2015 Manchester Forum on Data Science, Tech Mining and Innovation
29-30 October 2015

Full Paper Presentations

Title: Funding data: Collection, coding, and caveats

Author(s): Grassano N., Hopkins M., **Rotolo Danile**.

Author Affiliations: University of Sussex

Corresponding author e-mail: d.rotolo@sussex.ac.uk

Abstract:

The acknowledgements of scientific publications are a rich source of data that can inform us about how science is conducted. One application of this data is to link research inputs (grants) with research outputs (publications), in order to reveal details of the funding landscape without relying on thereports of individual funders. Yet, acknowledgement data do not conform to a single recognised standard. Instead they are reported in an unstructured manner, which complicates the identification and extraction of relevant funding data. Some commercial data providers provide extracted data on funders from publications but how well do they collate and aggregate data that relates to particular funders? Indeed how should such data be collected, extracted and coded to give an accurate picture of the funding landscape in a given field? To address these questions the paper proposes a method and a series of guidelines for the collection and processing of acknowledgements sections contained publications in order to extract and codify relevant data on research funding. We then compare our findings with results from alternative datasets provided by searching MEDLINE/ Pubmed and ISI Web of Science. We apply our approach to the field of cancer research in the UK. We identify a sample of 7,510 publications produced in the year 2011 that involved at least one author affiliated to a UK research host organisation. For each publication in thesample, we extracted the funding data from the acknowledgement section. We then developed a number of guidelines for funding data collection and coding. We describe the rationale for our approach with associated caveats and limitations. We use the processed data to explore patterns of co-funding among funding organisations in the setting of cancer research in the UK. The analysis shows the UK contributes 6.9% of global research publications in the field ‘neoplasms’ (a term that captures research related to cancer and pre-cancerous growths) that year. The results are informative about the main features of the UK research funding landscape. Funding sources were acknowledged in half of publications, and revealed that almost two thirds of these papers benefited from multiple funders with a mean of 3.3 funders per publication. The analysis also reveals not only the major UK funders but also shows the important role of overseas funding in supporting the work of UK authors, through the funding of their co-authors – just under half benefited from overseas funding. The role of industry is was shown to be considerable, with almost afifth were supported by firms in the UK or elsewhere. Finally this bottom up approach to identifying research funders reveals the contributions of myriad small charities to cancer funding. We also compare our data with that obtained by querying on the same set of publications in MEDLINE/PubMed and ISI Web of Science (WoS). Results of the comparison demonstrate the poor coverage of funders in

MEDLINE/PubMed, and the incomplete aggregation of funders in ISI Web of Science, relative to results obtain through manual collection, cleaning and aggregation of the data using a team approach, coordinated with the guidelines we describe in the paper.

Title: The SMS infrastructure for Science and Innovation Studies – a demonstrator

Author(s): Peter van den Besselaar, Ali Khalili, Al Issidrou, Antonis Lizou, Frank van Harmelen

Author Affiliations: VU University Amsterdam, Netherlands

Corresponding author e-mail: p.a.a.vanden.besselaar@vu.nl

Abstract:

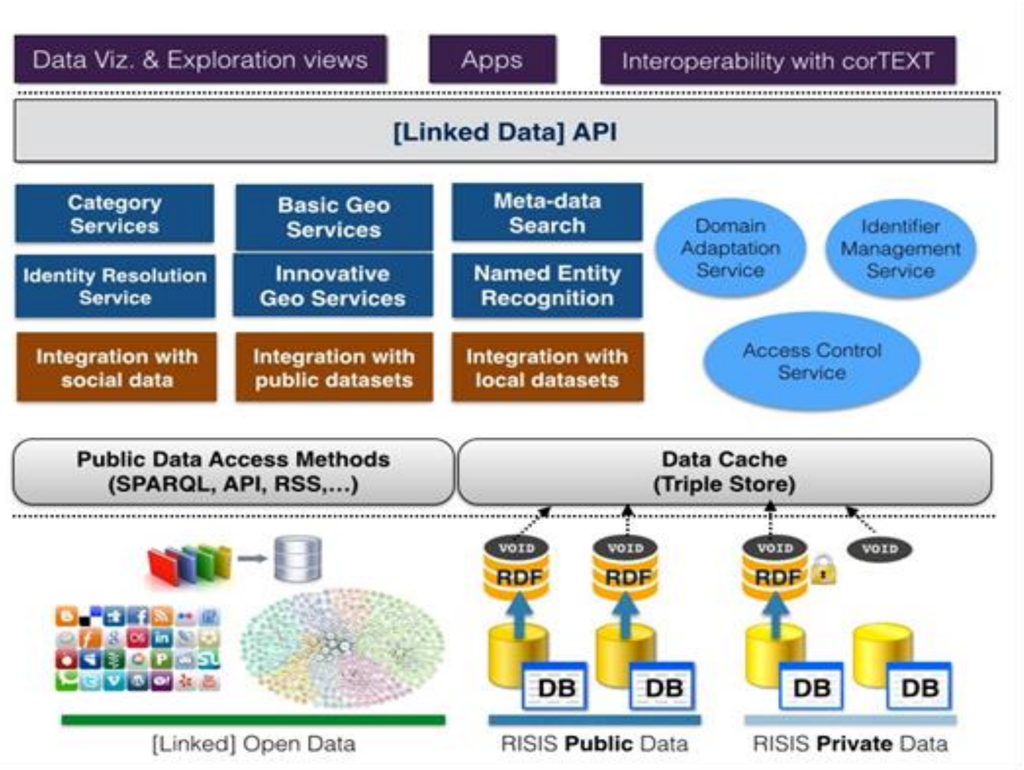
The Semantically Mapping Science (SMS) platform supports access to heterogeneous data on science and innovation, and supports combining, integrating and analyzing those Data. SMS is one of the facilities in the RISIS project, a distributed data infrastructure for research and innovation dynamics and policy studies.

