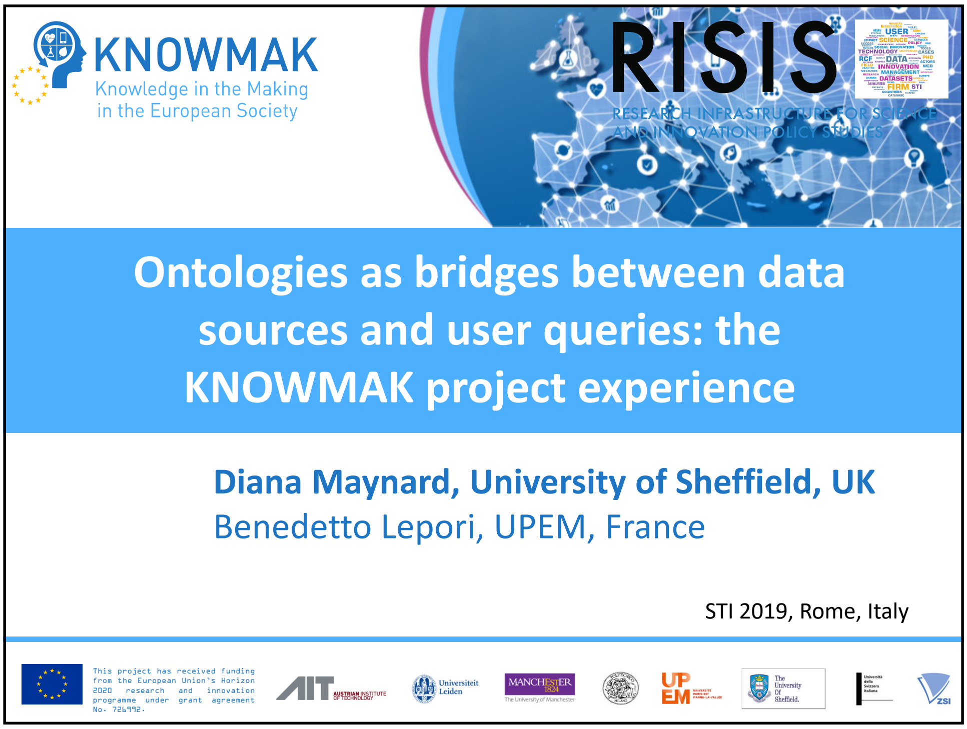


KNOWMAK
Knowledge in the Making
in the European Society

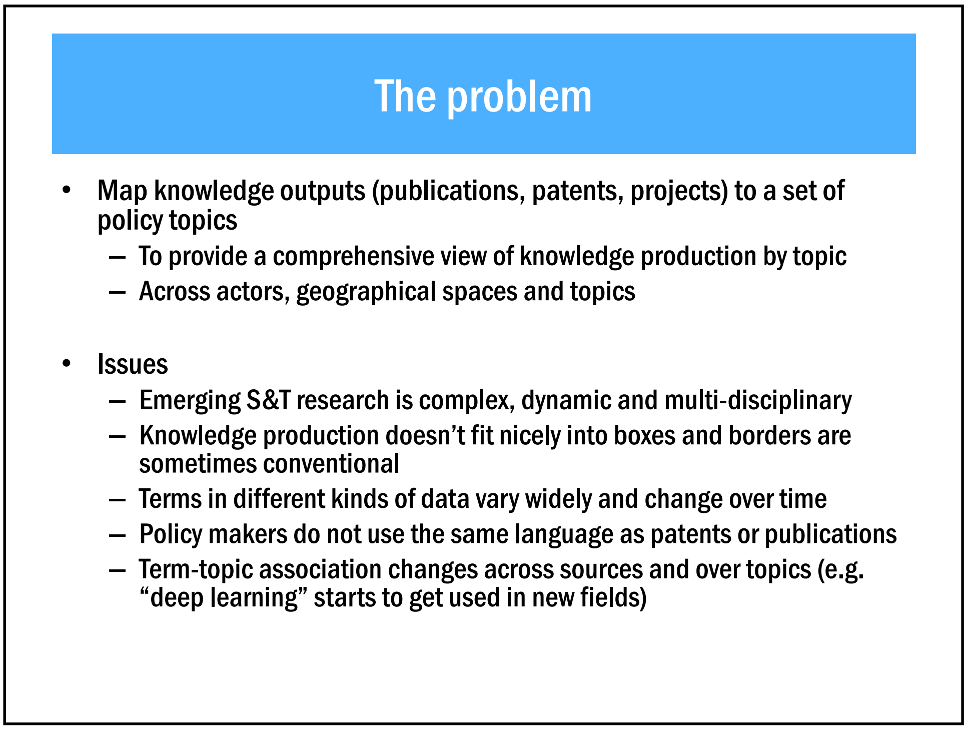
RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

