Using ontologies to map between research and policy data: opportunities and challenges

Diana Maynard[1], Benedetto Lepori[2] and Philippe Laredo[3]

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

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

[3]*philippe.laredo@enpc.fr*
University of Paris Est, 77454 Marne-la- Vallée Cedex 02, France

Abstract
Understanding knowledge co-creation in key emerging areas of European research is a critical issue for policy makers in order to analyse impact and make strategic decisions. However, current methods for characterising and visualising the field have limitations concerning the changing nature of research, the differences in language and topic structure between policies and scientific topics, and the coverage of a broad range of scientific and political issues that have different characteristics. In this work, we discuss the novel use of ontologies and semantic technologies as a way to bridge the linguistic and conceptual gap between policy questions and data sources. Our experience suggests that a proper interlinking between intellectual tasks and the use of advanced techniques for language processing is key for the success of this endeavour.

Introduction
Mapping diverse kinds of scientific output to key topics in the science policy debate is a central concern that still requires more research. Traditional methods to characterise and visualise the field of knowledge production have important limitations. The move towards open linked data in STI studies is generating new opportunities, but also new challenges that are at the core of this paper. The opportunities concern the ability to interlink different kinds of data sources, such as publications, projects and patents, in order to provide a richer view of knowledge production (Light *et al*, 2014); the challenges are related to the need for a robust approach to identify and model relevant topics, such as those associated with specific policy and scholarly questions (Cassi *et al*, 2017).

Traditional classification systems for characterising research, e.g. the Web of Science Journal classification (Leydesdorff et al., 2009) and the IPC codes for patent classification (Debackere and Luwel, 2005), are typically simple, stable, and have widespread coverage. However, combining such schemes in order to depict an overall view of scientific knowledge production that encompasses different data sources is inherently challenging: each scheme is closely related to a specific type of data source, and despite wide-ranging efforts to map different classification schemes (Schmoch *et al*, 2003), they remain largely incommensurable. Without this overall view of knowledge production, cross-field comparisons cannot be made. Furthermore, mapping these classifications of scientific basis to policy-oriented topics presents a further issue due to terminological and conceptual divergence.

We address these problems through the use of ontologies to drive the development of a web-based tool providing interactive visualizations on European research. The tool is designed to provide information to users wishing to understand the nature of, and connections between, key European research, focusing on two topic areas central to policy makers: Key Enabling Technologies (KET) and Societal Grand Challenges (SGC). Ontologies share with

classifications the fact that they are constructed upon some intellectual understanding of reality; while their creation can be assisted by all kinds of text-based methods, they ultimately require some method of expert-based arbitration and must rely on some kind of "shared vision of the structure of the domain of interest" (Daraio *et al*, 2016). Our experience shows that while natural language processing (NLP) techniques are critical for linking ontologies with large datasets and extracting from the latter robust evidence, nevertheless some key design choices on the ontology and its application to data are basically of an intellectual nature. This suggests that the design of robust interactions between expert-based priori knowledge and evaluation on the one hand, and the use of advanced data techniques on the other hand, is a key requirement for robust S&T ontologies.

Related Work

A large body of work has been developed to address the limitations of classification systems outlined above. These mostly rely on individual data items, and include citation analysis (Šubelj *et al*, 2016) and NLP (Van den Besselaar and Heimeriks, 2006). Recent NLP work has focused on extracting relevant information from scholarly documents[1], but this is primarily concerned with metadata and citation extraction. Other research has investigated keyword extraction from academic publications (Shah et al., 2003) and overlay maps (Rafols, 2010). The semantic web approach of Motta and Osborne (2012) in Rexplore takes scholarly data analysis a step further by examining research trends at different levels of granularity, and by finding semantic relations between authors, using relations such as co-citation, co-publication and topic similarity. However, this is again limited to only publication data, which is relatively cohesive.

