HOKKAIDO UNIVERSITY

Title	Comparative Analysis of Scientific Publications of Research Entities Using Multiple Disciplinary Classifications
Author(s)	Pitambar, Gautam
Citation	2016 5th IIAI International Congress on Advanced Applied Informatics, July 10-14, 2016, Kumamoto, Japan, 523-528 https://doi.org/10.1109/IIAI-AAI.2016.117
Issue Date	2016-07
Doc URL	https://hdl.handle.net/2115/62522
Rights	© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Type	conference paper
File Information	multidisc-class-IIAI2016-gautam-finalversion.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Comparative Analysis of Scientific Publications of Research Entities Using Multiple Disciplinary Classifications

Pitambar Gautam
Creative Research Institution (CRIS)
N21 W10, Kita-Ku, Sapporo 001-0021, Hokkaido, Japan
e-mail: pgautam@cris.hokudai.ac.jp

***Abstract*—This study attempts a comparative disciplinary analysis of a Web of Science (WoS) recorded dataset (ca. 15,000 core journal articles and reviews) of real publications for 5-years period by the research departments, dedicated to science, technology, engineering, mathematics, and medicine (STEMM), of a comprehensive university using different disciplinary schemes: Essential Science Indicators (ESI) 22 research fields, SCOPUS 27 subject areas and OECD Frascati 38 subordinate research fields. It demonstrates that assigning the publications to departments and disciplines followed by correspondence analysis and clustering of the contingency table comprising disciplinary share of publications results in enhanced understanding and visualization of the research output at different levels of disciplinary or research entities. Such exercise improves creation of publication subsets (for different research entities and disciplinary schemes) for further processing, visualization and interpretation with sophisticated bibliometric/scientometric tools including the science mapping techniques.**

Keywords-bibliometry; science map; correspondence analysis; clustering; research field; cross-disciplinarity; co-word; co-citation

I. INTRODUCTION

Research volume or the number of publications of a particular research entity and research quality, judged by the research impact measured by citations, have been increasingly used as the basic indicators of research performance at various levels: researcher, research unit, university or research institute as well as country or even regional geopolitical unit. Also, these are in wide use for research benchmarking such as university ranking, both global and regional, with varying weights along with the numerous indicators, either as raw values or after normalization for time, subject category, publication type, the size of research entities, research expenditure, and so on. Complex bibliometric indicators, such as citation percentiles (in particular, percentage of publications within the citation-based top 1% and top 10% percentile windows), citation indices (h-index and its likes), journal influence factors like the journal quality indicators and discipline-standardized citation indicators are also commonly used in research evaluation. Citation data currently comprise the 'big data' subjected to map the co-citation relationships of the global research publications and form the basis for delineating or mapping the research strengths (e.g., research fronts, and research competencies) of various research entities.

During the process of collection, calculation and interpretation of the research performance indicators, it is desirable to take into account the disciplinary classifications (research fields, subject areas, subject categories, etc.) that most suit the disciplinary orientations of the target research entities. Also, citation-based indicators normally involve the subject categories consideration that may technically depend on the database owing to the preference of producers to specific disciplinary classification schemes. In view of the ever-increasing use of the citation-based indicators in research performance assessment, and the increasing engagement of the university research administrators (URAs) in informing as well as assisting the executives in formulation, planning and designing the research strategies, it is now time to analyze the research publications considering all possible factors including the disciplinary or cross-disciplinary (including multidisciplinary, transdisciplinary and interdisciplinary) structure of the research units [1,2].

In the above-mentioned context, this study compares a set of publications data by departments engaged in science, technology, engineering, mathematics, and medicine (STEMM) in the following sequence: (i) Breakdown of publications to departments and 3 different disciplinary classes; (ii) Correspondence analysis and clustering to reveal department-discipline relationships; (iii) Regrouping of publications based on the departmental and disciplinary characteristics, especially for departments with distinct cross-disciplinary (including multidisciplinary and interdisciplinary) nature; and (iv) Further analysis of the regrouped datasets through more sophisticated mapping and visualization tool.

