Development of an Ontology for CMS Open Data Portal @ IFCA

Internship Report

INSTITUTO DE FÍSICA DE CANTABRIA
Departamento de Física de Partículas experimental
Av. de los Castros, 39005 Santander, Cantabria

Author:
Guadalupe Cañas Herrera

Directores:
Intership supervisor: Jesús Marco de Lucas
Academic supervisor: Alicia Calderón Tazón

BSc in Physics

Santander, August 15th 2015 - November 15th 2015

Índice

1. Personal Data

- Name: Guadalupe Cañas Herrera
- DNI: 72073192R
- Current studies: Bachelor in Physics, tracks in Fundamental Physics
- email: canasg@ifca.unican.es / gch24@alumnos.unican.es
- Start Date: August 15th 2015
- End Date: November 15th 2015
- Time Schedule: from 9:00 AM to 14:00 PM
- Hours: 450

2. Purpose of the Internship

In this report it is explained my experience as intern at Instituto de Física de Cantabria, in which I carried out my internship in the fields of Particle Physics and Computer Science. Moreover, I state how I accomplished my tasks and how they improved my learning in Physics and especially in research.

I decided to carried out an internship at IFCA in order to get experience in a research group and to learn how science is perform nowadays, although I did not need explicitly the credits ECTS to pass a course at the University of Cantabria. Furthermore, as I plan to study a master program, I expected this internship to clarify if being a researcher is a suitable career for me in the future. This opportunity of working in a physics department together with a research group make me realize that, in fact, I wish to pursue a master program destined to get enrolled in a PhD program.

3. About the Collaborating Entity and the Department

I developed my internship at Instituto de Física de Cantaria (IFCA), a joint centre between the University of Cantabria and the Consejo Superior de Investigaciones Científicas (CSIC) set up in 1995. Its orientation is mainly the research on basic science, whose more important research lines are Particle Physics, Astrophysics and Observational Cosmology and Statistical and Non-linear Physics. The IFCA also produces yearly more than 220 publications in the best journals in the respective fields and has nearly 15 active projects, providing an idea of how the impact this institute has in the scientific community as well as in the academic one at University of Cantabria is.

I joined the Particle Physics Department at IFCA, specifically the Experimental Group. This department works for the CMS Experiment (Compact Muon Chamber Experiment) placed at the LHC in the CERN, Switzerland. I worked for CMS Preservation Group, whose main responsibility is to guard and maintain CMS Data for the future years in such a way that the data can still being accessible for the rest of scientists and the general public.

4. About the tasks

4.1. General Description

I am currently in charge of creating, designing and implementing a didactic ontology on High Energy Physics with a focus on CMS Experiment and its public data from 2010 in the CMS Open Portal at IFCA.

The CMS Open Portal at IFCA, whose web-site is `https://cmsopendata.ifca.es/`, is an online virtual platform conceived to introduce Bachelor Students in Particle Collisions and the analysis of the outcoming data from CMS Experiment. At the moment, this portal contains:

1. Virtual Machine with CMS Environment Software and ROOT (specific programs from CERN)

2. Analysis Software written in Python Language designed by the scientist Ana Rodríguez

3. CMS Public Data from 2010

The initial aim was to upload this existing portal with a new analysis software (task developed by the bachelor student Palmerina González Izquierdo) and with a formal ontology definition in High Energy Physics suitable to be used with a didactic purpose in order to present the analysis steps and the common vocabulary used to describe the process.

4.2. Design of an Ontology

4.2.1. Introduction

An Ontology could be understood as *a set of concepts in a domain of knowledge linked by their relations through the use of a common vocabulary*. Therefore, an Ontology is basically a model that describes a concrete knowledge field and introduces the possibility of organizing and planning science research. This scheme can be understood both by people and machines, easing the way of communicating among them. At the same time, the communication allows to share knowledge among scientific groups.

Researchers works with already-defined steps in the procedure. Data Analysis processes generally consist in a sequence of time ordered steps over data suitable to be modelled using scientific workflows. Consequently, a scientific workflow is a description of a process for accomplishing a scientific objective. If we develop an ontology able to express the logic between concepts and step in the process, we will obtain an ideal way of implementing a workflow thanks to a general semantic framework; that is, a collection of common vocabulary and expressions in a knowledge field.