# Ontologies as bridges between data sources and user queries: the KNOWMAK project experience

**Diana Maynard, University of Sheffield, UK**
Benedetto Lepori, UPEM, France

STI 2019, Rome, Italy

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

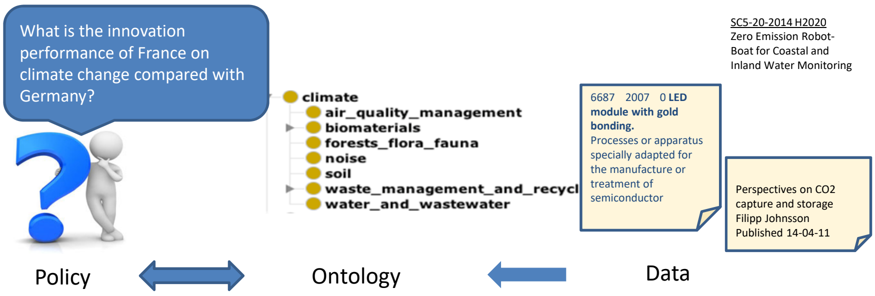


## The problem

- Map knowledge outputs (publications, patents, projects) to a set of policy topics
  - To provide a comprehensive view of knowledge production by topic
  - Across actors, geographical spaces and topics

- Issues
  - Emerging S&T research is complex, dynamic and multi-disciplinary
  - Knowledge production doesn’t fit nicely into boxes and borders are sometimes conventional
  - Terms in different kinds of data vary widely and change over time
  - Policy makers do not use the same language as patents or publications
  - Term-topic association changes across sources and over topics (e.g. “deep learning” starts to get used in new fields)

# Policy questions



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

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

Policy
Ontology
Data

**In a nutshell:**

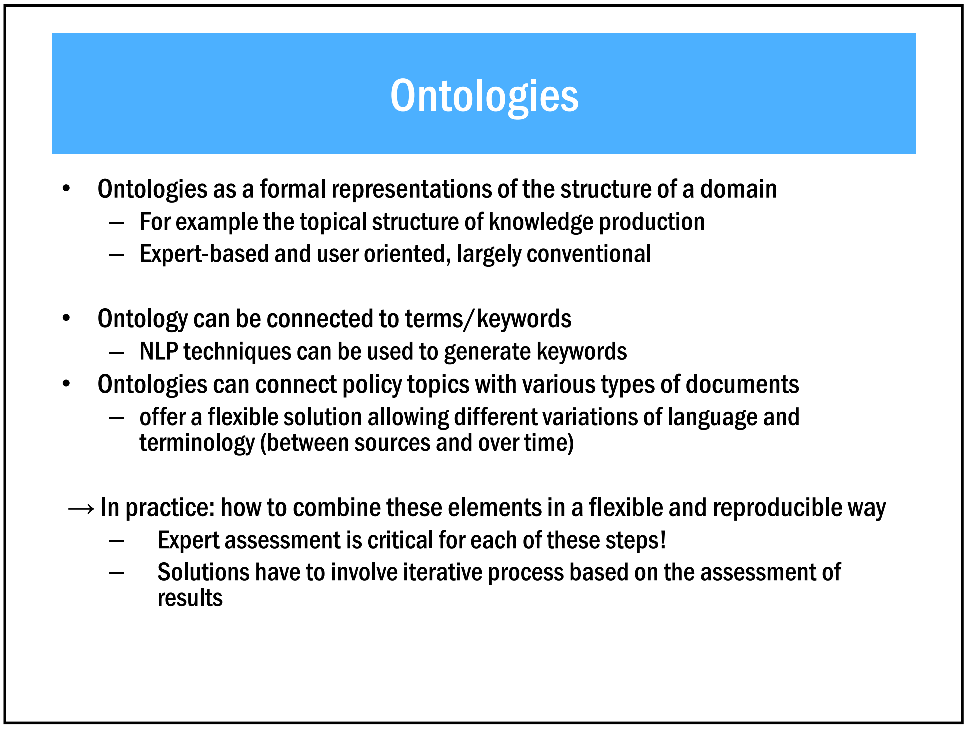
- We need to know which topics each document is talking about (multi-class classification)
- But we have to connect these topics together coherently

