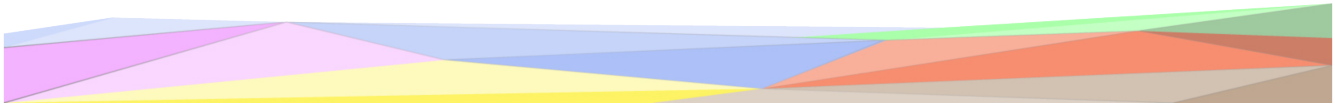
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2023 - Weather and Climate Change in Trentino

Document Data:	Reference Persons:
January 25, 2024	Hasan Aldhahi, Veronika Deketová



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 16, 2023	Hasan Aldhahi, Veronika Deketová	Document created
0.2	October 29, 2023	Hasan Aldhahi, Veronika Deketová	ER data model, Information Gathering
0.3	November 11, 2023	Hasan Aldhahi, Veronika Deketová	Language definition
0.4	December 4, 2023	Hasan Aldhahi, Veronika Deketová	Knowledge definition
0.5	January 5, 2023	Hasan Aldhahi, Veronika Deketová	Data definition
0.6	January 20, 2023	Hasan Aldhahi, Veronika Deketová	Evaluation and Metadata

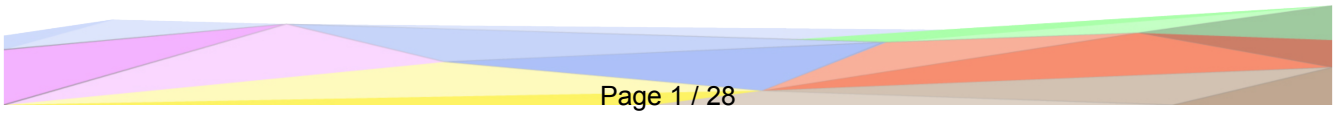
1 Introduction

Climate change and weather fluctuations are having significant impact on livability in many cities around the whole globe. Since industrial revolution, we observe a significant rise in the concentration of chemical substances which were not present in the atmosphere in such amounts before. The increase of air pollutant concentration can have a significant effect on human health, especially when taken in consideration more sensitive groups, such as elder people, children and those with respiratory or cardiovascular problems. Moreover, we can refer refer to monitoring of pollutant concentration as one of the manifestation of climate change, as together with temperate rise and higher occurrence of forest fires, it is one of the consequences of climate change. However, there are way more connections between air pollution and climate change – climate change can exacerbate air quality issues, and air quality can, in turn, contribute to climate change due to the emissions of air pollutants.
Therefore, we present here a KGE project, which allows to citizens or other tools to monitor situation regarding the air quality data in the connection with weather forecast and other climate situation indicators, to give a complex perception of a situation in given area. Users can then make conclusions based on the data together with consideration of their own personal and health interests.
The current document aims to provide a detailed report of the project developed following the iTelos methodology. The report is structured, to describe:

- Section 2: Definition of the project's purpose and its domain of interest in connection to the topic of climate change.
- Section 3: High level description of the project development, based on the two main sub process considered by iTelos, producer and consumer, respectively.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities, as well as considered from the point of view of the producer first, and the consumer later.
- Section 9: The description of the evaluation criteria and metrics applied to the project final outcome.
- Section 10: The description of the metadata produced for all (and all kind of) the resources handled and generated by the iTelos process, while executing the project.
- Section 11: Conclusions and open issues summary.

2 Purpose and Domain of Interest (DoI)

The goal of this project is to develop a Knowledge Graph (KG) that offers thorough data regarding the weather and air quality, as climate change factor, in Trentino. The final KG is a useful tool for anyone who is looking for details on different air quality monitoring locations and pollution



or weather forecasts throughout the Trentino geographical area. As well as the health impacts related to poor air quality, and the appropriate procedures to be taken to mitigate the risks.

- **Scope and Temporal Domain**:

  The following temporal domain concentrate on the following data within:

  – *Real-time measurement*: Air quality and weather measures taken in real-time starting at the time of the search, providing current data.

  – *24-hour forecast*: 24-hour air quality forecasts to assist users in making outdoor plans while taking future air quality conditions and weather situation into account.

  – *Historical data*: Lastly, historical data on air quality, referring to the time from 2013 to 2024, sheds the light on Trentino's mid-range trends in air quality, and the historical data from 2013 reflects on the weather trends and temperatures in the past.

- **User Demand**: The goal of the project is to provide a service that enables users across Trentino to access these different temporal data about different air quality observation sites and forecasts. By taking into account air quality and its historical trends as well as the forecasts with connection to the weather data, this service provided by the Knowledge graph will impact the users decision making process effectively in regards to outdoor activities, travel, and general well-being.

- **Resource Structure**: Historical air pollution data gathered from 2013 form the basis of our resource, together with weather data collected since 2013. In order to create fresh historical records of air quality measurements gathered over time, this historical archive serves as a foundation. Users are always provided access to the most recent data on air quality because of real-time data feeds that continually fill the database. This historical information allows for a thorough understanding of the air quality in Trentino, making it an asset for both locals and visitors. It also allows for real-time observations and forecasts.

3 Project Development

Project Development section describes how the project purpos is going to be satisfied. The section provides information from the perspective of Data Production and Data Composition.

3.1 Data Production

For our project of Weather and Climate Change in Trentino area, producer has to produce datasets to meet defined purpose and meet sufficient quality – this means creation of a new datasets if they do not exist, or do not meet required quality. In our case it is gathering information about weather and climate measures in Trentino area, obtained via weather and pollution stations with appropriate sensors. Measuring stations are represented via their name, location – GPS coordinates longitude and latitude (geometric features), and elevation. Quantity measurement should be represented with a quantity which is being measured (for example temperature,

wind speed, humidity, SO_2 concentration), its unit, value, and measurement timestamp. Similarly as quantity measurement, stations are characterized as well by predicted values for given location, which means what are expected values of given quantity for given station for next 24 hours.

3.2 Data Composition

Consumer providers a list of high quality data resources which are already available to be used for the project. In the current iteration of the project, non of the already created datasets already created by iTelos methodology or as a knowledge graph project are satisfying exact needs for our project.

4 Purpose Formalization

The main goal of purpose formalization is to describe and collect use cases for our resource, then extract the pertinent data to create a knowledge graph with an emphasis on air quality data. By following these steps, the resource's goal will be precisely defined, its target audience will be identified, and the issues it should cover will be listed.

4.1 Scenarios and Personas

- **Scenario 1:** The number of people with lowered immunity or suffering from breathing difficulties, or chronic health conditions is non-negligible even in the Trentino area.
- **Scenario 2:** With growing awareness about the possible impacts of air pollution on human health increases as well number of people who are worried about spending their free time outside during unpleasant air situations.
- **Scenario 3:** Trentino has stunning nature and many possibilities for spending time outdoors. For successful trip planning, it is useful to have information about short-term weather prediction.
- **Scenario 4:** Trentino area is very lucrative for organizing many sports events. Weather history together with air quality history can help while deciding about the ideal location and date for a sports race event.
- **Scenario 5:** Weather stability and air pollution can have a significant effect on produced agricultural products. Having long-term statistics for specific Trentino area can bring valuable information while making decisions in the agricultural field.

  For scenarios presented above, we describe now users with specific features:

- **Personas 1:** Matteo, a 35-year-old living in Borgo Valsugana, is interested in outdoor activities and sports, but his health is very sensitive to outer conditions, suffering from asthma and having lower immunity. He is therefore interested in the current concentrations

of air pollutants to make sure it is safe for him to perform sports outside.

- **Personas 2:** Lenka, a 24-year-old professional triathlete living close to Bolzano, is spending many hours outside every day training for her upcoming races. She values her health and well-being and is particularly concerned about air quality, as poor air quality can negatively affect her health and even athletic performance. Moreover, she wants to plan her training sessions based on the current weather forecast, as she cannot place her hard training during too hot days or during wet conditions.

- **Personas 3:** Sarah is a 29-year-old digital nomad from Norway planning her next destination in Trentino for a few months of remote work and exploration. She prioritizes her health, comfort, and productivity, making the selection of a city with stable temperature and air quality essential. She is therefore searching for a place with a stable average temperatures during summer and low values of O_3 and NO_2.

- **Personas 4:** Paolo is a 19-year-old student who just moved to Trento for his studies and his family is going to visit him. He wants to plan a trip to show his family the most out of the city. He wants to know what will be the weather forecast for the following day, so he can plan the activities for his family accordingly.

- **Personas 5:** Mario and Suse are family farmers from Rome family farmers aiming to set up a new ecological farm in the Trentino area. They are searching for the right place to do so. They are concerned about the weather stability in a given location together with the concentration of pollutants.

From the description of scenarios and personas, we define Competency Questions (CQ).

- **CQ 1:** Provide Matteo the most current air pollution data for a given day closest to Borgo Valsugana, so he can make a decision about the time he wants to spend outside given his health condition. The data are supposed to be from a station which is closes to Matteo's home. Return a list of all pollutants together with their concentration in $\mu g/m^3$.
- **CQ 2:** Give Lenka air pollution prediction for next day close to Bolzano, so she can adjust her training plan accordingly and potentially move some of her training sessions indoors. The data are suppose to be from a station which is closes to Lenka's home.
- **CQ 3:** Sarah is interested in the long-term history of temperature, rainy days, and wind speed in cities in the Trentino area.
- **CQ 4**: Provide Sarah with a list of the average values of O_3 and NO_2 for stations available through the history of the dataset, ordered by the smallest average value.
- **CQ 5**: Paolo is interested in the precipitation and temperatures status for current day for Trento.

