
Article:
Maynard, D., Lepori, B., Petrak, J. et al. (2020) Using ontologies to map between research data and policymakers' presumptions: the experience of the KNOWMAK project. Scientometrics, 125 (2). pp. 1275-1290. ISSN: 0138-9130

https://doi.org/10.1007/s11192-020-03664-6

Scientometrics
https://doi.org/10.1007/s11192-020-03664-6

Using ontologies to map between research data and policymakers' presumptions: the experience of the KNOWMAK project

Diana Maynard[1] · Benedetto Lepori[2,3] · Johann Petrak[1] · Xingyi Song[1] · Philippe Laredo[3]

Received: 13 December 2019

Abstract
Understanding knowledge co-creation in key emerging areas of European research is critical for policy makers wishing to analyze impact and make strategic decisions. However, purely data-driven methods for characterising policy topics have limitations relating to the broad nature of such topics and the differences in language and topic structure between the political language and scientific and technological outputs. In this paper, we discuss the use of ontologies and semantic technologies as a means to bridge the linguistic and conceptual gap between policy questions and data sources for characterising European knowledge production. Our experience suggests that the integration between advanced techniques for language processing and expert assessment at critical junctures in the process is key for the success of this endeavour.

Keywords Ontology · Natural language processing · Knowledge co-creation · Policymaking · Termextraction

Introduction

In recent years, a priori classification systems for science and technology, such as the Field of Science Classification (OECD 2015) and IPC codes for patents (Debackere and Luwel 2004), have been increasingly replaced by data-driven approaches, relying on the

✉ Diana Maynard
d.maynard@sheffield.ac.uk

Benedetto Lepori
benedetto.lepori@usi.ch

Philippe Laredo
philippe.laredo@enpc.fr

[1] Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield, UK

[2] Faculty of Communication Sciences, Università della Svisera italiana, 6904 Lugano, Switzerland

[3] Laboratoire Interdisciplinaire Sciences, Innovations et Sociétés (LISIS), University of Paris Est, 77454 Marne-la-Vallée Cedex 02, France

Published online: 17 August 2020

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 visualise 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 display limitations. Such methods provide very detailed views of specific knowledge domains, but are less suited to large-scale mapping across the whole S&T landscape. Furthermore, lacking a common ontology of S&T domains (Daraio et al. 2016), such mappings are largely incommensurable across dimensions of knowledge production. 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 2001).

In this paper, we describe how these challenges have been addressed to develop a web-based tool[1] providing interactive visualisations on European research, focusing on two key categories in the European research policy debate: Key Enabling Technologies (KET) and Societal Grand Challenges (SGC). These exemplify the delineation issues mentioned above, since they are political instruments established at a high level of decision and policymaking, and are deliberately designed to have a broad coverage, often also reflecting the interests of a diverse set of stakeholders. On the other hand, producing indicators about their elements requires some kind of common structure. We thus take a broad and encompassing vision of their semantic content, while this can nevertheless be narrowed by altering classification thresholds, as explained in "Ontology design and implementation" and "Results and evaluation" sections.

Our approach is based on two main elements: (a) the design of an ontology of the KET and SGC knowledge domains to make explicit their content and to provide a common structure across dimensions of knowledge production through a two-level structure where KET and SGCs are decomposed into a set of subclasses; and (b) the integration between natural language processing (NLP) techniques (to associate data sources with the ontology categories) and expert-based judgement (to make sensible choices for the matching process). This drove a recursive process where the ontology development and data annotation were successively refined based on expert assessment of the generated indicators.

Our experience with this specialised ontology and classification shows that while NLP techniques are critical for linking (policy-related) ontologies with large datasets, some key design choices about the ontology and its application to data are 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 the use of advanced data techniques is a key requirement for robust S&T ontologies. Our paper contributes to this endeavour 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 such an ontology. While the ontology is indeed tailored to highly specific policy topics, it covers a wide range of disparate subjects and data, and the approach is nevertheless still flexible and scalable.

[1] http://knowmak.eu.

