SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

Ontologies as bridges between data sources and user queries: the KNOWMAK project experience[1]

Diana Maynard* and Benedetto Lepori**

*d.maynard@sheffield.ac.uk
Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield, UK

** benedetto.lepori@usi.ch
Faculty of Communication Sciences, Università della Svizzera italiana, 6904 Lugano, Switzerland and Laboratoire Interdisciplinaire Sciences, Innovations et Sociétés (LISIS), University of Paris Est, 77454 Marne-la-Vallée Cedex 02, France

ABSTRACT

This paper describes ongoing work in the KNOWMAK project, which aims to develop a web-based tool providing interactive visualisations and state-of-the-art indicators on knowledge co-creation in the European research area. One of the main novel developments in this work is the use of ontologies to act as a bridge between the data sources (research projects, patents and publications) and user queries, in order to address the problems of mapping between heterogenous data sources with different vocabularies while still maintaining a level of standardization necessary for summarising the information required to provide informative views about the highly dynamic S&T landscape.

INTRODUCTION

The identification of research performed around a certain topic is a central issue for STI research – to map the structure and temporal dynamics of knowledge production, but also to analyse the impact of publicly-funded research towards addressing certain key issues, like the development of Key Enabling Technologies (KET[2]) and the solution of Societal Grand Challenges (SGC[3]).

The traditional approach relies on the creation of classification systems, where knowledge production activities are classified within a hierarchical structure of science or technology domains at different levels of disaggregation – examples are the fields of R&D classification (FORD) introduced by the Frascati manual for R&D expenditures and personnel (OECD, 2015), the Web of Science Journal classification (Leydesdorff and Rafols, 2009), and the IPC codes for patent classification (Debackere and Luwel, 2005). These classifications also come with rules to match data items, respectively based on the department where a researcher is working, the journal in which a paper is published, and keywords used in the patent description. Classifications present some desirable properties, which explain their wide usage: they are

[1] This work was partially supported by the European Union under grant agreement No. 726992 KNOWMAK. The authors want to thank the other members of the KNOWMAK team for useful comments and discussions on the subject.
[2] https://ec.europa.eu/programmes/horizon2020/en/area/key-enabling-technologies
[3] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges

simple, stable over time and, at least in principle, cover all fields of knowledge production, thereby allowing for cross-field comparisons. These characteristics lead also to a number of relevant limitations: classifications are slow to adapt to the changing landscape of science and, therefore, to identify emerging domains, which tend to remain hidden within existing categories; their hierarchical structure is not well suited to the interconnected nature of new sciences (Bonaccorsi, 2008); and more importantly, classifications are closely related to specific data sources and their respective structure, and therefore, despite wide-ranging efforts for mapping different classification schemes (Schmoch *et al*, 2003), they remain largely incommensurable. Finally, there are lasting issues in mapping existing classifications of the science basis with policy-oriented concepts, with their different focus and vocabulary.

To address these limitations, a large body of work has been developed based on more fine-grained techniques relying on individual data items, including the analysis of citation networks (Small et al., 2014) and co-word analysis to map scientific topics and their respective semantic linkages. The use of overlay maps (Rafols et al., 2010) to identify clusters of publications or patents associated with a specific topic has also been explored (Rotolo et al., 2014). National Language Processing (NLP) techniques have also been used to map topics and to enhance traditional sources of information about R&D activities, e.g. reported on company websites and in databases of patents and publications (Gok et al. 2015, Kahane et al., 2015), but the focus is typically on regular expression-based keyword search to group similar terms, rather than complex linguistic analysis.

While these techniques tend to provide in-depth and detailed views of specific knowledge domains and of their evolution over time, they are currently less suited for large scale mapping across the whole S&T landscape, where some level of standardization and summarization of the information is required to provide informative views. Moreover, such techniques rely heavily on the specific vocabulary and linguistic genre of data items, and therefore interlinking data sources like patents and publications is not straightforward.

In this paper, we will present some pathways to address these issues, as developed within the recently-launched H2020 project KNOWMAK (Knowledge in the Making in the European Society). The approach we are developing relies on the development of ontologies focusing on KET and SGC, as a way to achieve a balance between the construction of a sufficiently stable classification of topics and the flexibility to address different vocabularies (for example between policy documents and scientific publications) and their evolution over time. While keeping some basic features of classifications, like the presence of a core set of subjects organized in layers, ontologies are more flexible in their structure and, through instances, can be connected to (different and evolving) vocabularies across heterogeneous data sources.

ONTOLOGIES IN THE KNOWMAK PROJECT

The KNOWMAK project aims to develop a web-based tool providing interactive visualisations and state-of-the-art indicators on knowledge co-creation in the European research area. It will provide information to users wishing to understand the nature of and connections between key European research in particular topics, institutions and locations. At the heart of this is the quadruple helix model of innovation, linking government, industry, academia and civil participants.

