KNOWMAK

KNOWledge in the MAKing in European Society

Report on Ontologies and Tagging

December 2019

Diana Maynard, Johann Petrak, Xingyi Song and Adam Funk (USFD)

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

DELIVERABLE INFORMATION

Name of the deliverable	Report on Ontologies and Tagging
Number of the deliverable	D2.4
Related WP number and name	WP2 Ontologies
Related task number and name	T2.2 Methodology for ontology creation and semi-automatic tagging
Deliverable dissemination level	PU
Deliverable due date	31 December 2019
Main author(s)	Diana Maynard
Contributing partners	USFD

REVISIONS

Version	Date	Comments (Y/N)	Project partner
1			

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 Introduction

This document describes the final version of the ontology and the improvements since the first report, as well as some discussion on lessons learned and future work.

In the first version of the ontology, the design of the ontology structure was described, and preliminary population of the ontology with keywords related to the topics in the ontology was carried out. This enabled preliminary tagging of a set of publications, patents and projects with topics, by means of the keywords and a weighting mechanism. The extension to this deliverable described some improvements to the initial ontology population and tagging methodologies.

In this document, we summarise the role of ontologies and our experience of ontology development in the project, and describe substantial progress in: (1) refining the ontology structure (classes and subclasses, corresponding to topics); (2) improving the mechanisms for keyword generation; and (3) scoring the documents during the classification process. We also outline our plans for the remaining work to be done in the project. Improvements to the ontology have been made based on a continuous cycle of update and testing on the documents in our database. In practical terms, the resulting classification of documents is much improved. In technical and methodological terms, we have developed new NLP techniques for generation and scoring of keywords, and have demonstrated that the combination of NLP, deep learning and ontologies can enhance standard classification-based approaches typical of the STI field.

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

2 The role of ontologies

Our experience in the project has shown that while natural language processing (NLP) techniques are critical for linking ontologies with large datasets and extracting from the latter robust evidence, nevertheless some key design choices on the ontology and its application to data are basically of an intellectual nature. This suggests that the design of robust interactions between expert-based priori knowledge and evaluation on the one hand, and the use of advanced data techniques on the other hand, is a key requirement for robust S&T ontologies.

2.1 Approach

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 (ontology creation); second, assigning keywords to the topics (ontology population); and third, classifying documents based on the frequency of keywords (data annotation). All three steps require human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced language analysis techniques. Consequently, if any changes are deemed necessary, the process can easily be re-run and the data re-annotated within a short period of time and in a principled way.

In terms of development process, we utilize an interactive and staged approach that exploits the interaction with data sources to improve the system and relies on expert assessment at specific bridging points. In this way, we seamlessly integrate automated methods, based largely on NLP and Semantic Web technologies, with prior expert information. 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 by users.

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.

In more detail, the stages of the work consist of:

1. building a core ontology structure (with classes)
2. populating the ontology (with keywords)

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

   a. adding an initial set of keywords to classes, based on information in the ontology
   b. refining the ontology population (extending the initial set of keywords using other sources of information)
   c. weighting the keywords
3. creating the classifier (a GATE application)
4. running the classifier on the documents (annotation via a GATE web service)

In the next two sections, we describe in more detail the process undertaken to improve the ontology from the initial versions. A full description of the process, along with the ontology itself, is also provided in the publicly available technical documentation.[1] Then in Section 5, we describe the classification process.

[1] https://gate.ac.uk/projects/knowmak/

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

3 Ontology structure

The ontology is defined according to the two strands of KET and SGC. We take as a starting point some existing classifications, which we merge and map, such as the mappings between IPC (International Patent Classification) codes and both KETs (Van der Velde, 2012) and SGCs (Frietsch et al., 2016). For KETs, we also make use of the structure implemented in the nature.com ontologies portal (Hammond and Pasin, 2015). Some of these topics are already connected to DBpedia and MESH, which provides us with an additional source of information for keywords. Linking with the nature.com ontology helps with mapping the publications, and enables future extension of our ontology to other topics.

