

**Upcoming concepts in a specific scientific discipline:  
an analysis based on a categorisation of the related terminology**

Ivana Roche<sup>(1)</sup>, Marianne Hörlesberger<sup>(2)</sup>, Beatrix Wepner<sup>(2)</sup>, Christine Louala<sup>(1)</sup>,  
Claire François<sup>(1)</sup>, Nathalie Antonot<sup>(1)</sup>, Georg Vorlauffer<sup>(3)</sup>, Edgar Schiebel<sup>(2)</sup>, Dominique Besagni<sup>(1)</sup>

<sup>(1)</sup>INIST-CNRS, 2 allée du Parc de Brabois, 54519 Vandoeuvre-les-Nancy Cedex, France

<sup>(2)</sup>AIT, Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna, Austria

<sup>(3)</sup>AC2T Research GmbH, Viktor Kaplan-Strasse 2, 2700 Wr. Neustadt, Austria

The aim of this work is to introduce a methodology allowing the analysis of the evolution of a specific scientific domain by means of studying its terminology extracted from related specialized international literature. The developed approach is based on an analytical methodology produced in the framework of the PROMTECH project, financed by the European Commission [BES07]. This methodology operates a selection of dynamic fields on the basis of the growth of their annual productivity, followed by a categorisation of their respective terminologies [ROC07], [ROC10]. In this paper, we introduce a global characterisation of these terminologies at the level of the fields allowing to compare their evolution and study their relationships defined in terms of terminology exchanges.

The data set is extracted from the PASCAL database that is specifically adapted to the purpose of our approach. It provides broad multidisciplinary coverage of scientific publications and contains nowadays about 20 million bibliographic records from the analysis of the scientific and technical international literature published predominantly in journals and conference proceedings. The PASCAL records benefit from an indexing by both keywords and thematic categories of a classification scheme assigned to each individual publication, either manually by scientific experts or automatically based on a content analysis. The queries operated in this work are based on this information. After a verification step done by a scientific expert, this terminology is employed in our analysis.

In this work, we are particularly interested on the research produced within the last 10 years in the domain of "tribology", the science and engineering of interacting surfaces in relative motion, including the study and application of the principles of friction, lubrication and wear. Tribology, lubrication and surface technology are key technologies and utmost importance for all branches in industrial production. Innovations in these technologies require the best possible fundamental understanding of the complex processes taking place at the interfaces of a tribological system. Controlling the macroscopic effects of a tribosystem, the correlations between macro-, micro-, nano- and sub-nano scaled progress have to be known as well. Therefore the Xtribology centre in AC2T Research GmbH is confronted with the challenges of a highly interdisciplinary field. The centre would like to generate holistic knowledge in tribology on the one hand. On the other hand, it has to deal with high level research in the areas of smart materials, surfaces and coatings, lubricants and lubrication systems, high-resolution wear measurement systems, simulation and modelling of friction and wear processes, develop and extend standardization, test methods and databases concerning tribological test data, promote the latest knowledge to optimise friction and reduce wear, raw material consumption and environmental loading in industrial processing, and, in addition, take into account social, environmental and sustainable aspects.

Looking for published data from 2001 to 2010, the database query produced a corpus collecting of about 20,000 bibliographic records distributed almost regularly by publication year. The corpus is at 99 % constituted of articles published in journals. The main language of these papers is English (93.5 %), followed by German (3.2 %) and French (2.5 %). The affiliation countries of the authors are mainly the USA (24 %), China and Japan, each one with 11%, and UK, Germany and France, each one with 10%. The indexing present in the bibliographic records gathers more than 15,000 different keywords.

The most representative fields are identified by the distribution of the classification categories present in the corpus records. To select the dynamic fields, we are looking for fields showing a steady and consistent development over time. This guarantees the selection of reliable and interesting fields that are not "one-hit wonders" and that give the promise of growth in the future.

For this purpose, we have defined two types of indicators measuring the evolution of the annual productivity of each field: those based on a growth index and those based on the Sharpe Ratio. The growth index is the ratio of the number of publications in the most recent year to the number of publications in the oldest year. This indicator neglects any up and down in the development over time and therefore abstracts from all the changes in the years between. It is a straightforward comparison of the two endpoints of the period under observation. Based on the annual average growth rate, it is possible to take yearly changes into account. As the average growth rate can be highly influenced by outliers, we introduce the so-called Sharpe Ratio, an indicator which is generally used in the analysis of financial markets. This indicator takes the stability of growth into account, so that data with strong growth on average, but large annual variation are devalued. Each indicator emphasises a different aspect and their values are not directly comparable. So we have to take into account all of them for the assessment of the considered fields. For that purpose, we applied a simple ranking of the fields by the result of each indicator. The sum of these ranks build a new indicator which then can be ranked again and leads to a meta-ranking or a composite indicator of the ranks, implying the assumption that each indicator has the same weight in the composite.

After the identification of the most representative fields in the studied domain we introduce the methodology based on a diffusion model approach to analyse in-depth their evolution. In previous works, we presented its detailed description [SCH10], [ROC10]. This approach evaluates the term status in a considered field by measuring its degree of diffusion.

