

Semantic Integration and Interoperability

Ignacio Amaya and Philipp Eisen

Introduction

Nowadays the amount of data that is being generated is huge, but connecting this data is usually very time consuming because datasets are usually in different format and they don't share the same structure. That's why a lot of resources are being used for creating new ways of easing the integration of data sources. Ontologies provide a very effective way to solve that problem and link data from different systems or organizations. The ontologies define the way the data sources should be merged, so you don't need to perform additional ETL tasks. As data can be linked in very different ways, it is very important to specify which ones are the objectives of the ontology. There could be different ontologies linking two databases for different purposes. In this report we are going to focus on Open data in education and try to show the power of ontologies to solve the problems stated above.

Open data is very common in the educational field, but each country usually stores that data using different formats. The different educational systems make it difficult to compare two countries using this data. In this project we've solved a problem in this domain using a Linked Data approach. The process we have followed will be explained in detail in this report, but can be summarized into three basic steps:

- Selection of the educational data sources that will be converted into Linked Data
- Design of the ontology that will allow to map those two datasets
- Generation of the RDF data, linking it with other existing datasets, such as Dbpedia.

Purpose of the Developed Ontology

The purpose of this set of ontologies is to allow to compare the number of students coursing different university educational fields by countries and gender. Given two countries it can be known the percentage of students of a subject type by gender. This information can be useful to detect differences among countries for a given subject in the distribution of the gender of their students. Measures to tackle specific distribution problems could be taken based on that knowledge. For example, if Spain has a small amount of female students in Science, but germany the number is much higher that can mean that the education system in Spain could be adapted to balance the Science students. This could be done copying some aspects of the educational model in Germany in that area.

Data sources

We have selected two data sources that contain information about how many students are currently studying in their universities, which subjects they study and their gender. However, this information is presented in very different formats. Below we explain where we collected the datasets and how each of them are structured. The extension of both of them is csv.

Original data sources

- [Students enrolled in Germany by course](#)
- [Students graduated in Singapore](#)

Germany Format

Each row contains one subject in one semester and the number of male and female students. They have the total count of students enrolled, but also they differentiate among German and Foreign students. The semesters are encoded using a german format for administrative purposes.

Singapore Format

Each row contains the year, the sex of the students, the type of course and the number of graduates.

Differences

The time formats are different, because one uses semester and the other uses years. They also have different subjects. While the German dataset has a large number of them, the one from Singapore has only 15. This is because they are subject types denoting their educational field, while in the German data the subjects are not grouped into categories.

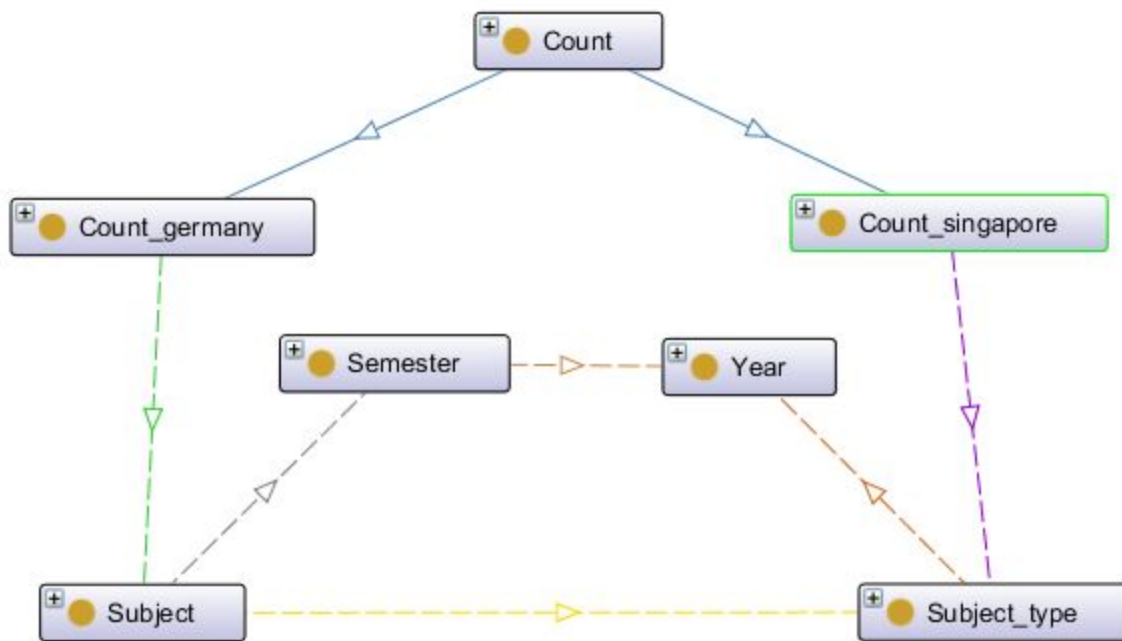
Methodology

First of all, as the subjects and subject types from Germany and Singapore were different it was necessary to manually map them. The German subjects were grouped depending on their specific field of study, but sometimes the mapping was not so clear, so there can be some mappings that are not completely accurate.

We have considered that the semesters when the students enroll can be mapped with the years they graduate. These two values can be compared, so we have made a relation among those concepts. We follow a linked data approach while merging the two datasets, which focus on reusability. We discarded the option of doing three ontologies (two source ones and a mapping

one), but this could be a good option to keep things organized in case datasets from more countries want to be linked.



The ontology was created using Protege. The main class where all the information is stored is the Count class. It has two subclasses: Singapore counts and German counts. Each of the counts has associated a subject or a subject type and those are from a year or a semester. These are the classes where the data is linked.



