Semantic Technologies in Scientometrics: the KNOWMAK project

Semantic Technologies in Scientometrics

Opportunities:

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

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)

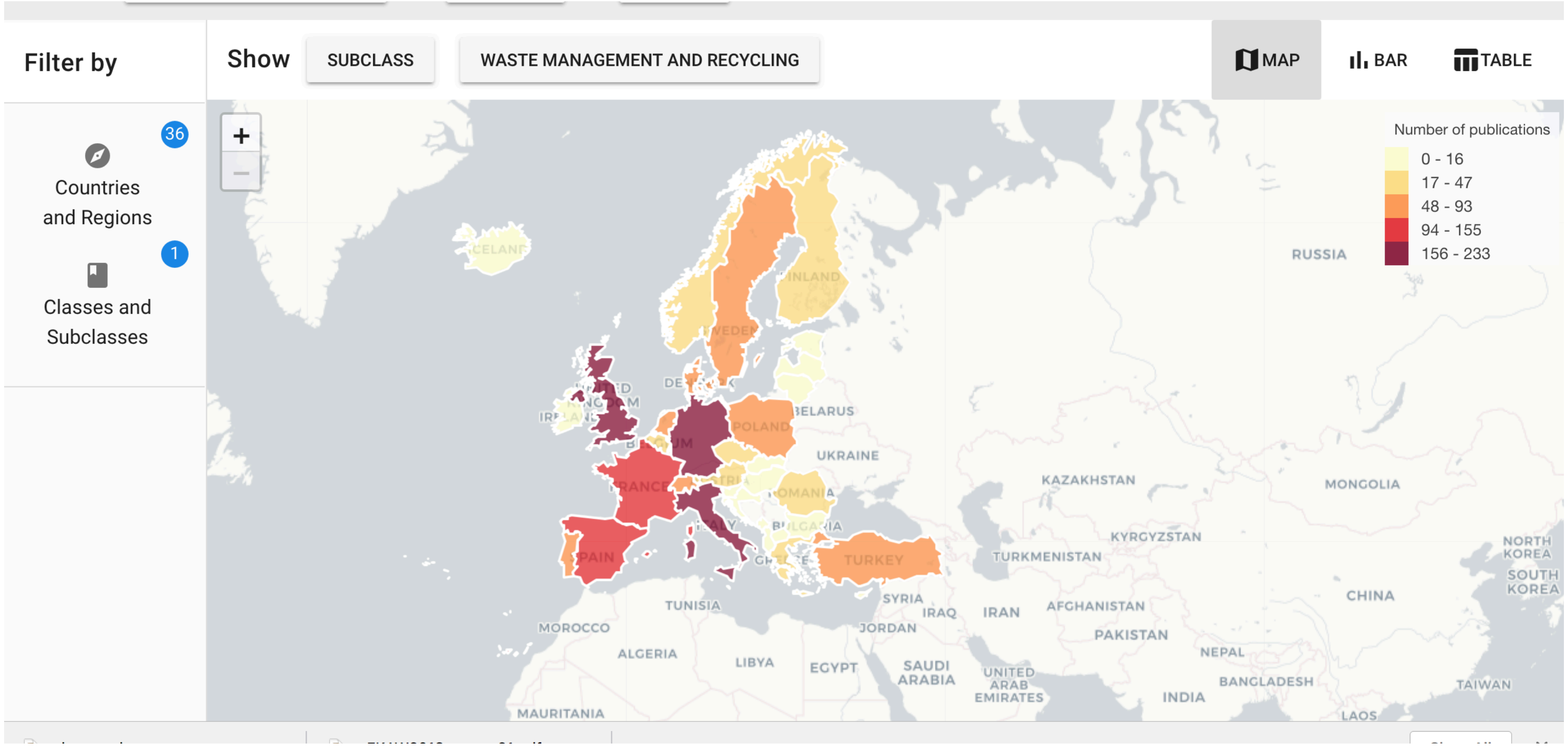
What kind of questions do we want to answer?

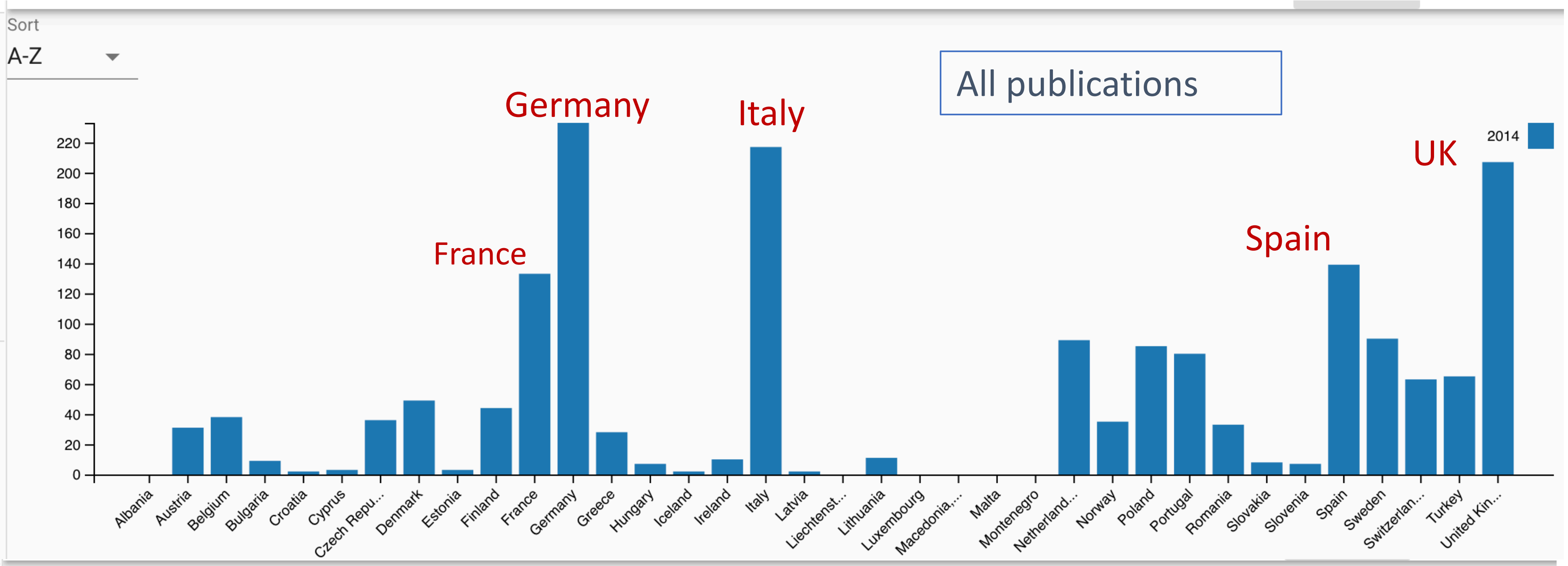
- Which country published most about waste management and recycling in 2014?
- What happens when you look only at the top 10% most cited?
- What kind of international collaborations do we see?
- What about patents?