- **CQ 6**: Provide Mario and Suse with data about areas in Trentino and history of temperatures, number of rainy days and levels of air pollution and which area has the best values. This means provide Mario and Suse a list of all historical reports for all weather and pollution stations in Trentino area.

These categorized use cases emphasize the relevance of air quality and climate stability information for both individuals and enterprises, allowing them to make informed decisions for various purposes.

Scenario	Personas	CQs	Entities	Properties	Focus	Popularity
1	Matteo	CQ1	CurrentPollutionReport, Station, Location	SO_2_value (float), NO_2_value (float), O3_value (float), PM25_value (float), PM10_value (float), datetime: str, station_name (str), longitude (float), longitude (float)	Core	Common
2	Lenka	CQ2	PollutionPrediction, Station	pred_SO2_value (float), pred_NO2_value (float), pred_O3_value (float), pred_PM10_value (float), pred_PM25_value (float), datetime: str, station_name (str)	Core	Common
3	Sarah	CQ3	HistoryWeatherReport	temp_max (float), temp_avg (float), wind_avg (float), rain_precipitation (float), datetime: str, station_name (str)	Core	Common
4	Sarah	CQ3	HistoryPollutionReport, Station	O3_value (float), No2_value (float), datetime: str, station_name (str), latitude (float), longitude (float)	Core	Common

Table 1: Persona's Competency Questions and the Expected Final Knowledge Graph's Output

From the CQs, referring to Personas and Scenarios, we extract Entities with properties. These entities are categorized as either Common, Core, or Contextual entities by considering Focus classification and Popularity classification. Details are captured in the Table 1 and

Scenario	Personas	CQs	Entities	Properties	Focus	Popularity
5	Paolo	CQ4	CurrentWeather, Station	temp_min (float), temp_max (float), precipitation (float), datetime: str, station_name (str)	Core	Common
6	Mario and Suse	CQ6	HistoryWeatherReport, HistoryPollutionReport, Station, Location	temp_avg (float), wind_avg (float), rain_precipitation (float), PM25_value (float), PM10_value (float), datetime: str, station_name (str), station_location_id (int), latitude (float), longitude (float)	Core	Common

Table 2: Persona's Competency Questions and the Expected Final Knowledge Graph's Output

Table 2.

The Entity-Relationship Model has in one level station, location, the prediction entities, and the report. The model was designed in such way to keep the current and the historical reports in the same sub level concerning the data from pollution and the weather. Also, prediction entities are separated as they are disjoint from each other in terms of the data that they produce. Each entity has its own data properties. The sub-classes inherits the properties of the child class. The relationship is in the shown in the Figure 1.

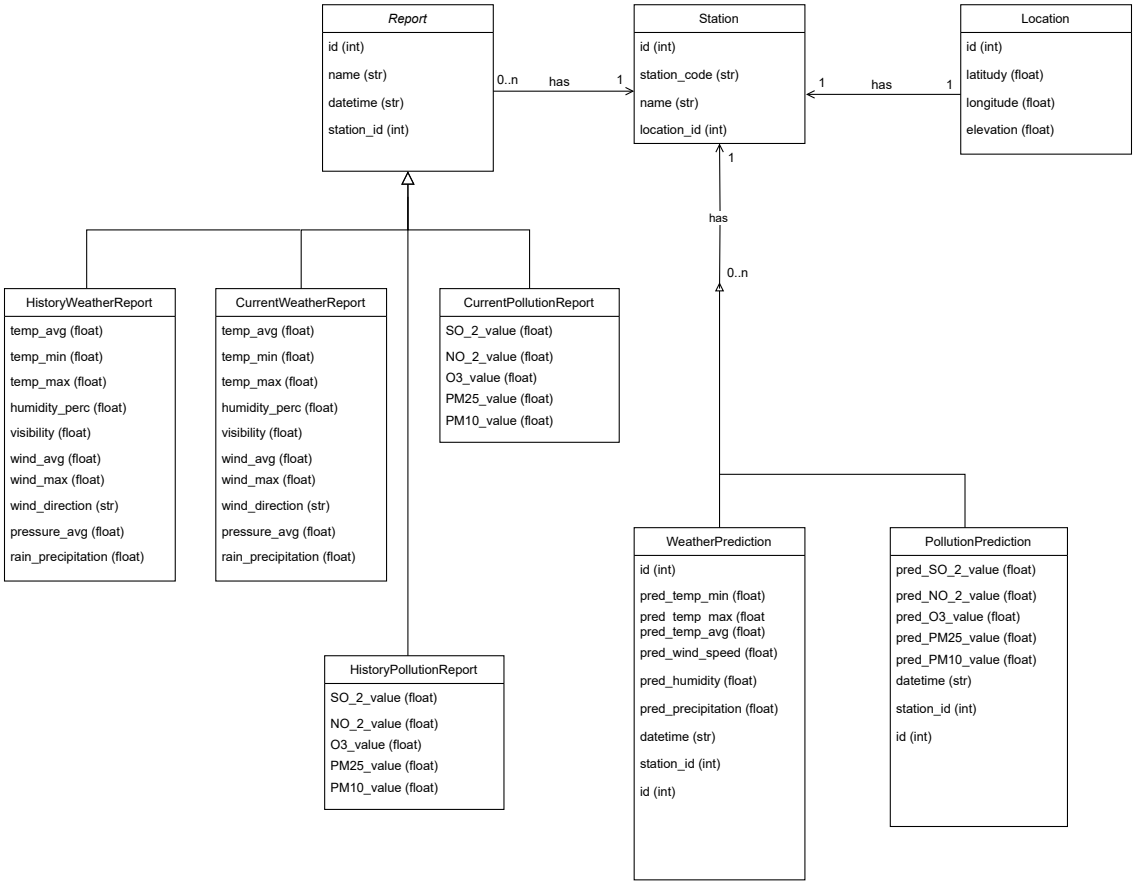


Figure 1: ER Diagram

5 Information Gathering

In this section, we describe sources identified by a consumer as formal data and knowledge resources. Furthermore, the producer aims to collect semi-formal data for integration into the project to achieve its objectives.

5.1 Data and Knowledge Source

5.1.1 Informal Data and Knowledge Source from Producer

We identified two informal resources. First of them represents Historical Air Quality data from the European Environmental Agency (see Table 3), Current Air Quality data from the European Environmental Agency (see Table 4). The reason why the data from the European Environmental Agency have been classified as informal is that they are missing exact information regarding name of the measuring station or the name of the pollutant, which had to be extracted via another European Environmental Agency's dataset. On the other hand, this dataset consist of daily data since 2013 until present for all different values of possible pollutants: PM2.5, PM10, NO_2, SO_2 and O_3.

The second informal data source is for predicted air pollution data, provided by the Copernicus Atmosphere Monitoring Service (CAMS) in cooperation with the European Centre for Medium-range Weather Forecasts (ECMWF)[1]. This service is globally available for any non-commercial usage after requesting and creating account. As a part of everyday service, CAMS provide up to four-day forecast of the EU-WHO regulated pollutants, together with pollen and aerosol tracers in a form of ensemble models[2]. This data source is described in the following Table 5.

As a semi-formal dataset was identified KGE22 – Trentino Weather, which includes the information about historical, current and predicted weather data in Trentino area (see Table 6). This project was created as a part of the university course as this report, so it is following equivalent methodology. However, some parts of the project needed to be transformed in order to fir our project's purpose.

[1] https://atmosphere.copernicus.eu/
[2] https://atmosphere.copernicus.eu/european-air-quality-forecast-plots
[3] https://www.eea.europa.eu/en/datahub/datahubitem-view/778ef9f5-6293-4846-badd-56a29c70880d
[4] https://link.springer.com/chapter/10.1007/978-3-030-71187-0_104
[5] https://www.intechopen.com/chapters/86794
[6] https://sdi.eea.europa.eu/catalogue/datahub/api/records/778ef9f5-6293-4846-badd-56a29c70880d/formatters/xsl-view?output=pdf&language=eng&approved=true
[7] https://www.eea.europa.eu/data-and-maps/explore-interactive-maps/up-to-date-air-quality-data
[8] https://link.springer.com/chapter/10.1007/978-3-030-71187-0_104
[9] https://www.intechopen.com/chapters/86794
[10] https://confluence.ecmwf.int/display/CKB/FTP+access+to+CAMS+global+data
[11] https://confluence.ecmwf.int/display/CKB/FTP+access+to+CAMS+global+data
[12] https://github.com/jclock98/Weather_Trentino/tree/main/Datasets/Data%20Integration
[13] https://jclock98.github.io/Weather_Trentino/Documentation/KGE_2022_Trentino_Weather_Lazzerini_Clocchiatti.pdf
[14] https://jclock98.github.io/Weather_Trentino/Documentation/KGE_2022_Trentino_Weather_Lazzerini_Clocchiatti.pdf