In order to insert the instances we created a Python script which automatizes that task. The years were linked with Dbpedia pages, so each of the years can be put into their context of what else happened in that year.

Testing

In order to ensure the correctness of our owl file generated we have tested the ontology with the RDF instances. We actually found that there were some mistakes in the counts that didn't have a value in the csv files. We didn't had them into account and although the ontology was valid in Protegé, the types were wrong. We fixed that problem thanks to the insights provided by the OWL2 validator from the University of Manchester.



OWL 2 Validation Report

Summary

The ontology and all of its imports are in the OWL 2 profile

Imports Closure

Ontology IRI	Physical URI
OntologyID(OntologyIRI(<http://www.semanticweb.org/educationMapping>))	

Some other modifications should be performed to obtain a better quality ontology, although it is not necessary for the correctness of the ontology. For that we used Opps ontology pitfall scanner. We should add disjoints, annotations and inverse relationships according to the good practices.

Evaluation results

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

- **Critical** 🚫 : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** ⚠️ : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** 🟡 : It is not really a problem, but by correcting it we will make the ontology nicer.

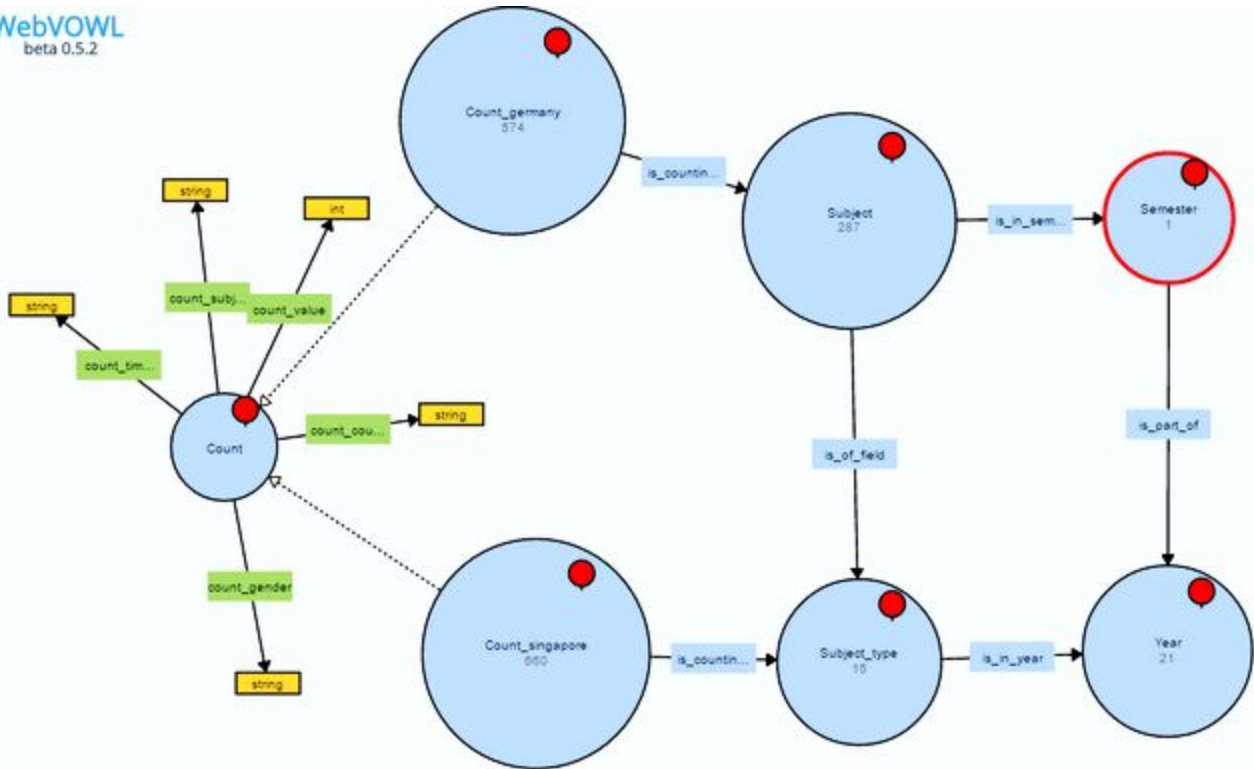
[Expand All] | [Collapse All]

Results for P08: Missing annotations.	18 cases Minor 🟡
Results for P10: Missing disjointness.	ontology* Important ⚠️
Results for P13: Inverse relationships not explicitly declared.	6 cases Minor 🟡

Results

All the results are presented in an owl file where all the definition of the classes, the object properties and the attributes are defined. This file also has all the RDF instances initialized using the data present in the datasets we used.

A visualization of the final ontology with all the RDF instances is presented below using WebOWL. We can notice that only a semester is present in the German dataset, so it will be advisable to find more datasets from Germany of other years to make our work more useful.



Conclusions

The mapping performed allows to access counts of the number of students from Germany and Singapore in a given year. Thanks to the created ontology it is easy to perform queries using SPARQL. Applications could use this data to analyze differences among those two countries. We've shown that linking ontologies is more reusable than the conventional ETL processes, because other countries could be linked into this data, while ETL processes are hardly ever reused because they are performed to solve very specific problems. We have eased the access to the selected data and researchers will be able to compare the counts according to their number of students in each subject each year without having to worry about integrating issues, which will save them a lot of effort and time.

We can conclude that ontologies have a huge potential for saving time and efforts in integration. If all the open data accessible today is linked, it will create new possibilities for researchers and businesses. We are creating huge amount of data every day, but most of it is useless if isolated. That's why this techniques for integrating data sources and enhancing interoperability are very important and should become a standard in some fields, such as education or medicine.