FET_TRACES

Tracing impacts of the FET programme

Project number. 665083
Programme: H2020-FETOPEN-2014-CSA
Contract type: CSA
Start date of project: July 15, 2015
Duration: 28 months

Deliverable D 5: Determining the level of novelty of FET projects (LDA analysis)

Authors: Ulrich Schmoch (Fraunhofer ISI),
Bernd Beckert (Fraunhofer ISI)
Dissemination level: public
Deliverable type: report
Version: 1.0

Submission date: 19 September 2017

About FET_TRACES

FET_TRACES is a research project for the European Commission which analyses and measures the impacts of the research funding scheme “Future and Emerging Technologies Open” (FET Open and FET Proactive). Within the European research funding landscape, the FET scheme acts as a pathfinder for new ideas and themes for long-term research in the area of information and communication technologies and beyond. Its mission is to promote high risk research, offset by potential breakthrough with high technological or societal impact (see http://cordis.europa.eu/fp7/ict/fet-open/home_en.html).

In the FET_TRACES project we will investigate and measure direct and indirect impacts of these two schemes on the science and technology landscape and its perception by individual researchers who are potential proposers for FET Open and FET Proactive projects. Results from innovation research will be used to develop a targeted indicator set covering central aspects of the FET mission (novelty, trans-disciplinarity, innovation-ecosystem). For the data collection we use sophisticated impact assessment methods like bibliometrics, patent analysis and online surveys. In addition to the impact assessment we will analyze selected breakthrough-projects to find out about necessary components for “breakthrough”-research. The study will also include insights from FET-like funders on national levels in Europe.

Terms of use

This document was developed within the FET_TRACES project (see www.fet-traces.eu), funded by the European Commission within Horizon 2020, by a consortium consisting of two partners, the Fraunhofer ISI in Karlsruhe, Germany (coordinator) and AIT in Vienna, Austria.

Document history

Version	Date	Changes
1	18 September 2016	

Contents

1 Introduction

"Novelty" of an idea, concept, method or approach followed in a FET project is an important impact dimension for the assessment of the FET programme. However, it is difficult to conceptualise and measure "novelty", and in fact an adequate description requires a multi-perspective approach. Thus, in the FET_TRACES project we have defined eight different indicators to describe novelty in FET projects (see table 1). In this report, we document the results of the LDA analysis. Novelty-assessments based on other methods are documented in the bibliometrics deliverable (D 6.1) and the survey deliverable (D 7.1).

Table 1: Indicators to cover "Novelty" in FET projects

Indicator		*Method*
Novelty		
1	New topics: Percentage of FET projects dealing with topics that were not present in the scientific literature before	LDA-Analysis
2	Off-mainstream research ideas: Number of FET projects that were rejected before by other funding institutions	Survey
3	Uptake in science: Number of articles from FET-projects which are highly-cited by other researchers	Bibliometrics
4	Outstanding excellence: Major scientific awards received for FET-related research	Survey
5	Novel combination of approaches: Novel composition of the interdisciplinary consortia	Survey
6	New research avenues for established researchers: Number of researchers pursuing new research directions	Survey
7	Novelty of outcome: Answers of researchers about the novelty of their results	Survey
8	High impact scientific publications: Number of publications in *nature* and *science*	Bibliometrics

Source: FET_TRACES Deliverable D3: List of indicators, p. 5

2 Method

In order to determine the “percentage of FET projects dealing with topics that were not present in the scientific literature before” (indicator definition, see table above) we used an adapted version of the Latent Dirichlet Allocation (LDA, see Mund 2014: 77ff) analysis, a bibliometric method which determines levels of similarity between scientific publications.

In a first step, for all our 224 FET projects in the sample, an early relevant scientific publication had to be determined. Here, we mainly used the early FET publications with the highest citation rates, because we have seen that these publications usually describe the basic concepts on which the projects build on (see bibliometrics deliverable 6.1, p. 17f).

In order to limit the search to a reasonable level, only publications with at least one identical keyword to the keyword of the analysed publications were examined.

For the remaining publications the similarity to publications in the period 5 years before the analysed publication were checked by bibliographic coupling, a method which is based on reference pattern analysis (see Glänzel and Czerwon 1996). Hereby, the measure for similarity ranges from 0 (no similarity) to 1 (very high similarity).

Whereas bibliometric coupling allows for a relatively good automated analysis in the lower levels of similarity, a specific determination of similarities becomes necessary as soon as a certain level is reached: Through experimenting with a smaller sample we determined the measure from which similarities need to be checked separately at 0.8.

In the following, all publications with a novelty level of 0.8 or higher were checked by an examination of the text-based similarity of the abstracts by the method of TF Idf[1]. For this analysis, the frequency of the words in the abstract is compared with a lower weight of less significant words such as ‘method’. The inverse frequency is taken as weight. Thus, a word appearing in many documents gets a lower weight, as it points not to the specific content of a text. Words appearing in fewer publications are considered as specific terms and get a higher weight.