Background

A large body of work has been developed to address the limitations of existing classification systems. These 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 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. While they provide detailed views of specific knowledge domains and of their evolution over time, they are less suited to large-scale mapping of broad policy themes. Connecting such topics with relevant themes at the policy level is far from simple, since

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

the associated terminologies are largely incompatible (Cassi et al. 2017). 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 a hierarchical representation of topics, 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 organised in layers, ontologies are more flexible in structure. Our KNOWMAK ontology operates as a bridge between (policy) questions and heterogeneous data sources (Maynard and Lepori 2017). For the audience, 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 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 (Estañol et al. 2017). Such techniques are not scalable and are not suitable for mapping large policy themes, such as those associated with KET and SGCs, to a broad set of knowledge outputs.

Ontology design and implementation

Our ontology development involves three 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; and 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 an iterative process, 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 expert assessment of the 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 stabilised (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 flexible, for example to respond to changes in policy interests (see "Discussion and conclusions" section), and to a certain extent scalable since

new data sources can be integrated within the process. All three steps require some human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced NLP techniques for the parts involving handling large data. The classification process is fully automated once the ontology is complete, so re-annotation can easily take place if the ontology is changed or new data is available.

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 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 practical, and because these affect the implementation of the document-topic mapping. Moreover, as already discussed, 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. Ontologies must be dynamic: new terms and definitions continuously emerge from researchers and standardisation groups, while other terms may become irrelevant or replaced by more popular synonyms. This means that 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. We take as a starting point some existing classifications, 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 provide 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.

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 distinctive enough to be interesting for policymakers without making the choices too specific. The latter has an impact on quality, because it is hard to allocate documents to topics at very precise levels, but also on usability of the system.

A key expert decision relates also to the conceptual overlap between classes. 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 documents can be classified according to multiple topics, but the topics themselves are as distinct as possible.

Ontology population

The ontology needs to be populated with instances (keywords) from various data sources, which help to: (1) match user queries to topics; 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 keywords used in the data sources, and even between the different types of data source, terms vary widely.

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 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, "intelligent transport" is a key term for the topic "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 GATE's Automatic Term Recognition tool TermRaider (Maynard et al. 2007; Zhang et al. 2018). 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. For example, "radar tracker" is a non-preferred term for the topic "intelligent transport". This term might be relevant here only if found in conjunction with another relevant term for the topic.

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. The enrichment process can be broken down into three main steps: corpus pre-processing, embeddings training, and embeddings-based term scoring[4]. First, we apply linguistic pre-processing to our training corpus to find: (1) all occurrences of original ontology keywords in corpus (both of which are lemmatised); and (2) single and multi-word term candidates in the corpus, filtering out any Named Entities (e.g. names of people, places etc.). Next, we merge the ontology matches and the term candidate, and create (potentially overlapping) keyword candidates. We then calculate the canonical lemmatised string for these candidates, and finally calculate term statistics for all term candidates (using tf, df, idf). This results in a set of 1.2 million keyword candidates in 180 million locations in the corpus.

Next, we train the embeddings (vector representations for single and multi-word terms) from our keyword candidates and corpus. These embeddings were 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.[5] Finally, we score the terms based on the embeddings, according to their "representativeness" of that class, and prior probabilities generated using Pointwise Mutual Information (PMI) for term combinations, based on frequency of

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

[4] For details readers should refer to technical documentation at https://gate.ac.uk/projects/knowmak/.

[5] The embeddings are available at http://downloads.gate.ac.uk/knowmak/embeddings201812.txt.gz.

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

Topic	Key	Preferred	Project	Generated	Enriched	Total
KET						
AdvancedManufacturing 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

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. We use a novel method we call *centrboth*. For each class, we calculate the average embedding for the set of preferred terms, and another average embedding for the set of non-preferred terms related to the class. The final embedding is the weighted average of both. We then use a method we term *simonly*. This is the 0/1 normalised cosine similarity between the embeddings representing the ontology class (*centrboth*) calculated in the previous step, and the embedding representing the candidate term. In both cases for *simonly*, we take the unweighted average, since using the weighted (tf, idf) average did not work well in early experiments.