Resource name	Air Quality Download Service
Domain	Europe – focused: Trentino-Alto Adige (Italy)
Keyword	Air Quality
Language	English
Provider	European Environmental Agency
Data URL	Air Quality Download Service[3]
Data format	.parquet files
Data description	Historical daily measurements going from 2013 until presence having an overview about all of measured air pollutant: PM2.5, PM10, NO_2, SO_2 and O_3
Knowledge URL	N/A possible knowledge: paper 1[4] paper 2[5]
Knowledge description	EEA Metadata Factsheet [6] - high level description of the historical data

Table 3: Air Quality Download service

Resource name	Up-to-date air quality data
Domain	Europe – focused: Trentino-Alto Adige (Italy)
Keyword	Air Quality
Language	English
Provider	European Environmental Agency
Data URL	Up-to-date air quality data[7]
Data format	.csv files
Data description	Up-to-date measured values of air pollutant: PM2.5, PM10, NO_2, SO_2 and O_3 with hourly refresh rate
Knowledge URL	N/A possible knowledge: paper1[8] paper 2[9]
Knowledge description	N/A

Table 4: Up-to-date air quality data

Resource name	CAMS global data – Air pollution prediction
Domain	Europe – focused: Trentino-Alto Adige (Italy)
Keyword	Air Pollution Forecast
Language	English
Provider	Copernicus Atmosphere Monitoring Service (CAMS) & European Centre for Medium-range Weather Forecasts
Data URL	Modeled prediction data are available via FTP server[10]
Data format	.grib files
Data description	Ensemble model for air pollution prediction: PM2.5, PM10, NO_2, SO_2 and O_3
Knowledge URL	N/A
Knowledge description	Data are described in the service manual[11]

Table 5: CAMS global data – Air pollution prediction

Resource name	KGE22 – Trentino Weather
Domain	Trentino (Italy)
Keyword	Weather Forecast, Weather History, Weather Current
Language	English
Provider	University of Trento
Data URL	Final KG construction[12]
Data format	.ttl files
Data description	Knowledge Graph giving data about weather in Trentino area.
Knowledge URL	Knowledge Graph project report[13]
Knowledge description	Knowledge Graph project report[14]

Table 6: Trentino weather data

5.1.2 Formal Data and Knowledge Source from Consumer

As mentioned in the section 3.2, consumer at this phase of the project did not identify any sources as formal ones which could be directly used for this project.

5.2 Resource Collection, Processing and Scraping

5.2.1 Informal Resource Collection, Processing and Scraping from Producer

Data from the European Environmental Agency has been fetched in two different ways. Historical data has been downloaded via data download service as one large archive containing files in parquet format. One parquet file represented the whole history for one pollution station for one specific pollutant. In order to collect the data, we needed to first convert them from the parquet format into csv format, identify which location ids are withing our desired region, and also identify which pollutant id represents which of the pollutants. This mapping was done via daily data, which were enhanced with data description and exact locations. However, not all pollution stations have a rich historical data for all mentioned pollutants – for some of the stations, this form of data is more sparse then what the current pollution data provides.

Current pollution data was possible to fetch via API requests with parameters of pollutant type and datetime for which we want to obtain the data – search via close history (up to 10 days) is also possible. This form of obtaining the data allowed for direct data export in the csv format without need for further processing. As historical data did not contain information about the name of the stations and its location, we had to do the mapping via *station_id* with current data, to identify which pollution measurement stations' data we can provide within Trentino area, together with the mapping of the *pollutant_id* to it's name – detailed description is in the Table 8.

Pollution prediction models provided by Copernicus Atmosphere Monitoring Service needed to be downloaded from their respective FPT server after granting access. All the models are stored in the grib file format, which allows querying the data for specific longitude and latitude. In case of 3D models, it also allow to obtain the data from specific altitude. For reading the data files, we used ecCodes package developed by ECMWF. This package provides an application programming interface and a set of tools for decoding and encoding various formats used for

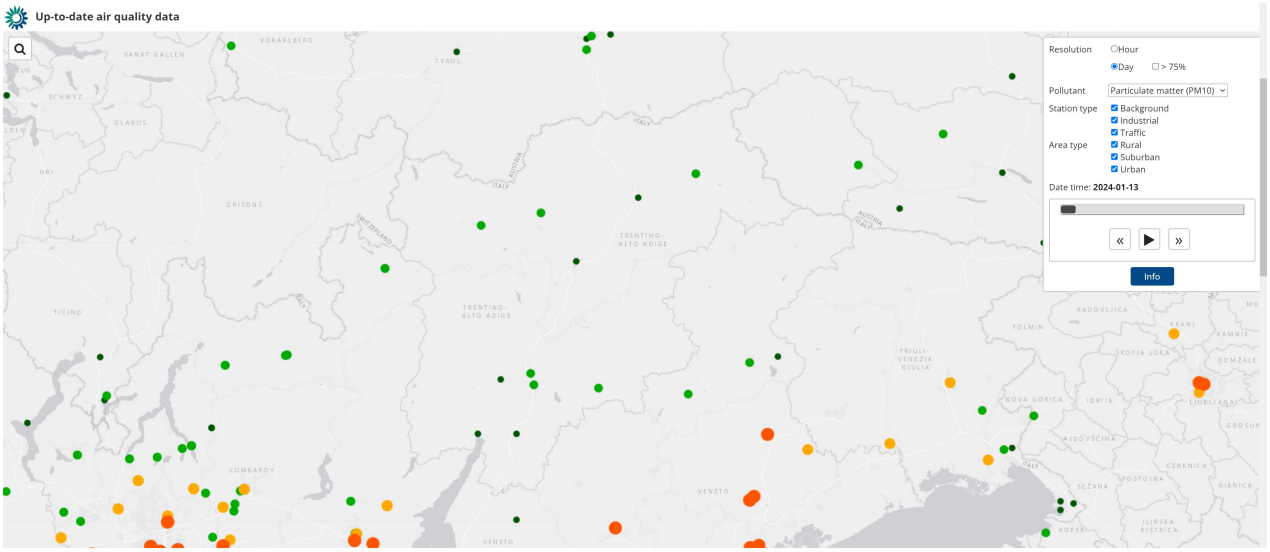


Figure 2: Example of daily pollution data from EEA

Files	Domain	Description	Source
history_pollution_stations.json, *station_id_pollutant_id*.csv	pollution stations	json file contains list of available pollution stations and their location, csv files contain list of all historical records for given station and given pollutant	EEA – Download Service
daily_pollution_stations.json, *pollutant_id_datetime*.csv	pollution stations	json file contains list of available current pollution stations and their location, csv files contain list of all values measured for a given pollutant for a given day on available stations	EEA – Up-to-Date data

Table 7: Fetched data from European Environmental Agency

weather data, including grep files[15]. When using ecCodes package, we were able to query data points closest to our pollution stations based on their longitude and latitude. For data having different elevation levels, we picked those ones in the lowest altitude, as further information about conversion to real altitude values was missing. As models are prepared with timestamp at midnight given day, we downloaded data models with the step 36 hours in order to obtain the prediction data for the next day (for noon). This step was identified to meet our requirements and purpose defined in the section 4.

Files	Domain	Description	Source
pollutant_id_datetime_step.csv	pollution stations	csv file containing prediction data for location points from the list of daily pollution stations, step represents number of hours in future	CAMS

Table 8: CAMS prediction data after transformation from grep file

Usage of KGE22 – Weather Trentino project way less manual processing. We were also able to use previous tools prepared in order to fetch historical, current, and predicted weather data, so it could be easily incorporated as a part of our data sources. The original weather data included as a source Meteotrentino[16], which is a platform focused on aspects of meteorology, snow science, and glaciology in Trentino province. This side also allows for historical data download, keeping the data consistency since 2013.

5.3 Integrate Scraped data with Scraped Schemas Using Karma

To make our Schema, we have used protege tool to define our entities, create the hierarchy, and connect them via Object properties. We also had to define the data properties in the schema for each entity and link each property to the corresponding entity as well as its data type as shown in the Figure 3.

Karma tool was used for integrating our scraped data sets with knowledge schema that was built from the protege tool. The tool allow us to edit transform some of the columns of the dataset and create linkage between multiple data sets and our schema as seen in the Figure 4.