In a last step, all publications with a novelty level of 1.0 had to be checked manually in order to determine whether publication is actually or almost identical with the FET publication.

---

1 TF Idf (Term frequency Inverse document frequency, see https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/).

This method, which is a specific combination of bibliometric coupling, TF ldf and manual cross-checks was developed in the course of attempting to implement the LDA analysis to our context.

It turned out that the originally proposed LDA method did not work in our context. The originally proposed LDA method would have analysed the novelty of new topics represented by a set of documents. However, even after trying various clustering strategies, this method overtaxed the available computing power of our servers and it did not lead to useable results. As we were in fact interested in determining the novelty level of single publications (and not topical clusters) we adjusted the method accordingly and applied the step-by-step procedure as described above.

3 Results

For 185 projects or 83% of our sample, a similarity below 0.8 was found. This means that the largest share of the analysed projects deal with ideas, concepts or approaches which were not present in the scientific literature in the 5 years before the first relevant FET publication. These projects thus could be labelled as "radically new" (see table 2 and figure 1).

For 9 projects (4% of our sample), a similarity of the bibliographic coupling of 0.8 or 0.9 was found. This category could be labelled "New solutions to similar debates".

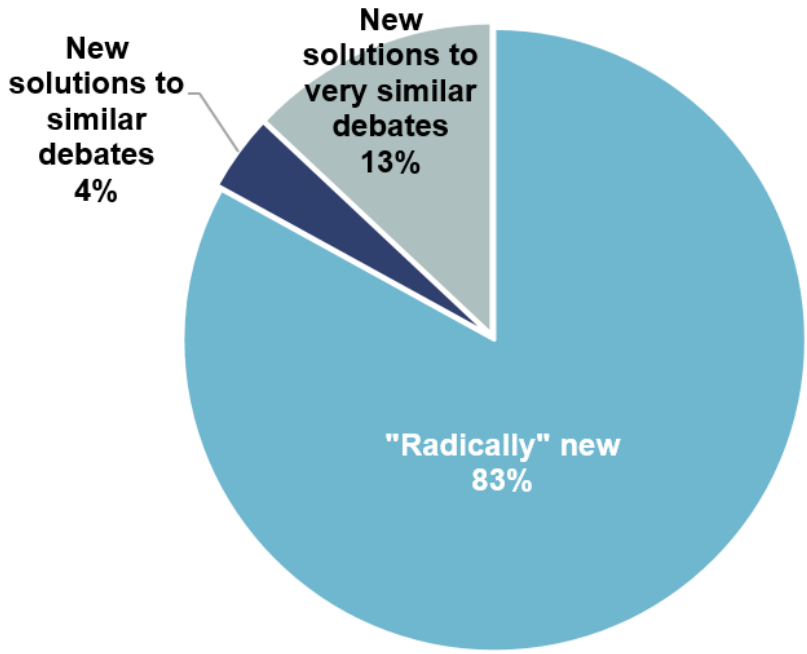
And for 30 projects (13% of our sample) a similarity level of 1.0 was found. These projects relied on existing strands of research but nevertheless introduced new solutions. At the level of 1.0 (high similarity) we often found not only one similar publication but several and in one case even nine publications with similar content. As described in the section above, we manually cross-checked all publications with this high value of similarity in order to identify duplications or almost identical publications. However, this manual similarity check which was based on the abstracts did not lead to identical pre-publications. Thus the text-based value of 1.0 indicates a high similarity, but not identity.

Table 2: Novelty in our sample of FET projects

Novelty level	**Similarity of precursor publications**	**percentage**	**# of projects**
Radically new	<0.8	83%	185
New solutions to similar debates (>0.8)	>0.8	4%	9
New solutions to very similar debates (1.0)	=1.0	13%	30
Total		100%	224