KNOWMAK builds extensively on the EU RISIS project, which is creating a distributed infrastructure for research and innovation data and policies, focusing on three main areas: creating and making available open datasets on research and innovation issues; developing open

platforms for building and manipulating such datasets; and developing tools to facilitate interconnections between heterogeneous existing datasets. One of the critical ways in which KNOWMAK goes beyond RISIS is the addition of an ontological element at its core, which forms a bridge between knowledge sources and user queries, and which, along with sophisticated NLP and term extraction techniques, enables better integration of the different resources, a more dynamic system that can cope with changing data, and improved handling of the complexities of language.

The use of sophisticated NLP techniques to model terms has a long-established history in the computational terminology field, though advances in machine learning and sheer computational power have enabled great strides in recent years (Francopoulo et al. 2016, Amjadian et al., 2016). In the Semantic Web field, ontologies have long been used to address policy issues, e.g. (Loukis 2007), and the addition of sophisticated NLP, in the form of semantic annotation tools which link texts to an ontology, is also far from new. A good overview of recent semantic annotation work, for example to discover and link information from large-scale documents such as patent data, health records, archived material and social media, is given in (Maynard et al. 2016). Other work has investigated the need for combining information from related fields to populate domain-specific ontologies, e.g. in the field of metabolomics (Spasic et al., 2008).

The creation and use of the KNOWMAK ontology for linking data is not a simple task. First, the areas of research covered in the project, KET and SGC, are themselves rather general, and a clear and consistent delineation from policy documents cannot be expected. Second, the ontology needs to connect user queries, mostly formulated in rather generic terms and variable over time and user groups, with the more specialized vocabularies used in scientific documents. To address these issues, we design a staged approach to ontology development, relying first on existing mapping studies and classifications and on expert knowledge in order to construct a basic ontology structure (Fraunhofer ISI, IDEA Consult, 2016; IDEA Consult, ZEW, NIW, TNO, CEA, Ecorys Fraunhofer ISI 2015), and second, enriching and populating the ontology vocabulary through text mining techniques on a selection of policy documents, publications, patents and descriptions of research projects.

LINKING DATA WITH ONTOLOGIES

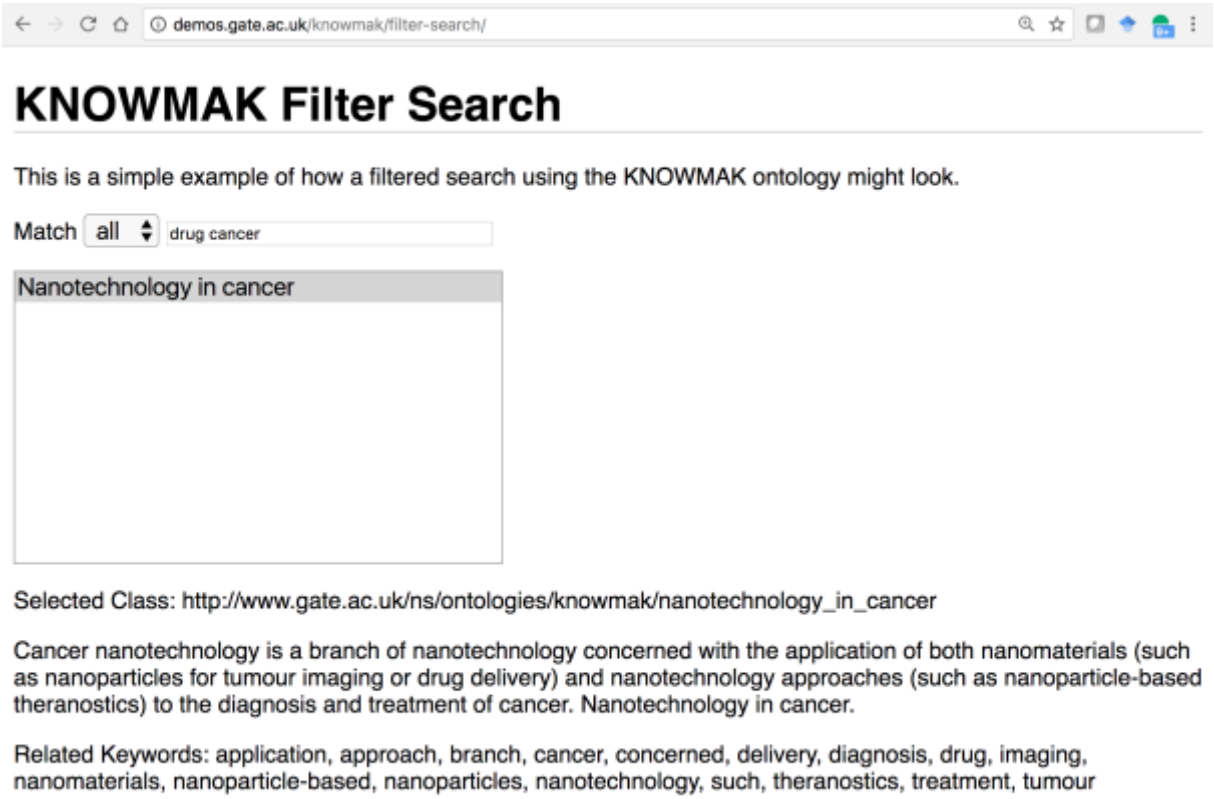
The challenge of dynamic linking between the users, topics and the data stored in our knowledge bases is addressed through the use of semantic annotation against our KNOWMAK ontology, based around the two research areas SGC and KET. Later, these could be extended to include other topics. Documents are annotated with one or more classes from the ontology, linking them with topics which, along with geographical and actor information, are used to build indicators about the data. We combine a traditional static classification of the data sources (mapping between datasets and topics via correspondence tables) with a more dynamic knowledge extraction process based on NLP. Fully bottom-up NLP is tricky due to the standardisation of information, i.e. a reasonably stable set of classifications is required (both to combine sources and to achieve consistency over time), but fixed classifications, on the other hand, are too rigid to deal with different ways of expressing concepts. Our approach combining classification and NLP takes place in 3 ways: (1) the design of the ontologies, which combines topics (classes) and vocabularies (instances); (2) the construction of the ontologies, which combines a classificatory approach with bottom-up ontology population (construction of vocabularies) from the data; and (3) the annotation process which links the data dynamically with the relevant ontological classes.