The problem:

- topics for different document types don’t match – different classification systems
- and they don’t correlate with EU policies

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





Sort

A-Z

All publications

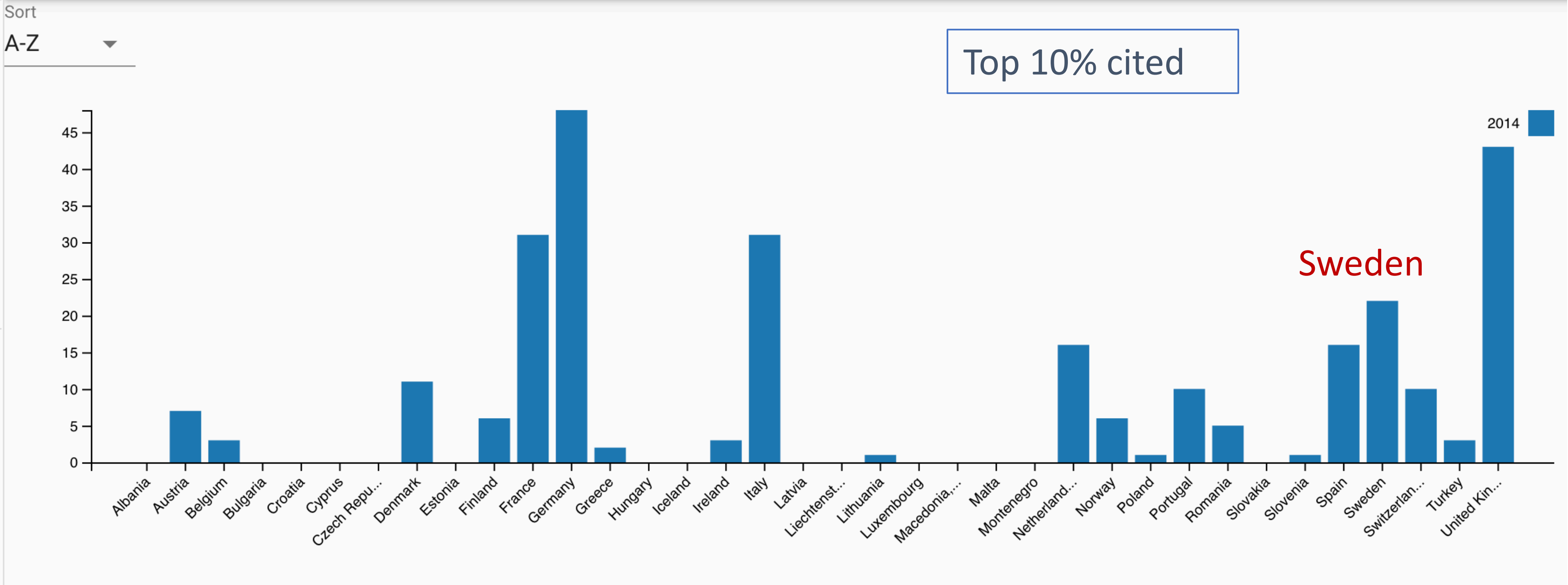