Source: FET Traces 2017, LDA analysis

Figure 1: Results of the novelty analysis



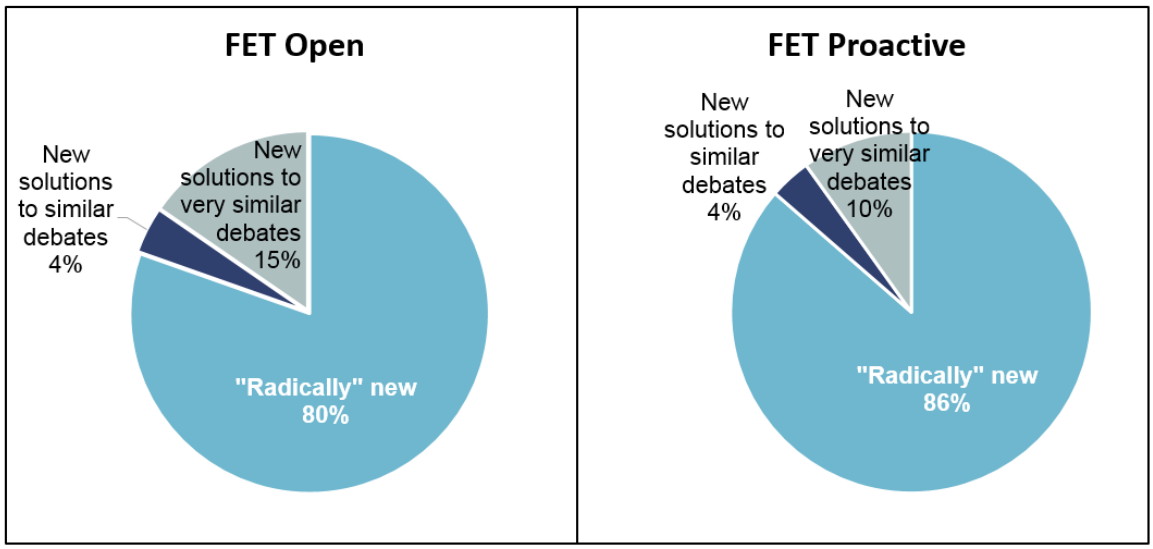
Source: FET Traces 2017, LDA analysis

Taking together the two categories which relied on existing debates, it can be said that in our sample there is a smaller share of projects (13%+4%=17%) in which the idea pursued was not completely new, but in which researchers started from an existing debate how to solve a certain problem. One or several similar publications in fact existed before, but the FET publication brought in new approaches.

In order to better understand the difference between the two “similar debates”-categories it could be instructive to introduce a term from patent law: Here, we speak of a “high inventive step” if a solution was found for a long-standing problem. In our context, it could be said that projects for which we found similar precursor documents with a similarity of 0.8 or 0.9 still had a relevant distance to the final solution whereas projects with precursor documents with a text-based similarity of 1.0 were closer to the final solution, but did not achieve the definite breakthrough.

A separate counting of novelty levels of FET Open and FET Proactive projects reveals that both, FET Open and FET Proactive projects, induce novel ideas to a very high degree (80,4% of FET Open and 86,4% of FET Proactive in figure 2).

Figure 2: Novelty levels of FET Open and FET Proactive



Source: FET Traces 2017, LDA analysis

Interestingly, FET Proactive projects less often build on very similar debates (10% in figure 2) than FET Open projects (15% in figure 2). In turn, this means that FET Proactive projects show a slightly higher level of novelty than FET Open projects. This finding seems counterintuitive at first glance because topics can be chosen freely in FET Open projects whereas FET Proactive projects usually relate to topics which had been specified in specific calls. Our interpretation of this finding is that the specific calls for FET Proactive indeed addressed new topics which were not present in the scientific community at that time, thus starting something new, whereas FET Open projects were more often tied to existing debates. However, it has to be noted that these findings are only valid on a comparative level and that the difference between the figures is relatively small.

4 Summary

According to our analysis all projects in our sample worked with new concepts, the majority introduced radically new concepts, some projects brought in new solutions to debates of the scientific community how to approach a certain problem.

In this context, it has to be kept in mind that novelty is only one dimension of success. In general, it requires many years for a really new scientific concept to get transformed into a marketable technology. For realizing this target, many other factors must come in.

5 Critical review of the method

As mentioned in the method-section, it was not possible to apply the LDA-method as originally planned. Instead, a multi-step-method was developed which includes two automated processes and a manual identification of departure publications as well as a manual cross-check of a sub-sample. It turned out that the project reviewers who found our approach to determine the level of novelty “very ambitious” were right, as we have spent many efforts to make the novelty analysis work.

Summarizing the experiences with the solution we finally found it can be said that the method found is well suited to identify levels of similarity in large datasets without requiring extreme computing power. In fact, the reference pattern analysis as well as the abstracts based similarity analysis did not require excessive computing power and could be done in relatively short time.

The constraints of this method, however, are the manual steps of identifying relevant publications for the analysis and the cross-check of almost identical publications in the final step. These working steps require high intellectual input and are very time-consuming.

Other restrictions relate to the explanatory power of the method. Whereas we have found a measure to separate the “radically new” ideas from the ideas which relate to existing debates, the large share of “radical” new ideas raises the question whether research generally deals with novel ideas. As such, there is no final answer to the question whether FET projects are in fact “newer” than projects funded by other funding schemes. However, a level of novelty as high as 83% is a very remarkable result and indicates a high level of achievement of one of the most important impact dimensions of the FET programme.

6 Literature

Glänzel Wolfgang; Czerwon, Hans-Jürgen (1996): A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. In: Scientometrics, October, Vol. 37, Issue 2, pp 195–221.

Mund, Carolin (2014): Identification of Emerging Scientific Topics in Bibliometric Databases. PH. D. Thesis at the Karlsruhe Institute of Technology, https://publikationen.bibliothek.kit.edu/1000042107