Dipartimento di Ingegneria e Scienza dell'Informazione

Knowledge and Data Integration

Project Report - Education in Trentino

Document Data:	Reference Persons:
January 12, 2022	Ali Hamza,Tecla Venturelli

Index:

Revision History:

Revision	**Date**	**Author**	**Description of Changes**
0.1	10.11.2021	Ali Hamza	Document created
0.2	13.11.2021	Ali Hamza	Inception Phase Report
0.3	17.11.2021	Tecla Venturelli	Informal Modeling Phase Report
0.4	04.12.2021	Tecla Venturelli	Formal Modeling Phase Report
0.5	10.01.2022	Ali Hamza	Data Integration
0.6	10.01.2022	Ali Hamza	Open Issues
0.7	10.01.2022	Tecla Venturelli	Exploitation

1 Introduction

Reusability has a vital role in the Data Integration process defined by the ITelos methodology. Detailed documentation that contains description of the steps taken along the process of data integration make huge difference in the reusability of the data resources produced as a result and also resources handled in the process. This report contains brief explanation of methodology used for project's purpose formalization including details of personas and scenarios leading to the competency questions, resource collection and categorization into the reusability categories , definition of the knowledge layer that includes collection of the teleologies needed and comprehensive description of the activities done in the inception phase.

2 Purpose and project's resources

2.1 Purpose Formalization

The first and the most important step of the data integration process and inception phase is formalization of purpose and domain of the project. Final outcome of the data integration process strongly depends upon clarity of the project purpose. We followed iTelos Data Integration methodology that since it is lead by purpose and also know as Purpose Driven Data Integration Methodology. Purpose formulation basically involves definition of Domain of Interest, Personas and Scenarios involving the personas in the domain of the project.

2.1.1 Project Purpose

we had a closer look at the data available across all available data sources. After knowing the nature of the available data, initial purpose of the project was defined as follows:

"A service that will help parent and student to find schools, including details about the school and courses offered, in the region of Trentino based on city, commune, school type, course duration and teaching activities schedules."

2.1.2 Project Domain Composition

Composition of domain is the first step that was taken while defining the purpose of the project since it might combine different domains for a new composed domain leading towards a specific purpose. In case of this project, education facilities in Trentino was composed domain involving geo-Space domain and time information domain.
Since the education in Trentino was a very broad domain of interest, it was necessary to set the boundaries within which information of the project would be considered. Education in Trentino involves all kind of education institutes like kindergarten, primary schools, high schools universities and other vocational training institutes, belonging to both public and private sectors, where each entity can have an infinite set of properties. Leaving the domain of the project wide open could have led us to never ending process of data collection and integration.

2.1.3 Personas

Personas are the possible actors that are supposed to exploit the final result of the data integration project. This section of the report list different kind of users that will be acting with the systems for different needs. Looking for schools in the region of Trentino is the common need of all type of users of the systems, despite having different factors that might lead to different result for each of them. For example one person could be looking for all schools in Trento city while other could be in search of all kindergarten schools that teach at day time. But in case of both, underlying purpose is similar.

Alex is 14 years old. He lives in the Trento city. He just finished his middle school and interested in taking admission in a science high school. Despite being native to the area, Alex does not know much about the high schools in the Trento city or nearby cities. So it is very necessary for him to search high schools in order to get admission. Since he lives in Trento so he would like to look for high schools that are in the same city. Alex has passion for science and technology, so he would like to search for only science school.

Chloe is 16 years old French girl who moved to Trento with her parents. She has passion for art, especially painting. To support her studies and hobbies related to her passion of painting, she started working part-time in a flower shop. She works 4 hours a day in the morning. She wants to take admission in an arts school where teaching activities are conducted only in the afternoon because she does not want to quit her job. In addition to that, she is in search of arts schools that are strictly in the city of Trento because she does not have time to travel to other town or city after the work.

Ali is 30 years old man. He just moved to the city of Rovereto with his wife. He is father of a 5 years old boy. He is not native Italian so he does not know anything about the city or the region. He is looking for a kindergarten school where he can admit his son. Since Ali runs his own business and also own a car, so he is looking for good kindergarten schools in the city of Rovereto and other nearby cities. He will prefer to send his son to a school where teaching is done during day time but also doesn't mind otherwise.

David is 35 years old man who lives in Switzerland. He works as a data scientist with the European Education Foundation. His job is to collect statistics related the educational activities conducted in the region of Trentino. He has been assigned a task to get details all the schools in the region of Trentino where at least one course has duration of 2 years. He wants name of the school, commune where the school is located, any contact information of the school and list of the courses that fall under his search criteria.

Mario is 47 years old man. He is a native Italian and lives in Bolzano. He is a government personal working in the Italian education department. His job is to allocate educational funds and build schools in the areas where no schools exist already. He has been given list of all cities in the region of Trentino. He needs a list of cities where more at least 5 schools exist which he with cross match with the list he already has to look for the cities where the work is need to be done.

Alina is 25 years old German girl. She is an Erasmus student in Trento. She is doing an internship with an organization that collaborates with the Italian education department. The organization works on improvement in the educational sector. Updates in the teaching methodologies are communicated to all schools on telephonic calls. Alina has been assigned a task to compose a list of all schools along with contact information especially phone number. She wants details of all kind of schools that operate in the Trentino region.

Mattia is a 22-year-old university student that needs a language certificate to leave for Erasmus. He has to obtain the certification as soon as possible, otherwise he cannot leave.

Marc in Rovereto wants to evaluate the quality of the education in the schools of the comune comparing the percentage of people that successfully completed an academic year in 2021 with respect to the previous year.

Chiara has a disabled child; he is 14 and he is about to start the high school. Chiara wants to know if there are in Trento schools that provide study plans for people with special needs and also with the appropriate infrastructure.

Sandro needs a mechanics high school degree to open his own business. He is a worker so he can attend the lecture only in the evening. He needs the degree as soon as possible so he wants to know how many worker-students graduated in time in the past 2 years in part time schools of mechanics, so that he can choose the school and has more probability to get the degree in time.

2.1.4 Scenarios

Below is the list of some of the most important possible scenarios how the mentioned personas could possibly interact with the outcome of this DI project:

1. A person living in Trento who just finished his middle school and is passionate about science. Would like to look for high schools that are in the Trento city, has passion for science and technology so would like to search for only science school.

2. A person who moved to city of Trento with parents, works part-time in a flower shop in the morning, interested in Arts, wants to take admission in an arts school where teaching activities are conducted only in the afternoon because does not want to quit the job. In addition to that, in search of arts schools that are strictly in the city of Trento because does not have time to travel to other town or city after the work.

3. A non-native Italian who just moved to Rovereto city with spouse and a 5 years old son, owns a business and a car, looking for good kindergarten schools in the city of Rovereto and other nearby cities, will prefer to send the son to a school where teaching is done during day time but also doesn't mind otherwise.

4. A person who lives in Switzerland, Works as data scientist with the European education foundation, working on statistics related to the education activities in Trentino and has been assigned a task to get details all the schools in the region of Trentino where at least one course has duration of 2 years, wants name of the school, commune where the school is located, any contact information of the school and list of the courses that fall under his search criteria.

5. A native Italian person working with the education department of Italy, job is to allocated funds to areas where there are no school, has been given list of all cities in the region of Trentino, needs a list of cities where more at least 5 schools exist which he with cross match with the already existing list to look for the cities where the work is need to be done.

6. A perons who is an Erasmus student in Trento, Works with a private organization in education sector, has been assigned a task to compose a list of all schools along with contact information especially phone number, wants details of all kind of schools that operate in the Trentino region.

7. A university student that needs a language certificate to leave for Erasmus. has to obtain the certification as soon as possible, otherwise cannot leave, searching for an English school to start the lecture as soon as possible.

8. A person wants to evaluate the quality of the education in the schools of the comune of Rovereto so needs to know the percentage of people that successfully completed an academic year in 2021 and the percentage of people that successfully completed an academic year in 2020 to compare the two.

9. A disabled child, who is about to start the high school and wants to know if there are schools in Trento that provide study plans for people with special needs and also with the appropriate infrastructure.

10. A person who needs a mechanics high school degree to open his own business, is a worker so can attend the lecture only in the evening. wants to choose the school in which the person has the highest probability to get the degree in time so wants to know how many worker-students graduated in time in the past 2 years in part time schools of mechanics.

2.2 knowledge Resources

Follwing were the reference teleologies initially collected to satisfy the purpose of the project along the integration process :

- https://schema.org/EducationalOrganization
- https://schema.org/School
- https://schema.org/ElementarySchool
- https://schema.org/HighSchool
- https://schema.org/MiddleSchool

- https://schema.org/Preschool
- https://schema.org/City
- https://schema.org/Person
- https://schema.org/Course

2.3 Data Resources

2.3.1 Data Collection

Collection of data that supports the purpose of the project was one of the challenging part of this work. OPENdata Trentino was use to collect the data sets related to the education data in Trentino. It is a web portal that contains huge amount of public datasets owned by Autonomous Province of Trento [1]. From the above mentioned portal, following datasets downloaded:

- First dataset that was downloaded is “Trentino educational institutions” [2]. The dataset is in XML format. It contains List of educational institutions in Trentino, with information such as, type of institution, site, training offer, contact information, address, school type, number and list of course offered. The dataset contains huge number of attributes that is the reason that we have listed only most important ones in the table below.

Dataset Attributes	
Attribute	**Description**
istituzione-scolastica:denominazione	Name of the educational institution
istituzione-scolastica:telefono	Telephone number of the institution
istituzione-scolastica:fax	Fax number of the institution
istituzione-scolastica:emailIstituzione	Email of the institution
istituzione-scolastica :responsabile:nominativo	Name of the Contact person of the educational institution
istituzione-scolastica :responsabile:ruolo:descrizione	Role of the Contact person of the educational institution
istituzione-scolastica:sede :descrizione	Address of the institution
istituzione-scolastica:sede :indirizzo	Street Addressof the institution
istituzione-scolastica:sede :comune:descrizione	Name of the comune where the institution is located
istituzione-scolastica:sede :comune:descrizione	Name of the comune where the institution is located

istituzione-scolastica :unita-didattica:denominazione	Name of the teaching unit
istituzione-scolastica:unita-didattica:tipo-scuola-riferimento	Type of the teaching unit
istituzione-scolastica:unita-didattica:tipologia-orario	Typology time (Daytime or Evening)
istituzione-scolastica:unita-didattica:tipologia-orario	Typology time (Daytime or Evening)
istituzione-scolastica:unita-didattica:offerta	Contains information of the courses being offered

- Second data set downloaded from the website was “Study courses of the Trentino schools” that contains list of study courses of Trentino schools with description and status of activity [3]. The dataset is in XML format. This dataset has extra information, like starting year and duration, about the study courses being taught at the school.

Dataset Attributes	
Attribute	**Description**
corso-studi:codiceOrigine	Code of the course
corso-studi:descrizione	Name of the study course
corso-studi:annoCorsoInizio	Starting academic year of the course
corso-studi:durataAnni	Duration of the course
corso-studi:attivo	Active/inactive status of the course

- Later after having look at the dataset, it was noticed that there were some cities data missing from the datasets. So list of all the communes in the province of Trento was manually downloaded from a website and cleaned [4].

Dataset Attributes	
Attribute	**Description**
ISTAT Code	Code of the municipality
Italian	Italian name of the commune
German	German name of the commune
Area	Area of commune KM squares

- Initially dataset of the degree programs offered university of Trento was also downloaded in the very initial phase but was later discarded during the purpose formalization phase. The reason was that there was not data available for the universities in the Trento region.

2.4 Metadata

This section of the document briefly describes metadata of the datasets collected during the dataset resource collection phase. It provided all necessary and meaningful infromation about the dataset.

2.4.1 Trentino educational institutions

Dataset Metadata	
Dataset Name	Trentino educational institutions
Identifier of the dataset	p_TN: 653ca277-0643-4834-badc-dc27ecc8e99e
Themes of the dataset	**Education, culture and sport** 3211 teaching 3206 education 3216 school organization
Dataset Publisher	**Name:** Vocational Training, Tertiary Training and System Functions Service **IPA / VAT Code:** W05601
Format	XML
Dataset Language	Italian
Release Date	10-09-2014
Modification Date	10-09-2014
Geographic coverage	Trento
URI of GeoNames	`http://www.geonames.org/3165241/`
Refresh Rate	Annual
Holder	**Name:**Autonomous Province of Trento
Author	**Name:** Vocational Training, Tertiary Training and System Functions Service
License	Creative Commons CCZero 1.0

2.4.2 Study courses of the Trentino schools

Dataset Metadata	
Dataset Name	Study courses of the Trentino schools
Identifier of the dataset	p_TN: 90e4dd5d-f967-4a0a-854a-a8bcef48114b
Themes of the dataset	**Education, culture and sport** 3211 teaching 3206 education 3216 school organization
Dataset Publisher	**Name:** Vocational Training, Tertiary Training and System Functions Service **IPA / VAT Code:** W05601
Format	XML
Dataset Language	Italian
Release Date	10-09-2014
Modification Date	10-09-2014
Geographic coverage	Trento
URI of GeoNames	`http://www.geonames.org/3165241/`

Refresh Rate	Annual
Holder	**Name:**Autonomous Province of Trento
Author	**Name:** Vocational Training, Tertiary Training and System Functions Service
License	Creative Commons CCZero 1.0

3 Inception

Inception phase is the very phase of the ITelos methodology. Below is the abstract level description of the activities performed in this phase:

- Purpose formalization.
- composition of competency questions from the purpose (CQs).
- List of datasets to support the purpose of the DI project.
- Selection of the datasets among the available datasets.
- Collection of knowledge resources.
- Classification of knowledge and data resources into the reusability categories.

Details of the activities performed including reasoning at each step is provided below.

3.1 Purpose Formalization

The first and the most important step of the data integration process and inception phase is formalization of purpose and domain of the project. Final outcome of the data integration process strongly depends upon clarity of the project purpose. We followed iTelos Data Integration methodology that since it is lead by purpose and also know as Purpose Driven Data Integration Methodology. Purpose formulation basically involves definition of Domain of Interest, Personas and Scenarios involving the personas in the domain of the project.

