

Dipartimento di Ingegneria e Scienza dell'Informazione

Knowledge and Data Integration

Project Report - Education in Trentino

Document Data:

December 7, 2021

Reference Persons:

Ali Hamza, Tecla Venturelli

© 2021 University of Trento
Trento, Italy

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.



Index:

1	Introduction	1
2	Purpose and project's resources	1
2.1	Purpose Formalization	1
2.1.1	Project Purpose	1
2.1.2	Project Domain Composition	1
2.1.3	Personas	2
2.1.4	Scenarios	3
2.2	knowledge Resources	4
2.3	Data Resources	5
2.3.1	Data Collection	5
2.4	Metadata	6
2.4.1	Trentino educational institutions	7
2.4.2	Study courses of the Trentino schools	7
3	Inception	8
3.1	Purpose Formalization	8
3.1.1	Domain Composition	8
3.1.2	Domain of Interest	8
3.1.3	Inception Sheet	9
3.2	Data and knowledge resource collection	12
3.2.1	Data Input Heterogeneity	12
3.2.2	Data Transformation Activity - DTA1	12
3.3	Inception Phase Evaluation	13
3.3.1	Schema Level Evaluation	13
3.3.2	Data Level Evaluation	13
4	Informal Modeling	14
4.1	Purpose Formalization and Modeling Sheet Description	14
4.2	ER Model Description	15
4.2.1	Schema Level EType and ERD Description	15
4.3	Informal Modeling Evaluation	17
4.3.1	Schema Level Evaluation	17
4.3.2	Data Level Evaluation	19
5	Formal Modeling	20
5.1	ETG Generation	20
5.1.1	ER Model Update	20
5.1.2	ETG Generation	22
5.2	Data management (syntactic heterogeneity)	22
5.2.1	Data Types Misalignment	29

5.2.2	Data Value Format Misalignment	29
5.2.3	Data Value Language Misalignment	29
6	Bibliography	29

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
0.3	04.12.2021	Tecla Venturelli	Formal Modeling Phase Report

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
<i>istituzione-scolastica:denominazione</i>	Name of the educational institution
<i>istituzione-scolastica:telefono</i>	Telephone number of the institution
<i>istituzione-scolastica:fax</i>	Fax number of the institution
<i>istituzione-scolastica:emailIstituzione</i>	Email of the institution
<i>istituzione-scolastica:responsabile:nominativo</i>	Name of the Contact person of the educational institution
<i>istituzione-scolastica:responsabile:ruolo:descrizione</i>	Role of the Contact person of the educational institution
<i>istituzione-scolastica:sede:descrizione</i>	Address of the institution
<i>istituzione-scolastica:sede:indirizzo</i>	Street Address of the institution
<i>istituzione-scolastica:sede:comune:descrizione</i>	Name of the comune where the institution is located
<i>istituzione-scolastica:sede:comune:descrizione</i>	Name of the comune where the institution is located

<i>istituzione-scolastica:unita-didattica:denominazione</i>	Name of the teaching unit
<i>istituzione-scolastica:unita-didattica:tipo-scuola-riferimento</i>	Type of the teaching unit
<i>istituzione-scolastica:unita-didattica:tipologia-orario</i>	Typology time (Daytime or Evening)
<i>istituzione-scolastica:unita-didattica:tipologia-orario</i>	Typology time (Daytime or Evening)
<i>istituzione-scolastica:unita-didattica:offerta</i>	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
<i>corso-studi:codiceOrigine</i>	Code of the course
<i>corso-studi:descrizione</i>	Name of the study course
<i>corso-studi:annoCorsiInizio</i>	Starting academic year of the course
<i>corso-studi:durataAnni</i>	Duration of the course
<i>corso-studi:attivo</i>	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
<i>ISTAT Code</i>	Code of the municipality
<i>Italian</i>	Italian name of the commune
<i>German</i>	German name of the commune
<i>Area</i>	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 information 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, High-school
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, Day-time, 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 previous 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 listed. 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. It often requires a lot of preprocessing in order to make it useful and ready for integration. Since data was downloaded from more than one data source, 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 formats:

- 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, High-School, MiddleSchool, Kindergarten, Student, City, Area, Place, Person, DisabledPerson, StudyCourse, HeadTeacher, Director, HeadManager, Coordinator, PrivateSchool, Public-School }