*Implemented within the KNOWMAK EU-FP project and further extended under the RISIS EU infrastructure project (risis.eu)*

# Usual approaches

- Link classifications to topics
  - For example IPC classes for patents or Fields of Science
  - Straightforward, but sometimes arbitrary and time-consuming
  - rigid and difficult to employ across different sources / publications vs. patents)
- Use NLP techniques to create maps of knowledge production
  - Very flexible, but sometimes too much detail, link with high-level topics unclear
  - Difficult to connect to policy language and across data sources with very different vocabularies

→ Need for an approach that combines expert-based classification and NLP techniques in a flexible manner

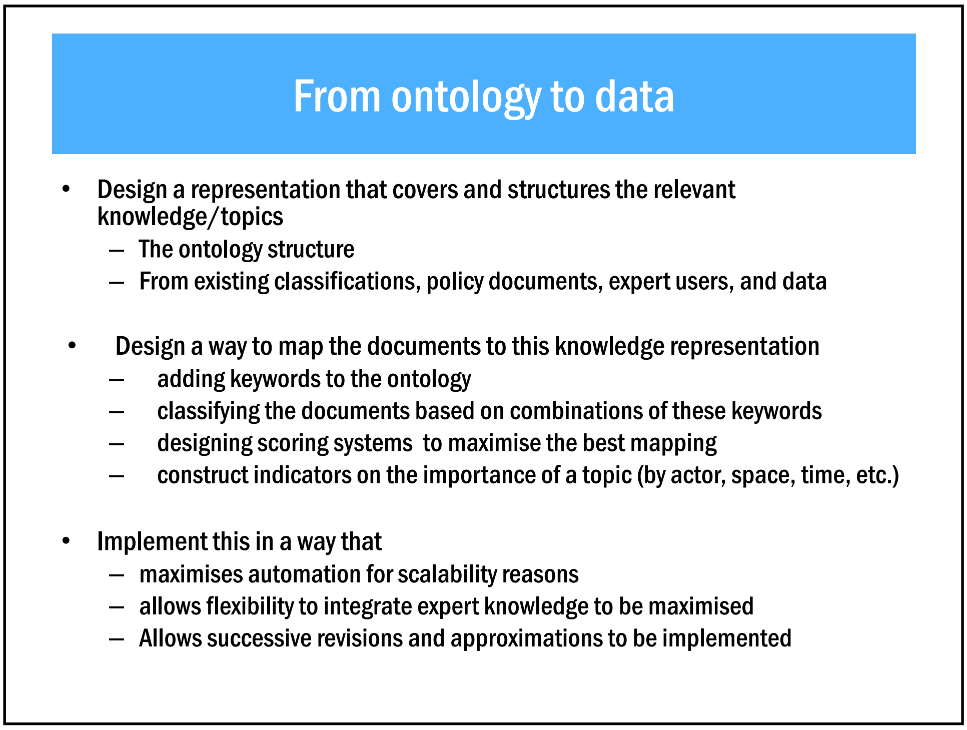


## Ontologies

- Ontologies as a formal representations of the structure of a domain
  - For example the topical structure of knowledge production
  - Expert-based and user oriented, largely conventional
- Ontology can be connected to terms/keywords
  - NLP techniques can be used to generate keywords
- Ontologies can connect policy topics with various types of documents
  - offer a flexible solution allowing different variations of language and terminology (between sources and over time)