3.1.1 Domain Composition

Composition of domain is the first step that was taken while defining the purpose of the project since it might combine different domains for a new composed domain leading towards a specific purpose. In case of this project, education facilities in Trentino was composed domain involving geo-Space domain and time information domain.

3.1.2 Domain of Interest

Since the education in Trentino was a very broad domain of interest, it was necessary to set the boundaries within which information of the project would be considered. Education in Trentino

involves all kind of education institutes like kindergarten, primary schools, high schools universities and other vocational training institutes, belonging to both public and private sectors, where each entity can have an infinite set of properties. Leaving the domain of the project wide open could have led us to never ending process of data collection and integration.
In order to avoid such situation, we had a closer look at the data available across all available data sources. After knowing the nature of the available data, initial purpose of the project was defined as follows:
"A service that will help parent and student to find schools, including details about the school and courses offered, in the region of Trentino based on city, commune, school type, course duration and teaching activities schedules."

3.1.3 Inception Sheet

Following section contains brief description of the inception sheet. The sheet provides meaningful combination of personas and scenarios in which the DI project outcome can be exploited. Different competency question involving variety of use cases also categorise the concepts involved in each use case into different re-usability categories defined in the knowledge and data integration context. Table below lists all important competency questions:

Competency Questions					
Person	**Scenario**	**Competency Question**	**Common Concept**	**Core Concept**	**Contextual Concept**
Alex	1	Are there any schools in Trento?	Educational Institute, Location	School, City	Trento, address
Alex	1	How many science schools are in Trento?	Educational Institute, Location	School, City	Trento, Science, Address
Alex	1	Are there any high schools in Trento?	Educational Institute, Location	School, City	Trento, Science, Highschool
Alex	1	Where are science high schools in Trento?	Educational Institute, Location	School, City	Trento, Science
Chloe	2	Are there any arts schools in Trento?	Educational Institute, Location	School, City	Trento, Arts, Address
Chloe	2	Are there any schools where classes are conducted in the afternoon?	Educational Institute, Location	School, City	Trento, Science, Afternoon, Teaching

Chloe	2	Are there any art school in my via/block where teaching activities are carried out in the evening?	Educational Institute, Location	School, Residential Block	Trento, Science, Afternoon, Teaching
Ali	3	Are there any schools in Rovereto?	Educational Institute, Location	School, City	Rovereto
Ali	3	Can I have list of kindergarten schools in Rovereto?	Educational Institute, Location	School, City	Trento, Science, Afternoon, Teaching
Ali	3	Which are the kindergarten schools that teach during the day ?	Educational Institute, Location	School, City	Trento, kindergarten, Daytime, Classes Schedule
Ali	3	Can I have list of kindergarten schools?	Educational Institute, Location	School	Kindergarten
David	4	How many schools teach more than 1 study courses in Trentino?	Educational Institute, Location	Schools, Province	Trentino, multiple study courses
David	4	Are there any study courses that are equal or longer than 2 years in schools of ?	Educational Institute, Location	Schools, Province	Count of courses longer or equal to 2 years, Trentino
David	4	What is name and contact information of the schools in Trentino that have at least course with duration of 2 years ?	Educational Institute, Location	Schools, Province, Study course	Study course, duration, Trentino
Mario	5	Is there any commune in Trentino where there is no school at all ?	Location	Commune, City	Address,Trentino
Mario	5	List of cities where there are more than 5 schools in Trentino?	Educational Institute, Location	Cities, Schools	Count, Trentino

Alina	6	How many schools are in the Trentino region that have contact information?	Educational Institute, Location	Schools	Contact information, Trentino
Alina	6	What is the list of contact numbers of all the schools in Trentino?	Educational Institute, Location	Schools, Province	Trentino, Email, name, Phone number
Mattia	7	Are there any English schools in Trento?	Location, Educational Institute	School	English course
Mattia	7	When do the courses start?		School	Starting date
Mattia	7	How long do the courses last?	Educational Institute	School	Duration
Mattia	7	How much does the registration cost?	Educational Institute	School	Registration cost
Marc	8	What are the schools in Rovereto?	Educational Institute	School	
Marc	8	Number of students graduated in time in 2020?	Educational institute, Person	School, Academic year	Graduation date
Marc	8	Total number of students studying in 2020 in schools of Rovereto?	Educational Institute	School student	enrollment status
Marc	8	Number of students graduated in time in 2020 in schools of Rovereto?	Educational institute, Person	School,student	Total number of students
Chiara	9	Are there schools in Trento providing study plans for people with special needs?	Educational Institute	School	level of education, infrastructure
Chiara	9	Are there schools in Trento with the requested infrastructures? Recently modernized	Educational institute	School	last renovation date

Sandro	10	Are there any Mechanics schools in Trento?	Educational Institute, Place	School	Mechanics Schools
Sandro	10	Are there any part time schools in Trento?	Educational institute	School	part-time education

3.2 Data and knowledge resource collection

In the previours phase, research on the datasets was completed and a list of related datasets was created. Most of the datasets were downloaded from Data Trentino[1] website. Since the number of datasets was not huge so no automation or script was done and all of the datasets were downloaded manually.

One of the datasets was in a raw format. That dataset contained information of all municipalities in Trentino and it was manually exported from Wikipedia[4]. Later this data was cleaned and formatted as CSV.

Schema.org was used as reference to collect all the knowledge resources. Initially all knowledge resources that could potentially be related to the purpose of the project were list. The list was later refined keeping only necessary knowledge resources.

3.2.1 Data Input Heterogeneity

Real life data does not come in a single pre-cleaned and ready to use format. If often requires a lot of preprocessing in order to make is useful and ready for integration. Since data was downloaded more than one data sources, so it was in different formats. Some of the datasets were chosen wisely to avoid hassle for conversion of data to a uniform format. The iTelos input format enforces the datasets to be integrated to have one of the below mentioned format:

- Knowledge Resources:
  - RDF-OWL
- Data Resources:
  - CSV
  - Excel Spreadsheet
  - JSON
  - XML

3.2.2 Data Transformation Activity - DTA1

So keeping in mind the above mentioned format, data transformation activity was necessary to be performed in our case. Datasets downloaded from OPENdata Trentino portal were both in the XML format which is supported by ITelos methodology. While the Dataset of communes in the province of Trento that was manually fetch from a web page was in an excel file. This data was cleaned manually to remove unnecessary commas and special characters. After cleaning of the datasets, it was converted to CSV (comma separated values) format that is also accepted by the ITelos supported input format.

3.3 Inception Phase Evaluation

In the inception phase, evaluation was done on both schema level and data level separately. Coverage and extensiveness was calculated for knowledge and data resources.

3.3.1 Schema Level Evaluation

This section gives an overview of the Etypes and properties extracted from the CQs and collected knowledge resources prior to the calculations.

Classes/Etypes in Ontology: *{ EducationalOrganization, School, ElementarySchool, HighSchool, MiddleSchool, Preschool, City, Person, Course}*

Classes/Etypes in CQs: *{ EducationalOraganization, School, HighSchool, MiddleSchool, PrimarySchool, comune, Person, DisabledPerson, Student, StudyCourse, PublicSchool, PrivateSchool, HeadTeacher, Director, HeadManager, Coordinator}*

In case of properties, since there are so many entities in both collected resources and each of them have larger number of properties. So only School Etype was selected to compute the coverage and extensiveness of the properties.

Properties in Ontology [total 113]: *{ address, amenity, name, branchCode, openingHours, telephone,latitude, longitude, email, gender .... parentOrganization}*

Properties in CQs: *{ name, address, email, phone, numberOfStudent, latitude, longitude, coordinates, typology, fieldOfStudy, levelOfEducation, startingDate, timetable }*

Evaluation Matrix	
EType Coverage:	0.44
EType Extensiveness:	0.34
Properties Coverage:	0.53
Properties Extensiveness:	0.75

3.3.2 Data Level Evaluation

This section gives an overview of the Etypes and properties extracted from the CQs and collected data resources prior to the calculations.

Classes/Etypes in Datasets: *{ EducationalOrganization, School, Address, City, Person, Course}*

Classes/Etypes in CQs: *{ EducationalOraganization, School, HighSchool, MiddleSchool, PrimarySchool, comune, Person, DisabledPerson, Student, StudyCourse, PublicSchool, PrivateSchool, HeadTeacher, Director, HeadManager, Coordinator}*

In case of properties, since there are so many entities in both collected data resources and each of them have larger number of properties. So only School Etype was selected to compute the coverage and extensiveness of the properties.

Properties in Dataset: [total 10] *{ name, address, email, telephone, fax, responsiblePerson, coursesOffered, timetable, typeOfSchool, LevelOfEducation}*

Properties in CQs: [total 13] *{ name, address, email, phone, numberOfStudent, latitude, longitude, coordinates, typology, fieldOfStudy, levelOfEducation, startingDate, timetable }*

Evaluation Matrix	
EType Coverage:	0.37
EType Sparsity:	0.36
Properties Coverage:	0.53
Properties Sparsity:	0.57

4 Informal Modeling

Following section of the report describes all the step done in the informal modeling phase of the iTelos Methodology.

4.1 Purpose Formalization and Modeling Sheet Description

From the CQs formalization step we extracted following main ETypes: Educational institute, School/Teaching Unit, Comune/City, Study course and Person/Student. We observed a hierarchy among some of these entities, in particular, different types of school or also known as teaching unit are subcategories of Educational institute, and High school is a subcategory of School which means that these entities will inherit the attributes of their parent Etypes. In a similar way, a person can be in-charge of an educational institution as well as all schools under that institution. In the datasets we found also other type of educational institute that can be added in the phase of building the ER model. The following key attributes for each entity were extracted:

- Educational institute: name, phone number, e-mail, street address,comune/city name, type (public, private).
- School: name, name, phone number, e-mail, street address,comune/city name, type (public, private), level of education (it can be kindergarten, elementary, middle school, high school), timetable, study-courses-offered.
- Comune: name, code, address
- Study Course: code, name, duration, starting-academic-year, duration, active status
- Person: name, role, profession,in-charge-of.
- Student: class, levelOfEducation, fieldOfStudy, enrolledInCourse

4.2 ER Model Description

4.2.1 Schema Level EType and ERD Description

All of the entities that were identified during the previous phases from the competency questions and modeling sheet are divided into three categories. The common category contains all of the ETypes that are shared by the elements of the core category. ETypes that are necessary to answer the competency questions are categorised as core. Then some of the ETypes that are not strictly required to get an answer for competency question but can bring more completeness to it are categorised as contextual ETypes. Important ETypes from each category are list below.

Common ETypes

- **Person** : a common type entity describing any living human being in the world
  Attributes:
  - first name : first name of the person
  - last name : last name of the person
  - date of birth : date birth of the person
  - age : age of the person
- **Area** : a common type entity describing any area of a geographical place
  Attributes:
  - area : area of a place in KM square
- **Place** : a common type entity describing any geographical location on the map
  Attributes:
  - latitude: latitude coordinates of the place
  - longitude: longitude coordinates of the place
  - address : street address of the location/place

- **Educational Institute** : a core type entity describing any educational institute
  Attributes:
  - denominazione : name of the institution
  - telefono : phone number of the institution
  - email : email of the institution
  - fax : fax number of the institution
  - indirizzo : street address of the institution

Core ETypes

- **School** : a core type entity describing any general school
  Attributes:
  - denominazione : name of the school
  - telefono : phone number of the school
  - email : email of the school
  - fax : fax number of the school
  - indirizzo : street address of the school
  - tipo-scuola-riferimento : Type of the teaching unit
  - tipologia-orario : Typology time (Daytime or Evening)
- **Commune** : a core type entity describing any general commune inside a city or geographical area
  Attributes:
  - name : name of the commune
- **Student** : a core type entity describing any school student
  Attributes:
  - class : grade in which the student studies
  - levelOfEducation : level of education of the student
  - field : field of study (science,arts etc)
  - enrolledIn: study course where student is enrolled
  - studiesIn: school where the student studies

Contextual ETypes

- **Study Course** : a contextual type entity describing any general study course offered in educational institutes or schools
  Attributes:
  - descrizione : name and description of the school
  - annoCorsoInizio : starting academic year of the course

    - durataAnni : duration of the course
    - attivo : active/in-active status of the course

- **Head teacher** : a contextual entity to describe head teacher of an educational institution or school.
  It does not have any attributes.

- **Pedagogical Coordinator** : a contextual entity to describe pedagogical coordinator of an educational institution or school.
  It does not have any attributes.

- **Director** : a contextual entity to describe director of an educational institution or school.
  It does not have any attributes.

- **Head Manager** : a contextual entity to describe head manager of an educational institution or school.
  It does not have any attributes.

The above described entities were modeled using appropriate relation between the entities. Each entity has certain properties that explain the role and key characteristic of the it.The model is provided as Figure[1] in the following section.
Some of the entities that inherit their properties from parent are conned to the immediate parent using is-a relationship. For example a person can be student,so the student entity inherits the properties of the person concept. Following this concepts avoid over-complication of the ER model.
During the analysis and entity extraction phase, it was noticed that no entities can be related to each other in many-to-many and one-to-many relationship. So such entities are bound to each other using appropriate relationship. For and example, a course can be offered in many school so it is associated to the school EType with one-to-many relationship. In a similar way, a school can offer multiple study course hence EType school make one-to-many relationship with the Study Course EType.

4.3 Informal Modeling Evaluation

In the informal modeling phase, evaluation was done on both schema level and data level separately. Coverage and extensiveness was calculated for knowledge and data resources.

4.3.1 Schema Level Evaluation

This section gives an overview of the Etypes and properties extracted from the CQs and modeled in the ER model, prior to the calculations.