However, initial experimentation (as reported previously) 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 make a more balanced classification system. A collection was 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. A key expert decision relates also to the extent of overlap between classes and subclasses, as some are intrinsically related. For example, the "Advanced Manufacturing" KET is problematic because it 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), and so on. It is very hard to define these related classes in such a way that they are distinct, and we therefore expect some overlap both in keywords and in data annotation, though we try to minimise this as much as possible.

Annotation with the second version of the ontology produced much higher quality results already. However, it became clear that some ontology classes were still problematic. This was largely due to the starting point of existing mapping schemes and the Nature.com ontology, which in some cases were flawed, and in other cases just did not fit our structure properly. We identified several particular issues.

First, some subclasses did not really fit the higher-level class description (such as security, which is specific to public security and safety, but had subclasses that were less relevant). Second, there was some confusion between related high-level classes such as "Energy" and "Climate", and some reorganisation was necessary. Third, the introduction of social innovation documents required some additional subcategories of SGCs. These issues have been addressed, and some reorganisation of the classes and simplification has ensued. A co-creative approach has been used, drawing on the expertise of various partners in the consortium, especially for the social innovation part.

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

Following an iterative refinement and validation process, through a combination of consulting experts, testing the results of the annotation, and testing the ontology design with real users in the KNOWMAK tool, it was decided to further simplify the ontology down to 2 basic levels: the top level of the 13 KETs and SGCs, and one level below this. This not only makes more sense conceptually to the users, and makes the KNOWMAK tool much easier to navigate and use, but also has led to great improvement in the annotation task. The reason for the latter is that previously, the semantic distance between different ontology classes was widely varied, which led to some problematic results when generating the extended keywords by means of word embeddings (see Section 4). This was because the methodology works by measuring the semantic proximity of each candidate keyword to the rest of the ontology and finding the best match to a class. If some classes are semantically very close, this reduces the accuracy of the method.

The final version of the ontology contains 150 topics based around the 6 KETs and 7 SGCs, and a total of 9076 keywords, of which 6790 are unique (because some are attached to more than one topic). This is depicted, along with the frequency of keywords per topic, in Table 1.

Finally, we experimented with an automated method of producing for each topic a description, which is used in the KNOWMAK tool for explaining to the user what the topic signifies. These descriptions are also used as a source for extracting keywords automatically. The automated method is only used in cases where the topic does not already have a description from other sources such as policy documents or the Nature.com ontology. The method aims to find the most relevant Wikipedia page for each ontology class, by training an LDA model (n topics) based on relevant training documents, and then extracting the "About" information for that page as the description, as shown in Figure 1 below.

Alternative fuels, known as non-conventional and advanced fuels, are any materials or substances that can be used as fuels, other than conventional fuels like; fossil fuels (petroleum (oil), coal, and natural gas), as well as nuclear materials such as uranium

Figure 1: Automatically generated description of the topic "Alternative Fuels" from Wikipedia

To train the model, we use a collection of abstracts from projects, patents, and publications, as well as policy documents (the same collection as used for the embeddings training). The training documents are pre-processed by removing stop words and stemming. To find the most relevant Wikipedia page, we first use the class label as a search query, and score the Wikipedia page with gestalt pattern matching between class label and the Wikipedia page. If the matching score is less than 0.7, then we reset the search queries as keywords, and use the most relevant page.

We evaluated this approach manually and replaced the ones which were erroneous. Out of the 147 topics in the ontology (excluding KET and SGC themselves), 99 had no topics. For these 99 topics, the method found 78 new descriptions. In total, 65.38% of these were correct, and an additional 17.95% partially correct (these were judged to be either right in principle, but wrong in this particular context, or not quite exact enough. 16.67% of the

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

pages found were incorrect. The incorrect and partially correct descriptions were manually corrected (in many cases, finding a synonym of the original topic as a starting point was sufficient; in some cases, however, no relevant Wikipedia page existed).

The use of Wikipedia in this way forms part of ongoing experimentation extending the use of these pages to help with the keyword population (see Section 7.1).