The Ontology Implementation may also promote data analysis preservation, including the data itself as well as the procedure to obtain the final result. It can be also useful for new students being introducted in a field where a workflow is already in use.

4.2.2. Learning the concepts

The first step in my internship was to study several concepts related to Ontologies. I needed to understand what the Unified Modelling Language (UML) is and its relation with the World Wide Web. Furthermore, I had to understand what the Semantic Web and Web of Data are.

Basically, computer resources, which are any physical or virtual components of limited availability within a computer or information management system, are identified through Uniform Resources Identifiers, or URIs, according to the Semantic Web and the World Wide Web Consortium. To organize these URIs, it is used the Resource Description Framework (RDF), that allows also to describe relations between URIs in the form of triples. The triplets are in the form of,

$$Subject + verb + complement$$

where the subject is a URI and the complement may be another URI or information related to the subject as data. In this last case, that data associated to the *subject* is known as *metadata*.

The organization of these triplets may be applied in a common way in order to share knowledge. In this case, the *verb* in the triplet structure is obtained from a common language. The most well-known are RDF-s and OWL. Furthermore, OWL introduces the possibility of organizing the triplets through logical closures so that we can include conditional situations. Therefore, RDF together with RDF-s and OWL provide the perfect mechanism for designing an ontology in a field of knowledge.

4.2.3. Search for previous examples

Secondly, once I had a basic background in Ontologies, I seek for previous examples of High Energy Physics Ontologies developed by other groups. In first instance I focused on DASPOS project (`https://daspos.crc.nd.edu`) as well as on CERN Library Project.

The first group, DASPOS, is currently developing an ontology in JSON format in order to organize the real data measured in CMS detector and preserve the whole process. The have attached several details in the measurement and analysis workflows but they lack in a didactic point of view in order to provide a general idea of the process as a whole.

On the other hand, CERN Library Project contains this didactic point of view although they missed the rigour and preciseness that should be required in order to implement a common Semantic Framework as they simple use html forms.

From that point on, I centred my attention in the creation of a ontology that fulfils the requisites by the World Wide Web Consortium and is also defined through the main purpose of outreaching High Energy Physics.

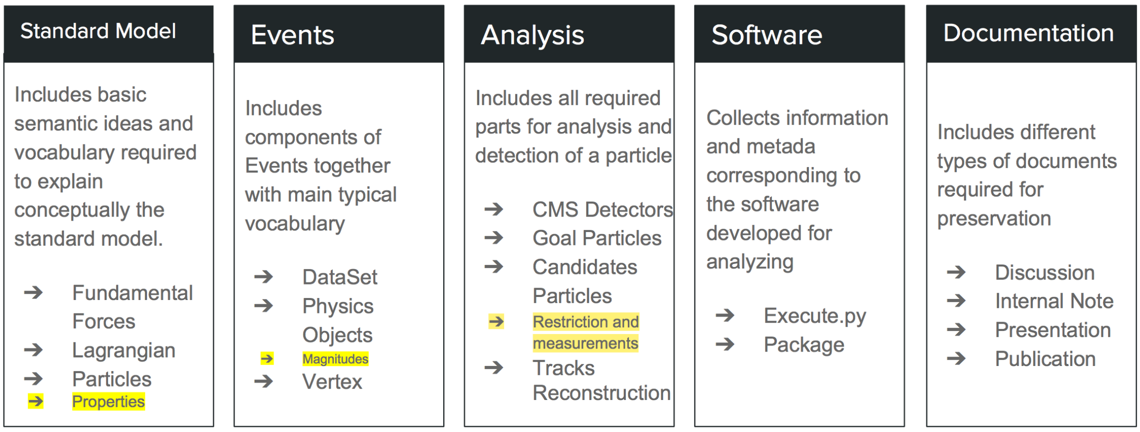
4.2.4. Defining an Implementing an Ontology Platform

The best (and easiest) way of designing an ontology is through the proper graphical tool. In this case, I went through ontologies thanks to *Protégé*. Protégé is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies. It was developed by Stanford University and has become one of the most famous ontologies' application.