II. PUBLICATIONS DATA AND METHODS OF ANALYSIS

This study utilized a dataset of 14,689 articles & reviews published during 2009-2013 by a comprehensive and research-intensive university obtained after cleaning the publications harvested from the Thomson Reuter's WoS Core Collection [3] on February 23, 2015. Each publication was assigned to the department(s) (research unit(s) such as faculty, graduate school, research institute or research center executing research and or education programs in independent manner) by identifying all possible affiliation variants using a semi-automatic worksheet-based matching approach, and simultaneously also to ESI research fields following Thomson Reuters' master journal list [4] using the whole-count method. Owing to poor

representation of the publications from the social sciences and arts & humanities departments, the subset for STEMM departments shown in Table I was used for further analysis using multivariate techniques (correspondence analysis, clustering) and visualization through stacked bar diagrams to reveal the details of the department-discipline relationships as demonstrated in ref. [5] for ESI fields. A total of 13,537 publications, i.e. ca. 92% of the primary WoS-based dataset overlapping with Scopus, was further analyzed for SCOPUS [6] and OECD Frascati schemes listed in Table 2. After thorough analysis of the department-discipline relationships following the procedure in ref. [5], the publications data were regrouped into three major disciplinary groups composed of publications from major departmental clusters combined with additional subsets discriminated by subfields specific to multidisciplinary departments (e.g., earth science subset from SCI department, as shown in Fig. 4 in ref. [5]). These groups were subjected to science mapping using more sophisticated techniques such as the Vosviewer v. 1.6.3 – a Leiden University Software for visualizing scientific landscapes [7].

I. Results And Interprettion

A. Department-Discipline Relationships from Correspondence Analysis

Correspondence diagrams [5, 8] shown in Fig.1, as 2D maps with Axis 1 (explaining 27-31% of variance) versus Axis 2 (explaining 20-22% of variance), for departments & disciplines separately for each of the 2 different schemes show three broad departmental clusters as follows: (i) departments related to physical sciences and engineering positioned close to the negative part of axis 1 at the left side; (ii) those related to earth and environmental sciences, agriculture and fisheries related in the upper right quadrant; and, (iii) departments related to medical and health sciences including dentistry, veterinary, and pharmaceutics related departments in the lower right quadrant). The position of individual discipline relative to the departments is indicative of the degree of department-discipline proximity. Four departments (SCI, CRIS, LFSC and PHARM) plot close to the origin (cross-section of the first two dimensions) but far from most disciplines showing the cross-disciplinary nature.

TABLE I. Departments and related academic disciplines

Depart-ment	Disciplinary Focus	**Depart-ment**	Disciplinary Focus
AGR	Agriculture	**IGM**	Genetic medicine
CIIS	Isotope science	**ILTS**	Low temperature science
CRC	Catalytic chemistry		
CRIS	Transdisciplinary sciences	**LFSC**	Life science
DENT	Dentistry	**MEDH**	Medicine & Hospital
EES	Environmental science	**MUSE**	Museum
ENGG	Engineering	**PHARM**	Pharmacology & pharmacy
FISH	Fisheries		
FLSC	Field science	**RCIQE**	Quantum electronics
GCSE	Chemical sciences & engineering	**RCZC**	Zoonoses
		RIES	Electronic science
GIST	Information science & technology	**SCI**	Natural sciences
		VETM	Veterinary medicine
HLTS	Health sciences		

TABLE II. Subject categories (areas, fields and subordinate fields) used to assign publications

ESI22 Fields	**SCOPUS Subject Areas**	**OECD (Frascati) Subordinate Fields****	
Agricultural sciences	Agricultural and biological sciences	Agriculture, forestry, fisheries	Other agricultural science
Biology & biochemistry	*Arts and humanities	Animal and dairy science	Other engineering and technologies
Chemistry	Biochemistry, genetics and molecular biology	*Art	*Other natural sciences
Clinical medicine		Basic medical research	*Other social sciences
Computer science	*Business, management and accounting	Biological sciences	*Philosophy, ethics and religion
*Economics & business	Chemical engineering	Chemical engineering	Physical sciences and astronomy
Engineering	Chemistry	Chemical sciences	*Political science
Environment/ecology	Computer science	Civil engineering	Psychology
Geosciences	*Decision sciences	Clinical medicine	*Social and economic geography
Immunology	Dentistry	Computer and information sciences	*Sociology
Materials science	Earth and planetary sciences	Earth and related environmental sciences	Veterinary science
Mathematics	*Economics, econometrics and finance		
Microbiology	Energy	*Economics and business	
Molecular biology & genetics	Engineering	*Educational sciences	
	Environmental science	Electrical engineering, electronic engineering	
*Multidisciplinary	Health professions		
Neuroscience & behavior	Immunology and microbiology	Environmental biotechnology	
Pharmacology & toxicology	Material science	Environmental engineering	
	Mathematics	Health sciences	
Physics	Medicine	*History and archaeology	
Plant & animal science	Multidisciplinary	Industrial biotechnology	
Psychiatry/psychology	Neuroscience	*Languages and literature	
*Social sciences, general	Nursing	*Law	
Space science	Pharmacology, toxicology and pharmaceutics	Materials engineering	
		Mathematics	
	Physics and astronomy	Mechanical engineering	
	Psychology	*Media and communication	
	Social sciences	Medical engineering	
	Veterinary	Nano-technology	

