The University Of Sheffield.

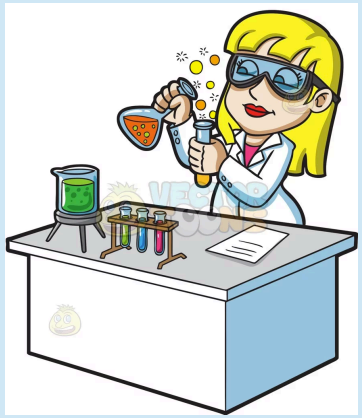
general architecture for text engineering GATE

Semantic technologies for mapping European research: bridging the gap between policy and research

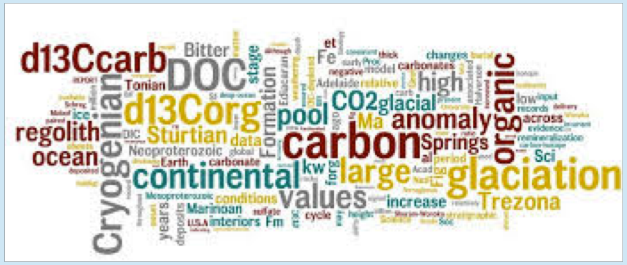
Dr. Diana Maynard
University of Sheffield, UK

KNOWMAK
Knowledge in the Making in the European Society

The KNOWMAK project



- 3-year EU H2020 project since January 2017
- Provide interactive visualizations and indicators on knowledge co-creation in European research
- Based around:
  - **Research Actors** (organisations)
  - **Research topics**
  - **Geographical spaces** (based on NUTS and FUA)



http://knowmak.eu

Classical Knowledge Production
Scientific
Technological
Project-based

Non-Classical Knowledge Production
Social Innovation
Citizens' Attention

KNOWLEDGE in the **MAK**ING

Central KNOWMAK Database

KNOWMAK Indicators System

Classical sources (publications, patens, projects)

Non-classical sources (social innovation, social media)

Topics
(KET, SGC)

Geography
Countries
Regions

Standardized
Key Actors

Supporting data

The KNOWMAK Tool
Indicators extraction and analysis


Retrieval


Explanation




Exploration

How is European knowledge distributed across regions?

- A composite indicator combining publications, patents and projects shows that:
  - the volume of knowledge production is highly concentrated in large metropolitan regions, e.g. Paris, London, Munich
  - some medium-sized regions are highly productive in terms of intensity (normalised by population), e.g. Eindhoven and Heidelberg
  - some smaller areas have high volume and intensity, e.g. Oxfordshire
  - Eastern Europe shows low volume and intensity, except major cities, but all have low intensity (except Ljubljana)

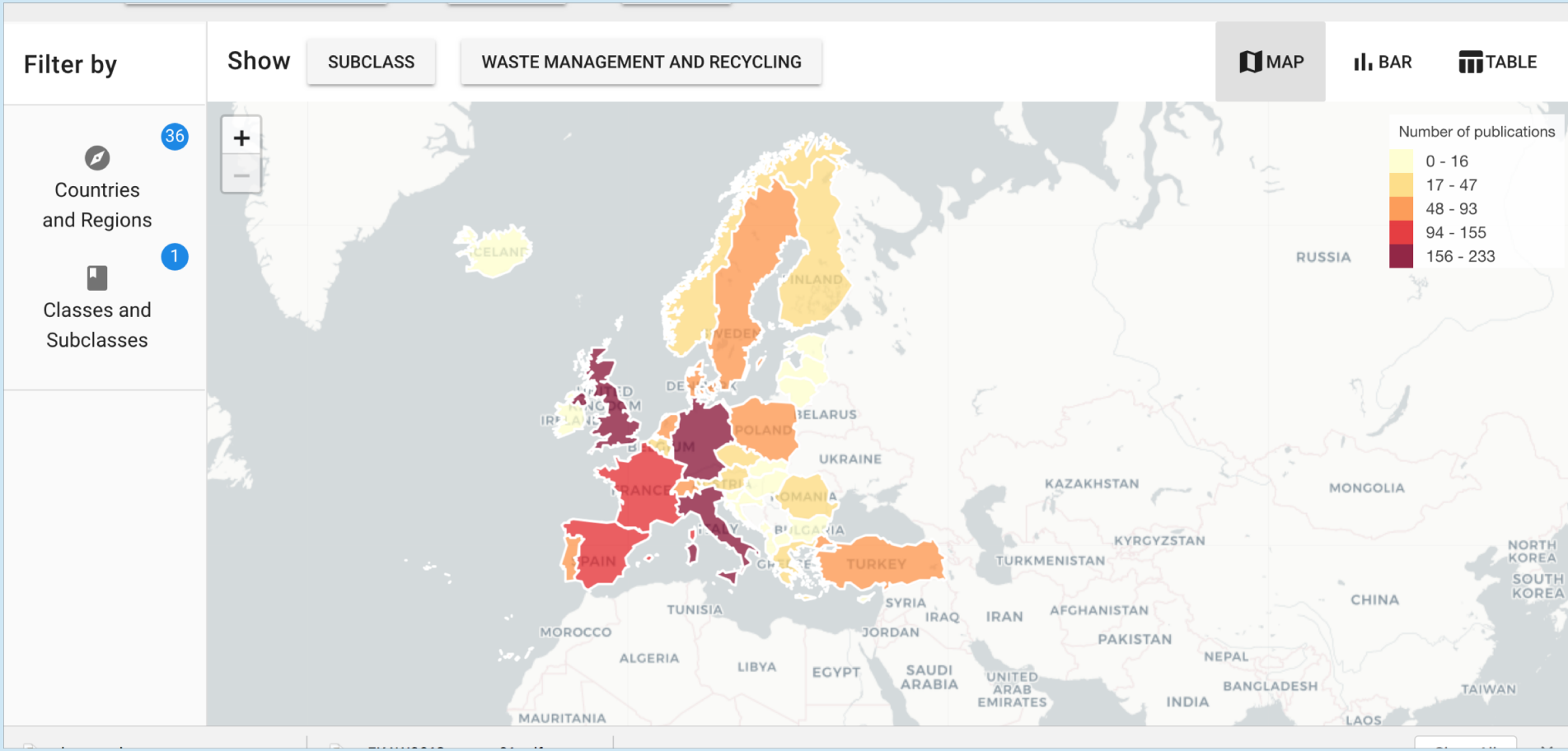
Technological vs scientific knowledge production in genomics