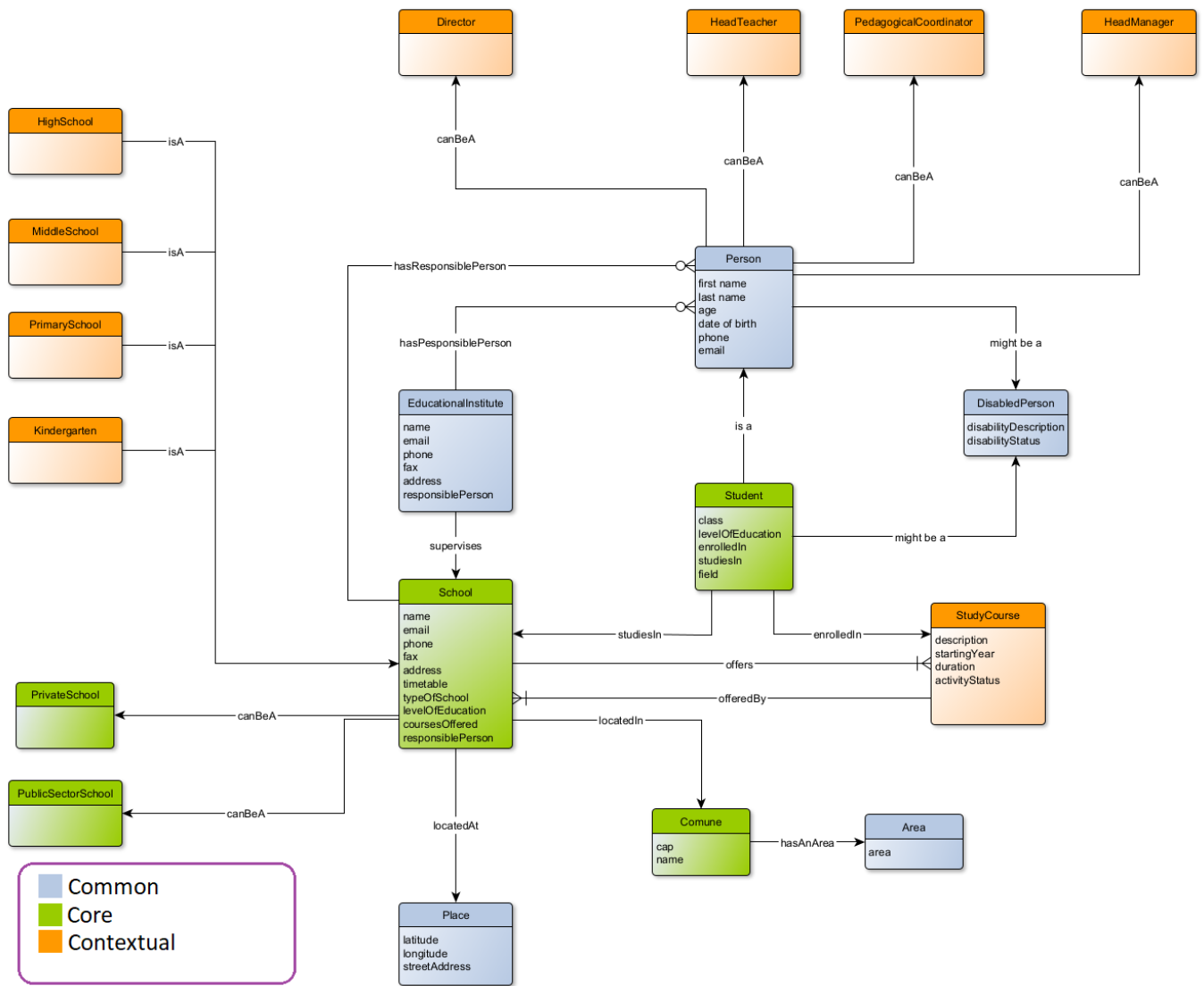


Figure 1: ER Model Diagram

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.