→ In practice: how to combine these elements in a flexible and reproducible way

- Expert assessment is critical for each of these steps!
- Solutions have to involve iterative process based on the assessment of results



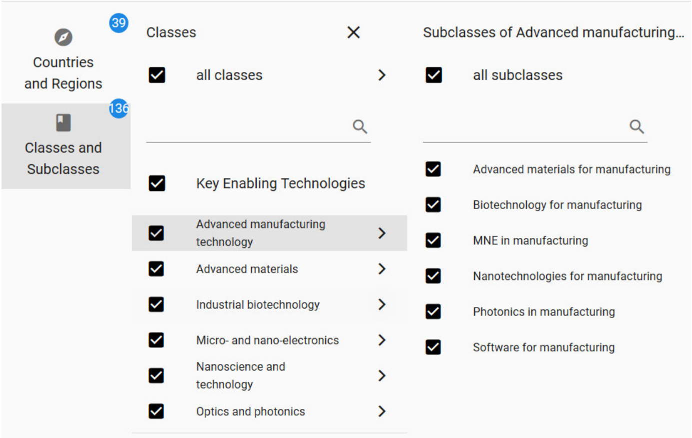
## From ontology to data

- Design a representation that covers and structures the relevant knowledge/topics
  - The ontology structure
  - From existing classifications, policy documents, expert users, and data
- Design a way to map the documents to this knowledge representation
  - adding keywords to the ontology
  - classifying the documents based on combinations of these keywords
  - designing scoring systems to maximise the best mapping
  - construct indicators on the importance of a topic (by actor, space, time, etc.)
- Implement this in a way that
  - maximises automation for scalability reasons
  - allows flexibility to integrate expert knowledge to be maximised
  - Allows successive revisions and approximations to be implemented

# Ontology structure

- There aren't any suitable ontologies already out there
  - The amount of data is too big to build them manually
  - But automated methods are problematic too
- Solution: create the initial structure manually based on existing representations
  - Nature.doc for technology
  - EU policy documents for SGCs
- Reducing the complexity: 2-level ontology, 13 KETs/SGCs and 135 subclasses
- Assessing  the structure
  - Reducing overlap between classes
  - Dropping 'rare' classes
  - Adding classes from data sources (social innovation)
- Expert knowledge is needed for fine-tuning

# Ontology structure



Classes ✕

☑ all classes >

☑ Key Enabling Technologies

☑ Advanced manufacturing technology >

☑ Advanced materials >

☑ Industrial biotechnology >

☑ Micro- and nano-electronics >

☑ Nanoscience and technology >

☑ Optics and photonics >

Subclasses of Advanced manufacturing...

☑ all subclasses

☑ Advanced materials for manufacturing

☑ Biotechnology for manufacturing

☑ MNE in manufacturing

☑ Nanotechnologies for manufacturing

☑ Photonics in manufacturing

☑ Software for manufacturing

http://demos.gate.ac.uk/knowmak/faceted-search/

http://demos.gate.ac.uk/knowmak/filter-search/

## Ontology population

- Source data comprises policy documents, topic descriptions, links to other knowledge sources etc.
  - For example the IPC patent vocabulary
- Apply NLP tools
  - Generate lists of terms associated with each class (gazetteers)
- Linguistic variants: more sophisticated NLP
  - “Similar” terms: word embeddings, additional info sources (DBpedia, terminologies, policy documents)