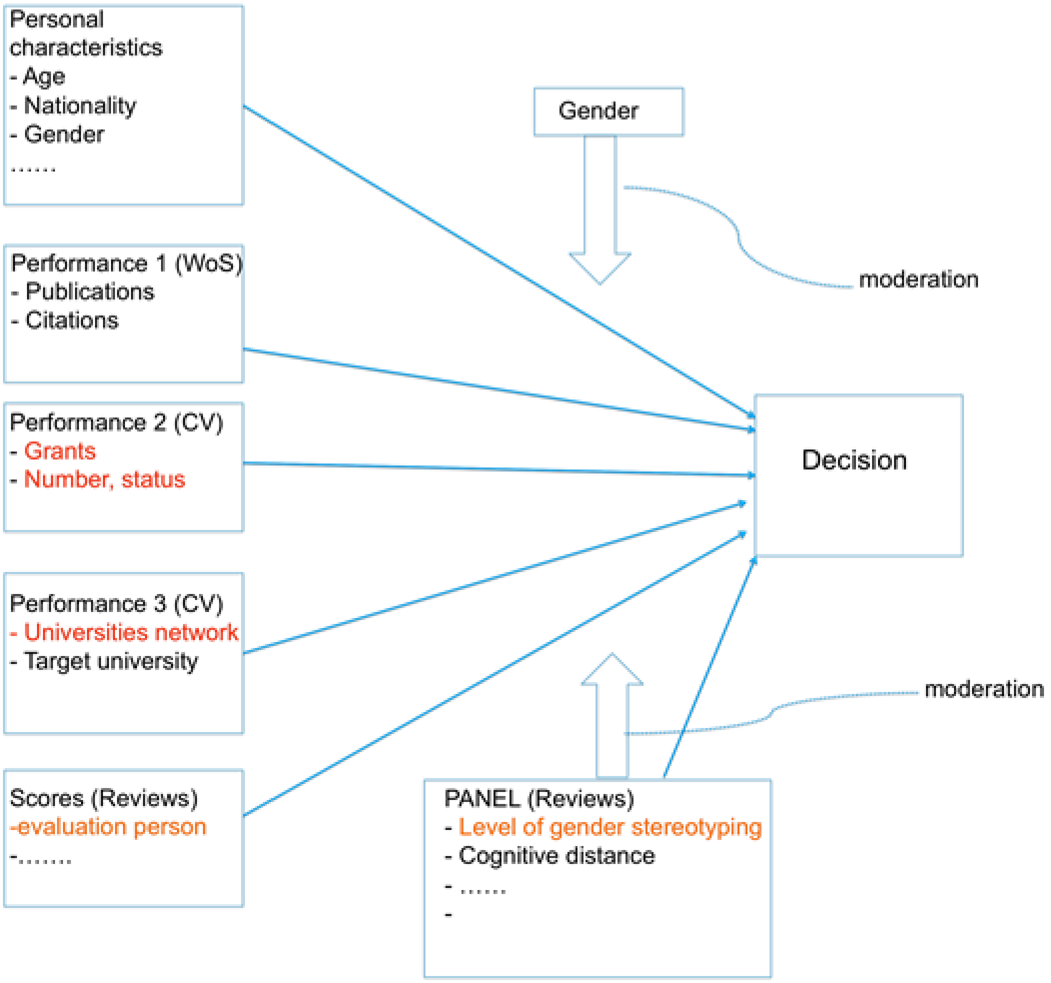
Integration of distributed and heterogeneous data is a crucial aspect of the SMS platform, as this helps to ask more complex research questions. To facilitate data integration, SMS employs a set of services to access, transform and process data from available public or private data repositories.

We will present the basic architecture, and a demonstration case to show the possibilities.

The platform supports preprocessing, and after that translation of data into RDF data representation model for efficient data integration and reasoning. Then a series of services can be deployed, in order to search the datasets (via the metadata system), to harmonize categories and attributes in specific formats, such as geo---location, detecting patterns (entities) in unstructured text and entity disambiguation. This brings the data into a form (i.e. Linked Data) that allows integration. To support integration, a series of Linked Data APIs are implemented. These APIs allow users to ask simple or complex queries (by combining multiple APIs) to get relevant data out of the platform (drawing on various sets) that are needed for the analysis. On top of the API level, a series of analytical and visualization tools can be deployed (or tailor---made by the user). A main focus is on open (Web accessible) data. The platform also supports the use of traditional bibliographic, administrative and research databases. Access to and integration of proprietary data of course require access rights. The platform can handle open data, and data with access restrictions. Figure 1 shows the basic architecture.

The use case shows a variety of datasets and queries that result in a dataset needed for a specific research project, focusing on gender bias in research councils' decision making: (i) Is the process characterized by gender bias, and (ii) what non---academic factors influence the grant decision?. Figure 2 shows the model of that study. For the demonstrator, we use confidential data from the council (application---data, CV---data, evaluation reports, administrative data), proprietary data (WoS) and open data (ETER, Leiden Ranking, DBpedia), preprocessing tools, and queries to generate the required dataset for the research project.


Data Viz. & Exploration views
Apps
Interoperability with corTEXT
[Linked Data] API
Category Services
Basic Geo Services
Meta-data Search
Domain Adaptation Service
Identifier Management Service
Identity Resolution Service
Innovative Geo Services
Named Entity Recognition
Integration with social data
Integration with public datasets
Integration with local datasets
Access Control Service
Public Data Access Methods (SPARQL, API, RSS,…)
Data Cache (Triple Store)
VOID
RDF
DB
[Linked] Open Data
RISIS Public Data
RISIS Private Data


Personal characteristics
- Age
- Nationality
- Gender
……
Gender
moderation
Performance 1 (WoS)
- Publications
- Citations
Performance 2 (CV)
- Grants
- Number, status
Decision
Performance 3 (CV)
- Universities network
- Target university
moderation
Scores (Reviews)
-evaluation person
-…….
PANEL (Reviews)
- Level of gender stereotyping
- Cognitive distance
- ……
-

Title: Assessing Connections between Cognitive Science and Education

Author(s): Jan Youtie, Gregg Solomon, Alan L. Porter, Stephen Carley

Author Affiliations: Georgia Institute of Technology

Presenting author e-mail: jy5@mail.gatech.edu

Abstract:

