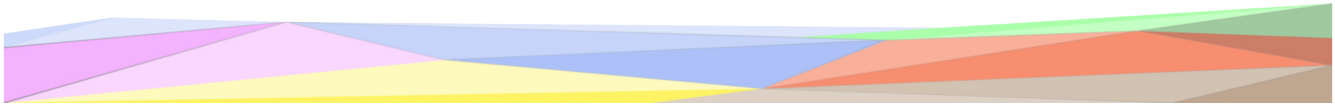
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Integration of medical data on Covid-19

Document Data:	Reference Persons:
October 21, 2020	Nisha Antony
	Daniel Gotca
	Maria Jyate
	Lorenzo Donini



Contents

1 Knowledge Graph Development Process

The goal of this part of the document is to describe the knowledge graph process. For its development, the iTelos methodology has been chosen, that guide us step by step through the process. First, it is important to focus on the contest and the problem we have to solve. Then, it comes the personas definition that can benefit from the knowledge graph. After that, the Inception phase has been described as well as the Competency Queries (**CQs** from now on) and the datasets used.

1.1 Scope Definition

This section aims to define the purposes for the creation of the Knowledge Graph, describing the context in which it has to live as well as the usage scenarios in which it can be involved. Moreover, a list of general questions regarding the objectives to achieve through the development of the Knowledge Graph, is reported here. More in general this section has to give all those information which allow to understand which is the problem to solve, and why we need a KG to solve it.

1.1.1 Problem Context Definition

The goal of "Integration of medical data on Covid-19" project is to group all the implicit and explicit data. Thanks to this integration, it would be possible to better understand the diffusion of the Covid-19 virus in the Trentino Region. Especially in this historical period, localizing new epidemic centers of the virus is a fundamental factor to limit the diffusion.

The solution that we propose is an integration of data regarding places and any kind of point of interest. In order to obtain a complete data collection, we want to integrate data about Trentino facilities as well as data about the situation in the neighbor countries.

Such countries has been considerate due to the fact that Trentino attracts tourist from all over the place, with a major presence of people coming from France, Swiss, Germany, and UK. The high number of tourist attraction lead to an higher risk of new infections: this means that it is also important to know how the pandemic situation is in those countries. It is relevant to understand not only the situation, but how other nations act in regard of Covid-19 virus[1]. In a region that thrives on tourism the economic losses would be high if the virus is underestimated.

1.1.2 Personas

Ranjith is a young man of 21 years old from India. He is a student at the University of Trento and loves to travel. Every time possible he takes a chance to go and find new cities to discover through art, museum, culture and food. Next week Ranjith want to go on trip but he has not decided yet where to go due to the health situation since he might get stuck in another city. He also want to keep monitored the situation in his home country to know if would be possible or not for home to go visits his family during the winter break.

Franco is a middle-aged man who owns a nice hotel near ski slopes. During the summer, his clients come to go hiking in the mountains while in the winter people come to ski. Usually the hotel guests come from all over Europe, from France, Swiss, Germany, Austria, and UK especially. Franco wants to keep monitored the situation

because he would like to have a clear idea on how he should manage the winter season. If the number of people infected grows, he might as well close for now in order to let the numbers go down and reopen at the beginning of 2021.

Gianni is a school principle of a complex of an elementary and a middle school. In this historical moment, schools have to be particularly cautious with the health situation. They have to monitor the incidence of the cases of Coronavirus in order to make decisions about the didactic system and the students and staff safety. Since the end of the summer, the number of infection and of false alarms caused by seasonal cold grow, school need to find a solution to allow all the student to follow the lessons without dangerous exposure to the virus. School also need to take action to avoid at all cost online classes since most of the students are children that need the supervision of the parents or tutors in order to follow correctly the lessons.

Marta is an employee at the local transportation agency and her job is to manage and guarantee the safety and security of all the staff and passengers. She has to verify if all the guidelines are respected among all the urban and suburban lines that connect the province of Trento. With the approaching of the winter season, and the increasing number of tourists in the province of Trento the public transportation load increases. There is the need to optimize the circulation of the vehicles in order to avoid excessive loads and to sanitize the busses and the trains.

Carla is the manager at a private RSA. She wants to monitor day by day the pandemic situation in her structure and outside to allow visits from family members. Since the guest of the structure are mainly elderly people, so the part of the population subject to the major threat from the COVID-19, Carla has to be particularly cautious. Some of the guest of the structure wants to be able to see their family members: this will be possible only if the regional situation for what concern the infections rate is low and Carla should be able to keep it monitored. Moreover, the staff component of the RSA has to be constantly under watch to keep track of their health state. The staff has to be tested once in a week or two depending on the pandemic state to avoid causing a coronavirus site.

1.2 Inception

This section is dedicated to the Inception phase description. Here are reported the initial definitions for CQs (Competency Queries), initial datasets collected and the relative metadata. For each of those elements the procedures and the tools adopted to achieve the results, have to be reported in the sections below.

1.2.1 CQs definition

This subsection is dedicated to the definition of the Competency Queries. They have to be listed and explained with details in order to have the information they bring, as clear as possible. This section plays a crucial role in the project description due to the fact that the CQs are the starting point to define the single objects/entities involved in the KG. For this reason the CQs will be used in the next phases as evaluation base to define the quality of the outcomes of each phase.

Persona	Number	Question	Action
Ranjith	1.1	Give the progress of Covid-19 cases in Sicily of the last days/weeks	Return the data of the last week of cases in Sicily
Ranjith	1.2	Give the progress of Covid-19 cases in Trento	Return the data of the last week of cases in the province of Trento

Ranjith	1.3	Give the risk of lockdown in Trentino	Return the estimation of infection in Trentino
Ranjith	1.4	Give the risk of lockdown in Sicily	Return the estimation of infection in Sicily
Ranjith	1.5	There are any travel limitations in Sicily?	Search and return if there are any limitations
Ranjith	1.6	How is the situation in India?	Return the data of infection in the last week in India
Ranjith	1.7	There are any travel limitations in india?	Return the list of information about travel policy for country outside EU
Franco	2.1	What is the progress of Covid-19 cases in trentino?	Return the data of the last 3 days of cases in trentino
Franco	2.2	Give the number of possible tourists in trentino	Return the prediction of cases and mobility in trentino
Franco	2.3	Give the progress of Covid-19 cases from tourist's country (es France)	Return the number of cases in the last week of the France
Franco	2.4	Give the progress of Covid-19 cases from tourist's country (es Germany)	Return the number of cases in the last week of the Germany
Franco	2.5	Give the progress of Covid-19 cases from tourist's country (es Austria)	Return the number of cases in the last week of the Austria
Franco	2.6	Give the progress of Covid-19 cases from tourist's country (es UK)	Return the number of cases in the last week of the UK
Franco	2.7	Are there any travel limitations in Italy?	Search and return if there are or not limitations in the Italy
Franco	2.8	Are there programmed lockdowns in Trentino for the next month?	Search and return if there are any lockdowns scheduled.
Gianni	3.1	How many cases in trentino?	Returns the number of cases in trentino
Gianni	3.2	How many cases in schools in trentino?	Given the region number of cases in schools
Gianni	3.3	Are there schools that are closing?	Return whether schools are closing in the region
Gianni	3.4	Is the situation growing fast?	Return the data are there lockdowns in other regions?
Gianni	3.5	Search and return if there are or not lockdowns will the cases grow in the future weeks in Trentino?	Return the prediction of cases in trentino for the next weeks
Gianni	3.6	Will the number of cases grow in italy?	Return the prediction of cases in italy
Marta	4.1	Which is the progress of covid cases in trentino?	Return the data of the last 3 days of cases in trentino
Marta	4.2	Do people move more or less?	Return mobility how many cases in italy?
Marta	4.3	How many cases in the border countries from which trains arrive?	Return the data of the last 3 days of cases in foreign countries

Marta	4.4	Are there travel limitations in italy?	Search and return if there are or not limitations in the country
Marta	4.5	Will the numer of cases grow in the future weeks in Trentino?	Return the prediction of cases in trentino for the next weeks
Carla	5.1	How is the situation growing in trentino?	Return the data of the last 3 days of cases in trentino
Carla	5.2	How many cases are in the RSA in trentino?	Return number of cases in RSA in trentino
Carla	5.3	Give the risk of lockdown in trentino	Return the estimation of infection
Carla	5.4	Are there travel limitations in trentino?	Search and return if there are or not limitations
Carla	5.5	Will the number of cases grow in the future weeks in Trentino?	return the prediction of cases in trentino for the next weeks

The following table aims to link the **CQs** to the type of data that has to be retrieved.

NUM	TYPE	PROPERTIES
1:1-2-6, **2:**1-3-4-5-6, **3:**1-2, **4:**1-3, **5:**1-3	Covid Status	Date, Total number of cases, Number of active cases, Number of new positive cases, Number of deaths, Number of recovered cases
1:5-7, **2:**7, **4:**2-4, **5:**4	Travel Restrictions	Location, Severe travel restriction period
1:3-4, **2:**2-8, **3:**4-5, **4:**2-4, **5:**4	Lockdown	Location, Date, Est.infections, Stay at home period
3:3, **4:**2-4, **5:**4	Institutional restrictions	Location, School/University closure period
4:2-4, **5:**4	Business Restrictions	Location, Business closure period
4:2-4, **5:**4	Non-essential Business Restrictions	Location, Non-essential Business closure period
4:2-4, **5:**4	Gathering restrictions	Location, Gathering restriction period
5:2	RSA Cases	Date, Number of cases in RSA, Number cases in Home care, Total number of cases in RSA
3:6, **4:**5, **5:**5	Case Projections	Location, Date, Mean of Est.infections, Lower bound of Est.infections, Upper bound of Est.infections

1.2.2 Initial Datasets description

Reported below as tables there are all the details for what concern the datasets used in this activity.