## Pitfalls and issues

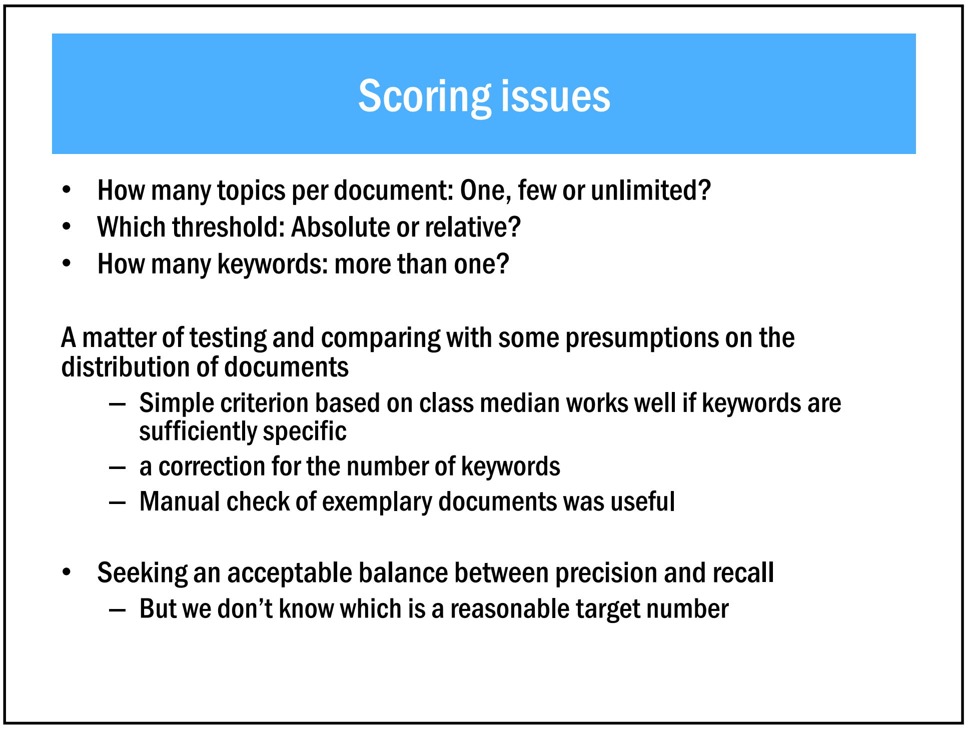
- Overlaps between classes lead to issues with automatic keywords generation
  - Manual cleaning needed
- Generic keywords creep in through automatic generation
  - List of stopwords critical and manual check ex-post from the scoring
  - But very specific keywords might lead to low recall if these are too few
- Very unequal number of keywords by class
  - Need to take into account in the scoring

## Keyword occurrences in patents

| | | |
|---|---|---|
| macromolecular | 58’772.00 | dna nantechnology |
| scaffold | 16’373.00 | dna nantechnology |
| dna | 15’226.00 | dna nantechnology |
| rna | 13’101.00 | dna nantechnology |
| macromolecule | 10’751.00 | dna nantechnology |
| surface | 345’986.00 | nanobiotechnology |
| interface | 49’989.00 | nanobiotechnology |
| concentration | 37’631.00 | nanobiotechnology |
| molecule | 30’961.00 | nanobiotechnology |
| array | 22’340.00 | nanobiotechnology |
| assay | 14’160.00 | nanobiotechnology |
| microorganism | 4’088.00 | nanobiotechnology |
| neural | 2’513.00 | nanobiotechnology |

## Annotating Data with Ontologies

- Data sources are annotated against the ontologies
  - each document is associated with one or more topics
- NLP matching of keywords in the documents (from titles, abstracts etc) with ontology
- Based on linguistic pre-processing, term recognition, frequency and some weighting mechanisms
- Higher priority (weights) allocated to a topic for that document if:
  - Multi-word term (vs single-word term)
  - It belongs to a more specific ontology class
  - It comes from a particular trusted source (e.g. IPC patent codes)
  - more matching terms associated with that class



