Towards an Infrastructure for Understanding and Interlinking Knowledge Co-Creation in European research

Diana Maynard[1], Adam Funk[1], and Benedetto Lepori[2,3]

[1] Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield, UK
[2] Faculty of Communication Sciences, Universit della Svizzera italiana, 6904 Lugano, Switzerland
[3] Laboratoire Interdisciplinaire Sciences, Innovations et Socits (LISIS), University of Paris Est, 77454 Marne-la-Valle Cedex 02, France

Abstract. This paper describes the initial development of an infrastructure for understanding and visualising knowledge co-creation in European research. Datasets containing information about projects, publications and patents are enhanced with semantic information enabling indicators to be generated, that are used to inform users about the state of art in the research area and new trends. This helps to resolve the problem of increasing complexity and multidisciplinarity of emerging scientific and technological research. Ontologies and Semantic Web techniques play a central role in mapping between topics, data, and user queries so that information can be enhanced, interlinked and aggregated.

Key words: ontologies, linked data, knowledge co-creation, NLP

1 Introduction

Mapping the development of Science and Technology (S&T) has become an issue of increasing societal and political relevance and, at the same time, of increasing complexity in terms of measures and indicators. On the one hand, EU policy is focusing on useful knowledge, in terms of (1) Key Emerging Technologies (KET), which foster growth and economic development, and (2) knowledge production to respond to so-called Societal Grand Challenges (SGC), i.e. those challenges considered critical for the future of European society [1]. On the other hand, knowledge production no longer comes from one source (i.e. universities), but emerges from multiple sources. Furthermore, knowledge production processes are becoming more and more hybrid, crossing disciplines (especially for key technologies), institutional borders (especially for societal challenges), and geographical borders (internationalisation).

These developments imply a number of challenges such as integrating a broader scope of data sources, the construction of more flexible classification systems, and more comprehensive coverage of actors involved in knowledge co-creation. Traditional indicators on S&T development fall short of addressing

these challenges for various reasons: they tend to be aggregated and focused on traditional research outputs, like publications and patents, and therefore cannot adequately identify new places of social innovation and the outreach of research; they adopt stable and, usually, rather coarse classification schemes (like field of science or ISI subject categories), which do not allow mapping the emergence of new knowledge domains at the intersection of existing scientific fields; third, by the nature of the data used, tend to be backward-looking and thus difficult to use for policy decisions.

Innovative approaches have been developed in order to overcome some of these limitations: this includes the use of linguistic techniques and overlay maps to identify clusters of publications or patents associated with a specific topic [14, 12]; the use of altmetric data to provide a broader view on research impact [17] and analyse user attention on scientific activities through social media [13]; and finally, content analysis performed on research project descriptions, as these are much nearer to the frontier of knowledge production.

However, one of the major problems lies in making connections between the different kinds of data, and in the provision of cross-cutting search facilities that enable information from them to be combined. Policy makers often struggle to get a good picture of the fast-changing nature of research, because information is dynamic, conflicting and hard to understand. In particular, the language used can differ widely, and this restricts search usefulness.

The goal of the recently-launched H2020 project KNOWMAK (Knowledge in the Making in the European Society) is to address these challenges by developing 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 domains (topics), institutions and locations. The central aspect of this system is a set of ontologies, which form the bridge between knowledge sources and user queries. These ontologies are focused on two research areas: Societal Grand Challenges (SGC), and Key Enabling Technologies (KET). At the heart of this is the quadruple helix model of innovation, linking government, industry, academia and civil participants.

On querying the KNOWMAK platform, users will be able to identify via the ontologies some related concepts, including those from the scientific and technological vocabulary, and to link them with the relevant markers of knowledge production, specifically publications, patents, research projects and social innovation projects. In this way, they will be also able to visualize maps of space and actors relevant for that topic, and to zoom in onto specific research projects and outputs relevant for their topic.

2 Related Work

KNOWMAK builds on work performed in the EU RISIS project[4], which is creating a distributed infrastructure for research and innovation data and poli-

[4] http://risis.eu

cies, 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. Critically, the addition of the ontological element, which is lacking in RISIS, will enable better integration of all the different resources, and improved handling of the complexities of language.

While there is a long tradition in using ontologies for addressing policy issues (see for example [7]), KNOWMAK is faced with a number of specific issues. First, KET and SGC are themselves rather general terms. and a clear and consistent delineation from policy documents cannot be expected. Second, KNOWMAK ontologies need 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, KNOWMAK is developing 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 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.

The use of text mining techniques for understanding and enhancing innovation has been studied extensively by [4] and [6], who found that traditional sources of information about R&D activities (reported on company websites and in databases of patents and publications) could be enhanced significantly by such methods. However, they only used quite simple methods of analysis, such as regular expression-based keyword search. In the highly dynamic field of NLP, [2] used predictive modelling to predict the key technical NLP terms of the future. Other work has investigated the need for combining information from related fields to populate domain-specific ontologies, e.g. in the field of metabolomics [15]. Previous semantic annotation work has demonstrated the power of combining text mining and ontologies to discover and link information from large-scale documents such as patent data [16], archived material [10] and social media [11].

