KNOWMAK

KNOWledge in the MAKing in European Society

Scientific paper on the KNOWMAK ontology

D2.5

December 2019

Diana Maynard (University of Sheffield) Benedetto Lepori (UPEM, Paris), Philippe Larédo (UPEM, Paris)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

Name of the deliverable	
Number of the deliverable	D2.5
Related WP number and name	WP2
Related task number and name	-
Deliverable dissemination level	European Commission
Deliverable due date	31 December 2019
Main author(s)	Diana Maynard, Benedetto Lepori, Philippe Larédo
Contributing partners	University of Sheffield, University of Paris Est

REVISIONS

Version	Date	Comments (Y/N)	Project partner
1			UFSD, UPEM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

1 Summary

Understanding knowledge co-creation in key emerging areas of European research is critical for policy makers wishing to analyse impact and make strategic decisions. However, current methods for characterising and visualising the field have limitations concerning the changing nature of research, differences in language and topic structure between policies and scientific topics, and coverage of a broad range of scientific and political issues that have different characteristics.

In this paper, we discuss the novel use of ontologies and semantic technologies as a means 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.

Reference: Maynard D., Lepori B., Larédo Ph. (2019), Using ontologies to map between research data and policymakers' presumptions. The experience of the KNOWMAK project, Scientometrics, draft paper.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

2 Introduction

In recent years, a priori classification systems for science and technology, such as the Field of Science Classification (OECD, 2002) and IPC codes for patents (Debackere and Luwel, 2004), have been increasingly replaced by data-driven approaches, relying on the automated treatment of large corpora, such as word co-occurrences in academic papers (Van den Besselaar and Heimeriks, 2006), clustering through co-citation analysis (Šubelj et al., 2016), and overlay maps to visualize knowledge domains (Rafols et al., 2010). These approaches have obvious advantages, since they are more flexible to accommodate the changing structures of science, and are able to discover latent structures of science rather than impose a pre-defined structure over the data (Shiffrin and Börner, 2004).

Yet, when the goal is to produce indicators for policymakers, purely data-driven methods also display limitations. On the one hand, such methods provide very detailed views of specific knowledge domains, but are less suited to large-scale mapping across the whole S&T landscape. On the other hand, lacking a common ontology of S&T domains (Daraio et al., 2016), such mappings are largely incommensurable across dimensions of knowledge production. Even more importantly, data-driven methods do not allow presumptions of categories used in the policy debate to be integrated in the classification process. Such presumptions are largely implicit and subjective, implying that there is no gold standard against which to assess the quality and relevance of the indicators, but these are inherently debatable (Barré, 2001a).

In this paper, we report on how these challenges have been addressed to develop a web-based tool providing interactive visualizations on European research and focusing on key categories in the European research policy debate, namely Key Enabling Technologies (KET) and Societal Grand Challenges (SGC)[1]. Our approach was based on two main elements: a) the design of an ontology of the KET and SGC knowledge domains, in order to make explicit their content and to provide a common structure across dimensions of knowledge production; and b) the integration between natural language processing (NLP) techniques to associate data sources with the ontology categories on the one hand, and expert-based judgement in order to make sensible choices for the matching process on the other hand. This drove to a recursive process where the development of the ontology and the process of data annotation were successively refined based on expert assessment of the generated indicators. In that respect, the decomposition of knowledge production indicators by geographical spaces (countries and regions) and research actors (public and private) played a central role, as it allowed for a fine-grained assessment of results.

Our experience shows that while natural language processing 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 and closely associated with specific user needs. This suggests that the design of interactions between expert-based a 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. Our paper contributes to this endeavor by providing an in-depth knowledge of how such interactions can be managed, as well as a more precise understanding of the key choices to be made in the design and implementation of the ontology.

[1] http://knowmak.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

3 Background

A large body of work has been developed to address the limitations of existing classification systems. These mostly rely on individual data items, and include citation analysis for publications (Šubelj *et al,* 2016) and NLP (Van den Besselaar and Heimeriks, 2006). Recent NLP work has focused on extracting relevant information from scholarly documents[2], but this primarily involves 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 publication data, which is relatively cohesive.

Shallow NLP techniques have also been used to map topics and to enhance traditional sources of information about R&D activities, e.g. those reported on company websites and in patents and publication databases (Gok et al. 2015, Kahane et al., 2015). However, the focus here was on using regular expression-based keyword search to group similar terms, rather than on complex linguistic analysis. The use of sophisticated NLP techniques to model terms has a long-established history in the computational terminology field, however, and advances in machine learning and computational power have enabled great strides (Amjadian et al., 2016). Predictive modelling has also been used with some success to predict the key technical NLP terms of the future (Francopoulo et al., 2016).