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 2122 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 9076 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.

Document classification

Our data sources comprise three major datasets on S&T made available within the RISIS H2020 infrastructure project[6]: the Web of Science version at CWTS, University of Leiden (about 30 m publications), the PATSTAT version at IFRIS in Paris (2.37 m patents), and the EUPRO database of European Framework Programme projects (67,475 projects), all from the period 2000–2017. The annotation links each data element (e.g. a project) with the relevant topic(s) in the ontology, so that indicators can be built around them. The amount of data that can be annotated is restricted only by time and processing power, and annotation time can be reduced by adding extra threads to the processing.

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.

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.

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 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

[6] http://risis2.eu.

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 Ontology Design Principles methodology for ensuring the quality and validity of an ontology (Suárez-Figueroa et al. 2012). According to these principles, the quality and effectiveness of an ontology should be considered primarily in the context of its intended use. Just as the notion of indicators has moved away from the traditional statistical 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 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 "Ontology design and implementation" section); assessing the relevance and coverage of the keywords attached to the classes ("Keyword evaluation" section); and a task-based assessment of the ontology ("Task-based evaluation" section), checking for minimal overlap between class assignment and ensuring that all classes have sufficient—but not too many—documents assigned.

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 6790 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.

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).

Since there are some closely related classes, 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 matching of keywords to classes is context-dependent, e.g. every time “DNA” is found in a text it will not

necessarily be matched with 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, 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. Automatically extending the set of keywords led to better recall but at the expense of poor precision and many errors (documents containing very popular keywords like "cell" were matching many classes). The enrichment helped somewhat with extending the recall, but only when rigorously policed to ensure that rogue keywords were not accidentally generated, and this is not scalable. We therefore considerably extended the corpus used for enrichment, and this could be further increased with newly emerging relevant data. However, this brings a tradeoff, as a larger corpus also contains more irrelevant documents, which bias the results unfavorably. This was confirmed with experiments using larger corpora of pre-trained embeddings such as Glove (Pennington et al. 2014).

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 (e.g. rail and road transport). Currently, some manual intervention is required in order to define a blacklist of topic-keyword combinations. However, we plan to automate this process. While expert intervention will always be required to some extent, this could be further minimised with additional term weighting techniques based on maximising the semantic distance between terms from closely related classes. This increased automation also makes adapting to a wider set of topics more feasible.

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 characterise geographical spaces (countries or regions) and actors (public research organisations 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 Fig. 1).

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 keywords, which strongly inflate individual classes. As shown in Fig. 2, the

