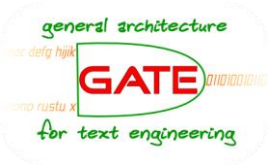
Semantic Technologies for Understanding Knowledge Co-Creation in European Research

Diana Maynard

University of Sheffield, UK



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 topics** (SGC, KET)
  - **Research Actors** (organisations)
  - **Geographical spaces** (Locations)
  - at varying levels of granularity
- 5 data sources:
  - Established: **publications**, **patents**, **projects**
  - New: **social innovation projects** and user attention based on **social media**.

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

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)

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?

The problem

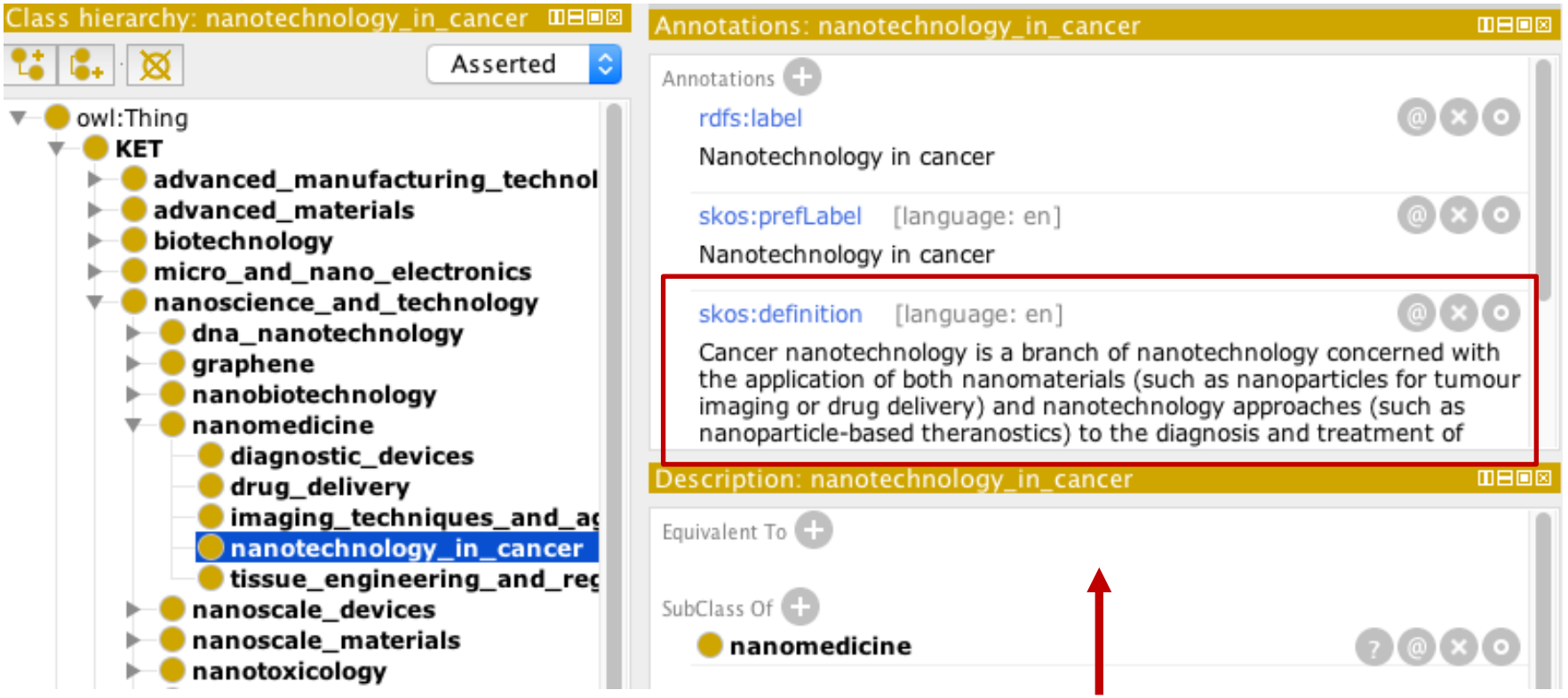
- Traditional STIs are too rigid, and difficult to use for policy decisions
- Emerging S&T research is complex, dynamic and multi-disciplinary
- Knowledge production doesn't fit nicely into boxes
- Terms in different kinds of data vary widely
  - Policy makers do not use the same language as patents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

The solution: ontologies

- Ontologies enable mapping between user queries, indicators and topics
- 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 connect information