- COVID-19 Coronavirus data [4]

Field name	Description
Title	Covid-19 Coronavirus data
Description	The European Centre for Disease Prevention and Control(ECDC) has created this dataset by collecting reports from health authorities worldwide ever since the Covid-19 outbreak. The data is updated on a daily basis and is available in CSV format. The dataset contains cumulative daily status regarding the pandemic across the world.
Category	Health
Keywords	COVID-19, disease outbreak, corona virus, SARS-CoV-2, coronavirus, severe acute respiratory syndrome coronavirus-2
Last update	2020-10-20
Publisher	European Centre for Disease Prevention and Control (ECDC)
Contact Information	ECDC European Centre for Disease Prevention and Control https://www.ecdc.europa.eu/en/about-ecdc
Frequency	Daily
Temporal Coverage	2019-12-31
Spatial Coverage	Europe, Asia, Africa, America, Oceania

- COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections [5]

Field Name	Description
Title	COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections
Description	IHME's COVID-19 projections were developed in response to requests from the University of Washington School of Medicine and other US hospital systems and state governments working to determine when COVID-19 would overwhelm their ability to care for patients. The forecasts show demand for hospital services, daily and cumulative deaths due to COVID-19, rates of infection and testing, and the impact of social distancing, organized by country and state (for select locations).
Category	Health
Keywords	COVID-19, disease outbreak, coronavirus
Last update	2020-10-15
Publisher	Institute for Health Metrics and Evaluation (IHME).
Contact Information	covid19@healthdata.org

Temporal Coverage	2020-02-04
Spatial Coverage	Global

- COVID-19 emergency health situation: Province of Trentino [2]

Field Name	Description
Title	COVID-19 emergency health situation: Province of Trentino
Description	The datasets include the clinical status and current situation of Trentino municipalities. The data is in Italian and is available in CSV format. The information is collected by the Trentino Digitale from the health centres to keep residents up-to-date regarding the Covid situation.
Category	Health
Keywords	COVID-19, coronavirus, Trentino data, Covid-19 Trentino
Last update	2020-10-20
Publisher	Trentino Digitale
Contact Information	https://www.trentinodigitale.it/
Spatial Coverage	Province of Trentino

1.2.3 Datasets metadata documentation

- COVID-19 Coronavirus data [4]

Variable	Description
dateRep	Date of report
day	Date element from the date of report
month	Month from the date of the report
year	Year from the date of report
cases	Number of new cases on the date of the report
deaths	Number of deaths on the date of the report
countriesAndTerritories	Country and territory name related to the records
geoId	Country code based on ISO alpha -2 STDs
countryterritoryCode	Country code based on ISO alpha -3 STDs
popData2019	Population of each country as of 2019
continentExp	Name of the continent to which the country belongs to
Cumulative_number_for_14_days_of_COVID-19_cases_per_100000	Cumulative number of Covid cases for 14 days per 100K

- COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections [5]

  1. Reference_hospitalization_all_locs dataset

Variable	Description
location$_name$	Name of the country or subnational location
date	Date
allbed_mean	Mean covid beds needed by day
allbed_lower	Lower uncertainty bound of covid beds needed by day
allbed_upper	Upper uncertainty bound of covid beds needed by day
ICUbed_mean	Mean ICU covid beds needed by day
ICUbed_lower	Lower uncertainty bound of ICU covid beds needed by day
ICUbed_upper	Upper uncertainty bound of ICU covid beds needed by day
InvVen_mean	Mean invasive ventilation needed by day
InvVen_lower	Lower uncertainty bound of invasive ventilation needed by day
InvVen_upper	Upper uncertainty bound of invasive ventilation needed by day
admis_mean	Mean hospital admissions by day
admis_lower	Lower uncertainty bound of hospital admissions by day
admis_upper	Upper uncertainty bound of hospital admissions by day
newICU_mean	Mean number of new people going to the ICU by day
newICU_lower	Lower uncertainty bound of the number of new people going to the ICU by day
newICU_upper	Upper uncertainty bound of the number of new people going to the ICU by day
bedover_mean	[covid all beds needed] - ([total bed capacity] - [average all bed usage])
bedover_lower	Lower uncertainty bound of bedover (above)
bedover_upper	Upper uncertainty bound of bedover (above)
icuover_mean	[covid ICU beds needed] - ([total ICU capacity] - [average ICU bed usage])
icuover_lower	Lower uncertainty bound of icuover (above)
icuover_upper	Upper uncertainty bound of icuover (above)
deaths_mean	Mean daily covid deaths
deaths_lower	Lower uncertainty bound of daily covid deaths
deaths_upper	Upper uncertainty bound of daily covid deaths
totdea_mean	Mean cumulative covid deaths
totdea_lower	Lower uncertainty bound of cumulative covid deaths
totdea_upper	Upper uncertainty bound of cumulative covid deaths
deaths_mean_smoothed	Mean daily covid deaths (smoothed)
deaths_lower_smoothed	Lower uncertainty bound of daily covid deaths (smoothed)
deaths_upper_smoothed	Upper uncertainty bound of daily covid deaths (smoothed)
totdea_mean_smoothed	Mean cumulative covid deaths (smoothed)

totdea_lower_smoothed	Lower uncertainty bound of cumulative covid deaths (smoothed)
totdea_upper_smoothed	Upper uncertainty bound of cumulative covid deaths (smoothed)
mobility_data_type	Indicator of whether mobility composite is observed/projected
mobility_composite	Mobility composite score
total_tests_data_type	Indicator of whether total tests composite is observed or projected
total_tests	Total tests
confirmed_infections	Observed data only (confirmed infections)
est_infections_mean	Mean estimated infections
est_infections_lower	Lower uncertainty bound of estimated infections
est_infections_upper	Upper uncertainty bound estimated infections

2. Summary_stats_all_locs.csv

Variable	Description
location_name	Name of the country or subnational location
peak_bed_day_mean	Mean peak bed use date
peak_bed_day_lower	Lower uncertainty bound of peak bed use date
peak_bed_day_upper	Upper uncertainty bound of peak bed use date
peak_icu_bed_day_mean	Mean peak ICU bed use date
peak_icu_bed_day_lower	Lower uncertainty bound of peak ventilator use date
peak_icu_bed_day_upper	Upper uncertainty bound of peak ventilator use date
peak_vent_day_mean	Mean peak ventilator use date
peak_vent_day_lower	Lower uncertainty bound of peak ventilator use date
peak_vent_day_upper	Upper uncertainty bound of peak ventilator use date
all_bed_capacity	Total number of beds that exist at that location
icu_bed_capacity	Total number of ICU beds that exist at that location
all_bed_usage	Average number of total beds used normally at that location
icu_bed_usage	Average number of ICU beds used normally at that location
available_all_nbr	All_bed_capacity - all_bed_usage: excess bed capacity at that location
available_icu_nbr	Icu_bed_capacity - icu_bed_ usage: ICU excess bed capacity at that location
travel_limit_start_date	Start date for Severe travel restrictions
travel_limit_end_date	End date for Severe travel restrictions
stay_home_start_date	Start date for People ordered to stay at home
stay_home_end_date	End date for People ordered to stay at home
educational_fac_start$_date$	Start date for Educational facilities closed

educational_fac_end_date	End date for Educational facilities closed
any_gathering_restrict_start_date	Start date for Any gathering restrictions
any_gathering_restrict_end_date	End date for Any gathering restrictions
any_business_start_date	End date for Any business closures
any_business_end_date	End date for Any business closures
all_non-ess_business_start_date	Start date of Non-essential businesses ordered to close
all_non-ess_business_end_date	End date of Non-essential businesses ordered to close

- COVID-19 emergency health situation: Province of Trentino [2] (Stato_clinica_td.csv)

Variable	Description
giorno	Date of the report
domicilio	Number of home isolated cases
infettive	Number of infectious cases
alta$_i nt$	Number of high-intensity cases
terapia$_i n$	Number of intensive care cases
guariti	Number of recovered cases
deceduti	Number of deaths
totale_pos	Total number of positive cases
pos_att	Number of active cases
rsa	Number of cases in rsa
tot_prec	Total number of cases in the previous day
incremento	New positive cases (difference between totale_pos & tot_prec)
casa_cura	Number of cases in homecare
strut_int	Number of high-intensity structures
tot_rsa	Total number of rsa cases
dimessi	Number of discharged cases
tot_dime	Total number of discharged and healed cases
nuovi	Number of new admissions
nuo_screen	Number of new cases under observation

1.2.4 Datasets collection process

The collection of initial datasets was fairly easy as a lot organisations, Universities and sectors are studying the Covid-19 data to derive patterns and insights. The data we gathered till now is available on the public domain in which some are synthetic data which is produced as a result of the study conducted by IHME research team [5]. The data forecasts the demand of hospital services, cumulative deaths, rates of infection, impact of social distancing and so on. The Covid-19 Coronavirus data is an open data created by The European Centre for Disease Prevention and

Control(ECDC) for conducting research on pandemic across the world [2]. The Covid-19 data from the Province of Trentino is also an open data collected by Trentino Digitale [4]. All of these datasets are available in the public domain and was downloaded from their respective portals.

The iterative process of data collection, verification and realization of entities are explained in the following sections.

1.2.4.1 Iteration Zero In the zeroth iteration, at the schema level, we defined the Common informal Etypes such as demographic information from the data objects identified inception. At the data level, we verified whether the datasets selected are meeting the requirements of the CQs. Additional datasets were thus collected from various sources to answer the queries.