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

4 Ontology population

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

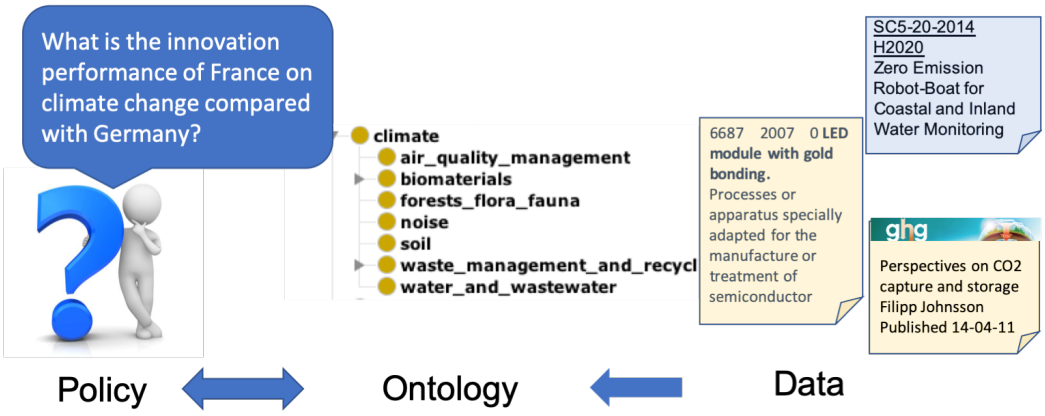


Figure 2: The role of an ontology in connecting policy-related questions from users with data sources

Our experience with annotating the documents in our collection with version 2 of the ontology showed substantial improvement over the first version, but still some issues remained. We have pursued two strands of work: first, on improving the initial keyword generation, and second, on specifically improving the methodology for creating the enriched keywords. Following a number of experiments and iterations, the final solution adopted involves multiple layers of keyword extraction and a mixture of automated techniques interspersed with expert knowledge at key junctures. A number of changes were made to the methodology. First, a stop list was manually created in order to prevent too generic keywords (e.g. “method”) being selected. Several iterations of this selection process were performed, and checks were made to prevent any keyword being assigned to too many topics. Furthermore, at every stage, multi-word terms are preferred, as these tend to be more specific and therefore are better at distinguishing between similar topics. Third, an automatic keyword enrichment method is used to boost the number of keywords.

4.1 Improving the initial keyword generation

Our analysis showed problems with certain topics that had too few, too many, or badly chosen keywords. We have worked on mitigating this in the following ways.

First, analysis of the distribution of documents to topics showed that some classes had very low numbers of documents assigned to them. Investigation showed that this was primarily due to either just low numbers of keywords, or to poor choice of keywords. This was particularly the case with topics such as “social inequality”, where the keywords were mostly too general to be useful. To resolve this, we manually reviewed the seed keywords for these classes, and added some new ones which were more targeted. This meant that in the enrichment process, a better set of additional keywords was then produced.

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

Second, the distribution analysis also showed that some classes had very high numbers of documents assigned to them, e.g. “public engagement”. Investigation showed that some very general keywords were produced by the enrichment process. To resolve this, we reviewed the results and added a blacklist of term-topic combinations that should not be generated by the enrichment (or indeed, at any other keyword generation stage). For example, “shipyard” is not a good keyword for the topic “aeronautics” but it is a good keyword for the topic “maritime transport”, so the combination “shipyard – aeronautics” appears in the blacklist.

In the final version, a small set of very specific but high-quality keywords is selected manually for each topic (typically around 5 per topic). These are known as *key* terms, and 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, known as *project* terms, comes from the subject index of the EUPRO project database, which we have mapped to our ontology.[2]

4.2 Keyword enrichment

We have also improved the generation of the enriched keywords. The basic workflow for this is as follows:

1. generate a set of seed keywords associated with each ontology class;
2. extend these keywords by finding semantically similar terms in a large corpus, using word embeddings trained on that corpus (extract a set of terms, then find the ones most similar to seeds);
3. score the keywords according to how representative they are of that class;
4. generate prior probabilities using PMI for term combinations, based on frequency of co-occurrence in the training data (this is used later in the classification tool – see Section 5).