* Subject areas/fields not considered in this study of departments dealing with natural sciences, engineering, technology, mathematics, and medicine **InCites OECD Category to WoS Category Mapping 2012

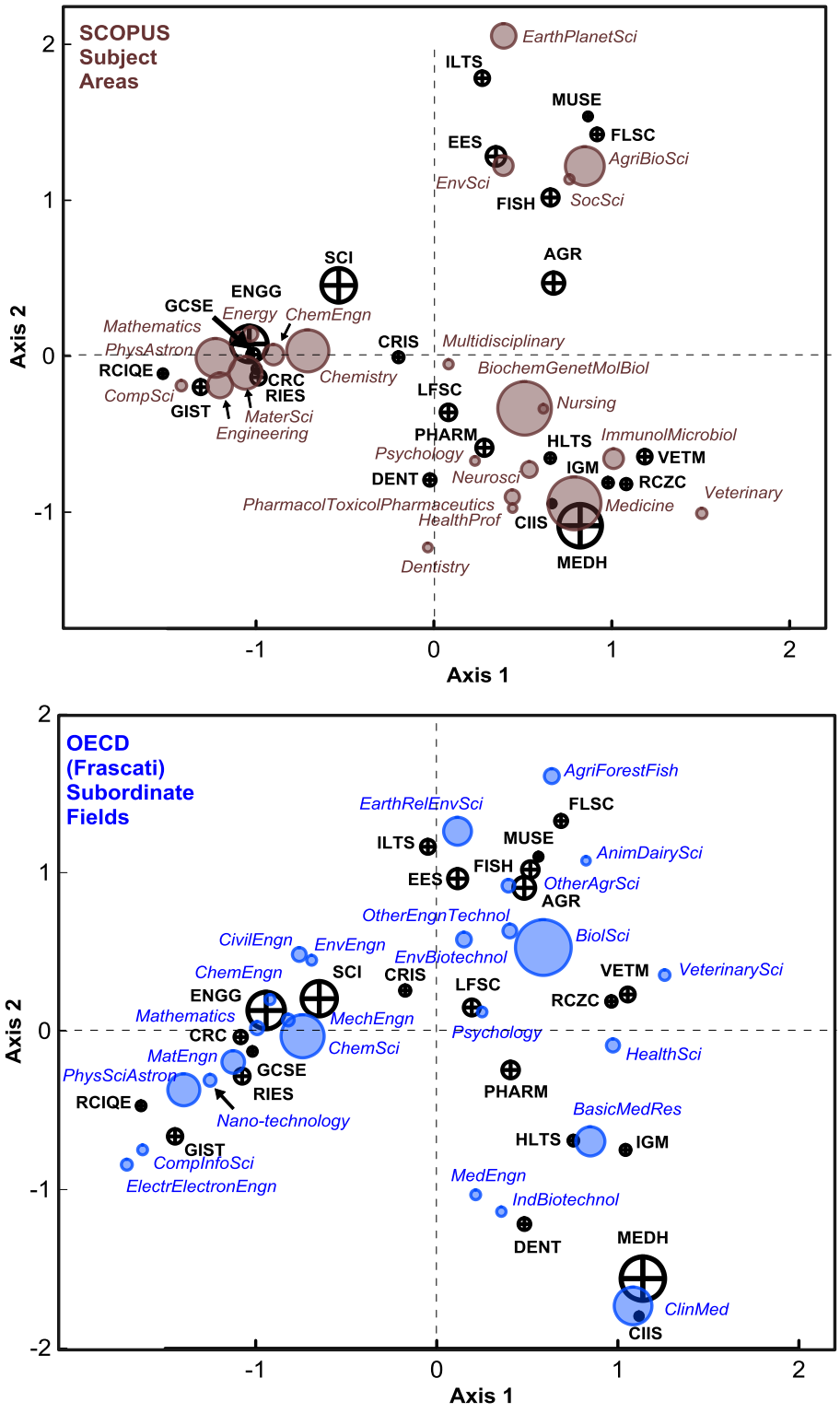

Fig. 1. Visualization of department/discipline contingency data in terms of two disciplinary schemes (above: SCOPUS Subject Areas; below: OECD Subordinate Fields) using correspondence analysis. Crossed-out and filled circles stand for departments and subject categories, while their size is proportional to the logarithm of the publications counts ranging from 50 for CIIS to 3383 for Biological Sciences (BiolSci). Coordinates for departments (same in both diagrams) and subject categories vary depending on the discipline-department relations in terms of 2 major axes (explaining 27-31% and 20-22% of total variance, respectively). For clarity, department and subject categories are abbreviated (Tables I & II). Three departmental clusters surrounded by disciplinary clusters, named as "engineering and information science"; "clinical & basic medicine including dentistry, health science and veterinary science"; "earth, environmental and agricultural sciences" can be clearly discriminated. Four departments, i.e., SCI, CRIS, LFSC and PHARM, plot close to the origin and far from any of the subject categories implying to their cross-disciplinary nature. For the "engineering and information science" cluster, the OECD scheme (left, lower diagram) reveals more details than the SCOPUS scheme (left, upper diagram). GIST is well characterized by proximity of "Computer and Information Science" and "Electrical & Electronic Engineering".

B. *Visualization of Department-Discipline Relationships using Stacked Bars*