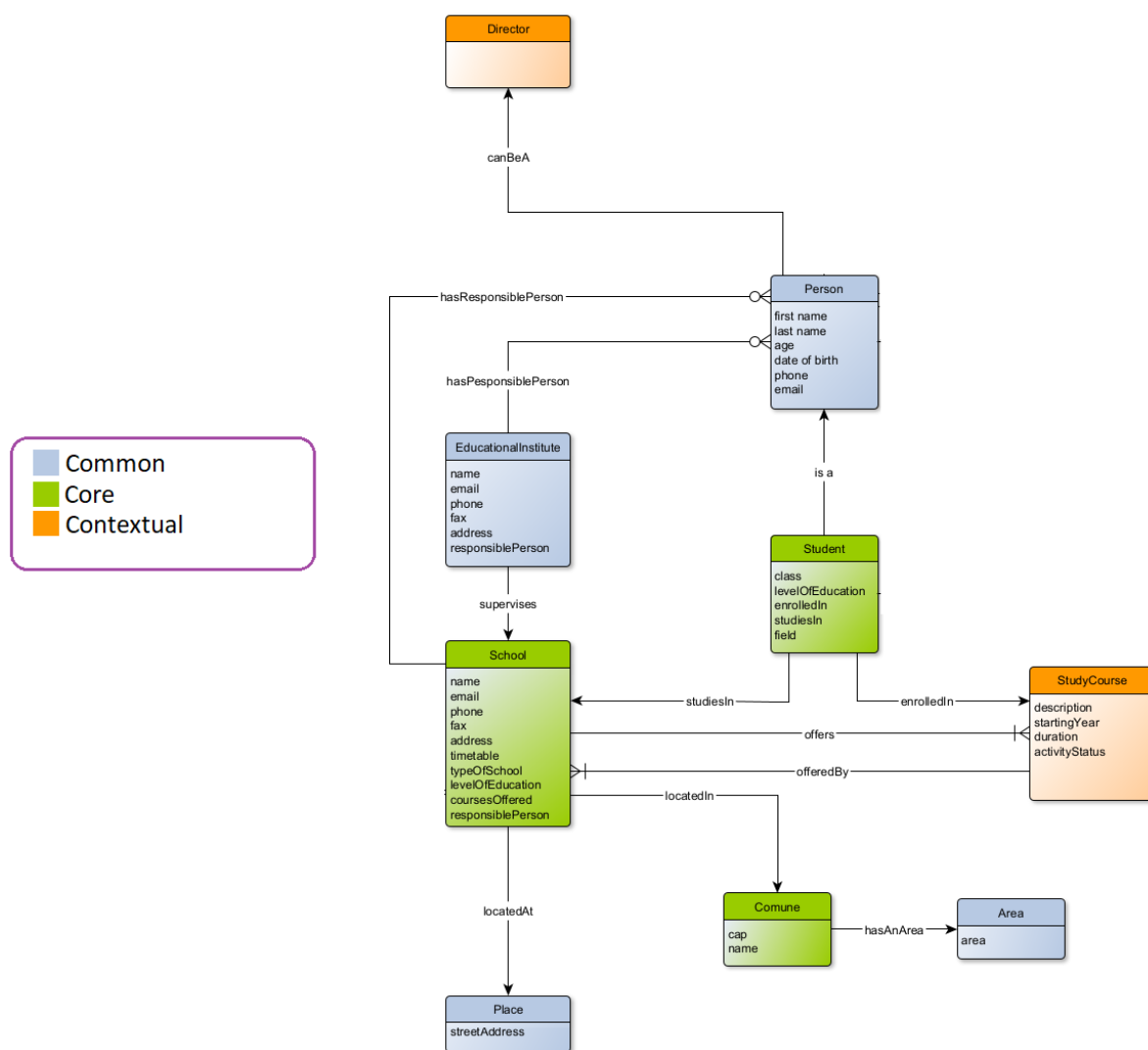


Figure 2: Updated ER Model Diagram

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].

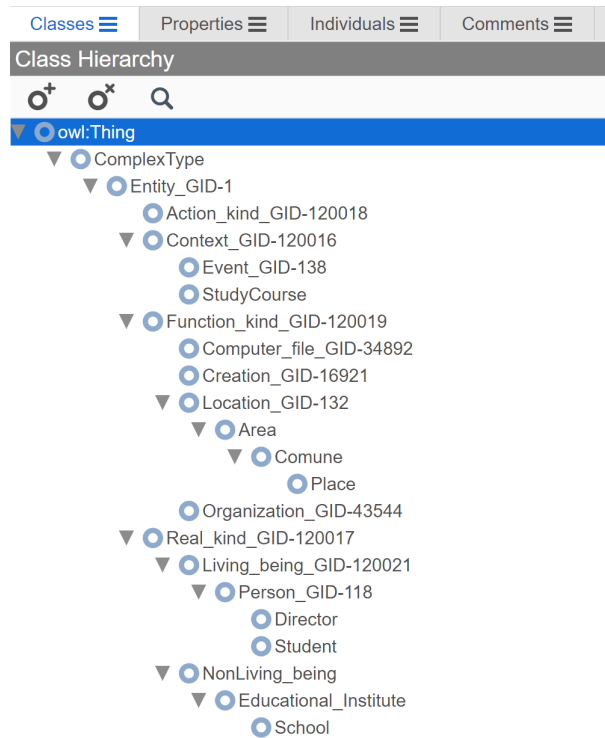


Figure 3: ETG - Entity Classes

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 than 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 on 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.