Due to the low quality of some of the keywords generated in the first stage, we revisited this process. We still automatically generate keywords from class names, descriptions, and related information (e.g. DBpedia, skos, etc.) using term recognition tools, but as mentioned above, these are now separated into:

- *preferred* (e.g. coming directly from class names or other “good” sources);
- *generated* (e.g. coming from descriptions – might not be so high quality).

Only preferred terms are used for the enrichment process, and they also get a higher weighting at document classification time (see Section 5). Figure 3 shows an example of a class description for the topic “smart cities and communities”, where relevant terms have been extracted automatically (highlighted in yellow) by NLP tools. These would be classified as “generated”, while terms derived from the class name itself (e.g. “smart cities”, “smart cities and communities”), would be classified as “preferred”.

[2] This mapping is publicly available at 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.

Sustainable development of urban areas is a challenge of key importance. It requires new, efficient, and **user-friendly technologies** and services, in particular in the areas of **energy**, **transport** and **ICT**. However, these solutions need integrated approaches, both in terms of research and development of advanced technological solutions, as well as deployment. The focus on **smart cities technologies** will result in commercial-scale solutions with a high market potential.

Figure 3: Keywords extracted automatically from a class description

The process is thus as follows:

1. create a set of additional keywords for each class in the ontology using an automatic unsupervised approach;
2. create a large corpus of patent, publication and project abstracts as well as relevant policy documents;
3. extract new candidate terms from this corpus;
4. train domain-specific word embeddings for these terms, in such a way that we can have vectors also for multi-word terms;
5. use the embeddings to find the similarity between the seed terms and the new terms;
6. use the similarity to decide which new terms to keep, and which concept to map them to.

The enrichment process itself can be broken down into three main steps: corpus pre-processing, embeddings training, and embeddings-based term scoring.

4.2.1 Corpus pre-processing

Our corpus consists of 2.6 million documents in total, comprising project, patent and publication abstracts, and a set of policy documents. Pre-processing consists of the following steps:

1. Run a GATE application for linguistic pre-processing, which consists of POS tagging, lemmatisation, entity finding, etc. This is used 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.).
2. Merge the ontology matches and the term candidate, and create (potentially overlapping) keyword candidates.
3. Calculate the canonical lemmatized string for these candidates.
4. 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.

4.2.2 Training the embeddings

This step generates embeddings (vector representations from our keyword candidates and corpus). The following steps are undertaken:

1. calculate a set of 330 stopwords (also used for scoring later on) and a set of unique multi-word terms from the original ontology keywords;

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

2. train the embeddings using Python GenSim, removing stopwords and single-letter words from the corpus
3. Use sentences as training examples, generated in the following way:
    - match each sentence against the list of multi-word terms;
    - for each multi-word term, create one sentence where each lexical unit is a separate multi-word term;
    - create one sentence also where all multi-word terms are single lexical units.

This process results in the generation of 591,526 embeddings, which we make publicly available.[3]

4.2.3 Scoring the terms based on embeddings

We have investigated various ways of calculating embeddings to represent ontology topics and measuring similarity between the keyword and class. Best results have been achieved so far with a method we term *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. We did experiment with some more complex methods, such as downweighting similarities to one class by how similar the term is to other classes similar to that class, but found they did not improve scores.

4.2.4 Analysis

One of the major challenges 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 which is the best method of defining the similarity and cut-off thresholds. Starting from a set of 2,122 ontology keyword/class pairs, 11814 new keyword/class pairs are generated, before a second stopword list is applied, to produce a final set of 9,076 pairs.

The result of the ontology population stage is thus a set of keywords associated with each class, each of which has a score indicating the degree of its relevance to that class. Table 1 shows the counts for the different types of keywords for each high-level topic, including those attached to subtopics, i.e. we include the keywords for all subclasses in the count. There is some overlap because occasionally, the same keyword can appear in both a higher-level class and one (or more) of its subclasses, though we aim to minimise this. As described above, 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 these 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. Generated terms are those created by the term extraction tool, while

[3] 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.

enriched terms come from the automatic enrichment process. Both of these may be of lower quality and get a lower weighting.