Classes/Etypes in ER Model [total 19]: *{ EducationalInstitute, School, Primary, HighSchool, MiddleSchool, Kindergarten, Student, City, Area, Place, Person, DisabledPerson, StudyCourse, HeadTeacher, Director, HeadManager, Coordinator, PrivateSchool, PublicSchool }*

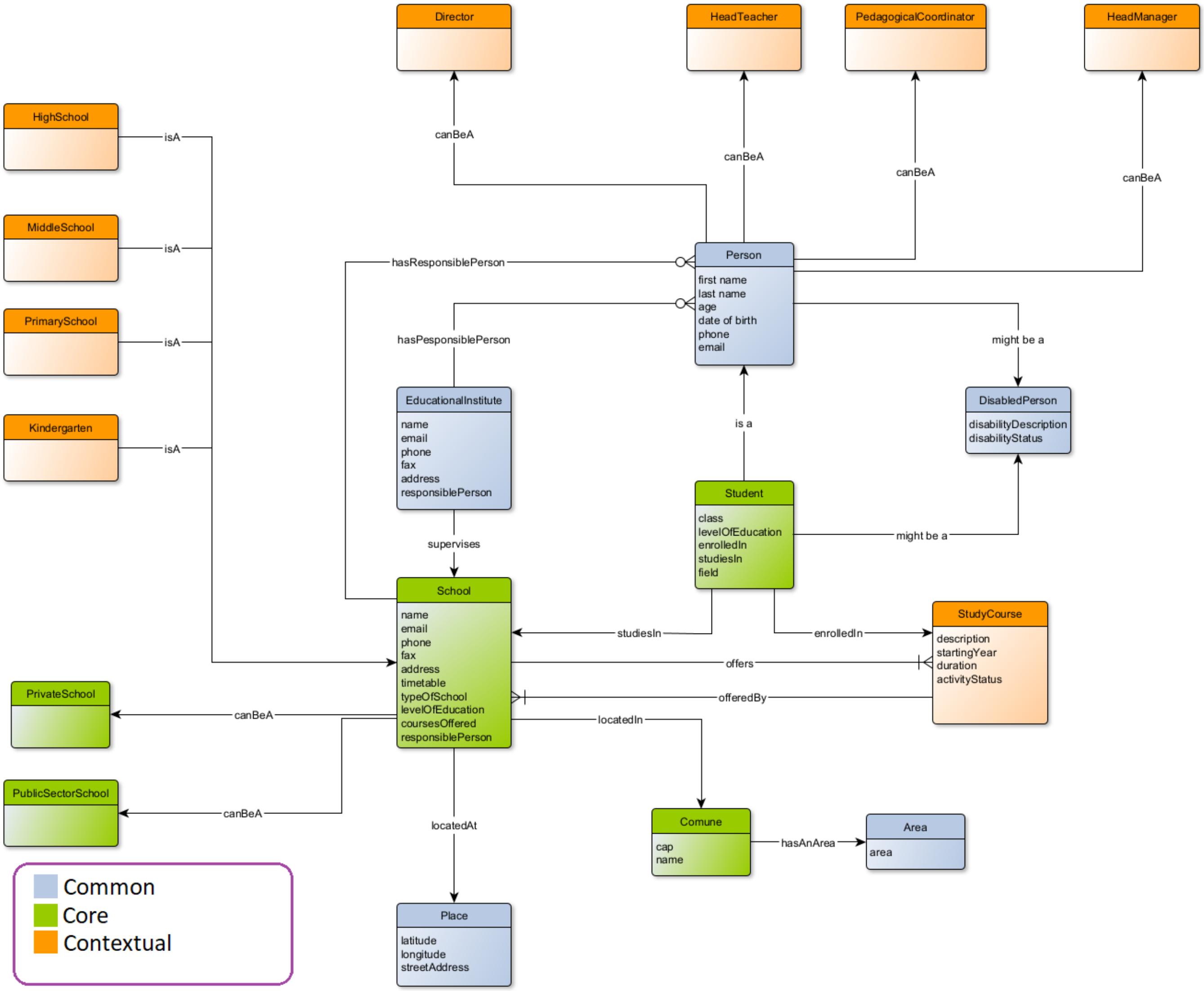


Figure 1: ER Model Diagram

Classes/Etypes in CQs [total 11]: *{ EducationalOraganization, School, HighSchool, HighSchool, comune, Person, DisabledPerson, Student, StudyCourse, PublicSchool, PrivateSchool}*

In case of properties, since there are so many entities in both collected resources and each of them have larger number of properties. So only School Etype was selected to compute the coverage and extensiveness of the properties.

Properties in ER Model [total 10]: *{ name, email, fax, phone, address, timetable, typeOfSchool, LevelOfEducation, coursesOffered, responsiblePerson}*

Properties in CQs: *{ name, address, email, phone, numberOfStudent, latitude, longitude, coordinates, typology, levelOfEducation, timetable }*

Evaluation Matrix	
EType Coverage:	0.81
EType Extensiveness:	0.42
Properties Coverage:	0.54
Properties Extensiveness:	0.05

4.3.2 Data Level Evaluation

This section gives an overview of the Etypes and properties extracted from the CQs and collected data resources prior to the calculations.

Classes/Etypes in Datasets: *{ EducationalOrganization, School, Address, City, Person, Course}*

Classes/Etypes in ER Model [total 19]: *{ EducationalInstitute, School, Primary, HighSchool, MiddleSchool, Kindergarten, Student, City, Area, Place, Person, DisabledPerson, StudyCourse, HeadTeacher, Director, HeadManager, Coordinator, PrivateSchool, PublicSchool }*

In case of properties, since there are so many entities in both collected data resources and each of them have larger number of properties. So only School Etype was selected to compute the coverage and extensiveness of the properties.

Properties in Dataset: [total 10] *{ name, address, email, telephone, fax, responsiblePerson, coursesOffered, timetable, typeOfSchool, LevelOfEducation}*

Properties in ER Model [total 10]: *{ name, email, fax, phone, address, timetable, typeOf-*

School, LevelOfEducation, coursesOffered, responsiblePerson}

Evaluation Matrix	
EType Coverage:	1.0
EType Sparsity:	0.68
Properties Coverage:	1.0
Properties Sparsity:	0.0

5 Formal Modeling

This part of the report explains the work done in the formal modeling phase. It contains description of the step taking during the phase and reasoning behind it. Evolution of the schema as well as the data-sets are explained in detail.

5.1 ETG Generation

During the ETG generation phase, some observations were made in the data-sets which led to some mandatory updates in the ER model generated in the informal modeling phase in order to create an ETG well aligned with the entities and the properties provided in the data-set.

5.1.1 ER Model Update

In the informal modeling phase, ER model was constructed from the concept and properties extracted by the competency questions. While creation of the ETG it was realised that some of the entities that are present in the ER model are not present in the data. So schema of project was updated in order to align it to the data-sets and also the ETG created. Following changes were made to the ER model:

- Private School and public School entities were removed from the ER model. Instead ”typeOfSchool” property was kept in the school entity to model the information that either a school is private or public.

- MiddleSchool, HighSchool, PrimarySchool and PreSchool entities were also removed from the ER model. Instead ”levelOfEducation” property was kept in the school entity to model the information about level of education provided by the school.

- HeadTeacher, HeadManager and PedagogicalCoordinator entities were also removed from the ER model. Instead only Director entity was kept to bind the information of responsible person for the school.

- DisabledPerson entity was totally removed from the ER model since the data-sets collected do not have information about such entity.

- latitude, longitude properties from the entity Place were also removed from the ER model since the data-sets collected do not have information about such entity.

- Area was also removed as the separate entity and placed under the comune entity as has_area property.

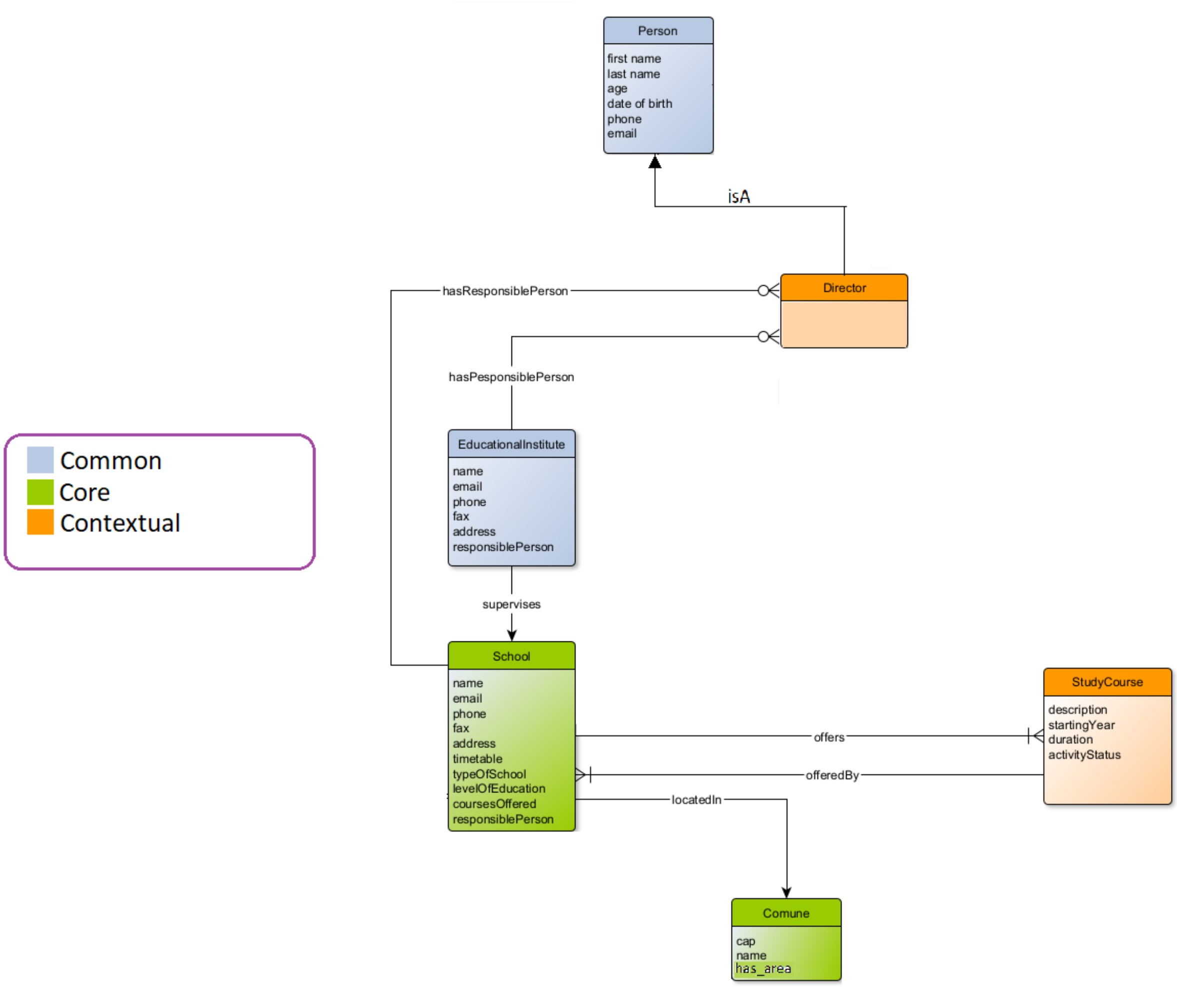
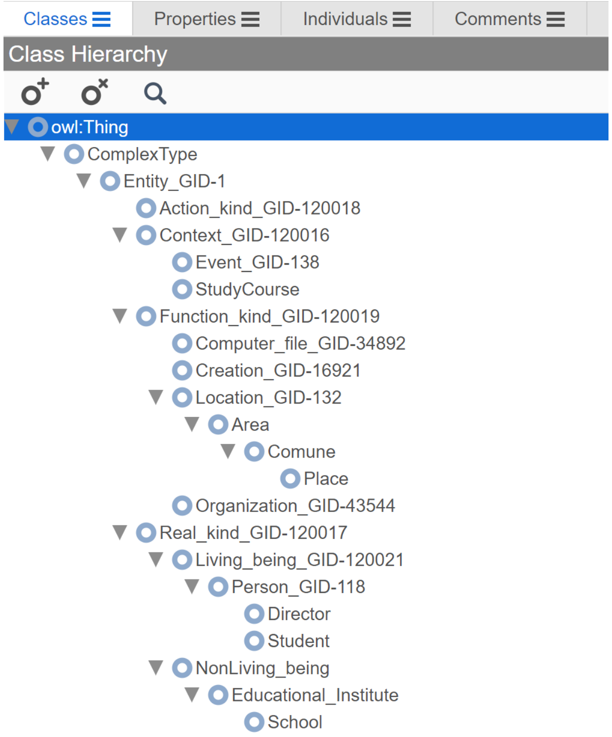


Figure 2: Updated ER Model Diagram

After the above mentioned updates, ER model look like Figure[2]. Having done the updates it the ER model, it was well aligned with the ETG and the information provided by the data-set.

5.1.2 ETG Generation

After the necessary updates in the ER model, ETG was generated in the Protege tool. First step taken was to search already existing ontologies that might be suitable for purpose of our project and align as much as possible with the ER model.
Classes were created for the entities not already present in the ontology. Visualization is given in Figure[3].



```
▼ owl:Thing
    ▼ ComplexType
        ▼ Entity_GID-1
            Action_kind_GID-120018
            ▼ Context_GID-120016
                Event_GID-138
                StudyCourse
            ▼ Function_kind_GID-120019
                Computer_file_GID-34892
                Creation_GID-16921
                ▼ Location_GID-132
                    ▼ Area
                        ▼ Comune
                            Place
                Organization_GID-43544
            ▼ Real_kind_GID-120017
                ▼ Living_being_GID-120021
                    ▼ Person_GID-118
                        Director
                        Student
                ▼ NonLiving_being
                    ▼ Educational_Institute
                        School
```

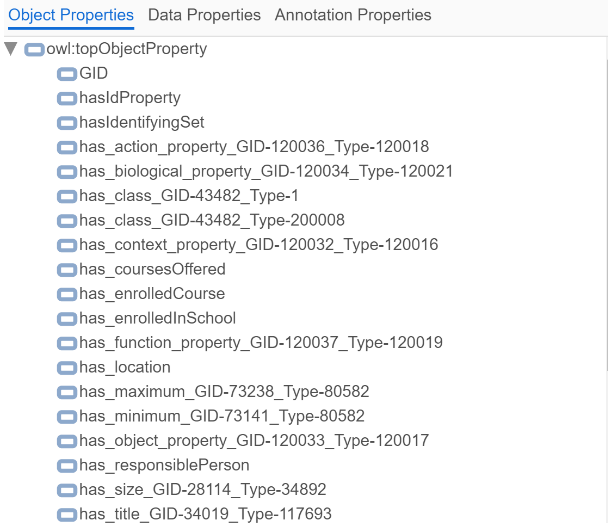
Figure 3: ETG - Entity Classes

In order to bind the classes together, necessary object properties were created. Visualization is given in Figure[4].