Germany

Italy

France

Spain

UK



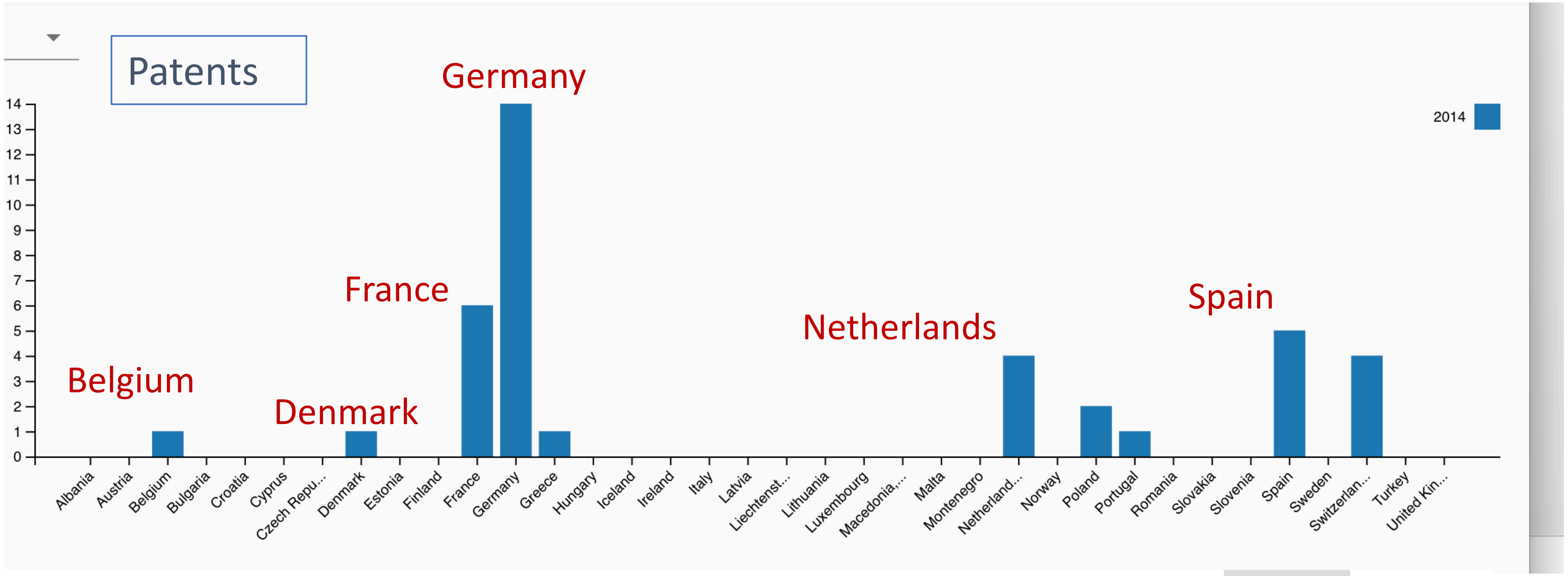
Sort

A-Z

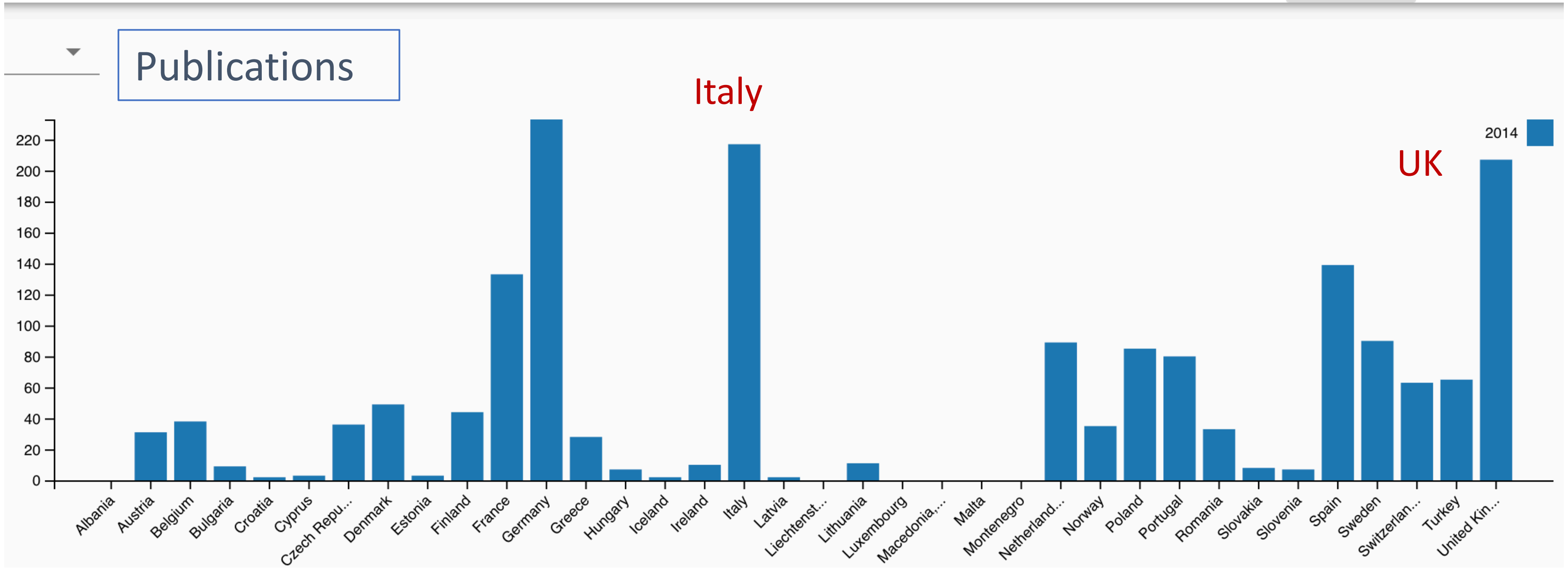
Top 10% cited

Sweden

Patents



Publications



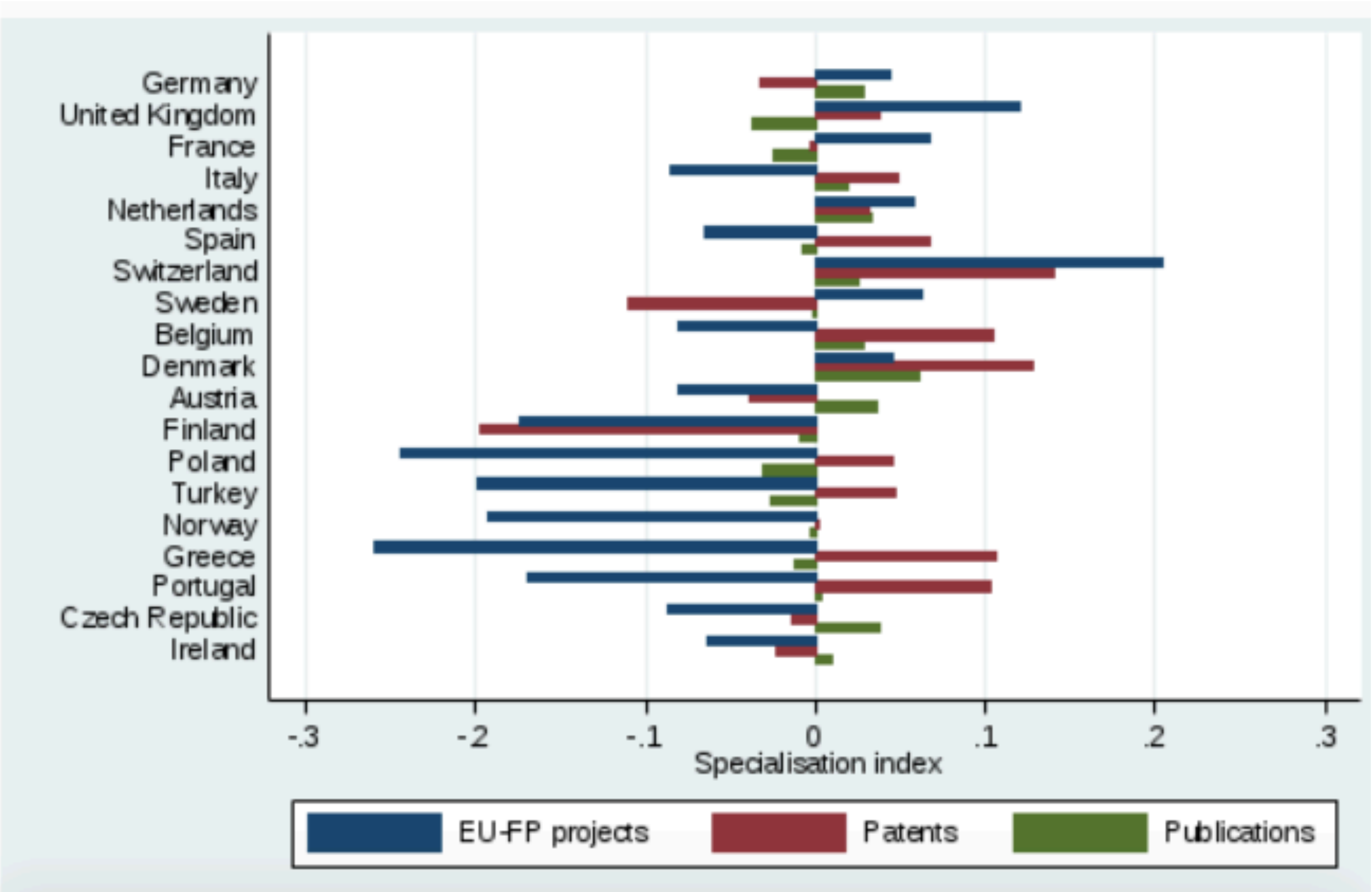
How is European knowledge distributed across regions?