Education Research and Cognitive Sciences have been observed to be working on the same topics in parallel, but using different terminology and methods. Encouraging greater connections between Education Research and Cognitive Science could yield benefits, and thus, beginning in the late 1990s, studies and funding programs have supported such cross-fertilization. However, there are substantial barriers to knowledge sharing across disciplines, which have been widely discussed in the literature on multidisciplinarity. This study addresses the extent of connections between these two fields by examining the level of knowledge sharing, as proxied by citations to "out-field" articles, i.e., Education Research citations to Cognitive Science articles, and vice versa. We analyze out-field citations of articles from the Web of Science in journals representing these two areas. We examine out-field citations between journals in these two fields in 1994, 1999, 2004, 2009, and 2014, years chosen to obtain snapshots before, during, and after the aforementioned studies and funding programs and we track the extent to which the communities and, more particularly, their literatures have come into systematic contact. Results suggest that the extent of cross-citations between the two fields significantly increased since 2004. Education Research articles were two-to-three times more likely to cite Cognitive Science articles since that time. Moreover, the presence of several specialized fields—Educational Psychology, Learning Technology and Human Computer Interaction, and Applied Linguistics—appear to serve as "border communities" in that they attract citations from both Education Research and Cognitive Science.

Title: How international is internationally collaborated research? A bibliometric study of collaboration networks of Russian surname holders

Author(s): Maria Karaulova, Abdullah Gök, Philip Shapira

Author Affiliations: MIoIR

Corresponding author e-mail: maria.karaulova@postgrad.mbs.ac.uk

Abstract:

The rapid growth of international collaboration among the developed as well as the developing countries has become increasingly recognised as an important capacity-building factor of domestic research indicating the increase in research quality (Bornmann et al., 2015). Recent studies have increasingly studied volumes (Wagner and Leydesdorff, 2005), patterns (Shapira and Wang, 2010) and effects of international collaboration.

However, little attention has so far been paid to the principle of international collaboration since Katz and Martin (1997) problematised the conventional approach of measuring collaboration using co-authorship. We highlight potential problems that emerge when international collaborations are used as a proxy to measure international cooperation. It is tacitly understood that the increasing volume of international collaboration is meant to indicate exchange of ideas, knowledge and technologies between the two national systems. This approach has become central in science policy for countries that heavily embedded in international research networks, such as South Korea (Park et al., 2014; Yoon and Jeong, 2015), where bibliometric indicators of international collaboration become important for policy decision making.

This paper tests a hypothesis that a significant share of international collaborations of a national research system occurs with researchers previously affiliated with this system. As globalisation processes intensify, more scientists, especially in highly sought after domains of emerging technologies, become mobile and choose to relocate abroad of their original countries. At the same time, they may maintain links and even affiliations of countries they previously lived in.

This research uses typical Russian surnames active in nanotechnology research as the basis of inquiry. As well as a bibliometric analysis of 33,538 publications, this research is also informed by qualitative fieldwork conducted in Russia, USA, Germany and the UK. The case study was selected on the premise of very low mobility to and from the Soviet Union during the larger part of the 20th century, which limited Russian surname distribution outside Russia. After the freedom of movement was reinstated in the 1980s, Russia and other post-Soviet countries experienced exodus of population, including significant shares of research personnel (Graham and Dezhina, 2008).

A two-step procedure is used in this research. At the first stage, a lexicological query is constructed on the basis of common patronymic suffixes of Russian surnames (Unbegaun, 1972). This query is then used to map the geographic dispersal of Russian surname holders. As a second step, the resulting data is compared with the pattern of international collaborations of domestic researchers in Russia (Karaulova et al., 2015). The focus on emerging technology, such as nanotechnology, provides a selection of frontier research for this study.

Surname data has been used in bibliometric analyses to determine contribution of recognisable ethnic groups that are distributed across many nation states, such as Jewish surname holders, to the development of particular discipline (Kissin and Bradley, 2013; Kissin, 2011), to determine effects of inter-ethnic collaboration on quality of publication (Freeman and Huang, 2014), or to highlight the

contribution of ethnic and gender minorities in research publications (Lewison and Markusova, 2011; Lewison, 2001; Webster, 2004). Such studies take a technology-level, rather than a country-level focus, because they assume wide geographic dispersion of ethnicity-specific surname holders. However, the literature utilising surname data has not shed as much light on international collaboration and cooperation.

This paper contributes to further understanding of international collaboration and cooperation by exploring what the notion of 'international collaboration' actually means in the globalised age where researchers of various backgrounds are distributed across the globe. It also employs the surname-based approach to the area dominated by research based on geographic affiliation of publications, thus bringing methodological novelty to the field.

Title: Analysing the Influence of Renewable Energy R&D Projects on Local Clusters in Europe

Author(s): Jaso Larruscain-Sarasola, Rosa Río-Belver Ernesto Cilleruelo-Carrasco and Gaizka Garechana

Author Affiliations: University of the Basque Country, Spain

Corresponding author e-mail: jaso.larruscain@ehu.es

Abstract:

The objective to increase the percentage of Renewable Energy (RE) in the European energy mix to 20% by 2020 has caused the boost of a new multi-discipline industrial sector. The need to incorporate a greater percentage of different technologies (including not so well-developed ones such as wave or tidal energies) within the new structures of energy generation and distribution in cities makes local participation increasingly important in terms of industrial development, creating new organizational structures such as local clusters or innovation networks. Whilst national and regional RE clusters have been studied in detail by the scientific community, no specific studies have been carried out at a local level where clusterization is an industrial hallmark and its success is largely determined by the growth potential of its SME industries. The European Union finances most R&D collaborative technological projects in Europe in RE, characterized mainly by the slow innovation cycles, the long lead time ventures, the relatively weak position newcomers with a high percentage of public support and increasing regulatory constraints. It is essential to analyze in detail the relationship between the roadmap of RE knowledge of RE technology (at NUTS3 level) generated through the information from R&D projects and the potential development of new local clusters. For this purpose, 4,772 projects and 10,786 partners for Wind, Biomass, Solar, Geothermic and Sea sectors were identified using Text Data Mining process (Vantage Point) over CORDIS database for the period of 2000-2013. Then, a new analytical approach of Social Network Analysis (with Pajek) is developed to examine 90,058 relationships, focused mainly on the potential influence of partners on the network. The results show that R&D projects influence on the likely development of local RE clusters, although to varying degrees depending on the heterogeneity of partnership and the status of development of the subsector