The kinds of language used, as well as the typical classification structures, vary considerably across the different data types. For example, while policy documents may deal with usage of various technologies, research publications tend to deal with techniques for their development. Patent documents are classified by numeric codes and have complex labels, e.g. "*Medicinal preparations containing antigens or antibodies*", which tend not to map neatly to a single ontological class or to a keyword typically used in a project or publication. Users will never express a query in such a way either. The ontology classes form a bridge between the different data sources, and also enable linking at different levels of specificity. This also means that the ontology can be refined depending on new information, without affecting any of the mappings already performed (Maynard et al., 2007; Maynard, Funk & Peters, 2009): the databases get updated, terms are dynamic, and user needs change.

In the area of KET, the topics covered are not mutually exclusive. Aside from the cross-cutting KET AMT (Advanced Manufacturing Techniques), there are also many overlaps between topics in Photonics (e.g. *miniaturisation issues with modules for fibre-optics*), Advanced Materials (e.g. *miniaturisation of various AM for Nano- and Micro-Electronics*); and Advanced Manufacturing Techniques (e.g. *semiconductor manufacturing processes*). These are modelled in our ontology through multiple inheritance (e.g. AMT_Advanced_Materials is a subclass of both AMT and Advanced Materials). Similarly in the SGC area, *Biotech* is modelled in our ontology as a subclass of both the Health and Bioeconomy SGC top-level classes.

The ontology is designed via an expert process, taking into consideration the needs of the users (following dedicated user workshops) and previous existing classifications and protocols, following a principled development process (Maynard, Funk and Lepori, 2017). The ontology is then enriched with instances representing topic vocabularies, based on keywords from the relevant databases, term extraction from the class descriptions and labels in the ontology, term extraction from patent code labels, and term extraction from the actual publications, patents and project descriptions in the databases. Simple examples of this can be seen in the online search demo[i], where keywords are connected with topics (note that the functionality is still quite preliminary at this stage – there are clearly a few erroneous terms currently generated). A sample demo session is shown in Figure 1. Here, the user has searched for the keywords "drug" and "cancer" occurring together, which have been associated with the ontological class "nanotechnology in cancer". Below, one can see the description of this concept as well as other instances (keywords) associated with it.

Figure 1: Example of a filter search showing the keywords and related classes in the ontology



By populating the ontology with instances from different data sources, with their specific vocabularies, it will then be possible to associate publications, patents and European projects to a specific subject and to construct summary indicators – like the number of patents for a given subject in a European region. Our approach is therefore combining some features of traditional classifications, like the possibility of producing summary indicators to compare subjects or regions, with the flexibility of NLP techniques. While the subject structure might remain rather stable over time, the associated vocabularies can be adapted to changes in knowledge dynamics in a straightforward way.

