Nobel Ontology

Group A3D

Andrea Bruttomesso, Alessandro Corrò, Davide Seghetto, Andrea Stocco

January 14, 2025

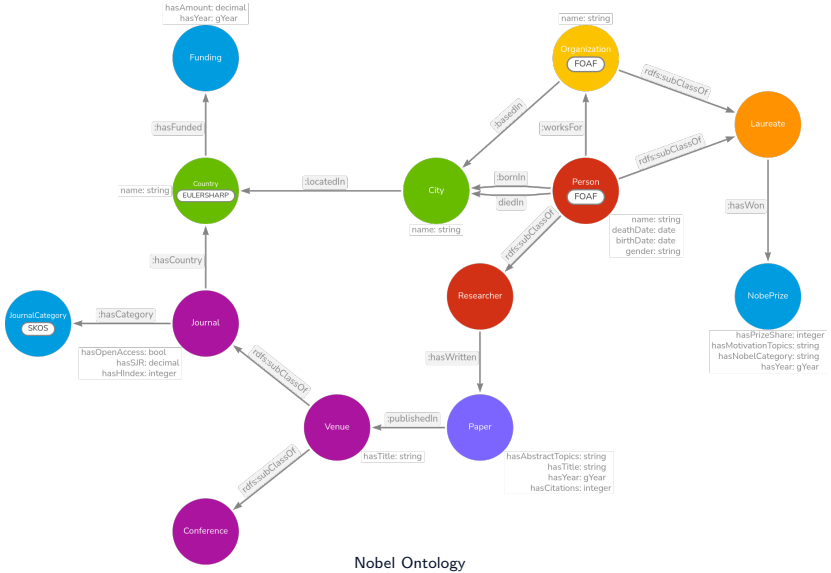
Overview

1. **Domain of Interest**
2. **Ontology Design**
3. **Problems**
4. **Analytics**

Domain of Interest

We have chosen the domain of scientific research. Specifically, we aim to analyze potential correlations among Nobel Prize winners, their publications, and the research funding invested by various countries. This domain was selected because it allows us to reveal potential historical and geographical patterns in scientific research.

Nobel Ontology



Problems

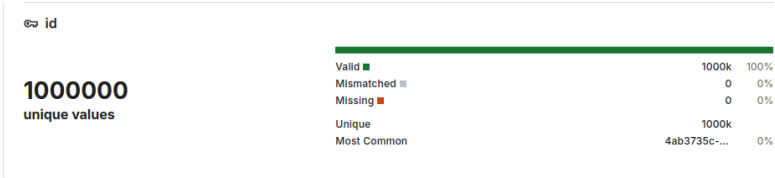
- Errors in Nobel laureates dataset
- Subset of papers dataset
- Researchers and Nobel laureates matching

Errors in Nobel-Laureate dataset

Table: Example of dataset error

Year	Category	Prize Share	Full Name
1908	Medicine	1/2	Ilya Ilyich Mechnikov
1908	Medicine	1/2	Paul Ehrlich
1908	Medicine	1/2	Paul Ehrlich

Subset of papers dataset



The Papers Dataset originally contained 1 million rows, which made it necessary to filter the data. After careful consideration, we decided to retain only the papers authored by a Laureate or published in a venue included in the Venues Dataset. Even with this filter applied, the dataset still contained approximately 300,000 rows so we further reduced it by selecting only the first 50,000 rows.

Researchers and Nobel laureates matching

The Laureates and Papers Datasets often represent names differently, such as "Antoine Henri Becquerel" and "Antoine H. Becquerel." To effectively link these datasets, we utilized a library that implements a fuzzy matching algorithm with a similarity threshold of 90%. This approach allowed us to identify and match names that were not identical but sufficiently similar, ensuring a robust connection between the datasets.

Relationship between Nobel Prize winning ideas and published studies

Table: Number of papers per Nobel topic in 2004

Nobel topic	Nobel Prize	Number of papers
protein	Chemistry 2004	28
development	Peace 2004	13
flow	Literature 2004	8
interaction	Physics 2004	4
discovery	Chemistry 2004	3
discovery	Physics 2004	3
degradation	Chemistry 2004	3
asymptotic	Physics 2004	2
forces	Economics 2004	1
cycles	Economics 2004	1
olfactory	Medicine 2004	1
organization	Medicine 2004	1

The topic "protein" appeared in 28 papers. The high number of papers mentioning this Nobel topic suggests that it was widely discussed or relevant in 2004.

Most active research areas in a year

Table: Number of papers for each journal subcategory in 2004

Journal Subcategory	Number of papers
Biochemistry Genetics Molecular Biology	418
Social Sciences	344
Decision Sciences	125
Arts Humanities	74
Business Management Accounting	68
Physics Astronomy	65
Neuroscience	55
Health Professions	34
Psychology	27
Earth Planetary Sciences	22
Economics Econometrics Finance	14
Materials Science	12
Environmental Science	12
Agricultural Biological Sciences	10
Energy	2
Pharmacology Toxicology Pharmaceutics	2

Molecular biology was the most active research area in 2004.

Questions?