1.2.4.2 Iteration First In the first iteration, at the schema level, we defined the Core informal Etypes such as the Covid cases and realized it further into atomic entities. In the data level, the collected datasets are verified for the Common Etypes from schema level output. The Demographic Covid dataset and the IHME's projection dataset collected for the Global, Country-level and sub-region level meet these requirements.

1.2.4.3 Iteration Second In the third iteration, at the schema level, we identified and defined the Contextual informal Etypes. The location type and the restriction type were identified as the contextual entities. In the data level, the datasets collected are verified for the Core Etypes from schema level output. The dataset regarding Trentino records is collected to meet these requirements.

1.2.4.4 Iteration Third In the third iteration, at the schema level, we check for any missing informal Etypes or related attributes. In the data level, the datasets collected are verified for the Contextual Etypes from schema level output. The dataset collected in the previous iterations accounts the information need at this level.

1.2.5 Inception level evaluation

The last section of the Inception phase report the evaluation of the outcomes obtained in this phase, through specif evaluation metrics.

1.3 Informal Modeling

This section is dedicated to the Informal Modeling phase description. The section is divided in Schema and Data level in order to report the details of the elements involved in the generation of the schema, as well as the description of the datasets evolution in this phase. Moreover a specif section, one for each level, reports the difference between the elements defined in this phase and the definitions in the previous phase, analyzing in this way the variance in the different phases.

1.3.1 Schema level

The goal of this section is to provide a conceptual data model called EER (enhanced entity relationship) model. To do that we have used the yEd software. We started from Table 2 that contains the information about the type of data that has to be retrieved from the competency queries. The column of the "type" represents the "entities" while the column of the "properties" represents the "attribute" of the entities.

The EER model is an expanded version of the ER model: this is very helpful for the design phase of databases with high-level models. Moreover, with the enhanced features, it is possible to plan databases more thoroughly by delving into the properties and constraints with more precision. The elements of an EER diagram are:

- Entities objects or concepts represent important data. The entity is represented by a rectangular and can be of three types: core, common and contextual. The core entities are the most important entities regarding the project's solution. These entities are in the color blue. The common entities have the strongest impact in terms of dependencies and links core entities to each other. These entities are represented in yellow in the diagram. The contextual entities (red in the diagram) are the specification of the core entities.
- Attributes are characteristics of an entity and are written in the bottom part of the rectangular.
- Relationships are associations between entities and can be represented by a line or by an arrow and usually described through a verb.

EER diagrams[1] are perfect for taking a more detailed look on this information.
In our case, the two most important entities of this EER model, which are also both core entities, are "Restriction" and "Cases Information".
The entity "Restriction" has four attributes: "restriction" of type "**RestrictionType**" that can select six different type of restriction, the date when the quarantine started and the period of it ("closure_start" of type "Date", and "period" of type int) and the attribute "location" of type "**Location**". In general, this entity represents the period, the location and the type of restriction that affected part of the population.
The contextual entity "**RestrictionType**" has an attribute that can be of six types and describe which are the activity with some limitation: for the activity, "*Essential Business*", "*Non Essential Business*", "*Travel*", "*Institutional*", "*Gathering*", and "*Lockdown*" for common people.
The core entity "**Restriction**" is then linked to two entities of common type: "**Location**" and "**Date**". The entity "**Location**" is represented by a "*name*", the "*restriction*" and then the attribute "*location*" of the type "**LocationType**". This entity is of type contextual since it explain the typology of the location: it has an attribute and it can be of four different categories: "*Nation*", "*Region*", "*City*", and "*Institution*".
The other common entity is "**Data**" and it is represented through three attributes: "*Day*", "*Month*" and "*Year*". These two common entities, that are linked to "**Restriction**" entity, are also linked to the core entity called "**Cases Information**". This entity has five attribute: "num_cases", "num_deaths", the "location" of the type "**LocationType**" and the "date" of the type "Date".
The contextual entity "**Covid Status**", in addition to including the attributes of entity "**Cases Information**", contains the total number of positive "num_ active", the number of new positive "num_new_pos" and the number of recovered "num_ recovered)". The core entity "**School Cases**", in addition to including the attributes of entity "*Cases Information*", contains the identification "school_id", the name "school_name", the actual number of cases in the school "total_cases_schools", and the total number of classes in quarantine in the schools "num_classes_cases".
The core entity "**RSA Cases**", in addition to including the attributes of entity "*Cases Information*", contains the identification "RSA_id", the name "RSA_name".
The contextual entity "**Cases projection**", in addition to including the attributes of entity "*Cases Information*", contains the lower bound estimation "lower_bound_est" and the upper bound estimation of cases "upper_bound_est" for Covid-19 virus.

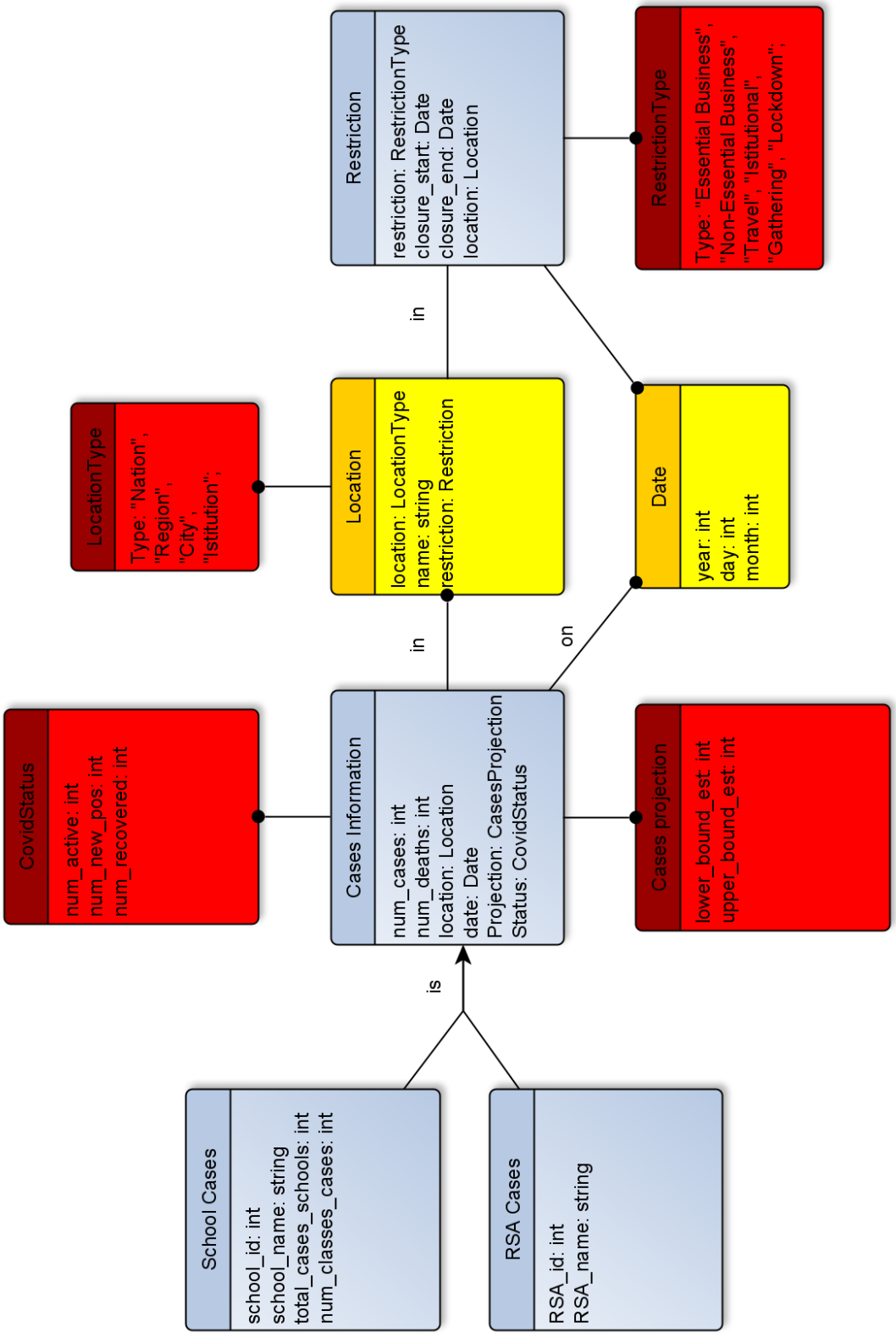


Figure 1: EER Diagram of the data used inside the project

1.3.1.1 Variance respect CQs definition

This section aims to define the variance between the schema elements produced in this phase, and the definition of the CQs reported in the previous phase. This a way to define the quality of the outcomes for the current phase as well as the alignment of the overall project development process.

1.3.2 Data level

The data level section in this phase reports the evolution of the datasets collected previously, reporting the metadata information for each new data, or new version of data, obtained.

1.3.2.1 Datasets management process

During the Informal Modeling phase the datasets collected in the previous phase are filtered and managed in order to obtain more suitable sets of data.

As we know the datasets collected from various resources can be really dirty. It will have both relevant and irrelevant information. When you have enormous amount of data to play with and integrate, the performance of our system will hinder because of the size. In short, yes the size matters and it is wise to get rid of the useless information that we have in our datasets. The filtration process is done by taking the CQs and corresponding outcomes into consideration.

The datasets we have are in Excel and CSV formats which can be easily handled and filtered out by the using Python programming language. We used Pandas package to manipulate the structured datasets in a dataframe and filtered out the irrelevant attributes. The resultant dataframe was again stored as a CSV file in the repository so that it can be used as an input in the upcoming phase.

