

López-Illescas, C. & Noyons, E.C.M. & Visser, M.S. & de Moya-Anegón, F. & Moed, H. F. (2009) *Scientometrics* in press

Tenner, H. & Thurmayr, G.R. & Thurmayr, R. (2003) *http://www.med-ai.com/meva/*

Shifts in Knowledge Production Patterns of Mexican Physics in Particles and Fields

Collazo-Reyes, F[1], Luna-Morales, ME[2], Russell, J[3] &Pérez-Angón, MA[1]

[1] fcollazo@fis.cinvestav.mx, mperez@fis.cinvestav.mx
Centro de Investigación y de Estudios Avanzados del IPN, Departamento de Física, Av. IPN 2508, Col. San Pedro Zacatenco, 07360, (México)

[2] *elena@csb.cinvestav.mx*
Centro de Investigación y de Estudios Avanzados del IPN, Unidad de Servicios Bibliográficos, Av. IPN 2508, Col. San Pedro Zacatenco, 07360,(México).

[3] jrussell@servidor.unam.mx
UNAM (Universidad Nacional Autónoma de México), Centro Universitario de Investigaciones Bibliotecológicas, Ciudad Universitaria, 04510 (México)

Introduction

Mexican research in the physics area of particles and fields (MPPF) has evolved by diversifying ways of generating knowledge and thus enriching organizational, production and scientific communication structures. Its first significant achievement was to gain continuity in the production of results. This was accomplished on the basis of theoretical research and to a lesser extent, phenomenological, steadily showing more consistent scientific practice which served as a basis for the formation of new research groups and the development of other types of research within the discipline (Collazo-Reyes, Luna-Morales & Russell, 2004). In order to examine this change, we analyze in detail the scientific production of this community in the period 1948-2007 through two different but complementary bibliographic systems: the traditional *Science Citation Index Expanded* (SCIE) framework and the *Stanford Public Information REtrieval System (SPIRES)*.

Material and Methods

Table 1 shows the retrieved papers and citations organized in the following way. The 120 papers and the 793 citations found in SCIE, were used to complete the analysis of the period not covered by SPIRES. The 3,278 papers found in both databases received, 33,387 citations from published papers and 21,803 citations from unpublished work in SPIRES and 40,942 citations in SCIE. The 1,784 unpublished papers in SPIRES received 2,894 citations from published sources and 2,185 citations from unpublished sources. The 677 papers identified in SPIRES published in sources not covered by SCIE, generated 1,315 citations from published sources and 712 from unpublished sources.

Table 1. Papers and citations in SCIE and SPIRES].

		Citations in SPIRES			
Paper Type	**Nos. of Papers**	**From Published Papers**	**From Unpublished Papers ***	**Total SPIRES**	**Citations In SCIE**
Available in SCIE	120				793
Available in SCIE and SPIRES	3278	33387	21803	55190	40942
Unpublished papers SPIRES	1784	2894	2185	5079	
Published not covered in SCIE	677	1315	712	2027	
Totals	5859	37596	24700	62296	41735

* e-prints, conference papers, conferences, theses.

The 3, 278 papers found in both databases were disaggregated into the five distinct types of research assigned in SPIRES system: (1) theoretical, (2) phenomenological, (3) experimental, (4) cosmological and (5) other. The publications and citations retrieved from the SPIRES were classified

into three types: 1) published in journals also covered by SCIE, 2) published in sources not covered by SCIE, and 3) unpublished.

Results

The 3,278 papers generated 40,942 citations according to SCIE and 55,190 according to SPIRES. Table 2 shows the distribution of these papers and citations by type of research and the respective percentages. We have included both SCIE and SPIRES citations to each type of research. In the latter case we have considered also the citations given in sources not included in the SCIE. The theoretical papers represent the largest share in production but they acquired only 27.7% of the SCIE citations, while the experimental papers represent only 16.1% of total production but received 25.2% of the SCIE citations. On the other hand, "other" papers generated 15.1% of the SCIE citations with just 0.9% of the production. With the sole exception of the experimental papers, publications received more citations in SCIE than in SPIRES.

The first publications in the 70s involved theoretical and phenomenological studies, but in the 80's most of the publications were only on theory. By the early 90´s research had diversified: theoretical, phenomenological, experimental, cosmological, and other types.

Table 2. MPPF: papers, citations and averages in SCIE and SPIRES].

SPIRES-SCIE	THE	% THE	PHE	% PHE	EXP	% EXP	COS	% COS	OTH *	% OTH	TOT
Papers: 1971-2007	1370	41.8	868	26.5	527	16.1	483	14.7	30	0.9	3378
Citations SCIE											
Citations	11353	27.7	8888	21.7	10328	25.2	4211	10.3	6162	15.1	40942
Average	8.3		10.2		19.6		8.7		205.4		12.5
Citations SPIRES											
Total citations	11323	20.5	11246	20.4	19746	35.8	5265	9.5	7610	13.8	55190
From published papers	8595	25.7	6560	19.6	9950	29.8	3716	11.1	4566	13.7	33387
From unpublished papers	2728	12.5	4686	21.5	9796	44.9	1549	7.1	3044	14	21803
Total average	8.3		13		37.5		10.9		253.7		16.8
From published papers average	6.3		7.6		18.9		7.7		152.2		10.2
From unpublished papers average	2		5.4		18.6		3.2		101.5		6.7

Columns: The = Theoretical; Phe = Phenomenology; Exp = Experimental; Cos = Cosmology; Oth = Other; Tot = Total.

***** Incudes review articles written in collaboration by the Particle Data Group wich are the most cited in this subject area.

Discussion and Conclusions

MPPF was developed under two distinct knowledge production structures. (I) A steady state where institutions, academic profiles, financing, professional training, and teaching programmes, all operated within the conditions surrounding the theoretical and phenomenological modes of knowledge production. (II) A period of growth and expansion of the organizational, subject and cognitive boundaries of the scientific practice; characterized by diversification of the types of research and the incorporation of a new generation of Mexican experimental physicists to the collaborative practices of elite, multi-institutional groups (Jones, Wuchty, & Uzzi, 2008) which complemented and enriched the knowledge production patterns of the MPPF.

References]

Collazo-Reyes, F., Luna-Morales, M.E. & Russell, J.M. (2004). Publication and citation patterns of the Mexican contribution to a Big Science discipline. *Scientometrics*, 60 (2), 131-143.

Jones, B.F., Wuchty, S. & Uzzi, B. (2008). Multi-University Research Teams: Shifting, Impact, Geography, and Stratification in Science. *Science*, 322(21), 1259-262.