The second main strand of related research involves modelling topics and domains in order to gain an overview of S&T fields. Here, techniques such as LDA (Blei et al., 2003), PLSA (Blei, 2012) and KDV (Börner *et al.,* 2003) are used for mapping 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 contain multiple topics according to a probabilistic distribution. Topics are based on clusters of terms, and thus documents can also be clustered together according to similarity of the topics exhibited. However, the drawback is that it can be hard to make sense of the resulting information and to understand the nature of clusters and topics, and this work often has to be done manually. Unlabelled clusters can group together similar documents, but these cannot be automatically mapped to a set of specific and stable topics. This is critical for producing suitable end-user visualisations and addressing policymakers' needs – a too large set of topics that is not properly structured will be unusable. Furthermore, if new documents are added to the system, there is a risk that the clusters will change, and documents may be classified differently, leading to an instability which is incompatible with our goals. Finally, these methods do not deal well with topics outside a core subject domain, since they are designed to work on homogenous datasets, and clustering within a broad domain may result in sets of 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 they provide detailed views of specific knowledge domains and of their evolution over time, they are currently less suited to large-scale mapping across the whole S&T landscape. 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).

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

An alternative approach is to rely on ontologies, defined as the "explicit formal specification of the terms in the domain and relations among them" (Gruber ,1993). 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 kind of expert-based arbitration relying on a "shared vision of the structure of the domain of interest" (Daraio et al., 2016).

An ontology is essentially a hierarchical representation of topics, but with the possibility of multiple inheritance (a topic can be represented as a subclass of more than one class). While keeping the presence of a core set of subjects organized in layers, ontologies are more flexible in structure. On the audience side, ontologies are a means to translate questions of interest, frequently expressed in generic terms in policy documents, into a formal structure of classes and keywords. On the data side, through instances (keywords), ontologies can be connected to different and evolving vocabularies across data sources. Ontologies thus effectively work as a bridge between (policy) questions and heterogeneous data sources (Figure 1).

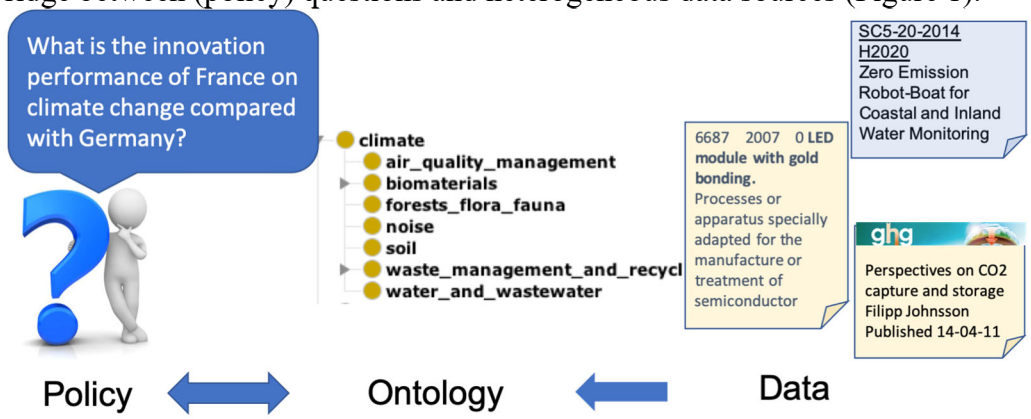


Figure 1: the role of an ontology in connecting policy-related questions from users with data sources

Ontologies have long been used to address policy issues, e.g. (Loukis, 2007), and the addition of semantic annotation tools which link texts to an ontology is also far from new (Maynard et al., 2016). Other Semantic Web research has also 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). Previous work using semantic annotation has demonstrated the power of combining text mining and ontologies to discover and link information from large-scale documents such as patent data (Tablan et al., 2015), archived material (Maynard and Greenwood 2012), and social media (Maynard et al., 2017). Attempts have also been made to use ontologies for mapping research to more generic societal problems, but these have typically focused on small hand-crafted ontologies in a particular domain (Estanol et al., 2017).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

4 Ontology design and implementation

Ontology development in our application involves three major aspects: first, the design of the ontology structure, consisting of a set of related topics and subtopics in the relevant subject areas; second, populating the ontology with keywords; third, classifying documents based on the frequency of keywords.

The mapping process can be seen as a problem of multi-class classification, with a large number of classes, and is achieved by relying on source-specific vocabularies and mapping techniques that also exploit (expert) knowledge about the structure of individual data sources. This is not a one-off process, but an iterative one, based on co-dependencies between data, topics, and the representation system. Our initial ontology derived from policy documents was enriched and customised, based on the outcome of the matching process and on expert assessment of the matching results. Eventually, the original ontology classes may also be adapted based on their distinctiveness in terms of data items. Such a staged approach, distinguishing between core elements that are stabilized (the ontology classes) and elements that are dynamic and can be revised (the assignment of data items to classes), is desirable from a design and user perspective. Therefore, the approach is highly flexible, for example to respond to changes in policy interests, and scalable since new data sources can be integrated within the process whenever required.

All three steps require human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced NLP techniques. Consequently, if changes are deemed necessary, the process can easily be rerun and the data re-annotated within a reasonable period of time.