Title: A Workflow for Doing Tech Mining in Python

Author(s): Scott W. Cunningham

Author Affiliations: Delft University of Technology, Netherlands

Corresponding author e-mail: S.Cunningham@tudelft.nl

Abstract:

Can there be a standard process and set of information products for doing tech mining? In this talk we compare and contrast a standard rational process of tech mining with a number of emerging perspectives on performing data science, knowledge discovery, and data mining studies. Any such process is changing in at least three ways. The analyst is increasingly working in a mixed environment where some parts of the work can be template and reused, and other parts need new and creative input. Second, there are a greater variety of parties now involved in the mining process than ever before. Third, there is a greater variety of information products produced as intermediate products from the tech mining process. The result is that tech mining processes (as experienced by real users) are more open and more porous than ever before. In Python, a major language for data science, a new and emerging consensus of data scientists points towards the creation of notebooks as a functional means of addressing an increasingly dynamic and porous operating environment. These Python notebooks are self-documenting, and are easily portable from user to user. The notebooks also easily adaptable to new sources of data, creating a general and reproducible template for analysis. These Python notebooks may also contain rich and interactive animations of data, permitting effective visual communication. While discussing solutions for tech mining workflow, this talk will also give a quick, but informative overview over major analysis and visualization packages in Python including scipy, networkx, matt plot scikit learn, and NLTK.

Title: Novel Approach for Depicting Emerging Technologies: The Case of Cloud Computing

Author(s): Iñaki Bildosola, Rosa Río-Belver, Ernesto Cilleruelo

Author Affiliations: University of the Basque Country, Spain

Corresponding author e-mail: inaki.bildosola@ehu.eus

Abstract:

This work aims at proposing a novel approach to gathering and structuring information concerning an emerging technology, generating a relevant profile, identifying its past evolution, forecasting the short and medium-term evolution and integrating all of the elements graphically into a hybrid roadmap. The approach combines four families of technological forecasting methods, namely: Statistical Methods in terms of Bibliometrics and Text Mining; Trend Analysis in terms of emerging keywords and clusters; Descriptive Methods in terms of Technological Roadmapping; and Expertise. The approach consists in 8 specific steps, from the retrieving of the raw data and cleaning it, to the final reconstruction of the outcome based on the expertise. Thus, steps one to three are focused on obtaining the first main outcome of the approach: a relevant profile of the technology being depicted. This will allow answering questions such as who/where/when/how are researching on a specific technology. The second part of the approach, steps four to eight, aims at generating an ontology of the technology and finally obtaining a technological roadmap where the past, the present and the future evolution are represented. This will allow realizing which sub-technologies within the main technology are already in a mature stage, which are a dead end and which can be those which will lead the field in a medium-term future. The approach make use of Science Citation Index (SCI) and the Scopus database tools in order to retrieve and filter the data; VantagePoint in order to clean the data and perform text mining in terms of listing, clustering and discovering emerging terms; R programming language to perform text mining and combining the results obtained from VantagePoint. The whole approach has been applied to Cloud Computing and positive feedbacks have been received from experts when it came to evaluate the technological roadmap.

Title: Towards a novel method of research evaluation: Tracking Translation in biomedical research using a proximities framework

Author(s): Frederique Lang, Michael Hopkins, Jordi Molas-Gallart, Ismael Rafols, and Puay Tang

Author Affiliations: University of Sussex

Corresponding author e-mail: f.lang@sussex.ac.uk

Abstract:

The present research aims at understanding how links emerge between different scientific specialisations through funded research projects. It particularly focuses on the evaluation of translational biomedical research projects and their ability to create such links. By definition, translational research aims to connect scientific and clinical findings to the development of new therapies and medical practices which could have results on health impacts (Zerhouni 2003, MRC 2012).

Translational research has been an area of increasing priority for biomedical funders, as in many cases the innovation process is characterised as obstructed or inefficient. Biomedical breakthroughs with apparently great potential have led to disappointingly slow progress in the clinic (Hopkins et al. 2006; Hopkins et al. 2007).

Policy makers and funders increasingly need tools to assess if the translational drive is achieving its aims. While previous research has already identified a number of barriers and gaps for translational work (Wanless 2002; Bioscience Innovation and Growth Team 2003; House of Lords S&T Committee 2009), this research aims at understanding how researchers from diverse backgrounds can overcome translational barriers through building relationships and a shared knowledge base.

Swan et al. (2007) have identified that biomedical innovation is achieved through successful collaboration that is not only based upon cognitive aspects but also through shared incentives. Thus the research evaluation framework developed here must take into account important dimensions that are necessary for individuals to collaborate, and learn from each other. The proximity dimensions described by Boschma (2005) are believed to be critical for successful collaboration leading to knowledge creation, innovation and have the potential for assessing translational research (Molas-Gallart, 2014). These are namely: cognitive, organizational, social, institutional and geographical proximities. It states that although the cognitive aspect is important in building collaboration that leads into innovative products, other dimensions should also be considered. It is also concerned with the institutional component which deals with incentives, the degree to which people are acquainted with each other through the social dimension, organisational arrangements or even the degree of face to face interaction between researchers involved in the collaboration. All of these dimensions are believed to improve communication and learning. We therefore build upon these dimensions to develop our method.

This research focuses on tracking changes in the 5 proximities outlined above in order to assess whether individual researchers have been influenced by other researchers from different backgrounds, if their common work or future individual work is affected by this collaboration, and if their research becomes 'more translational'. The proposed method will assess translational projects along these 5 dimensions using both quantitative and qualitative tools.

