KNOWMAK
Knowledge in the Making in the European Society

Creating a large topic ontology for policymakers

Dr. Diana Maynard

University of Sheffield, UK

London, 29 Sept 2017

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

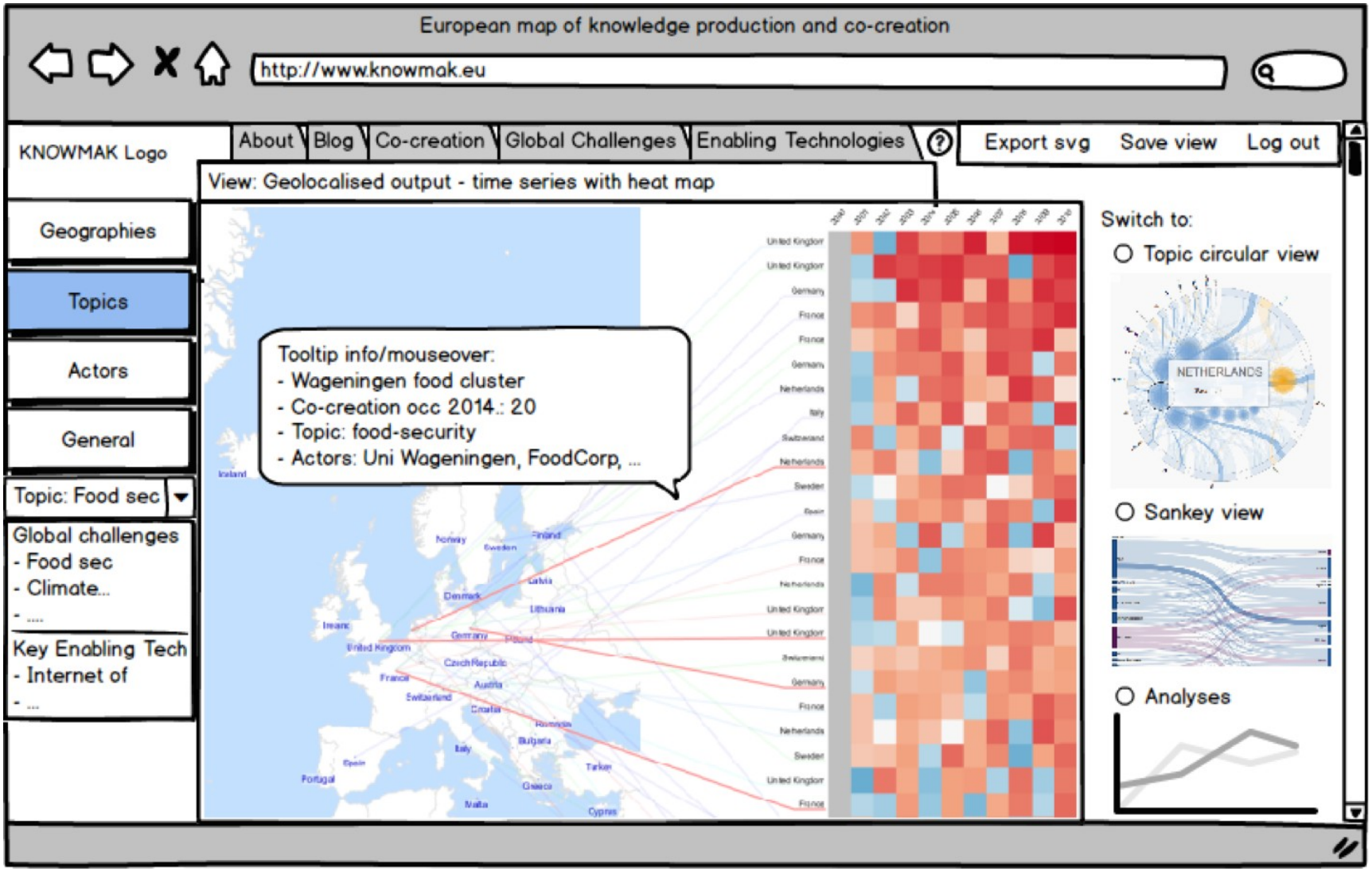
Aims of the KNOWMAK project

- 3-year EU H2020 project since January 2017
- 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?
- What is the innovation performance of a region compared to other regions?
- How diversified is a region's knowledge base?

