Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2024 - Trentino Climate Change Causes

Document Data:	Reference Persons:
November 25, 2024	Lorenzo Fumi, Giacomo Tezza

Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 21, 2024	Lorenzo Fumi, Giacomo Tezza	Document created
0.2	October 29, 2024	Lorenzo Fumi, Giacomo Tezza	Purpose Definition
0.2	November 10, 2024	Lorenzo Fumi, Giacomo Tezza	Source Identification
0.2	November 11, 2024	Lorenzo Fumi, Giacomo Tezza	Dataset Collection
0.2	November 12, 2024	Lorenzo Fumi, Giacomo Tezza	Dataset Cleaning
0.2	November 13, 2024	Lorenzo Fumi, Giacomo Tezza	Information Gathering Complete

1 Introduction

This document presents the foundational elements, purpose, and methodology behind the creation of a Knowledge Graph (KG) focused on climate change in the autonomous Province of Trento. The primary goal of this Knowledge Graph is to map and understand climate-related data specific to this region, enabling deeper insights into both the environmental impact and the potential drivers behind observed climate changes. By integrating geospatial data from sources such as OpenStreetMap, historical meteorological data (including temperature, air quality, and other indicators over time), and demographic data (such as population density, energy and gas consumption, water usage, and waste production), this Knowledge Graph seeks to create a rich, interconnected representation of climate dynamics and their underlying causes within Trento.

The purpose of this documentation is to ensure not only the precise definition and execution of this project's objectives but also the reusability of the Knowledge Graph for future applications and analyses. Reusability is a core principle in the Knowledge Graph Engineering (KGE) process as defined by iTelos, which emphasizes that detailed and clear documentation serves as a vital resource for external audiences. This document thus aims to provide a transparent account of the project's conceptualization, resources, and methodologies, aiding in future adaptations, extensions, and exploitation of the Knowledge Graph's insights.

The resources, files, and data associated with this Knowledge Graph project are available in the project's GitHub repository. This repository includes the datasets, code, and documentation referenced throughout this document, as well as the final Knowledge Graph. For those interested in further exploring, utilizing, or contributing to the project, the repository provides open access to all materials and version histories, facilitating transparency, collaboration, and reusability. You can access the repository at https://github.com/DeeJack/WeatherTrentino.

The report is structured as follows:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: Description of the Information Gathering phase, that aims at identify, collect, clean and standardize data.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities.
- 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 Definition

2.1 Informal purpose

The aim of the project is to create a Knowledge Graph capable to establish a comprehensive and interconnected representation of the factors contributing to climate change within the autonomous province of Trento. This purpose centers on capturing and linking various types of data (geospatial, meteorological, and demographic) over historical time frames, with the goal of illuminating the relationship between human activities, environmental conditions, and resulting climate impacts in this specific region.

By integrating geospatial data, such as that from OpenStreetMap, with historical meteorological records, including temperature trends, air quality metrics, and other environmental indicators, the Knowledge Graph aims to reflect the environmental transformations that have occurred over the years. Additionally, historical demographic data, such as population density, energy consumption, gas usage, water consumption, and waste production, will provide insight into local patterns of resource utilization and waste generation. These variables will collectively inform an understanding of how these human and environmental factors interlink to drive climate change in Trento.

Ultimately, the Knowledge Graph will serve as a tool to clarify the underlying causes of climate change within the context of this region, allowing users to explore the interconnected nature of socio-environmental impacts and aiding in the development of targeted actions or policies to address these issues.

2.2 Context definition

The Context of this Knowledge Graph project encompasses geographical, temporal, and domain-specific boundaries that shape and constrain the analysis of climate change in the autonomous province of Trento, Italy. This project primarily reflects a Reference Context Purpose, aimed at examining the broader environmental and socio-economic factors in Trento without a focus on individual user perspectives. The context, however, implicitly considers the interaction between human activities and the reference environment, making it relevant for understanding how individuals and communities are affected by and contribute to climate change within this regional setting.

2.3 Domain of Interest (DoI)

The Domain of Interest for this Knowledge Graph project is climate change within the autonomous province of Trento. This domain encompasses both geographical and temporal dimensions that are essential to understanding the climate dynamics and human interactions in the region.