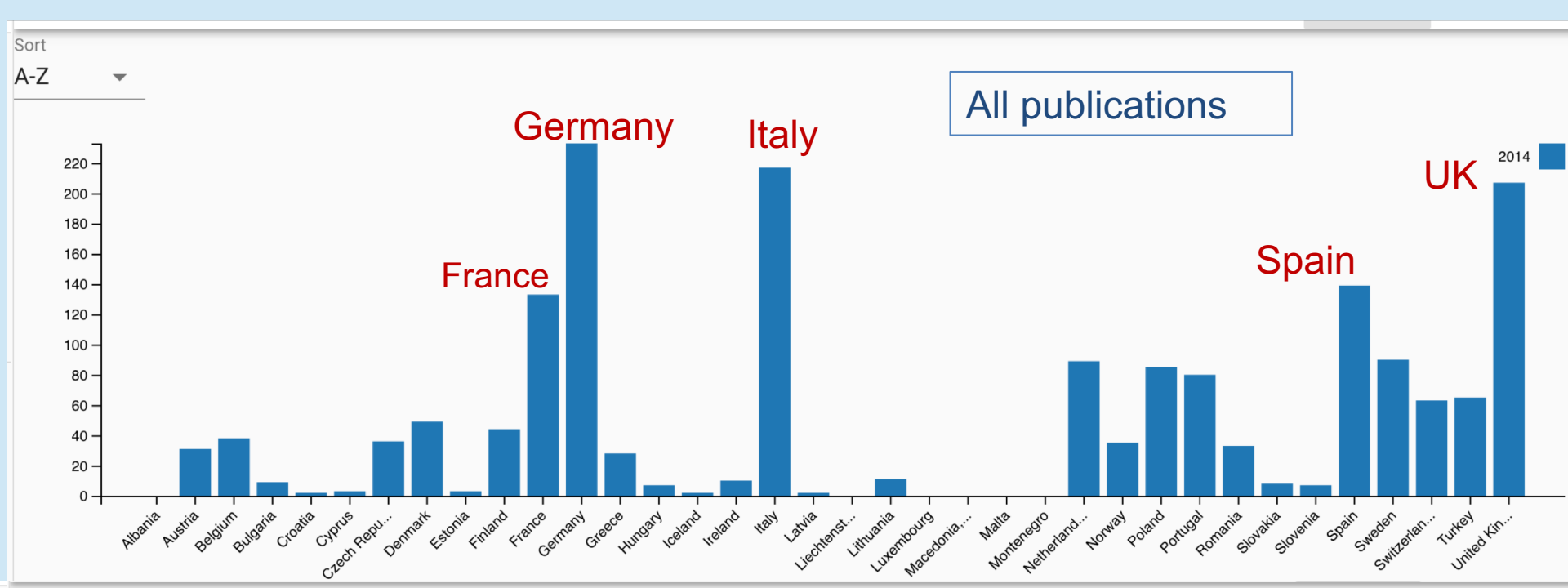
- **Technological** production is measured by **patents**
- **Scientific** production is measured by **publications**
- These 2 types show different geographical distributions: technological are more concentrated in space
- In terms of volume, Paris is the biggest cluster for both types
- Within regions, production varies a lot: London is the biggest producer of both types, while Eindhoven is key in terms of technological knowledge (both for volume and intensity)
- These findings reflect the different structure of public and private knowledge

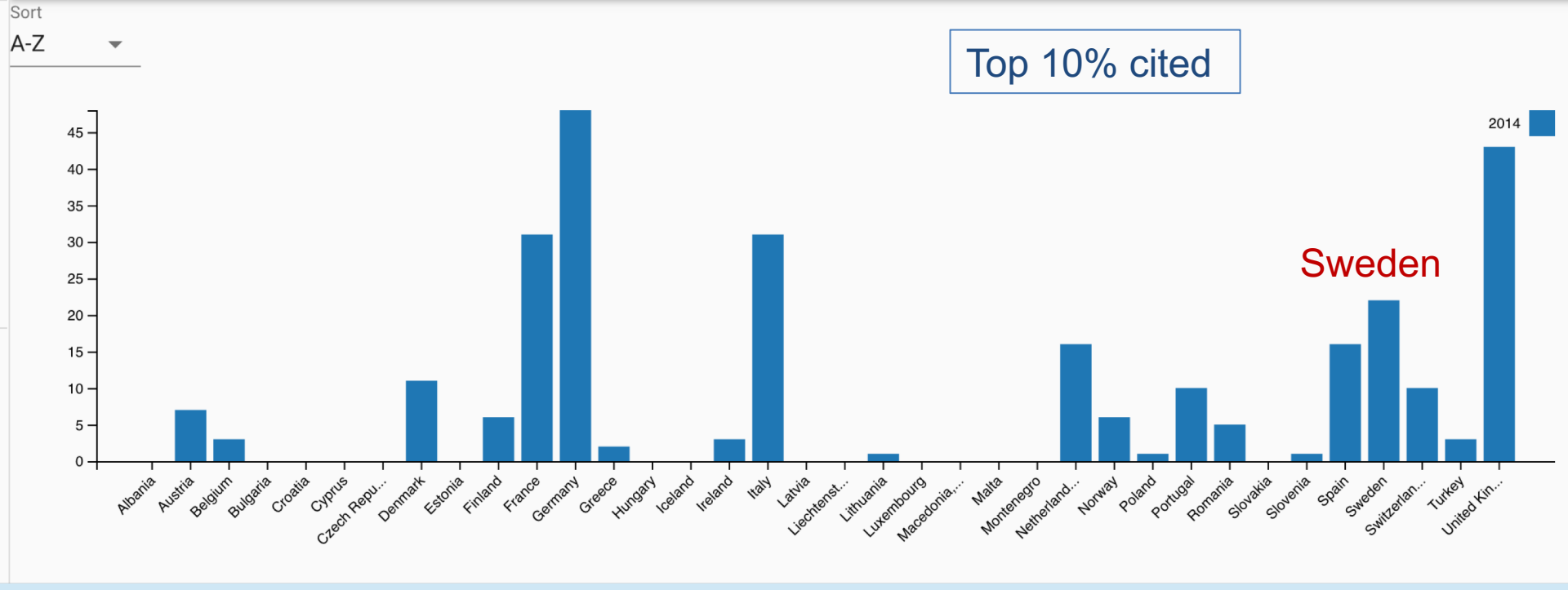
What kind of questions can we answer?