4.1 Ontology design

The ontology is defined according to the two strands of KET and SGC. This has implications because there is inherent overlap, not only between these two domains, but also within them. For example, within SGC, the topics of energy and climate change are closely intertwined, while much current research on transport is connected with sustainability. While KET topics focus primarily on technological research, there are clear overlaps with the “social” topics of SGCs, which often require technological solutions.

Therefore, a good structure is hard to define because it is not clear what level of precision is necessary and practical, and because these affect the implementation of the later stage of document-topic mapping. Moreover, the intrinsic vagueness of the notion of KETs and especially SGCs means that the topics are hard to define, and there is no gold standard against which to evaluate.

The structure must also be intuitive for human users to navigate, and this is perhaps the most challenging component. Moreover, ontologies must be dynamic: new terms and definitions continuously emerge from researchers and standardization groups, while other terms may become irrelevant or replaced by more popular synonyms. This means that continuous updating of existing ontologies is required, through reference to new documents.

We have attempted to mitigate these problems 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.

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

topics are already connected to DBpedia and MESH, which provides us with an additional source of information for keywords. Linking with the nature.com ontology helps with mapping scientific publications, and enables future extension of the ontology to other topics. A collection was also made of relevant EU policy documents, which describe how the KETs and SGCs are structured (Maynard and Lepori, 2017), followed by an iterative process of annotating documents and looking for missing topics.

However, initial experimentation made it clear that relying heavily on pre-existing classifications was impractical – not only due to the huge number of topics, but more importantly because these classifications were very different (and no single classification covered all topics), so that the classes in the ontology were unevenly distributed and varied greatly in coverage. Furthermore, aligning elements from different origins led to a number of inconsistencies and duplications. We therefore manually refined this initial structure, removing the lower levels, reconfiguring branches, and adding additional topics where needed, in order to develop a more balanced classification system and to cover expert-based assessment of the relevant topic. For example, the inclusion in the KNOWMAK tool of a set of social innovation projects led to an expansion of the relevant topics in areas such as 'education' or 'employment', as those inherited from EU policy documents were not considered to cover social innovation adequately.

The first version of the ontology contained 4 levels of categorisation and a total of 457 topics, which is impractical for user selection. The refinement process has left us with a set of 150 topics in 3 levels - the first containing the distinction between KET and SGC, the second containing the major 13 topics belonging to them, and the third containing the major subtopics - e.g. "society" is divided into topics such as "housing", "education" and "employment". This classification is deemed distinctive enough to be interesting for policymakers without making the choices too specific. The latter has an impact on quality, because it is harder to allocate documents to topics at very precise levels, but also on usability of the system.

A key expert decision relates also to the extent of overlap between classes and subclasses, as some are intrinsically related. For example, the KET Advanced Manufacturing. is deliberately designed to be crosscutting across the other 6 KETs, so its direct subclasses include "Advanced Materials for Manufacturing" (which overlaps with the "Advanced Manufacturing" KET). While the use of an ontology in some sense fundamentally addresses this problem of overlap, on the other hand the topic classification method essentially relies on matching each document with the best fit to a class. For this to work effectively, classes must be as distinct as possible. We aim for a middle ground whereby we enable the possibility for classification in multiple topics where required, but minimize the degree of overlap in the ontology itself.

4.2 Ontology population

The ontology needs to be populated with instances (keywords) from various data sources, which help to: (1) match user queries to topics in the ontology; and (2) match documents from the various databases to these topics.

In the KET domain, until now topic definitions have been mostly based on keywords in papers; however, this is not sufficient and these definitions need to consider also other kinds of documents and references. Furthermore, terms used by policymakers may not correspond to the actual keywords used in the data sources, and even between the different types of data source, terms vary widely. For this reason, we develop a series of constraints in order to mitigate this.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

SGCs offer a particular set of terminology-related problems, because keywords are often less technical and more ambiguous than those belonging to KET topics. For example, a related keyword for the topic of "education" could be "learning" but this occurs frequently in relation to other topics; similarly, "skill" is indicative of the "employment" topic but occurs in many unrelated documents.

Concerning the mapping of data sources to the ontology, differences in vocabularies within academia, industry and society mean that the same concepts are typically expressed in different ways, especially in patents, which are extremely technical. Existing attempts at classification, as described earlier, have highlighted these issues. Our solution 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.

Following a series of initial experiments, the solution adopted involves multiple layers of keyword extraction and a mixture of automated techniques interspersed with expert knowledge at key junctures. First, a small set of specific high-quality keywords is selected manually for each topic (typically around 5 per topic). These *key* terms are used, together with the *preferred* terms for each class (automatically derived from the class name or a linguistic variant) as seed terms for the expansion stage later. For example, a key term for the topic "intelligent transport" is "intelligent navigation". An additional source of keywords comes from the subject index of the EU-FP project database, which we have mapped to our ontology.[3]

The next stage consists of automatically generating further terms from the ontology class names and associated information, such as class descriptions, using Automatic Term Recognition techniques (Maynard et al., 2007). These terms are known as *generated* terms, and are only used for the matching stage later, where they have a lower weighting, since we are less confident about their relevance or because they may be ambiguous. An example of a non-preferred term for the topic "intelligent transport" is "radar tracker". This term might be relevant if found in conjunction with another relevant term for the topic, but not necessarily on its own.