DISCUSSION AND CONCLUSION

It has become clear from our previous experience and from discussions with the user community that, while the work on ontologies is very challenging, it is in high demand due to numerous issues in analysing and linking different kinds of European research activity, and the dynamic nature of these emerging technologies. While the project is still in the early stages of development, it builds on a solid foundation of established work, and thus the way forward is relatively clear, even if many problems remain to be solved. Indeed, much of the design of the system and the core strategy is already in place. However, while for example the principles for the automatic tagging of data sources are defined, only experimentation will show which exact methodologies are most successful, and how best to tune the well-known tradeoff between precision and recall (Buckland and Gey, 1994). It is clear that the design of the ontological element and the mapping between different topics is one of the most challenging elements of the system, but also one of the most exciting in terms of its potential, and thus one of the most critical in terms of technological advancement.

References

Amjadian, E., Inkpen, D., Paribakht, T. S., & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. *5th International Workshop on Computational Terminology (Computerm 2016*), (pp. 2-11). Osaka, Japan.

Bonaccorsi, Andrea 2008. Search regimes and the industrial dynamics of science. Minerva, 46, 285-315.

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45, pp. 12-19.

Debackere, Koenraad and Marc Luwel 2005. Patent data for monitoring S&T portfolios. In Handbook of Quantitative Science and Technology Research, ed. Anonymous , pp. 569-585. Springer.

Francopoulo, G., Mariani, J., Paroubek, P. & Vernier, F. (2016). Providing and Analyzing NLP Terms for our Community. *5th International Workshop on Computational Terminology (Computerm 2016*), (pp.94-103). Osaka, Japan.

Fraunhofer ISI, IDEA Consult (2016), Study on EU Positioning. An Analysis of the International Positioning of the EU Using Revealed Comparative Advantages and the Control of Key Technologies, European Commission, DG RTD, Brussels.

Frietsch, R./ Neuhäusler, P./ Rothengatter, O./ Jonkers, K. (2016): Societal Grand Challenges from a technological perspective – Methods and identification of classes of the International Patent Classification IPC (= Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis No. 53). Karlsruhe: Fraunhofer ISI.

Gok, A., Waterworth, A. & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometric*s 102(1), pp. 653–671.

IDEA Consult, ZEW, NIW, TNO, CEA, Ecorys Fraunhofer ISI (2015): Key Enabling Technologies– First annual report: European Commission, DG Growth, Brussels.

Kahane, B., Mogoutov, A., Cointet, J.P., Villard, L. & Laredo, P. (2015). A dynamic query to delineate emergent science and technology: the case of nanoscience and technology. Technical report *Content and technical structure of the Nano S&T Dynamics Infrastructure* pp. 47–70.

Leydesdorff, Loet and Ismael Rafols 2009. A global map of science based on the ISI subject categories. J.Am.Soc.Inf.Sci.Technol., 60, 348-362.

Loukis, E.N. (2007). An ontology for G2G collaboration in public policy making, implementation and evaluation. *Artificial Intelligence and Law* 15(1), pp. 19–48.

Maynard, D., Bontcheva, K. & Augenstein, I. (2016). *Natural Language Processing for the Semantic Web*. Morgan and Claypool.

Maynard, D., Funk, A. & Lepori, B. (2017). Towards an Infrastructure for Understanding and Interlinking Knowledge Co-Creation in European research. Proceedings of *ESWC 2017 Workshop on Scientometrics*, Portoroz, Slovenia.

Maynard, D., Funk, A. & Peters, W. (2009). NLP-based support for ontology lifecycle development. *Proceedings of ISWC Workshop on Collaborative Construction, Management and Linking of Ontologies (CK 2009)*, Washington.

Maynard, D., Peters, W., d'Aquin, M. and Sabou, M. (2007). Change Management for Metadata Evolution. In *Proceedings of ESWC International Workshop on Ontology Dynamics (IWOD)*, Innsbruck, Austria.

OECD. 2015. Frascati Manual 2015. Guidelines for Collecting and Reporting Data on Research and Experimental Development. Paris: OECD.

Rotolo, D., Rafols, I., Hopkins, M.M., & Leydesdorff, L. (2014). Scientometric mapping as a strategic intelligence tool for the governance of emerging technologies. *SPRU Working paper Series* SWPS-2014-10.

Schmoch, Ulrich, Francoise Laville, Pari Patel and Rainer Frietsch 2003. Linking technology areas to industrial sectors. Final Report to the European Commission, DG Research, 1, 100.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 43(8), pp. 1450-1467.
Spasic, I., Schober, D., Sansone, S.A., Rebholz-Schuhmann, D., Kell, D.B., Paton, N.W. (2008): Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC bioinformatics* 9(5), S5.

[i] http://demos.gate.ac.uk/knowmak/filter-search/