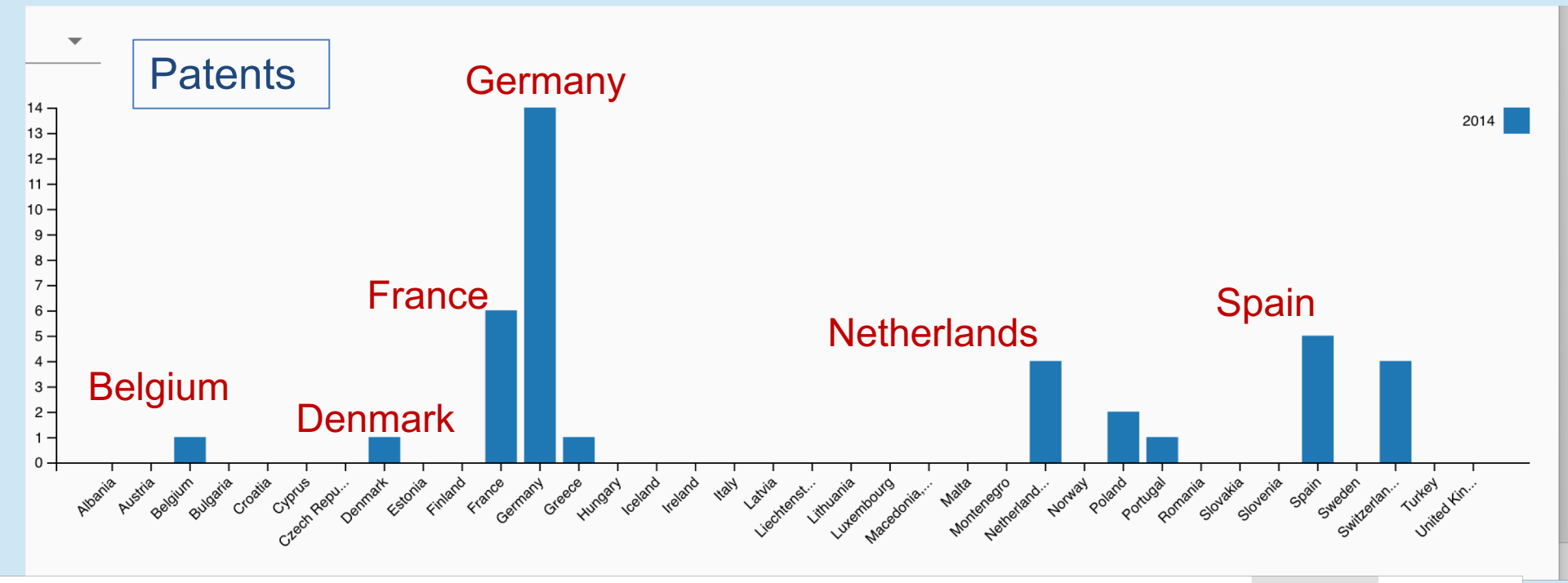
- Which country published most about waste management and recycling in 2014?
- What happens when you look only at the top 10% most cited?
- What kind of international collaborations do we see?
- What about patents?

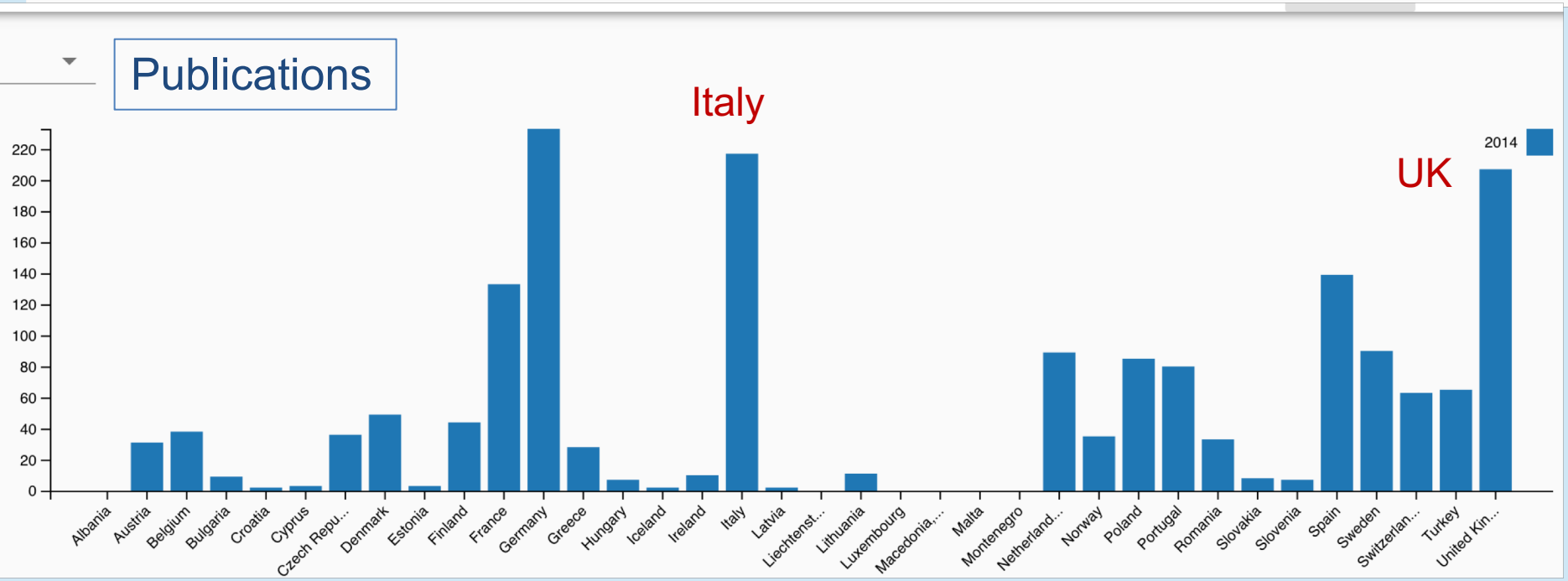
Which countries published most about waste management and recycling in 2014?