Based on the ER model, data properties were also created other then the ones already present in the classes. Visualization is given in Figure[5].
Following are the classes in the current ETG with set of data and object properties:

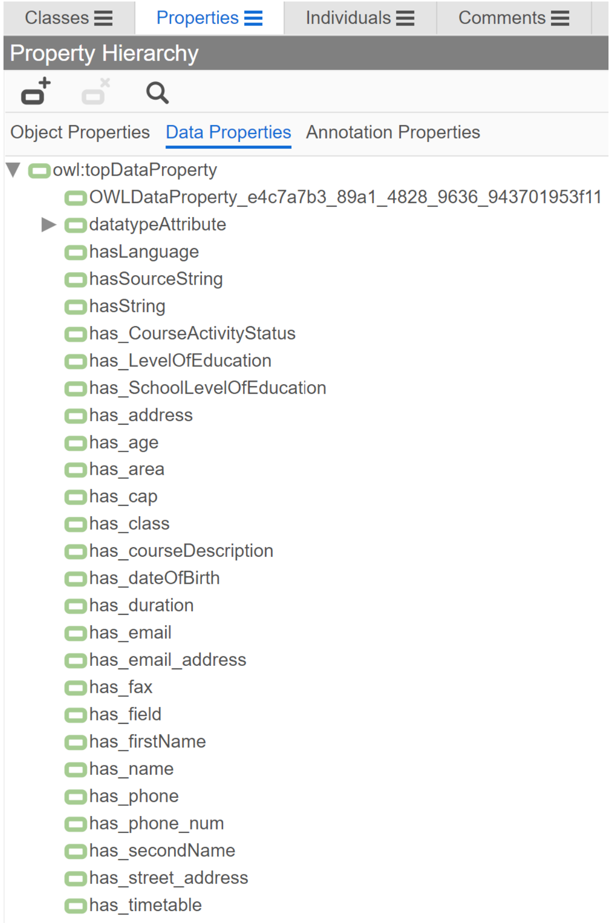
5.2 Data management (syntactic heterogeneity)

This section of the report explains the work done one the data level while managing the data in order to resolve syntactic heterogeneity. The activity was conducted at the data level over all the short listed and final selection of the data-sets.



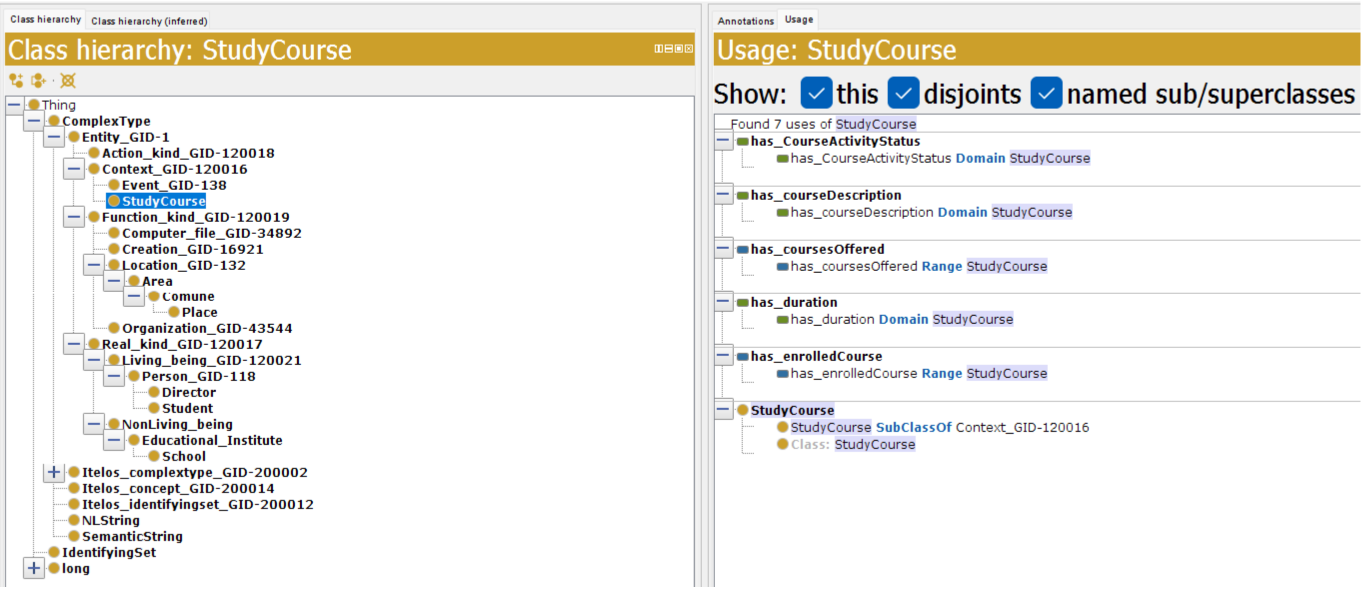
owl:topObjectProperty
GID
hasIdProperty
hasIdentifyingSet
has_action_property_GID-120036_Type-120018
has_biological_property_GID-120034_Type-120021
has_class_GID-43482_Type-1
has_class_GID-43482_Type-200008
has_context_property_GID-120032_Type-120016
has_coursesOffered
has_enrolledCourse
has_enrolledInSchool
has_function_property_GID-120037_Type-120019
has_location
has_maximum_GID-73238_Type-80582
has_minimum_GID-73141_Type-80582
has_object_property_GID-120033_Type-120017
has_responsiblePerson
has_size_GID-28114_Type-34892
has_title_GID-34019_Type-117693

Figure 4: ETG - Object Properties



owl:topDataProperty
OWLDataProperty_e4c7a7b3_89a1_4828_9636_943701953f11
datatypeAttribute
hasLanguage
hasSourceString
hasString
has_CourseActivityStatus
has_LevelOfEducation
has_SchoolLevelOfEducation
has_address
has_age
has_area
has_cap
has_class
has_courseDescription
has_dateOfBirth
has_duration
has_email
has_email_address
has_fax
has_field
has_firstName
has_name
has_phone
has_phone_num
has_secondName
has_street_address
has_timetable

Figure 5: ETG - Data Properties



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: StudyCourse

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

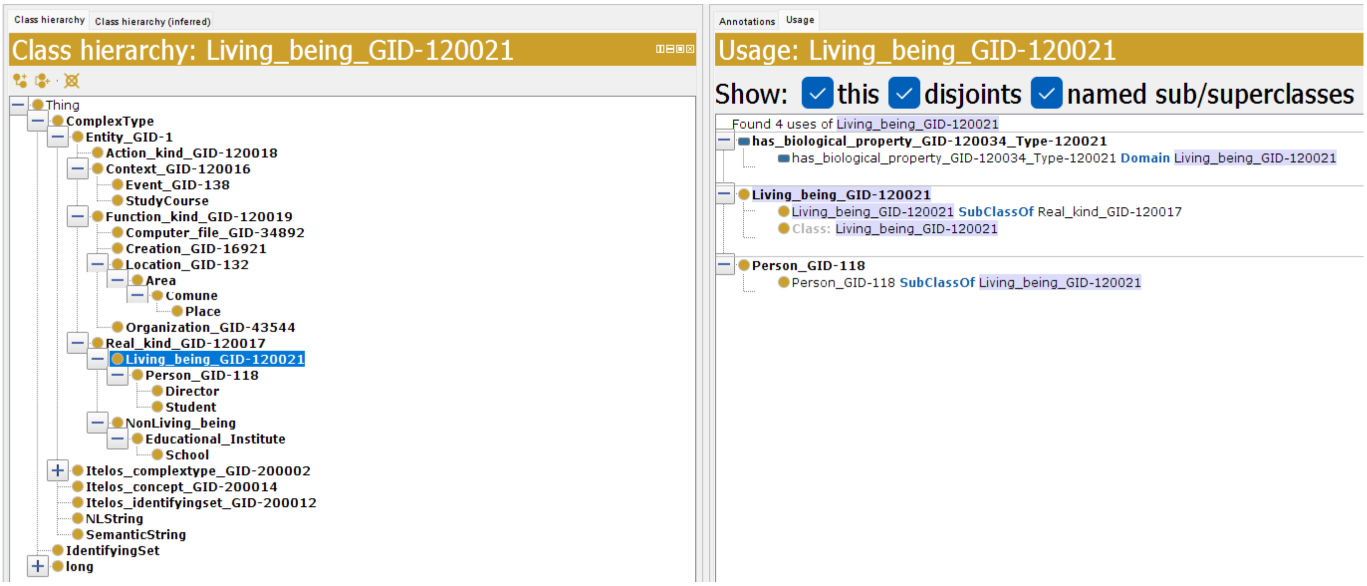
Annotations | Usage

Usage: StudyCourse

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 7 uses of StudyCourse
has_CourseActivityStatus
    has_CourseActivityStatus Domain StudyCourse
has_courseDescription
    has_courseDescription Domain StudyCourse
has_coursesOffered
    has_coursesOffered Range StudyCourse
has_duration
    has_duration Domain StudyCourse
has_enrolledCourse
    has_enrolledCourse Range StudyCourse
StudyCourse
    StudyCourse SubClassOf Context_GID-120016
    Class: StudyCourse
```

Figure 6: Study Course Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Living_being_GID-120021

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

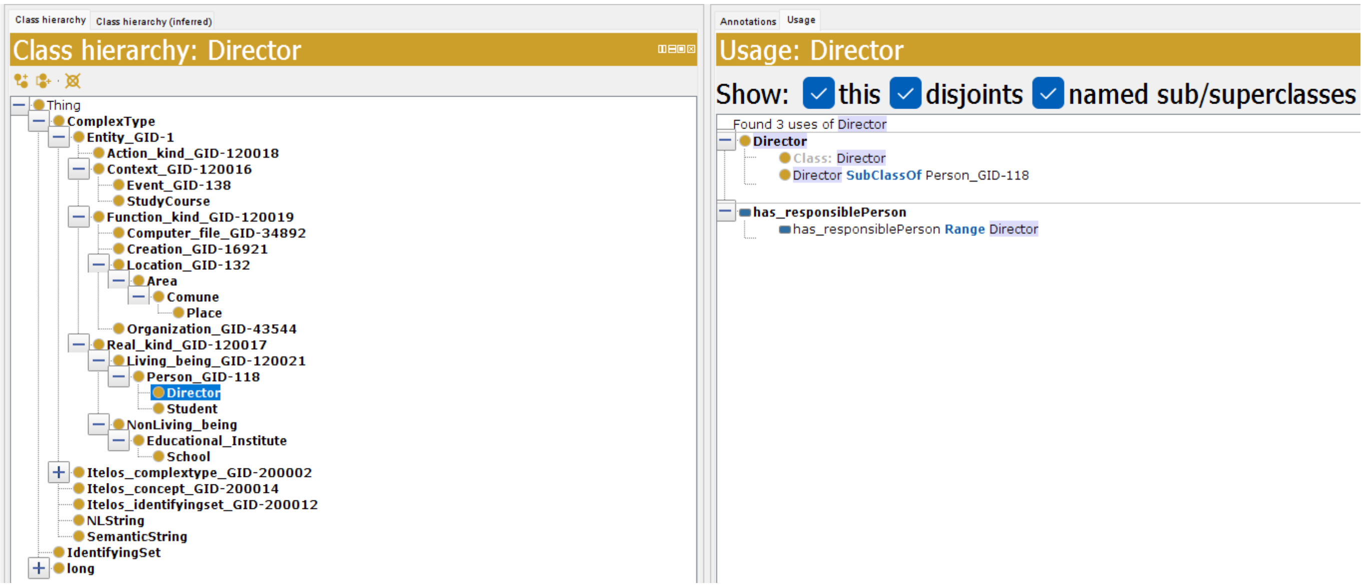
Annotations | Usage

Usage: Living_being_GID-120021

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 4 uses of Living_being_GID-120021
has_biological_property_GID-120034_Type-120021
    has_biological_property_GID-120034_Type-120021 Domain Living_being_GID-120021
Living_being_GID-120021
    Living_being_GID-120021 SubClassOf Real_kind_GID-120017
    Class: Living_being_GID-120021
Person_GID-118
    Person_GID-118 SubClassOf Living_being_GID-120021
```