As it is been already mentioned, an ontology is a formal explicit description of concepts in a domain of discourse in the form of **classes**, **properties** (remind the *verbs* in the triplets structure) of each class describing various features and attributes of the concept, and restrictions on properties known as **facets**. An ontology together with a set of individual instances of classes constitutes a knowledge base.

The concept of ontology is quite similar to Object-Oriented programming. If you are provided with a general class, you may create objects that inherit class-properties and also include their own properties' values. Analogously, in an Ontology you can find general classes, and their individuals instances that resemble the objects in Object-Oriented programming.

In this case of a High Energy Physics Ontology, I defined five different general classes as it is shown in Fig. 1.



Standard Model

Includes basic semantic ideas and vocabulary required to explain conceptually the standard model.

➔ Fundamental Forces
➔ Lagrangian
➔ Particles
➔ Properties

Events

Includes components of Events together with main typical vocabulary

➔ DataSet
➔ Physics Objects
➔ Magnitudes
➔ Vertex

Analysis

Includes all required parts for analysis and detection of a particle

➔ CMS Detectors
➔ Goal Particles
➔ Candidates Particles
➔ Restriction and measurements
➔ Tracks Reconstruction

Software

Collects information and metada corresponding to the software developed for analyzing

➔ Execute.py
➔ Package

Documentation

Includes different types of documents required for preservation

➔ Discussion
➔ Internal Note
➔ Presentation
➔ Publication

Figura 1: Ontology Class Diagram showing the main classes and their description and subclasses.

From each class I created different individuals (see Fig. 2). Individuals are related with another individuals or metadata through properties that are classified as,

- **object properties**: for instance,

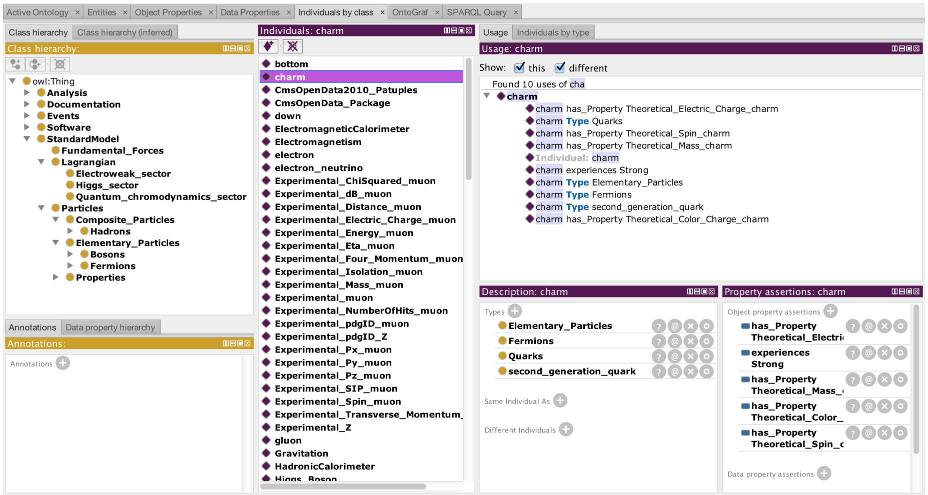
  experimental_muon is_detected_in muon_chambers

  where experimental_muon is a individual from the class Events and muon_chambers is a individual from the class Analysis.

- **data properties**: for instance,

CMSOpenData_Analyzer has_Version "1.0.0"

where CMSOpenData_Analyzer is a individual from the class Software and "1.0.0ïs simply a string value attached to the individual.