Sort
A-Z
All publications
Germany
Italy
France
UK
Spain
2014
0
20
40
60
80
100
120
140
160
180
200
220
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Repu...
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenst...
Lithuania
Luxembourg
Macedonia,...
Malta
Montenegro
Netherland...
Norway
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerlan...
Turkey
United Kin...


Sort
A-Z
Top 10% cited
Sweden
2014
0
5
10
15
20
25
30
35
40
45
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Repu...
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenst...
Lithuania
Luxembourg
Macedonia,...
Malta
Montenegro
Netherland...
Norway
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerlan...
Turkey
United Kin...


Patents
Germany
France
Belgium
Denmark
Netherlands
Spain
2014
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Repu...
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenst...
Lithuania
Luxembourg
Macedonia,...
Malta
Montenegro
Netherland...
Norway
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerlan...
Turkey
United Kin...


Publications
Italy
UK
2014
0
20
40
60
80
100
120
140
160
180
200
220
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Repu...
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenst...
Lithuania
Luxembourg
Macedonia,...
Malta
Montenegro
Netherland...
Norway
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerlan...
Turkey
United Kin...

Semantic Technologies in the Science and Technology Landscape

Opportunities:

- Ability to link different kinds of data sources to provide a richer view of knowledge production

Challenges

- Need for a robust approach to identify and model relevant topics
- **Language** (connect different kinds of data due to terminology differences)
- **Commensurability** (cannot connect different kinds of classifications)
- **Flexibility** (model changes over time and space)

Semantic Approach

What is the innovation performance of France on climate change compared with Germany?

- climate
  - air_quality_management
  - biomaterials
  - forests_flora_fauna
  - noise
  - soil
  - waste_management_and_recycl
  - water_and_wastewater

6687 2007 0
LED module with gold bonding.
Processes or apparatus specially adapted for the manufacture or treatment of semiconductor

SC5-20-2014
H2020
Zero Emission Robot-Boat for Coastal and Inland Water Monitoring

ghg
Perspectives on CO2 capture and storage
Filipp Johnsson
Published 14-04-11

Policy ⟷ Ontology ← Data

In a nutshell:

- We need to know which topics each document is talking about (multi-class classification)
- But we have to connect these topics together coherently

The datasets

CWTS-WoS

Enhanced version of Thomson Reuters publication and citation indexes, convering almost 13,000 current international peer reviewed journals and around 15 million publications and all their references

IFRIS-PATSTAT

Global patent data recorded in PATSTAT (patent holders, inventors, technological classification, fine grain patents type selection, etc.), enriched by external data sources and cleaned/standardized information.

EUPRO

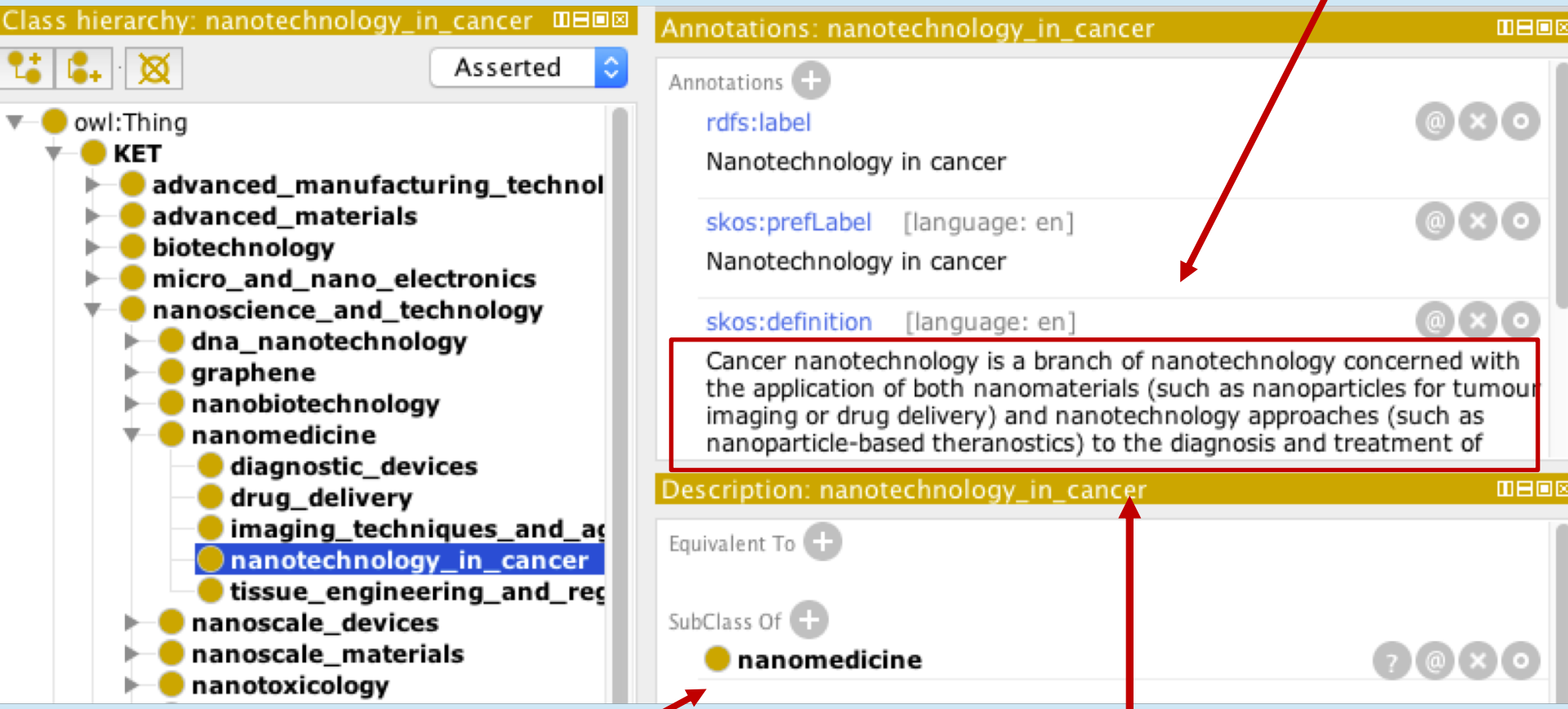
Systematic information on R&D projects and all participating organizations funded by the European Framework Programmes (EU-FPs). EUPRO covers information on projects and participations (FP1-H2020)