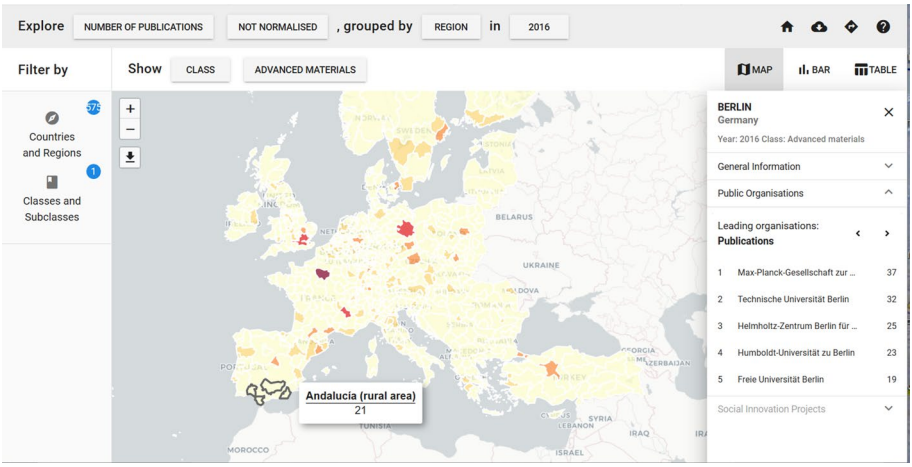

Fig. 1 The KNOWMAK tool interface and indicators

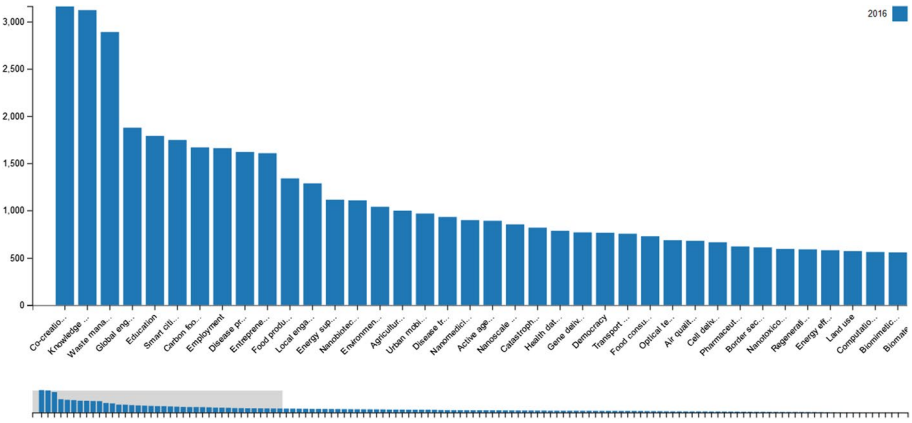

Fig. 2 Number of European projects by subtopic

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 100–1000 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.

Third, the tool allows also for a fine-grained disaggregation at the level of research organisations, since it is possible to single out for each region and topic the top five organisations in terms of numbers of publications, patents and EU-FP projects (see Fig. 1). 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 organisation 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 analyse the knowledge production profile for individual organisations, 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.

Finally, we have noted that since KETs and SGCs are not mutually exclusive, our approach could potentially prevent multiple classification of a document in both branches of the ontology. We found that around 17% of patents and 15% of projects were classified in both a KET and SGC topic, while for publications this figure was lower at 4%. This indicates that at least to some extent, the approach deals with this topical overlap. The risk of independent classification of the two branches is that less relevant topics in each branch might get artificially promoted resulting in spurious classifications. However, in future work we aim to investigate also this approach (see "Discussion and conclusions" section).

The common feature of these task-based evaluations is that they check that aggregated figures are deemed reasonable by experts in the field. Such an approach is more parsimonious than a systematic evaluation of document assignments, and allows successive revisions of the ontology to be implemented. Thus rather than seeking to develop a 'perfect' annotation method at once—an impossible task given the lack of a gold standard—we improved the ontology stepwise by designing more complex and fine-grained tasks at each step. 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). 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, formalising the expert feedback process and implications for the ontology.

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 around KETs and SGCs. This is different from the general problem of science mapping where data-driven classification approaches can be used, because in our case a fixed classification system is required in order to make comparisons over time and across different kinds of data. 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. An ontology for mapping policy classifications such as ours is also very different from maps of web science and from the small ontologies that describe a narrow topic area, since we cover a broad and disparate, yet defined set of topics. 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. Furthermore, this work is not small-scale: almost a million documents are classified (excluding the irrelevant documents processed, which are not assigned to a topic from the ontology); the ontology is wide-ranging and contains 50 classes with an average of 46 keywords per class, totalling around 10,000 keywords.

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 maximise 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 maximise 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.

The work naturally has some 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, and numerous issues in terminology extraction still need to be solved globally. 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 be applied to new kinds of documents and new geographical boundaries, with little adaptation, since these processes are all automated. Introducing new languages would require some manual intervention, but the automated parts of the process will require little beyond re-training on a suitable corpus. We have already adapted the ontology structure to take into consideration recent changes to KET topics and to migrate from SGCs to SDGs. Much of this process was automated, for example by automatically extracting relevant topics and keywords from descriptions of the new topic, although expert intervention was required to check the final structure and make some small adjustments. The process of determining seed keywords is the most critical part: the experience described in this paper shows that if this part is done well, the next step (regeneration of the enriched keywords, which cannot easily be checked manually due to their large number) will be high quality. We envisage making further improvements to the methodology in terms of the enrichment process in particular, however, with more complex topic-aware deep learning methods. Finally, the classification process itself requires no adaptation for the new ontology, though it will of course be prudent to verify the results with experts.