Another starting point is the work of [3], who constructed a mapping between the IPC patent classification codes and 6 of the 7 Societal Grand Challenges. This mapping forms the basis for part of our ontology. Furthermore, they developed a set of subcategories for each SGC, through a comprehensive expert-based process, including feedback and validation from the European Commission. However, this creates new problems, since the mapping is still evolving, and the subclasses generated for the SGCs do not necessarily match our needs. Our ontology needs to relate to the language and needs of the users, the EU policies, and the projects, patents and publications data. It also needs to be easily extendable as new kinds of information sources are added: later in the project, information collected from social media will be added, which will almost certainly be expressed in different kinds of language. Furthermore, beyond the life of the project, the kinds of information covered could extend both beyond European research to international research, and beyond the SGC and KET areas to other fields of research.

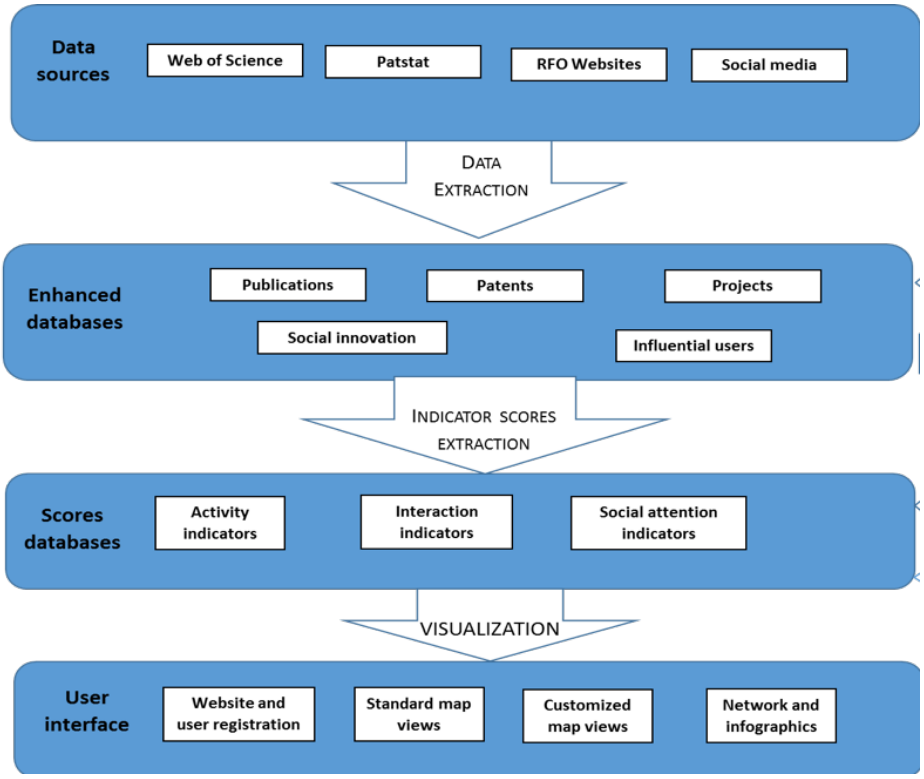

Fig. 1. KNOWMAK platform architecture

3 The KNOWMAK platform

Figure 1 shows the basic architecture of the framework, connecting the user queries (at the bottom of the diagram) to the information in the databases (at the top of the diagram) via a set of enhanced databases and indicators (in the middle) which allow users to explore the data. The enhanced databases connect relevant information from the projects, patents, publications and policies to the ontologies, which in turn allows indicators to be created that can provide answers to users' questions and a means to visualise associated data.

One of the major problems with linking the data sources to the ontologies is that the language is very different in each of these kinds of data. This means that finding relevant instances of the ontology classes in the data (semantic annotation) is tricky. Standard NLP techniques for ontology population operate on principles such as term recognition [8], but terms in these documents may be completely different and may also change over time. We will discuss this further in Section 4. Another problem is that the information in the original databases is not always standardised or correct. For example, the same name or term is written differently, mistakes occur and so on. In both patent and publication

databases, this is a long-standing issue, and although attempts have been made to address this through (a) database cleaning and enhancement and (b) use of NLP and ML techniques to match and link names and terms, it is far from resolved.