- 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

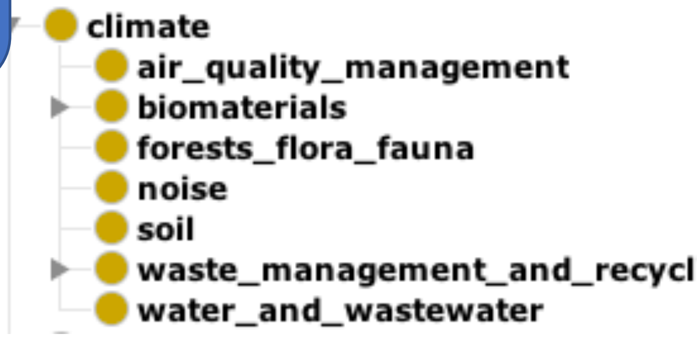
- **Technological** production is measured by **patents**
- **Scientific** production is measured by **publications**
- These 2 types show different geographical distributions: technological are more concentrated in space
- In terms of volume, Paris is the biggest cluster for both types
- Within regions, 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

Specialisation Indexes in Biotechnology around Europe



The 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

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

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

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

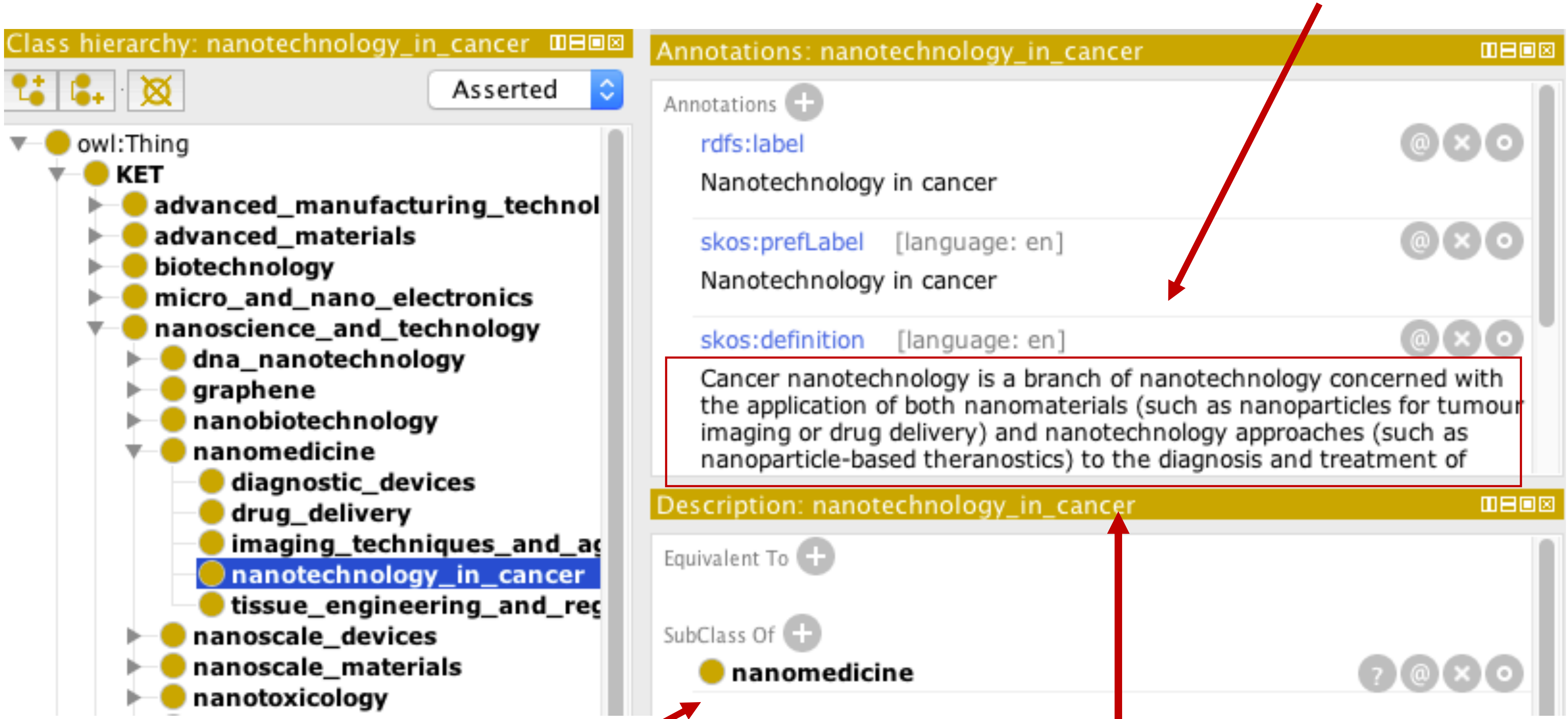
Policy ⟷ Ontology ⟵ Data

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

Ontologies connect information

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



Class hierarchy: nanotechnology_in_cancer

Asserted

- owl:Thing
  - KET
    - advanced_manufacturing_technol
    - advanced_materials
    - biotechnology
    - micro_and_nano_electronics
    - nanoscience_and_technology
      - dna_nanotechnology
      - graphene
      - nanobiotechnology
      - nanomedicine
        - diagnostic_devices
        - drug_delivery
        - imaging_techniques_and_ag
        - nanotechnology_in_cancer
        - tissue_engineering_and_reg
      - nanoscale_devices
      - nanoscale_materials
      - nanotoxicology

Annotations: nanotechnology_in_cancer

Annotations

rdfs:label
Nanotechnology in cancer

skos:prefLabel [language: en]
Nanotechnology in cancer

skos:definition [language: en]
Cancer nanotechnology is a branch of nanotechnology concerned with the application of both nanomaterials (such as nanoparticles for tumour imaging or drug delivery) and nanotechnology approaches (such as nanoparticle-based theranostics) to the diagnosis and treatment of