6 Language Definition

A specification for the language concepts that will be used to describe the data, included in the final Knowledge graph, needs to be established. In order to fulfill the purpose of the knowledge graph the data consumers and providers must have their language and concepts selected. Universal language core (UKC) was used for already defined concepts to use already existing structure. Concepts missing in UKC were newly created along with their specific identifier. Language Definition sub activities:

[15] https://confluence.ecmwf.int/display/ECC/What+is+ecCodes
[16] https://www.meteotrentino.it/index.html#!/home

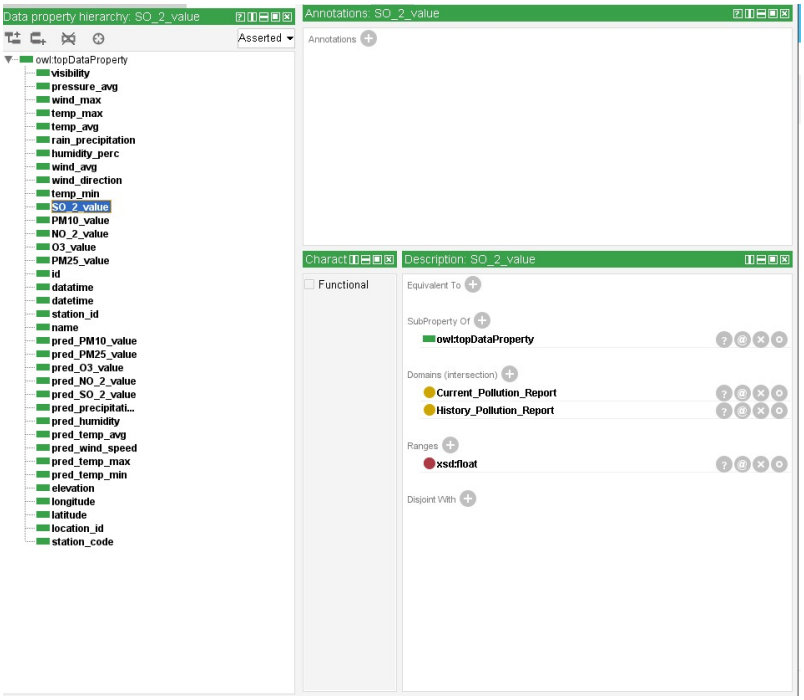
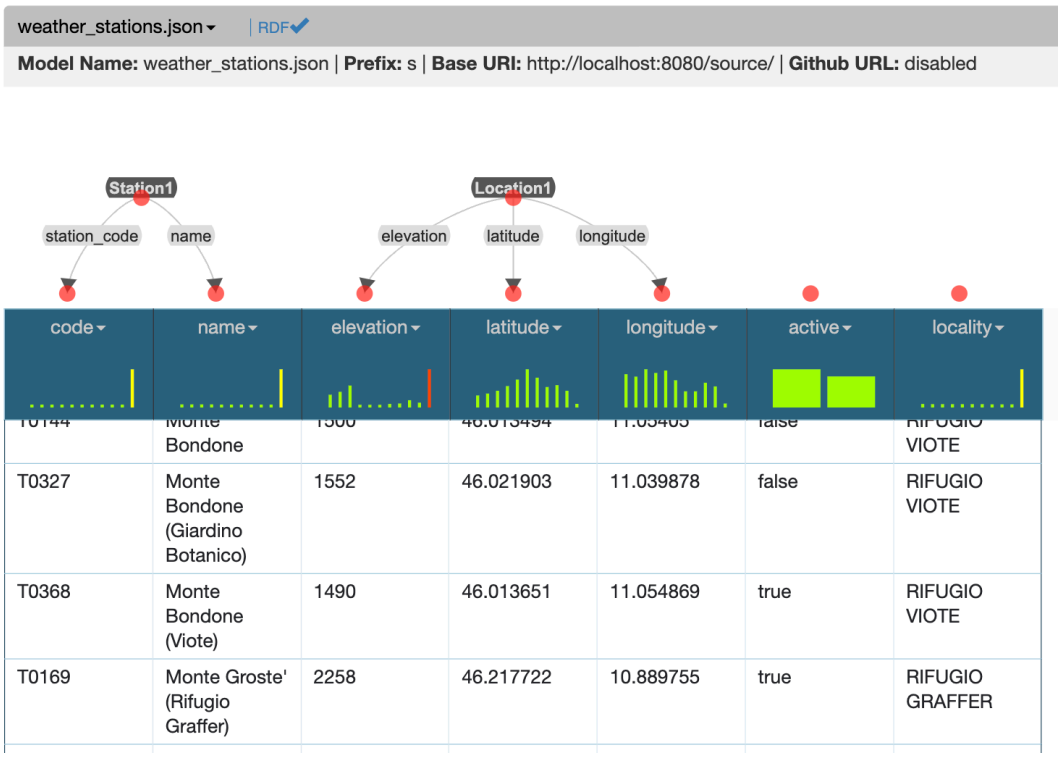


Figure 3: Protege 5 for modeling the schema of our Ontologie



T0144	Monte Bondone	1500	46.013494	11.05405	false	RIFUGIO VIOTE
T0327	Monte Bondone (Giardino Botanico)	1552	46.021903	11.039878	false	RIFUGIO VIOTE
T0368	Monte Bondone (Viote)	1490	46.013651	11.054869	true	RIFUGIO VIOTE
T0169	Monte Groste' (Rifugio Graffer)	2258	46.217722	10.889755	true	RIFUGIO GRAFFER

Figure 4: Karma tool for Merging the dataset with the Schema to formulate the Knowledge graph

- **Producer activities**: The concepts that need to be used to fulfill our purpose were selected and defined accordingly. In the Universal Knowledge Core (UKC) alignment, concepts were filtered and defined or uniquely defined into the knowledge core. From the producer activities perspective, fields connected with current and history reports were formalized, together with descriptions for stations and locations.
  - Knowledge layer:
    * Concept identification
      All the entities and data properties shown in the ER model (Figure 1) for the project were defined and conceptualized.
    * UKC alignment
      If the selected concept has not been defined within the formal definition of the UKC, a unique UKCIdentifier was given to the label. For a given label, an apposite definition was created, giving all potential users specific context and necessary domain knowledge for understanding modeled data (see in Figure 5).
  - Data layer:
    * Dataset filtering
      Based on the scenarios defined in the section 4.1 and together with the assistance of the data consumer, we filtered out the data that were not relevant to our purposes or were not defined by any of the concepts formalized. Those fields were mostly connected with special further undescribed identificators or unclear datetime specification, such as: DATETIME_BEGIN, DATETIME_END, SAMPLINGPOINT_LOCALID, DataCapture, FkObservationLog.

Concept Labels	Description
forecast_GID-36179	a prediction about how something (as the weather) will develop
Report_GID-12001	Its measrurment that could be in the present or past or future
CurrentWeatherReport_GID-12002	A real-time report providing up-to-the-minute data on current weather conditions, including average temperature, temperature range, humidity, visibility, wind speed and direction, atmospheric pressure, and precipitation.
temp_avg _GID-12003	The average temperature, typically in degrees Celsius, at the time of the report.
temp_min_GID-12004	The minimum temperature, typically in degrees Celsius, recorded at the time of the report.
temp_max _GID-12006	The maximum temperature, typically in degrees Celsius, recorded at the time of the report.
humidity_perc _GID-12007	The percentage of relative humidity in the air at the time of the report.
visibility_GID-26112	capability of providing a clear unobstructed view
wind_avg_GID-12008	The average wind speed, often in meters per second, at the time of the report.
wind_max_GID-12009	The maximum wind speed, often in meters per second, observed at the time of the report.
wind_direction_GID-12010	The cardinal or compass direction from which the wind is blowing.
pressure_avg_GID-12011	The average atmospheric pressure, often in millibars or hectopascals, at the time of the report.
rain_precipitation_GID-12012	The amount of precipitation, typically in millimeters, that has fallen or is expected to fall at the time of the report.
CurrentPollutionReport_GID-12013	A real-time report that provides current data on pollution levels, including measurements of sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), fine particulate matter (PM2.5), and coarse particulate matter (PM10) in the environment.
SO2_value_GID-12014	The current level of sulfur dioxide (SO2) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³).
NO2_value_GID-12015	The current level of nitrogen dioxide (NO2) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³).
O3_value_GID-12016	The current level of ozone (O3) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³).
PM25_value_GID-12017	The current concentration of fine particulate matter (PM2.5) in the air, which are the following elements Si, Ca, Al, Na, Mg, K, and Fe, typically measured in micrograms per cubic meter (µg/m³).
PM10_value_GID-12018	The current concentration of coarse particulate matter (PM10) in the air, which are the following elements Fe, K, Ca, Na and Al, typically measured in micrograms per cubic meter (µg/m³).
Station_GID-12019	A physical location equipped with monitoring equipment to collect and report environmental data
Location_GID-12020	Geographical coordinates represented by latitude and longitude, along with elevation data, providing the precise geographic position associated with monitoring stations and facilitating accurate location identification.
id_GID-12021	A unique identifier for a geographical location.
latitude_GID-46263	the angular distance between an imaginary line around a heavenly body parallel to its equator and the equator itself
longitude_GID-46270	the angular distance between a point on any meridian and the prime meridian at Greenwich
elevation_GID-49932	a raised or elevated geological formation
HistoryWeatherReport_GID-12022	A report that contains historical data on past weather conditions, including average and extreme temperatures, humidity, visibility, wind characteristics, atmospheric pressure, and precipitation recorded at a specific point in time in the past.
station_id_GID-12023	An identifier, typically in integer format, used to associate a report or data entry with a specific monitoring station within the system, indicating the source or origin of the data.
datetime_GID-12024	A textual representation of the date and time when a report or data entry was generated or recorded, typically in a specific format to convey temporal information.
HistoryPollutionReport_GID-12025	A report that contains historical data on past pollution levels, including measurements of various pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), fine particulate matter (PM2.5), and coarse particulate matter (PM10) recorded at a specific point in time in the past.