Sample user interface



Connecting the queries with the data

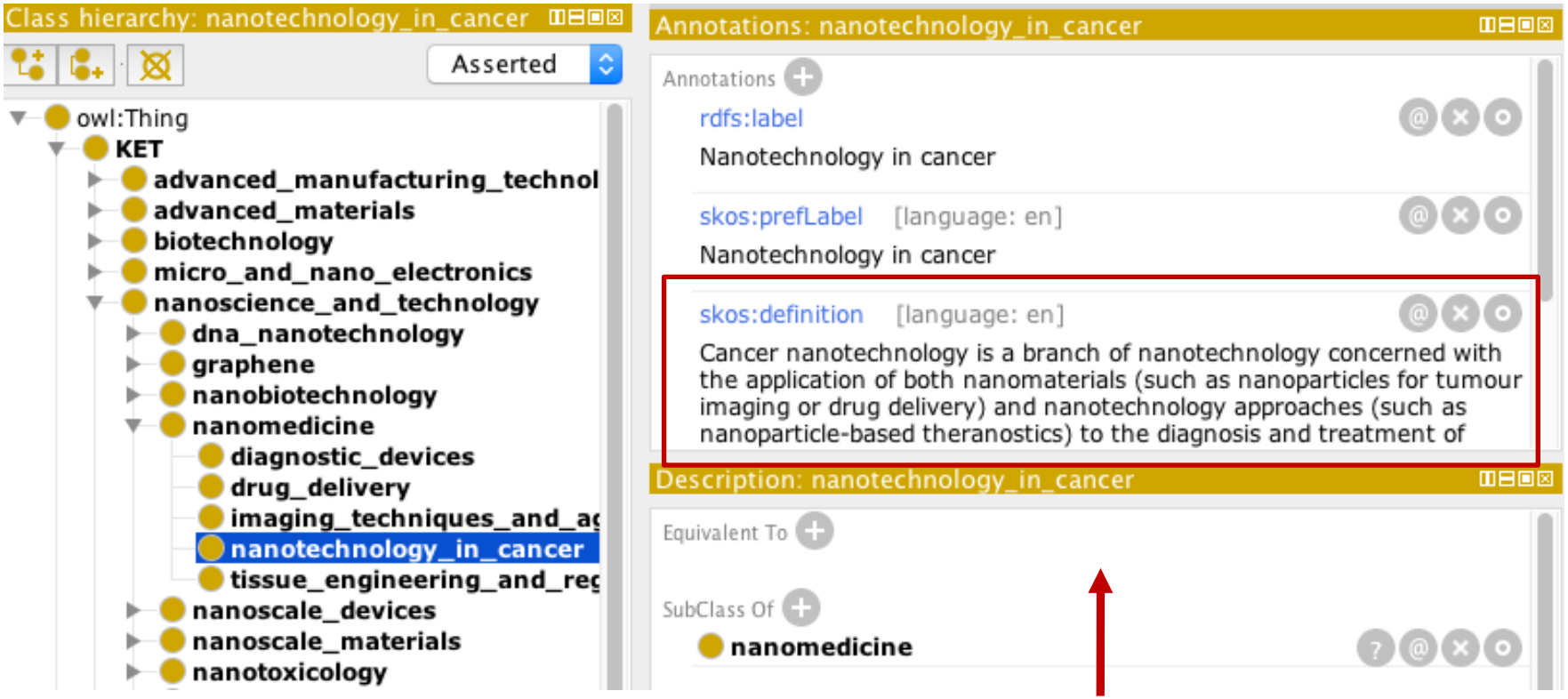
- We need to connect the user queries with the data sources (projects, patents, publications)
- But the users are policy makers who use different kinds of language and terminology
- How do we build indicators on the data that can help answer the user queries?