Another strand of research that moves away from traditional classification systems involves modelling topics and domains in order to gain an overview of the field. Here, techniques such as LDA (Blei et al., 2003), PLSA (Blei, 2012) and KDV (Börner *et al,* 2003) are used extensively for understanding and mapping large research areas, for example to understand the evolution of topics over time (Chen et al., 2017). These techniques essentially model the distribution of topics, based on the principle that documents typically contain multiple topics according to a probabilistic distribution. However, the drawback is that it can be hard to make sense of the resulting information and to understand the nature of these unlabelled clusters and topics, and this work often has to be done manually. Furthermore, these methods do not deal well with topics outside a core subject domain, and clustering may result in multi-disciplinary topics without strong internal cohesion (Boyack, 2017).

All these techniques extract topics in a bottom-up manner from structural (in the case of citation analysis) and linguistic (in the case of NLP and topic modelling) features of documents. However, while these provide 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. Methods like LDA also work well on homogenous kinds of data, such as a large collection of publication abstracts, but cannot extract good topics that encompass the diversity of more disparate datasets. Finally, connecting such topics with relevant themes at the policy level is far from simple, since the associated terminologies are largely incompatible (Cassi *et al*, 2017).

Task

In this work, we use an ontology-based approach to address these issues. An ontology is defined as the "explicit formal specification of the terms in the domain and relations among them" (Gruber 1993), and in our case, acts as a bridge between the different and evolving vocabularies across heterogenous data sources. Our ontology is essentially a hierarchical

[1] http://csxstatic.ist.psu.edu/about/scholarly-information-extraction

representation of topics, as seen in many traditional classification systems, but with the possibility of multiple inheritance. 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 (keywords), can be connected to (different and evolving) vocabularies across heterogeneous data sources. An ontology thus offers a formal representation of a domain of knowledge that is shared by a specific group of "experts", based largely on a priori conceptual knowledge, while clustering approaches represent a more informal, data-driven view.

Challenges

Implementing an ontology involves 2 major aspects: first, the design of the ontology structure, consisting of a set of related topics and subtopics in the areas of KET and SGC; and second, a way to map relevant documents to topics in this structure (which can be seen as a problem of multi-class classification, with a large number of classes).

There are several major challenges associated with the design and use of ontologies in this scenario: both conceptual and practical. First, the ontology structure is hard to define because it is not clear what level of precision is both needed and practical, and because these affect the implementation of the document-topic mapping. The structure must also be intuitive for human users to navigate, and this is perhaps the most challenging component: it must reflect both the needs of the policy makers, but also the variety of ways in which information is portrayed in the data sources (in our case comprising publications, project abstracts, patents, and descriptions of social innovation projects). This latter is also critical for the data classification task.

We have attempted to mitigate this problem by consulting experts at every stage of the process, holding workshops with policy makers from a variety of fields in order to understand their needs, and following a principled development process. However, the intrinsic vagueness of the notion of Key Enabling Technologies and especially Societal Grand Challenges means that the topics are hard to define, and there is no gold standard against which to evaluate.

Second, differences in vocabularies within academia, industry and society mean that the same concepts are typically expressed in very different ways, especially in patents which are extremely technical. Existing attempts at classification, as described earlier, have highlighted these issues. Our solution to this lies in the use of sophisticated techniques from NLP and Machine Learning, where this kind of language variation is a common problem and techniques go far beyond the simple keyword matching approach used in other work. For example, word embeddings and a plethora of similarity and distance measures are used to determine possible mappings between data and classes in the ontology.

Finally, there are numerous issues related to evaluation of such a large-scale classification. It is impossible to know if every document has been correctly classified, and almost certainly there will be errors. We mitigate this by testing in different ways and at a variety of stages in the process, checking a sample of annotated documents, looking at the global picture for incongruities (such as topics with an unexpectedly high or low number of documents), evaluating different keyword generation strategies, and tweaking the ontology where needed.

Approach

The method we adopt in this work comprises 3 steps: ontology creation, ontology population, and ontology-based classification (data annotation). All three steps require human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced language analysis techniques. Consequently, if any changes are deemed necessary, the process can easily be rerun and the data re-annotated within a short period of

time and in a principled way. The current version of the ontology contains 150 topics based around the 6 KETs and 7 SGCs, and around 8,700 unique keywords.