Figure 7: Person Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Director

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

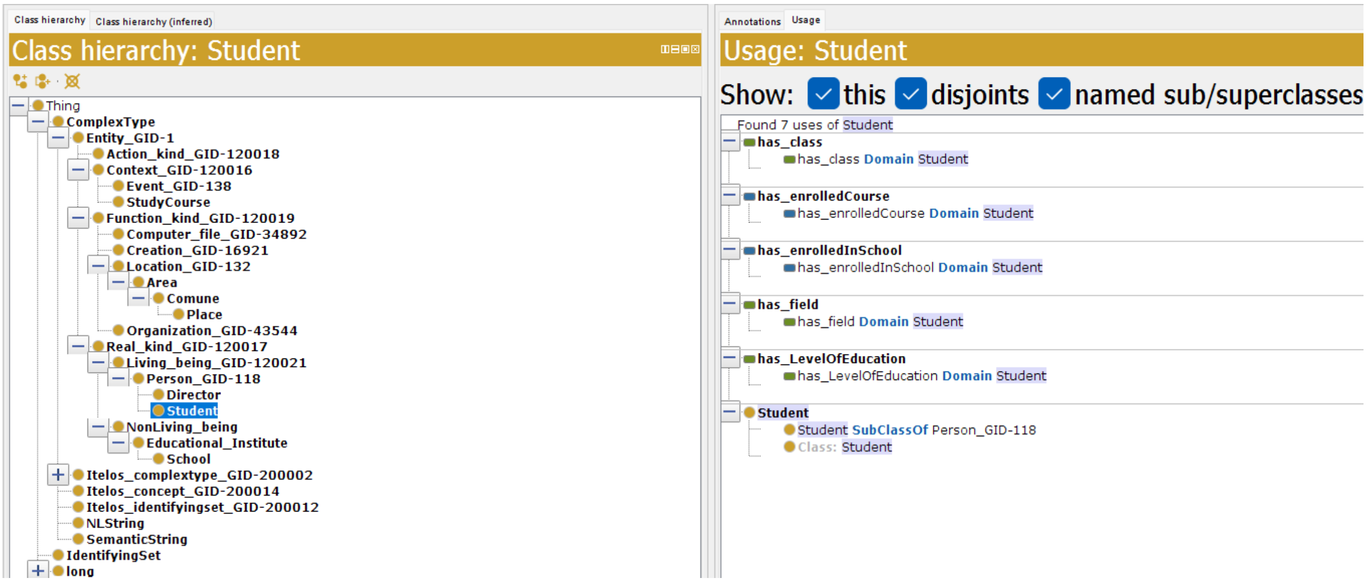
Annotations | Usage

Usage: Director

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 3 uses of Director
Director
    Class: Director
    Director SubClassOf Person_GID-118

has_responsiblePerson
    has_responsiblePerson Range Director
```

Figure 8: Director Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Student

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

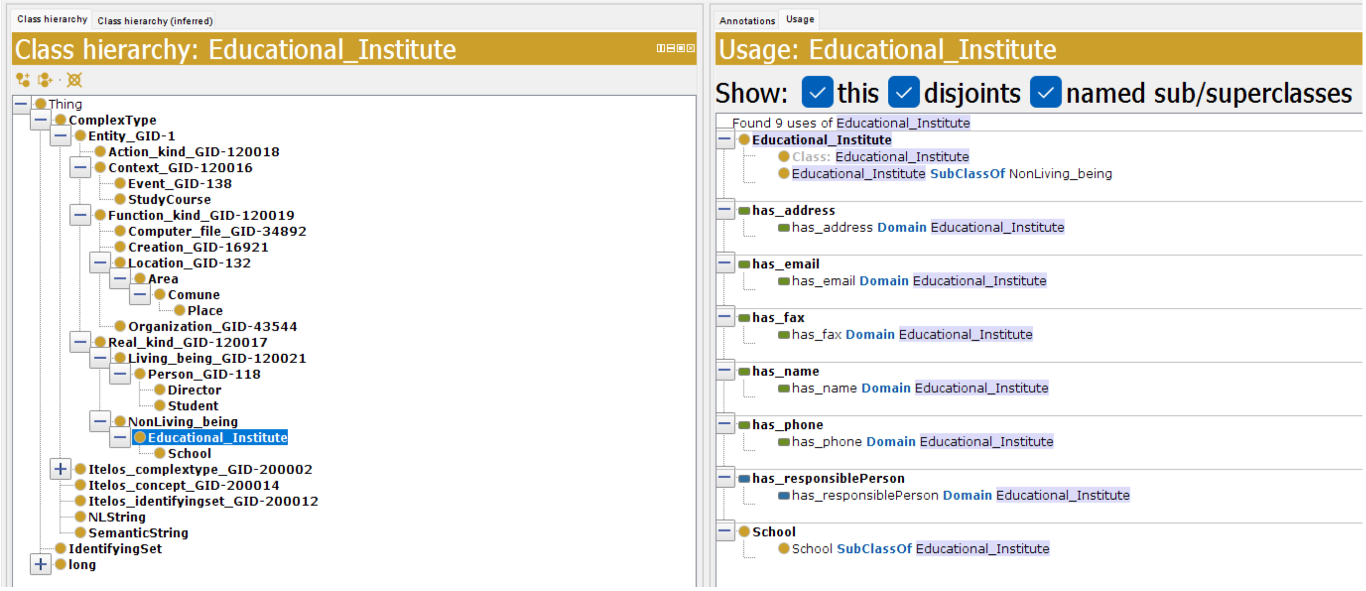
Annotations | Usage

Usage: Student

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 7 uses of Student
has_class
    has_class Domain Student

has_enrolledCourse
    has_enrolledCourse Domain Student

has_enrolledInSchool
    has_enrolledInSchool Domain Student

has_field
    has_field Domain Student

has_LevelOfEducation
    has_LevelOfEducation Domain Student

Student
    Student SubClassOf Person_GID-118
    Class: Student
```

Figure 9: Student Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Educational_Institute

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

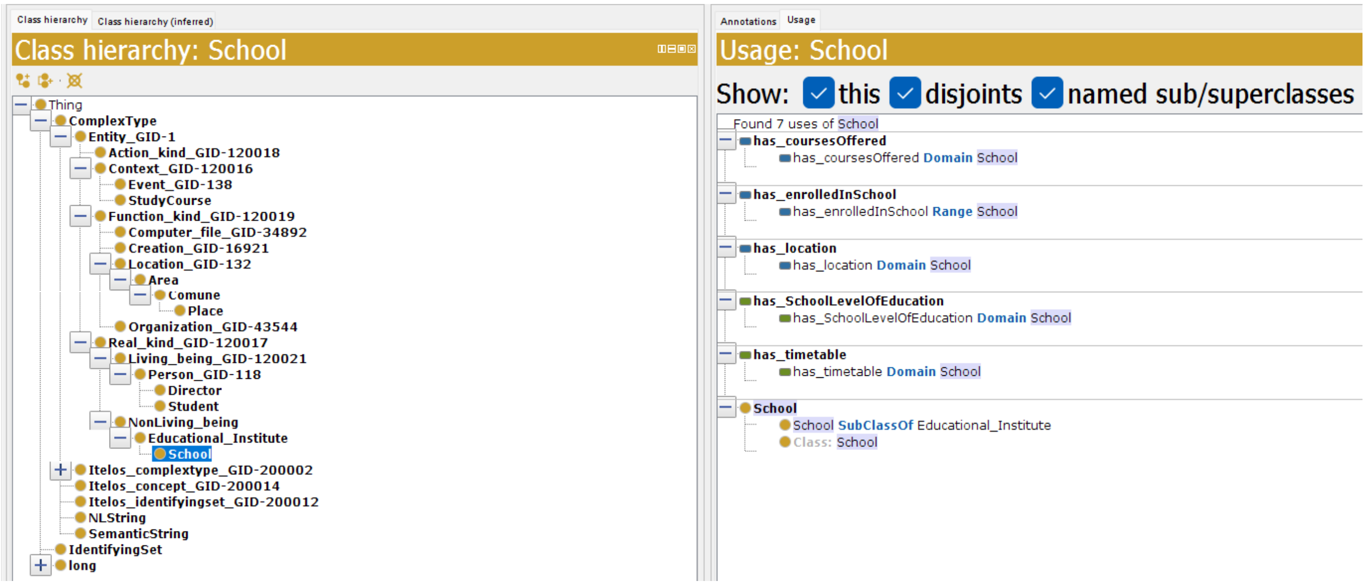
Annotations | Usage

Usage: Educational_Institute

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 9 uses of Educational_Institute
Educational_Institute
    Class: Educational_Institute
    Educational_Institute SubClassOf NonLiving_being
has_address
    has_address Domain Educational_Institute
has_email
    has_email Domain Educational_Institute
has_fax
    has_fax Domain Educational_Institute
has_name
    has_name Domain Educational_Institute
has_phone
    has_phone Domain Educational_Institute
has_responsiblePerson
    has_responsiblePerson Domain Educational_Institute
School
    School SubClassOf Educational_Institute
```

Figure 10: Educational Institute Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: School

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

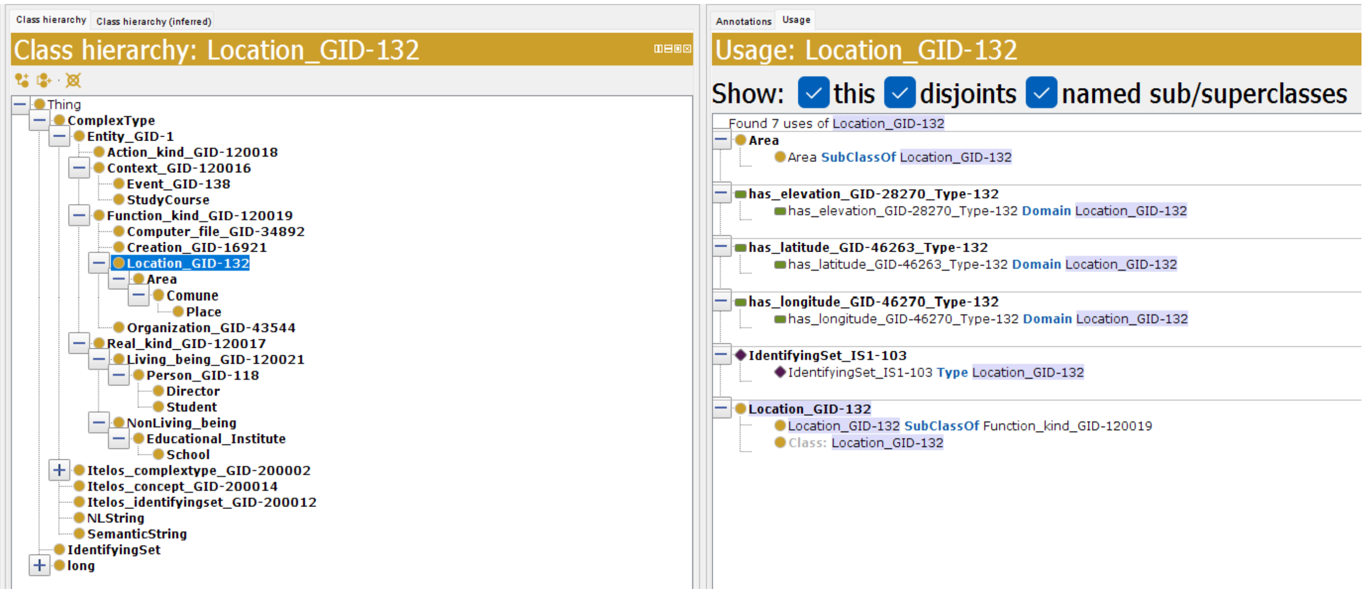
Annotations | Usage

Usage: School

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 7 uses of School
has_coursesOffered
    has_coursesOffered Domain School
has_enrolledInSchool
    has_enrolledInSchool Range School
has_location
    has_location Domain School
has_SchoolLevelOfEducation
    has_SchoolLevelOfEducation Domain School
has_timetable
    has_timetable Domain School
School
    School SubClassOf Educational_Institute
    Class: School
```

Figure 11: School Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Location_GID-132