Department-Discipline relationships are shown in Figs. 2&3 as stacked bars depicting the percentage contribution of disciplines (along the vertical) to the departments (along the horizontal). The sequences of disciplines and departments are based on the results of hierarchical clustering done following the procedure given in ref. [5]. The stacked bars enable further appreciation of the clusters of departments engaged in similar disciplines. One distinct feature is the departmental ubiquity of "Biochemistry, genetics and molecular biology" of SCOPUS and "Biological sciences" of OECD schemes, whereas no ubiquitous field occurs in ESI scheme in Fig. 3 of ref. [5]. A comparison of these stacked plots suggests that seemingly similar fields in the different schemes may have largely varying disciplinary shares at the level of individual department and such behavior is expected to be valid for any research entity differing in scale. Due to the differences in the ways of assigning publications to disciplines (unique journal-based assignment to ESI field but multiple journal-based assignments of a single publication to several subject areas in SCOPUS and OECD schemes), such comparisons are difficult to support by sound statistics.

C. *Construction of Custom Publication Subsets and Mapping of Science*

In order to test construction of training datasets and illustrate the science mapping, the entire dataset was subdivided into three main subsets: (i) "earth & environmental sciences, agriculture and fisheries" comprised by publications assigned to AGR, FISH, FLSC, MUSE, EES, ILTS and the biology and geoscience- related partial dataset from SCI, and "biology" dataset from LFSC; (ii) "medical, health and life sciences" comprised by publications by MEDH, HLTS, DENT, IGM, RCZC, VETM, LFSC, PHARM and the partial subset from SCI related to life science; and, (iii) "physical sciences, engineering and computer science" comprised by publications by ENGG, GCSE, GIST, RCIQE and partial subset from SCI excluding publications related to geology, life sciences and biology. Discrimination of the partial disciplinary subsets for cross-disciplinary department followed the procedure outlined in ref. [5].