Class hierarchy:

owl:Thing
Analysis
Documentation
Events
Software
StandardModel
Fundamental_Forces
Lagrangian
Electroweak_sector
Higgs_sector
Quantum_chromodynamics_sector
Particles
Composite_Particles
Hadrons
Elementary_Particles
Bosons
Fermions
Properties

Annotations ⊕

Individuals: charm

bottom
charm
CmsOpenData2010_Patuples
CmsOpenData_Package
down
ElectromagneticCalorimeter
Electromagnetism
electron
electron_neutrino
Experimental_ChiSquared_muon
Experimental_dB_muon
Experimental_Distance_muon
Experimental_Electric_Charge_muon
Experimental_Energy_muon
Experimental_Eta_muon
Experimental_Four_Momentum_muon
Experimental_Isolation_muon
Experimental_Mass_muon
Experimental_muon
Experimental_NumberOfHits_muon
Experimental_pdgID_muon
Experimental_pdgID_Z
Experimental_Px_muon
Experimental_Py_muon
Experimental_Pz_muon
Experimental_SIP_muon
Experimental_Spin_muon
Experimental_Transverse_Momentum_
Experimental_Z
gluon
Gravitation
HadronicCalorimeter
Higgs_Boson

Usage: charm

Show: ☑ this ☑ different

Found 10 uses of cha

charm
charm has_Property Theoretical_Electric_Charge_charm
charm Type Quarks
charm has_Property Theoretical_Spin_charm
charm has_Property Theoretical_Mass_charm
Individual: charm
charm experiences Strong
charm Type Elementary_Particles
charm Type Fermions
charm Type second_generation_quark
charm has_Property Theoretical_Color_Charge_charm

Description: charm

Types ⊕
Elementary_Particles
Fermions
Quarks
second_generation_quark

Same Individual As ⊕

Different Individuals ⊕

Property assertions: charm

Object property assertions ⊕
has_Property Theoretical_Electri
experiences Strong
has_Property Theoretical_Mass_
has_Property Theoretical_Color_
has_Property Theoretical_Spin_c

Data property assertions ⊕

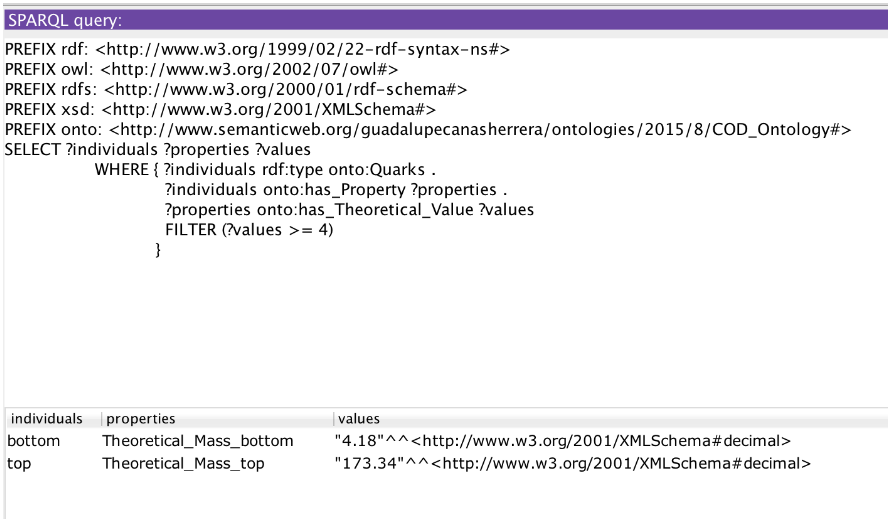
Figura 2: Screen Capture of Protege App at the individuals tab view for the High Energy Physics Ontology

4.3. SPARQL Use Cases

Once the Ontology was designed in Protégé, I checked how it works through SPARQL queries. SPARQL protocol is a programming language precisely thought for asking information to RDF triplets.