1.3.2.2 Datasets metadata documentation

In this section is reported a list of new metadata in order to describe the modification performed on each datasets and attribute, to achieve the new version of the datasets.

1.3.2.3 Variance respect Inception datasets

This section aims to define the variance between the data elements (datasets and attributes within them) produced in this phase, and the initial datasets collected in the previous phase. This a way to define the quality of the outcomes for the current phase as well as the alignment of the overall project development process.

In this phase, we filtered out several irrelevant attributes from the collected datasets. This filtration is done by taking the CQs and corresponding outcomes into consideration. The data we have right now is far more clean and useful. We also transformed some of the excel files into CSV format as it is more flexible in the working environment. By removing those irrelevant attributes, now our datasets are more aligned to extract the knowledge required for answering the CQs. The result of this phase has not only cleansed datasets but also reduced the size to a great amount.

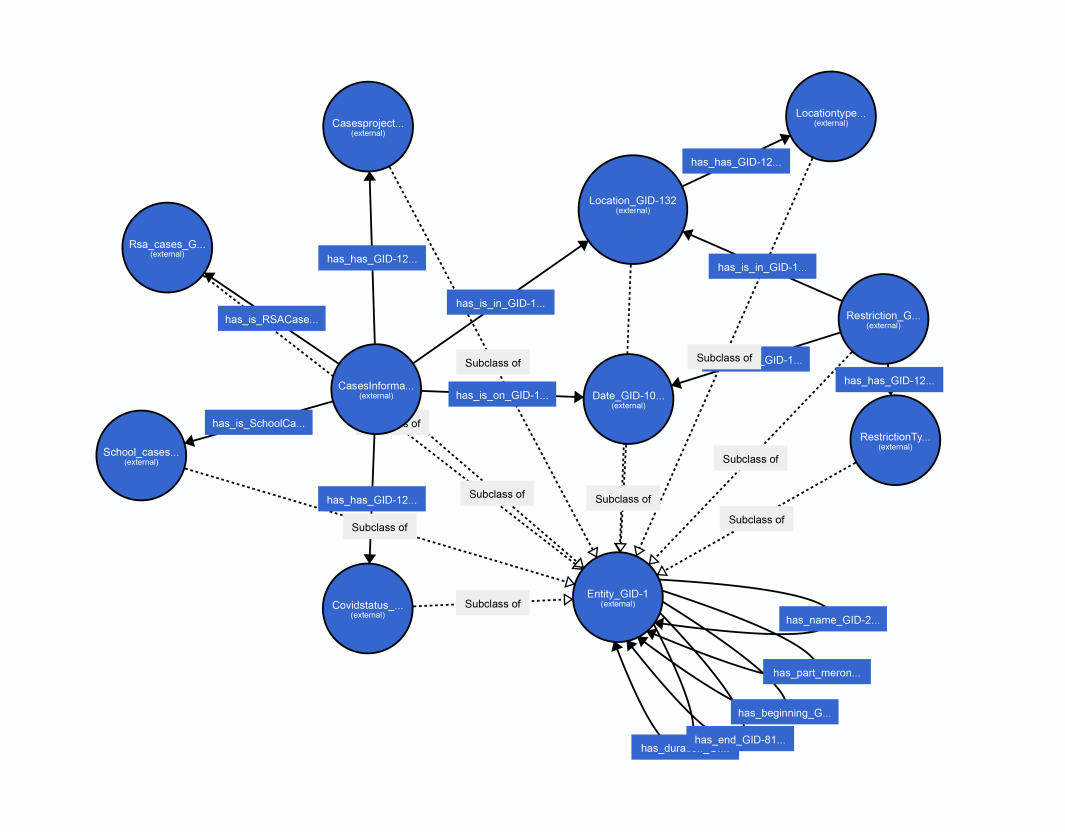


Figure 2: SKG Schema representing the ontology created

1.3.3 Informal Modeling Evaluation

The last section of the Informal Modeling phase report the evaluation of the outcomes obtained in this phase, through specif evaluation metrics.

1.4 Formal Modeling

This section is dedicated to the Formal Modeling phase description, which is the fourth phase of the iTelos Methodology. The main task of this phase consist of formalizing the whole model into an ontology in order to correctly represent the data that we want to work with.

The Section is divided in Schema and Data level in order to report the details regarding both the ontology generated and the datasets version in the current phase.

1.4.1 Schema level

The schema level section in the current phase, reports the detailed description of the ontology generation.
Starting from the EER model, we want to obtain the Schema Knowledge Graph (SKG Figure 2). The main objective is to relate on the iTelos teleology in order to help support and to ease the construction of the final formal model. The Schema Knowledge Graph is based on a data-driven requirement and uses standards already implemented thanks to the teleologies. The graph-like structure allow us to clearly see how the data interact with

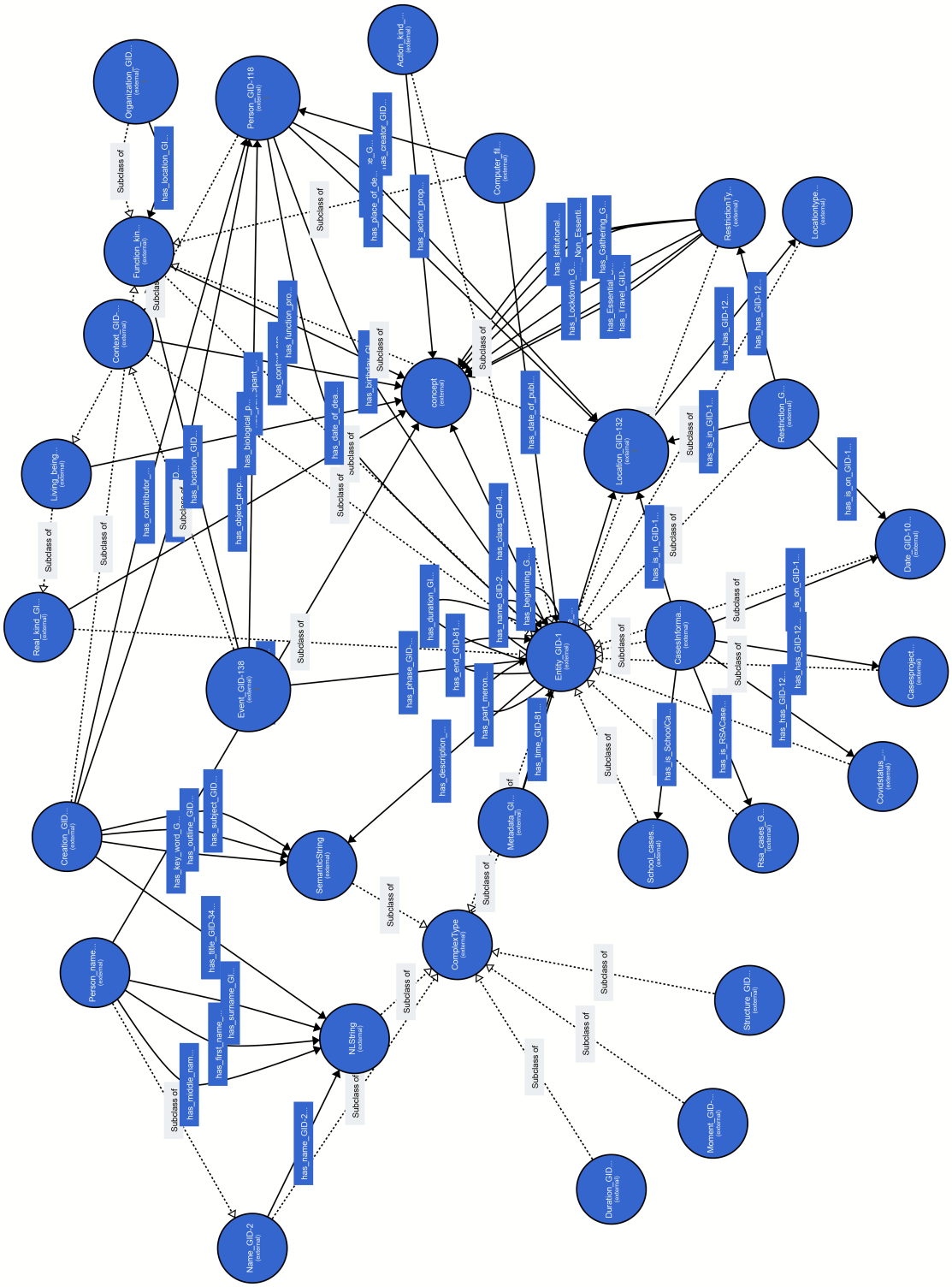


Figure 3: SKG Schema with the predefined ETypes

each other, with the nodes representing the entities and their attributes while the relations between these elements are viewed as edges. Both the entities and the relations are constrained by properties and logics already defined at the EER schema level.

1.4.1.1 Ontology definition

This section reports in details how the ontology is generated stating from the informal schema of the previous phase, which tools are used to do that, as well as usage of external ontology resources adopted to obtain the final KG ontology. Moreover a list of metadata is reported in this section, in order to describe all the elements of the ontology defined.

As mentioned above, this section aims to illustrate how the ontology was created. In order to do that tools such as Protege have been used to deliver the final result (more on that in the section below) . Starting from the EER model, which give a schematic representation of the structure, we want to represent the same data in a more complete and detailed way by building the ontology to entirely represent the database.

Tools

To build the ontology, several tools have been used.

