Generating structured Profiles of Linked Data Graphs

Besnik Fetahu\textsuperscript{1}, Stefan Dietze\textsuperscript{1}, Bernardo Pereira Nunes\textsuperscript{1,2}, Davide Taibi\textsuperscript{3} and Marco A. Casanova\textsuperscript{2}

\textsuperscript{1}L3S Research Center, Leibniz University Hannover (Germany), \textsuperscript{2}Pontifical Catholic University of Rio de Janeiro (Brazil), \textsuperscript{3}Italian National Research Council, Institute for Educational Technologies (Italy)

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

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.

\[
\text{score}(t) = \frac{\phi(t, D)}{\Phi(t, D)} + \frac{\phi(t, \cdot)}{\Phi(t, \cdot)} \quad \forall t \in T \land D \in \mathcal{D}
\]

Approach: Generating structured Profiles

To address the 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 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/