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

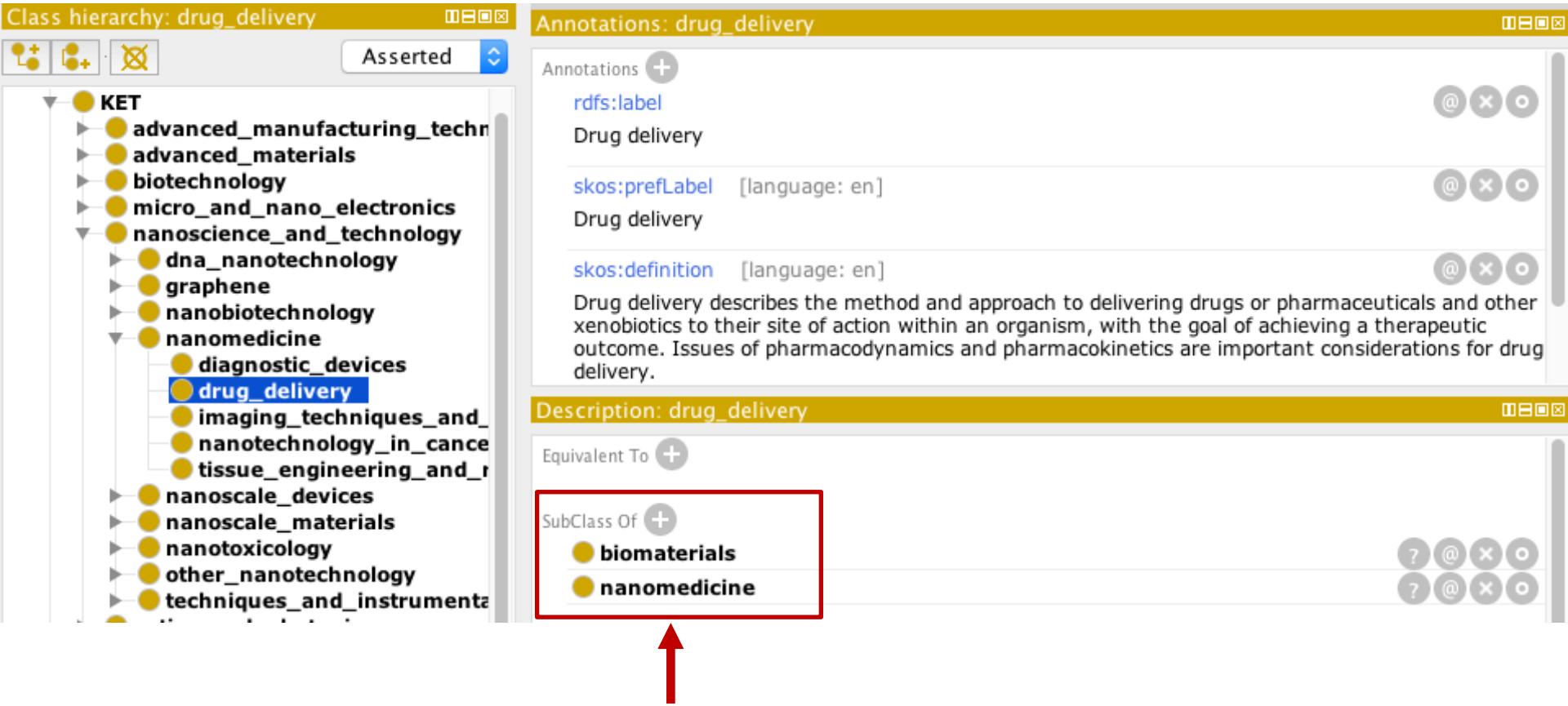
Link with information from other sources
(Nature.com, skos, DBpedia…)