Initial experiments with generating keywords automatically were largely unsuccessful for two reasons: first, this information was very inconsistent (some classes had detailed descriptions while some had none), and second, many important keywords were missing, even with the addition of information extracted from external knowledge sources such as Wikipedia. Furthermore, term extraction tools could not sufficiently distinguish between high quality (specific and distinct) keywords from more general ones, resulting in the same keywords being extracted for a large number of classes. Previous approaches to mapping documents to topics based on keywords, especially in the patent domain (e.g. Gok et al., 2015), have been focused on a very specific domain and thus the keywords have been manually selected, which is not feasible here. It is clear that some expert intervention is necessary in order to ensure high quality.

To resolve these issues, first, a stop list was manually created in order to prevent generic keywords (e.g. "method") being selected. Furthermore, at every stage, multi-word terms are preferred, as these are better at distinguishing between similar topics. Then, an automatic keyword enrichment method was used to boost the number of keywords, based on a large collection of training material (2.6 million documents containing a mixture of patent, project and publication abstracts as well as EU policy documents), from which we extracted new candidate terms. A set of domain-specific word embeddings was trained for these terms, with

[3] This mapping is publicly available: https://gate.ac.uk/projects/knowmak/mappings-eupro-knowmak-ontology.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

vectors for both single-word terms and multi-word terms. These embeddings were then used to find the similarity between the seed terms and new terms, and to decide which new terms to keep, as well as which topic to map them to.[4] Finally, the terms were scored according to their "representativeness" of that class, and prior probabilities generated using Pointwise Mutual Information (PMI) for term combinations, based on frequency of co-occurrence in the training data. These were used in the final classification stage, in order to ensure that more representative terms got a higher weighting, and to avoid outliers getting ranked too highly: some keywords are only good indicators when they occur together in the same document as another keyword. For example, the term "packaging" could refer to many topics, but when found with the term "microelectronics" it is a good indicator of various subtopics of Micro- and Nano-Engineering. A major challenge with the keyword enrichment process is that there is no gold standard with which to compare the results, so manual judgements must be made about the best method of defining the similarity and cut-off thresholds. Starting from a set of 2,122 ontology keyword/class pairs, 11,814 new keyword/class pairs were generated, before a second stopword list was applied, to produce a final set of 9,076 pairs. This stopword list was developed based on manual judgement and contains keyword-concept pairs which should not be matched (for example, "shipyard" is not a good keyword for the topic "aeronautics" but it is for "maritime transport").

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 (see Table 1). There is some overlap because occasionally, the same keyword appears in a higher-level class and one (or more) of its subclasses. *Preferred* terms are automatically generated from the class label and are usually similar to, or the same as, the class name itself. *Key* terms are the additional terms manually generated by experts, or which come from other knowledge sources such as DBpedia. Both are considered to be high quality (though they are also manually checked), are used as input for the term enrichment process, and are given a higher weighting during the annotation process. *Project* terms come from existing project keyword classifications. *Generated* terms are those created by the term extraction tool, while *enriched* terms come from the automatic enrichment process. These may be of lower quality and get a lower weighting.

[4] http://downloads.gate.ac.uk/knowmak/embeddings201812.txt.gz

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

	Topic	Key	Preferred	Project	Generated	Enriched	Total
KET	Advanced Manufacturing Technology	40	15	0	7	33	95
	Advanced Materials	39	8	0	28	583	658
	Industrial Biotechnology	110	35	2	852	1515	2514
	Micro- and Nano-electronics	35	22	0	12	378	447
	Nanoscience and technology	105	15	0	291	535	946
	Optics and photonics	85	15	0	249	689	1038
SGC	Bioeconomy	78	15	7	0	431	531
	Climate change and the environment	151	16	4	0	316	488
	Energy	30	25	1	6	330	392
	Health	81	22	4	10	446	563
	Security	36	11	0	0	376	423
	Society	289	29	7	5	916	1246
	Transport	57	14	2	0	202	282
	Total	1136	242	27	1460	6750	9076

Table 1: Number of each type of keyword for the high-level topics

4.3 Document classification

Our data sources comprise three major datasets on S&T made available within the RISIS Horizon 2020 infrastructure project[5]: the Web of Science version at CWTS, University of Leiden (about 30m. publications), the PATSTAT version at IFRIS in Paris (2.37m. patents), and the EUPRO database of European Framework Programme projects (67,475 projects), all from the period 2000-2017. 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 can be built around them.
Due to availability and licensing restrictions, we only have access to titles, abstracts and some internal classification (such as IPC classes for patents). This limits data available for training, and might affect the matching of keywords, as previous findings have shown that, while the abstract has the best ratio of keywords, neglecting the rest of the paper might lead to the omission of important relevant terms (Shah et al. 2003). We also currently only consider documents in English, which limits the patent collection.
Our classifier takes documents as input and returns information about the class(es) to which each is linked, along with a score, based on (i) the weight of that keyword for that class (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; (iii) subclass boosting, whereby keywords belonging to a more specific class in the ontology are preferred over more general ones.