Figure 5: Formalized concepts for the Trentino Climate Change project – producer part

- **Consumer activities**: As the biggest part was covered by producer activities, on the consumer part there were only activities regarding future predictions for both weather and air pollution, which needed to be properly formalized during the language definition phase.
  - Knowledge layer:
    * Concept identification
      From the knowledge layer for concept identification, also the defined ER model was used (Figure 1). For potentially missing terms, we used the Home Weather ontology[17].
    * UKC alignment
      For missing concepts in the UKC definition system, we created and formally defined new one having specific label with id and necessary description (6).
  - Data layer:
    * Dataset filtering
      For the part of the data regarding predictions, there were no redundant fields that would not match any of our formalized concepts.

WeatherPrediction_GID-12026	A report that provides forecasts and predictions of future weather conditions, including parameters such as minimum and maximum temperatures, average temperature, wind speed, humidity, and precipitation, typically associated with a specific date and time.
pred_temp_min _GID-12027	The forecasted minimum temperature, typically measured in degrees Celsius, for a specific date and time
pred_temp_max_GID-12028	The forecasted maximum temperature, typically measured in degrees Celsius, for a specific date and time.
pred_wind_speed_GID-12030	The forecasted wind speed, often measured in meters per second, for a specific date and time.
pred_humidity_GID-12031	The forecasted humidity level, typically represented as a percentage, for a specific date and time.
pred_precipitation_GID-12032	The forecasted amount of precipitation, typically measured in millimeters, expected for a specific date and time.
PollutionPrediction_GID-12033	A report that offers forecasts and predictions of future pollution levels, including measurements of pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), fine particulate matter (PM2.5), and coarse particulate matter (PM10), typically associated with a specific date and time.
pred_SO2_value_GID-12034	The forecasted level of sulfur dioxide (SO2) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³), for a specific date and time.
pred_NO2_value_GID-12035	The forecasted level of nitrogen dioxide (NO2) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³), for a specific date and time.
pred_O3_value_GID-12036	The forecasted level of ozone (O3) in the environment, typically measured in parts per million (PPM) or micrograms per cubic meter (µg/m³), for a specific date and time.
pred_PM25_value_GID-12037	The forecasted concentration of fine particulate matter (PM2.5) in the air, which are the following elements Si, Ca, Al, Na, Mg, K, and Fe, typically measured in micrograms per cubic meter (µg/m³), for a specific date and time.
pred_PM10_value_GID-12038	The forecasted concentration of coarse particulate matter (PM10) in the air, which are the following elements Fe, K, Ca, Na and Al, typically measured in micrograms per cubic meter (µg/m³), for a specific date and time.

Figure 6: Formalized concepts for the Trentino Climate Change project – consumer part

Activities described above from the producer and consumer perspective helped to create and formalize specific concepts for further knowledge graph processing and improve general understanding for used elements.

7 Knowledge Definition

This section is dedicated to the description of the formal modeling phase. We create a Teleontology by extending our existing Teleology with the reference ontologies following the language alignment. kTelos 3-level approach allows us to do knowledge modeling. This section will be dedicated to describing these three approaches: Top-Down, Bottom-Up, and Middle-out. In all these approaches reuse the concepts from existing Knowledge resources. The main goal in iTelos process is re-usability and share-ability. Then Language alignment is used for semantic interoperability enhancement. From the producer and the consumer task point of view, they

[17] https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/WeatherOntology.owl

Figure 7: Ontology

are based on the re-usability of existing formal (standard) ontologies. The consumer has the modeling of singular unique ontology. in contrast, the producer has the whole ontology for each dataset to model.

In this section, we performed both tasks, making sure our Language was aligned, and each entity type had its Language definition. We also pinpoint the main Ontologies that we have and the other sub-entities in that ontology:

7.1 EER Model

We have added hierarchical entities in which we can describe our entities by reusing already built-in schemas. We wanted to aggregate our schema by adding more description and differentiating the spatio-temporal entities from each other. We have also modified our relationship by adding one object property ”part_of” to serve as a connection between entities such as Location connected to the station, similarly with all the reports. and the final output would like this as shown in the figure 8

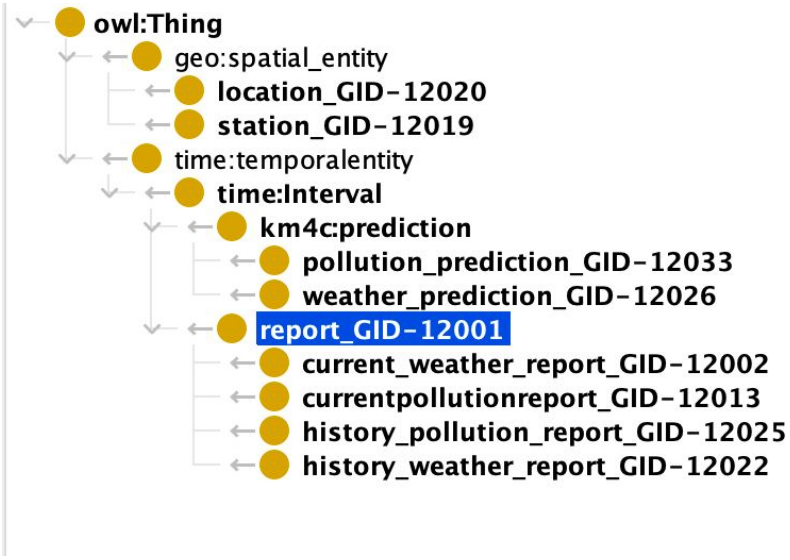


Figure 8: EER model

7.2 Top-Down: Ontology

We have decided to introduce hierarchy in our ontology in which we know there is no perfect ontology that can represent our structure. We have separated the etypes based on the temporal and spatial entities. Time elements and location elements are important in our schema. Re-usability is an important aspect of defining our ontology. Therefore, we have used already built schema, and the following choices were made from different schemas

- **geo: SpatialEntity** Weather Station and Locality are both places with coordinates that describe their position in space. W3C Geospatial Ontology defines spatial things as “Anything with spatial extent, i.e. size, shape, or position.”. We decided to add this entity to extend our teleontology. ISA: Locality, Weather Station.
- **time:temporalEntity** Weather and pollution is measured by time. this entity serves its purpose to give the overall weather and pollution phenomena measurements.
- **km4c: Prediction** The km4city knowledge model facilitates the illustration of smart cities, taking advantage of its interconnection, and storage of data from many different data sources. We have used Prediction as we have this etype of pollution prediction that predicts the pollutants, such as O_3 and NO_2, in the air.
- **time: interval** In our teleology, we have time intervals that describe the weather and pollution reports, as well as historical weather and pollution reports and their structure, so this definition serves the ontology description the best imported from the time schema.

7.3 Bottom-up : Teleology

For the teleology, we revised our 6 competency questions and revised the Etypes and properties relevant to be modeled in Protege:

- What is the current situation of concentration of NO_2, SO_2, O_3, PM10 and PM2.5 in $\mu g/m^3$ in Borgo Valsugana?
- What is the prediction for the next day’s noon concentration of pollutants: NO_2, SO_2, O_3, PM10 and PM2.5 for LS1 Laives station?
- Query the list of historical weather values: average temperature, average pressure, max wind speed, precipitation and humidity for Trentino area.
- What are the cities with the lowest average values of O_3 and NO_2 in Trentino area?
- What is the weather like outside at this moment: precipitation and temperature?
- Query the history of the average daily temperature, pressure, maximum wind speed, precipitation and humidity, and also provide the history of pollutants: NO_2, SO_2, O_3, PM10 and PM2.5 for Trentino area.

7.4 Middle-out : Teleontology

Combining the previous two approaches into one to yield the final schema that will be used to generate our final knowledge graph. In this step, we make sure to align all the hierarchical entities and ensure they are aligned with the language alignment via knowledge annotation, checking every notion with the (UKC-based) language annotations spreadsheet. For the teleology, we revised our 6 competency questions and the Etypes and properties relevant to be modeled. This step was done using Protege. We can see in figure **??** that the number of axioms has increased as well as the entities, and the number of entities that will be used will be dependent on mapping them with real data in the Data Definition section.

