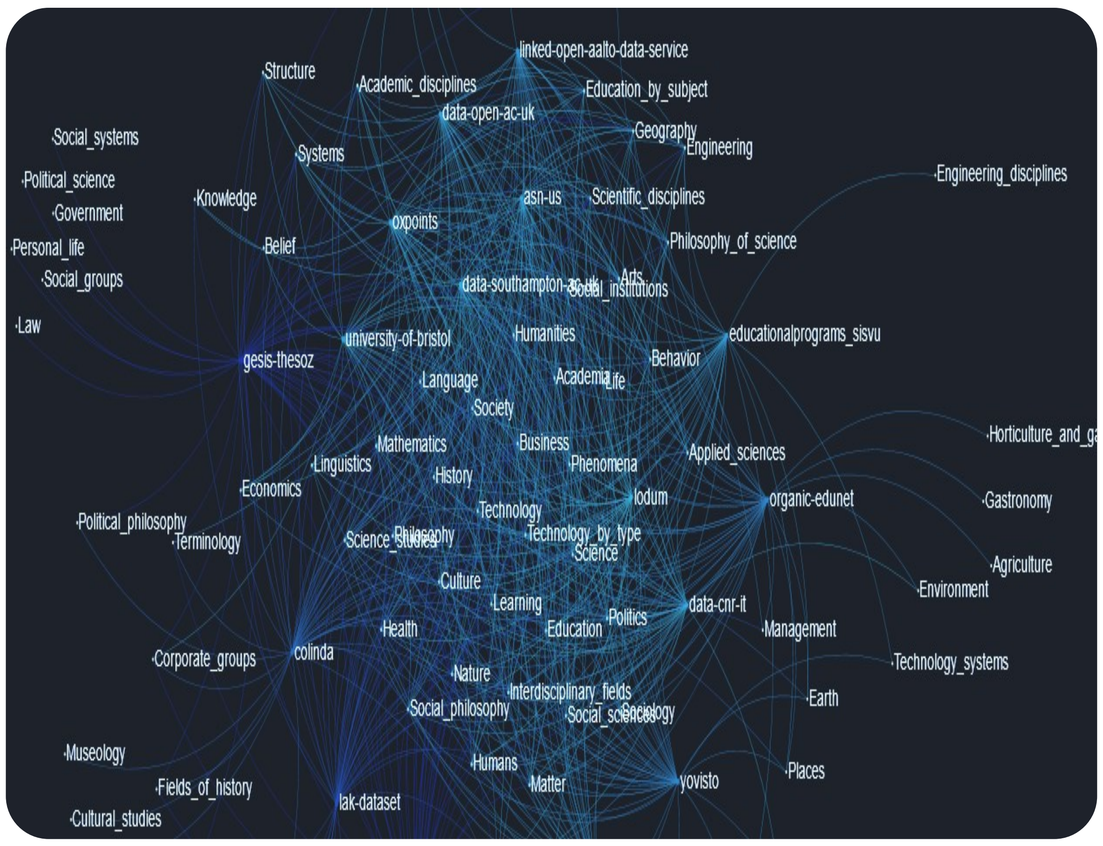
Generating structured Profiles of Linked Data Graphs

Besnik Fetahu[1], Stefan Dietze[1], Bernardo Pereira Nunes[1,2], Davide Taibi[3] and Marco A. Casanova[2]

[1]L3S Research Center, Leibniz University Hannover(Germany), [2]Pontifical Catholic University of Rio de Janeiro (Brazil),

[3]Italian National Research Council, Institute for Educational Technologies (Italy)



Graph representation of example generated structured profiles of Linked Data Graphs.

Abstract

The profiling pipeline automatically assesses, annotates and indexes available linked datasets. The generated profiles embed datasets into an interlinked data-graph of datasets based on shared topics and vocabularies.

The pipeline for generating structured profiles of linked data graphs considers the following:

① Incremental sampling of resources

② Entity Recognition

③ Category Extraction and Normalisation

④ Automated Validation & Filtering

⑤ Explorable structured dataset profiles

Approach: Generating structured Profiles

To address the of lack of descriptions of Linked Datasets in DataHub, we provide an automatic mechanism for generating structured profiles captured as part of a VoID dataset.

Indexing: automatically index subset of *resource instances* for all existing *resource types* from datasets of interest in DataHub.