[5] http://risis2-eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

The classification process assigns multiple possible topics to each document. Thresholds are used to decide which of the topics are most relevant, as the ontology is used to build aggregated indicators at the regional and/or topical level. This is a typical expert-based task that involves manual checking of classified documents and distribution analysis to find a reasonable balance between recall and precision. Different approaches for thresholding have been tested, resulting in a simple criterion assigning documents to classes with a score above the median of the whole set of documents, which works reasonably well, but there is admittedly room for fine-tuning the scoring approach in the future.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

5 Results and evaluation

Lack of suitable frameworks within which to evaluate topic classification methods and tools is a well-known problem, since gold standards cannot easily be produced for the massive datasets typically used. As discussed by Velden et al. (2017), there is also a general lack of understanding of how different methods affect the results obtained. We cannot directly compare our ontology or classification tool with others, since there are no other tools able to classify the same set of topics and document types, and it is impossible to know if every document has been correctly classified.

We have followed the methodology for ensuring the quality and validity of an ontology known as Ontology Design Principles (Suárez-Figueroa et al., 2012). This comprises the following steps: (1) select the most suitable ontological resources to be reused; (2) carry out the ontological resource re-engineering process to modify the selected ontological resources; (3) assess if the modified/new ontology fulfils the ontology requirement specifications.

According to both these principles, the quality and effectiveness of an ontology should be considered primarily in the context of its intended use, rather than in isolation. This helps avoid the inevitable subjectivity and/or inherent biases: there is no use to an ontology except within an application. Just as the notion of indicators has moved away from the traditional statistical fixed approach, and is now widely adopted as a social construct composed of customised, interoperable, and user-driven components (Lepori et al., 2008), so the notion of ontologies should be interpreted within the wider framework of the actors in the policy debate.

In practical terms, we have assessed whether the ontology fulfils the requirements by involving experts at the key stages of the development and testing process. This includes checking that users understand and are satisfied with the ontology structure and iteratively refining it according to their needs (as described in Section 3); assessing the relevance and coverage of the keywords attached to the classes (as described below in Section 4.2); and a task-based assessment of the ontology (described below in Section 4.3), involving checking that there is minimal overlap between class assignment and that all classes have sufficient – but not too many - documents assigned.

5.1 Keyword evaluation

The quality of keywords is critical for the working of the annotation process. To evaluate them, we consider (1) statistical representation of topics and keywords; and (2) intrinsic keyword quality evaluation, by manually checking the quality of a selection of the keywords, representatively sampled.

We look first at the distribution of keywords to class, which shows how well the class is represented (the more keywords, the better the chance of a match, but this leads to inaccuracies if the keywords are not of adequate quality). In the first version of the ontology, there were 3,854 unique keywords. With 448 unique classes in the ontology, this gave an average 8.6 keywords per class. The distribution was extremely uneven, however: some classes had only 1 or 2 keywords, while others had many more. In the final version of the ontology, there are 6,790 unique keywords. With 148 keyword-containing classes (the 2 top-level KET and SGC classes themselves do not have keywords), this gives an average of just under 46 keywords per class. The distribution follows a fairly standard bell curve, with the majority of classes having 20-100 keywords. However, the range is somewhat greater than ideal, with 10 classes having fewer than 10 keywords, and 26 classes having more than 100 keywords, both of which are potentially problematic.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

By looking at the distribution of classes to keywords, we see that 78% of keywords are only associated with one class, and more than 92% are associated with fewer than 3 classes. This means that our keywords are extremely distinctive of a topic. For comparison, in previous iterations of the ontology, the keyword "DNA" was assigned to 41 different classes (now assigned to only 7), while "gene" was assigned to 38 (now 5).

As we have mentioned already, there are a number of closely related classes, particularly in the KET area, so we should not expect all keywords to be unique. Recall also that keywords are weighted, with higher weights given to preferential terms, e.g. those which were manually produced and validated, those which score highly on similarity to the topic in the enrichment process, and those which co-occur in a document with strongly related terms (via the PMI weight). Moreover, the appearance of a single keyword in a text is not necessarily sufficient to match a document to that class, so this does not mean that every time "DNA" is found in a text it will automatically classify that document into all 7 classes.

When it comes to the final document annotation, the weights are critical in determining which topics should be allocated. In future versions of the ontology, we plan to fine-tune the weighting system for the keywords further, for example by ensuring that certain kinds of more general terms will only get scored when they occur in a document in conjunction with more specific terms related to the same topic. This is implicit in some of the weighting mechanisms already, but could be reinforced.