- Protege, a free and open sourced ontology editor, developed by the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. [11, 10]
- KOS, a specilized tool that offers the representation, the visualization, and the properties in the study domain. [8]
- SWEB, a set of API that include the KB Importer and the RDI importer. [9]
- Karmalinker, a designing tool for data linking that allow us to align data to the ontology generated. [7]

In order to generate the ontology, process started by comparing the concept present in our EER schema with the ones already present in the KOS tool. Then, a list of the ETypes and the attributes that were missing in the KOS was generated as an excell file by using the API provided. After that, the already prehexisting ontology was downloaded as an RDF/XML file, alwasy from the API. We used Protégé [11] to add and modify the necessary classes with the corresponding concepts. The next stage was to add the required data and object properties regarding the Etypes present with the procedure already mentioned above. To add the enumeration we adopted a different method from the one published in this tutorial [6] since that methodology was not supported by the KOS system: so a different methodology has been used, the Enumeration Types were defined as a class while the Individuals were defined as the class attributes.

1.4.1.2 Classes

- **Cases information** is one of the two core entities, in fact as we can see from protege, it has several links (object properties) to other classes, starting from the two contextual entities (has has GID...) Covid Status and Cases Projection which eventually will work as additional attributes of this class, then the two sub classes (has is RSA...) RSA Cases and School Cases and finally a link to Location "has_is_in_..." and date (has is on...). The data properties of this class are number of deaths and number of cases.

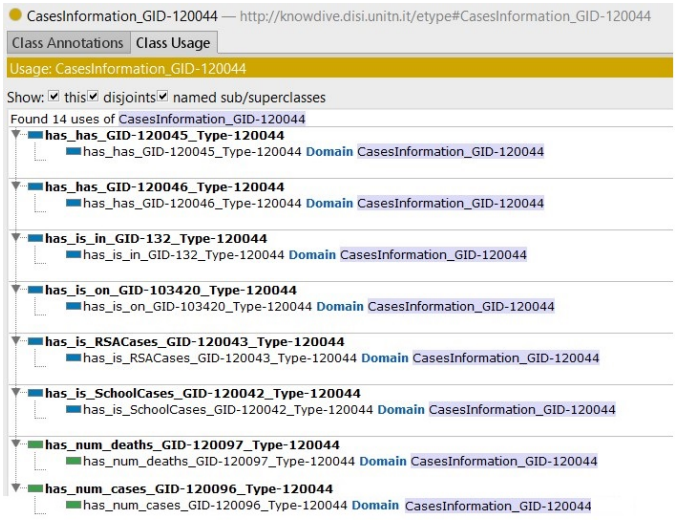


Figure 4: Cases Information in Protégé

- **Cases Projection** Is the contextual class which is needed to add information to Cases Information, it's attributes or data properties are lower and upper bound of the projection and it's linked to Cases Information

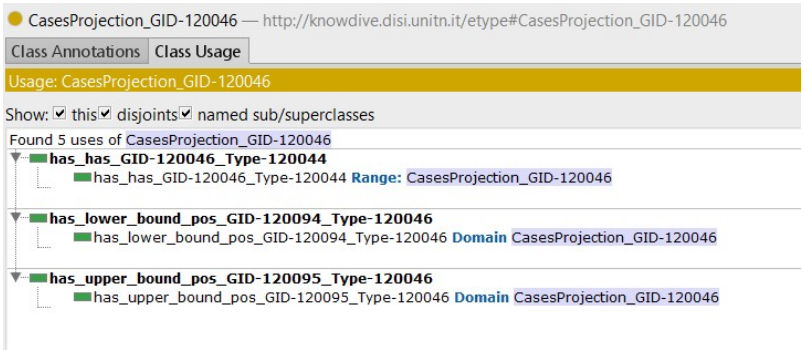


Figure 5: cases Projection in Protégé

- **RSACases** In addiction to Case Information data properties also contains the identification RSAid, the name.

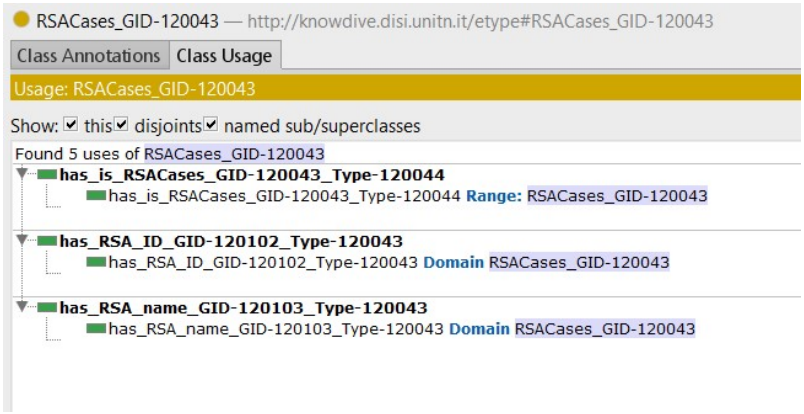


Figure 6: RSACases in Protégé

- **SchoolCases** inherits the attributes from Cases Information (so it obviously links to the core class) and in addiction it contains the identification schoolid, the name , the actual number of cases in the school, and the total number in a class in the school.

Figure 7: SchoolCases in Protégé

- **CovidStatus** is the other one contextual entity needed to complete Cases Information it's data properties are: number of active, number of new positives and number of recovered

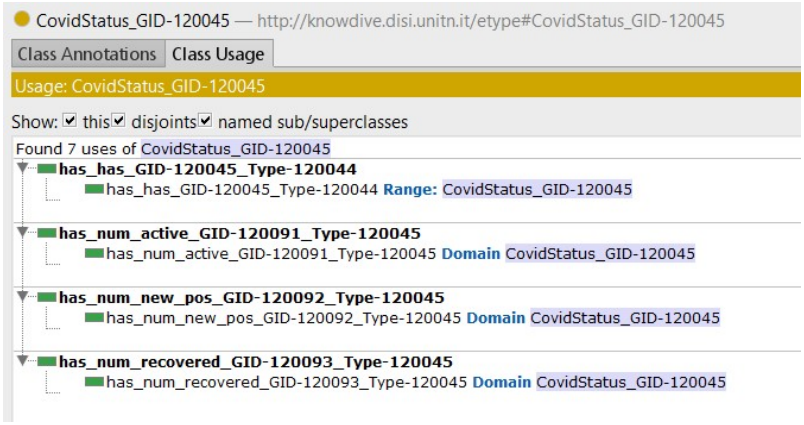


Figure 8: CovidStatus in Protégé

- **Date** is one of the two common entities since it's defined in many applications. It's linked to Covid Information and Restriction, the data properties are day month and year.

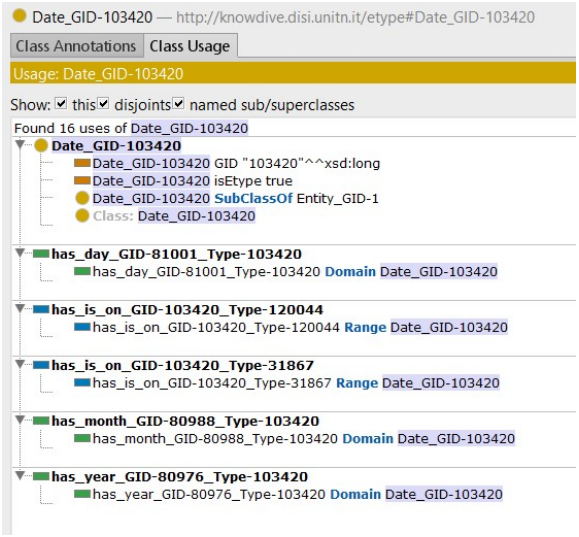


Figure 9: Date in Protégé

- **Location** is the entity that represents one of the link between the entities Restriction and Cases Information. It stands for the place.

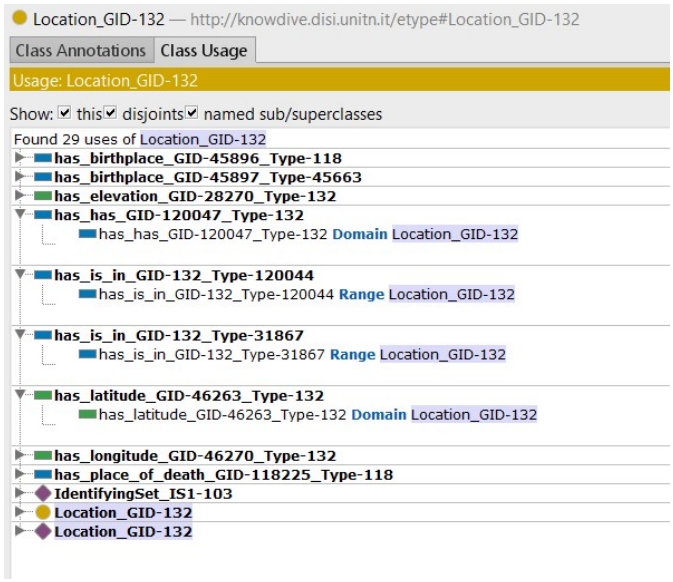


Figure 10: Location in Protégé

- **LocationType** is an EnumerationType of class that allow us to specify the place under certain categories. Its data properties are Nation, Region, City, and Institution.