Spatially, the domain is limited to the province of Trento, situated in Northern Italy and characterized by diverse topography that includes mountainous landscapes, forests, rivers, and valleys. These geographic features play a critical role in shaping the local climate and contribute to the unique environmental challenges faced by the area. Trento's natural landscape, alongside its urban areas and agricultural zones, is integral to the Knowledge Graph, as it provides the physical context in which climate-related variables and human activities interact.

Temporally, the project focuses on historical data extending over the past several decades, capturing long-term changes and trends in climate, population, resource consumption, and environmental quality. This temporal scope is chosen to reveal patterns and variations that highlight how climate change has evolved in Trento over time, influenced by both natural and human factors. By studying these historical trends, we aim to identify and understand the temporal progression of climate effects and contributing factors specific to this region.

2.4 Scenarios definition

These scenarios illustrate how users might interact with the Knowledge Graph to analyze, interpret, and respond to climate-related information, providing a foundation for further definition of user personas in the next section.

2.4.1 Scenario 1: Policy Analysis and Environmental Planning

In this scenario, the Knowledge Graph supports policy analysts and environmental planners seeking to assess the impact of regional policies on climate change factors, such as emissions, energy consumption, and waste production. By exploring the relationships between policy implementations and climate indicators over time, users can evaluate the effectiveness of past policies and shape future strategies for mitigating climate impacts.

2.4.2 Scenario 2: Community Awareness and Education

This scenario focuses on enhancing community awareness about climate change in Trento. Educators, community leaders, and local activists can use the Knowledge Graph to access and visualize data on local climate conditions and human activities, creating educational materials or interactive sessions for residents. This scenario helps raise awareness of the impact of individual and collective actions on the environment, fostering informed community discussions and behavior change.

2.4.3 Scenario 3: Urban and Rural Development

In this scenario, urban planners and development specialists leverage the Knowledge Graph to analyze the effects of urbanization and rural land use on climate change in Trento. They can investigate factors like population density, energy use, and waste production within urban and rural areas to assess environmental impacts and guide sustainable development practices that align with climate objectives.

2.4.4 Scenario 4: Environmental Research and Analysis

Researchers and scientists studying climate change can use the Knowledge Graph to analyze historical climate and socio-economic data in Trento, identifying correlations and trends within this specific regional context. This scenario enables deep analysis of local climate change phenomena, serving as a foundation for scientific studies and publications that contribute to understanding and addressing climate change at the regional level.

2.4.5 Scenario 5: Infrastructure and Resource Management

This scenario addresses the needs of infrastructure managers and resource planners who require insight into the demand and usage patterns of resources like energy and water. By using the Knowledge Graph, they can assess how climate change and demographic shifts affect resource consumption, helping to optimize infrastructure planning and ensure the resilience of critical resources in the face of environmental change.

2.5 Personas definition

These semi-fictional characters represent key groups of end-users who will interact with the Knowledge Graph within the various scenarios defined. Each persona embodies distinct goals, motivations, and challenges, aligning with the core purpose of the Knowledge Graph project and reflecting broader user needs within the climate change domain. These personas, along with their associated use cases, provide a robust framework to ensure that the Knowledge Graph design addresses a range of user perspectives.

2.5.1 Persona 1: Giulia Rossi, Policy Analyst

Background: Giulia is a 34-year-old environmental policy analyst working for the provincial government in Trento. She has a master's degree in environmental science and specializes in evaluating the effectiveness of local and regional environmental policies.

Goals: Giulia wants to assess the impact of Trento's climate policies on reducing emissions, improving air quality, and managing resources. She's also interested in identifying data-supported recommendations for future policy adjustments.

Challenges: Her main challenge is accessing up-to-date, interconnected data that shows how past policies have influenced climate indicators. She needs reliable, actionable insights to back her policy recommendations, especially when presenting findings to decision-makers.

Use Case: Giulia uses the Knowledge Graph within the Policy Analysis and Environmental Planning scenario 2.4.1, accessing historical data on climate indicators and policy impacts. She generates reports that link policy decisions to measurable outcomes, allowing her to advocate for evidence-based policy adjustments.

2.5.2 Persona 2: Marco Bianchi, Science Teacher

Background: Marco, 45, teaches high school science in Trento. Passionate about climate education, he seeks engaging ways to make complex climate concepts accessible and relatable for his students.