There are a number of important considerations concerning both the assignment of keywords to the ontology, and their role in the classification process. During various iterations of the ontology, a variety of methods was tested. Initially, the set of keywords was designed to be small but relatively precise, but this led to poor annotation results as some topics were not well captured. Extending the set of keywords led to better recall but at the expense of poor precision and many erroneous classifications (for example, very popular keywords like "cell" were matching documents to a large number of classes). The enrichment process helped somewhat with extending the recall further, but only when rigorously policed to ensure that rogue keywords were not accidentally generated. The initial corpus used for the enrichment process was also too small, and was therefore extended in a second iteration with a much larger dataset. This could be further extended as additional relevant data becomes available. However, this in itself brings a tradeoff – while larger corpora may provide better training material, they tend to contain more irrelevant documents which bias the results unfavourably. This was confirmed with some early experiments we performed using larger corpora of pre-trained embeddings on more general kinds of text, e.g. Glove (Pennington et al., 2014).

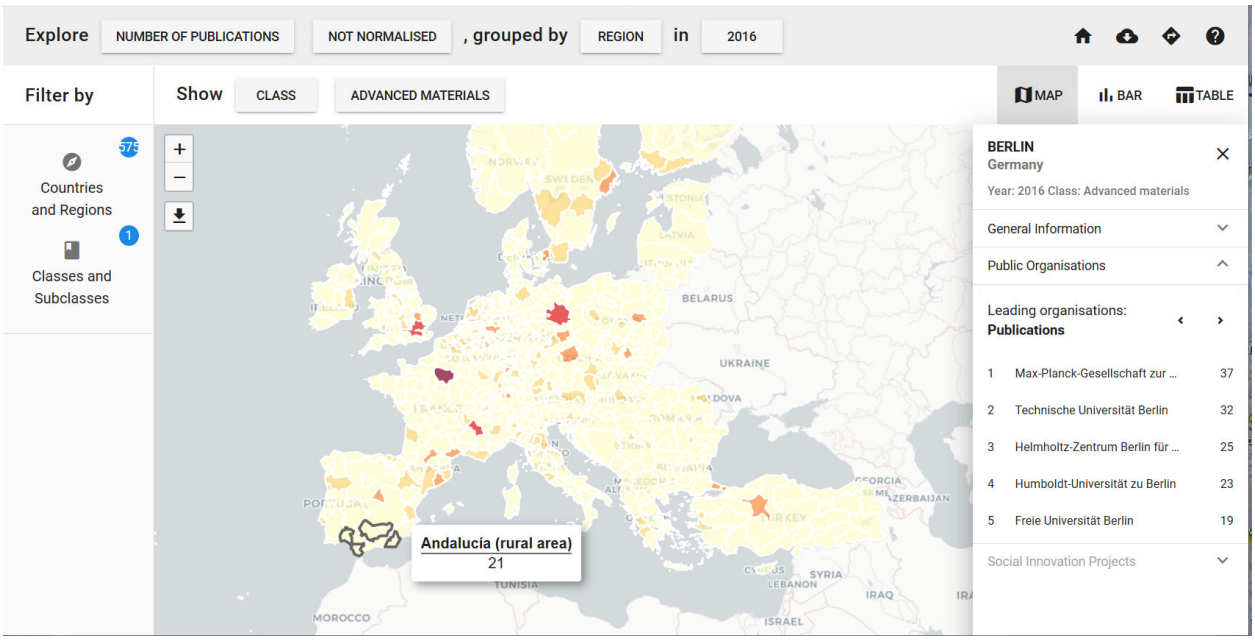
In general, the implementation of the ontology population process has demonstrated that the use of automatic techniques enables the generation of a large number of keywords, but becomes problematic when two subclasses share some similar terms (like rail and road transport). Currently, manual intervention is required in order to define a blacklist of topic-keyword combinations, which is a non-negligible amount of effort. The blacklist is reusable for future iterations of the enrichment process, but if the enrichment process produces a substantially new set of terms from the previous iteration, the manual verification process is required again. While we believe that expert intervention will always be required to some extent, this could be minimised further in future with additional statistical techniques to further weight terms based on maximising the semantic distance between terms from such closely related classes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

5.2 Task-based evaluation

The ontology should be evaluated against the specific tasks for which it has been designed. Specifically, the goal of KNOWMAK is to generate aggregated indicators to characterize geographical spaces (countries or regions) and actors (public research organizations and companies) in terms of various dimensions of knowledge production. For each topic or combination of topics, the mapping of documents enables the generation of indicators such as the number of publications, EU-FP projects and patents, as well as various composite indicators combining dimensions, such as the aggregated knowledge production share and intensity, publication degree centrality (see Figure 2).

Figure 2. The KNOWMAK tool interface and indicators



This specific task had several implications on the evaluation of the ontology.

First, it implied that a balance should be sought between recall and precision in the annotation process in order to get reasonable aggregated figures. This is obviously tricky to assess precisely without large-scale evaluation; the simple approach adopted was to test on samples of documents, and for selected classes to test that the proportion of false positives was not too large, while also ensuring that classes were sufficiently well populated. For example, this led to the rejection of document scoring criteria that were clearly too restrictive, such as imposing that documents were assigned to classes only when multiple keywords were matched. Since annotated texts are very short (as we do not have access to full-texts), this strategy strongly favoured classes with many keywords, generating huge imbalances in the indicators.