4.3.1. Ask to the Standard Model Class

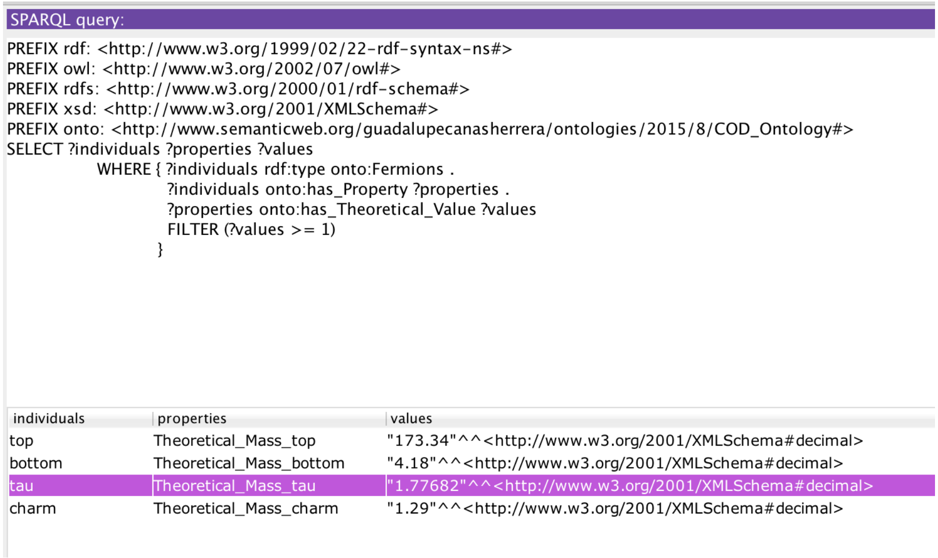
These queries have been posed to the Ontology main class based on the Standard Model



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://www.semanticweb.org/guadalupecanasherrera/ontologies/2015/8/COD_Ontology#>
SELECT ?individuals ?properties ?values
        WHERE { ?individuals rdf:type onto:Quarks .
                ?individuals onto:has_Property ?properties .
                ?properties onto:has_Theoretical_Value ?values
                FILTER (?values >= 4)
              }
```

```
individuals | properties           | values
bottom        Theoretical_Mass_bottom   "4.18"^^<http://www.w3.org/2001/XMLSchema#decimal>
top           Theoretical_Mass_top      "173.34"^^<http://www.w3.org/2001/XMLSchema#decimal>
```

Figura 3: Example of SPARQL Query filtering possible Quarks' theoretical values.



SPARQL query:

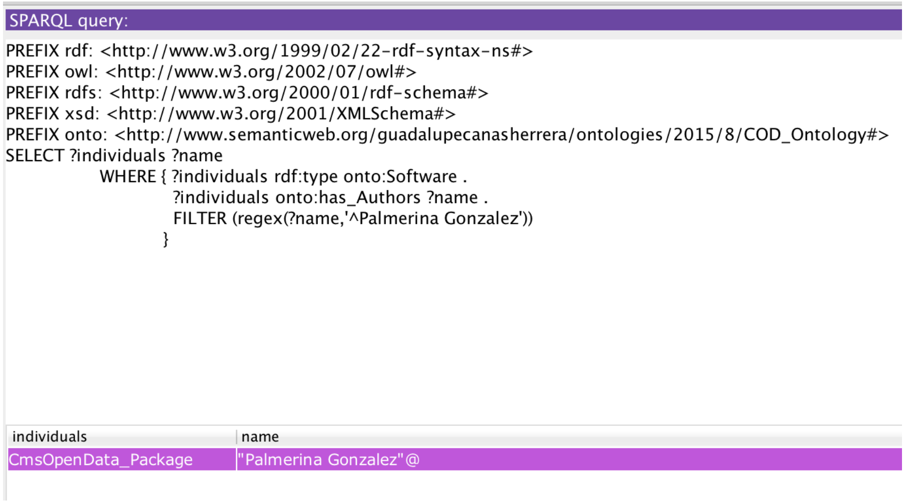
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://www.semanticweb.org/guadalupecanasherrera/ontologies/2015/8/COD_Ontology#>
SELECT ?individuals ?properties ?values
        WHERE { ?individuals rdf:type onto:Fermions .
                ?individuals onto:has_Property ?properties .
                ?properties onto:has_Theoretical_Value ?values
                FILTER (?values >= 1)
              }
```

individuals	properties	values
top	Theoretical_Mass_top	"173.34"^^<http://www.w3.org/2001/XMLSchema#decimal>
bottom	Theoretical_Mass_bottom	"4.18"^^<http://www.w3.org/2001/XMLSchema#decimal>
tau	Theoretical_Mass_tau	"1.77682"^^<http://www.w3.org/2001/XMLSchema#decimal>
charm	Theoretical_Mass_charm	"1.29"^^<http://www.w3.org/2001/XMLSchema#decimal>

Figura 4: Example of SPARQL Query filtering possible Fermions' theoretical values.