The role of ontologies

- Translate generic user queries related to policy-making into a formal structure of classes and keywords linked to data sources
- Ensures that queries are restricted to relevant topics and that relevant topics are covered
- Offer a flexible solution allowing
  - variations of language and terminology
  - connections between concepts (at both the topic and keyword level)
  - adaptability over time and topics of interest
  - different levels of aggregation
  - minimal user input when changes are required

Ontologies connect information

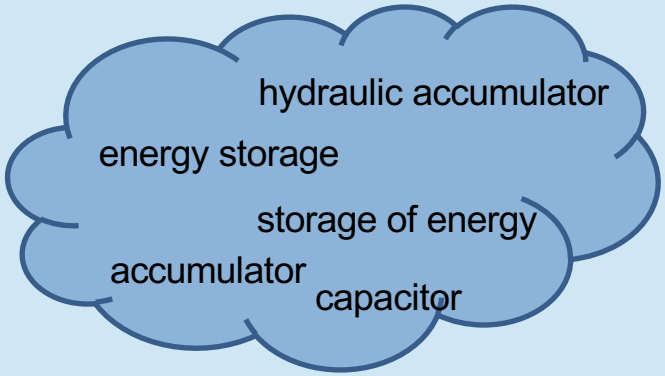
Link with other sources (Nature.com, skos, DBpedia…)



Link related topics

Find more information about the topic

From ontology to data



1. Create ontology of topics representing KET and SGC
    - From existing classifications, policy documents, expert users, and data
2. Automatically generate collections of keywords
    - NLP techniques (term extraction, word embeddings) from large training dataset
    - Ranking and scoring algorithms to decide:
        - Which topic(s) to match the keywords to?
        - Which are the best keywords?
        - Which are the best keyword combinations?
3. For each document, decide which topics best fit it
    - based on keywords and scoring algorithms

Topics: Societal Grand Challenges

Health	Health, demographic change and wellbeing
Bioeconomy	Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the bio-economy
Energy	Secure, clean and efficient energy
Transport	Smart, green and integrated transport
Climate	Climate action, environment, resource efficiency and raw materials
Security	Secure societies - protecting freedom and security of Europe and its citizens
Society	Europe in a changing world - inclusive, innovative and reflective societies