Description: nanotechnology_in_cancer

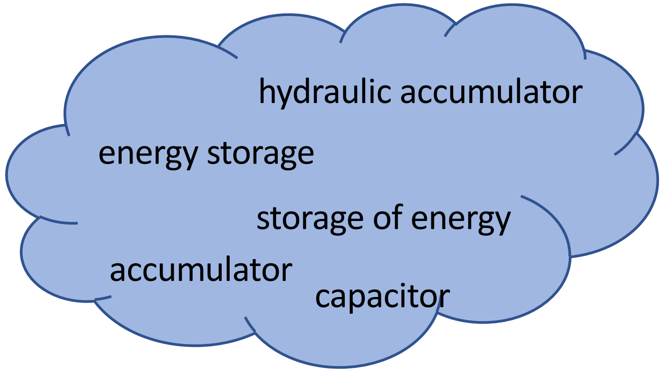
Equivalent To

SubClass Of
nanomedicine

Link related topics

Find more information about the topic

From ontology to data



hydraulic accumulator

energy storage

storage of energy

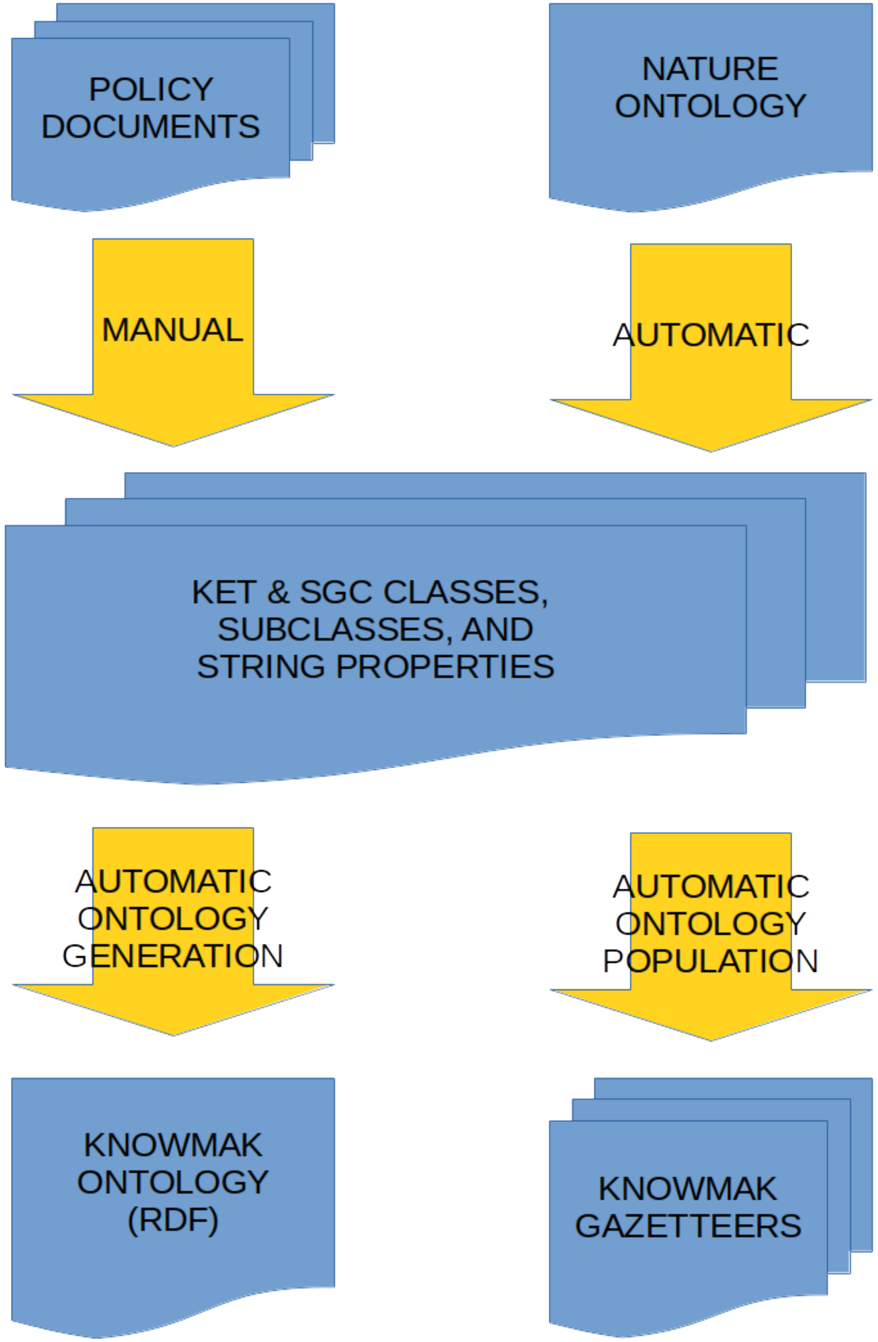
accumulator

capacitor

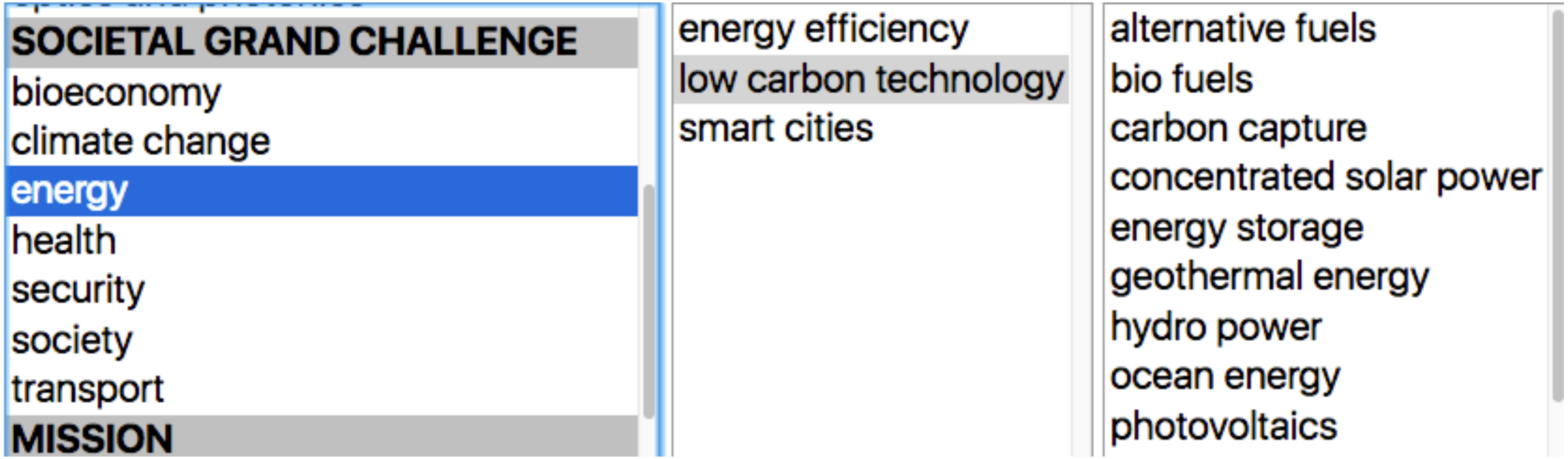
1. Create ontology of topics representing KET and SGC
   - From existing classifications, policy documents, expert users, and data
2. Automatically generate collections of keywords
   - NLP techniques (term extraction, word embeddings) from large training dataset
   - Ranking and scoring algorithms to 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