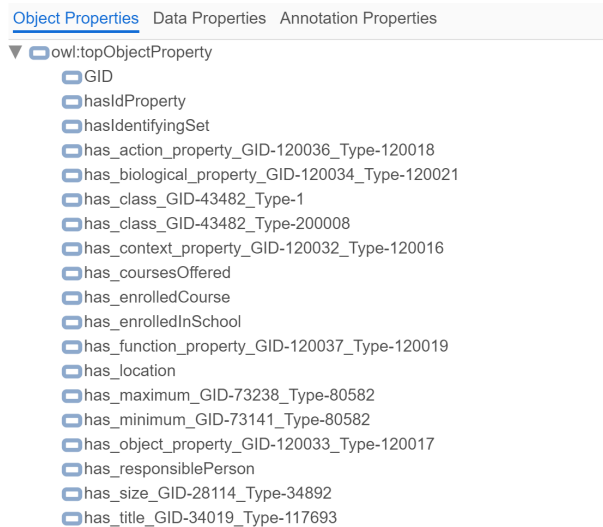


Figure 4: ETG - Object Properties



Figure 5: ETG - Data Properties

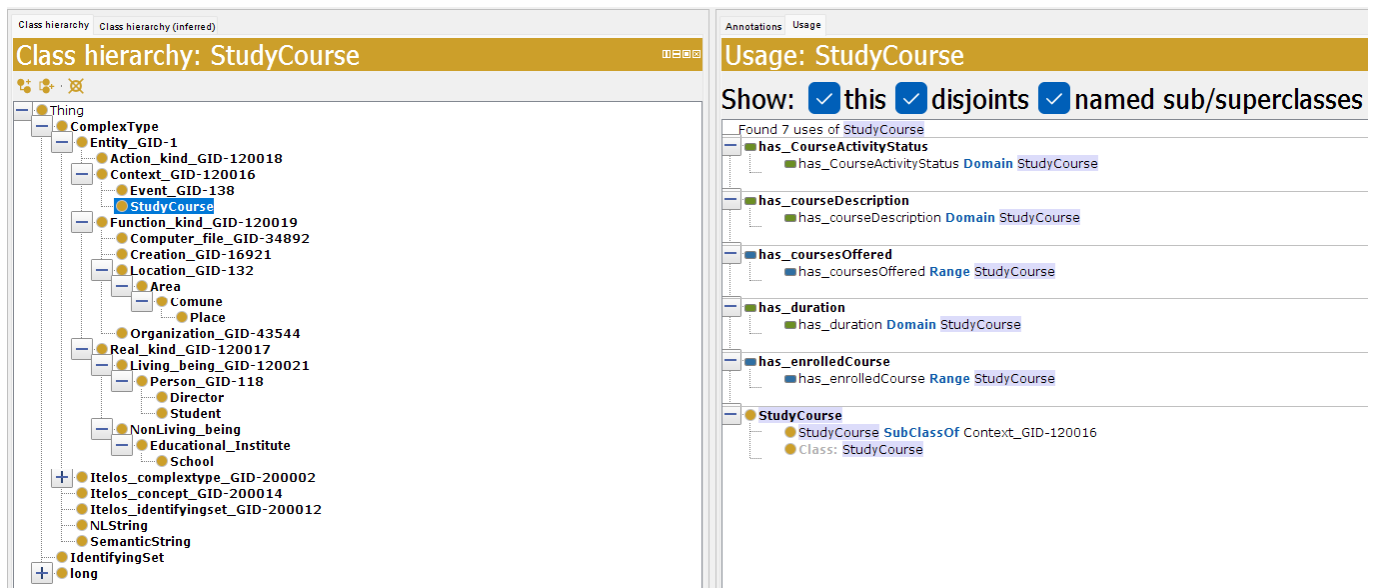


Figure 6: Study Course Class

