KNOWMAK

Knowledge in the Making
in the European Society

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

Diana Maynard, University of Sheffield, UK

Benedetto Lepori, UPEM, France

STI 2017, Paris, France

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

Aims of the KNOWMAK project

- Develop a web-based tool providing interactive visualizations and indicators on knowledge co-creation in the European research area
- Based around:
  - **Research Actors** (organisations)
  - **Research topics** (based on SGC, KET)
  - **Geographical spaces** (based on NUTS and FUA)

Potential user queries

- What kinds of research topic does a region specialise in?
- Who are the main actors on a particular topic in a particular region?
- How are they connected?
- How diversified is a region’s knowledge base?
- What is the innovation performance of a region compared to other regions?
- How diversified is a region’s knowledge base?

Connecting the queries with the data

- We need to connect the user queries with the data sources (projects, patents, publications)
- But the users are policy makers who use different kinds of language and terminology
- How do we build indicators on the data that can help answer the user queries?

The problem

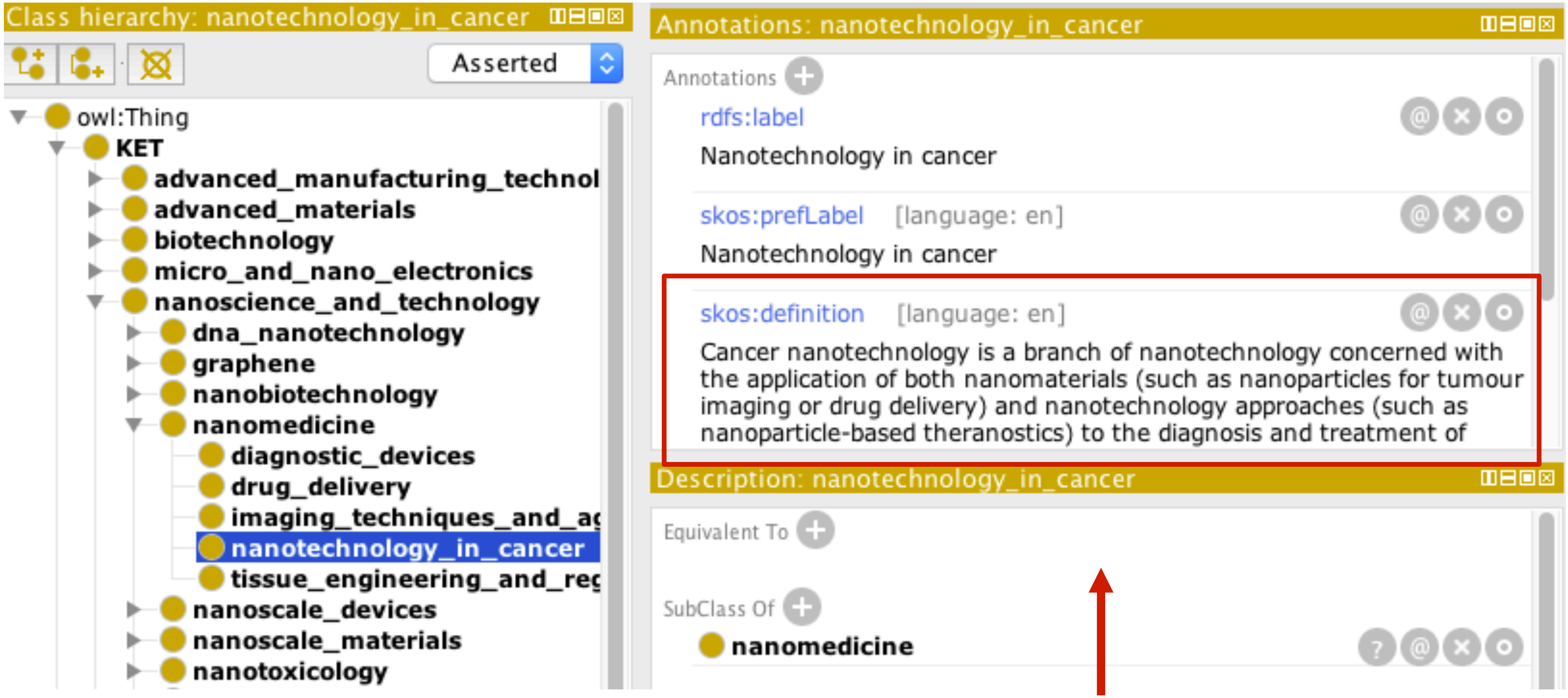
- Traditional STIs are too rigid, and difficult to use for policy decisions
- Emerging S&T research is complex, dynamic and multi-disciplinary
- Knowledge production doesn’t fit nicely into boxes
- Terms in different kinds of data vary widely
  - Policy makers do not use the same language as patents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