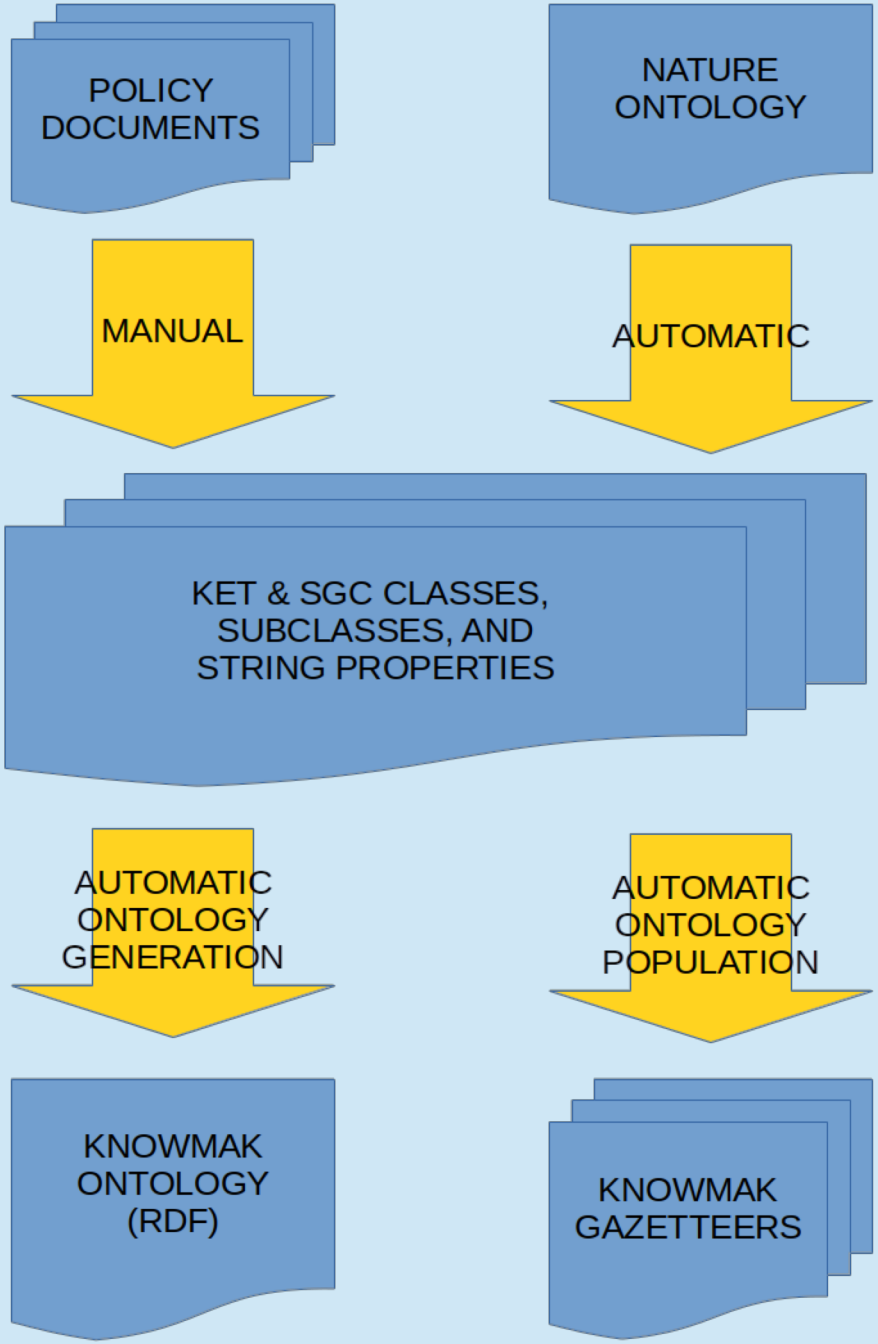
Topics: Key Emerging Technologies

IB	Industrial Biotechnology
NANO	Nanotechnologies
PHOT	Photonics
AMT	Advanced Manufacturing Technology
NME	Micro- and Nano-Electronics
AM	Advanced Materials

- Overlap between the topics
- AMT is designed to be cross-cutting over the other 5
- Problems for ontology design (and topic assignment)
- Different vocabulary is used in each

Creating and populating the ontology

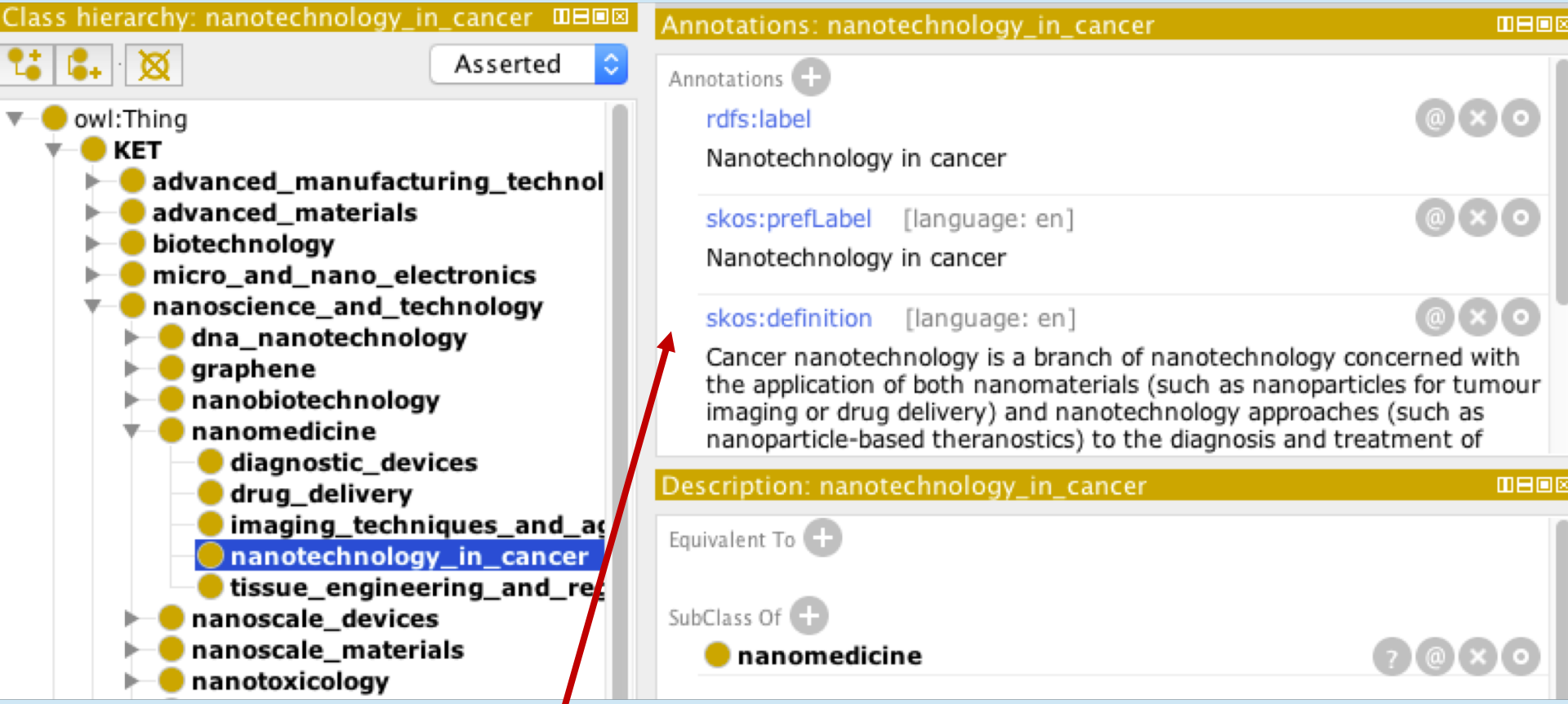
1. Create ontology structure (classes & subclasses)
2. Add extra information (descriptions, links, alternate class names)
3. Ontology population: generate lists of terms associated with each class (gazetteers)