	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

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

5 Classification

The annotation tool classifies documents according to the best matching topics from the ontology. Each topic is matched based on a number of keywords, but a complex process defines how to rate the "quality" of these keywords (how well they indicate a particular topic) and how to combine the various keyword scores for each topic. This has been significantly enhanced since D2.3. For example, a single keyword might be relevant for more than one topic, but it might be more relevant for one topic than another, so it would get a higher score. At the classification stage, all relevant keywords and topics are scored, resulting in a list of topics and scores, together with the keywords they comprise (and their scores). At a later stage, cut-off thresholds are established for each document type, so that only the highest and most relevant topics will be added to the database for that document and used for the indicators (see Section 5.3).

5.1 Scoring process

The original scoring process was based on the number of matching keywords in a document and topic, normalised by document length, with some additional weighting for longer terms (which are thought to be more specific). We have enhanced the scoring process in a number of ways.

First, we deal with the problem of ambiguity. Some keywords are good indicators of a topic only when they appear in the same document as another keyword. For example, "packaging" could relate to many topics, but if it appears in the same document as the term "microelectronics" (not just as a multi-word term such as "microelectronics packaging", but as two distinct terms), it can be considered to be a good indicator of the topic of MNE (micro- and nano-electronics). We therefore want to weight more strongly terms that appear together with some specific other terms. The question is then how to find which are these "specific other terms".

PMI (Pointwise Mutual Information) is an indicator of lexical cohesion which considers terms to be more closely related the more often they occur together in a large training corpus. We therefore use this to "boost" the score of certain keyword pairs that occur together in a document – terms with high PMI will be more strongly boosted. To do this, we pre-calculate on our training corpus pairwise collocation statistics for all term candidates. We then select only those where the normalised PMI value >0 and the minimum occurrence frequency is 20 (based on heuristic experimentation). This gives us 309,932 pairs from an initial 2.8 million pairs.

From our enrichment process, each keyword already has a keyword score (kw). In the document classification stage, we generate a base score from this keyword score, as follows:

- kw score is multiplied by 2 if it fulfils certain criteria (e.g. a patent classification keyword is matched in a patent document, or a project classification keyword is matched in a project document);
- kw score is multiplied by 1.1 if it is a preferred term.

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

Next, the base score is boosted by *0.5 * the score of the highest scoring direct superclass* (if any). This accounts for the fact that terms belonging to a superclass should theoretically be less specific, though still relevant, so we want to boost the more specific terms over the more general ones so that we annotate the document at the most specific level possible.

Next, we integrate the PMI boosting. All keywords for a document are looked up in the matching pairs PMI table generated previously. We use the highest value of any matching pair, and boost the score by *1+PMI value*.

We finally generate two scores for each topic:

- unboosted: *100*base score / doc length*;
- boosted: *100*base score + PMIboost / doc length*.

5.2 The classification tool

Separate documentation for users of the classification tool is available publicly.[4] This explains the technical details of how to run the web service. In summary, the software provides a REST service on the USFD servers which accepts documents, classifies them according to the topics in the ontology, and returns classification and keyword information in JSON. This information is fed back into the KNOWMAK database.

Since the previous version of the classification tool, some improvements have been made – mainly in terms of providing additional output. Instead of producing a single score for a topic, multiple scores have been produced so that the different mechanisms can be evaluated. These match the different scores described in the previous section: standard score and boosted score, each with or without PMI boosting. For the boosted score, the URI of the topic which boosted it is also given. This also means that if necessary, different scoring mechanisms can be used for different kinds of documents.