Our approach: ontologies

- Ontologies enable mapping between user queries, indicators and topics
- Built around the KETs and SGCs
- Handle user searching by topic / keywords
- Allow user exploration of knowledge around topics
- Enable creation of indicators around topics
- Act as a bridge between user queries and information in the databases
- Ontologies offer a flexible solution allowing different variations of language and terminology

Ontologies connect information

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



- 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_ac
        - nanotechnology_in_cancer
        - tissue_engineering_and_reg
      - nanoscale_devices
      - nanoscale_materials
      - nanotoxicology

rdfs:label
Nanotechnology in cancer

skos:prefLabel [language: en]
Nanotechnology in cancer

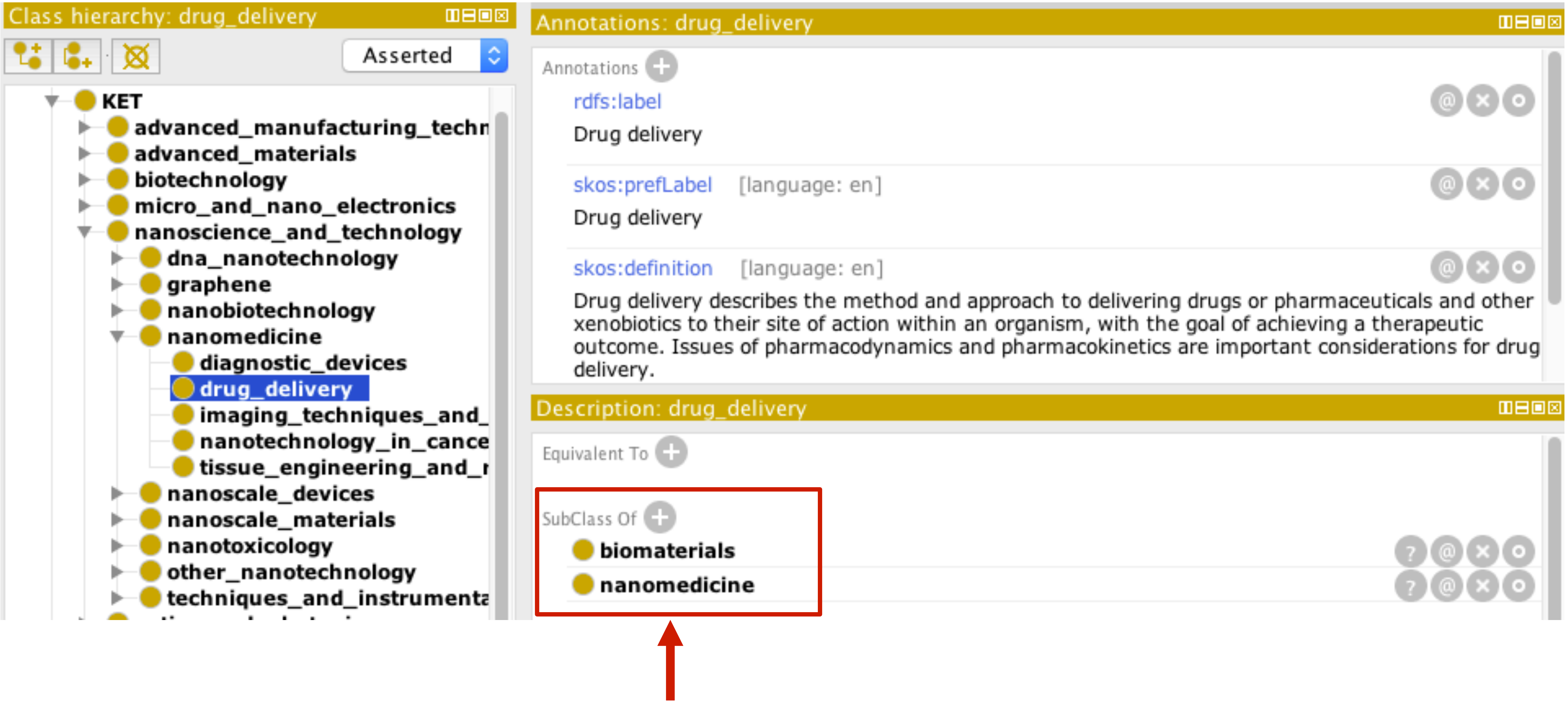
skos:definition [language: en]
Cancer nanotechnology is a branch of nanotechnology concerned with the application of both nanomaterials (such as nanoparticles for tumour imaging or drug delivery) and nanotechnology approaches (such as nanoparticle-based theranostics) to the diagnosis and treatment of