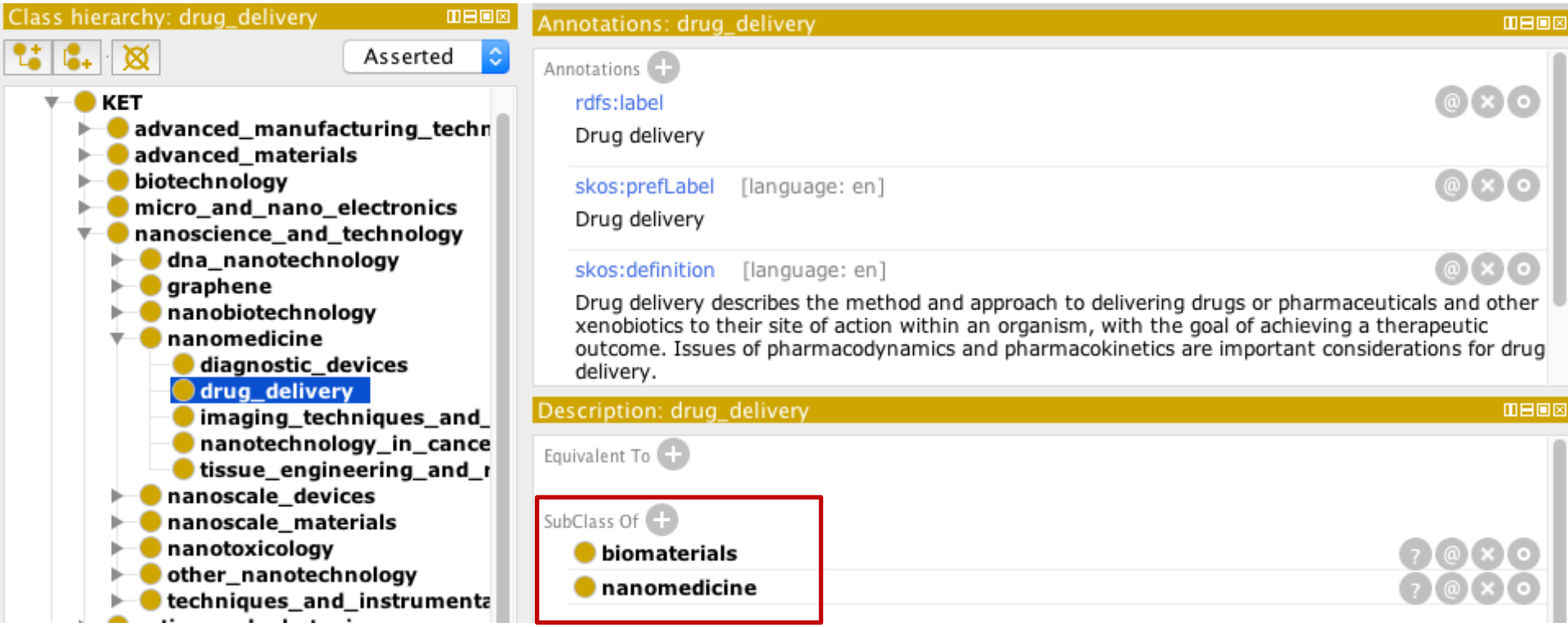
Link with information from other sources
(Nature.com, skos, DBpedia…)



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

KNOWMAK Filter Search

This is a simple example of how a filtered search using the KNOWMAK ontology might look.

Match all drug cancer

Nanotechnology in cancer

Selected Class: http://www.gate.ac.uk/ns/ontologies/knowmak/nanotechnology_in_cancer

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 cancer. Nanotechnology in cancer.

Related Keywords: application, approach, branch, cancer, concerned, delivery, diagnosis, drug, imaging, nanomaterials, nanoparticle-based, nanoparticles, nanotechnology, such, theranostics, treatment, tumour

KNOWMAK Faceted Search

This is a simple example of how a faceted search using the KNOWMAK ontology might look.

KEY EMERGENT TECHOLOGY	Applied Immunology	Antagomir and rna sponge
Advanced Manufacturing Technology	Assay systems	Antibody fragment therapy
Advanced Materials	Biologics	Antibody therapy
Biotechnology	Biomaterials	Antisense oligonucleotide therapy
Micro and Nano electronics	Biomimetics	Cell therapies
Nanoscience and technology	Cell delivery	Dna vaccines
Optics and photonics	Environmental biotechnology	Gene therapy
SOCIETAL GRAND CHALLENGE	Expression systems	Locked nucleic acid
Bioeconomy	Functional genomics	Meganucleases
Climate	Gene delivery	Nucleic acid therapeutics

Selected Class: http://www.gate.ac.uk/ns/ontologies/knowmak/dna_vaccines

Dna vaccines. A DNA vaccine is a substance that is composed of deoxyribonucleic acid (DNA) and encodes antigens. After administration of the DNA, antigens are produced and stimulate an immune response. DNA vaccines.

Related Keywords: acid, administration, antigen, composed, deoxyribonucleic, dna, immune, produced, response, substance, vaccine

Ontology Design

- Built around the KET and SGC
- Topics based on existing principled classifications
  - KET/SGC subclasses in policy documents
  - Nature.com ontology
- Linked to the primary data sources (patents, publications, projects)
  - Mappings to topics in patents/publications/projects
- Keywords associated with topics
  - created from a combination of policy documents and primary data sources

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
- Multi-word terms are more important than single-word terms, e.g. “vapor deposition” is more useful than “vapor”
- Annotated data sources are then used to build indicators
  - e.g. for each topic, how many publications are there and in which region?

Example: Project Abstract and Topics

I propose to investigate a new research frontier on spin physics at the boundaries (surfaces) of materials with strong spin-orbit interaction (SOI). Although the properties of these materials have been studied for more than half a century, researchers are just starting to grasp the richness of SOI phenomena that occur at them. SOI leads to surface and boundary states with unusually large spin splitting in simple heavy elements. It can also produce a nontrivial topology in band insulators that brings about metallic surface states with exotic spin textures that are protected by time reversal symmetry.
I plan to use our cutting-edge expertise on all-electrical lateral spin injection and detection methods to unravel the spin dynamics in them, providing a wealth of information that could not be otherwise obtained. A comprehensive set of objectives will include material integration with ferromagnets and insulators, and innovative devices and measurement protocols.

- Optical data storage
- Biotechnology
- Advanced materials

Summary

- Project only started 6 months ago, so very early stages
- Ontologies and topics are the core of the system, but the hardest to develop
- Many problems with ontology population, annotation, ambiguity, and different use of language in different data sources
- Continuing development of ontologies and annotation methodologies
  - more sophisticated techniques
  - integrating with other knowledge sources
- High-risk but highly exciting!

More information

- 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.