4.3.2. Ask to the Documentation Class

This query is related to the Documentation Class



SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://www.semanticweb.org/guadalupecanasherrera/ontologies/2015/8/COD_Ontology#>
SELECT ?individuals ?name
        WHERE { ?individuals rdf:type onto:Software .
                ?individuals onto:has_Authors ?name .
                FILTER (regex(?name,'^Palmerina Gonzalez'))
              }
```

individuals	name
CmsOpenData_Package	"Palmerina Gonzalez"@

Figura 5: Example of SPARQL Query filtering the name of the Software Package Author.

4.4. User's Guide

During the development of the whole project (uploading the existing portal with CMS Open Data from 2010 at IFCA), Palmerina González and I started to write a User's Guide indicated to Bachelor Students in Physics in order to explain how the portal works, and what it is behind the software and the ontology. The main contents of the User's Guide is the list of requirements in order to understand and analyse data, the explanation of how to access to the Virtual Machine through an SSH protocol (needed to create public keys), the SPARQL queries examples and the structure of the Analysis Software.

The guide is currently under construction together with the virtual portal, waiting to solve diverse issues before its actualization.

Introduction to CMS Open Data Analysis

Guide of User

G. Cañas / P. González

3. Requirements

CMS Data Analysis is very complex due to the large set of information collected in every collision. Therefore, it is necessary to use computers and software tools able to make theoretical and numerical operations easier. A good knowledge in Computing is highly recommended.

3.1 What you know

These are some explanations of basic concepts that most phyisicist have studied previously during some programming courses. In case you have studied in *Universiry of Canabria*, you can skip this chapter as these notions have been already tought in the *Programación* course. In case you are not familiar with the following topics, please take your time to be confident with them before keeping reading.

3.1.1 What an Operating System is

An Operating System (OS) is a computer software designed to operate and control the computer hardware and to provide a platform for running application software.

Nowadays, most people are *Microsoft Windows* users, which is a MS-DOS Operating System developed by Microsoft. However, when you are being introducted in science analysis, you realize that a great amount of scientists use UNIX based Operating Systems. The Unix-like family is a diverse group of operating systems with several major sub-categories including the famous *Linux* and its main distributions as *Ubuntu* or *Debian*. Apple's OS X Operating System is also Unix-like.

If the reader has never been in contact with an UNIX-based OS, starting to have a look at it is suggested. Most UNIX distributions are completely free and can be easily installed in every personal computer. It is convenient to remark that, for a non-experienced computer user, the main difference between Microsoft Windows and Ubuntu is the syntax of the commands written in the terminal (see next section).

Figura 6: Picture of the User's Guide main age and one of its section.

4.5. CMS Preservation Group Workshop

The 14th October, Palmerina González and I were invited to present the results of the project to the CMS Preservation Group Workshop that took place at CERN. We were connected to the Workshop through teleconference in order to analyse what should be the best scenario to conclude our project. Moreover, we tried to figure out how to work out diverse issues that appeared during the process (see section 5 for further information about this topic).

At this Workshop I had the opportunity to hear DASPOS group presentation, finding out new clues perfect to be included into the ontology project.

The presentation slides as well as the currently ontology code are at the github repository `https://github.com/gcanasherrera/CmsOpenData_Ontology/`.

5. Problems and Solutions

In the course of the project, I had to deal with several issues. In this section, these problems are explained and further solutions are presented.

5.1. Move the Ontology to a Database

As it has already been mentioned, the ontology was conceived and designed using *Protégé*. This software is linked to its own database in order to provide the different OWL and RDF-s schemas to query the ontology in SPARQL. However, as the main purpose of this ontology is its didactic aim, the ontology should be move from *Protégé* to a database able to save information in the form of triplets.

Several databases were taken in account, including some graph databases. Yet, only one those found databases was already prepare to save triplets and included the possibility of asking in SPARQL to the ontology. This database is free and open source, and its name is *Stardog*. The implementation of the ontology in Stardog was performed without any trouble, although some new issues appeared when trying to connect this database to the open portal.

5.2. Future Database Connection

The database including the Ontology has to be connected to the Open Portal at IFCA to make the Bachelor Students work with them without going through *Protégé*. Nevertheless, this step seemed to be more tough that it was thought before.