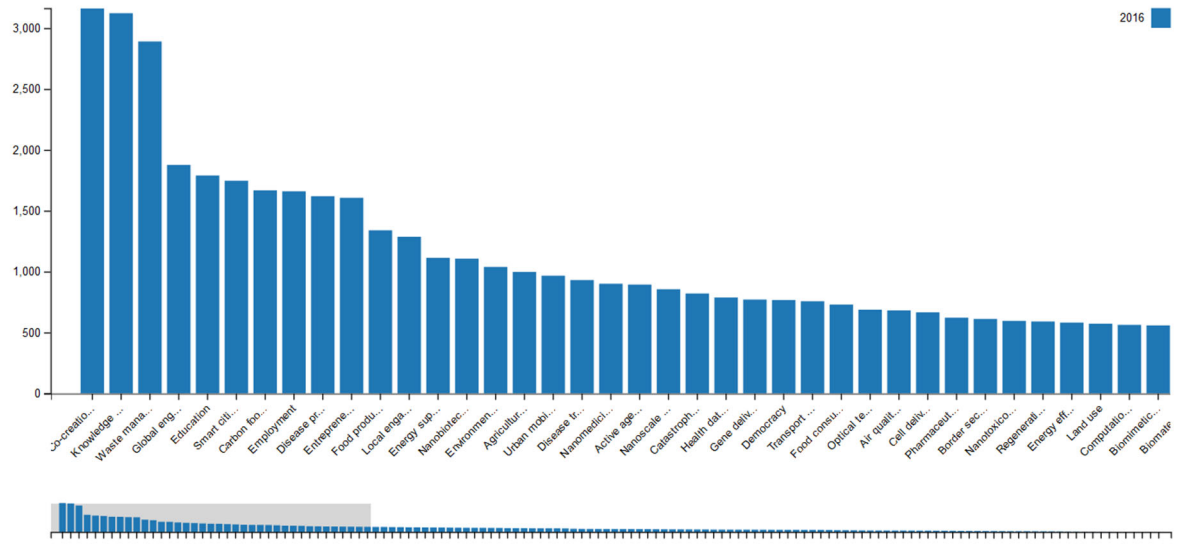
Second, the focus of the tool is on comparing the *relative indicators* across topics and geographical spaces. Examples of relevant questions are therefore to discover the regions with more publications or EU-FP projects on a specific topic, rather than to measure the absolute value. We expect that such comparisons are less sensitive to some characteristics of the annotation process, such as the exact scoring method, while they are more strongly impacted by the design of the ontology structure and the delineation of topics.

Accordingly, a major focus of the evaluation was checking the distribution of data items by ontology subclass in order to detect issues such as irrelevant classes and the presence of generic

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

keywords, which strongly inflate individual classes. As shown in Figure 3, the current distribution looks fairly reasonable: the few very populated classes are expected, such as knowledge transfer, which is a major focus of many European projects, while most subclasses are in the range of 100-1,000 projects. This analysis allows also the identification of subclasses with very few projects, which might necessitate either removal since they are not very relevant, or improvement in terms of delineation and keywords. While there is of course some arbitrariness in these judgements, this can be mitigated by discussion with external experts when presenting the results. For instance, experts quickly agreed that the adopted method for patent thresholding provided too low figures by class, and this led to a revision of the method.

Figure 3. Number of European projects by subtopic



Third, the tool allows also for a fine-grained disaggregation at the level of research organizations, since it is possible to single out for each region and topic the top-five organizations in terms of numbers of publications, patents and EU-FP projects (see Figure 2). In this respect, one can check for differences in the top knowledge producers by topic. For example, technical schools and research institutes are expected to be top in microelectronics; research hospitals in some medical topics; and generalist universities in many societal grand challenges. In previous versions of the ontology, this test did not provide satisfactory results, as in many cases the same organization had the largest output in all topics, as an outcome of the presence of very generic keywords. This situation clearly improved with the last version of the ontology. Moreover, it becomes possible to analyze the knowledge production profile for individual organizations, such as universities, by looking at the importance of dimensions (for example science vs. technology) and to the portfolio in terms of topics. At this very fine-grained level, experts and research managers of the relevant organizations are likely to own precise information to compare with the outcome of the tool.