The quantitative side of the method relies on bibliometric tools using publication data. Publications can be seen as traces of collaboration generated by individual researchers, which can generate information about the cognitive background, formal relationships, institutional background and geographical location. Using overlay mapping methods (Rafols et al. 2010) we aim at understanding the change in the cognitive background of the research team before and after the project started. We also use publication data to build networks of collaboration that emerged as part of the research project.

However, publications are only traces of research activities. It only provides a snapshot of information about the research conducted. Publications may appear with a lag compared to when the research was conducted, and may not be fully inclusive of all the research activities conducted by the researchers. Secondly, this type of data provides little indication about some of the dimensions we wish to study, such as the organisational and institutional aspects.

For these reasons the method we develop also relies upon qualitative tools such as semi-structured face to face interviews with people involved in the case study project, and also phone interviews based on a shorter version of the questionnaire. One novel feature of the method is the emphasis placed on drawing a map of interactions between project partners during the interviews in order to have a systematic way to record organisations (recording types of institutions and their geographical locations), individuals involved within these organisations and their interactions with other people within the project (looking at social and institutional proximity). The various aspects of proximities and evolution of relationships are then further discussed in order to get a sense of how relationships have evolved and how the individuals have learned from each other and overcome potential barriers of communication.

The paper presents work in progress on the implementation of the method in a test case. The paper therefore provides an opportunity to reflect on the advantages / disadvantages of the method, as well as the challenges involved in the implementation of mapping approaches and development of proxy-measures employed for each of the 5 proximities to be studied

Title: Analytics for decision making in development: Scientific Landscaping in Africa

Author(s): Paul Oldham

Author Affiliations: One World Analytics

Corresponding author e-mail: poldham@mac.com

Abstract:

This presentation will present work in progress on the novel use of scientific landscaping to inform decision-making on investments in development projects in several African countries. Decision-making on development projects is typically based on identified country needs and the priorities of donors. However, this raises difficult practical questions about how to identify projects and whom to invest in to realise the goals of development cooperation. In this presentation I report on the use of a combination of Web of Science, VantagePoint, Gephi, RStudio and Tableau Public to develop scientific landscapes for eleven African countries. The research consists of mapping the overall scientific landscape for a country of interest using journal publications. The analysis is refined for a thematic area, in this case genetic resources, and key organisations, actors and issues are mapped. The results are then published online using Tableau Public (e.g. https://public.tableau.com/profile/poldham#!/). The results are used by teams of researchers to organise country level interviews and as tools for engagement with government officials and researchers in writing country diagnostic reports to inform decision-making on project selection. The landscapes are being developed with the German Technical Cooperation (GIZ) as part of a longer term programme to support the practical implementation of the Nagoya Protocol on Access to Genetic Resources and Benefit-Sharing. However, the use of scientific landscaping to inform development planning appears to be novel and is likely to be of interest to the wider Data Science and Tech Mining community.

Title: A method to identify computer-implemented inventions at the EPO

Author(s): Rainer Frietsch, Peter Neuhäusler

Author Affiliations: Fraunhofer Institute for Systems and Innovation Research, Germany

Corresponding author e-mail: rainer.frietsch@isi.fraunhofer.de

Abstract:

Software as such is not patentable at the European Patent Office (EPO). Computer pro-grams and processes are only patentable if they solve technical problems with technical means, which says nothing about the context of the technical solution that is claimed, but expresses that an invention containing an element in the form of software is a computer implemented invention (CII). Embedded software is one group within CII, for example.

In effect, CII cannot be identified using the IPC or by simply using a few keywords. Xie and Miyazaki (2013)1 had suggested a keyword-based method to identify software patents in the automobile industry. We employ this method and develop it further so that it is applicable to all patents at the EPO. The method uses a set of keywords, which were qualified by their recall and precision (indicators). Next to the title and abstract, the keywords were used for the claims, adding a considerable amount of reasonable patents to the group of CII, while using the description of the patents leads to imprecise results.

In this presentation, we will explain the method, its implications and shortcomings. We will give some results and provide the Oracle SQL scripts to implement the method to anyone who is interested.

We apply a rather “conservative” definition of Computer-implemented inventions (CII), strictly excluding “software as such” – in consequence, our data are minimum numbers. Empirical results based on this method show that since about 2002, more than 35% of total filings at the EPO are CII patents. Non-EPC members have higher shares of CII patents at the EPO than EPC member countries. CII more frequently target international markets, thereby securing international competitiveness – this holds at least for German applicants. CII plays a more important role in a number of sectors in Germany than in many other countries.

Title: Customising VP scripts for 'one-click' automated patent analysis

Author(s): Peter Keefe

Author Affiliations: UK- Intellectual Property Office

Corresponding author e-mail: peter.keefe@ipo.gov.uk

Abstract:

In 2015 the patent informatics team at the UK-IPO were asked by the Government to produce patent analysis on over 45 different technology areas of interest to policy advisors. With a four-week deadline, such large volume patent analysis required a rethink in the way we produce analysis. The team have previously experimented with creating a number of VantagePoint scripts to help speed up specific steps within our processes for patent analysis but this project required a major step-change in how we use VantagePoint; a 'one-click' patent analysis that takes raw imported patent data and automates the creation of lists, applies thesauri, creates time-sliced subsets and populates a spreadsheet. It would have been impossible to deliver this work on time without automation and this talk will focus on how the UK-IPO utilised the power of VantagePoint to produce quick-look automated patent analysis with a single click; we shall discuss how we developed in-house expertise in writing custom VantagePoint scripts, how the script was designed based on the required output and the resulting challenges we encountered during development. We will also discuss future work to improve the script by allowing the user to customise the output format and level of analysis presented, and improving the robustness and efficiency of the script.