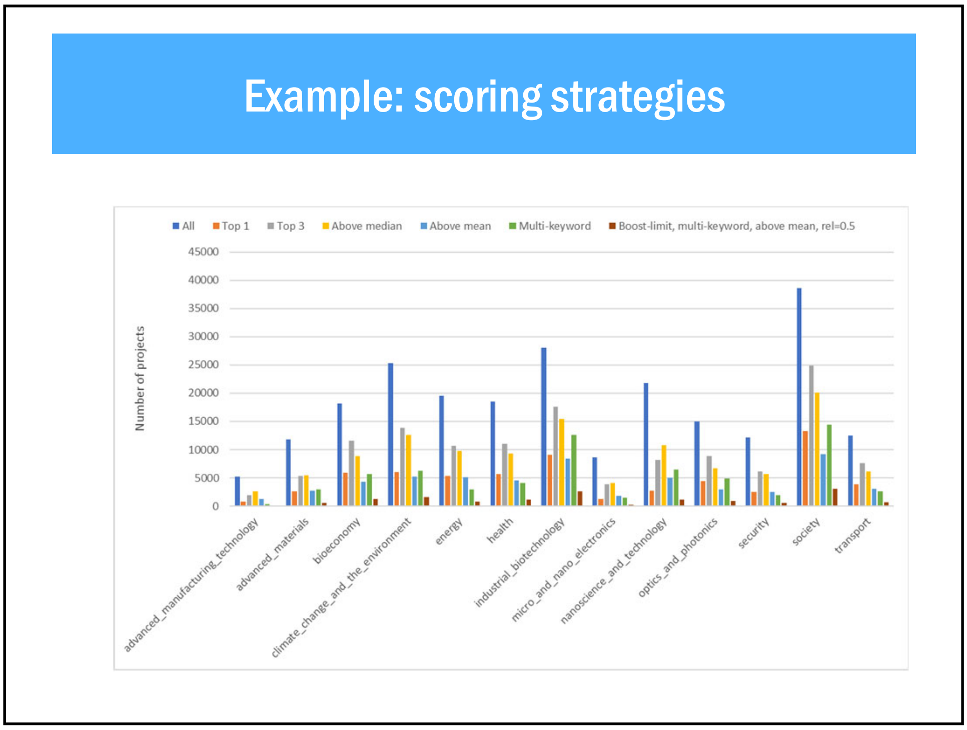
# Scoring issues

- How many topics per document: One, few or unlimited?
- Which threshold: Absolute or relative?
- How many keywords: more than one?

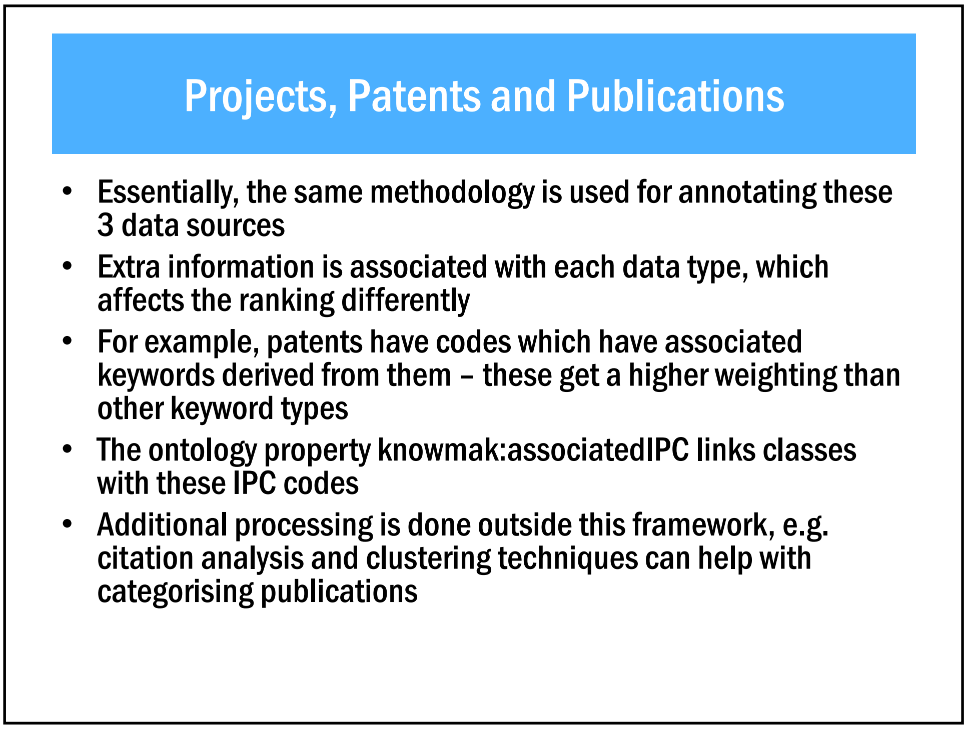
A matter of testing and comparing with some presumptions on the distribution of documents