Acknowledgements This work was partially supported by the European Union under Grant Agreement No.726992 KNOWMAK and Grant Agreement No. 825091 RISIS.

References

Amjadian, E., Inkpen, D., Paribakht, T. S., & Faez, F. (2016). Local-global vectors to improve unigram terminology extraction. In *5th international workshop on computational terminology (Computerm 2016)* (pp. 2–11). Osaka, Japan.

Barré, R. (2001). Sense and nonsense of S&T productivity indicators. *Science and Public Policy, 28*(4), 259–266.

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Boyack, K. (2017). Investigating the effect of global data on topic detection. *Scientometrics, 111*(2), 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). Dordrecht: Springer.

Estañol, M., Masucci, F., Mosca, A., & Ràfols, I. (2017). *Mapping knowledge with ontologies: The case of obesity*. arXiv:1712.03081.

Francopoulo, G., Mariani, J., Paroubek, P., & Vernier, F. (2016). Providing and analyzing NLP terms for our community. *Computerm, 2016,* 94.

Frietsch, R., Neuhausler, P., Rothengatter, O., & Jonkers, K. (2016). Societal grand challenges from a technological perspective: Methods and identification of classes of the international patent classification IPC. Technical 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. (2015). Use of web mining in studying innovation. *Scientometrics, 102*(1), 653–671.

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

Hammond, T., & Pasin, M. (2015). The nature.com ontologies portal. In *5th workshop on linked science*, 2015.

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

Lepori, B., Barré, R., & Filliatreau, G. (2008). New perspectives and challenges for the design and production of S&T indicators. *Research Evaluation, 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.

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

Maynard, D., Bontcheva, K., & Augenstein, I. (2016). *Natural language processing for the semantic web*. San Rafael: Morgan and Claypool.

Maynard, D., & Greenwood, M. A. (2012). Large scale semantic annotation, indexing and search at the national archives. In *Proceedings of LREC 2012*, May 2012, Istanbul, Turkey.

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

Maynard, D., Li, Y., & Peters, W. N. L. P. (2007). Techniques for term extraction and ontology population. In P. Buitelaar & P. Cimiano (Eds.), *Bridging the gap between text and knowledge: Selected contributions to ontology learning and population from text*. Amsterdam: IOS press.

Maynard, D., Roberts,I., Greenwood, M.A., Rout, D., Bontcheva, K. A. (2017). Framework for real-time semanticsocial media analysis. Web semantics: Science, services and agents on the WorldWide Web, 2017.

Motta, E., & Osborne, F. (2012). Making sense of research with Rexplore. In *Proceedings of the* 2012*th international conference on posters & demonstrations track* (Vol. 914, pp. 49–52). http://ceur-ws.org/.

OECD. (2015). *Frascati manual 2015. Guidelines for collecting and reporting data on research and experimental development*. Paris: OECD.

Pennington, J., Socher, R., & 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. (2010). 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.

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. (2008). Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics, 9*(5), S5.

Suárez-Figueroa, M. C., et al. (Eds.). (2012). *Ontology engineering in a networked world*. Berlin: Springer.

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. (2015). Mimir: An open-source semantic search framework for interactive information seeking and discovery. *Journal of Web Semantics, 30,* 52–68.

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. (2012). *Feasibility study for an EU monitoring mechanism on key enabling technologies*. Bruxelles: IDEA Consult.

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

Zhang, Z., Petrak, J., Maynard, D. (2018). Adapted TextRank for term extraction. In *Proceedings of semantics 2018*, Vienna, Austria, 10–13 September, 2018.