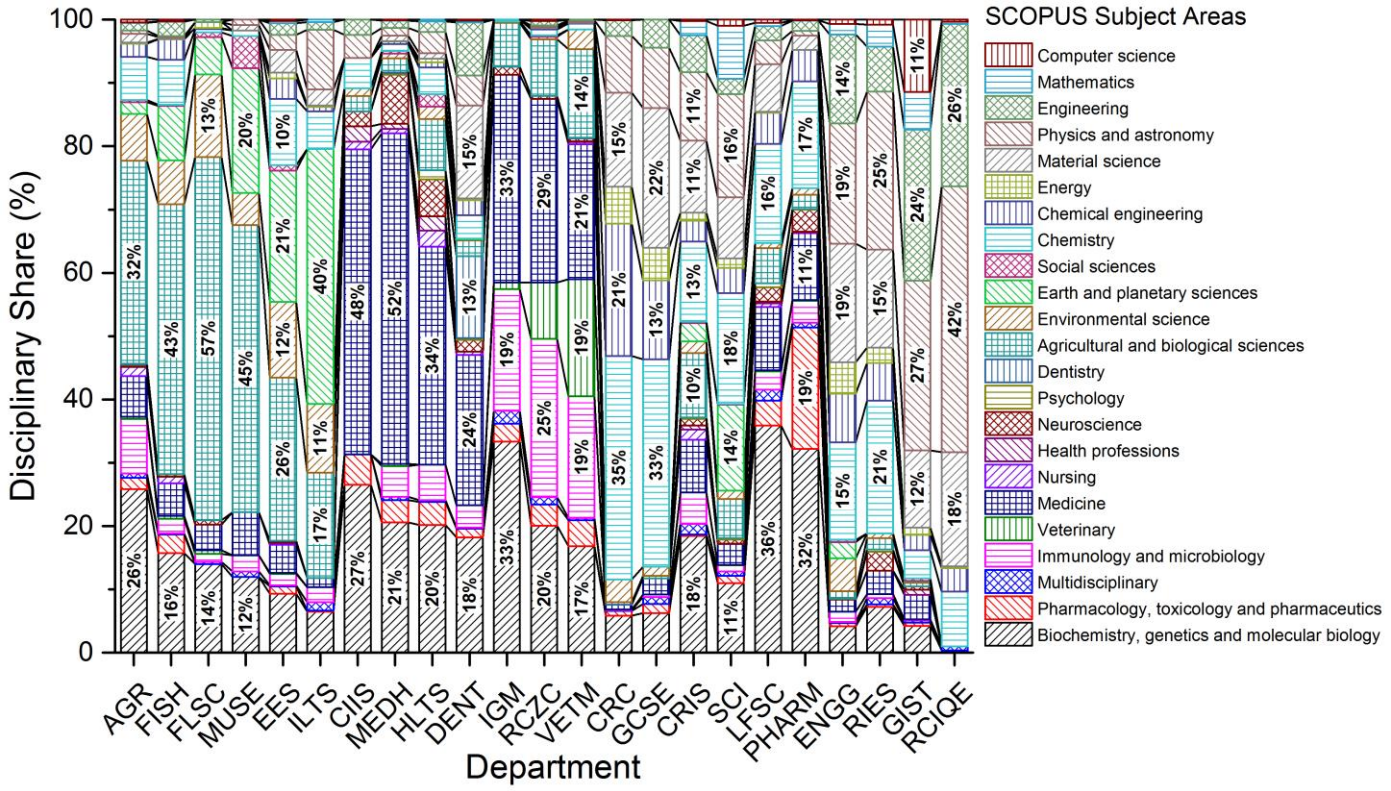

Fig. 2. Visualization of department and disciplines (SCOPUS Subject Areas) relationships as stacked bars. For each department placed along the horizontal axis, the percent share of disciplines is shown by vertical bars. The most obvious contributors, from left to right are: (i) Agricultural and biological sciences and Earth and planetary sciences (joint cluster of AGR, FISH, FLSC, MUSE, EES, ILTS), in the left; (ii) Medicine (CIIS, MEDH, HLTS ,DENT, IGM, RCZC, VETM); (iii) Chemistry (cluster of CRC, GCSE); Physics and astronomy (joint cluster of ENGG, RIES and GIST, RCIQE); and, (iv) Biochemistry, genetics and molecular biology, prominent in IGM, LFSC and PHARM but encompassing all but RCIQE. Widely spanning cross-disciplinary research characterizes 3 departments (SCI, ENGG and CRIS) for which the share of single subject category is less than ca. 25%.