Step 1: Ontology generation

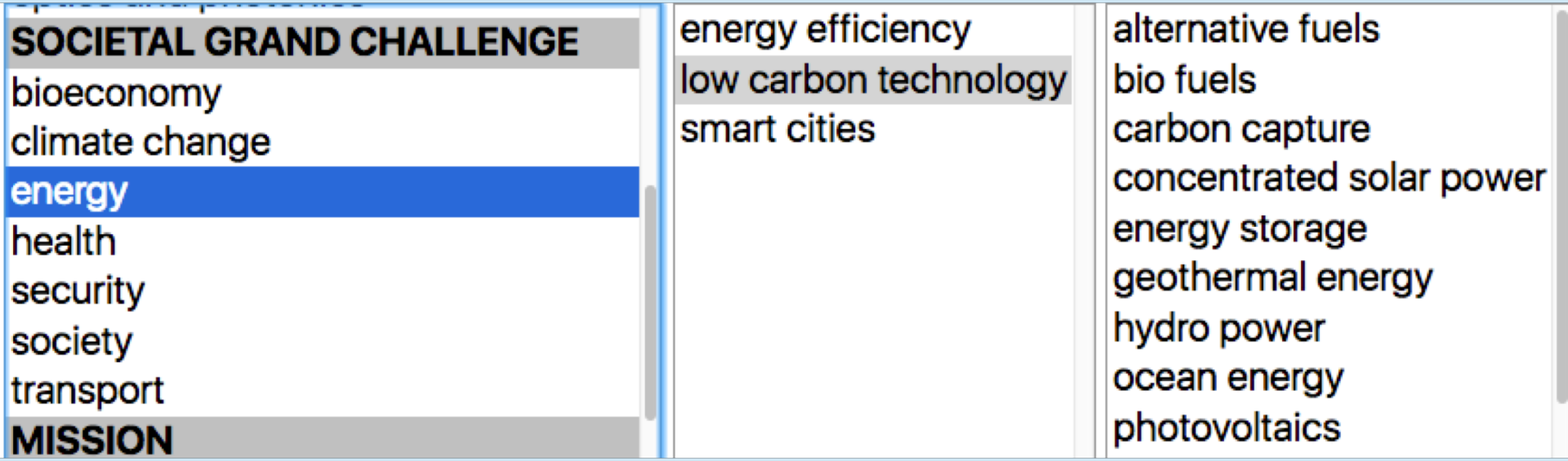
- Mixture of manual and automatic methodology
- Start with high-level KETs and SGCs
- For KETs, reuse existing Nature classification where relevant
- This includes useful extra information (links to other data sources, definitions etc.)
- Add subtopics based on definitions of KETs and SGCs in policy documents
- Add links to patent classification hierarchy and project classification topics

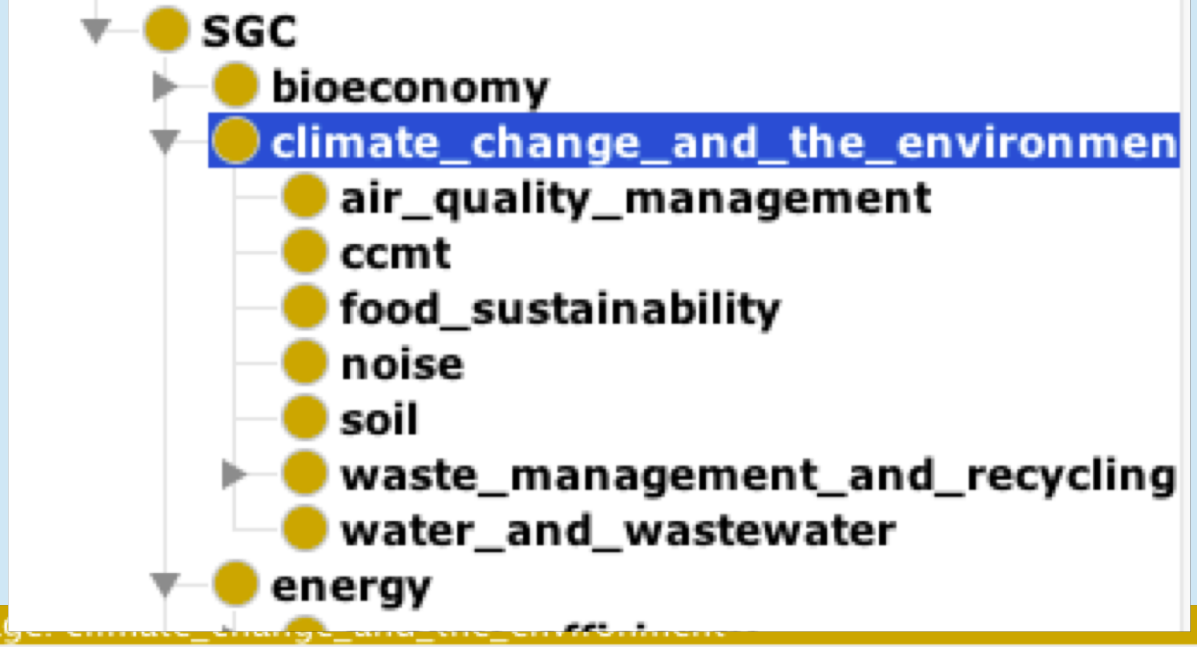
Linking information from Nature.com



Link to more information

SGC Topics and SubTopics





how: this disjoints named sub/superclasses

climate_change_and_the_environment

- climate_change_and_the_environment provenance "SGC-IPC-mapping.xlsx + ipc.xlsx"
- climate_change_and_the_environment rdfs:label "Climate change and the environment"
- climate_change_and_the_environment projectKeywords "Climate change and carbon cycle research"
- climate_change_and_the_environment skos:prefLabel "climate change"
- climate_change_and_the_environment topicID 99
- climate_change_and_the_environment provenance "Fraunhofer"
- climate_change_and_the_environment rdfs:label "environmental protection"
- climate_change_and_the_environment description "Climate change and carbon cycle research. Climate c
- climate_change_and_the_environment provenance "eupro-classes.xlsx"

food_sustainability

food_sustainability SubClassOf climate_change_and_the_environment

Step 2: Ontology population

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

Generation of keywords from the ontology

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.