Ontology creation

The ontology is defined according to the two strands of KET and SGC. We take as a starting point some existing classifications, which we merge and map, such as the mappings between IPC (International Patent Classification) codes and both KETs (Van der Velde, 2012) and SGCs (Frietsch et al., 2016). For KETs, we also make use of the structure implemented in the nature.com ontologies portal (Hammond and Pasin, 2015). Some of these topics are already connected to DBpedia and MESH, which provides us with an additional source of information for keywords. We manually refine this structure, removing the lower levels, to make a slightly more generic set of topics. We also create subclasses based on EU policy documents, which describe how the KETs and SGCs are structured. A key expert decision relates to the extent of overlap between classes and subclasses, as some KETs are intrinsically related.

Ontology population

Having created an initial structure containing the concepts (topics and sub-topics), the ontology then needs to be populated with instances (keywords) from various data sources. These instances help us to: (1) match user queries to topics in the ontology; and (2) match documents from the various databases to these topics. These two issues form the crux of the system.

The first stage consists of automatically generating key terms from the ontology class names and associated information, such as class descriptions, using Automatic Term Recognition techniques. Additional terms are manually generated by experts where information is sparse or where there is possible ambiguity. Terms in which the experts are highly confident are designated "preferred" and are used as seed terms for the expansion stage. These are typically the topic name itself, synonyms or linguistic variants of it, and additional manually generated terms. For example, one preferred term for the topic "intelligent transport" is "intelligent navigation". The remaining (non-preferred) terms are the automatically generated ones, and are only used for the matching stage later. These have a lower weighting during the matching, since we are less confident about their relevance. An example of a non-preferred term for the topic "intelligent transport" is "radar tracker" (which is somehow connected with the topic but is not a close synonym). This term might be relevant if found in conjunction with another relevant term for the topic, but not on its own.

The second stage involves the generation of additional keywords. First, our preferred terms are used to generate a seed set of initial keywords associated with each ontology class. We then find semantically similar terms to these using word embeddings trained on a large corpus of just over 8.3 million documents, comprising a mixture of publications, project descriptions, policy documents and patent abstracts. This corpus will be extended periodically as additional data becomes available – while larger corpora may provide better training, there is a tradeoff between this and the relevance of the documents (our previous experiments showed that using larger corpora of pre-trained embeddings on more general corpora gave worse results). The method consists of extracting a set of possible terms from that corpus using Automatic Term Recognition and NLP techniques, and then finding the ones most similar to the seeds. Finally, the terms are scored according to how "representative" they are of that class, and prior probabilities are generated using PMI for term combinations, based on frequency of co-occurrence in the training data.

The implementation of this process showed that automatic techniques enable the generation of a large number of keywords, but become problematic when two subclasses share some similar terms (like rail and road transport). Additional statistical techniques can be used to further weight terms based on maximising the semantic distance between terms from such closely

related classes, but some level of expert intervention is nevertheless required in order to delimit the subclasses and to attribute a sufficient number of distinct terms to each of them.
The result of the ontology population stage is thus a set of keywords associated with each class, each of which has a score indicating the degree of its relevance to that class. These keywords are used for the mapping between documents and topics in the final data annotation stage.

Data annotation

The data sources take the form of 4 databases containing information about projects, patents, publications and social innovation respectively. The idea of the annotation is to link each data element (e.g. a project) with the relevant topic(s) in the ontology, so that indicators (and from there visualisations) can be built around these. For example, in order to know how many EU projects there have been about "gene therapy" in a particular year and location, we must first know which projects should be associated with this topic.
We have developed a classifier which takes documents as input and returns information about the class(es) to which each is linked, and a score for it. The scores are based on (i) the weight of that keyword for that class (e.g. preferred terms have a higher score, as do terms ranked close in similarity to these); (ii) the combination of keywords found in the document using PMI calculations from the ontology population stage (on the assumption that term combinations with high PMI are better indicators that the class is relevant for that document); (iii) subclass boosting, whereby keywords belonging to a more specific class in the ontology are to be preferred over more general ones.
The process of classification thus assigns multiple possible topics to each document, not all of which are likely to be useful as some will be low-scoring. Thresholds are used to decide which of the topics are most relevant, based on analysis of distributions and some inspection of results. This is a typical expert-based task that involves manual checking of classified documents to find a reasonable balance between recall and precision.