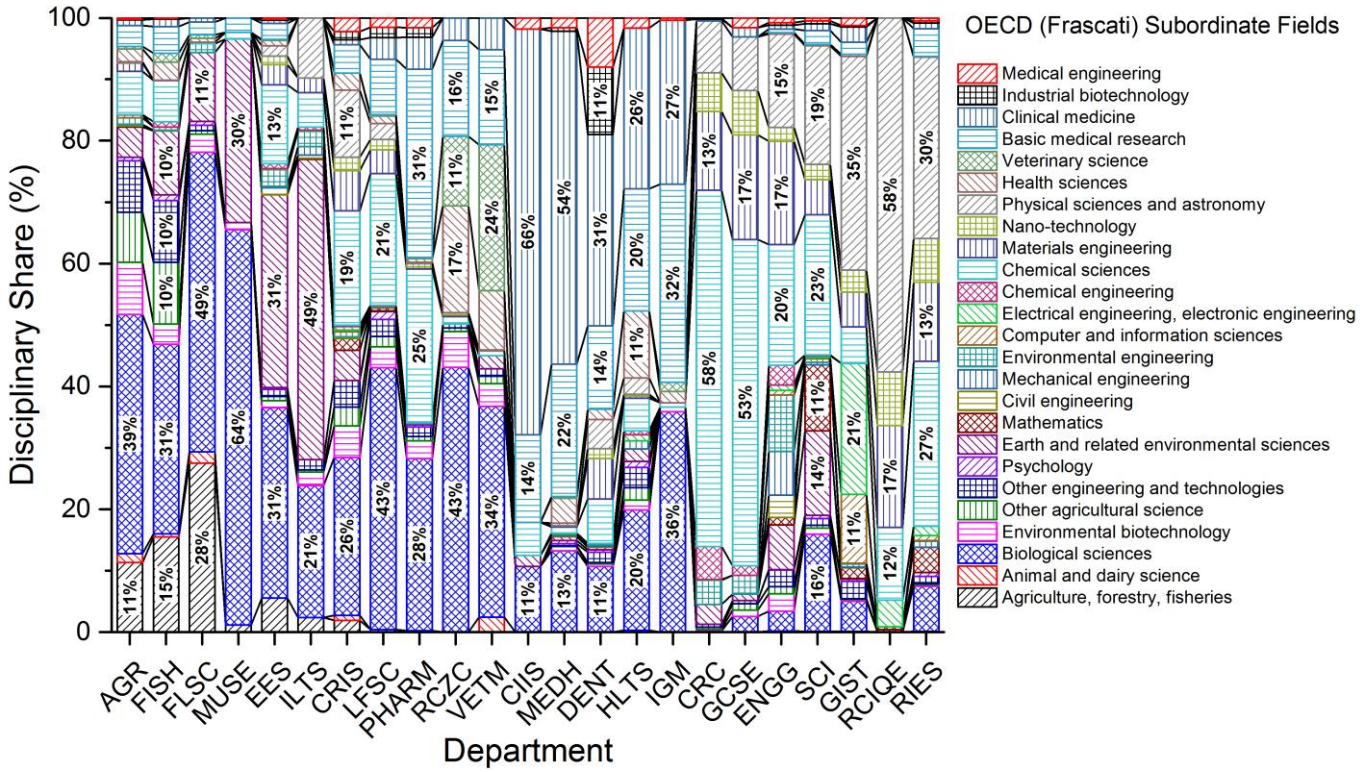

Fig. 3. Visualization of department and disciplines (OECD (Frascati) Subordinate Fields) relationships as stacked bars with plotting conventions as in Fig. 2. The most obvious disciplinary contributors (departments), from left to right, are: (i) Biological sciences (AGR to VETM); (ii) Earth and related environmental sciences (MUSE, EES and ILTS); (iii) Clinical medicine (joint cluster of CIIS, MEDH, DENT, HLTS, IGM); (iv) Chemical sciences (CRC, GCSE as well as ENGG and SCI); Physical sciences and astronomy (somewhat lower share of ENGG and SCI, but prominent in the cluster of GIST, RCIQE and RIES). Three departments, i.e., SCI, ENGG and CRIS, with the share of individual subject category being always less than ca. 25%, are more cross-disciplinary than other departments.

Dataset (i) "earth & environmental sciences, agriculture and fisheries" with 3,725 papers (ca. 25% of all) was subjected to further analysis by various methods. The following section (with Figs. 4 and 5, in the following page) describes the results of mapping for co-words in titles and abstracts of publications and co-cited sources (journals) as the representative examples.

The co-word map in Fig. 4 was constructed using 428 nodes (terms and noun phrases) occurring more than 20 times in the text comprising title and abstract of each publication (article or review) and representing 60% of the most relevant terms with highest scores using binary counting method. Clustering resolution and minimum clusters size were set as 1.2 and 10 resulting in 5-clusters solution. The clusters (sets of closely related nodes) shown by different color shades correspond to areas of intense research identified tentatively as follows: (i) biodiversity; (ii) molecular biology and microbiology; (iii) aerosols & particulate matters; (iv) forestry; and (v) earth-ocean-atmosphere studies including earthquake and tsunami disasters. In the map, the size of a node is proportional to the frequency of its occurrence, strongly related nodes are located close to each other, and the edges or lines connecting two nodes indicate the co-occurrence relations [7].

The co-citation map in Fig. 5 shows the relationship among 251 journals cited >100 times each and strongly linked through joint citations. Joint analysis of such map and accompanying analytical data reveals 9 journals (Nature, Science, Geophysical Research Letters, Journal of Geophysical Research: Atmospheres, Proceedings of the National Academy of Sciences of the United States of America, Journal of Geophysical Research: Oceans, Atmospheric Environment, Ecology, and Limnology and Oceanography) to have highest citations and co-citations. Also, the most influential journals in each cluster (e.g., Ecology, Evolution, Oecologia, Molecular Ecology, and New Phytology for environment/ecology cluster) can be readily identified. Such information on the high-impact journals in each discipline can be used to formulate the future publication strategy so as to achieve maximum impact of research through citations.

