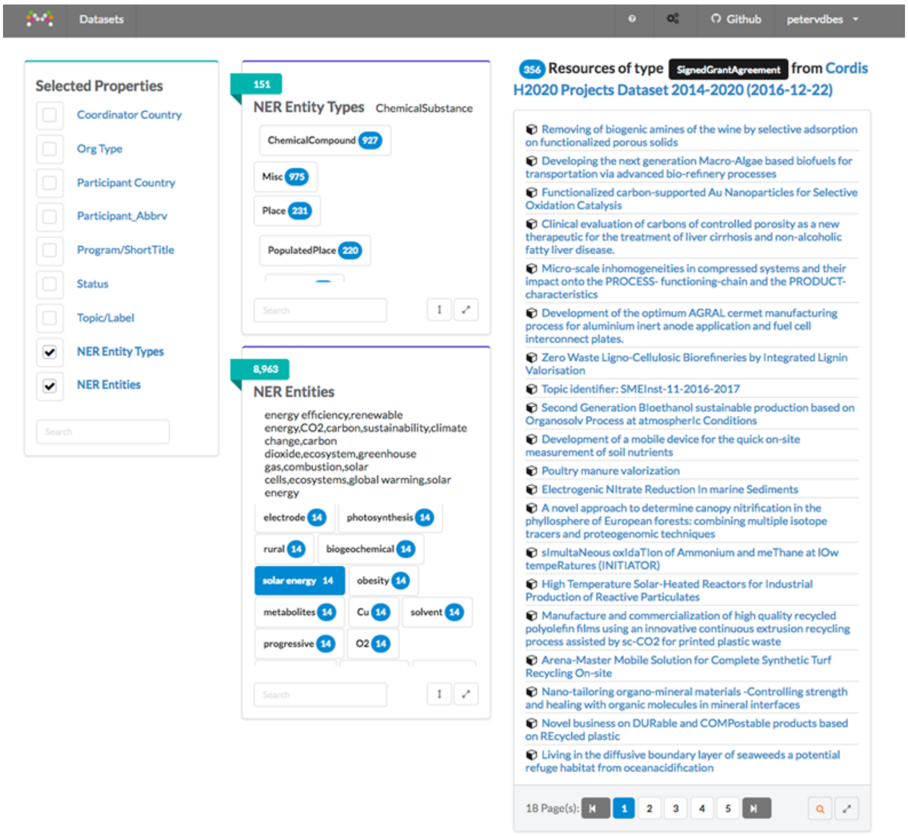
Evaluating research portfolio's through ontology based text annotation

Evaluating funding programs, or research output, has at least two dimensions: is the portfolio adequate in (i) scientific and in (ii) societal terms? A way to test this could be a double annotation process, where project descriptions or academic papers are annotated using a knowledge base with an ontology for the science fields involved, and a knowledge base with an ontology for societal challenges that should be addressed by the portfolio. Using those knowledge bases – which are generally not an individual but a collective product – overcomes the problem that individual experts annotating projects or papers are always biased and may select a biased set from the list of terms extracted from the material. This hold for technical keywords related to research fields as well as for technical terms relating to the societal challenges. Furthermore, annotating by experts is a time-consuming task, and therefore an automatic procedure would be helpful. Our approach makes use of the increasingly rich sphere of Linked (Open) Data.

We developed (as part of the SMS platform) the *annotation tool* that can be used to annotate projects using (existing) knowledge bases. Currently we deploy the *DBpedia Spotlight* tool which contains a few knowledge bases, such as DBpedia, Yago and Schema.org, but we are planning to have more knowledge bases integrated, with rich concept taxonomies for different knowledge domains. The more specific the taxonomies, the better one can assess the content of the portfolio. The tool is part of the SMS *faceted browser* in which the annotator is embedded, which enables the user to browse the linked data. For evaluation both tools are useful. The browser helps to get acquainted with the data, and gives a first qualitative idea about the project portfolio, the research topics and the societal issues addressed. By selecting the annotations, a *SPARQL query* is produced to retrieve the relevant data for further analysis and visualization.