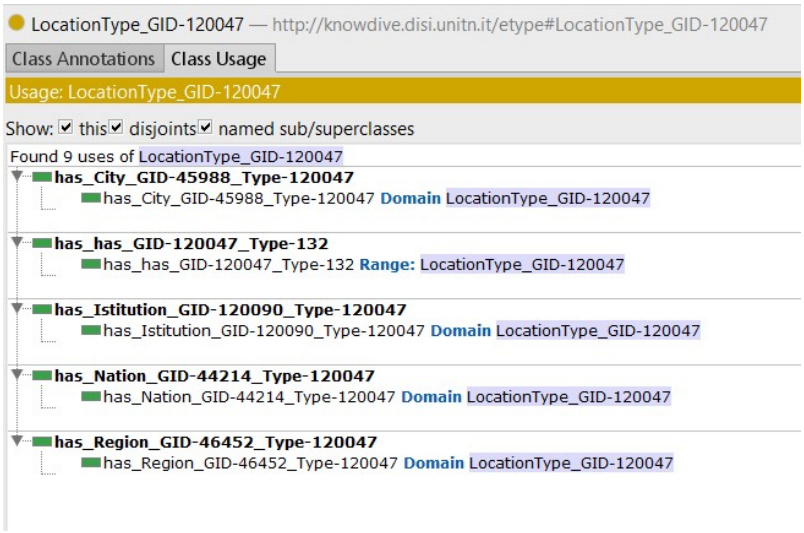


Figure 11: LocationType in Protégé

- **Restriction** represent all the limitations present in a given place. It is characterized by a date of start and a date of ending, representing a period of time.

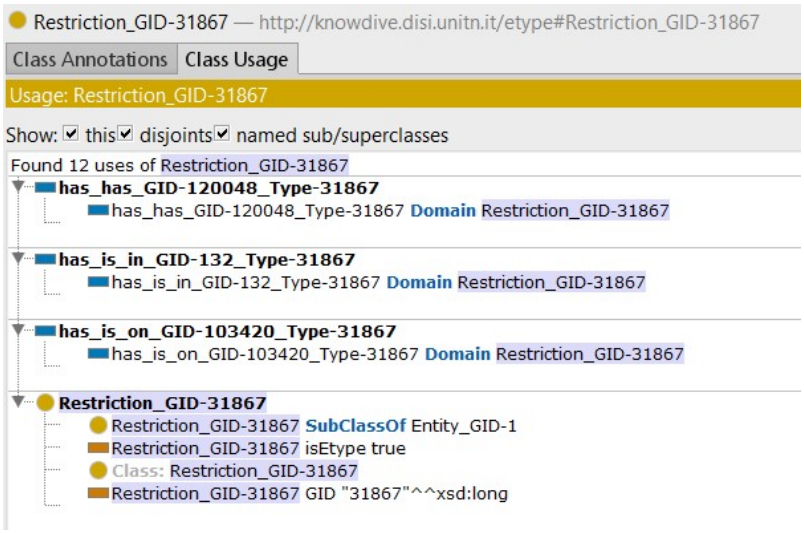


Figure 12: Restricion in Protégé

- **RestrictionType** is an EnumerationType of class that allow us to specify the type of restriction under certain categories. Its data properties are Essential Business, Non-essential business, Travel, Institutional, Gathering, and Lockdown.

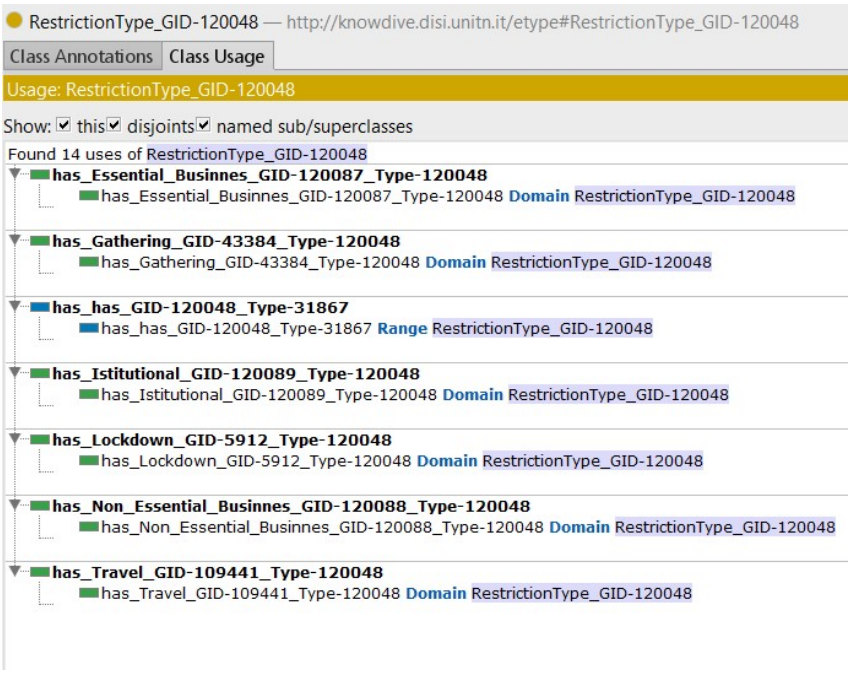


Figure 13: ResctricionType in Protégé

1.4.1.3 Data properties

- **Cases information**
  - num_cases
  - num_deaths
  - location
  - date
- **Cases projection**
  - lower_ bound_est
  - upper_bound_est
- **RSA cases**
  - RSA_id
  - RSA_name
- **School cases**
  - school_id
  - school_name
  - total_cases_schools
  - num_classes_cases
- **Covid status**
  - num_active
  - num_new_pos
  - num_recovered
- **Date**
  - year
  - day
  - month
- **Location**
  - location: **LocationType**
  - name

  - resctriction: **Restriction**

- **LocationType**
  - Nation
  - Region
  - City
  - Institution

- **Restriciton**
  - restriction: **RestrictionType**
  - closure_start: **Date**
  - closure_end: **Date**
  - location: **Location**

- **RestrictionType**
  - Essential business
  - Non-essential business
  - Travel
  - Institutional
  - Gathering
  - Lockdown

	Present	Added by us	Total
Classes	3	7	10
DataProperties	6	17	23
ObjectProperties	0	10	10

KOS CONCEPT	KOS CONCEPT ID	COMMENT
School_Cases	120042	representation of covid-19 cases in school
RSA_Cases	120043	representation of covid-19 cases in residence healthcare assistant
Cases_Information	120044	representation of covid-19 information in general
CovidStatus	120045	representation of the actual covid-19 situation
CasesProjection	120046	representation of the projection of cases of covid-19 in the future
LocationType	120047	specification of the location investigate
RestrictionType	120048	specification of the type of restriction
Location	132	a point or extent in space
Date	103420	assign a date to
Restriction	31867	a principle that limits the extent of something

Table 10: Classes present in the ontology

CONCEPT NAME	CONCEPT GID	CLASS	CLASS ID
has_day	81001	Date	103420
has_year	80976	Date	103420
has_elevation	28270	Location	132
has_latitude	46263	Location	132
has_longitude	46270	Location	132
has_period	80647	Restriction	31867
has_Essential_Business	120087	RestrictionType	120048
has_Gathering	43384	RestrictionType	120048
has_Istitutional	120089	RestrictionType	120048
has_Lockdown	5912	RestrictionType	120048
has_Non_Essential_Business	120088	RestrictionType	120048
has_Travel	109441	RestrictionType	120048
has_Region	46452	Locationtype	120047
has_Nation	44214	Locationtype	120047
has_Istitution	120090	Locationtype	120047
has_City	45988	Locationtype	120047
has_num_active	120091	Covidstatus	120045
has_num_new_pos	120092	Covidstatus	120045
has_num_recovered	120093	Covidstatus	120045
has_lower_bound_pos	120094	Casesprojection	120046
has_upper_bound_pos	120095	Casesprojection	120046
has_RSA_ID	120102	Rsa_cases	120043
has_RSA_name	120103	Rsa_cases	120043
has_school_id	120098	School_cases	120042
has_school_name	120099	School_cases	120042
has_total_cases_schools	120100	School_cases	120042
has_num_classes_cases	120101	School_cases	120042
has_num_deaths	120097	CasesInformation	120044
has_num_cases	120096	CasesInformation	120044
has_num_deaths	120097	CasesInformation	120044

Table 11: Data properties

CONCEPT NAME	DOMAIN	DOMAIN GID	RANGE	RANGE ID
has_has	CasesInformation	120044	Covidstatus	120045
has_has	CasesInformation	120044	Casesprojection	120046
has_has	Location	132	Locationtype	120047
has_has	Restriction	31867	RestrictionType	120048
has_is_in	CasesInformation	120044	Location	132
has_is_in	Restriction	31867	Location	132
has_is_on	CasesInformation	120044	Date	103420
has_is_on	Restriction	31867	Date	103420
has_is_RSACases	CasesInformation	120044	Rsa_cases	120043
has_is_SchoolCases	CasesInformation	120044	School_cases	120042

Table 12: Object properties

1.4.1.4 Variance respect to the EER Model

Once the ontology has been built, this section report the differences, and so the variance, respect the EER model defined in the previous phase. This a way to define the quality of the outcomes

for the current phase as well as the alignment of the overall project development process.
By checking the variance respect to the EER we can check the quality of our ontology in this phase. During the generation of the ontology it was necessary to modify the links between classes (Object properties): that is because the KOS logic has been followed. The Data Property names were deleted in the class Location since LocationType was added as a class that give us information about the location type. Attributes like *Nation*, *Region*, *Institution*, and *City* were inserted in a way such that if the nation "Italy" is added, it automatically includes all the regions inside that nation.

1.4.2 Data level

As in the previous phase the data level section here, reports the description of the new version of the datasets, after formatting operations.