nanomedicine

Link related topics

Find more information about the topic

Topics can belong to multiple classes



Class hierarchy: drug_delivery

Asserted

- KET
  - advanced_manufacturing_techn
  - advanced_materials
  - biotechnology
  - micro_and_nano_electronics
  - nanoscience_and_technology
    - dna_nanotechnology
    - graphene
    - nanobiotechnology
    - nanomedicine
      - diagnostic_devices
      - drug_delivery
      - imaging_techniques_and_
      - nanotechnology_in_cance
      - tissue_engineering_and_r
    - nanoscale_devices
    - nanoscale_materials
    - nanotoxicology
    - other_nanotechnology
    - techniques_and_instrumenta

Annotations: drug_delivery

Annotations

rdfs:label
Drug delivery

skos:prefLabel [language: en]
Drug delivery

skos:definition [language: en]
Drug delivery describes the method and approach to delivering drugs or pharmaceuticals and other xenobiotics to their site of action within an organism, with the goal of achieving a therapeutic outcome. Issues of pharmacodynamics and pharmacokinetics are important considerations for drug delivery.

Description: drug_delivery

Equivalent To

SubClass Of
- biomaterials
- nanomedicine

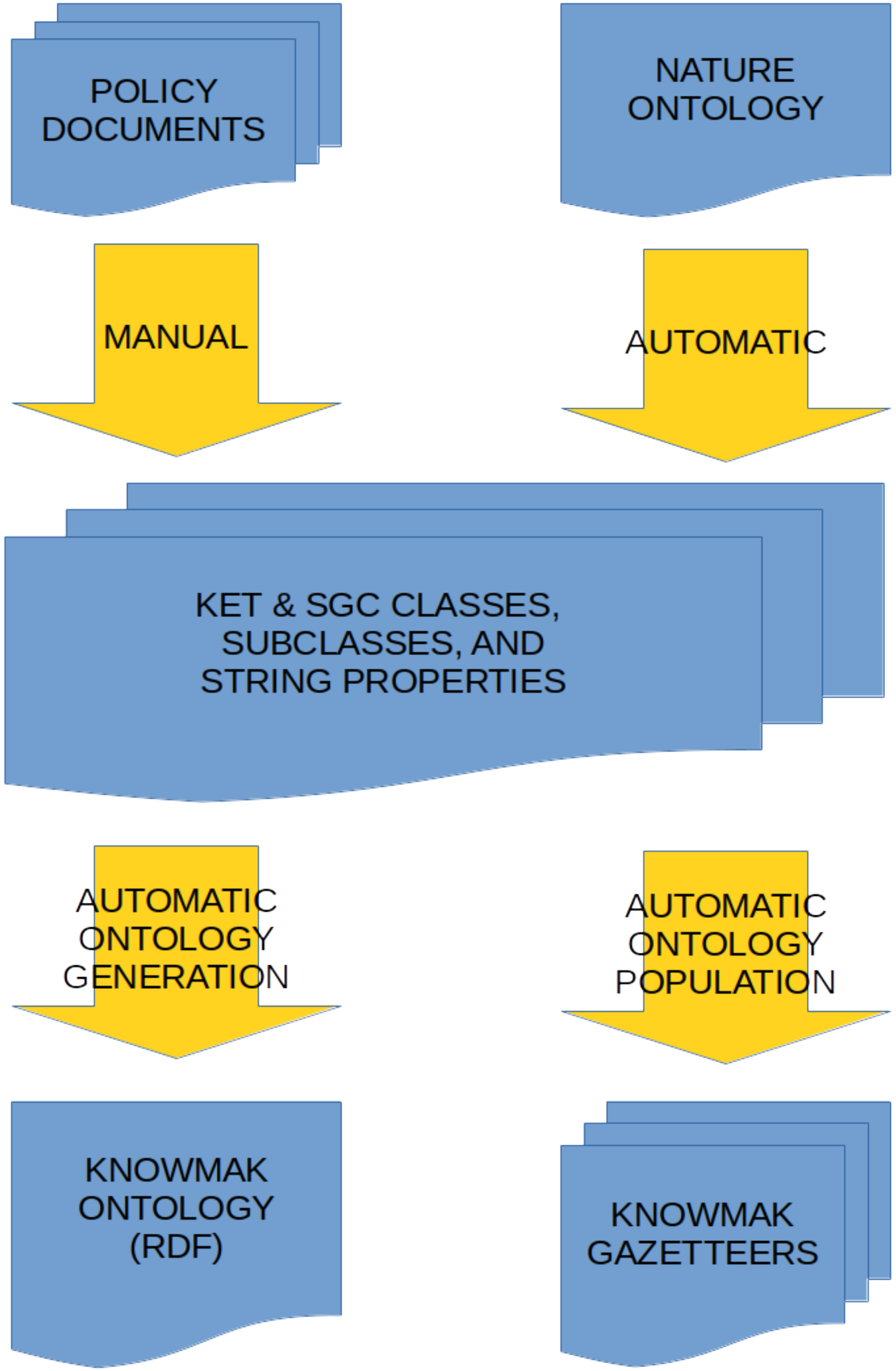
We can now look at both biomaterials and nanomedicine to find related information