In this paper we show an example of the tool. We use the *Cordis* open dataset with H2020 projects (version December 2016). The data were downloaded from the EC website, and converted into RDF format – the standard for linked data. This enables us to inspect and analyse the data. The browser shows the relevant characteristics of the projects, such as organizations involved, the organization type, and the program the project belongs to (figure). The CORDIS dataset contains among others a text summarizing the content of the project. This is a relatively short text, but it would not be difficult at all to couple full project descriptions (e.g., all full text granted applications) to the SMS platform. We will experiment with different texts, and try to find out what textual information leads to the most accurate representation of the projects. We use for annotation *DBpedia* as the knowledge base, but we are planning to add specific field related taxonomies.



Selected Properties

- ☐ Coordinator Country
- ☐ Org Type
- ☐ Participant Country
- ☐ Participant_Abbrv
- ☐ Program/ShortTitle
- ☐ Status
- ☐ Topic/Label
- ☑ NER Entity Types
- ☑ NER Entities

151 NER Entity Types ChemicalSubstance

ChemicalCompound 927
Misc 975
Place 231
PopulatedPlace 220

8,963 NER Entities

energy efficiency,renewable energy,CO2,carbon,sustainability,climate change,carbon dioxide,ecosystem,greenhouse gas,combustion,solar cells,ecosystems,global warming,solar energy

electrode 14 | photosynthesis 14 | rural 14 | biogeochemical 14 | solar energy 14 | obesity 14 | metabolites 14 | Cu 14 | solvent 14 | progressive 14 | O2 14

356 Resources of type SignedGrantAgreement from Cordis H2020 Projects Dataset 2014-2020 (2016-12-22)

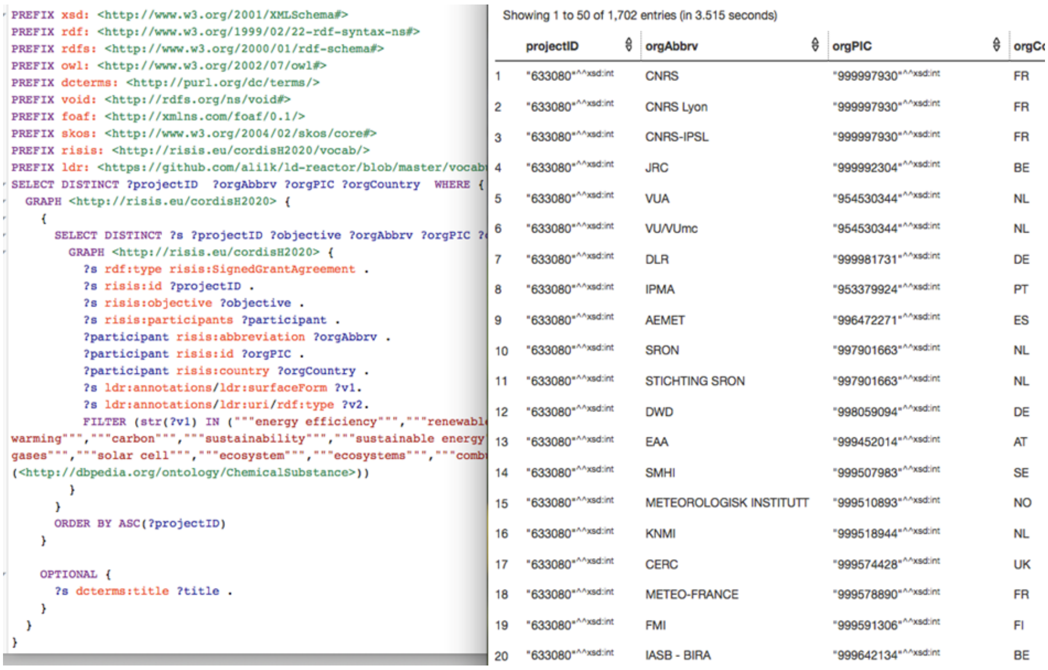
- Removing of biogenic amines of the wine by selective adsorption on functionalized porous solids
- Developing the next generation Macro-Algae based biofuels for transportation via advanced bio-refinery processes
- Functionalized carbon-supported Au Nanoparticles for Selective Oxidation Catalysis
- Clinical evaluation of carbons of controlled porosity as a new therapeutic for the treatment of liver cirrhosis and non-alcoholic fatty liver disease.
- Micro-scale inhomogeneities in compressed systems and their impact onto the PROCESS- functioning-chain and the PRODUCT-characteristics
- Development of the optimum AGRAL cermet manufacturing process for aluminium inert anode application and fuel cell interconnect plates.
- Zero Waste Ligno-Cellulosic Biorefineries by Integrated Lignin Valorisation
- Topic identifier: SMEInst-11-2016-2017
- Second Generation BIoethanol sustainable production based on Organosolv Process at atmospherIc Conditions
- Development of a mobile device for the quick on-site measurement of soil nutrients
- Poultry manure valorization
- Electrogenic NItrate Reduction In marine Sediments
- A novel approach to determine canopy nitrification in the phyllosphere of European forests: combining multiple isotope tracers and proteogenomic techniques
- sImultaNeous oxIdaTIon of Ammonium and meThane at lOw tempeRatures (INITIATOR)
- High Temperature Solar-Heated Reactors for Industrial Production of Reactive Particulates
- Manufacture and commercialization of high quality recycled polyolefin films using an innovative continuous extrusion recycling process assisted by sc-CO2 for printed plastic waste
- Arena-Master Mobile Solution for Complete Synthetic Turf Recycling On-site
- Nano-tailoring organo-mineral materials -Controlling strength and healing with organic molecules in mineral interfaces
- Novel business on DURable and COMPostable products based on REcycled plastic
- Living in the diffusive boundary layer of seaweeds a potential refuge habitat from oceanacidification