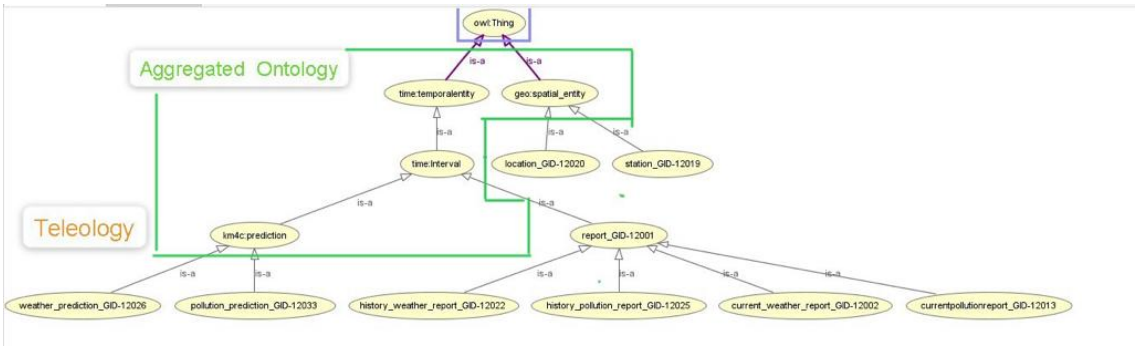


Figure 9: Teleontology

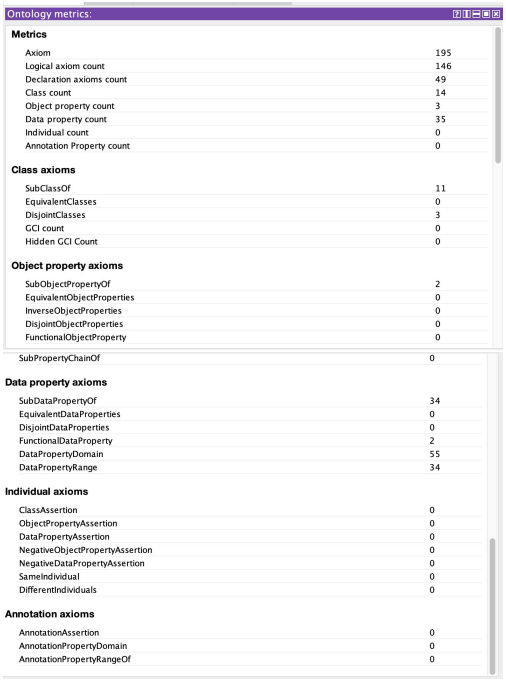


Figure 10: Characteristics of the teleology

8 Data Definition

In this section, we will continue our stages to build the final knowledge graph. After building the schema structure, producing teleontology in the last section. We will continue with the language definition and also we will clean and transform our real-world data in such a way that it will be easier to identify and map correspondingly with our produced teleontology. The main objective of this stage is to obtain for each dataset a knowledge graph. Obtaining a single structure out of the data and the schema as well as merging all the knowledge graphs into one file was the main objective of the consumer. Dealing with heterogeneity in the data will also be another objective of this phase.

On the **Producer activities**, both the data layer from one dataset will be mapped to the values presented in the knowledge layer in our produced schema that has etypes and its corresponding

data and object properties.

- **Entity identification:**
  In order for the mapping to work, each dataset should have an identifier that reflects on each entity on our schema to be able to traverse the data row by row and able to identify the data properties and the connections that this entity has. Therefore; For each etype, a special URI (Unique Resource Identifier), of the type URN. The only exception to this is the entity type tourist, whose identifier is the name of the tourist. The structure for the naming is as follows ”nu_name”, starting at number one. For the historical weather entity, we have used the locality for its detection and also the station entity has a station code for it to identify in the datasets.

Concept	Identifier
location	n_location
daily pollution	n_daily_pollution
historical_pollution_report	n_historical_pollution
daily Weather	n_daily_weather

- **Data mapping:**
  In this section, we have completely managed to link most of the data that we have in the data layer to the teleontology schema via the Karma Linker tool. We also mapped most of the entities with each other through object properties. Regarding the entity matching problem, we have used the made-up columns for identifiers to map each entity to be identified. The output was distinct 7 (RDF Turtle) TTL files. These files will be kept separate to be used for other purposes.

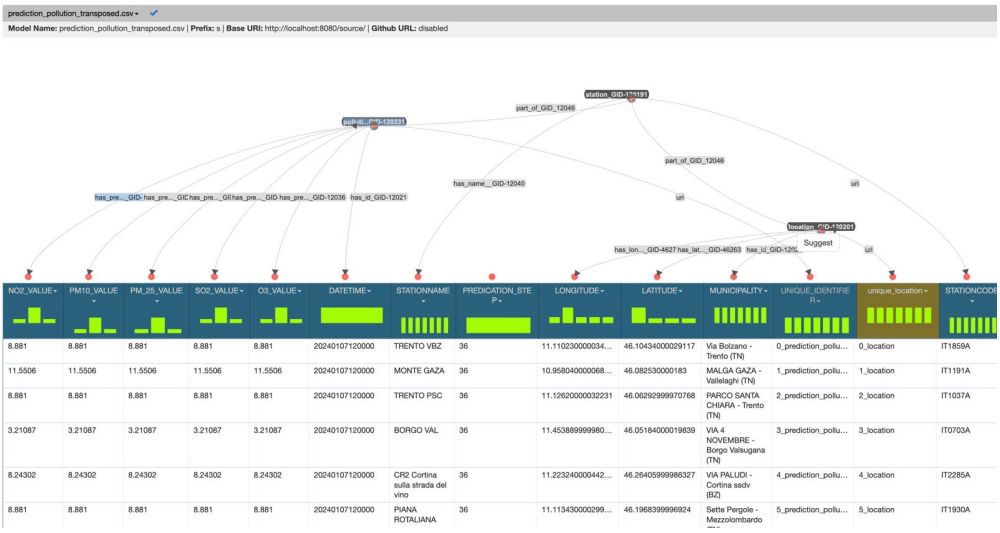
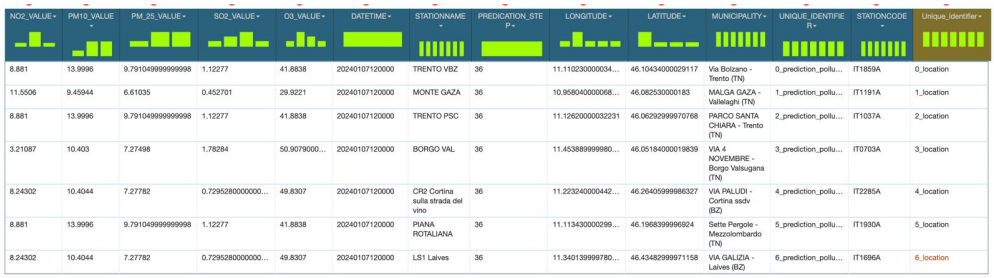


Figure 11: Example of data mapping

On the **consumer side**, the data definition phase aims at producing the final KG, suitable to satisfy the requirements extracted from the user purpose (Competency Questions). Our final

KG will be both highly reusable and purpose-specific, due to the language alignment with the UKC and the generation of the teleontology adopted to structure its information, respectively.

- **Entity matching:**
  - **Schema layer:** An update for the schema was made to match the real-world data in terms of mapping each entity together. Some of the schemas already ”has” links from the nature of hierarchical structure. We have added the object property ”part_of” to connect the disjoint entities such as current_pollution_report and Station entity. By this, we managed to link entities together and be able to access each other’s properties.
  - **Data layer:** We had to filter and merge some of the datasets to be able to produce one data set that can reflect upon one entity mostly and the other entities to be linked. Previously, we had 4 datasets reflecting the value of one individual pollutant. Merging all these data to have one dataset that gives us all the columns of the pollutant was an important step. Also, when it comes to Data mapping. It enabled to linking of multiple entities together. It made the process more neat in terms of defining one etype per dataset as shown in the figure of the merged dataset.



NO2_VALUE	PM10_VALUE	PM_25_VALUE	SO2_VALUE	O3_VALUE	DATETIME	STATIONNAME	PREDICATION_STEP	LONGITUDE	LATITUDE	MUNICIPALITY	UNIQUE_IDENTIFIER	STATIONCODE	Unique_identifier
8.881	13.9996	9.7910499999999998	1.12277	41.8838	20240107120000	TRENTO VBZ	36	11.110230000034...	46.10434000029117	Via Bolzano - Trento (TN)	0_prediction_pollu...	IT1859A	0_location
11.5506	9.45944	6.61035	0.452701	29.9221	20240107120000	MONTE GAZA	36	10.958040000068...	46.082530000183	MALGA GAZA - Valsalagni (TN)	1_prediction_pollu...	IT1191A	1_location
8.881	13.9996	9.7910499999999998	1.12277	41.8838	20240107120000	TRENTO PSC	36	11.12620000032231	46.06292999970768	PARCO SANTA CHIARA - Trento (TN)	2_prediction_pollu...	IT1037A	2_location
3.21087	10.403	7.27498	1.78284	50.9079000...	20240107120000	BORGO VAL	36	11.453889999980...	46.05184000019839	VIA 4 NOVEMBRE - Borgo Valsugana (TN)	3_prediction_pollu...	IT0703A	3_location
8.24302	10.4044	7.27782	0.72952800000000...	49.8307	20240107120000	CR2 Cortina sulla strada del vino	36	11.223240000442...	46.26405999986327	VIA PALUDI - Cortina ssdv (BZ)	4_prediction_pollu...	IT2285A	4_location
8.881	13.9996	9.7910499999999998	1.12277	41.8838	20240107120000	PIANA ROTALIANA	36	11.113430000299...	46.1968399996924	Sette Pergole - Mezzolombardo (TN)	5_prediction_pollu...	IT1930A	5_location
8.24302	10.4044	7.27782	0.72952800000000...	49.8307	20240107120000	LS1 Laives	36	11.340139999780...	46.43482999971158	VIA GALIZIA - Laives (BZ)	6_prediction_pollu...	IT1696A	6_location

Figure 12: Example of data matching

- **Entity identification:** Same as the one on the producer side, the identifier table was kept the same on the consumer side.
- **Entity mapping:** All the final 7 RDF-Turtle (TTL) files were merged into one single file that represents the final knowledge graph and is specific to its usage.