But how on earth do we build suitable ontologies?

- 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
  - not very reliable
  - we might miss lots of topics depending on our source data
  - we can’t easily represent different variations of the same term
  - terms change over time and between data sources
- **Solution**: create the initial structure manually based on existing representations where possible, and populate automatically
- For the linguistic processing, we use GATE, an open source infrastructure for NLP developed at Sheffield

Creating and populating the KNOWMAK 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)



POLICY DOCUMENTS

NATURE ONTOLOGY

MANUAL

AUTOMATIC

KET & SGC CLASSES, SUBCLASSES, AND STRING PROPERTIES

AUTOMATIC ONTOLOGY GENERATION

AUTOMATIC ONTOLOGY POPULATION

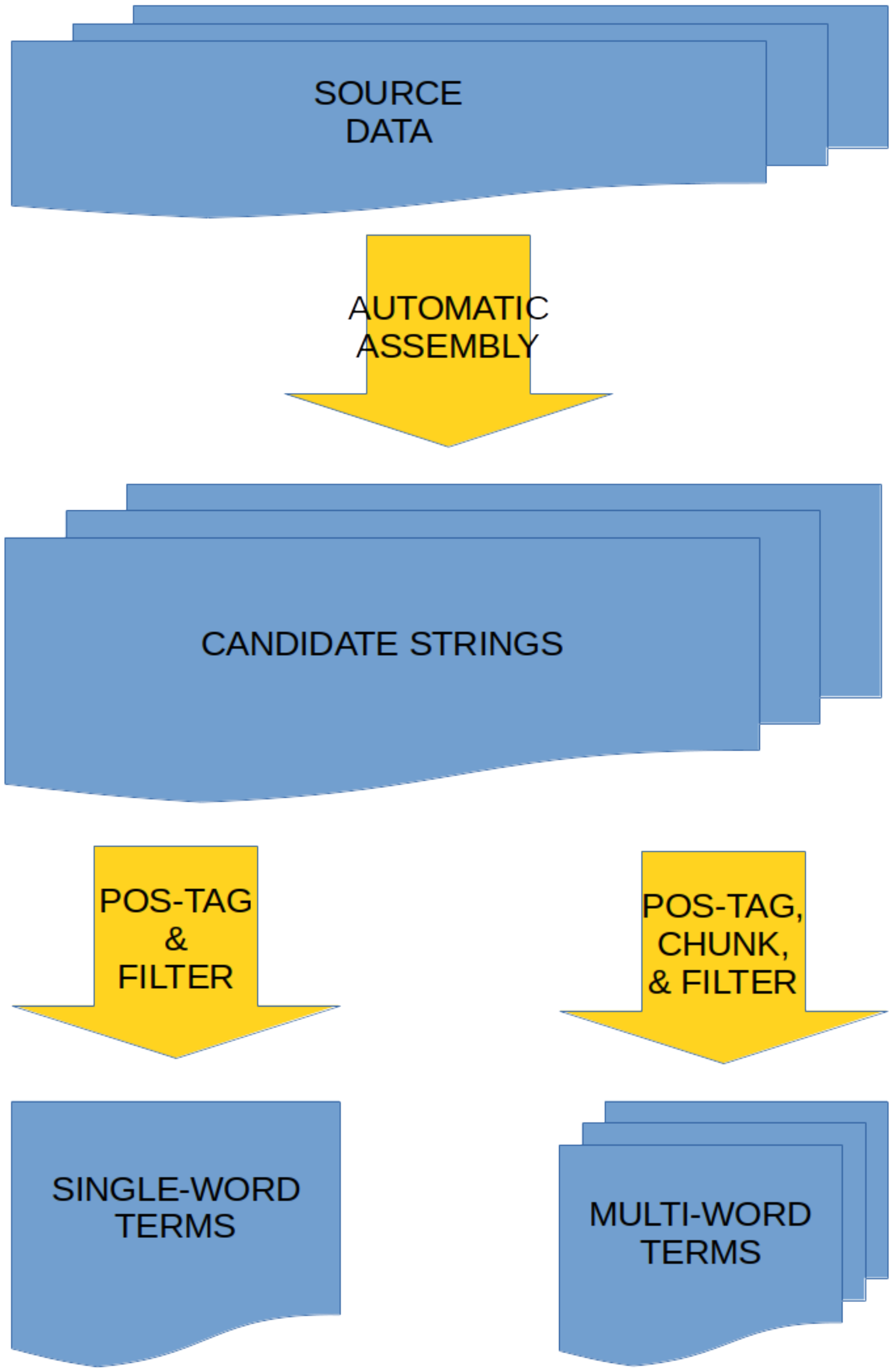
KNOWMAK ONTOLOGY (RDF)