REFERENCES

[1] L. Leydesdorff, "An evaluation of impacts in "Nanoscience & nanotechnology": steps towards standards for citation analysis," Scientometrics, 94(1), pp. 35-55, 2013.

[2] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W.. Boyack, J. Keyton, I. Rafols, and K. Börner, "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature," Journal of Informetrics, 165, pp. 14-26, 2011.

[3] Thomson Reuters, IP & Science. http://ipscience.thomsonreuters.com/product/

[4] Master Journal List – IP & Science - Thomson Reuters, http://ip-science.thomsonreuters.com/mjl/

[5] P. Gautam, "Deciphering the Department-Discipline Relationships within a University through Bibliometric Analysis of Publications Aided with Multivariate Techniques," 2015 IIAI 4th International Congress on Advanced Applied Informatics, Okayama, Japan: 468-471, 2015. DOI: 10.1109/IIAI-AAI.2015.212

[6] Elsevier, Scopus. http://www.elsevier.com/online-tools/scopus

[7] N. J. Van Eck, L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, 84(2), 523-538, 2010.

[8] S-E. Clausen, Applied Correspondence Analysis: An Introduction. SAGE Publications, 1998.

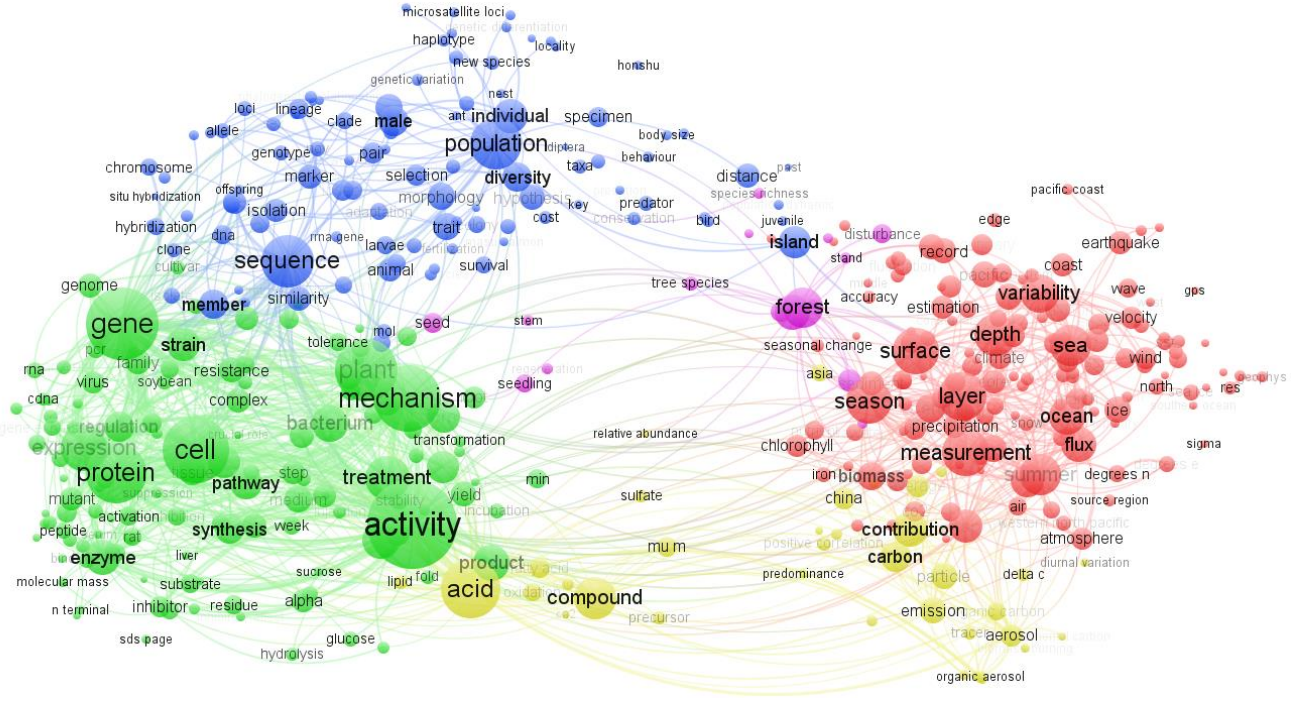

Fig. 4. Co-word map constructed from the frequency of co-occurrence of terms and noun phrases in the text corpus of articles and abstracts of 3,725 publications comprising the "earth & environmental sciences, agriculture and fisheries" area of the university. The 5-clusters solution, based on words co-occurring at least 20 times in the text corpus, coincide with clusters that could be loosely described as biodiversity (top), molecular biology and microbiology (the lower left), aerosols & particulate matters (the lower middle), forestry (center to left at the middle), and earth-ocean-atmosphere studies including earthquake and tsunami disasters (right). These clusters represent the most prolific research areas for the period 2009-2013. Map generated by Vosviewer version 1.6.3. [7]