An example of the output (in JSON format) is shown in Figure 4. This can be interpreted as follows:

- **Classification** shows the topic URL (in this case *antibiotics*).
- **Boosted by** shows the topic that boosted the score (in this case, *antimicrobials*). This means that keywords were found for both these two topics, but since *antimicrobials* is a superclass (more general) than *antibiotics*, the latter gets a score boost from the former.
- **Keywords** shows the relevant keywords found in the document that match this topic (in this case, *antibiotics* and *bacteria*).
- For each keyword, there are features **kind** and **score**
    - **Kind** shows the provenance of the keyword (was it automatically **generated** by the NLP tools, was it a **preferred** term (generated directly from a topic name, or manually added, and thus thought to be highly correct), or an **enriched** term (generated via the keyword enrichment techniques). This is

[4] https://gate.ac.uk/projects/knowmak/GATE-classification-tool-user-doc.pdf

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

useful to understand how a keyword was generated in case of a bad match (or indeed, a good one).
    - **Score** shows the score for that keyword, based on the scoring procedure described previously.
- **Score** (for the topic) shows the aggregated score for all keywords matching that topic, including the boosting process.
- **topicID** shows the number of the topic in the ontology, which is later added to the database (rather than the topic name) along with the document, once cutoff thresholds have been applied.
- **Unboosted score** (for the topic) shows the aggregated score for all keywords matching that topic, without the boosting process.

5.3 Topic assignment

The annotation tool assigns to each document as many topics as it finds matches for, without making any decision about which ones are valid. For example, some topics might have a very low score, and are clearly not relevant, but are not excluded at that stage. Instead, a further topic assignment stage is defined after the annotation. The reason for this is that this strategy might vary for different kinds of document. Furthermore, the strategies are based on an analysis of the entire set of annotated documents, and thus cannot be done at annotation time where each document is processed individually.

The strategies for each document type were originally determined independently, in case different strategies were required for different document types. In general, we found that out of the 4 scoring options (with or without PMI, and with or without boosted scoring, both PMI and the boosted score gave best results. After extensive experimentation, it was resolved that a single strategy could be used successfully for all document types. This consists of the following:

1. Suppress all class assignments for which the boosted score > 2*unboosted score (without PMI)
2. Then, using the PMI boosted score, compute the mean for a class on all documents assigned to that class.
3. Keep assignments for which the boosted class score / (keywords found in the document) > class mean

In previous versions, we had also removed classes for which there was only one matched keyword in a document, but this was found to be too restrictive, so we rejected this strategy. In Section 6, we describe further some of the testing methods we used to adopt the final strategy.

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

```
{
  "classification": {
    "http://www.gate.ac.uk/ns/ontologies/knowmak/antibiotics": {
      "boostedBy":
"http://www.gate.ac.uk/ns/ontologies/knowmak/antimicrobials",
      "keywords": {
        "antibiotics": {
          "kinds": [ "generated", "preferred" ],
          "score": 1.1527377521613833
        },
        "bacteria": {
          "kinds": ["generated"],
          "score": 0.5763688760806917
... },

      "score": [ 4.322766570605188, 4.4159785333 ],
      "topicID": "38",
      "unboostedScore": [ 2.5936599423631126, 3.75354899915 ],
    },
    "http://www.gate.ac.uk/ns/ontologies/knowmak/antimicrobial_resistance": {
      "boostedBy":
"http://www.gate.ac.uk/ns/ontologies/knowmak/antimicrobials",
      "keywords": {
...
} },

      "score": [ 8.069164265129682, 9.12545454545 ],
      "topicID": "42",
      "unboostedScore": [ 6.340057636887607, 7.35454545454 ]
    },
    "http://www.gate.ac.uk/ns/ontologies/knowmak/antimicrobials": {
      "keywords": {
...
} },

      "score": [ 3.4582132564841506, 4.54545452388 ],
      "topicID": "43",
      "unboostedScore": [ 3.4582132564841506, 4.54545452388 ],
}, },

  "doc_type": "publication",
  "doc_type_applied": "publication",
  "error": "_none_",
  "identifier": "12348874",
  "internalID": "4dec1ce0-cead-4a94-b16c-6fada1a26f49"
`
```


Figure 4: Example of JSON output from the classifier

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

6 Ontology 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.

As described in D2.2, 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 earlier in Section 3); assessing the relevance and coverage of the keywords attached to the classes (described below in Section 6.1); and a task-based assessment of the ontology (described below in Section 6.2), involving checking that there is minimal overlap between class assignment and that all classes have sufficient – but not too many - documents assigned.