Title: The Patent Guide: A Handbook for Analysing and Interpreting Patent Data

Author(s): Chris Harrison, Peter Evans

Author Affiliations: UK- Intellectual Property Office

Corresponding author e-mail: peter.keefe@ipo.gov.uk

Abstract:

There is a growing consensus that "innovation" is a key driver of productivity and economic growth. Innovation occurs away from the public eye, behind closed doors and in the minds of entrepreneurs and academics. This makes understanding innovation difficult, as there is no data to determine what is being created and where, and the impact it might have. The research community has been forced to seek alternative measures to proxy for innovation; patent data is one of the most popular. The concept of a patent is simple to understand; the underlying legal and procedural process is not. This creates a contradiction for patent analysis. Datasets can be relatively straightforward, proving easy-to-use for the experienced and inexperienced analyst alike. What is not as straightforward is the actual meaning and significance of the data; importance is placed on findings that have been misinterpreted and/or are wholly incorrect. At present there are clear differences in perspective between professional patent experts, researchers undertaking analysis and the audience for this research, including governments, the press, businesses and individuals. Such differences increase the possibility for incorrect analysis or inappropriate interpretation of analysis. Decisions based upon this could be incorrect and potentially harmful. In 2015 the UK-IPO launched 'The Patent Guide', a handbook created to improve shared understanding between patent experts and those undertaking or using patent research. It bridges the disconnect between those who understand the data and those who use it by addressing common misconceptions and the appropriate use of patent data for analysis. Hot topics discussed include patents as a proxy for innovation, the nationality of the patent, how to appropriately analyse citations and claims, and what to consider when analysing patent quality and value. A recent update has new chapters covering how to use and analyse patent families and patent classification.

Title: Semantic patent analysis – how to use modern methods for monitoring technological convergence and strategic technology planning

Author(s): Prof. Dr. Martin G. Moehrle, M.Sc. **Michael Wustmans**, M.Sc. Jonas Frischkorn, M.Sc. Frank Passing

Author Affiliations: University of Bremen

Corresponding author e-mail: michael.wustmans@innovation.uni-bremen.de

Abstract:

The increasing amount of data in the databases available worldwide lead to considerable problems in monitoring technology convergence and strategic technology planning. This problem of information overflow is not only vivid in the field of scientific literature but also in the knowledge discovery from patent information. The use of patents as a source of information is important for companies in order to distinguish themselves from competitors as well as to mitigate risks. To support this analytical process a semantic patent analysis process was developed at the Institute of Project Management and Innovation (IPMI) at Bremen University in cooperation with the spin-off company i3-Management GmbH for efficient analysis of text-based patent information. This so-called PatMining method provides the ability to extract semantic structures from patents as well as to filter linguistically and evaluate the data structure. With two case studies we would like to present our current research in (i) monitoring technology convergence and (ii) strategic technology planning. (i)Investigations have shown that especially patent data is a vital source for the analysis of measuring technological convergence. Although academics have paid much attention to classification and citation analyses these approaches have shortcomings in terms of e.g. delays and misclassifications. To overcome these shortcomings we will focus on a novel approach for measuring technological convergence by semantic similarity of patents using technology corpora. For this purpose, we test our semantic measurement approach by the case of smart grids which we analyze by the means of a technology complex. (ii)In the context of strategic technology planning it is a necessity for companies to observe the evolution of technology fields. This consideration involves thematic developments as well as player-related developments. Based on the semantic analysis of patents, we have devised an analytical method called Patent Lane Diagrams which permits a content-wise deployment of the technology field in a given time-frame. Through connection to further data, such as applicant or inventor details, information on patent classifications and the combination with informetric key figures, the Patent Lane Diagrams enable deep insights into the structure of technology fields, and may thus lead toward conclusions and recommendations for corporate action. The possibilities offered by Patent Lane Diagrams are demonstrated on the basis of a case study from the technology field of carbon fiber reinforced synthetics in the automotive industry.

Title: More precise impact metrics for citation data: The importance of the geometric mean

Author(s): Mike Thelwall

Author Affiliations: University of Wolverhampton

Corresponding author e-mail: m.thelwall@wlv.ac.uk

Abstract:

This talk will give a range of examples to demonstrate how a simple switch from citation indicators based on the arithmetic mean to indicators based on the geometric mean gives results that are more precise and for which confidence intervals can easily be calculated. The arithmetic mean is not optimal for highly skewed data sets, such as those based on citation counts, because individual very high values can have a substantial influence on the overall result. As previously argued by Michel Zitt, the geometric mean is a better option that is much less susceptible to individual high values. Confidence intervals for the results can also be calculated because one method of calculating the geometric mean for citation data involves a transformation to a distribution that is close to normal, and so standard formulae for confidence intervals can be used.

Title: Telling the whole story: seeing past indicators to the signals in bibliometric information using PCA

Author(s): Keith Julian, John Rigby

Author Affiliations: MIoIR

Corresponding author e-mail: John.Rigby@mbs.ac.uk

Abstract:

As bibliometric data is multidimensional, its study, by means of an index number, especially an index number in the form of a ratio, rarely captures all the information. Across the bibliometric and evaluation literatures there is increasing scepticism about the value that single index numbers give to the understanding of scientific behaviour and its consequences, including impact. The authors propose and demonstrate the use of a multivariate approach – principal component analysis - that gives greater insight into the aspects of scientific publication. Principal component analysis is put forward as the most suitable multivariate method available as it does not emphasise any variable and identifies important signals in the data that may not be observed with univariate methods. A data set analysed in a previous piece of work on double-dipping is used here and is subject to PCA which reveals three components, an input related component, an output related component and an outcome (impact) related one. Importantly, citation is shown to be of limited significance in explaining overall variability within the data set.