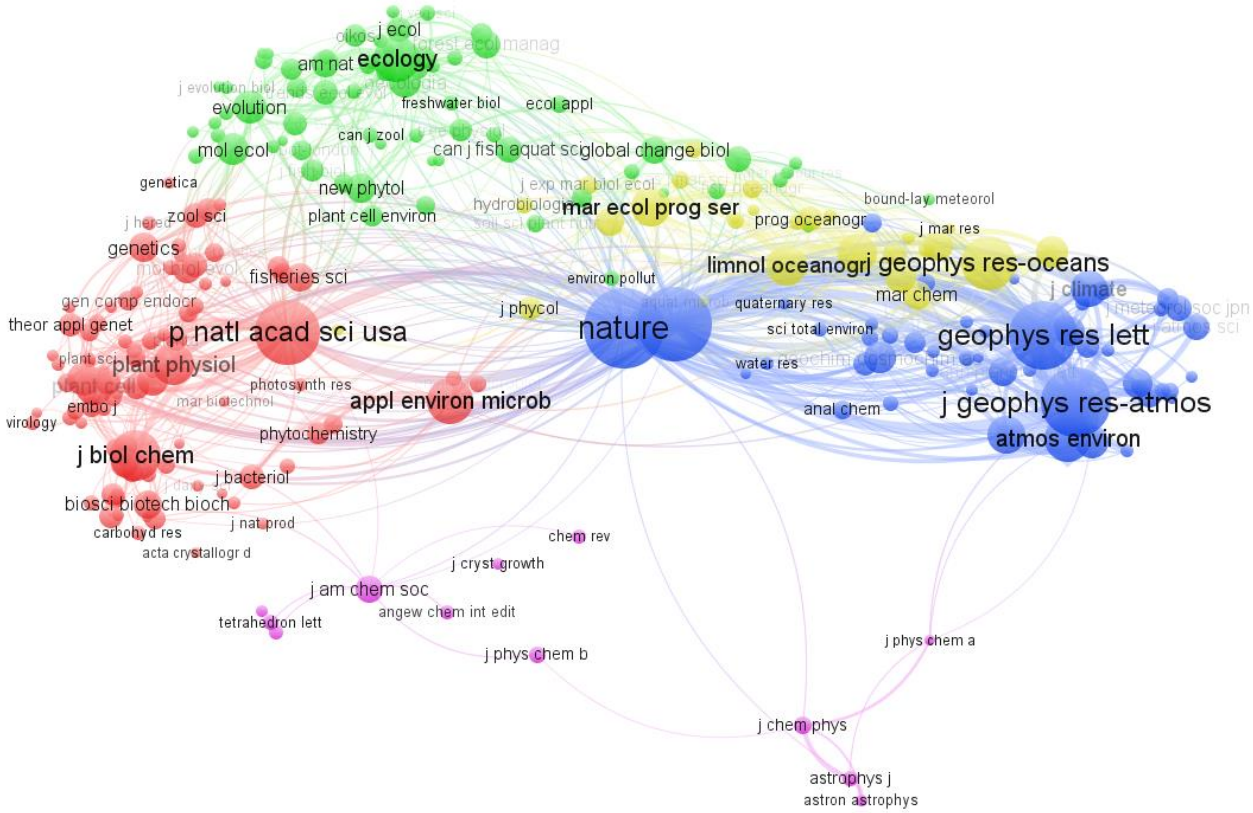

Fig. 5. Map depicting the co-citation relationship of sources (journals) as the units of analysis for the publications dataset used to construct the co-word map shown in the previous figure. The 5-clusters solution groups journals, cited more than 100 times each, based on their disciplinary focus: environment/ecology (top), microbiology, molecular biology and genetics (left), physical and chemical sciences (the lower middle), geophysics and geospheres (the middle right), and ocean science including fisheries (the right middle). Nature and Science (the large circle situated besides nature) known as the high impact science magazines are the most cited sources occupying the central and near neutral (connections to journal from almost all clusters) positions despite their greater affinity to and hence inclusion into the geophysics and geospheres cluster.