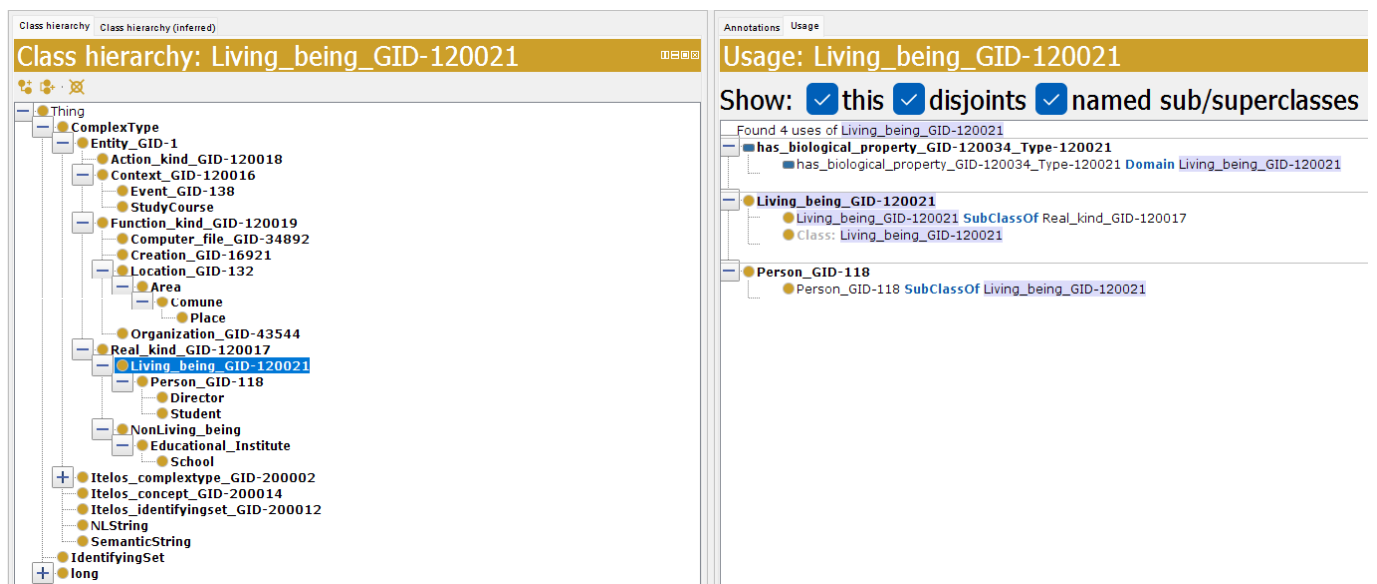


Figure 7: Person Class

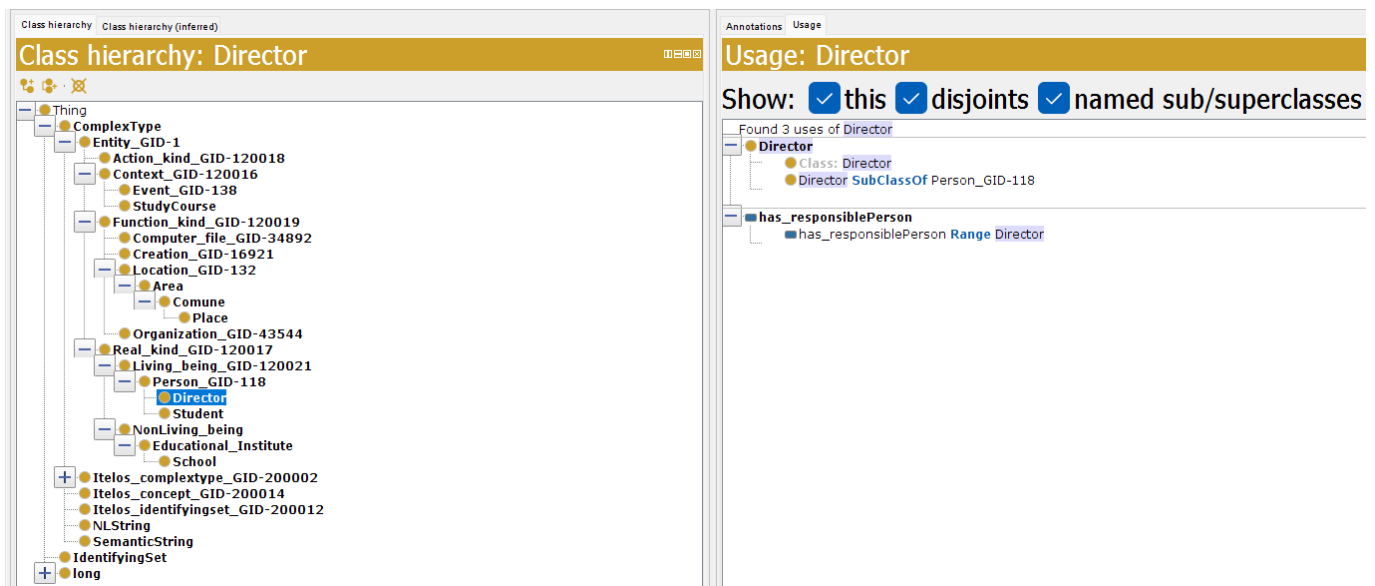


Figure 8: Director Class

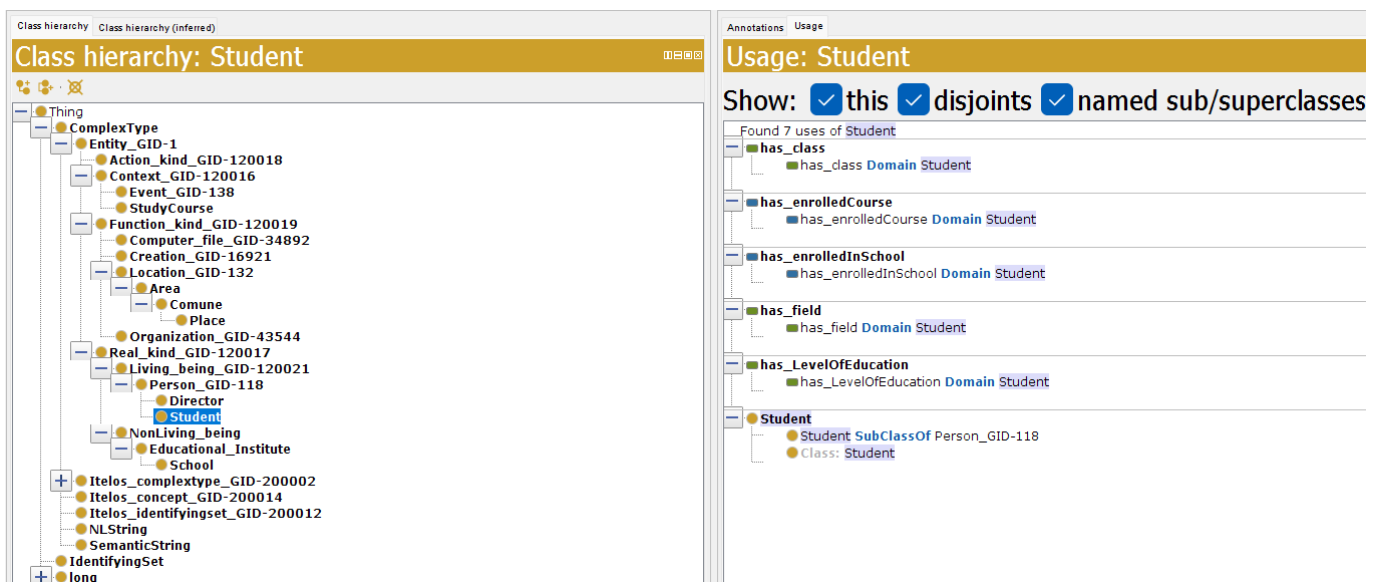


Figure 9: Student Class

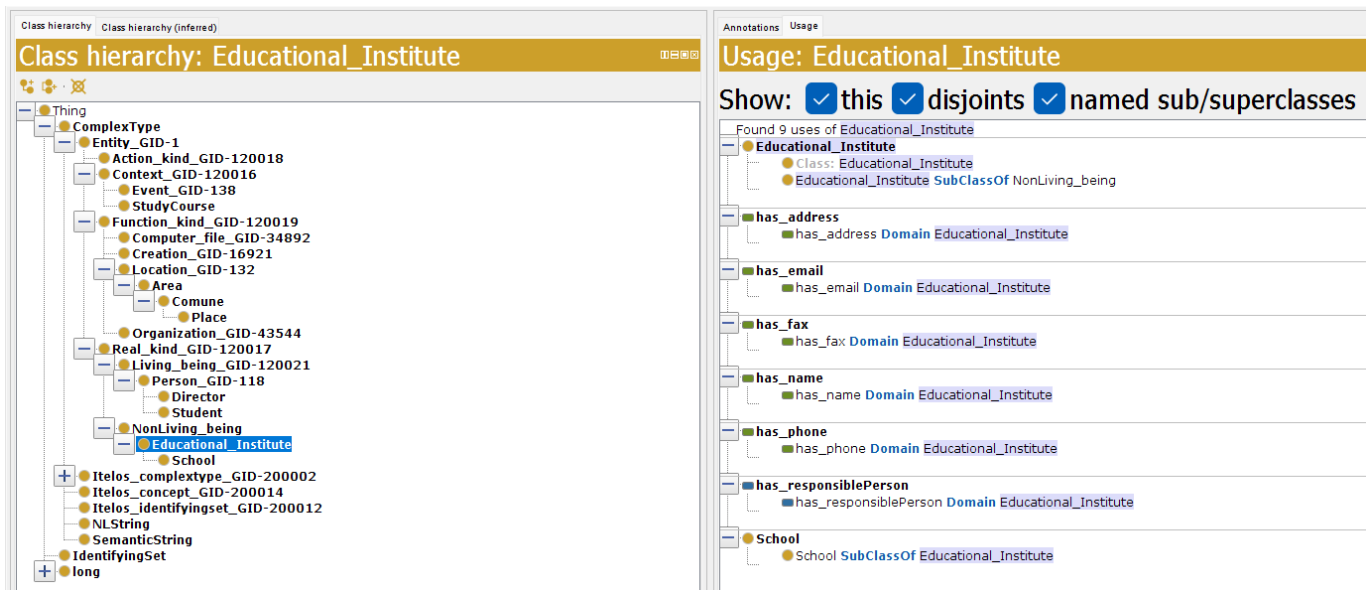