Link related topics

Find more information about the topic

Topics can belong to multiple classes



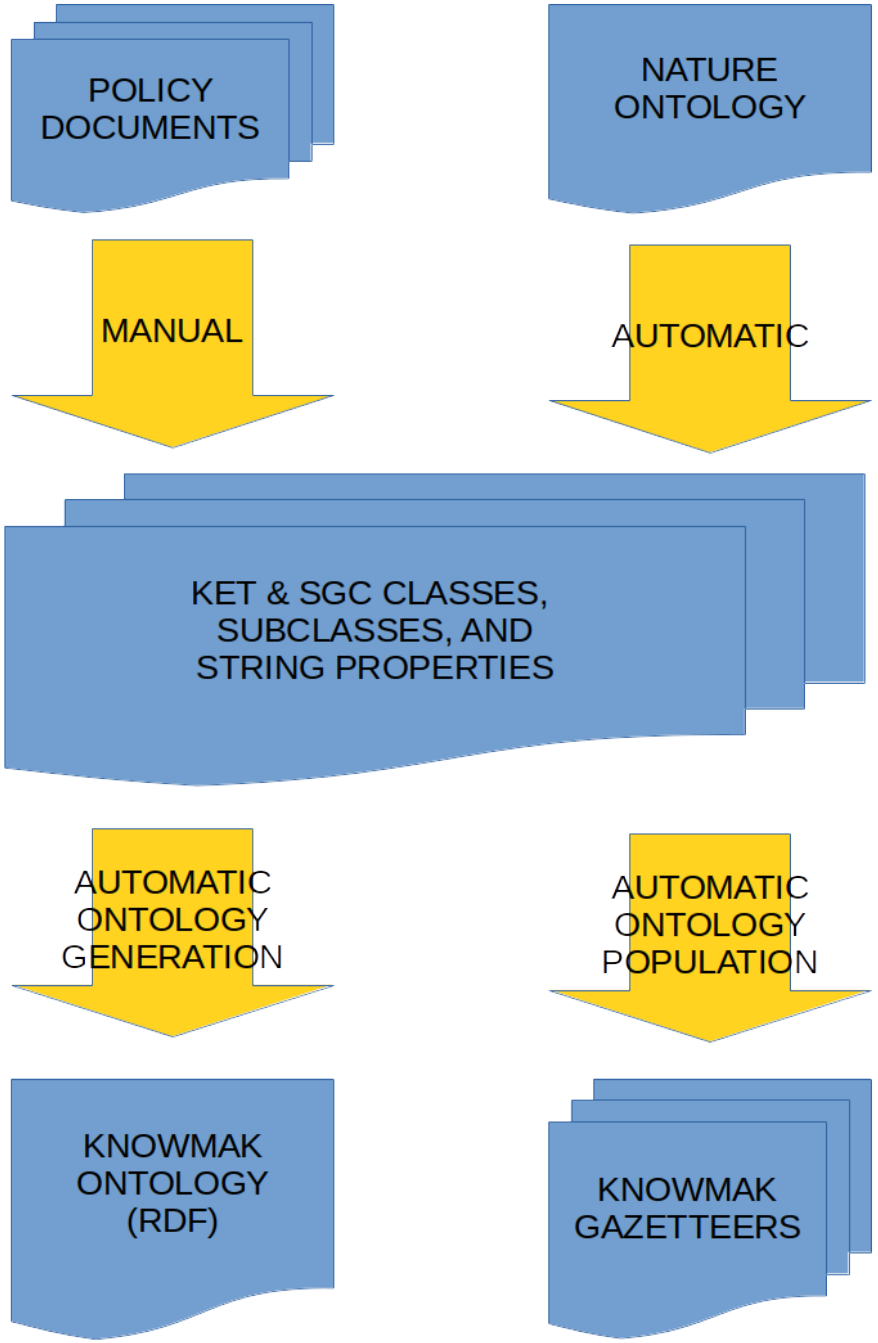
We can now look at both biomaterials and nanomedicine to find related information

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 term variation
  - 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

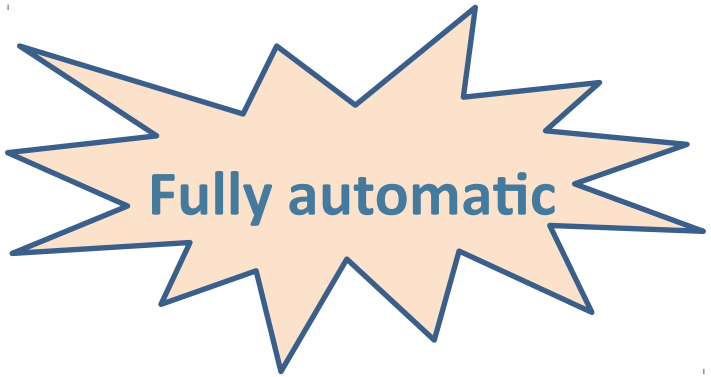
Creating and populating the KNOWMAK ontology