KNOWMAK GAZETTEERS

Ontology population

1. Source data comprises policy documents, topic descriptions, links to other knowledge sources etc.
2. Apply NLP tools
3. Generate lists of terms associated with each class (gazetteers)

Fully automatic



SOURCE DATA

AUTOMATIC ASSEMBLY

CANDIDATE STRINGS

POS-TAG & FILTER

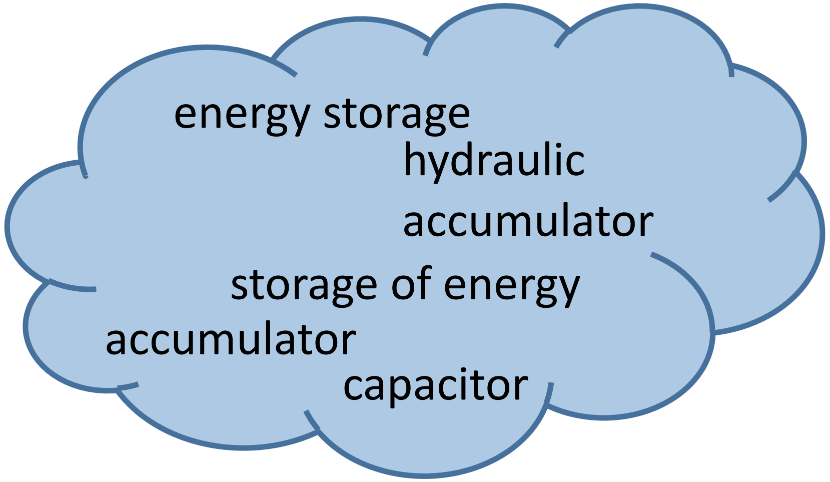
POS-TAG, CHUNK, & FILTER

SINGLE-WORD TERMS

MULTI-WORD TERMS

Extending the ontology population

- First version doesn’t have that many instances
- We need to find variants of existing terms



energy storage
hydraulic accumulator
storage of energy
accumulator
capacitor

- Linguistic variants: more sophisticated NLP
- “Similar” terms: word embeddings, additional info sources (DBpedia, terminologies, policy documents)

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
- 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
- Annotated data sources are then used to build indicators
  - e.g. for each topic, how many publications and in which region?

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

Annotation of a project document

Project ID: 51797
Program Type: FP5-LIFE QUALITY
Project name: Extracting products of high added value from vegetal species of the mediterranean basin using non-organic solvents

Project objective: The extraction of products derived from vegetal species used in the food, pharmaceutical, and cosmetic industries are heavily dependent on the use of organic solvents such as hexane and dichloromethane….The alternative technology proposed would obtain high quality natural products using non-toxic solvents and is based on the capacity of supercritical fluids, and mainly CO2 to dissolve natural products in a very selective form based on precise combination of pressure and temperature. The development of this technology coupled with the careful choice of the raw materials used (organically grow plants) would result in the production of extracts with minimal alterations in the colours, scent and flavour and free form toxic residues to the benefit of consumers and the mentioned industries.

Classes: advanced_manufacturing_technology (7.18); optofluidics (4.97); advanced materials (4.97)

Summary

- 3-year project, started in January 2017
- Ontologies and topics are the core of the system, but probably the hardest to develop
- Major issues:
  - sufficient coverage of ontology population
  - how to map between different language terminologies
  - term ambiguity and variation
- Continuous process of development and testing with real users
- Evaluation of ontologies is tricky – we concentrate mainly on functionality (does it enable us to perform the task well?)
- High-risk but highly exciting!

KNOWMAK

Knowledge in the Making
in the European Society

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.