- – Simple criterion based on class median works well if keywords are sufficiently specific
- – a correction for the number of keywords
- – Manual check of exemplary documents was useful

- Seeking an acceptable balance between precision and recall
  - – But we don't know which is a reasonable target number

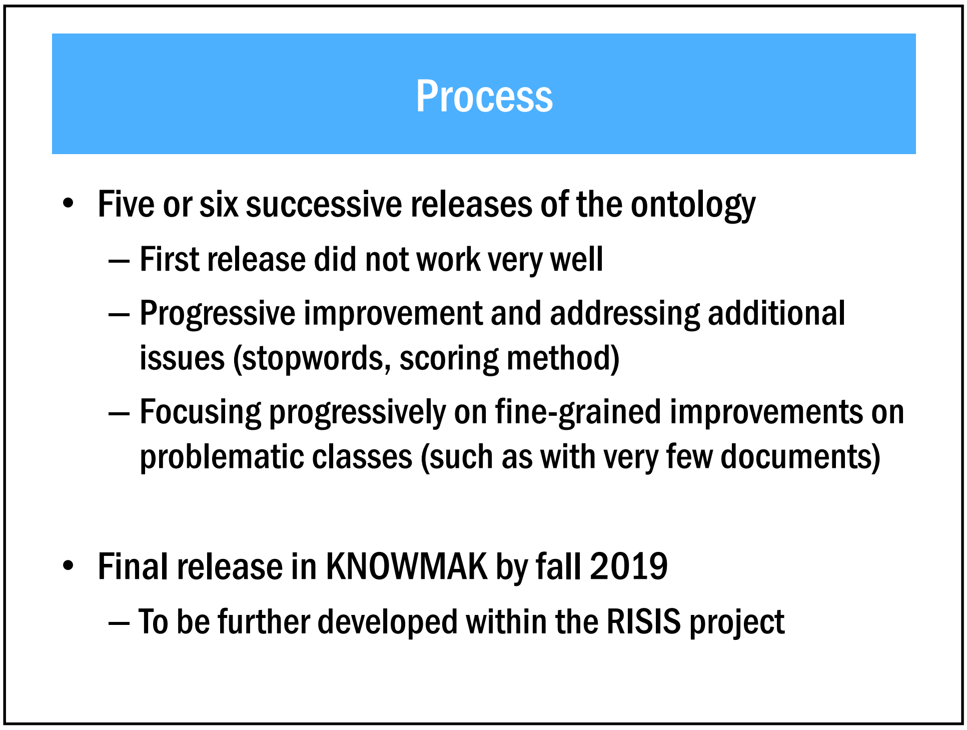


# Example: scoring strategies



## Projects, Patents and Publications

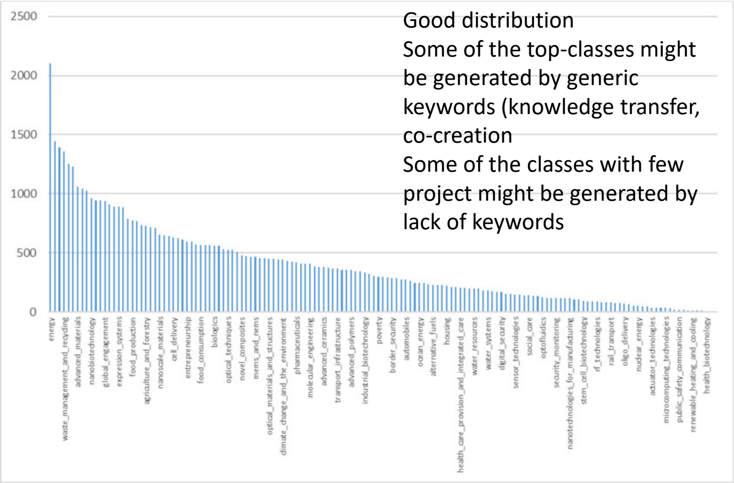
- Essentially, the same methodology is used for annotating these 3 data sources
- Extra information is associated with each data type, which affects the ranking differently
- For example, patents have codes which have associated keywords derived from them – these get a higher weighting than other keyword types
- The ontology property knowmak:associatedIPC links classes with these IPC codes
- Additional processing is done outside this framework, e.g. citation analysis and clustering techniques can help with categorising publications



## Process

- Five or six successive releases of the ontology
  - First release did not work very well
  - Progressive improvement and addressing additional issues (stopwords, scoring method)
  - Focusing progressively on fine-grained improvements on problematic classes (such as with very few documents)
- Final release in KNOWMAK by fall 2019
  - To be further developed within the RISIS project