1.4.2.1 Formal Modeling datasets management

In this section are reported the operations and the tools adopted to format the dataset collected, in order to align them to the ontology definitions generated at schema level. The filtered datasets are now to be aligned to the schematic entities and attributes. In order to carry out the proper alignment we had to transform certain attributes in the dataset based on the model. Apart from that, we also had to remove some columns as those attributes were no longer a part of the model.

Output from the informal modeling is served as the input to this phase. The datasets after preliminary filtration now has to be aligned to the schematic model. There were some common and specific transformations carried out in the input. In the schema we have several entities such as *Case Information, Covid Status, Cases projection, School Cases, RSA Cases, Location, LocationType, Date, Restriction and RestrictionType*. While most of these entities have a similar type and attributes as that of our datasets, entities such as *Date, Restriction and RestrictionType* needed to be transformed for better alignment. The transformation and final filtration operations were done in Python with the help of Pandas package.

Date is one of the common attribute that is present in all the datasets we used. In this model, Date is considered as a common entity with *year, month and day* as it's attributes whereas in the datasets it is residing as a Date object type. In order to align this, we extracted the day, month and year from the Date object and added it as additional attributes in the dataset. This operation was performed in all the datasets as it was a common attribute.

The Summary dataset we use holds the information about the restrictions across the world. This dataset had seperate attributes to keep track of all possible restrictions along with corresponding

start and end dates. Based on the schematic model, we had to shrink down the data corresponding to each of those under the attributes: restrictionType, closure start date and closure end date. In effect, instead of having separate columns for all kinds of restrictions and their corresponding periods, we reduced the dimensions of the dataset to align it with the model.

After all the possible transformations in the datasets, we then dropped the irrelevant attributes from each. Now we have a properly aligned datasets as the output of the formal modeling phase.

1.4.2.2 Datasets metadata documentation

In this section eventually new metadata information are added in order to describe the evolution of the datasets. We created the metadata documentation for the Datasets used in this project. The metadata description follows DCAT standards which is a formal vocabulary used to define the data catalogs. Here we create the formal machine understandable metadata description. It not only increases the visibility, but also allows execution of complex queries. The standardized metadata description is based on the filtered/cleaned datasets.

The documentation includes information about dataset level description and attribute level description such as the data resources, publishers, domain specifications, attribute names, number of entries and so on. The DCAT dataset description is saved in JSON format.

1.4.2.3 Variance respect Informal Modeling datasets

This section aims to define the variance between the data elements (datasets and attributes within them) produced in this phase, and the initial datasets collected in the previous phase. This a way to define the quality of the outcomes for the current phase as well as the alignment of the overall project development process.

In the previous phase, we partially aligned the collected datasets by performing basic data cleansing. In this part, we already had the EER model as a reference to perfectly align the attributes in our cleansed datasets. We transformed and again filtered out several certain attributes in order to achieve the appropriate alignment. Some new attributes were added to the datasets while some were found unwanted and has been removed. But the additions made didn't impact much of the size of our input datasets. In fact, the size was again reduced to a considerable amount making the data manipulation much faster.

1.4.3 Formal Modeling Evaluation

The last section of the Formal Modeling phase report the evaluation of the outcomes obtained in this phase, through specif evaluation metrics. Before going into the actual evaluation a premise is

needed. The COVID-19 is a new phenomena for the research and the data analysis. In order to obtain a better evaluation it would have been necessary to base the work done on a preexisting ontology on the same topic. Many ontologies on the matter treat the argument from different points of view, focusing mainly on the number of deaths, and number of infected people, number of people in intensive care while our work treats this elements in a very different way.
By confronting our project ontology with the one proposed in *"An Ontology for Collection and Analysis of Covid-19 Data"* [3], we decided to analyze a subgroup of this ontology since its large dimensions and the topic variation ([**?**] analyze also the patient diagnoses, clinical state, etcetc). The subgrutp analyzed comprehend the following entities:

- Statistics
- Place
- Exposure to Covid-19
- Status

CLASS OF THE REFERENCE ONTOLOGY (α)[3]	24
CLASS OF OUR ONTOLOGY (β)	10
COMMON CLASS (B)	6

Table 13: Number of classes of our and the ontology took in consideration [3]

Coverage	25,00%
Flexibility	16,60%
Extensiveness	11,76%
Sparsity	53,00%

Table 14: Evaluation results

As it is possible to see in the table 14, there are some consideration to be made. By the value obtained for the *coverage* it has to be said that the domain targeted by the knowledge graph is mostly unexplored: this is because the reference ontology is larger than the one obtained by us. As far as it is concerned for the *flexibility*, low values indicate that the reference schema has been extended by a small amount, and also for the *extensiveness* values we can say that the contribution of the created knowledge graph is limited. For the *sparsity*, we know that high values mean that there is an important difference between the considered type of elements defined in and the ones defined in , while low values indicates a good match. Since the results obtained is of 53,00%, the knowledge graph obtained sits in the middle.

1.5 Data integration

This section is dedicated to the Data Integration phase description.

1.5.1 Data integration operations and tool

This section is dedicated to the description of the usage of the data integration tool that allows to map the datasets generated and well formatted in the previous phases, with the final ontology generated. The last datasets adaptation performed using the tool, as well as the mapping operation are here detailed.

The well formatted datasets are ready to be mapped into the ontology created by the Knowledge Engineers. In order to carry out the integration process, we use Karmalinker tool. The integration is pretty much easy once you understand the relationships between each entities and their associated attributes. The Karmalinker V2.4 provide the interface to load, map the ontology and datasets, and publish the models.

The data integration part starts by importing the ontology into Karma workspace. Once it is imported successfully, we can create models for each datasets. The formatted datasets from the Formal modeling phase is served as second input in the data integration tool. The mapping is done based on the entity relationships specified in the ontology. The detailed description of this procedure is given below.

Data Integration of COVID-19 Coronavirus data

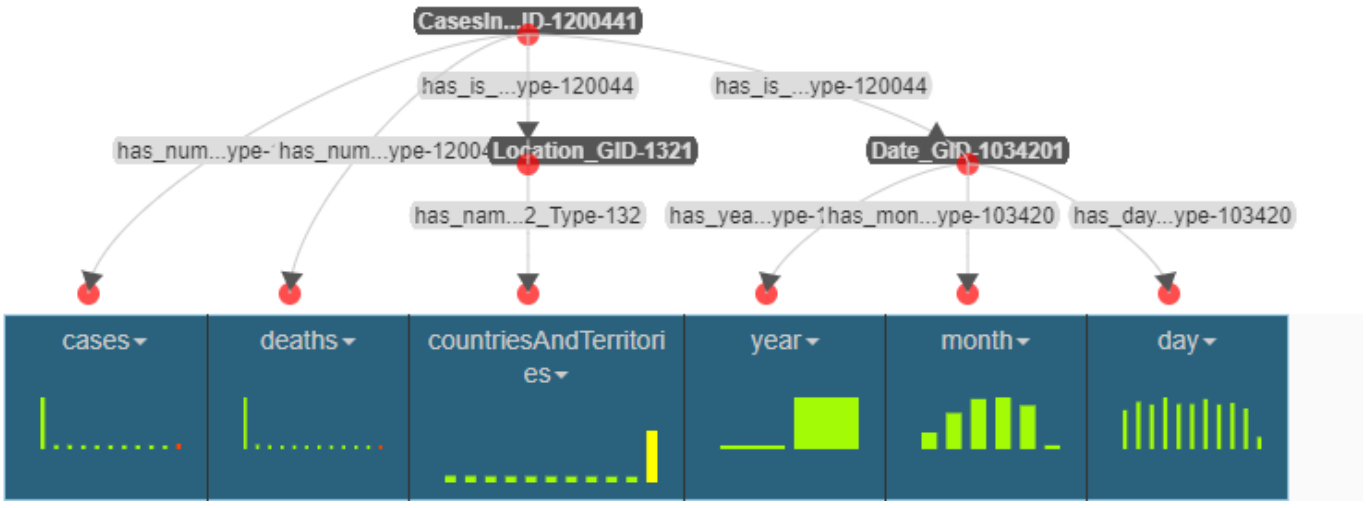


Figure 14: Mapping of Coronavirus dataset to Ontology

The formatted Covid dataset has daily records of Covid cases and deaths in every country. The dataset has columns: *cases, deaths, countriesAndTerritories, year, month, day.* Due to the memory limitations of this tool, the dataset has been partitioned into small chunks of about 10000 records and was mapped to the model. The file partitioning was done by a user defined batch function which is available in the repository. The mapping process was performed once for the first partition and model thus obtained was reused to complete the mapping for the rest of our dataset. The columns: cases and deaths were mapped to the Cases Information entity whereas the columns: year, month and day are mapped to attributes of type Date. The countriesAndTerritories column was mapped to LocationType class. The class Cases Information has outgoing links to Location and Date to indicate the ***'has_is_in' and 'has_is_on'*** relationship between those entities.

