Papazoglou, Theodore, "IT-Based Approaches in Support of ERC's Mission to Support 'Frontier Research': First Experiences"

IT-based approaches in support of ERC's mission to support "frontier research": first experiences

A contribution to the NSF-OISE Workshop "changing the Conduct of Science in the Information Age", November 12, 2010

The present note is addressing one of the main objectives of the workshop from the research funding agency perspective, namely the need to assess whether the projects for which financial support is requested, as well as those already funded, are compatible with the strategic aims of the organisation. The note shares, as example, the methodology that is currently followed by the European Research Council (ERC) in order to be informed on whether, firstly, its review system is successful in identifying proposals that address "frontier research" and, secondly, the projects that are being funded do correspond to new and emerging research areas. The underlying concepts of this approach is that information related to the proposals/projects is in principle readily available via the presence of the Principal Investigator and his/her research group on the WWW and the research literature (publications, conferences, data etc.), and consequently there is no real need to "harass" him/her with additional requests to the standard reporting obligations. On the other hand there is a plethora of (mainly bibliographic) tools that could help identify "emerging fields" of research. The challenge therefore will try to match ("correlate") these two categories of data and use this information to assess the implementation of the scientific strategy. To note that the purpose of this exercise was not to measure the direct structural impact of ERC-funded activities to areas such as job creations, build-up of infrastructure etc. This is expected to be mainly done via the analysis of the periodic reporting of the grantees and the assistance of a similar set of ancillary studies.

The ERC work programmes 2008 and 2009 made provisions to fund Coordinated and Support Actions (a 7th Framework Programme term to describe ancillary to the main instruments projects/initiatives/studies) to support the monitoring and evaluation strategy of the ERC Scientific Council. The calls were launched as "open call for proposals". The reason for this decision was the intention to explore new approaches in the evaluation, as suggested by the relevant scientific community. Two projects that use bibliometric tools were selected in the framework of this exercise:

- **DBF: Development and Verification of a Bibliometric model for the Identification of Frontier research** that started in October 2009 (3 years project, ARC systems research and Institute for Scientific and Technical Information-CNRS Nancy), aiming to provide a bibliometric monitoring of the peer review process of the ERC grant schemes. Particular interest is devoted to the extent the grant applications fulfil attributes of frontier research and the influence of these attributes on the decision of the panels. For this purpose, bibliometric parameters will be elaborated and applied on the relevant information available in the grant applications as well as in the relevant publications authored by the applicants prior to their submission of their grant application: *Novelty* (citations, "recentness", link to ERACEP-see below); *Risk* ("Market"-Share), *Pasteuresqueness* (presence of industry); *Interdisciplinarity* (variation of evaluation panels proposals were submitted).

- **ERACEP: Emerging Research Areas and their Coverage by the ERC-supported protects** that started in October 2009 (3 years project, Fraunhofer Institute Systems

and Innovation research and Leuven University – Faculty of business and economics), attempting to identify emerging research areas and analyse to what extent the ERC grants cover and contribute to these research areas. It intends to investigate how ERC is performing in respect to its basic mission: "stimulate scientific excellence by supporting and encouraging the very best, truly creative scientists, scholars and engineers to be adventurous and take risks in their research". The project uses the following methodology (extract from ERACEP's 1st periodic report): "*A set of ISI Subject Categories in the sciences, social sciences and humanities with remarkable growth in the last decade are defined. Twenty categories have been selected to undergo further structural analysis; the objective is to identify new and/or emerging topics within these subject matters. In particular, 13 fields have been selected from the sciences, 5 from the social sciences and 3 from the humanities. The underlying data have been retrieved from Thomson Reuters' Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI) for the period 1998-2007. Cluster analysis in bibliometrics is traditionally based on both citation links (bibliographic coupling, cross-citation, co-citation analysis) and textual links (co-word analysis, term representation). Both approaches have advantages and shortcomings. The main advantage of citation-based methods is their discriminative power. This is contrasted by a serious disadvantage: Citation-link matrices are extremely sparse and citation-cased methods tend to "underestimate" links among documents. Furthermore, citation links generate binary measures which are based on value 0 or 1 according as there is a citation link between two documents or not. By contrast, text-based measures are based on term frequencies in documents, which as such provide a natural weight underlying the similarity/distance measures used for the analysis. Link matrices are furthermore less sparse than their counterparts in the citation space. These advantages are cancelled out by two serious problems: The lower discriminative power, which results in "overestimating" links among documents and the dimensionality problem. At least the latter problem can be compensated by Singular Value Decomposition (SVD) or directly by Latent Semantic Indexing (LSI), which uses the first-mentioned algorithm*".

Some initial technical obstacles as well as legal constraints will be reported during the workshop in order to assist the participants to appreciate that even when "enabling technologies" are in the disposition of a research funding agency, the challenges are still significant.