## Current distribution by classes for EU-FP projects

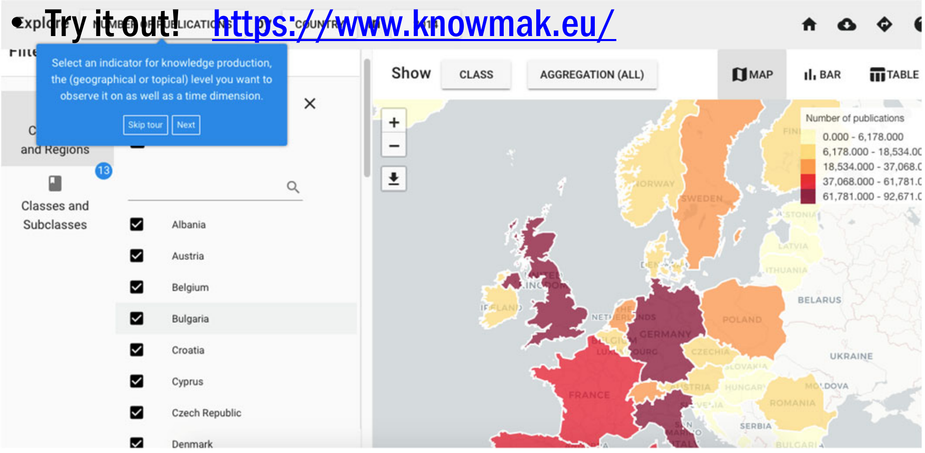


Good distribution
Some of the top-classes might be generated by generic keywords (knowledge transfer, co-creation
Some of the classes with few project might be generated by lack of keywords

## Work ahead in RISIS2

- Systematic evaluation of precision and recall
  - By scoring manually a sufficiently large number of documents
  - Needed for validity of the ontology
- Investigate better methods for automatically generating topic keywords that are both relevant and representative of a class
  - factoring in negative as well as positive feedback mechanisms
  - Adjustment of similarity thresholds and investigation of scoring metrics based on gold standard data
- Potential incorporation of new topics with minimal human expert intervention for a more sustainable approach long-term

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


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

KNOWMAK
Knowledge in the Making
in the European Society

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES
USER
SCIENCE
TECHNOLOGY
DATA
INNOVATION
MANAGEMENT
DATASETS
FIRM
STI

## THANK YOU FOR LISTENING!

Main project website
Sheffield's KNOWMAK work
RISIS project
GATE tools

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

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