The common feature of these task-based evaluations is therefore that they do not check whether all documents have been classified correctly, but rather that aggregated figures are deemed reasonable by experts in the field. On the one hand, such an approach is more parsimonious than a systematic evaluation of document assignments and allows for successive revisions of the ontology to be implemented in a reasonable time. In other words, rather than seeking to develop a 'perfect' annotation method at once – an impossible task given the lack of a gold

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

standard - we improved the ontology stepwise by designing more complex and fine-grained tasks at each step, a process that can be further extended in the future as the usage of the tool develops. On the other hand, this approach is consistent with an epistemological conception of indicators as (partially arbitrary) figures, which nurture the policy debate and include some level of arbitrariness (Barré, 2001). We notice that such a historical contingency is common to all existing S&T classifications, but it is usually black-boxed within a general claim of objectivity (Godin, 2001). Admittedly, there is scope for designing more systematically this process of debate and refinement, by identifying key tasks to be performed, formalizing the expert feedback process and the implications for the ontology.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

6 Discussion and conclusions

In this work, we aim to address some of the limitations in applying traditional classifications to a science policy domain for the purposes of mapping scientific research. We do this through the use of ontologies, in an effort to extend the reach of existing text-based classification methods while still maintaining the power and rigour of classification systems. In particular, we have attempted to overcome the problems in connecting policy-based topics with science-based topics, which require dealing with not only differences in the language and terminology used, but also in the topic structure itself.

In striving to find the balance between data-driven and user-driven approaches to the design and application of ontologies, we have uncovered insights into which processes have to be mostly driven by users, and which can be managed through automated approaches, as well as the best ways to involve users in the assessment and feedback. The methodology and tools in our approach have been designed in such a way as 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 and topics change over time, the methodology enables greater flexibility than many existing classification-based systems allow. Changes to the ontology or the input of new research data can be handled in an automatic way, and updates pushed to the central databases from which indicators are generated. On the other hand, these are tempered by expert intervention at critical stages in order to maximize accuracy and ensure suitability. We strongly assert that, in contrast to the growing trend for data-driven classification techniques, the ontology structure itself should be designed primarily in a top-down expert-based manner in order to meet the principal requirements of flexibility, commensurability and temporal stability.

This is not to say that the work does not have limitations. In particular, rigorous evaluation is difficult and requires manual intervention, which is time-consuming and subjective. The use of NLP techniques also brings its own issues, since language is complex to understand and process, which is why a certain amount of expert intervention is required at every step. 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 pathways for STI technologies, which open up avenues for a number of future directions of research.

We envisage a number of ways in which this work could be advanced. Beyond the methodological improvements already listed, our ontology has been designed for a specific use case: the mapping of the European research domain in the critical areas of KETs and SGCs, in order to assist policymakers with decision making and strategic planning by helping them to understand the nature of the field. The methods and tools presented could equally be applied to other research areas, new kinds of documents, new languages, and new geographical boundaries, with little adaptation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

7 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.

Barré, R., 2001. Sense and nonsense of S&T productivity indicators. The contribution of European Socio-Economic Resea

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. *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.

Estañol, M., Masucci, F., Mosca, A. and Ràfols, I., 2017. Mapping knowledge with ontologies: the case of obesity. *arXiv preprint arXiv:1712.03081*.

Francopoulo, G., Mariani, J., Paroubek, P., Vernier, F.: Providing and Analyzing NLP Terms for our Community. *Computerm 2016* p. 94 (2016)

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).

Godin, B., 2001. Tradition and Innovation: The Historical Contingency of R&D Statistical Classifications. Project on the History and Sociology of S&T Statistics Paper No. 11.

Gok, A., Waterworth, A., Shapira, P.: Use of web mining in studying innovation. Scientometrics 102(1), 653–671 (2015)

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

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

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)

Lepori, B., Barré, R., Filliatreau, G., 2008. New perspectives and challenges for the design and production of S&T indicators. Res Eval 17, 33-44.

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.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

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.

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

Maynard, D., Bontcheva, K., Augenstein, I. Natural Language Processing for the Semantic Web. Morgan and Claypool, December 2016. ISBN: 9781627059091

Maynard, D. and Greenwood, M.A. Large Scale Semantic Annotation, Indexing and Search at The National Archives. In Proceedings of LREC 2012, May 2012, Istanbul, Turkey.

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

Maynard, D., Li, Y. and Peters, W. NLP Techniques for Term Extraction and Ontology Population. Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text, P. Buitelaar and P. Cimiano (editors). IOS Press, 2007.

Maynard, D., Roberts, I., Greenwood, M.A., Rout, D., Bontcheva, K. A Framework for Real-time Semantic Social Media Analysis. Web Semantics: Science, Services and Agents on the World Wide Web, 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.

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

Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*(pp. 1532-1543).

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.

Shiffrin, R.M., Börner, K., 2004. Mapping knowledge domains. PNAS 101, 5183-5185.

Spasic, 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)

Suárez-Figueroa, Mari Carmen, et al., eds. Ontology engineering in a networked world. Springer Science & Business Media, 2012.

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.

Tablan, V., Bontcheva, K., Roberts, I., Cunningham, H.: Mimir: 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

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 enabling technologies. IDEA Consult (2012).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

Velden, T., Boyack, K.W., Gläser, J., Koopman, R., Scharnhorst, A. and Wang, S., 2017. Comparison of topic extraction approaches and their results. *Scientometrics*, *111*(2), pp.1169-1221.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.