Goals: Marco aims to help students understand the local effects of climate change and encourage eco-conscious behaviors. He wants to create interactive lessons and demonstrations using real-world data from their region.

Challenges: Marco struggles to find localized data that visually represents climate change impacts in an easily understandable way for teenagers. He needs a tool that translates raw data into a story that resonates with his students.

Use Case: In the Community Awareness and Education scenario 2.4.2, Marco uses the Knowledge Graph to pull localized climate data and display visualizations in his classroom. He develops lesson plans that show the impact of specific actions, like waste reduction, on the environment, inspiring students to participate in local environmental efforts.

2.5.3 Persona 3: Sara Verdi, Urban Planner

Background: Sara is a 39-year-old urban planner with the municipality of Trento, specializing in sustainable development. She has experience in balancing urban growth with environmental conservation and climate resilience.

Goals: Sara is focused on planning urban and rural areas that are environmentally sustainable. She aims to understand the interplay between population density, resource consumption, and climate impact to develop long-term plans that align with climate goals.

Challenges: Sara faces the challenge of accessing integrated datasets that link urbanization patterns to climate data, such as emissions or energy consumption, over time. She needs insights that help her balance growth and sustainability effectively.

Use Case: Within the Urban and Rural Development scenario 2.4.3, Sara uses the Knowledge Graph to analyze relationships between urban density and resource consumption, identifying areas where sustainable practices can be implemented. This enables her to propose zoning and infrastructure plans that support climate resilience.

2.5.4 Persona 4: Luca Ferro, Environmental Researcher

Background: Luca, 29, is an environmental researcher pursuing a Ph.D. in climate science with a focus on regional climate impacts. He is keen on studying how climate change affects ecosystems and communities in specific regions, including Trento.

Goals: Luca aims to conduct in-depth research into the correlation between human activities and climate changes in Trento. He wants access to comprehensive datasets to explore cause-effect relationships between environmental changes and socio-economic factors.

Challenges: Luca's research requires high-quality, localized datasets that are often fragmented. He needs a consolidated data source that supports complex queries and data exploration to support his academic research.

Use Case: In the Environmental Research and Analysis scenario 2.4.4, Luca utilizes the Knowledge Graph to access interlinked environmental and socio-economic data over time. This allows him to explore and test hypotheses on the impacts of human activities, contributing to publications that inform the scientific community.

2.5.5 Persona 5: Antonio Russo, Resource Management Specialist

Background: Antonio, 52, is a senior infrastructure and resource manager at a public utility company. He oversees resource planning, including water and energy supply, and ensures resilience against climate impacts.

Goals: Antonio's goal is to optimize resource management strategies in response to evolving climate patterns and demographic changes. He needs accurate, predictive data to assess future resource demands.

Challenges: He often contends with incomplete data, making it difficult to accurately project future resource requirements. He requires data that factors in both current usage trends and anticipated climate impacts to effectively plan for the long term.

Use Case: Within the Infrastructure and Resource Management scenario 2.4.5, Antonio leverages the Knowledge Graph to study historical trends in resource consumption linked to climate patterns, allowing him to prepare adaptive strategies that anticipate resource needs under various climate scenarios.

2.6 Competency Question definition

These critical questions defines what the Knowledge Graph (KG) must answer to meet the needs of the personas within each scenario. These CQs represent the functional requirements of the KG, providing a guide to the entities and relationships that should be modeled. Each question reflects a specific need of a persona in a scenario and indicates the diverse information paths the KG must support.

2.6.1 Policy Analysis and Environmental Planning

CQ1 How have specific policies influenced emission levels in Trento over the past decade? (2.5.1, 2.4.1)

CQ2 Which regions within Trento have seen the most significant reductions in air pollutants following policy implementations? (2.5.1, 2.4.1)

CQ3 What relationships exist between waste reduction initiatives and air quality metrics over the years? (2.5.1, 2.4.1)

CQ4 What are the policy measures that have shown the most significant impact on reducing energy consumption in urban versus rural areas? (2.5.1, 2.4.1)

2.6.2 Community Awareness and Education

CQ5 What is the historical trend of energy consumption in Trento, and how does it correlate with population density? (2.5.2, 2.4.2)