4 Ontology development

The ontology acts as a bridge between the users and the data stored in our knowledge bases, enabling dynamic linking between them. Documents are annotated with one or more classes from the ontology, linking them with topics which are used, along with geographical and actor information (at which location was the research done, and by which institution(s)) to build indicators about the data. The ontology is designed via an expert process, taking into consideration the needs of the users (following dedicated workshops) and previous existing classifications and protocols, following a principled development process. The ontology may be constantly refined depending on new information: the databases get updated, terms are dynamic, and user needs change. A first version of the ontology is available online[5], although it is in continuous update.

From a methodological perspective, two main issues need to be addressed. First, the structure of the ontologies is problematic. Users may search for quite general topics like "Societal Grand Challenges" which need to be associated with a more fine-grained vocabulary to allow for the identification of relevant data. Since the two areas of KET and SGC are rather different, they are developed independently, though mappings are created between relevant classes. Because the ontologies form the central hub of the system, they must cater to the needs of both the users (i.e. to cover the topics that they might want to search for) and the research domain. Through the use of sophisticated NLP and term recognition technologies linking the sources to the ontologies, we aim to help mitigate the problem of different terminologies. In the past, only use of keywords and simple matching algorithms to the text has typically been used. More complex document similarity, domain modelling and linguistic techniques will therefore be used to complement these strategies.

The second issue is the annotation of the data with the ontologies (semantic annotation). The difference between the language of the policies and data is again critical. Related terms are identified using methods for capturing lexical ambiguity and variation [11], and for the identification of related terms (e.g. semantic frames, word embeddings and ontology generation techniques [9]). Additionally, the documents retrieved by applying ontologies will be used as seeds in order to extend the search, either by textual similarity or by exploiting the internal structure of the data, for example with the use of overlay maps to identify clusters of publications or patents associated with a specific topic [14].

We take as a starting point some existing classifications, which we merge and map. For example, there exist already mappings between IPC (International Patent Classification) codes and both KETs and SGCs [18, 3]. This has

[5] `https://gate.ac.uk/projects/knowmak/`

the advantage that when new patents are added to the database, they can be automatically connected to the ontology. However, as [18] discovered, these are not sufficient to cover all cases, so an annotation process is still necessary to supplement this.

For the KET ontology, we also make use of the structure implemented in the nature.com ontologies portal [5]. This provides a repository for the semantic schemas driving the Nature.com publishing platform and datasets, comprising a common network of inter-related and constantly evolving ontologies. Three of the KETs are represented here: nanotechnologies, industrial biotechnology and photonics. Some subjects are also connected to DBpedia and MESH. Linking with the nature.com ontology helps with mapping publications, and enables future extension of our ontology to other topics. For the other 3 KETs, we manually create subclasses based on the EU policy documents which describe how the KETs are structured.

One major issue is that the topics covered in the 6 KETs overlap considerably. 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 (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).

For the projects data, we make use of the existing subject index[6] developed in the RISIS project to map some documents to our ontology (for example, "materials technology" can be mapped to the Advanced Materials KET, while "nanotechnology and nanoscience" can be mapped to the Nanotechnologies KET. Again, for the rest, we need to use text mining tools to annotate documents against the relevant ontology class.

The EU H2020 policy documents divide each of the 7 SGCs into sub-categories. For example, bioeconomy has the 4 subcategories: food security, blue growth, agricultural research and innovation, and bio-economy. Figure 2 shows a portion of the SGC ontology, including some of these mappings. For example, "A61K 39" is the patent code referring to "Medicinal preparations containing antigens or antibodies". We can also see in the ontology that "Biotech" is a subclass of both the Health and Bioeconomy SGC top-level classes.

5 Challenges and Future Directions

The role of the ontology in the KNOWMAK framework is critical, yet it is also hard to evaluate. While many ontology evaluation methodologies have been proposed, which validate the rigour, correctness and coverage of the ontology, by far the most useful method is a task-based evaluation. In other words, does the ontology fulfil the needs of the users? In terms of rigour and correctness,

[6] https://ec.europa.eu/growth/tools-databases/kets-tools/sites/default/files/library/final_report_kets_observatory_en.pdf