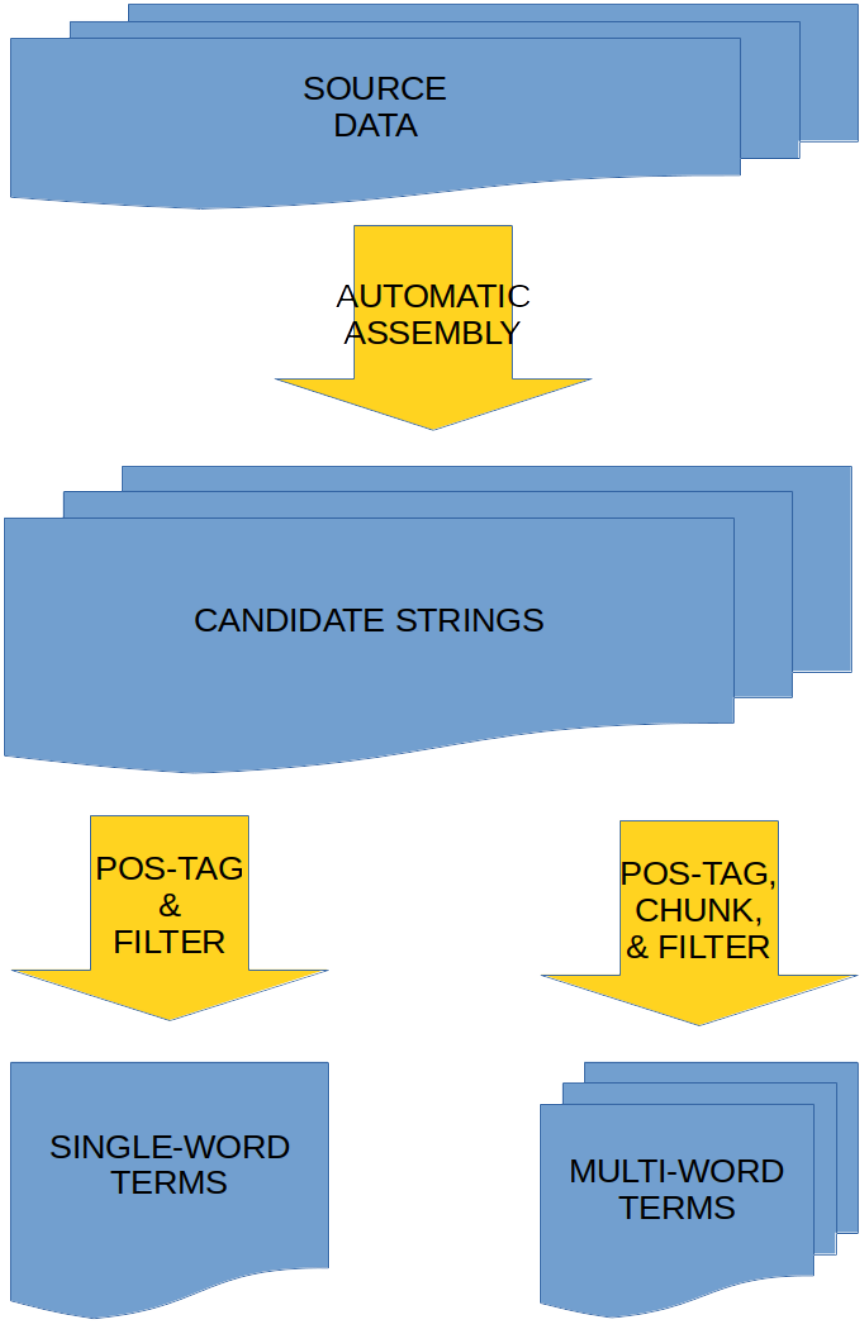
1. Create ontology structure (classes & subclasses)
2. Add extra information (descriptions, links, alternate class names)
3. Ontology population: generate lists of terms associated with each class (gazetteers)



Ontology population

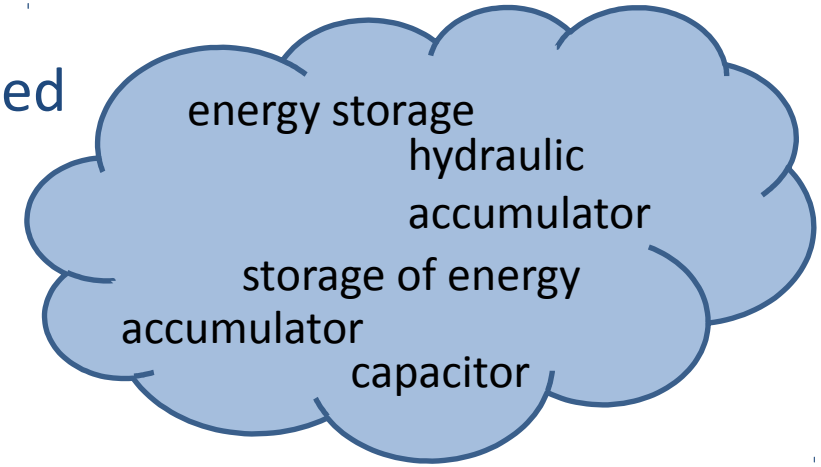
1. Source data comprises policy documents, topic descriptions, links to other knowledge sources etc.
2. Apply NLP tools
3. Generate lists of terms associated with each class (gazetteers)





Extending the ontology population

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



- Use DBpedia abstracts as extra sources of relevant terms:
  - For known DBpedia URIs, fetch abstracts
  - For unknown URIs, calculate semantic similarity:
    - generate possible list of matching URIs
    - calculate best matching abstract (using entity linking techniques from YODIE)

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

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

- 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: we need your help!
- Evaluation of ontologies is tricky – we concentrate mainly on functionality (does it enable us to perform the task well?)

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.