CQ6 How does waste production in Trento vary by neighborhood, and what seasonal patterns are observed? (2.5.2, 2.4.2)

CQ7 Which areas have experienced the most pronounced changes in air quality, and what are the potential causes? (2.5.2, 2.4.2)

CQ8 What are the projected environmental impacts if current trends in waste production continue? (2.5.2, 2.4.2)

2.6.3 Urban and Rural Development

CQ9 How does resource consumption (water, gas, energy) differ between urban and rural areas of Trento? (2.5.3, 2.4.3)

CQ10 What is the relationship between urban density and waste production over time in Trento? (2.5.3, 2.4.3)

CQ11 Which urban areas are at the greatest risk of environmental degradation due to current resource consumption rates? (2.5.3, 2.4.3)

CQ12 How have infrastructure improvements influenced energy efficiency and climate resilience in different parts of Trento? (2.5.3, 2.4.3)

2.6.4 Environmental Research and Analysis

CQ13 What are the trends in average temperature changes across Trento over the last 50 years? (2.5.4, 2.4.4)

CQ14 How does population density correlate with air quality indicators in urban and rural zones of Trento? (2.5.4, 2.4.4)

CQ15 What are the statistical correlations between gas consumption and emission levels? (2.5.4, 2.4.4)

CQ16 What patterns emerge when comparing climate indicators (temperature, precipitation) with demographic changes? (2.5.4, 2.4.4)

CQ17 What seasonal variations are observed in air quality, and how do these relate to energy consumption patterns? (2.5.4, 2.4.4)

2.6.5 Infrastructure and Resource Management

CQ18 How do seasonal shifts impact water and energy demand across Trento? (2.5.5, 2.4.5)

CQ19 What is the projected resource demand if population growth continues at the current rate in Trento? (2.5.5, 2.4.5)

CQ20 Which regions are projected to face resource scarcity, and how can management practices be adapted? (2.5.5, 2.4.5)

CQ21 How does residential versus industrial gas consumption vary by region, and what trends are projected? (2.5.5, 2.4.5)

CQ22 What historical data on resource consumption align with trends in urban development? (2.5.5, 2.4.5)

2.7 Concepts Identification

This section organizes the essential entities (ETypes) and their properties necessary to satisfy the Competency Questions (CQs) and meet the purpose of the Knowledge Graph (KG) project; categorizing concepts into three levels of importance: Common, Core, and Contextual, which indicate the relevance and potential reusability of each entity and property within the KG. This categorization not only shapes the KG for this specific purpose but also enhances its reusability and interoperability for broader applications.

The Focus Categories are defined as follows:

- Common: Entities and properties broadly applicable to the purpose and potentially reusable across multiple contexts.
- Core: Entities and properties essential to the purpose and central to addressing the specified CQs, though possibly more domain-specific.
- Contextual: Highly specific entities and properties that address niche or unique aspects of the purpose, likely less reusable outside this context.

2.8 ER modeling

The competency questions are used to design the ER model, ensuring that entities have the right properties to analyze climate change data in the right way.

This model maintains historical data in various environmental measurements to track climate change and understanding its impact within Trentino.

At the top level, the Location entity represents the geographic area associated with each measurement, then we have Measurement, the super-type for data measurement like air quality level, pollution levels, temperature, and so on.

The Weather Station entity links each station to its Location, with additional details like altitude and station type, specifying whether it measures weather, air quality, or pollution.