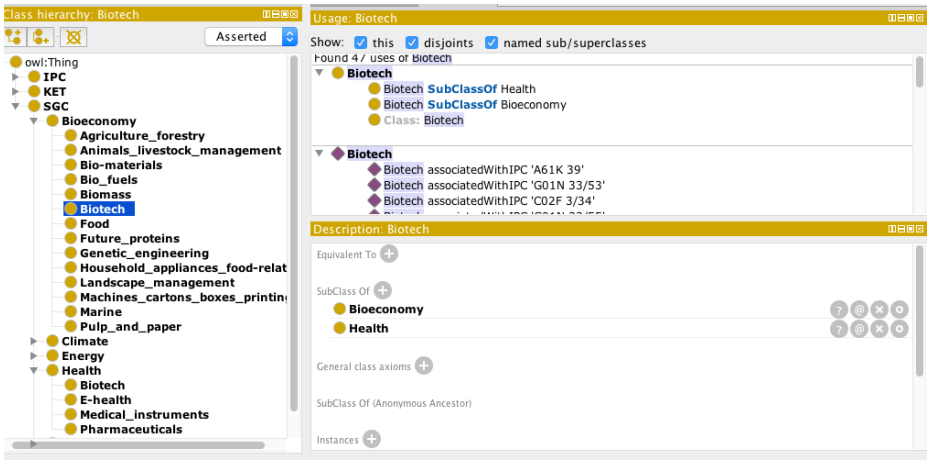

Fig. 2. Part of the SGC ontology

relying on existing established ontologies such as nature.com and the various mapping efforts goes some way to validating this. The construction process also relies heavily on continuous feedback from real users and knowledge experts: the ontology must cover the kinds of questions they want to ask, and must return useful results. This will be monitored throughout the course of the project, but it remains a risk.

In summary, the KNOWMAK project, while founded on some robust principles and rooted in some established previous initiatives, nevertheless attempts to address some very real and challenging problems. Indeed, the need for this type of work is precisely because there are so many issues in opening and linking different kinds of European research activity, and because of the dynamic nature of these emerging technologies. The framework is currently in early stages of development, and there remains much still to be done, so this paper serves primarily to set out the key motivations, general architecture design, and challenges to be overcome, and to position the work within a wider context. It is clear that the Semantic Web community can offer much in the way of existing solutions to the problems of ontology linking and population, which can help with the problem of bridging unstructured and structured content using semantic annotation techniques, and which will form part of our future work.

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

References

1. (EC), E.C.: Europe 2020: A strategy for smart, sustainable and inclusive growth. Working paper {COM (2010) 2020} (2010)
2. Francopoulo, G., Mariani, J., Paroubek, P., Vernier, F.: Providing and Analyzing NLP Terms for our Community. Computerm 2016 p. 94 (2016)

3. Frietsch, R., Neuhäusler, P., Rothengatter, O., Jonkers, K.: Societal grand challenges from a technological perspective: Methods and identification of classes of the international patent classification ipc. Tech. rep., Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis (2016)
4. Gök, A., Waterworth, A., Shapira, P.: Use of web mining in studying innovation. Scientometrics 102(1), 653–671 (2015)
5. Hammond, T., Pasin, M.: The nature.com ontologies portal. In: 5th Workshop on Linked Science (2015)
6. Kahane, B., Mogoutov, A., Cointet, J.P., Villard, L., Laredo, P.: A dynamic query to delineate emergent science and technology: the case of nano science and technology. Content and technical structure of the Nano S&T Dynamics Infrastructure pp. 47–70 (2015)
7. Loukis, E.N.: An ontology for g2g collaboration in public policy making, implementation and evaluation. Artificial Intelligence and Law 15(1), 19–48 (2007)
8. Maynard, D., Funk, A., Peters, W.: SPRAT: a tool for automatic semantic pattern-based ontology population. In: International Conference for Digital Libraries and the Semantic Web. Trento, Italy (September 2009)
9. Maynard, D., Funk, A., Peters, W.: Using Lexico-Syntactic Ontology Design Patterns for ontology creation and population. In: WOP 2009 – ISWC Workshop on Ontology Patterns. Washington, USA (October 2009)
10. Maynard, D., Greenwood, M.A.: Large Scale Semantic Annotation, Indexing and Search at The National Archives. In: Proceedings of LREC 2012. Turkey (2012)
11. Maynard, D., Greenwood, M.A., Roberts, I., Windsor, G., Bontcheva, K.: Real-time social media analytics through semantic annotation and linked open data. In: Proceedings of WebSci. Oxford, UK (2015)
12. 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 information Science and Technology 61(9), 1871–1887 (2010)
13. Robinson-García, N., Torres-Salinas, D., Zahedi, Z., Costas, R.: New data, new possibilities: exploring the insides of altmetric.com. arXiv preprint arXiv:1408.0135 (2014)
14. Rotolo, D., Rafols, I., Hopkins, M.M., Leydesdorff, L.: Scientometric mapping as a strategic intelligence tool for the governance of emerging technologies. (2014)
15. Spasić, I., Schober, D., Sansone, S.A., Rebholz-Schuhmann, D., Kell, D.B., Paton, N.W.: Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. BMC bioinformatics 9(5), S5 (2008)
16. Tablan, V., Bontcheva, K., Roberts, I., Cunningham, H.: Mímir: an open-source semantic search framework for interactive information seeking and discovery. Journal of Web Semantics 30, 52–68 (2015), http://dx.doi.org/10.1016/j.websem.2014.10.002
17. Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C.R.: Do altmetrics work? twitter and ten other social web services. PloS one 8(5), e64841 (2013)
18. Van de Velde, E.: Feasibility study for an EU monitoring mechanism on key enabling technologies. IDEA Consult (2012)