The tool does *entity recognition* using DBpedia. Entities included in DBpedia are recognized in the project descriptions. This may need some pre-processing as processes are case sensitive. As DBpedia is a knowledge graph, the projects are linked to specific places in the knowledge graph, and though the graph systematically related to each other. In the current version of SMS, entities are partly subsumed under higher level *Entity Types*. We now can use the knowledge graph to select projects. For example, by selecting a main category, e.g., *chemical compounds*, we only get projects that have in their description a term referring to chemistry (figure). As these projects are also annotated with other terms, one may add a different dimension, e.g., the application domain. By selecting within chemical compounds the annotations related to sustainability (the NER entities in the figure), we get a 'cross section' of projects within a research domain and within a societal priority.

We take the 976 chemistry projects as starting point and then select NER entities. The browser gives than all (8963) NER entities that are linked to these chemistry projects. These are listed from those that occur most often to those that occur only one time. As the NER Entities are no isolated terms but (at least partly) in a 'semantic hierarchy', it is useful to browse the NER Entities menu from top to bottom (figure). When selecting, one quickly experiences that less frequent sub-categories are not adding any projects to the list, as they are already included in higher level categories. In the menu, one sees which sub-categories are selected: Energy efficiency: Renewable energy; CO2; Carbon; Sustainability; climate change; Carbon dioxide; Greenhouse gas; Combustion; Solar cells; Ecosystems; Global warming; Solar energy. The browser (top of the right-hand window) shows that about 40% of the chemistry projects (356) is focusing on sustainability, which can be further analysed, for example in relation to *Org(anization) Type*, and *Participant Country*. One can now starts to formulate questions on how portfolios are distributed over countries, and over types of organizations. And is this distribution related to the problems of states or regions?

This combining of terms has a great advantage, as we can combine *technical research* terms and *policy related terms* to retrieve the relevant projects. This may solve the problem of finding how research links to the grand societal challenges. This is a core problem in assessing relevance of research (described in technical terms and policy related terms). Because the resulting set for a very specific topic is generally not too large, we can even manually inspect the science link.