Discussion and Conclusions

In this work, we aim to address some of the current limitations in applying traditional classifications to a science policy domain, through the use of ontologies, thereby extending the reach of existing text-based methods while still maintaining the power and rigour of classification systems. In particular, we overcome the problems in connecting policy-based topics with science-based topics. This wide-ranging view of the research domain requires the focus to shift from static maps and detailed analyses towards indicators that can be compared temporally, geographically, and across topics.
Our approach is designed to maximize automated processes wherever possible, which is not only critical for dealing with massive volumes of data, but also lends itself to domain and topic adaptation. Since research is not static – topics change over time, new terminology comes to the fore, and even geographical boundaries do not remain the same – this enables much greater flexibility than many existing classification-based systems. Changes to the ontology or the input of new research data can easily be handled automatically, and updates pushed seamlessly to the central databases from which visualisations are generated. On the other hand, these are tempered by expert intervention at critical stages.
There are, however, limitations. Rigorous evaluation is always difficult, and requires manual intervention, which is time-consuming and subjective. The use of NLP techniques brings its own problems: language is tricky for machines to understand, and tools will never be 100% accurate. Numerous issues in terminology extraction still need to be solved globally: many terms are ambiguous and require at the least context, and in some cases, only the kinds of world knowledge that humans can provide. Nevertheless, this work provides some critical new pathways for STI technologies, which open up avenues for future research directions.

Acknowledgements

This work was partially supported by the European Union under grant agreement No.726992 KNOWMAK and grant agreement No. 825091 RISIS.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, *55*(4), pp.77-84.

Börner, K., Chen, C. and Boyack, K.W., 2003. Visualizing knowledge domains. *Annual review of information science and technology*, *37*(1), pp.179-255.

Boyack K (2017) Investigating the Effect of Global Data on Topic Detection. In l¨aser, J., Scharnhorst, A. & Gl¨anzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, *Scientometrics*, *111*(2), 2017, pp.999-1015.

Cassi, L., Lahatte, A., Rafols, I., Sautier, P., & De Turckheim, E. (2017). Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics*, *11*(4), 1095-1113.

Chen, C. (2017). Expert review. Science mapping: a systematic review of the literature. *Journal of Data and Information Science*, 2(2), 1-40

Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). Data integration for research and innovation policy: an Ontology-Based Data Management approach. *Scientometrics*, *106*(2), 857-871.

Debackere, K., & Luwel, M. (2004). Patent data for monitoring S&T portfolios. In *Handbook of Quantitative Science and Technology Research*(pp. 569-585). Springer, Dordrecht.

Frietsch, R., Neuhausler, P., Rothengatter, O., Jonkers, K.: Societal grand challenges from a technological perspective: Methods and identification of classes of the international patent classification IPC. Tech. report. Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis (2016).

Gruber, T. (1993). What is anOntology. *http://www-ksl. stanford. edu/kst/whatis-an-ontology. Html*

Hammond, Tony, and Michele Pasin. "The nature. com ontologies portal." *5th Workshop on Linked Science,* 2015.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362.

Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, *101*(2), 1535-1551.

Maynard, D. and Lepori, B. Ontologies as bridges between data sources and user queries: the KNOWMAK project experience. *STI 2017*, Paris, France, September 2017.

Motta, E. and Osborne, F. Making sense of research with Rexplore. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track*-Volume 914 2012 Nov 11 (pp. 49-52). CEUR-WS. org.

Rafols, I., Porter, A.L., Leydesdorff, L.: Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for in- formation Science and Technology* 61(9), 1871–1887 (2010).

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, *1*(0), 100.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords?. *BMC bioinformatics*, *4*(1), 20.

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PloS one*, *11*(4), e0154404.

Van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377-393.

Van de Velde, E.: Feasibility study for an EU monitoring mechanism on key en- abling technologies. IDEA Consult (2012.