Figure 10: Educational Institute Class

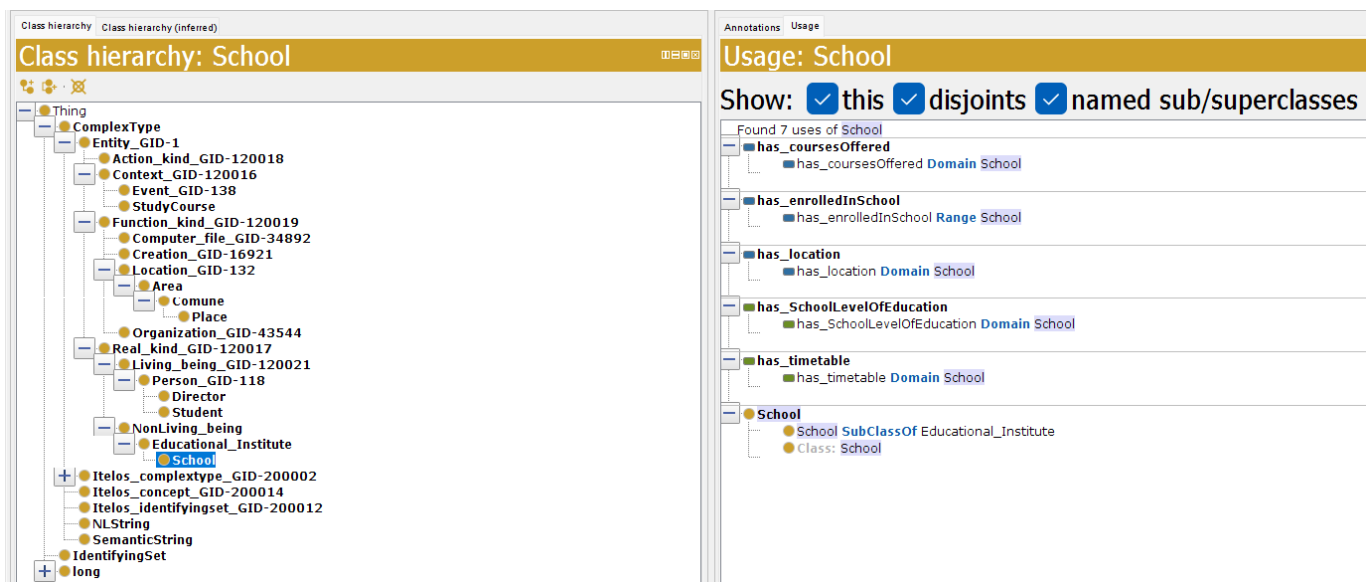


Figure 11: School Class

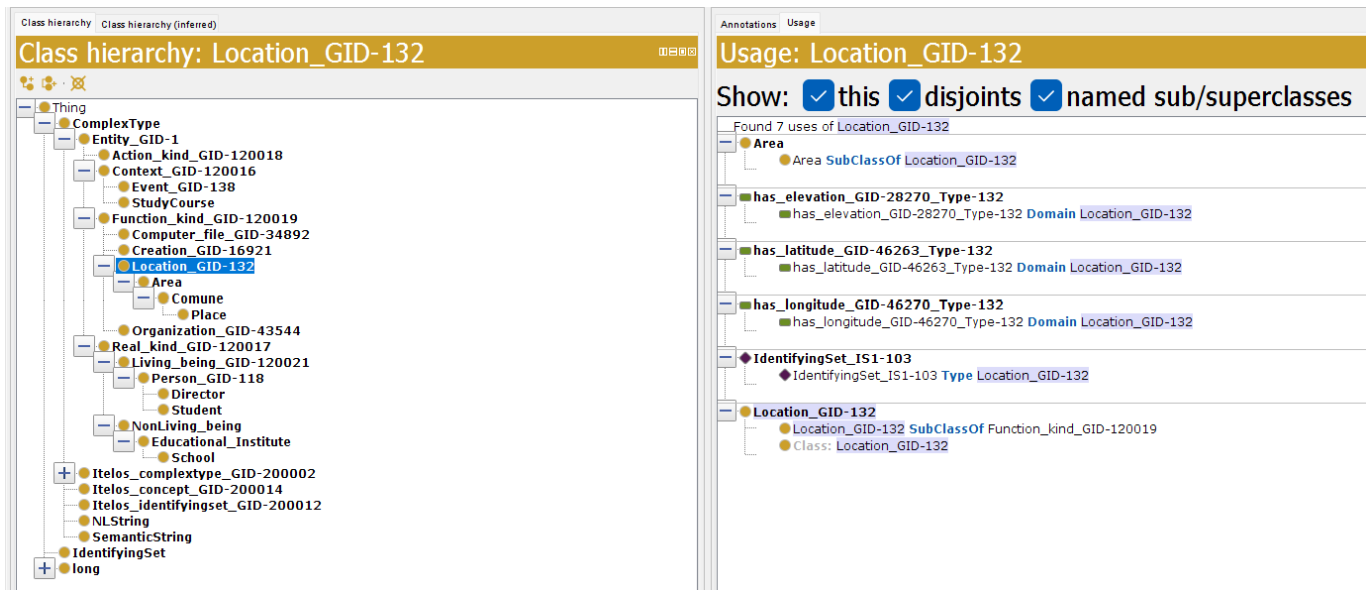


Figure 12: Location Class