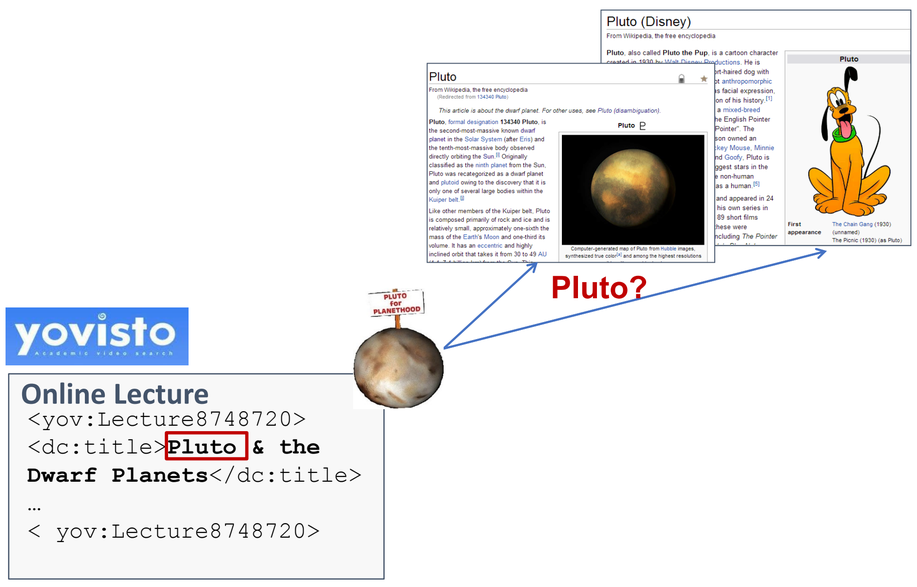
Named Entity Recognition & Disambiguation

Analyse textual content assigned to datatype properties like:

- `rdfs:label`
- `rdfs:comment`
- `teach:courseTitle`
- `---------------`

Incremental Annotation:

- *Pool* of extracted entities
- Similarity of *entity description* and resource's *textual content*
- Assign to resources entities above a pre-defined *threshold* of similarity



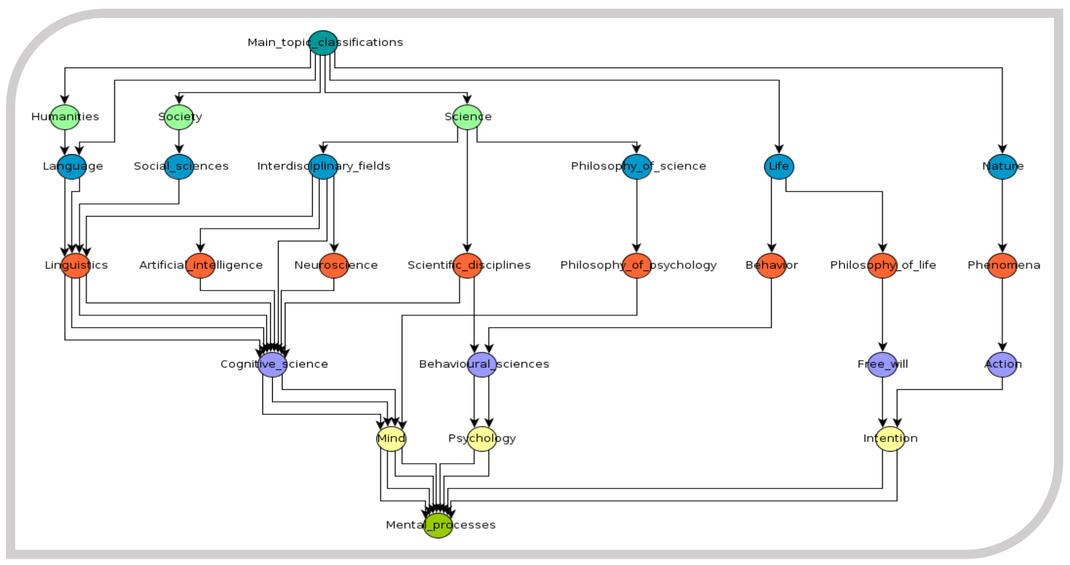
Category Extraction and Normalisation

Profiles as a set of ranked DBpedia *categories*. Assess the DBpedia sub-graph of directly related and broader topics (up to four levels) from *extracted entities* using datatype properties:

- `dcterms:subject`
- `skos:broader`

Category Ranking & Normalisation: measure the representativeness of a category for a dataset and how well it distinguishes from other datasets.

$$score(t) = \frac{\Phi(t,D)}{\Phi(\cdot,D)} + \frac{\Phi(t,\cdot)}{\Phi(\cdot,\cdot)},\ \ \forall t \in \mathcal{T} \wedge D \in \mathcal{D}$$



Category Graph of example entity <Mental_process>

Explorable structured dataset profiles

The structured profiles, currently are generated for the **linked-education** data group in DataHub.
The data can be accessed via *SPARQL endpoint* or via the *exploratory search interface* provided. For more visit the web-site of the demo:

http://data.linkededucation.org/

Contact info:
Besnik Fetahu
L3S Research Center
Appelstr. 9a,
30167 Hannover, Germany
Email: fetahu@L3S.de

Interested on similar projects?

We recommend:

-TRT

- Cite4Me

Vote for us!