9 Evaluation

In this section, we describe and conclude the final evaluation of the knowledge graph created. Usually, this is done as a part of the iterative process. In our solution, it is as a last step to evaluate the work done.

9.1 Knowledge Graph information statistics

As defined in a material for the KGE course[18], we are going to use coverage in order to explore how much of portion of knowledge (etypes and properties) is covered within created knowledge

[18] https://unitn-knowledge-graph-engineering.github.io/KGE2023-website/material/slides/L12.pdf

graph. Lower values for coverage can mean that reference schema is not appropriate for given domain, or that the targeted domain is not adequately explored. The overview of obtained statistics for knowledge graph is following:

Variable	Concept	Value	Reference
Cov_E	Number of etypes extracted from the CQs	7	The entities excluding reference ontology entities excluding Report and weather prediction Fig. 8
T_E	Number of etypes of the Teleontology	14	Number of classes that can be counted in Fig.9
Cov_P	Number of properties extracted from the CQs	32	We haven't used property wind direction, pred max wind, temp min, wind max, pressure_avg, for the object properties we haven't used "include" property
T_P	Number of properties of the Teleontology	38	Number of object (3) and data properties (35) in Fig.10
RO_E	Number of etypes extracted from the ROs	607	It can be seen here from the time schema (582) and from km4c ontology (23) as well as geo (2). Rersource:1[19] and Resource:2 [20]and Resource:3 [21]
RO_P	Number of properties extracted from the ROs	0	We haven't used any properties from the the reference ontology

9.2 Knowledge layer evaluation

9.2.1 Primary objective

Primary objective for evaluation of the knowledge layer is described as how the Teleontology covers the entities and properties obtained from CQs. For set of CQs (CQ), etype coverage COV_E of Teleontology T is define as following

$$COV_E(CQ_E) = \frac{|CQ_E \cap T_E|}{CQ_E} = \frac{7}{7} = 1,$$

where CQ_E is the number of etypes derived from the CQs and T_E is the number of etypes of the Teleontology. Property coverage COV_P is defined similarly as for etypes

$$COV_P(CQ_P) = \frac{|CQ_P \cap T_P|}{CQ_P} = \frac{32}{32} = 1,$$

where CQ_P is the number of properties obtained from the CQs and T_P is the number of properties from the Teleontology.

[19] https://lov.linkeddata.es/dataset/lov/vocabs/km4c
[20] https://lov.linkeddata.es/dataset/lov/vocabs/time
[21] https://lov.linkeddata.es/dataset/lov/vocabs/geo

9.2.2 Secondary objective

The secondary objective is supposed to describe how much the Teleontology covers etypes, and properties, extracted from the reference ontologies. Here we now talk about reusability of the final Knowledge graph. Definition of etype coverage COV_E given reference ontology (RO) of the Teleontology (T) is calculated

$$COV_E(RO_E) = \frac{|RO_E \cap T_E|}{RO_E} = \frac{7}{607} = 0.01.$$

Similarly defined also for properties

$$COV_P(RO_P) = \frac{|RO_P \cap T_P|}{RO_P} = \frac{38}{0} = undefined,$$

where RO_P is the number of properties extracted from the reference ontologies, T_P is the number of properties in Teleontology.

9.3 Data layer evaluation

Data layer evaluation is focused on the how are different parts of the knowledge graph connected and how dense the graph is. When evaluating the connectivity of the knowledge graph, we can focuse on connectivity from the entity perspective, as well as from the property perspective. Entity connectivity evaluates the grade of connection between different entities in the KG. On the other hand property connectivity is focused on the grade of connection between each single knowledge graph's entity and it's property values.

In our knowledge graph implementation, we have a close connection between the properties of the given graph's entity. However, other entities, such as different kind of pollution or weather report, are not that strongly connected.

9.4 Query execution

This section is devoted to the SPARQL query execution with the aim to answer competency questions defined in section 4.1.

- **CQ 1:** Provide Matteo the most current air pollution data for a given day in Borgo Valsugana, so he can make a decision about the time he wants to spend outside given his health condition. The data are supposed to be from a station which is closes to Matteo's home. Return a list of all pollutants together with their concentration in $\mu g/m^3$.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

SELECT ?no2value ?so2value ?so3value ?pm10value ?pm25value ?name ?Datetime
WHERE {
     ?daily_pollution rdf:type disi:currentpollutionreport_GID-12013 .

  ?daily_pollution disi:schema:has_NO2_value_GID-12015 ?no2value .
  ?daily_pollution disi:schema:has_SO2_value_GID-12014 ?so2value .
```

```
    ?daily_pollution disi:schema:has_O3_value_GID-12016 ?so3value .
    ?daily_pollution disi:schema:has_PM10_value_GID-12018 ?pm10value .
    ?daily_pollution disi:schema:has_PM25_value_GID-12017 ?pm25value .
    ?daily_pollution disi:schema:has_datetime_GID-12024 ?Datetime .

      ?daily_pollution <http://schema.org/part_of_GID_12046> ?station .
      ?location <http://schema.org/part_of_GID_12046> ?station .

      ?station disi:schema:has_name__GID-12040 ?name .

      FILTER(STRSTARTS(?name, "BORGO")).

  }
```

Listing 1: SPARQL query CQ1

	no2value	so2value	so3value	pm10value	pm25value	name	Datetime	long	lat
1	"16.76524"	"14.23249"	"11.87905"	"16.30048"	"9.755714"	"BORGO VAL"	"20231026000000"	"11.453889999980522"	"46.05184000019839"

Figure 13: Query output: Current pollution data for the station closest to Borgo Valsugana

- **CQ 2:** Give Lenka air pollution prediction for next day close to Bolzano, so she can adjust her training plan accordingly and potentially move some of her training sessions indoors. The data are suppose to be from a station which is closes to Lenka's home.

```
  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
  PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

  SELECT ?no2value ?so2value ?o3value ?PM10value ?name
  WHERE {
    ?prediction_pollution rdf:type disi:pollution_prediction_GID-12033 .

    ?prediction_pollution disi:schema:has_pred_NO2_value_GID-12035 ?no2value .
    ?prediction_pollution disi:schema:has_pred_SO2_value_GID-12034 ?so2value .
    ?prediction_pollution disi:schema:has_pred_O3_value_GID-12036 ?o3value .
    ?prediction_pollution disi:schema:has_pred_PM10_value_GID-12038 ?PM10value .

    ?prediction_pollution <http://schema.org/part_of_GID_12046> ?station .

    ?station disi:schema:has_name__GID-12040 ?name .

      FILTER(STRSTARTS(?name, "LS1")).
  }
```

Listing 2: SPARQL query CQ2

	no2value	so2value	o3value	PM10value	name
1	"8.24302"	"0.7295280000000001"	"49.8307"	"10.4044"	"LS1 Laives"

Figure 14: Query output: Pollution prediction for the closest station from Bolzano, next day noon prediction data

- **CQ 3:** Sarah is interested in the long-term history of temperature, rainy days, and wind speed in cities in the Trentino area.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

SELECT ?weather_history ?temp_avg ?pressure_avg ?wind_max ?rain_precipitation ?
    humidity_perc ?datetime
WHERE {
  ?weather_history rdf:type disi:history_weather_report_GID-12022 .

  ?weather_history disi:schema:has_temp_avg__GID-12041 ?temp_avg .
  ?weather_history disi:schema:has_pressure_avg_GID-12011 ?pressure_avg .
  ?weather_history disi:schema:has_wind_max_GID-12009 ?wind_max .
  ?weather_history disi:schema:has_rain_precipitation_GID-12012 ?rain_precipitation
      .
  ?weather_history disi:schema:has_humidity_perc__GID-12007 ?humidity_perc .
  ?weather_history disi:schema:has_datetime_GID-12024 ?datetime .

   # FILTER(CONTAINS(STR(?weather_history), "MORI"))
}
```

Listing 3: SPARQL query CQ3

	weather_history	temp_avg	pressure_avg	wind_max	rain_precipitation	humidity_perc	datetime
1	http://localhost:8080/source/ISERA	"11.0"	"1012.0"	"11.0"	"0"	"79.0"	"2018-01-02"
2	http://localhost:8080/source/ISERA	"11.0"	"1009.0"	"11.0"	"0"	"79.0"	"2018-01-02"
3	http://localhost:8080/source/ISERA	"11.0"	"1007.0"	"11.0"	"0"	"79.0"	"2018-01-02"
4	http://localhost:8080/source/ISERA	"11.0"	"1013.0"	"11.0"	"0"	"79.0"	"2018-01-02"
5	http://localhost:8080/source/ISERA	"11.0"	"1019.0"	"11.0"	"0"	"79.0"	"2018-01-02"
6	http://localhost:8080/source/ISERA	"11.0"	"1022.0"	"11.0"	"0"	"79.0"	"2018-01-02"
7	http://localhost:8080/source/ISERA	"11.0"	"1018.0"	"11.0"	"0"	"79.0"	"2018-01-02"
8	http://localhost:8080/source/ISERA	"11.0"	"1015.0"	"11.0"	"0"	"79.0"	"2018-01-02"
9	http://localhost:8080/source/ISERA	"11.0"	"1021.0"	"11.0"	"0"	"79.0"	"2018-01-02"
10	http://localhost:8080/source/ISERA	"11.0"	"1004.0"	"11.0"	"0"	"79.0"	"2018-01-02"
11	http://localhost:8080/source/ISERA	"11.0"	"1011.0"	"11.0"	"0"	"79.0"	"2018-01-02"
12	http://localhost:8080/source/ISERA	"11.0"	"1023.0"	"11.0"	"0"	"79.0"	"2018-01-02"

Figure 15: Query output: historical weather data for Trentino area

- **CQ 4**: Provide Sarah with a list of the average values of O_3 and NO_2 for stations available through the history of the dataset, ordered by the smallest average value.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

SELECT ?name (AVG(xsd:float(?o3value)) AS ?avgo3value)
WHERE {

  ?history_pollution rdf:type disi:history_pollution_report_GID-12025 .
  ?history_pollution disi:schema:has_O3_value_GID-12016 ?o3value .