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


POLICY DOCUMENTS
NATURE ONTOLOGY
MANUAL
AUTOMATIC
KET & SGC CLASSES, SUBCLASSES, AND STRING PROPERTIES
AUTOMATIC ONTOLOGY GENERATION
AUTOMATIC ONTOLOGY POPULATION
KNOWMAK ONTOLOGY (RDF)
KNOWMAK GAZETTEERS

SGC Topics and SubTopics



SOCIETAL GRAND CHALLENGE

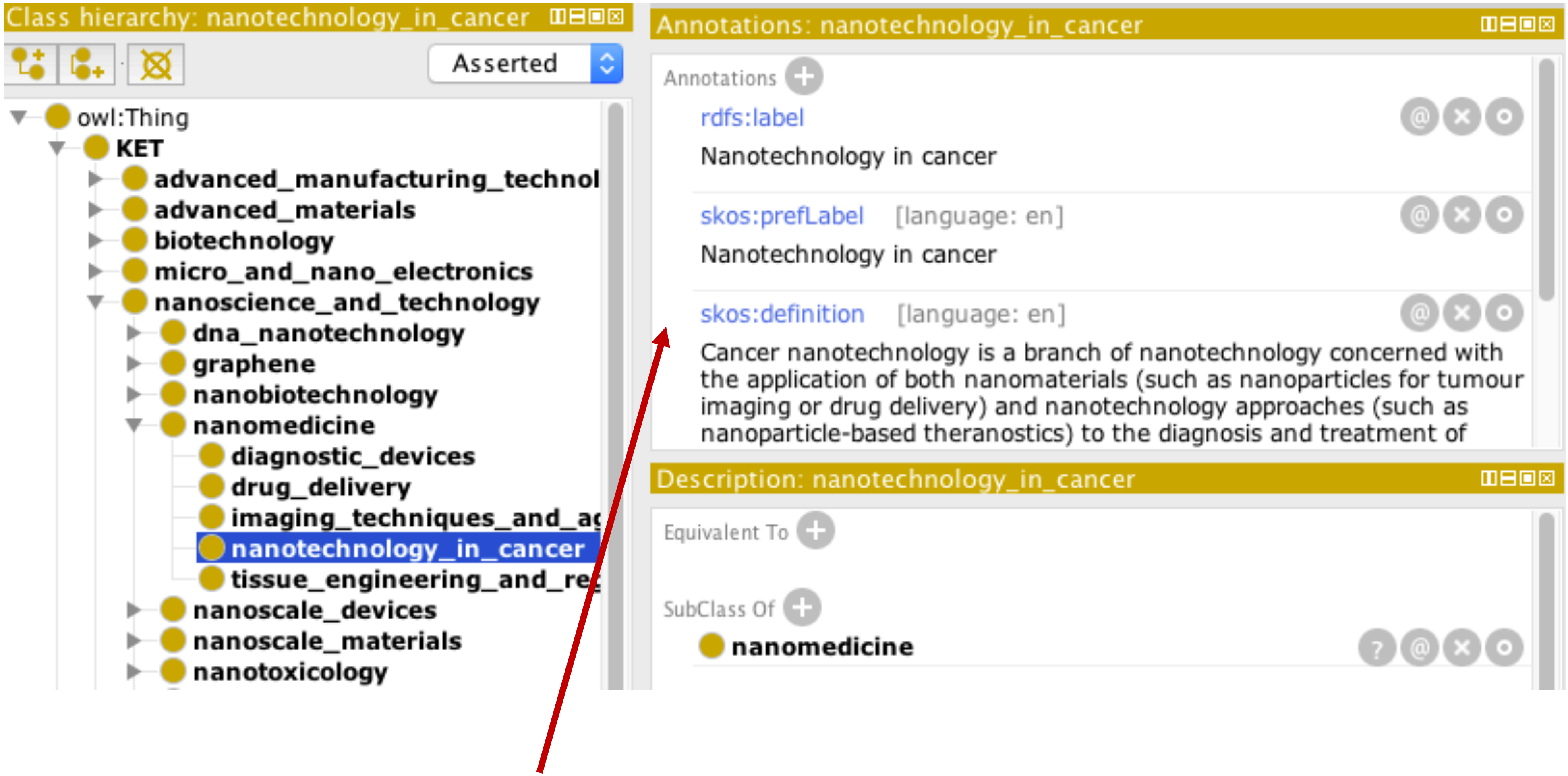
- bioeconomy
- climate change
- energy
- health
- security
- society
- transport

MISSION

- energy efficiency
- low carbon technology
- smart cities

- alternative fuels
- bio fuels
- carbon capture
- concentrated solar power
- energy storage
- geothermal energy
- hydro power
- ocean energy
- photovoltaics

Linking information from external sources



Class hierarchy: nanotechnology_in_cancer

Asserted

- owl:Thing
  - KET
    - advanced_manufacturing_technol
    - advanced_materials
    - biotechnology
    - micro_and_nano_electronics
    - nanoscience_and_technology
      - dna_nanotechnology
      - graphene
      - nanobiotechnology
      - nanomedicine
        - diagnostic_devices
        - drug_delivery
        - imaging_techniques_and_ag
        - nanotechnology_in_cancer
        - tissue_engineering_and_re
      - nanoscale_devices
      - nanoscale_materials
      - nanotoxicology

Annotations: nanotechnology_in_cancer

Annotations

rdfs:label

Nanotechnology in cancer

skos:prefLabel [language: en]

Nanotechnology in cancer

skos:definition [language: en]

Cancer nanotechnology is a branch of nanotechnology concerned with the application of both nanomaterials (such as nanoparticles for tumour imaging or drug delivery) and nanotechnology approaches (such as nanoparticle-based theranostics) to the diagnosis and treatment of