There are no previous examples connecting triplets database with an Ontology into a web-portal suitable to be ask in SPARQL. Furthermore, the portal at IFCA was designed using a open source Python tool called Django, which is a ”web framework that encourages rapid development and clean, pragmatic design so that you can focus on writing your web app without needing to reinvent the wheel”. Django allows you to connect your app forms into a database, and include some already-written codes to perform this procedure with some of the most famous database as MySQL. Still, there are no written patch to connect Django to Stardog.

As my academic track is no related with Computer Science and I lack diverse knowledge in web design, I had some difficulties to work our this problem. No solution has been provided for the moment.

5.3. Avoid to ask in SPARQL

The ontology has to be didactic and user-friendly, and these two principles do not get along with the use of SPARQL. This query language is messy and complicated for student who has no previous knowledge in RDF documentation, and explain SPARQL to those student in order to use the ontology miss this didactic and user-friendly characteristics that we want to achieve. For all these reasons, we should be able to suppress the use of SPARQL to query the ontology.

If SPARQL is avoided, we may create a question form suitable to be fulfilled in order to obtain information from the ontology. Nevertheless, there are a variety of problems in this process:

1. We first should design a form that contains all required fields to obtain information from the ontology, so that we should first think about the use cases the student may give to the ontology.

2. RDF documents only provide their information under SPARQL queries. Therefore, the information introduced to the form should be translated into SPARQL in order to ask the ontology at the end. This translation has to be perform through a program that takes the information from the form field and fulfils SPARQL queries instead.

3. In the hypothetical case that the SPARQL query successes, we have to think how we want to give the information back to the students (using tables, plots...).

4. We need to program the form in Django through html language or JAVA Script.

At the moment, we are still thinking the use cases in SPARQL in order to provide them to a Computer Sciences Group belonging to another project to be able to connect the Stardog database and the Portal.

6. Contribution to learning

Thanks to this internship I incremented my knowledge in several disciplines such as Particle Physics and Computer Science.

6.1. Computer Sciences

1. **Protégé Software**: I learned how to use Protégé at user level and its main features.

2. **RDF and SPARQL Languages**: I am able to understand RDF and SPARQL syntax, and I can perform complex queries to the ontology.

3. **Databases**: I learned what database are and their main types.

4. **Python Language with a focus on Django**: I improved my knowledge of Python including the new package Django for web-design.

5. **Latex Documentation**: I improved my Latex skills in order to write the user's guide.

6. **Github**: I progressed in Github commanding and repositories control.

6.2. Particle Physics

1. **Analysis Magnitudes**: I am able to understand the different magnitudes that take part in the analysis procedure of particle collisions, improving my knowledge of the particle physics course I took last semester.

2. **Cuts and restrictions**: I can preform quality cuts to the magnitudes in order to analyse a goal particle from candidates ones.

6.3. Others

During my internship I had also the chance of practising my English as most of the bibliography and the Preservation Workshop were in this language. Besides, I was able to learn how the Scientific method is applied in High Energy Physics Groups and how we should work when we take part in a project. Moreover, I could attend to meetings and conferences related with the project topic improvement my knowledge in this field.

7. Assessment of the internship

In general, I consider that the internship had a great benefit not only my student career but in my research experience. As an intern, I had now previous experience that would facilitate me the election of a master and its track in the future. I could obtain also experience in Computer Science as IT, which is always helpful for physicists and the enterprise sector.

I has also contact with students who pursue their PhDs, and this make me realize that their PhDs tasks do not defer from my internship ones (except in the difficulty level). The internship allowed me to have the experience of how pursuing a PhD would be in the next years.

I consider that the internship was well organized and my internship supervisor and my academic one helped me in everything they could. However, as the ontology topic is not currently a research field at the IFCA, I had to perform a long lasting bibliographical research. I would appreciate further help in the ontology implementation, although this lack did not discourage in my work but inspirit me in making a greater effort.

8. Suggestions for improvement

If I had to point out just one improvement in the internship period would be to perform more meetings of the project's members in order to organize and plan properly the development of the work and emphasize what the milestone are.