Group A3D: RDF Dataset

Group members: Andrea Bruttomesso 2120933
Alessandro Corrò 2125034
Davide Seghetto 2122548
Andrea Stocco 2108885

Link to the GitHub repository: https://github.com/andreastocco01/a3d

Ontology domain

The domain of interest for our ontology focuses on scientific research. In particular, we aim to analyze potential correlations between Nobel Prize winners, their publications, and the research funding allocated by nations. Figure 1 shows a schema of our ontology comprehensive of all the classes and the properties modeled. For data integration we also imported well-known ontologies available on the web, such as *SKOS*, *FOAF* and *EulerSharp countries*.

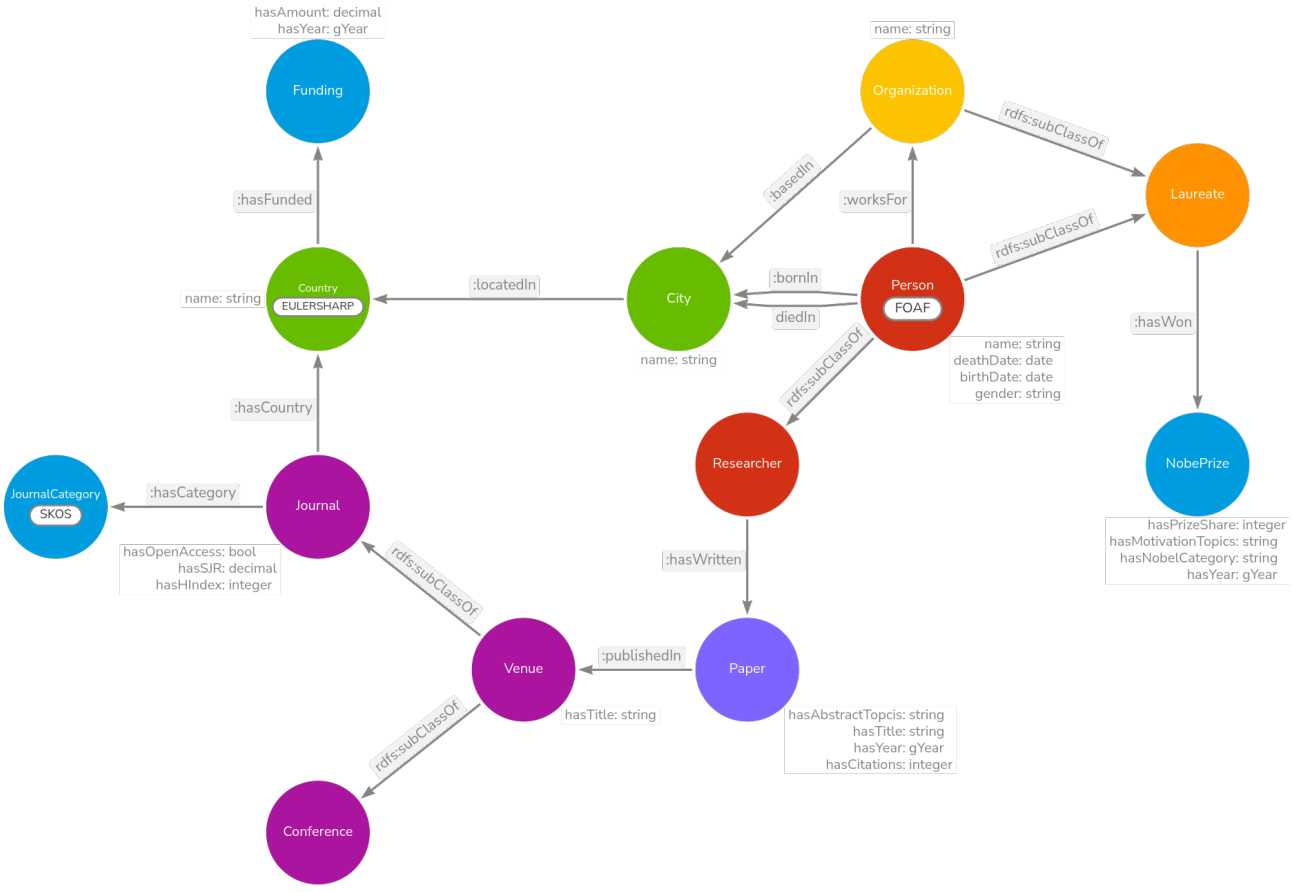

Figure 1: Nobel Ontology

Datasets and Population

To populate our ontology, we used four distinct open datasets containing information about Nobel Prizes, scientific papers, publication venues and national funding. The use of multiple datasets required us to address various inconsistencies.
Specifically, to resolve the discrepancies in names across the datasets, we used the Python library *rapidfuzz* for fuzzy matching the laureates' names with those of paper authors. We also ensured the absence of duplicates to

maintain data integrity.
Additionally, we encountered discrepancies with the EulerSharp ontology. For instance, some country names in our datasets differed from those used in the EulerSharp ontology, such as "United States" and "United States of America". These cases were manually addressed to ensure consistency.
Another critical aspect was filtering the information provided by various datasets. Initially, the Paper dataset contained over one million rows, making it necessary to reduce its size to avoid an excessively large and imbalanced database, particularly favoring the Researchers in terms of class population. To address this, we first selected only papers authored by Laureates or published in Journals also included in the dedicated dataset. Subsequently, we further narrowed it down to the first 50,000 rows of the already filtered dataset for the reasons outlined above.
Furthermore, we wrote a small script (*topic_extraction.py*) that uses *gensim* to extract the most significant topics from lengthy textual properties. For example, long descriptions such as "in recognition of the extraordinary services he has rendered by the discovery of the laws of chemical dynamics and osmotic pressure in solutions" were reduced into more focused terms like "chemical dynamics osmotic pressure solutions." This approach enhances the dataset's efficiency and significance while preserving its key concepts.

Main Statistics

The following statistics summarize the key features of the RDF dataset.

- **Triples Count:**
  - Total number of RDF triples: 3298916
- **Entities:**
  - Number of unique entities:
    - Person: 110981
    - Researchers: 110160
    - Laureates: 904 (23 Organizations, 881 Persons)
    - Organizations: 339
    - Papers: 53492
    - Journals: 18466
    - Conferences: 309
    - Cities: 874
    - Countries: 246
    - Nobel Prizes: 579
    - Funding: 1373
- **Laureates and Nobel Prizes:**
  - Researchers that won a Nobel Prize: 60
  - Total Nobel Prizes awarded: 969
  - Prize categories: 6 (Physics, Chemistry, Medicine, Literature, Peace, Economics)
- **Most Funding Countries:**
  1. United States: $4128 bn
  2. Japan: $1189 bn
  3. Germany: $1056 bn