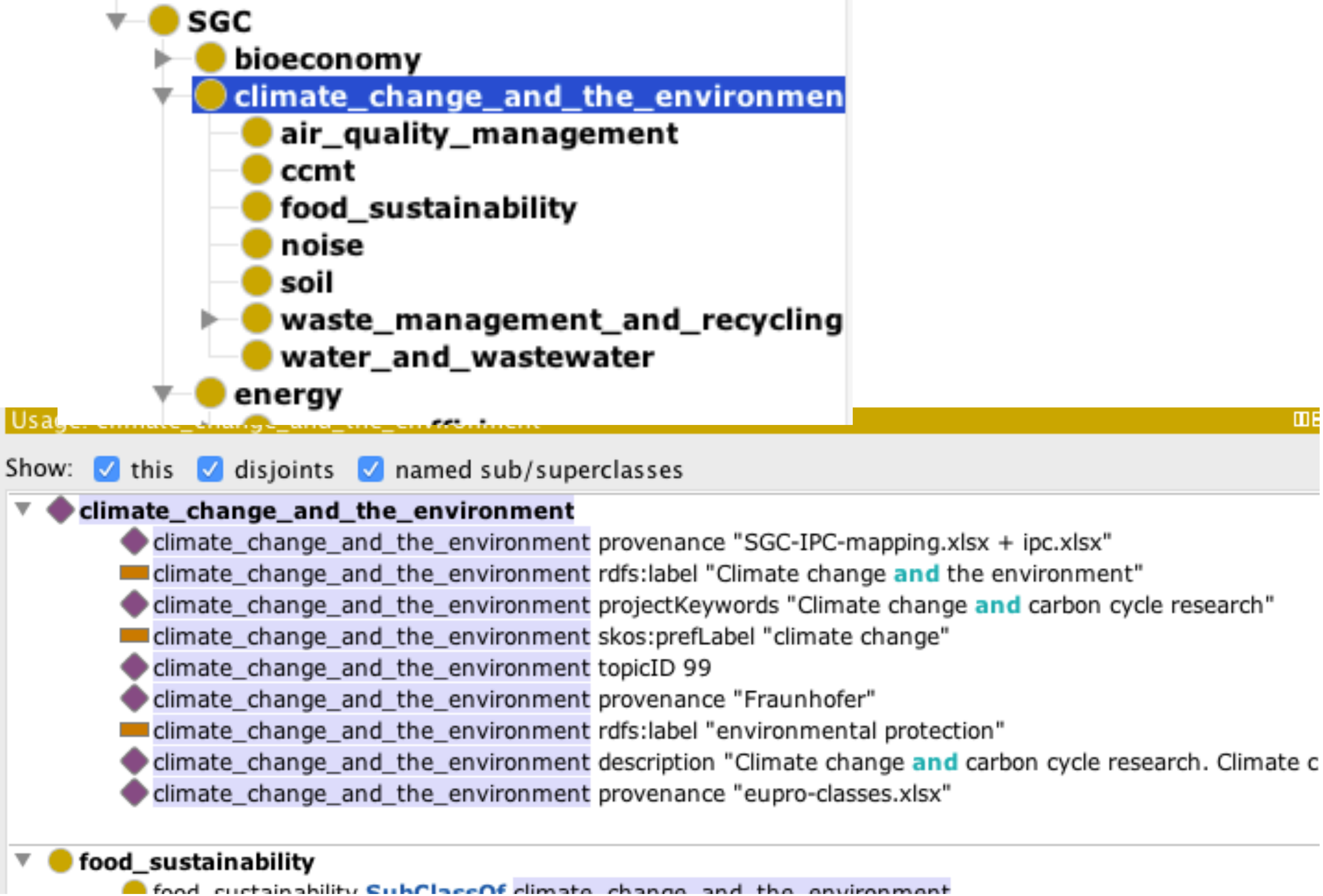
Description: nanotechnology_in_cancer

Equivalent To

SubClass Of

nanomedicine

Link to more information



SGC
bioeconomy
climate_change_and_the_environmen
air_quality_management
ccmt
food_sustainability
noise
soil
waste_management_and_recycling
water_and_wastewater
energy

Usage: climate_change_and_the_environment

Show: ☑ 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

Ontology population

> **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.

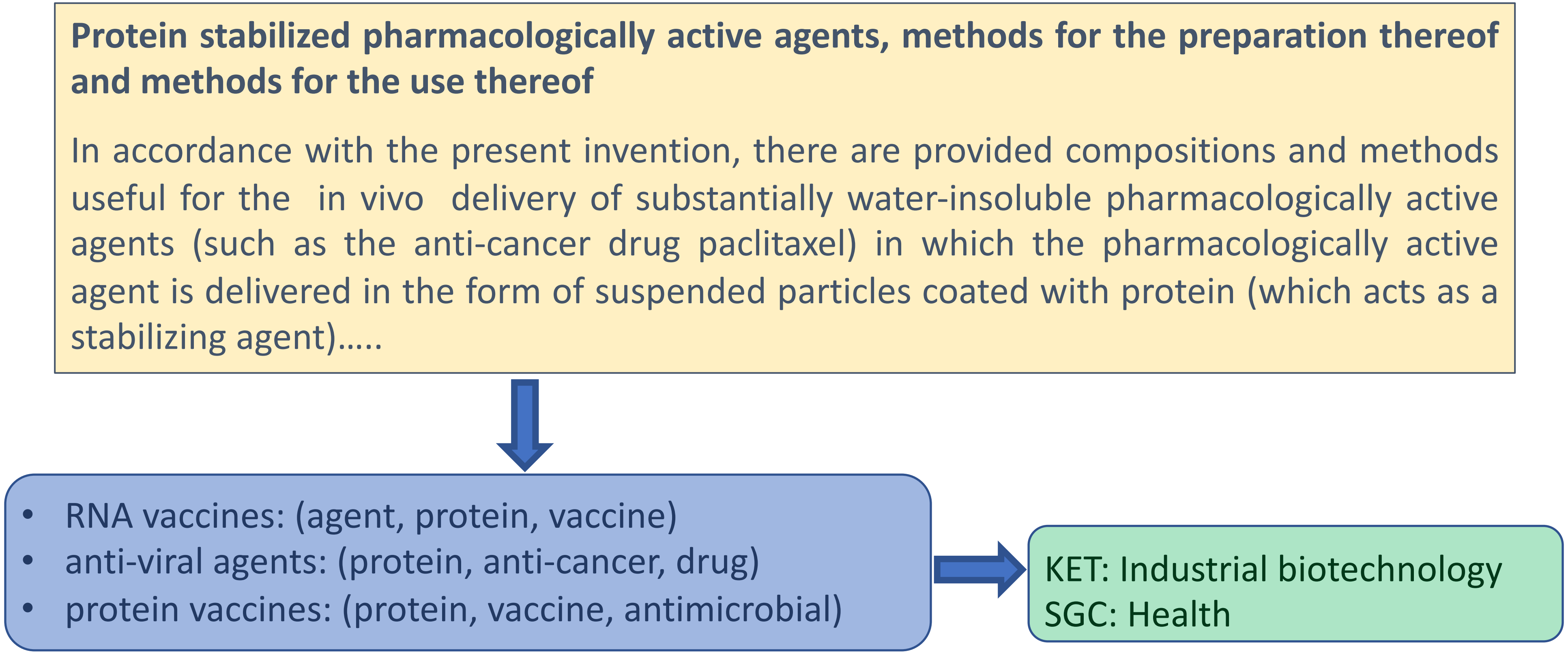
1. Automatically generate keywords from class names, descriptions, and related information (e.g. DBpedia, skos, etc.) using term recognition tools
2. Enrich using word embeddings
3. Score the keywords according to how representative they are of that class
4. Generate prior probabilities using PMI for term combinations, based on frequency of co-occurrence