Research in Progress Presentations

Title: Profiling Green Goods Sector Enterprises in China: An Exploratory Alibaba Web Mining Analysis

Author(s): Alec Waterworth, Abdullah Gök and Philip Shapira

Author Affiliations: MIoIR

Corresponding author e-mail: alec.waterworth@postgrad.mbs.ac.uk

Abstract:

This working paper presents findings from the web mining analysis of Chinese green goods firms. The analysis draws on the data collected from the English-language Alibaba (www.alibaba.com) web pages of a sample of 300 Chinese small and medium-size enterprises (SMEs). In this study, SMEs are enterprises with 1000 employees or less, established in 2002 or later, that are involved in the production of manufactured 'green goods'. The focus on enterprises established in 2002 or later is consistent with the project objective of examining the growth trajectories of relatively recent start-up firms.Firms were identified through a comprehensive set of sector-spanning green goods search terms (Shapira et al., 2014). Webpages for the years 2006-2013 were accessed through the Internet Archive Wayback Machine (web.archive.org), which is online archive of historical website content. The web mining approach used here was previously applied in a similar analysis of 300 UK green goods SMEs and is discussed in Gok et al., 2014. Whereas the previous UK analysis had focussed on company websites, for the Chinese firms, we chose instead to mine the Alibaba webpages for these companies. This decision was initially made due to a lack of English-language webpages for our sample of 300 firms. However, a unique opportunity quickly arose from this. Whereas a selection of 300 company websites will vary enormously in the content available and how that content is organised and arranged across the website's many pages, Alibaba webpages are in a regularised format, with a set of standardised fields. The paper emphasises the opportunities of mining websites such as Alibaba, which offers a semi-structured dataset, over traditional company websites, which typically produce a very large, unstructured dataset. This aligns web mining of Alibaba with more traditional innovation studies quantitative methods, such patent analysis and bibliometrics (i.e. semi-structured approaches).

Title: Study of the changes in innovation output of companies before and after getting certified under the R&D&i management standard UNE 166002. Use of web mining tools to characterize stages in innovation management proficiency

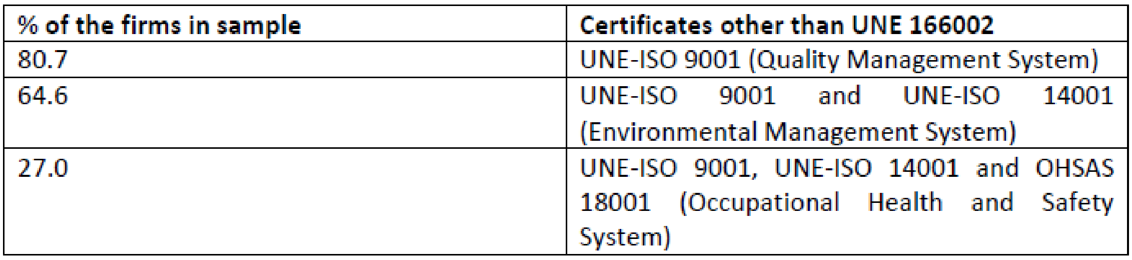
Author(s): Rosa Rio-Belver

Author Affiliations: University of the Basque Country, Spain

Corresponding author e-mail: rosamaria.rio@ehu.eus

Abstract:

Our sample consists of 243 businesses (most of them SMEs) certified under standard UNE 166002. Most of these firms have also certified other management areas under international standards, as shown in the table below. Other standards such as ISO 27001 (Information Security Management System) are also present in the sample.



% of the firms in sample	Certificates other than UNE 166002
80.7	UNE-ISO 9001 (Quality Management System)
64.6	UNE-ISO 9001 and UNE-ISO 14001 (Environmental Management System)
27.0	UNE-ISO 9001, UNE-ISO 14001 and OHSAS 18001 (Occupational Health and Safety System)

The aim of this study is to characterize the effect of the implementation of standard UNE 166002 in the innovative activities conducted by the firms in the sample. A set of well-known indicators can be used to compare the innovative performance before and after certifying under UNE 166002:

- Number of research projects.

- Amount of research collaborations.

- Winning of awards.

- Granting of patents.

We would like to find a way to further characterize the changes in the innovative behavior of these companies by web scraping their websites. New indicators could be developed showing the evolution of main concepts and corporate identity elements present in company's "gateway to the world". Main maturity stages in innovation management could also be characterized by using this information.

This approach would use the following information sources:

- Business databases: Sabi (Dun & Brasteed)

- Patent databases: Global Patent Index (GPI).

- Websites: Prior versions available with Wayback Machine. Webscraping tasks will be executed using IBM Watson software, and Vantage Point software will be used for text mining analysis.

Title: 3-D analytics technique for research options comparison

Author(s): Fernando Palop, Scott W. Cunningham and Blanca de-Miguel-Molina

Author Affiliations: Universitat Politècnica de València

Corresponding author e-mail: fpalop@ingenio.upv.es

Abstract:

This technique allows an easier way of comparison of different research options based on a 3-D analytics technique that integrates the "consensus of interest" of the research team and the management, the functionality or effect and the source. Once a manager decides to dig into particular research field, it is not uncommon, that there are different research options (metabolites, cells, substances, etc) to base and develop a research project. So the cost of opportunity to select one is high. The authors mined literature and patents contents and upon their results developed a 3-D analytics technique to map bioactive or technical effects, the agents or promoters of those effects and the sources that contain those agents or promoters. This technique has been already validated through a practical application. The output of this technique display automatically into the matrix cells the results of a third dimension choosed. It is not more about quantitative data results upon the co-occurrence between two fields or dimensions. It is about presenting the terms contained in a third field and their frequency. This automatic output process allows getting results upon the analysis of large volumes of scientific articles or patents records structured texts. The third field terms results displayed in the matrix cells allows an easy way to get the whole landscape of options. This presentation facilitates the researchers and decision makers, knowledge about the source (on the matrix rows) of the results displayed in the matrix and at the same time their type of activity or effcts (columns header) deployed in the application. So the 3-D matrix would facilitate the discussions and the consensus establishment about the best choices.