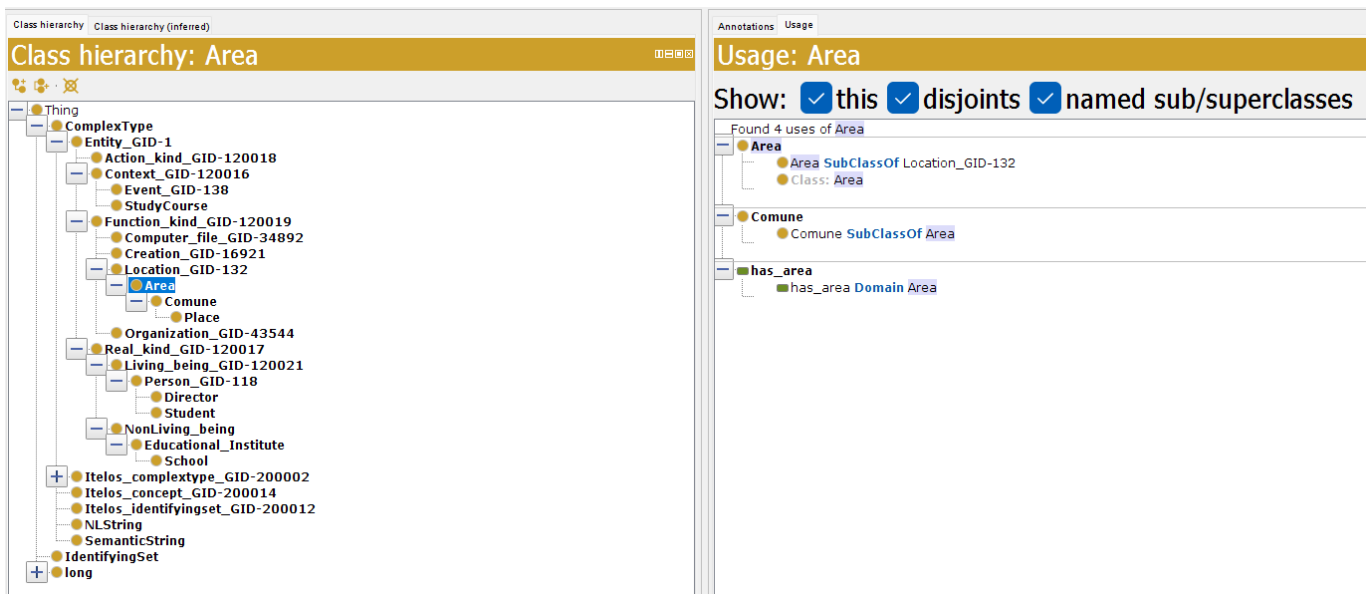


Figure 13: Area Class

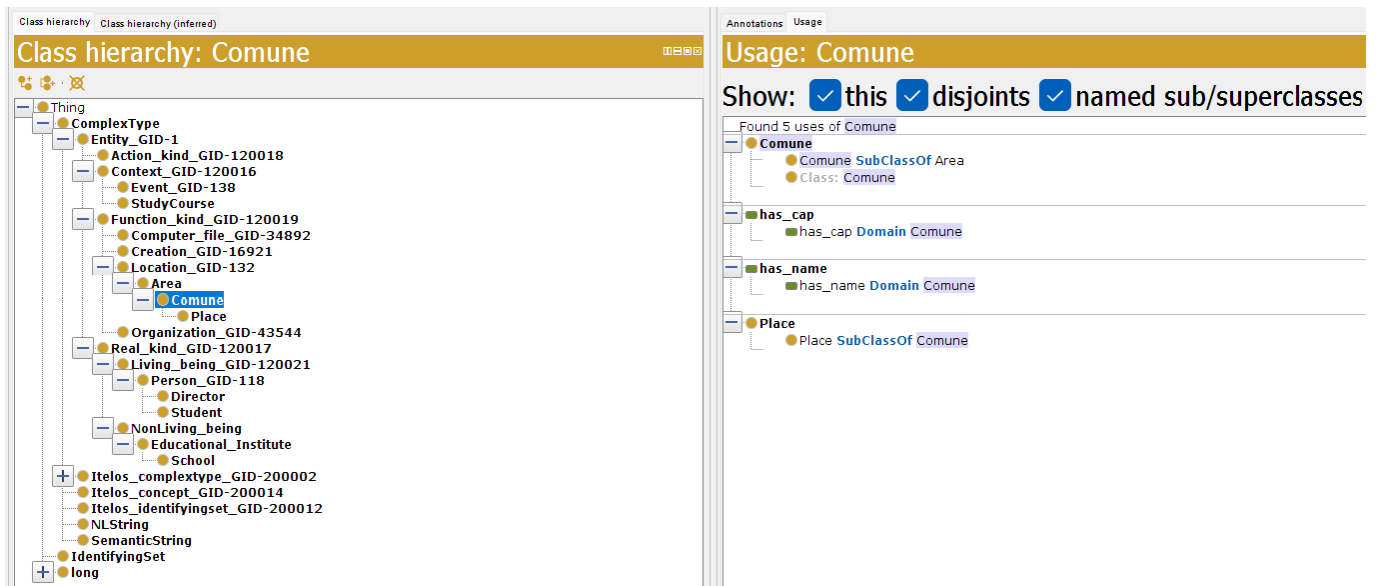


Figure 14: Comune Class

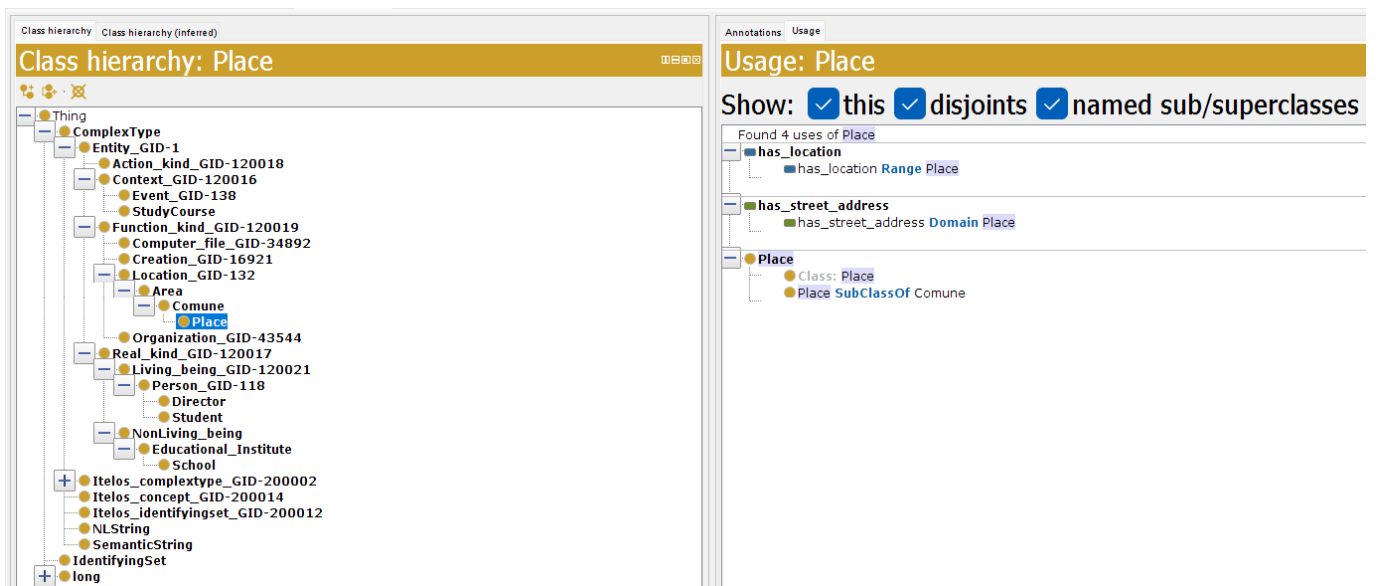


Figure 15: Place Class

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 without any pair of data-sets containing same piece 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.

6 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.