The faceted browser produces on the background a *sparql query* which can be used to retrieve the selected data (in this case project data) for further analysis (figure left). This needs some editing and therefore some computer skills. We did this for the chemistry for sustainability portfolio, and then it is possible to use the existing tools for analysis and visualization to come to an assessment in terms of fields covered and societal issues addressed – and in terms of gaps in the portfolio.



```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX risis: <http://risis.eu/cordisH2020/vocab/>
PREFIX ldr: <https://github.com/ali1k/ld-reactor/blob/master/vocab
SELECT DISTINCT ?projectID  ?orgAbbrv ?orgPIC ?orgCountry  WHERE {
  GRAPH <http://risis.eu/cordisH2020> {
    {
      SELECT DISTINCT ?s ?projectID ?objective ?orgAbbrv ?orgPIC ?
        GRAPH <http://risis.eu/cordisH2020> {
          ?s rdf:type risis:SignedGrantAgreement .
          ?s risis:id ?projectID .
          ?s risis:objective ?objective .
          ?s risis:participants ?participant .
          ?participant risis:abbreviation ?orgAbbrv .
          ?participant risis:id ?orgPIC .
          ?participant risis:country ?orgCountry .
          ?s ldr:annotations/ldr:surfaceForm ?v1.
          ?s ldr:annotations/ldr:uri/rdf:type ?v2.
          FILTER (str(?v1) IN ("""energy efficiency""","""renewabl
warming""","""carbon""","""sustainability""","""sustainable energy
gases""","""solar cell""","""ecosystem""","""ecosystems""","""comb
(<http://dbpedia.org/ontology/ChemicalSubstance>))
        }
      }
      ORDER BY ASC(?projectID)
    }

    OPTIONAL {
      ?s dcterms:title ?title .
    }
  }
}
```

Showing 1 to 50 of 1,702 entries (in 3.515 seconds)

	projectID	orgAbbrv	orgPIC	orgC
1	"633080"^^xsd:int	CNRS	"999997930"^^xsd:int	FR
2	"633080"^^xsd:int	CNRS Lyon	"999997930"^^xsd:int	FR
3	"633080"^^xsd:int	CNRS-IPSL	"999997930"^^xsd:int	FR
4	"633080"^^xsd:int	JRC	"999992304"^^xsd:int	BE
5	"633080"^^xsd:int	VUA	"954530344"^^xsd:int	NL
6	"633080"^^xsd:int	VU/VUmc	"954530344"^^xsd:int	NL
7	"633080"^^xsd:int	DLR	"999981731"^^xsd:int	DE
8	"633080"^^xsd:int	IPMA	"953379924"^^xsd:int	PT
9	"633080"^^xsd:int	AEMET	"996472271"^^xsd:int	ES
10	"633080"^^xsd:int	SRON	"997901663"^^xsd:int	NL
11	"633080"^^xsd:int	STICHTING SRON	"997901663"^^xsd:int	NL
12	"633080"^^xsd:int	DWD	"998059094"^^xsd:int	DE
13	"633080"^^xsd:int	EAA	"999452014"^^xsd:int	AT
14	"633080"^^xsd:int	SMHI	"999507983"^^xsd:int	SE
15	"633080"^^xsd:int	METEOROLOGISK INSTITUTT	"999510893"^^xsd:int	NO
16	"633080"^^xsd:int	KNMI	"999518944"^^xsd:int	NL
17	"633080"^^xsd:int	CERC	"999574428"^^xsd:int	UK
18	"633080"^^xsd:int	METEO-FRANCE	"999578890"^^xsd:int	FR
19	"633080"^^xsd:int	FMI	"999591306"^^xsd:int	FI
20	"633080"^^xsd:int	IASB - BIRA	"999642134"^^xsd:int	BE

The procedure shown in this paper would enable an evaluation of the H2020 (or any other program – or publication corpus) in terms of its scholarly and its societal focus. How many projects are devoted to specific research fields (the goal of stimulating excellent research) and how many to specific societal challenges? As the research fronts and the societal challenges change over time, one may do the analysis for time slices of projects and evaluate the change of the quality of the portfolio over time.

What are the next steps?

- Extending the tool with other knowledge graphs: specialized ontologies or vocabularies;
- Full text use for annotating, and testing which parts of the text are important;
- Standard queries for retrieving parts of the portfolio for further inspection (e.g. using statistics or visualization); these standard queries would help the user without the computer skills needed to edit the automatically generated queries;
- When the dataset is very large, selecting in the browser takes time; further work on increasing the speed of querying in the browser.