Annotating Data with Ontologies

- Data sources are annotated against the ontologies
  - each document is associated with one or more topics
- Sophisticated NLP matching and scoring of terms in the documents with ontology
- A REST service accepts documents, scores and 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?

```
{"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 ],
   },
```

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

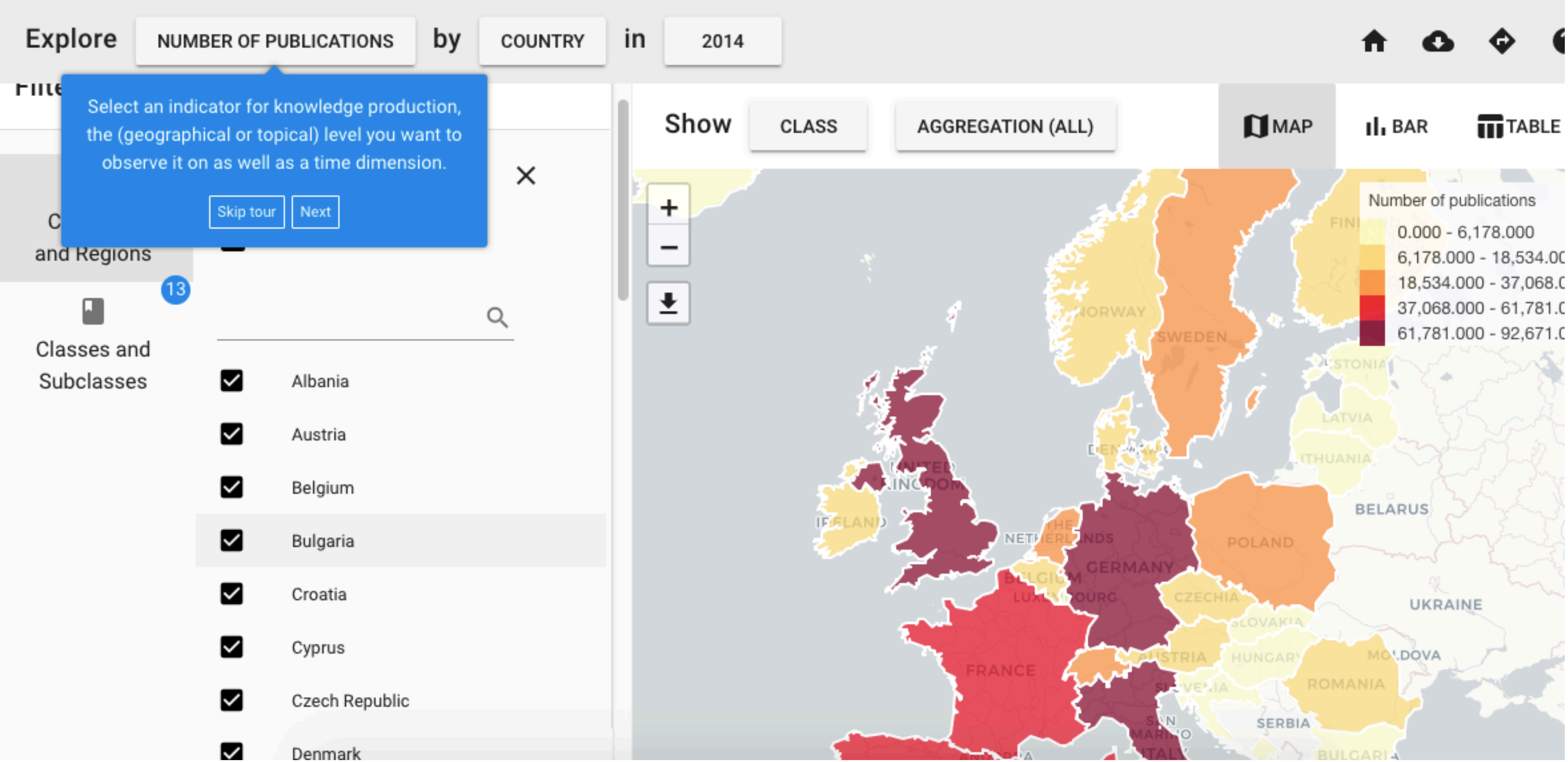
Evaluation

- 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

Try it out! https://www.knowmak.eu/



Explore NUMBER OF PUBLICATIONS by COUNTRY in 2014

Select an indicator for knowledge production, the (geographical or topical) level you want to observe it on as well as a time dimension.

Skip tour Next

☑ Albania
☑ Austria
☑ Belgium
☑ Bulgaria
☑ Croatia
☑ Cyprus
☑ Czech Republic
☑ Denmark

Show CLASS AGGREGATION (ALL)

Number of publications

▨	0.000 - 6,178.000
▨	6,178.000 - 18,534.00
▨	18,534.000 - 37,068.0
▨	37,068.000 - 61,781.0
▨	61,781.000 - 92,671.0

These technologies and ontologies are also being used in the RISIS project as a way to understand and integrate these datasets and many more in science and innovation

https://www.risis2.eu/

ACCESS TO RISIS CORE FACILITY

DATASETS

RISIS project gives access to 13 RISIS datasets

for studying science and innovation