Title: Global trends on additive manufacturing: latest inventions on ophthalmology field

Author(s): Marisela Rodríguez Salvador, Ana Marcela Hernández de Menéndez

Author Affiliations: Tecnológico de Monterrey

Corresponding author e-mail: marisrod@itesm

Abstract:

This research identifies global trends on additive manufacturing including research focus, organizations, latest inventions, and priority countries involved in emerging applications in ophthalmology. For this purpose, a scientometric patent analysis was developed using datamining software. Patents were retrieved from 19 authorities covering a time period from 1782 to 2015 (April 29). Results indicate that despite the big benefits that additive manufacturing will bring in ophthalmology this application is now in its infancy. A total of 12 family patents were obtained. Main insights show a global research trend toward the development of lenses, followed by prosthesis and implants. Essilor International leads inventions applying additive manufacturing by developing ophthalmic lenses. In regard to prosthesis, Liu Qinghuai (individual), Manchester Metropolitan University and Fripp Design Ltd. have developed and patented inclusive artificial eyes. On the other hand, Becker Hartwig (individual) has developed intraocular lens. Results show that the first country of protection is France, followed by United Kingdom. Most inventions have been developed to be used outside the human body. This fact could be related to the novelty of the technology. Results of this research offer valuable knowledge about emerging technologies and future innovations in ophthalmology.

Title: The effects of organizational and geographical diversity on EU-FP7 funded ICT research

Author(s): Jojo Jacob, **Ad Notten**, Bulat Sanditov,Alex Surpatean

Author Affiliations: UNU-MERIT

Corresponding author e-mail: notten@merit.unu.edu

Abstract:

We carry out an in depth examination of patent and publication data reported in the annual surveys of DG CONNECT from 2007 to 2013. We began by validating the raw publication and patent data by first matching and validating them with data in independent bibliographic databases. For publications, this was Elsevier's Scopus database and a Scopus Custom Data set, while for patents, EPO's PATSTAT database and Espacenet were used. For the validated patents, we collected additional information (from the PATSTAT database and Espacenet), such as the location of inventors, applicants, and IPC classes which were used in the subsequent analysis. The publications reported were validated using DOI in the first instance and the author-title pair in the second instance. We retained all publications that were either articles or proceedings and which had a validated publication year greater than or equal to 2007. Additional bibliographic information was generated using text-mining software.

Next publication and patent analyses are performed to generate a number of indicators which we will use later on as variables in our econometric model. We focus these analyses on the geographical and organizational distribution. In terms of internationalization of research, we find as expected that the cooperation among the EU-28 countries engaged in ICT research under FP7 is more evenly distributed than for the control sample. We can infer that the FP7 project participants display a significantly higher degree of internationalization than their control sample counterparts, whom are rather nationally focused. We also undertake a citation analysis to assess the impact of the research funded under FP7-ICT. We find that, for the total stratified sample, the relative citation rate is well above the world average for the years examined. From the citation analyses on a more disaggregate level we furthermore find that the public research and the research published in conference proceedings is of higher quality as compared to the control.

Lastly we undertake an econometric analysis in order to gain more fine grained insights on the factors contributing to the scientific and technological performance of the FP7 projects. In addition to using project-related variables such as funding, number of participants involved, the type of financing instrument, etc., we introduced a novel measure to capture the diversity in the scientific capabilities at the national level of organizations involved in each project using the data as found in the preliminary publication and patent analyses. Recent studies have highlighted the importance of a geographically diverse knowledge sourcing strategy that would enhance the possibilities for knowledge recombination—a vital ingredient to innovation. In line with this view, our analysis of the publication productivity revealed that scientific diversity of project teams played a significant role in enhancing their scientific productivity. We also found evidence on the positive contribution of funding duration and total funding, as well as the number of participants involved in a project and funding per participant, for publication productivity. We further found that the patent output of projects has a complementary effect on their publication output, and that projects with a greater share of university partners tend to perform better.

We carried out a similar analysis for patents. In over 14,000 project-year observations only about 170 observations report positive patents, making the results not entirely generalizable. Among the key variables, the diversity variable has a positive coefficient, which however is significant in only about

half of the models. Some consistent results stand out, however. Patent output is positively shaped by funding per person, publication output, and a high share of industry participants. Comparing the results on the factors shaping publications and patents, we can say that these two performance outcomes feed each other’s growth. Furthermore, while participants from universities tend to spur growth in publications, those from industry are critical for growth in patents. In sum, our analyses revealed a significantly high internationalization of the inventive activities in FP7 projects in the field of ICT. This is in sharp contrast to the general, non-collaborative, trend in the industry globally, and in the EU-based industry. Our econometric analysis confirmed that the diversity of the project team in terms of the scientific capabilities of the countries in which they are embedded plays a critical role in enhancing the productivity of the projects.

Title: Synthetic Biology and Interdisciplinarity – early insights from the UK's new synthetic biology research centres

Author(s): Yanchao Li, Abdullah Gok, Philip Shapira

Author Affiliations: MIoIR

Corresponding author e-mail: Yanchao.li@mbs.ac.uk

Abstract:

Synthetic biology has been recognized by policymakers across the world as a strategic, emerging technological area. It is believed that integration of various disciplines is crucial to promote the flourishing of this area. The UK's largest synthetic biology research funder Biotechnology and Biological Sciences Research Council (BBSRC) explicitly stated that '...proposals will require strong multidisciplinary partnerships between bioscientists and researchers in engineering, the physical sciences and information technology disciplines' in its responsive mode funding priorities. The establishment of the six synthetic biology research centres (SBRCs) in the UK offers an opportunity to investigate issues of interdisciplinarity within a clearly defined synthetic biology research community. All these SBRCs claim an emphasis on interdisciplinarity from their centre description and job vacancies. Yes little knowledge has been developed as per how interdisciplinarity is reflectedand advanced in those centres. This paper makes an early attempt to develop measures to examine interdisciplinarity and profile the six SBRCs' interdisciplinary characteristics. A methodology built on bibliometrics and documentation analysis has been adopted. Disciplinary compositions of those SBRCs, their respective intellectual focuses and their patterns of disciplinary integration are analysed. Documentation data sources include BBSRC funding announcements, SBRC proposals and other key documents, biographies of SBRC investigators and SBRC job announcements. Bibliometric data included publication records according to funding sources, and publication track records of investigators collected from Web of Science Core Collection in August 2015. Preliminary findings include that...Nottingham features biomedicine and pharmacy, Manchester focus on chemistry... The involvement of social scientists also varies, with Edinburgh and Manchester featuring the most proactive involvement of social scientists, while Bristol and Cambridge's profiles do not contain social science elements.