Data Integration of Reference hospitalization data of all locations

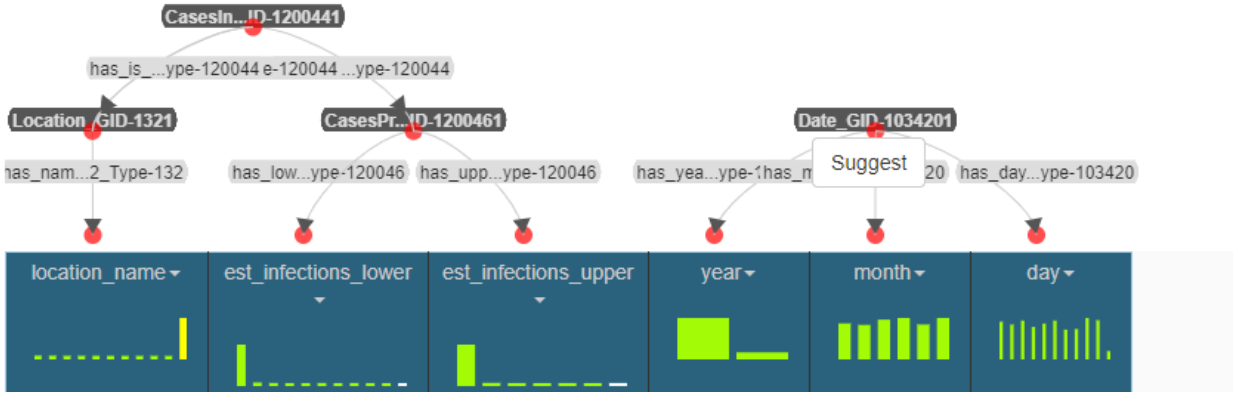


Figure 15: Mapping of Reference hospitalization all locations dataset to Ontology

The formatted Reference hospitalization dataset has information regarding the hospitalization and projections across the world. The dataset has columns: *location_name, est_infections_lower, est_infections_upper, year, month, day.* Since this dataset is huge, we partitioned this csv file into chunks, each of 10000 records. Just like the previous case, the model was created from the initial partition and was reused to map the rest of the records. The columns: est_infections_lower & est_infections_upper which provides the estimated infection rates, were mapped to the CasesProjection entity whereas the columns: year, month and day are mapped to attributes of enumerated type Date. The location_name column was mapped to Location class. The class CasesProjection have has_class relationship with Cases Information class. Hence we used an incoming link from Cases information to CasesProjection to indicate their relationship. Similarly, class Cases Information have outgoing links to Location and Date to indicate the ***'has_is_in' and 'has_is_on'***

relationship between those entities.

Data Integration of Summary stats data of all locations

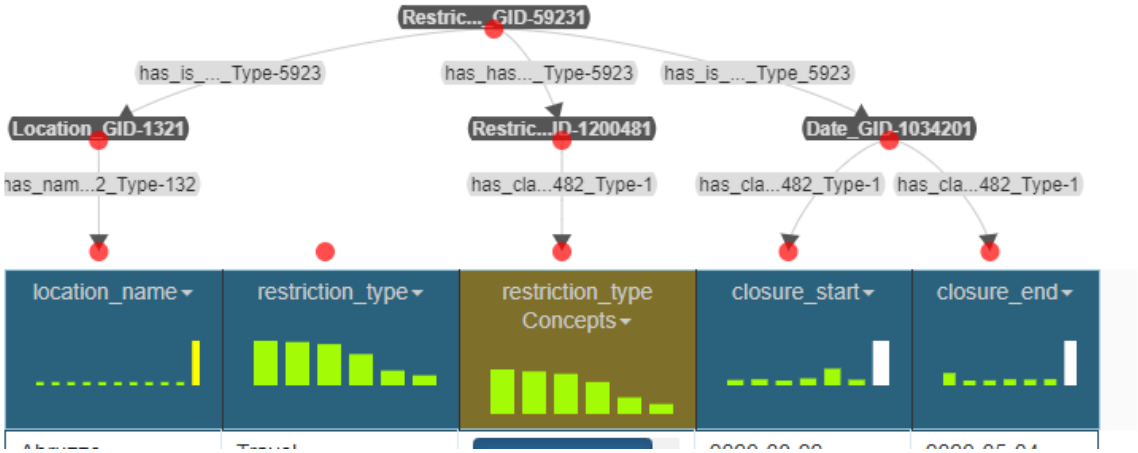


Figure 16: Mapping of Summary stats of all locations dataset to Ontology

The formatted Summary stats dataset has information regarding the social distancing norms in various countries and territories across the world. The dataset has columns: *location_name, restriction_type, closure_start, closure_end*. Since the dataset was within the size limits of Karmalinker, we didn't split the data into batches. The column restriction_type, gives the kind of restriction imposed in certain regions and the value can be from a pre-defined set. Thus we extracted the concepts associated with restriction_type. This restriction_type concept is mapped to class RestrictionType, which is has_class the common entity: Restriction. In order to indicate this relation, we specified an outgoing link from Restriction to RestrictionType. The location_name column was mapped to Location class and has an incoming link from Restriction to indicate the ***has_is_in*** relationship between them. The columns: closure_start, closure_end were mapped to the Date class and has an incoming link from Restriction to indicate their ***has_is_on*** relationship.

Data Integration of Trentino COVID-19 data

The formatted Trentino COVID-19 dataset has information regarding the Covid cases in Trentino region. The dataset has columns: *guariti, deceduti, pos_att, rsa, year, month, day*. Since the dataset was within the size limits of Karmalinker, we didn't split the data into batches. The column *guariti* indicates the number of recovered cases and was mapped to CovidStatus class. Similarly, the column pos_att, which indicates the number of positive cases was also mapped to CovidStatus whereas, *deceduti*, which indicates the number of deaths and was mapped to CasesInformation class and rsa was mapped to RSACases. The columns: year, month and day are mapped to attributes

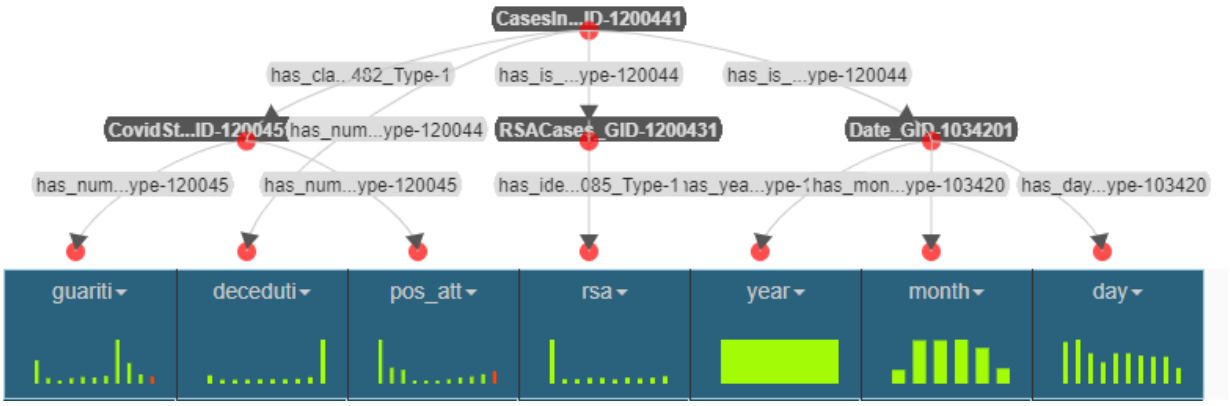


Figure 17: Mapping of Trentino COVID-19 dataset to Ontology

of enumerated type Date. The classes CovidStatus, RSACases and Date has ***has_class, has_is, has_is_in*** relationships with CasesInformation class. In order to indicate these, we added outgoing links from CasesInformation to all three classes.

1.5.2 Variance respect Formal Modeling datasets

The last section of the data integration phase aims to describe the variance, analyzing the differences, between the datasets integrated with the ontology, in the data integration platform which contain the KG, and the datasets collected in the previous phase. This analysis can highlight the results of the operations performed during the final phase of the data integration process.

The datasets from the formal modeling along with the ontology was the inputs in this phase of the project. The big datasets were split into batches for the data integration purposes as Karmalinker, the data integration tool we used had certain limitations.

1.5.3 Limits

The Karmalinker tool wan not able to handle big datasets (more than 10000 rows). The frequent session timeout and gateway errors has also been a blocker during this phase. The big datasets were partitioned into smaller chunks and then loaded in the Karmalinker workspace, to create the mappings and the EML and RDF files. But the crashing of this tool during the concept extraction and EML generations made the data integration part harder.

1.6 Conclusion and results

The project developed following the proposed methodology can be considered successful in its entirety since the resulting knowledge graph.
The KG produced is well populated and contains all the information needed in order to answer to the CQs initially defined for our personas. Moreover, the ontology generated can be easily embedded into a bigger ontology containing the different aspect that all the groups analyzed due to the presence of entities such as the Location, information about the restriction related to the transportation and travel, and in general health related content. In addition to that, the ontology can be further expanded with more specific information about the COVID-19 such as, for example, diagnoses and symptomatic information.

References

[1] Zachary Desson, Lisa Lambertz, Jan Willem Peters, Michelle Falkenbach, and Lukas Kauer, *Europe's covid-19 outliers: German, austrian and swiss policy responses during the early stages of the 2020 pandemic*, Health Policy and Technology **9** (2020), no. 4, 405 – 418, The COVID-19 pandemic: Global health policy and technology responses in the making.

[2] Trentino Digitale, *Covid-19 emergency health situation: Province of trentino.*

[3] B. Dutta and M. DeBellis, *Codo: Codo: an ontology for collection and analysis of covid-19 data.*, 2020, In Proc. of 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD), 2-4 November 2020 (accepted).

[4] European Centre for Disease Prevention and Control (ECDC), *Covid-19 coronavirus data.*

[5] Institute for Health Metrics and Evaluation (IHME), *Covid-19 mortality, infection, testing, hospital resource use, and social distancing projections.*

[6] Matthew Horridge, Simon Jupp, Georgina Moulton, Alan Rector, Robert Stevens, and Chris Wroe, *A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1. 2*, The university of Manchester **107** (2009).

[7] KnowDive, *Karmalinker.*

[8] _____, *Kos.*

[9] _____, *Sweb api.*

[10] Stanford University, *Protégé.*

[11] _____, *Webprotégé.*