- Automatically generate keywords from class names, descriptions, and related information (e.g. DBpedia, skos, etc.) using term recognition tools
- Separate 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)

Problems with basic keyword generation

- Not enough keywords for classification – many documents don't get annotated with a topic
- Keywords not relevant for policies / user queries
- Keywords not relevant for patents
- Some topics have many more keywords than others – inconsistencies lead to annotation bias
- Some keywords are too general to be useful
- Some keywords are too ambiguous (connected with multiple topics)
- Many keywords are correctly related to topic but not *indicative of it,* e.g. there could be some connection with the topic, but it's not necessarily the best topic fit

Solution: corpus-based ontology enrichment

- Create a set of additional keywords for each class in the ontology using an automatic unsupervised approach
- Create a large corpus of patent, publication and project abstracts as well as relevant policy documents
- Extract new candidate terms from this corpus
- Train domain-specific word embeddings for these terms
- Train in such a way that we can have vectors also for multi-word terms
- Use the embeddings to find the similarity between the seed terms and the new terms
- Use the similarity to decide which new terms to keep, and which concept to map them to

Steps for enrichment process (1)

- Corpus pre-processing (2.6 million documents in total)
- GATE application for linguistic pre-processing (POS tagging, lemmatisation, entity finding, etc.)
  - Find all occurrences of original ontology keywords in corpus (both lemmatised)
  - Find single and multi-word term candidates in the corpus (filter out NEs)
  - Merge ontology matches and term candidates and create (potentially overlapping) keyword candidates; calculate canonical lemmatized string for them
  - Calculate term statistics for all term candidates (tf, df, idf)

➢ 1.2 million keyword candidates in 180 million locations

Training the embeddings

- Calculate a set of 330 stopwords (also used for scoring) and a set of unique MWTs from the original ontology keywords
- Embeddings trained using Python GenSim, removing stopwords and single-letter words from the corpus
- Sentences used as training examples.
  - Match each sentence against the list of MWTs
  - For each MWT, create one sentence where each lexical unit is a separate MWT
  - Create one sentence also where all MWTs are single lexical units

➢ 591526 embeddings generated

Scoring (1)

- Investigated various ways of calculating embeddings to represent ontology topics and measure similarity between the keyword and class.
- Best results with:
  - *Centrboth*: for each class, calculate average embedding for set of preferred terms and another average embedding for set of non-preferred terms related to the class.
  - Final embedding is the weighted average of both (.0.75 preferred, 0.25 non-preferred)
  - *Simonly*: 0/1 normalised cosine similarity between the embeddings representing the ontology class (centrboth) and the embedding representing the candidate term
  - In both cases for simonly, we take the unweighted average since weighted (tf, idf) did not work well in early experiments

Scoring (2)

- For keywords already in the ontology:
  - Assign a preferred term a score of 1.0
  - Assign a non-preferred term a score of *simonly(t,c)*
- For new candidate keywords:
  - Select those with doc frequency between 20-100,000
  - Select those with term frequency between 50-500,000
  - For each of these, take top 10 closest classes according to simonly metric
- Filter both ontology keywords and candidate keywords by threshold of sim > 0.76

➢ 2122 ontology keyword/class pairs, 11814 new keyword/class pairs

PMI Boosting

- Due to ambiguity, some keywords are good indicators of a topic only when they appear in the same document as another keyword
  - e.g. “packaging” could relate to many topics, but in conjunction with “microelectronics” (not just an MWT), it’s a good indicator of MNE topics
- We use PMI to “boost” the score of certain keyword pairs that occur together in a document
- Calculate pairwise collocation statistics for all term candidates over the corpus
- Select only those where normalised PMI value >0 and minimum occurrence frequency is 20

➢ 309932 pairs from 2.8 million

Step 3: Annotating Data with Ontologies

- Data sources are annotated against the ontologies
  - each document is associated with one or more topics
- Sophisticated NLP matching of keywords in the documents (from titles, abstracts etc) with ontology
- A REST service accepts documents, classifies them according to the ontology, and returns classification and keyword information
- Several million documents can be processed in about a week (using around 12 threads)
- Annotated data sources are then used to build indicators
  - e.g. for each topic, how many publications and in which region?

Scoring the keywords and topics

- Base score is generated from Keyword (kw) score (from previous step):
  - multiplied by 2 if it fulfils certain criteria (e.g. a patent classification keyword is matched in a patent doc)
  - multiplied by 1.1 if it's a preferred term
- Base score is boosted by 0.5* score of the highest scoring direct superclass (if any)
- All keywords for a document are looked up in the matching pairs PMI table generated previously
  - Highest value of any matching pair is used
  - PMIboost score is 1+PMI value
- Final score is 100*base score / doc length
- Final boosted score is 100*base score + PMIboost / doc length

Classifier output

- **Classification**: topic URL e.g. *antibiotics*
- **Boosted by**: topic that boosted the score, e.g. *antimicrobials* (*antibiotics* is more specific, so it gets boosted by the keyword belonging to a more general topic)
- **Keywords**: all keywords in the document that match one in the ontology (e.g. *antibiotics*, *antimicrobials*)
  - **Kind**: provenance of the keyword (preferred, generated, etc.)
  - **Score**: score for that keyword
- **Score:** (for the topic) is the aggregated score of all the keywords, including boosting if applicable
- **Topic ID**: for use in the database
- **Unboosted score**: as above, without the boosting

{"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
…SEP},
"score": [4.322766570605188, 4.4159785333],
"topicID": "38",
"unboostedScore": [2.5936599423631126, 3.75354899915],
},

Example of patent annotation

Protein stabilized pharmacologically active agents, methods for the preparation thereof and methods for the use thereof

In accordance with the present invention, there are provided compositions and methods useful for the in vivo delivery of substantially water-insoluble pharmacologically active agents (such as the anti-cancer drug paclitaxel) in which the pharmacologically active agent is delivered in the form of suspended particles coated with protein (which acts as a stabilizing agent)…..

- RNA vaccines: (agent, protein, vaccine)
- anti-viral agents: (protein, anti-cancer, drug)
- protein vaccines: (protein, vaccine, antimicrobial)

KET: Industrial biotechnology
SGC: Health

Ongoing Challenges

Inconsistencies

- ontology design has to be tailored to user needs, but these are not uniform

Automation

- keyword-based approach still requires some manual intervention for best results

Accuracy

- language processing is never 100% accurate

Evaluation

- how do we know if/when it's good enough?
- Determine weighting mechanisms; cut-off thresholds…

The future?

- integration of existing classification and modelling approaches with our semantics

Acknowledgements

Johann Petrak did all the complicated stuff ☺

Adam Funk did the integration

This work supported by the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework and H2020 Programmes for R&D KNOWMAK (726992) http://knowmak.eu

- KNOWMAK website and tool: http://knowmak.eu
- Our work on KNOWMAK (demos, publications etc): http://gate.ac.uk/projects/knowmak