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

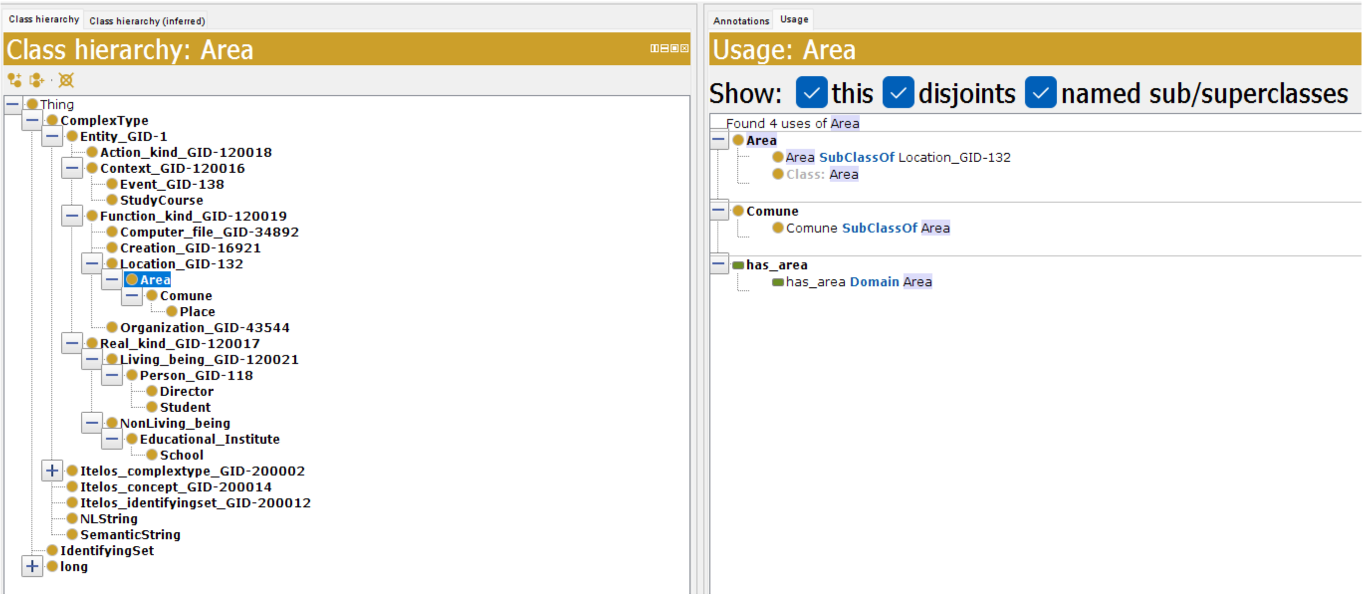
Annotations | Usage

Usage: Location_GID-132

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 7 uses of Location_GID-132
Area
    Area SubClassOf Location_GID-132
has_elevation_GID-28270_Type-132
    has_elevation_GID-28270_Type-132 Domain Location_GID-132
has_latitude_GID-46263_Type-132
    has_latitude_GID-46263_Type-132 Domain Location_GID-132
has_longitude_GID-46270_Type-132
    has_longitude_GID-46270_Type-132 Domain Location_GID-132
IdentifyingSet_IS1-103
    IdentifyingSet_IS1-103 Type Location_GID-132
Location_GID-132
    Location_GID-132 SubClassOf Function_kind_GID-120019
    Class: Location_GID-132
```

Figure 12: Location Class



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Area

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

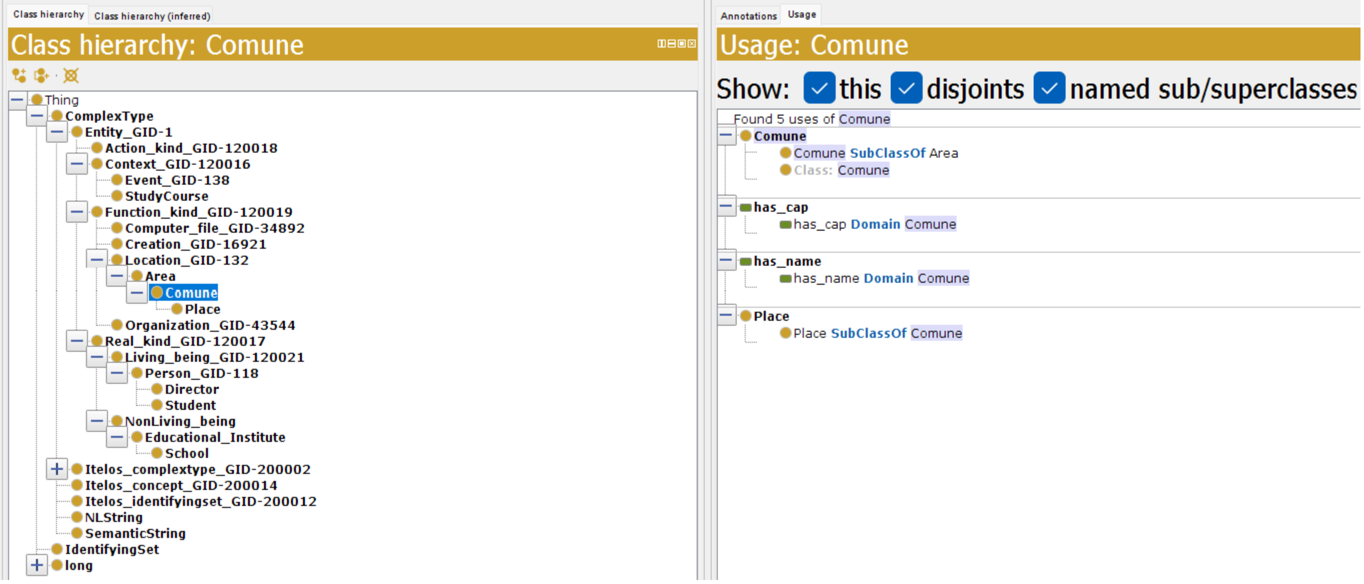
Annotations | Usage

Usage: Area

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 4 uses of Area
Area
    Area SubClassOf Location_GID-132
    Class: Area
Comune
    Comune SubClassOf Area
has_area
    has_area Domain Area
```

Figure 13: Area Class



Class hierarchy: Comune

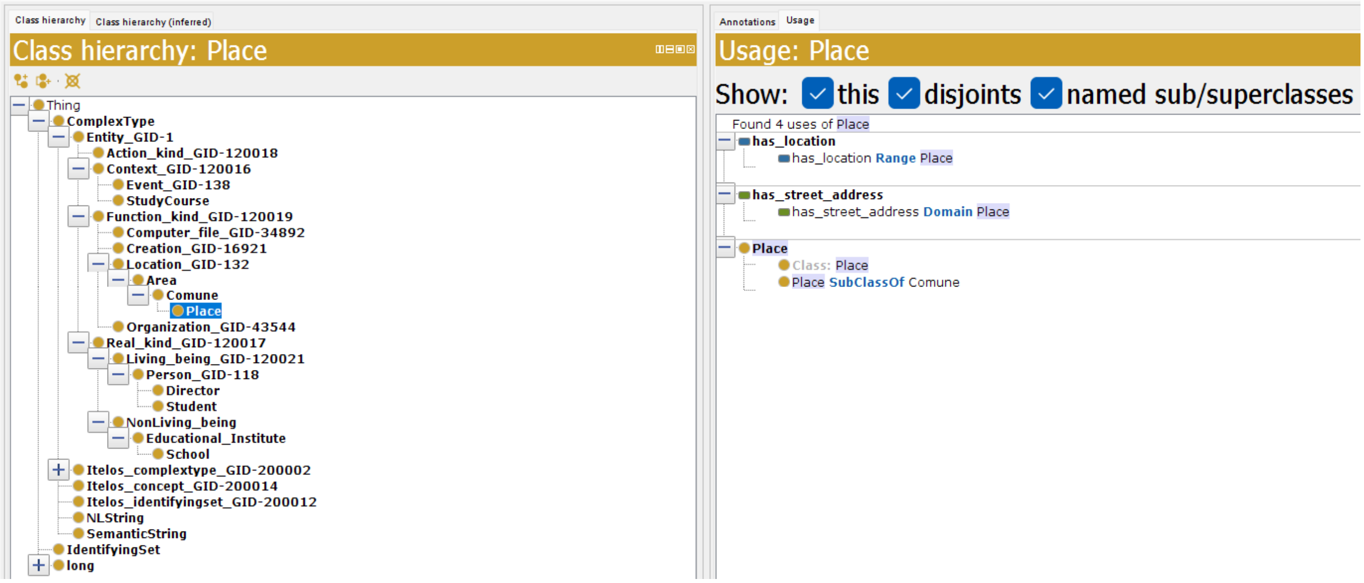
```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

Usage: Comune

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 5 uses of Comune
Comune
    Comune SubClassOf Area
    Class: Comune
has_cap
    has_cap Domain Comune
has_name
    has_name Domain Comune
Place
    Place SubClassOf Comune
```

Figure 14: Comune Class



Class hierarchy: Place

```
Thing
  ComplexType
    Entity_GID-1
      Action_kind_GID-120018
      Context_GID-120016
        Event_GID-138
        StudyCourse
      Function_kind_GID-120019
        Computer_file_GID-34892
        Creation_GID-16921
        Location_GID-132
          Area
            Comune
              Place
        Organization_GID-43544
      Real_kind_GID-120017
        Living_being_GID-120021
          Person_GID-118
            Director
            Student
        NonLiving_being
          Educational_Institute
            School
    Itelos_complextype_GID-200002
    Itelos_concept_GID-200014
    Itelos_identifyingset_GID-200012
    NLString
    SemanticString
  IdentifyingSet
  long
```

Usage: Place

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

```
Found 4 uses of Place
has_location
    has_location Range Place
has_street_address
    has_street_address Domain Place
Place
    Class: Place
    Place SubClassOf Comune
```

Figure 15: Place Class

5.2.1 Data Types Misalignment

First of all, data-sets were deeply analyzed in order to see any data type misalignment. Since the project contains only three data-sets and all of the data-sets contain different information with out any pair of data-sets containing same piece of of information, so no data type misalignment was identified.

5.2.2 Data Value Format Misalignment

Both of the main data-sets, educational institutes in Trentino and study courses offered in Trentino are Provided by the same publisher. Both were scanned manually to identify any data value format difference between the values of the most important properties like school name, courses name, course duration, start-date etc. Since the data-sets are collected from a very reliable resource so no discrepancy was found and the data value language was already well aligned and did not involve an further processing.

5.2.3 Data Value Language Misalignment

Both of the main data-sets, educational institutes in Trentino and study courses offered in Trentino are in the Italian language. Both were scanned manually to identify any language difference between the values of the most important properties like school name, courses name. Since the data-sets are collected from a very reliable resource so no discrepancy was found and the data value language was already well aligned and did not involve an further processing.

5.3 Import into the KOS System

After construction of the ontology in the protege modeling all the entities and the properties formulated from the ER diagram, the next step was to import the ontology in the KOS system. A key point to note here is that even thought the values of the data-sets are in the Italian language but the names of the entities and the data and object properties associated to each of the entity were kept in the English language to keep the overall modeling of the system uniform since most of the concepts are defined in English in the KOS system.

When we imported the ontology owl file in the KOS system during the completion of the formal modeling phase, many conceptual issue arose. The convention followed to names the entities and the properties came handy here as it was very easy to explore KOS in order to find already defined concept since the names in English were trivial and familiar to the system. As an example, has_area property of the comune was pretty easy to search in KOS and many predefined similar defined were present which were eventually associated to the particular property.

Only couple of the properties were not present there which we then defined providing the proper definition. The ’levelOfEducation’ property is the example of this scenario. KOS system provided support of creation the ETG after resolving the conceptual issues related the imported ontology. This version of the ETG was then used in the data integration phase that was carried out on the karmalinker tool.

5.4 Datasets Refinement

Up to this point, three data sets were finalized to be used in the system. After defining the necessary entities and their relative properties, another cleaning pass of the the data-sets was done. All of the unnecessary values from the data-sets were removed. Only the data-sets attributes that associate to the properties of the ETG were spared.

5.5 ETG Model Update

Although initially all of the entities in the ontology was designed with the intention the align it well with ER model but later in the validation phase, it was realized that some of the choices made in the formulation of the ETG do not align well.

To quote an example, a visible clash can be observed between the comune-location-area entities. As the school is supposed to be located in a comune and area should be associated to the comune but the previously defined ETG does not depict the same relation among these entities. So to resolve this issue, place entity was removed from the ETG and area was removed as entity and placed under the comune entity. To model this relation, has_area was introduced in the comune entity.

Similarly place class was also removed from the ER model and the ontology since it was not making sense anymore. Director entity was use to link the responsible person relation of the school and educational institution. Role of the responsible person was supposed to be saved in the educational institution class.

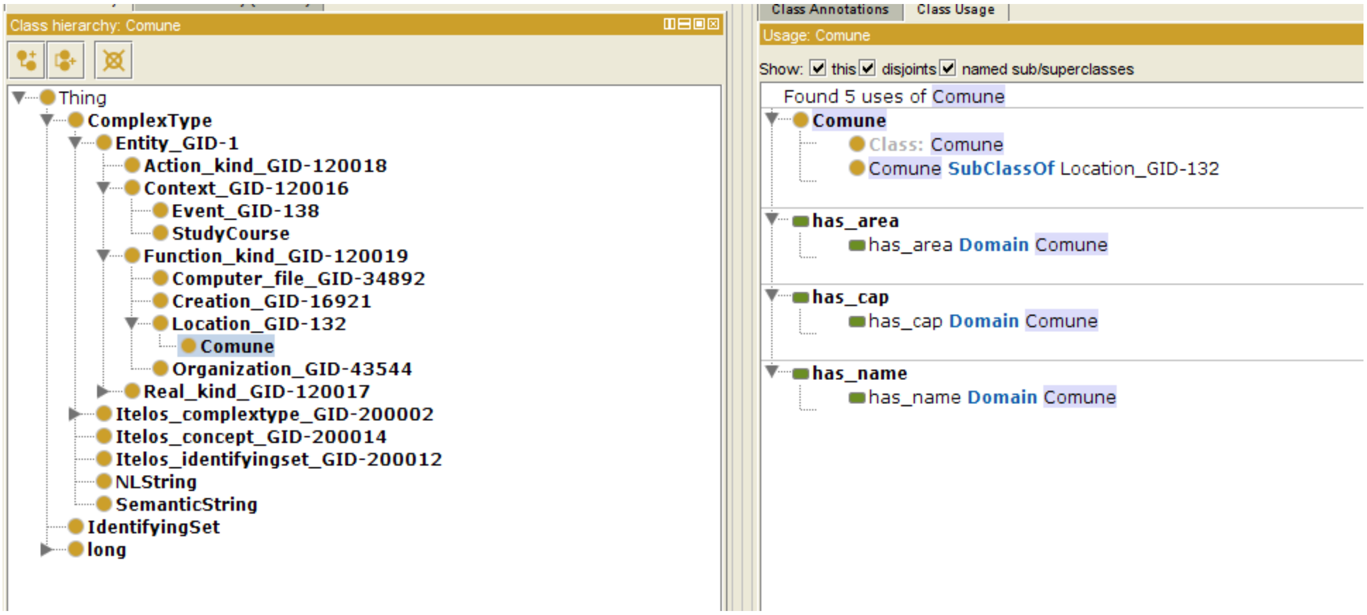
Figure[16]. provided below to give the better under standing of the situation discussed above in context of comune.

6 Data Integration

This section of the report is dedicated to the description of the data integration phase. It aims to describe the different sub activities performed as well as the phase outcomes produced.

6.1 Data Management (semantic heterogeneity)

Following phase of the process is about resolution the problem related the the semantic heterogeneity present in the data-sets. Well in case of this work,luckily there was no semantic heterogeneity identified in the data-sets.



```
▼ Thing
  ▼ ComplexType
    ▼ Entity_GID-1
        Action_kind_GID-120018
      ▼ Context_GID-120016
          Event_GID-138
          StudyCourse
      ▼ Function_kind_GID-120019
          Computer_file_GID-34892
          Creation_GID-16921
        ▼ Location_GID-132
            Comune
          Organization_GID-43544
      ▶ Real_kind_GID-120017
    ▶ Itelos_complextype_GID-200002
      Itelos_concept_GID-200014
      Itelos_identifyingset_GID-200012
      NLString
      SemanticString
    IdentifyingSet
  ▶ long
```

```
Found 5 uses of Comune
▼ Comune
    Class: Comune
    Comune SubClassOf Location_GID-132
▼ has_area
    has_area Domain Comune
▼ has_cap
    has_cap Domain Comune
▼ has_name
    has_name Domain Comune
```

Figure 16: Comune Updated Class

6.2 Entity Alignment

This section of the reports explains how the finalized ontology will be mapped with the well formatted data-sets. karmalinker is the tool that was used to map data-sets to the finalized ontology. The tool allows us to align reference data to the ontology created in the formal modeling phase of the data integration project. Three different format outputs can be generated from the tool that are: RDF file, a EML file and model file. Each of the mentioned format server a different purpose.