The diffusion model is based on the assumption that new findings in a research field are published in articles. Keywords that describe the innovative results occur in the first stage in an unusual manner. In the second stage the research intensifies and established keywords are used. In later stages, the results cross the disciplinary barrier by diffusing to other research fields where they follow a similar evolution cycle. Consequently, the diffusion status is obtained by the calculation for each keyword of a diffusion degree that can be either “unusual”, “established” or “cross-section”.

Two pragmatic approaches are successively employed to realize this categorisation. Firstly, the so-called *Home Technology terms (HT terms)* are defined. We assumed keywords which are specific for a field occurred with a higher probability in that field rather than in others. The probability is defined by the frequency of one term in a field divided by the number of articles in this field. For a term, the field with the highest probability is declared to be its *Home Technology*. So after this assignment we obtain for each field the list of its *HT terms*. Therefore the complete terminology associated to a field consists of the union of its *HT term* list and the set of terms imported from the other fields.

Secondly, we use the Gini index [GIN08], a measure of inequality in a distribution, to define the diffusion degree of a keyword in a field. The Gini index varies from 0 to 1, for which 0 means a completely uniform distribution and indicates that the term occurs in all the considered fields. Conversely, a Gini index of 1 tells us that the term is very specifically limited to the only field where it appears.

For each field, the *HT terms* are distributed in 4 categories:

- the terms occurring once,
- the terms with at least 2 occurrences and whose Gini index is lower or equal to a fixed threshold, considered as **cross-section**,
- the remaining terms whose relative term frequency (rtf) is greater than a fixed threshold, considered as **established**,
- the last terms, considered as **unusual**.

The *HT terms* occurring once are analyzed by introducing the notion of term age based on the publication year of the article indexed by this term. This diachronic approach allows to distinguish the “old” concepts appearing in the beginning of the studied period from the very “new” ones appearing in its latest years.

The other three keyword lists are analysed in-depth by a scientific expert in order to characterize the evolution of the fields:

- unusual terms index few publications. New terms and terms well-known in other fields form a set of strongly exotic terms,
- established terms can occur together with established methods, materials, tools, and applications from other fields. They begin to diffuse to other fields,

- cross-section terms, highly established, show a broad diffusion in other fields. This is the stage with the highest maturity.

In order to characterize the vocabulary used for each field, we define indicators: productivity, diversity, specificity, singularity, and normalized trade balance.

Let us consider: BN = number of bibliographic records in the field; KW = total number of keywords indexing bibliographic records in the field; HT = number of *HT terms* in the field; EX = number of *HT terms* exported by the field to other fields and IM = number of terms imported by the field from other fields.

The calculated indicators are:

- productivity = BN
- diversity = KW / BN  
the higher this value, the more diverse the terms used. Lower consistency in the terminology could indicate a young field. Conversely, a stable terminology of a field could express its establishment.
- specificity = HT / KW  
the lower this value, the bigger the part of the field terminology coming from abroad.
- singularity = [KW - (EX+IM)] / KW  
the weaker this value, the lower the number of so-called lonesome terms, neither exported nor imported, occurring exclusively in the field.
- normalized trade balance = (EX - IM) / (EX + IM)  
the range of values goes from -1 (the field exports none of its *HT terms*) to 1 (the field imports no term) and a zero value means a trade balance in perfect equilibrium.

We introduce also a measure of the field diffusion capacity: the average of the Gini index of all the *HT terms*. A Gini index of 0 means a completely uniform distribution of all *HT terms* and indicates that each HT term occurs in all the other fields. A Gini index of 1 tells us that all *HT terms* are very specific and occur only in this field.

Finally, a field impact is defined as a Hirsch index: a field has an h-index of X if X of its keywords is imported by at least X other fields. These X keywords form the h-core list of the field. We can then consider by analogy that this Hirsch index can help appraise the influence of the considered field on the others.

A cross analysis of these field indicators and the field terminology categorisation let us verify the convergence between these two approaches and contributes to a better understanding of the studied discipline evolution. The results produced in the context of this work are available on a web server to facilitate their assessment by scientific experts from the Austrian Competence Center for Tribology (AC2T).

[BES07] Besagni D., François C., Frietsch R., Hörlesberger M., von Oertzen J., Pretschuh J., Roche I., Schiebel E., Schmoch U. (2007) Final Report (Deliverable 06) for project PROMTECH - Contract N° 15615. 89 pages + 3 Appendixes

[ROC07] Roche I., François C., Besagni D. (2007) Les méthodes bibliométriques en soutien d'une approche expert dans la détection de technologies prometteuses, VSST'2007 – Veille Stratégique Scientifique & Technologique, Marrakech, 21-25 octobre 2007

[GIN08] Gini (2008). [http://en.wikipedia.org/wiki/Gini\\_coefficient](http://en.wikipedia.org/wiki/Gini_coefficient)

[ROC10] Roche I., Besagni D., François C., Hörlesberger M., Schiebel E. (2010) Identification and characterisation of technological topics in the field of Molecular Biology, *Scientometrics*, 82, 3, pp. 663-676

[SCH10] Schiebel E., Hörlesberger M., Roche I., François C., Besagni D. (2010) An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices, *Scientometrics*, 83, 3, pp. 765-781