  ?history_pollution <http://schema.org/part_of_GID_12046> ?station .
  ?station disi:schema:has_name__GID-12040 ?name .

  FILTER(xsd:float(?o3value) > 0).
}
```

```
        GROUP BY ?name
        ORDER BY ?avgo3value

    ### second query

    SELECT ?name (AVG(xsd:float(?no2value)) AS ?avgNo2Value)
    WHERE {
      ?history_pollution rdf:type disi:history_pollution_report_GID-12025 .

      ?history_pollution disi:schema:has_NO2_value_GID-12015 ?no2value .
      ?history_pollution <http://schema.org/part_of_GID_12046> ?station .

      ?station disi:schema:has_name__GID-12040 ?name .

      FILTER(xsd:float(?no2value) > 0).
    }
        GROUP BY ?name
        ORDER BY ?avgo3value
```

Listing 4: SPARQL query CQ4

	name	avgo3value
1	"LS1 Laives"	"34.984333"^^xsd:float
2	"MONTE GAZA"	"95.226295"^^xsd:float

Figure 16: Query output: comparison for lowest average pollutants, O_3

	name	avgNo2Value
1	"MONTE GAZA"	"5.1643105"^^xsd:float
2	"TRENTO PSC"	"45.9863"^^xsd:float

Figure 17: Query output: comparison for lowest average pollutants, NO_2

- **CQ 5**: Paolo is interested in the precipitation and temperatures status for current day for Trento.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

SELECT ?rain_precipitation ?temp_avg ?name ?Datetime ?long ?lat
WHERE {
     ?weather_current rdf:type disi:current_weather_report_GID-12002 .

  ?weather_current disi:schema:has_rain_precipitation_GID-12012 ?rain_precipitation .
  ?weather_current disi:schema:has_temp_avg__GID-12041 ?temp_avg .
  ?weather_current disi:schema:has_datetime_GID-12024 ?Datetime .

    ?weather_current <http://schema.org/part_of_GID_12046> ?station .
    ?location <http://schema.org/part_of_GID_12046> ?station .

    ?station disi:schema:has_name__GID-12040 ?name .
    ?location disi:schema:has_longitude_GID-46270 ?long .
```

```
    ?location disi:schema:has_latitude_GID-46263 ?lat .

   FILTER(STRSTARTS(?name, "Trento")).

}
```

Listing 5: SPARQL query CQ5

	rain_precipitation	temp_avg	name	Datetime	long	lat
1	"0.0"	"0.6"	"Trento (Laste)"	"2024-01-23T00:00:00+01"	"11.135703"	"46.071801"
2	"0.0"	"0.6"	"Trento (Laste)"	"2024-01-23T00:15:00+01"	"11.135703"	"46.071801"
3	"0.0"	"0.2"	"Trento (Laste)"	"2024-01-23T00:30:00+01"	"11.135703"	"46.071801"
4	"0.0"	"0.2"	"Trento (Laste)"	"2024-01-23T00:45:00+01"	"11.135703"	"46.071801"
5	"0.0"	"-0.4"	"Trento (Laste)"	"2024-01-23T01:00:00+01"	"11.135703"	"46.071801"
6	"0.0"	"-0.2"	"Trento (Laste)"	"2024-01-23T01:15:00+01"	"11.135703"	"46.071801"
7	"0.0"	"-0.4"	"Trento (Laste)"	"2024-01-23T01:30:00+01"	"11.135703"	"46.071801"
8	"0.0"	"-0.1"	"Trento (Laste)"	"2024-01-23T01:45:00+01"	"11.135703"	"46.071801"
9	"0.0"	"-0.4"	"Trento (Laste)"	"2024-01-23T02:00:00+01"	"11.135703"	"46.071801"

Figure 18: Query output: current weather for Trento

- **CQ 6**: Provide Mario and Suse with data about areas in Trentino and history of temperatures, number of rainy days and levels of air pollution and which area has the best values. This means provide Mario and Suse a list of all historical reports for all weather and pollution stations in Trentino area.

```

    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

    SELECT ?weather_history ?temp_avg ?pressure_avg ?wind_max ?rain_precipitation ?
        humidity_perc ?datetime
    WHERE {
      ?weather_history rdf:type disi:history_weather_report_GID-12022 .

      ?weather_history disi:schema:has_temp_avg__GID-12041 ?temp_avg .
      ?weather_history disi:schema:has_pressure_avg_GID-12011 ?pressure_avg .
      ?weather_history disi:schema:has_wind_max_GID-12009 ?wind_max .
      ?weather_history disi:schema:has_rain_precipitation_GID-12012 ?rain_precipitation
          .
      ?weather_history disi:schema:has_humidity_perc__GID-12007 ?humidity_perc .
```

```
      ?weather_history disi:schema:has_datetime_GID-12024 ?datetime .

    }

    ### for pollution

    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX disi: <http://knowdive.disi.unitn.it/etype#>

    SELECT ?name ?no2value ?o3value ?pm10value ?datetime
    WHERE {

        ?history_pollution rdf:type disi:history_pollution_report_GID-12025 .

      ?history_pollution disi:schema:has_NO2_value_GID-12015 ?no2value .
      ?history_pollution disi:schema:has_O3_value_GID-12016 ?o3value .
      ?history_pollution disi:schema:has_PM10_value_GID-12018 ?pm10value .
      ?history_pollution disi:schema:has_datetime_GID-12024 ?datetime .

      ?history_pollution <http://schema.org/part_of_GID_12046> ?station .

      ?station disi:schema:has_name__GID-12040 ?name .

        FILTER(xsd:float(?no2value) > 0).
    }
```

Listing 6: SPARQL query CQ6

	name	no2value	o3value	pm10value	datetime
1	"MONTE GAZA"	"1.6"	"96.0"		"2013-01-03 16:00:00"
2	"MONTE GAZA"	"1.4"	"98.0"		"2013-01-03 17:00:00"
3	"MONTE GAZA"	"1.2000000000000002"	"99.0"		"2013-01-03 18:00:00"
4	"MONTE GAZA"	"1.3"	"97.0"		"2013-01-03 19:00:00"
5	"MONTE GAZA"	"1.4"	"96.0"		"2013-01-03 20:00:00"
6	"MONTE GAZA"	"1.4"	"94.0"		"2013-01-03 21:00:00"
7	"MONTE GAZA"	"1.3"	"89.0"		"2013-01-03 22:00:00"
8	"MONTE GAZA"	"1.5"	"88.0"		"2013-01-03 23:00:00"
9	"MONTE GAZA"	"1.3"	"88.0"		"2013-01-04 00:00:00"

Figure 19: Query output: pollution history

10 Metadata Definition

- Language resources metadata description

DatLicense	Apache-2.0 licence
DatKeyword	airpollution, weather, trentino
DatPublisher	Veronika Deketova, Hasan Aldhahi
DatCreator	Veronika Deketova, Hasan Aldhahi
DatOwner	Veronika Deketova, Hasan Aldhahi
DatLanguage	English
DatSize	47
DatName	Language_Definition_KGE.xlsx
DatPublication	22/01/2024
DatDescription	Description of language used in the Trentino Weather and Climate Change Knowledge Graph
DatVersion	1
DatDomain	weather, air pollution
DatDFileFormat	xlsx

Table 9: Language resources metadata description

11 Open Issues

Regarding the data utilization and data incompleteness, especially regarding historical pollution data and weather prediction data, further improvements in this aspect could help fulfill the initial purpose with better information enrichment. The data we were able to collect, had still a limited amount of coverage, and further mapping of weather data would be welcomed. On the other hand, all the other data sources for both pollution and weather were challenging to tackle and we have presented utilization of many different sources. Especially for pollution prediction and history, data were coming from different sources and their formats required special software. Mostly for historical data, better coverage could help our personas to answer further or more complicated competency questions.

Furthermore, all of the pollution stations are not very scattered around the whole Trentino area, but they are all focused close to Trento. Further work could focus on obtaining more quality data sources, however, this is still limited by hardware and physical measuring stations available in the given area. As we believe that the European Environment Agency has a reliable overview of pollution stations available, this could be replaced by modeled data for more cities in the Trentino area. Similarly for historical data, but here we do not usually have models simulating the situation back in history.

Last, but not least, the data in the constructed knowledge graph are static, which means the snapshot of data was created at a specific time in history without further updates. This was one of the constrained we put on the created knowledge graph, but for ideal usage and for better utilization, data would have to be up to date and regularly refreshed.