6.2.1 Dataset Format Modification

When the ontology and the data-sets were imported in the karmalinker system, it was realized that the mapping entities and the related data and object properties was super complicated in case of some of the data-sets. Especially in the case of the data-set 'Trentino educational institutions', the data-set was in XML format and the structure of the data-set was hard to understand when imported in the karmalinker. To ease the situation, a critical decision of reformatting the data-set and splitting bigger data-set into small data-set was taken.

So the mentioned data-set was broken into 3 smaller data-set. All of the data-sets were converted to CSV to keep format consistent throughout the application. The new data-sets and their key attributes were following:

Dataset - courses_offered_in_trentino_school	
Attribute	**Description**
codiceOrigine	unique code of the study course
descrizione	description/name of the study course

durataAnni	course duration
attivo	activity status of the study course

Dataset - educational_institutions	
Attribute	**Description**
institute_name	name of educational institute
institute_phone	phone number of the educational institute
institute_fax	fax of the educational institute
institute_email	email of the educational institute
institute_address	address of the educational institute
responsible_person	responsible person/director of the educational institute

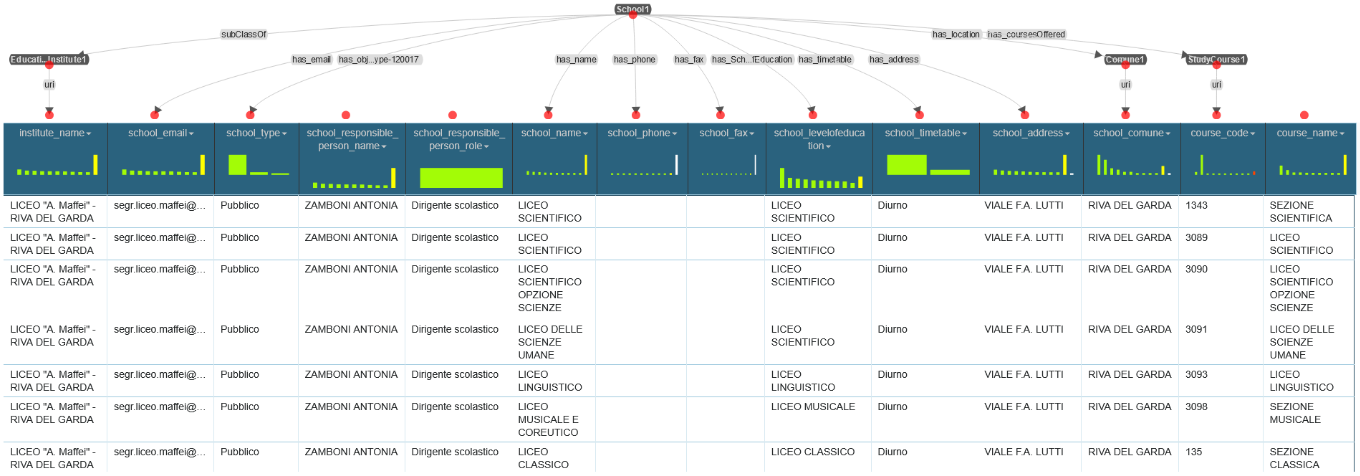
Dataset - schools_in_trentino	
Attribute	**Description**
institute_name	name of educational institute which the school is supervised by
school_phone	phone number of the school
school_fax	fax of the school
school_email	email of the school
school_address	address of the school
school_type	type of school
course_code	code of the course offered by the school
school_levelofeducation	level of education provided by the school
school_comune	name of the comune that the school belongs to

Dataset - responsible_person_educational_institutions	
Attribute	**Description**
institute_name	name of educational institute which the person is responsible of
responsible_person	name of the person
responsible_person_role	role of the person in the educational institution

Dataset - Trentino Commune List	
Attribute	**Description**
comune_name	name of comune
area	area of the comune
comunecode	zip code of the comune

6.2.2 Entity Matching

Having done above splitting and reformatting of data-sets, all the data-sets were imported in the karmalinker along the ontology. Entity matching was done in then between the data-sets and the ontology elements. Details of the mapping is provided in form of the Figure[17,18,19,20,21]. that contains information about how data was linked to the ontology for each of the above mentioned 5 datasets.



institute_name	school_email	school_type	school_responsible_person_name	school_responsible_person_role	school_name	school_phone	school_fax	school_levelofeducation	school_timetable	school_address	school_comune	course_code	course_name
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO SCIENTIFICO			LICEO SCIENTIFICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	1343	SEZIONE SCIENTIFICA
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO SCIENTIFICO			LICEO SCIENTIFICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	3089	LICEO SCIENTIFICO
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO SCIENTIFICO OPZIONE SCIENZE			LICEO SCIENTIFICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	3090	LICEO SCIENTIFICO OPZIONE SCIENZE
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO DELLE SCIENZE UMANE			LICEO SCIENTIFICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	3091	LICEO DELLE SCIENZE UMANE
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO LINGUISTICO			LICEO LINGUISTICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	3093	LICEO LINGUISTICO
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO MUSICALE E COREUTICO			LICEO MUSICALE	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	3098	SEZIONE MUSICALE
LICEO "A. Maffei" - RIVA DEL GARDA	segr.liceo.maffei@...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	LICEO CLASSICO			LICEO CLASSICO	Diurno	VIALE F.A. LUTTI	RIVA DEL GARDA	135	SEZIONE CLASSICA

Figure 17: Schools in Trentino Dataset Mapping

Upon successful completion of the entity matching process, PDF graph models were downloaded from the karmalinker which can imported into any graph visualization tools like GraphDb and SPQSQL.

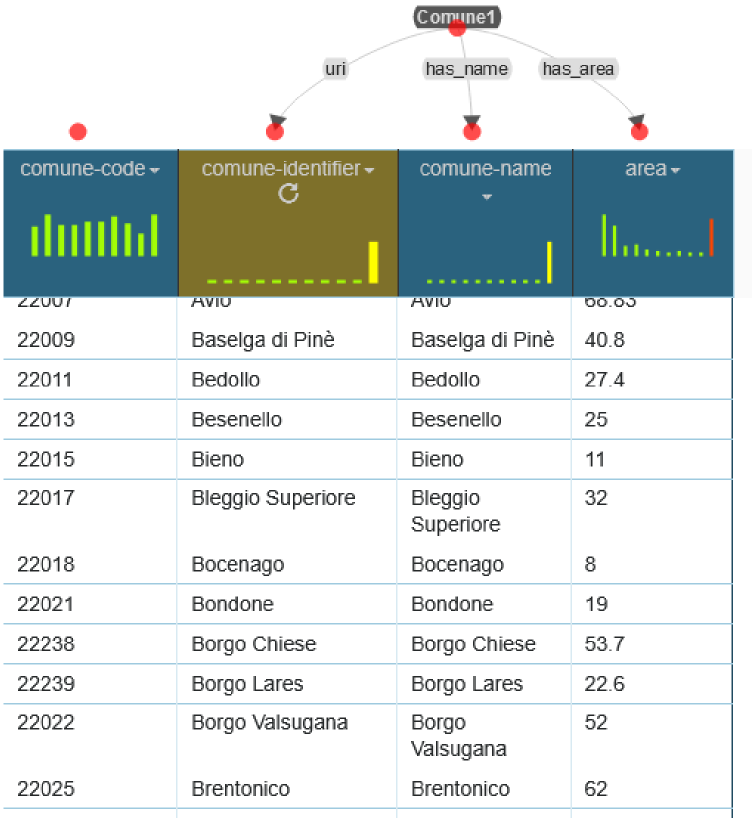
6.3 Data Integration Phase Evaluation

Having done all the necessary work, validation phase of the final graph constructed was started. Different type of evaluation of the data integration phase was done in order to ensure that system works in desired way.

6.3.1 Answering the CQs

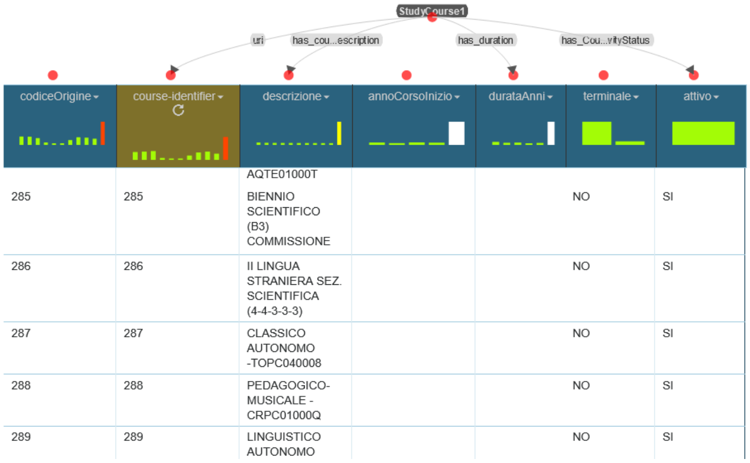
Since the goal of the data integration project was to construct a solution that can answer CQs (competency questions). Unfortunately due to some unforeseen errors, we were not able to explored the final graph in the GrapghDb or any other similar tool for graph exploration. Still the validation was done on the conceptual bases following the modeling/entity matching done in the karmalinker.

We tried to answer CQs following the incoming/outgoing links and in most of the cases we were able to find the desired data. For example, if someone wants to search all schools in Trento, we can very well start from the comune dataset looking for Trento comune and then following its conceptual link in the schools dataset to find out link of all school that are located



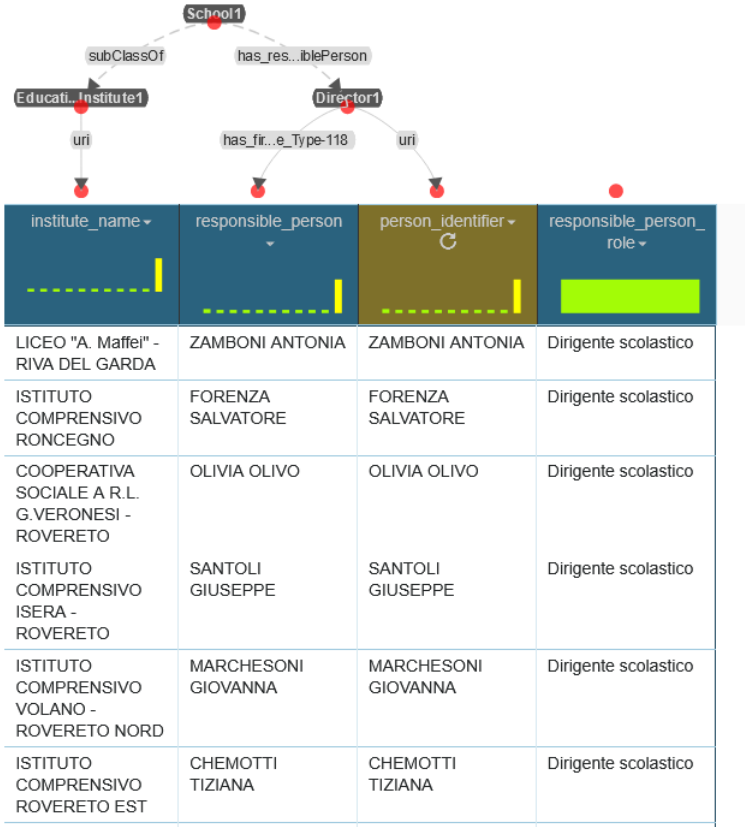
comune-code	comune-identifier	comune-name	area
22007	Avio	Avio	68.83
22009	Baselga di Pinè	Baselga di Pinè	40.8
22011	Bedollo	Bedollo	27.4
22013	Besenello	Besenello	25
22015	Bieno	Bieno	11
22017	Bleggio Superiore	Bleggio Superiore	32
22018	Bocenago	Bocenago	8
22021	Bondone	Bondone	19
22238	Borgo Chiese	Borgo Chiese	53.7
22239	Borgo Lares	Borgo Lares	22.6
22022	Borgo Valsugana	Borgo Valsugana	52
22025	Brentonico	Brentonico	62

Figure 18: Comune in Trentino Dataset Mapping



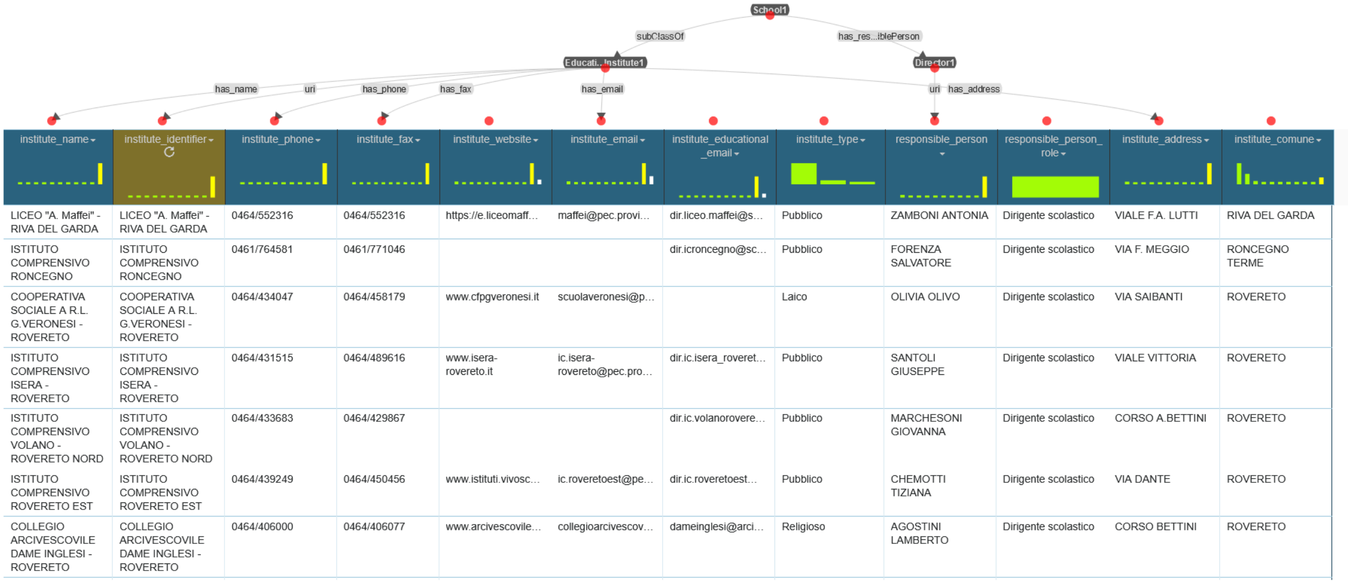
codiceOrigine	course-identifier	descrizione	annoCorsoInizio	durataAnni	terminale	attivo
		AQTE01000T				
285	285	BIENNIO SCIENTIFICO (B3) COMMISSIONE			NO	SI
286	286	II LINGUA STRANIERA SEZ. SCIENTIFICA (4-4-3-3-3)			NO	SI
287	287	CLASSICO AUTONOMO -TOPC040008			NO	SI
288	288	PEDAGOGICO-MUSICALE - CRPC01000Q			NO	SI
289	289	LINGUISTICO AUTONOMO			NO	SI