6.1 Keyword evaluation

The quality of keywords is critical for the success of the annotation. 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:

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

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.

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).

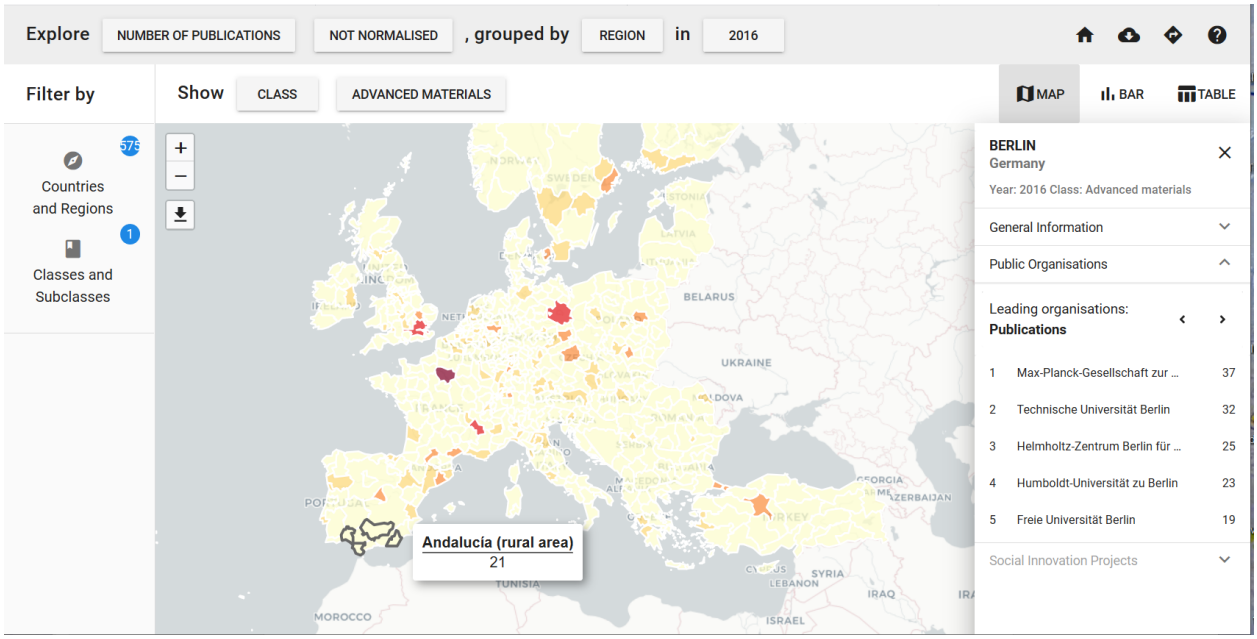
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

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.

6.2 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 5).

