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

Mapping of the open data in RDF

We modeled an ontology that captures the relationships among three key domains: Nobel Prize winners, academic papers published in journals or conferences, and R&D (research and development) budgets allocated by nations over the years. By integrating these domains, the ontology provides a comprehensive framework to explore connections between outstanding scientific discoveries, scholarly output, and the financial investments that drive global innovation. This integrated approach enables a deeper understanding of the dynamics between scientific achievement and the resources that fuel it, highlighting how funding decisions may influence groundbreaking research and its impact on society. Through this process, we aim to create a valuable resource for researchers and scholars, offering a clearer view of the intersection between scientific achievements, academic output, and the financial investments that drive progress in various fields of knowledge.
In order to model our ontology, we utilized four distinct open datasets, which contain information regarding Nobel Prize winners, scientific papers, scientific journals, and investments in R&D by countries.
We mapped the data into RDF format in the following way.

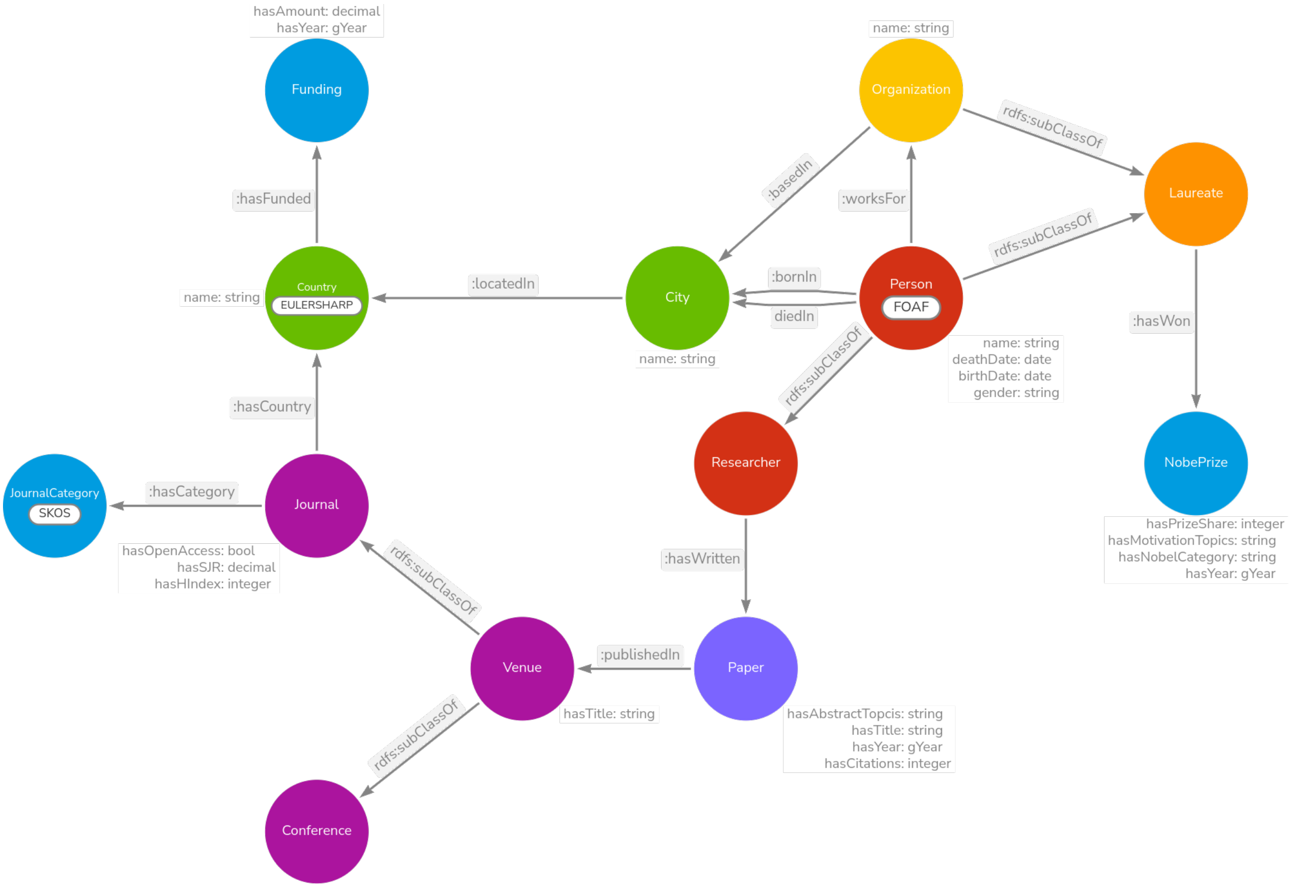

Figure 1: nobelOntology

We had to address the issue of duplicates to ensure that there were no resources in the RDF dataset with the same URI representing the same entity. This is a critical aspect of maintaining data integrity and avoiding inconsistencies in the ontology. Therefore, before adding a new resource to the RDF dataset, such as a scientific journal, we performed a thorough check to confirm that the resource was not already present in the graph. To facilitate this, we defined a series of functions, such as *handle_city* and *handle_org*.

In addition to addressing duplicates, we also encountered the challenge of discrepancies in naming conventions between the data in our open datasets and the EulerSharp ontology. In particular, some country names in our datasets differed from those used in the EulerSharp ontology (e.g., "United States" vs "United States of America"), so we had to manually handle these specific cases.

Moreover, we wrote a script (*topic_extraction.py*) to extract the most significant words from the texts defining the abstract of a scientific paper and the motivation for the award of a Nobel Prize. The script was developed to generate shorter, more concise values for the properties *motivationAbstract* and *motivationTopics*. For instance, 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" are transformed into more focused terms like "chemical dynamics osmotic pressure solutions." This approach enhances the dataset's efficiency by reducing the length of textual values while preserving the key concepts relevant to the Nobel Prize motivation and the research topics.

Main statistics