Figure 19: Course offered in Trentino School Dataset Mapping



institute_name	responsible_person	person_identifier	responsible_person_role
LICEO "A. Maffei" - RIVA DEL GARDA	ZAMBONI ANTONIA	ZAMBONI ANTONIA	Dirigente scolastico
ISTITUTO COMPRENSIVO RONCEGNO	FORENZA SALVATORE	FORENZA SALVATORE	Dirigente scolastico
COOPERATIVA SOCIALE A R.L. G.VERONESI - ROVERETO	OLIVIA OLIVO	OLIVIA OLIVO	Dirigente scolastico
ISTITUTO COMPRENSIVO ISERA - ROVERETO	SANTOLI GIUSEPPE	SANTOLI GIUSEPPE	Dirigente scolastico
ISTITUTO COMPRENSIVO VOLANO - ROVERETO NORD	MARCHESONI GIOVANNA	MARCHESONI GIOVANNA	Dirigente scolastico
ISTITUTO COMPRENSIVO ROVERETO EST	CHEMOTTI TIZIANA	CHEMOTTI TIZIANA	Dirigente scolastico

Figure 20: Responsible Person of Institutions Dataset Mapping



institute_name	institute_identifier	institute_phone	institute_fax	institute_website	institute_email	institute_educational_email	institute_type	responsible_person	responsible_person_role	institute_address	institute_comune
LICEO "A. Maffei" - RIVA DEL GARDA	LICEO "A. Maffei" - RIVA DEL GARDA	0464/552316	0464/552316	https://e.liceomaff...	maffei@pec.provi...	dir.liceo.maffei@s...	Pubblico	ZAMBONI ANTONIA	Dirigente scolastico	VIALE F.A. LUTTI	RIVA DEL GARDA
ISTITUTO COMPRENSIVO RONCEGNO	ISTITUTO COMPRENSIVO RONCEGNO	0461/764581	0461/771046			dir.icroncegno@sc...	Pubblico	FORENZA SALVATORE	Dirigente scolastico	VIA F. MEGGIO	RONCEGNO TERME
COOPERATIVA SOCIALE A R.L. G.VERONESI - ROVERETO	COOPERATIVA SOCIALE A R.L. G.VERONESI - ROVERETO	0464/434047	0464/458179	www.cfpgveronesi.it	scuolaveronesi@p...		Laico	OLIVIA OLIVO	Dirigente scolastico	VIA SAIBANTI	ROVERETO
ISTITUTO COMPRENSIVO ISERA - ROVERETO	ISTITUTO COMPRENSIVO ISERA - ROVERETO	0464/431515	0464/489616	www.isera-rovereto.it	ic.isera-rovereto@pec.pro...	dir.ic.isera_roveret...	Pubblico	SANTOLI GIUSEPPE	Dirigente scolastico	VIALE VITTORIA	ROVERETO
ISTITUTO COMPRENSIVO VOLANO - ROVERETO NORD	ISTITUTO COMPRENSIVO VOLANO - ROVERETO NORD	0464/433683	0464/429867			dir.ic.volanorovere...	Pubblico	MARCHESONI GIOVANNA	Dirigente scolastico	CORSO A.BETTINI	ROVERETO
ISTITUTO COMPRENSIVO ROVERETO EST	ISTITUTO COMPRENSIVO ROVERETO EST	0464/439249	0464/450456	www.istituti.vivosc...	ic.roveretoest@pe...	dir.ic.roveretoest...	Pubblico	CHEMOTTI TIZIANA	Dirigente scolastico	VIA DANTE	ROVERETO
COLLEGIO ARCIVESCOVILE DAME INGLESI - ROVERETO	COLLEGIO ARCIVESCOVILE DAME INGLESI - ROVERETO	0464/406000	0464/406077	www.arcivescovile...	collegioarcivescov...	dameinglesi@arci...	Religioso	AGOSTINI LAMBERTO	Dirigente scolastico	CORSO BETTINI	ROVERETO

Figure 21: Institutions Dataset Mapping

in Trento comune. All of this reasoning was done manually without use of any tool due to the reason mentioned above.

6.3.2 Consistency Dimension

Evaluation was also done in the consistency dimensions. During this phase of validation,following conditions were verified:

- No cycles are used in a class hierarchy.
- Avoid usage of polysemous term.
- Only one domain and range for each property.
- Do not use different term to refer same element.

6.3.3 Accuracy Dimension

Following conditions were verified while validating the KG in the accuracy dimensions:

- Avoid using relation name like isA or type. All of the relationship in the system had 'has_a' relationship.
- Discard any leaf class for which there is no instances.
- Do not use miscellaneous or other as a class name. Proper naming convention for the classes was followed.
- Avoid to keep isolated elements. All of the elements in the ontology were linked to each other.

6.3.4 Completeness Dimension

Last dimension in the validation phase was completeness dimension. During this,it was made sure that all of the classes have proper domain and range assigned.

7 Open Issues/Challenges

This section aims to describe any issues/problems remained open along the DI process. One of the issues that still exist in the final output is that the some of the data-sets might have some values that were left unused while entity matching. Although almost all of the unnecessary properties were removed from the data-sets in the cleaning phases but still some of the attributes of the data-sets were not utilized in the entity matching. This can cause unnecessary increase in the size of the final KG without improving its utilization. So a potential solution that is moved focused on the process would be to very deeply identify such attributes in the data-sets that do not have any potential use towards the Data integration end product and their removal to make the KG on point and more efficient, especially in terms of performance.

A critical problem was encountered while mapping data-set attributes to the final ontology in the karmalinker. The nature of the issue was structural rather than technical. One of the data-sets was very complex to understand when imported in the karmalinker. It was nearly impossible to map it to the ontology. To resolve this issue, we did split that data-set in to multiple smaller data-sets that made it easier to manage and map them.

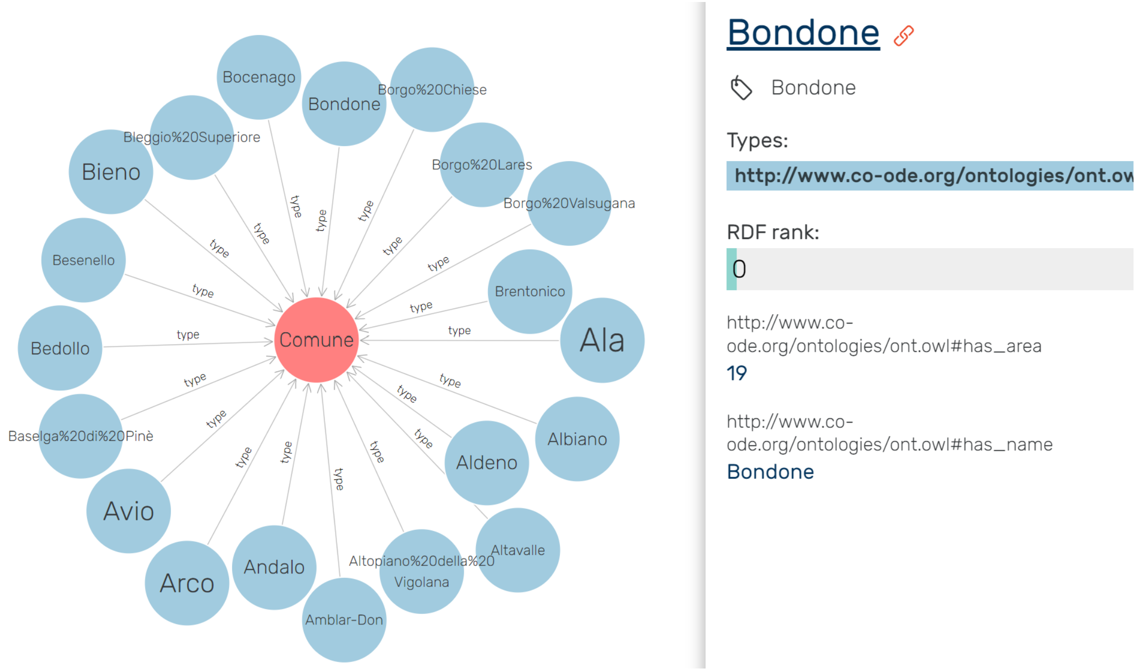
Another problem that occurred and still exists is while importing the final RDF files in the graph exploration tools. Invalid Base IRI error was encountered while importing the files in the GraphDb. We tried to redo the mapping in the karmalinker and again generated the RDF graph files but the issue still persisted and that is the reason that we were not able to properly explore the finally produced Knowledge Graph.

8 Exploitation

The goal of the data integration project was to construct a solution that can answer CQs (competency questions). Unfortunately due to some unforeseen errors, we were not able to explored the final graph in the GrapghDb or any other similar tool for graph exploration. Invalid Base IRI error was encountered while importing the files in the GraphDb. We tried to fix the issue by redoing the mapping in the karmalinker and again generated the RDF graph files but this did not fix the issue. Despite doing the necessary research, we could not track down the issue. As result of this we were not able to properly explore the finally produced Knowledge Graph. We were able to import only part of graphs in the GraphDB that are also shown in the figures[22,23] below.

Associated properties to each of the concepts can also be seen on the detailed widget on the right side.

Despite this major setback, the analysis of the potential usage of the KG can be done on the conceptual bases following the modeling/entity matching done in the karmalinker. We tried to answer CQs following the incoming/outgoing links and in most of the cases we were able to find the desired data. For example, if someone wants to search all schools in Trento, we can very well start from the comune data-set looking for Trento comune and then following its conceptual link in the schools dataset to find out link of all school that are located in Trento comune. All of this reasoning was done manually without use of any tool due to the reason mentioned above. A possible solution of this problem could be by making sure that there is no entity that is un-mapped or improperly mapped in the karmalinker and then generate the RDF graphs. We could not do this as per time limitation and limited availability of the tools involved these phases of the project.



Bondone

Bondone

Types:

http://www.co-ode.org/ontologies/ont.ow

RDF rank:

0

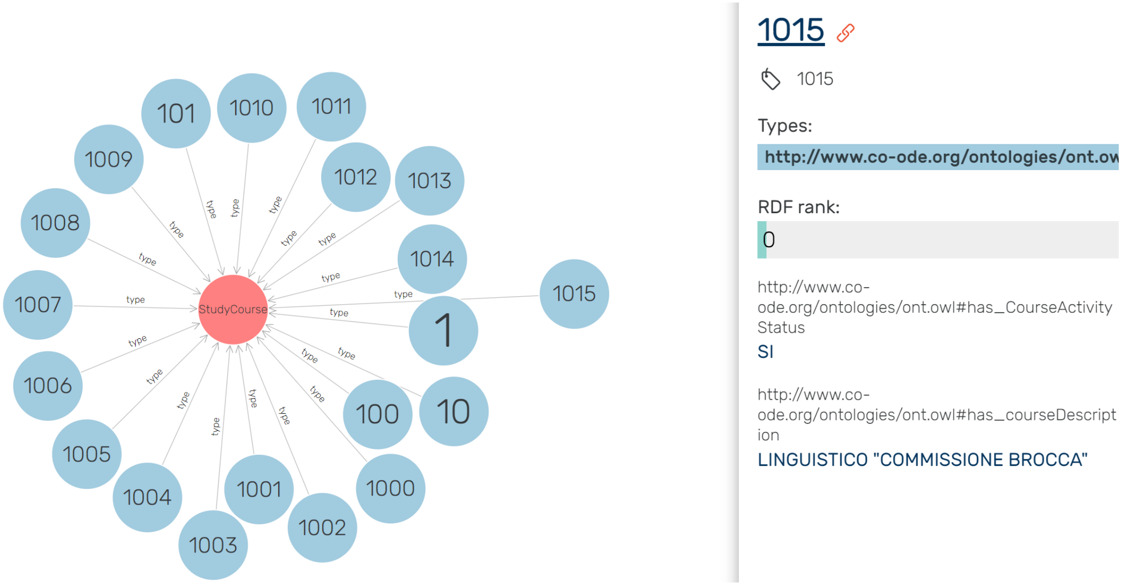
http://www.co-ode.org/ontologies/ont.owl#has_area

19

http://www.co-ode.org/ontologies/ont.owl#has_name

Bondone

Figure 22: Comune Concept Graph Visualization



1015

1015

Types:

http://www.co-ode.org/ontologies/ont.ow

RDF rank:

0

http://www.co-ode.org/ontologies/ont.owl#has_CourseActivityStatus

SI

http://www.co-ode.org/ontologies/ont.owl#has_courseDescription

LINGUISTICO "COMMISSIONE BROCCA"

Figure 23: Study Course Concept Graph Visualization

9 Bibliography

References

[1] "Benvenuti - dati trentino," https://dati.trentino.it/, accessed: 2021-10-14.

[2] "Istituzioni scolastiche del trentino - dati trentino," https://dati.trentino.it/dataset/istituzioni-scolastiche-trentino, accessed: 2021-10-20.

[3] "Corsi di studio delle scuole trentine - dati trentino," https://dati.trentino.it/dataset/corsi-di-studio-delle-scuole-trentine, accessed: 2021-10-22.

[4] "Municipalities of trentino - wikipedia," https://en.wikipedia.org/wiki/Municipalities_of_Trentino, accessed: 2021-10-26.