Figure 5: 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

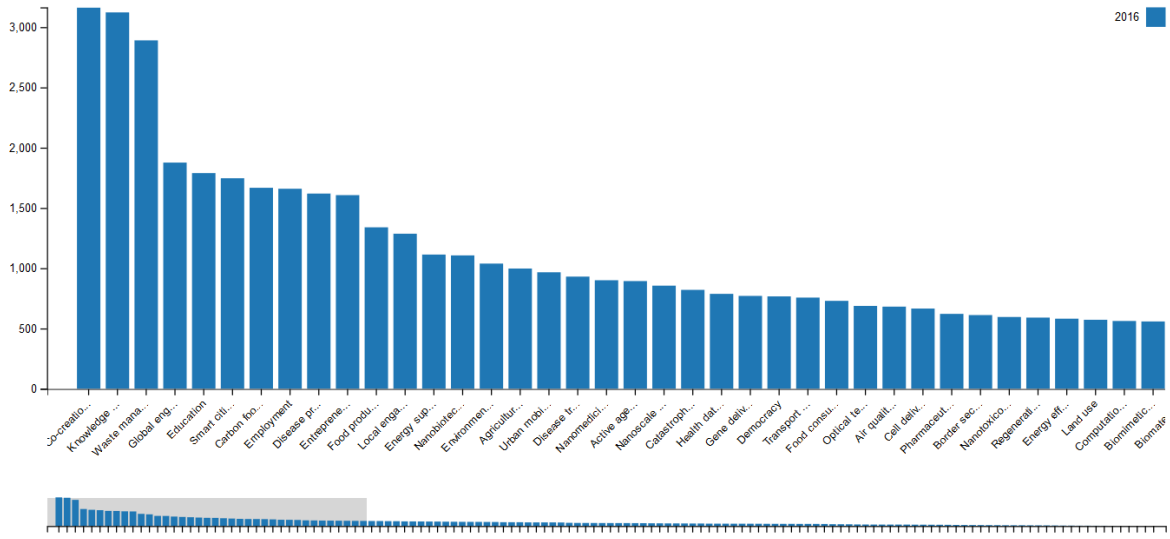
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 726992.

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 keywords, which strongly inflate individual classes. As shown in Figure 6, 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 6: 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 5). In this respect, one can check for differences in the top knowledge producers by topic.

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

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 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.

7 Discussion and future work

In this work package, 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.

7.1 Future work

The work presented in this report has demonstrated the feasibility of the ontology-based approach to document tagging with topics, but it has also highlighted an important number of issues. There are a number of ways in which the work can be extended, some of which are being actively pursued in the RISIS project.

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

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

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. The ontology structure could also be refined or further extended, for example to incorporate Missions. We will continue to test this with users in order to get further feedback on the structure, and a mechanism could be set up whereby users can suggest potential new keywords.

Additionally, we are currently experimenting with an alternative way of annotating documents with ontology topics. This is ongoing experimental work which we have not yet integrated into the classification tool. Experimentation to understand if this approach is feasible, and how it can be integrated with the existing keyword-based classification approach, is still to be done, and will be pursued within the RISIS project. This will be combined with further evaluation of the classification process generally.

The basic idea is that in addition to the keyword-based classification approach, topic modelling could also be used in order to rank the topics according to similarity with each document. Latent Dirichlet Allocation (LDA) is the most widely used topic modelling algorithm in natural language processing. LDA assumes that the document is a consistent mixture of *k* topics. Therefore, we can use LDA to classify the input documents by comparing the similarity between the probability distribution of the input document topics to the distribution of the ontology class topics.

The current approach includes the following steps. At training time, we:

1. train an LDA model (*n* topics) based on relevant training documents;
2. find the most relevant Wikipedia page as ontology class documents;
3. calculate the topic distributions for each class, based on the class document and trained model.

To train the model, we use a collection of abstracts from projects, patents, and publications, as well as policy documents (the same collection as used for the embeddings training). The training documents are preprocessed by removing stop words and stemming. To find the most relevant Wikipedia page, we first use the class label as a search query, and score the Wikipedia page with gestalt pattern matching between class label and the Wikipedia page. If the matching score is less than 0.7, then we reset the search queries as keywords, and use the most relevant page. Once we have the trained topic model and class documents, we can then calculate the distribution on each topic of the document, and this will return *n* element vectors for each class.

At application time, we:

4. calculate the input document topic distributions using the trained model;
5. calculate the cosine similarity between ontology class and document topic distributions;
6. return *k* most similar ontology class ids.

The development is still underway, but the idea eventually is to apply the ontology class keywords as a guide to train the LDA topic model, and then improve the ontology class documents by improving the string match algorithm; searching on a

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

different database besides Wikipedia; and combining the topic modelling algorithm with the existing classification algorithm. If successful, this methodology will address some of the limitations previously discussed around the reliance on keywords, and facilitate the transition to new data and languages, but it will still need to be tempered with expert verification at every stage of the process.

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

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.

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

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

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.