Scenarios	Personas	CQs	Entities	Properties	Focus
all	all	CQ2, CQ4, CQ6, CQ7, CQ9, CQ11, CQ20, CQ21, CQ22	Location	location_id, name, latitude, longitude, altitude, type	Common
S1, S2, S4	P1, P2, P4	CQ1, CQ2, CQ7, CQ13, CQ14, CQ15, CQ16, CQ17	Weather Station	station_id, location_id, name, type	Core
all	all	all	Measurement	measurement_id, station_id, timestamp	Core
S1, S2, S4	P1, P2, P4	CQ3, CQ7, CQ14, CQ17	Air Quality	air_quality_id, value, timestamp	Core
S1, S4	P1, P4	CQ1, CQ2, CQ15	Pollution	pollution_id, station_id, polluting, value, measurement_unit, timestamp	Core
S1, S2, S3	P1, P2, P3	CQ3, CQ6, CQ8, CQ10	Waste	waste_id, location_id, differentiated, undifferentiated, timestamp	Core
S1, S2, S3, S4, S5	P1, P2, P3, P4, P5	CQ4, CQ5, CQ9, CQ11, CQ12, CQ15, CQ17, CQ18, CQ19, CQ20, CQ21, CQ22	Consumption	consumption_id, value, timestamp	Core
S2, S3, S4	P2, P3, P4	CQ8, CQ11, CQ12, CQ13, CQ16	Weather	weather_id, station_id, temperature_min, temperature_max, temperature_avg, rainfall, rain_mm, snowfall, humidity, visibility_avg, wind_avg, timestamp	Common
S2, S3, S4, S5	P2, P3, P4, P5	CQ5, CQ10, CQ14, CQ16, CQ19	Demographic	demographic_id	Common

Table 1: Entity Types (ETypes) and Properties

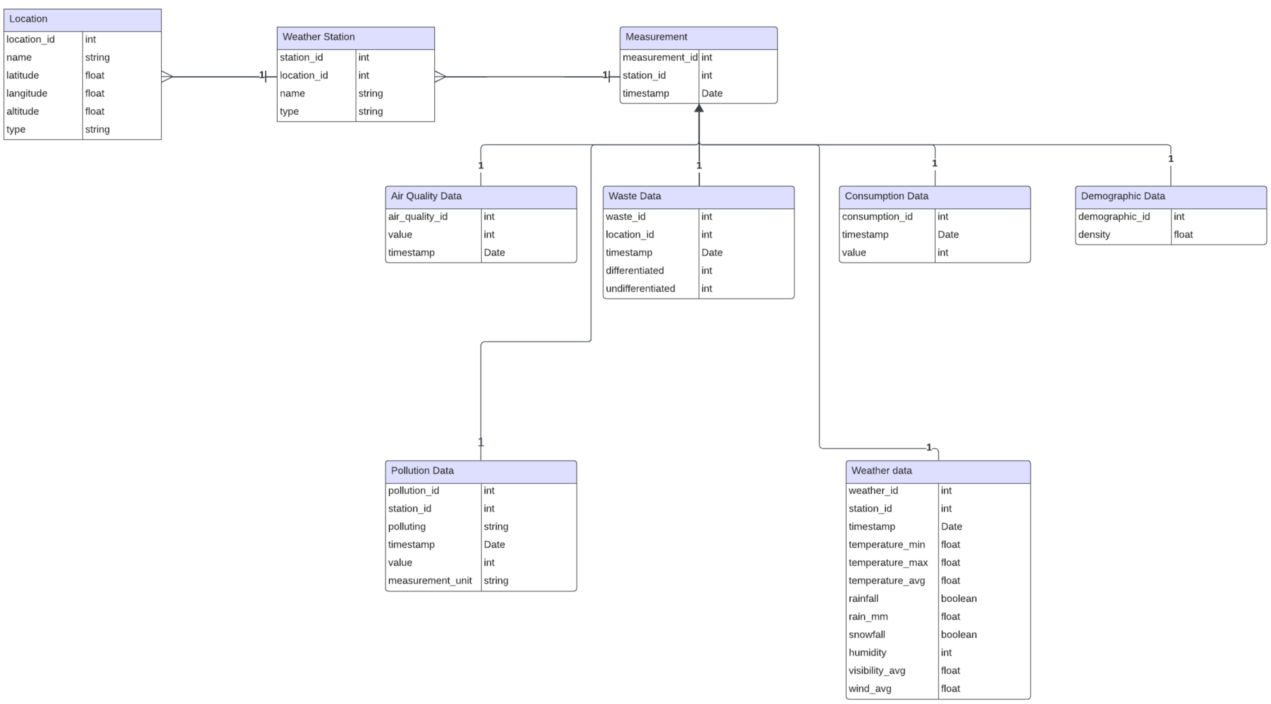

Figure 1: ER diagram

3 Information Gathering

3.1 Source Identification

The Source Identification activity is crucial for locating, evaluating, and selecting the datasets that will feed into the knowledge graph construction process. This process aims to not only leverage the sources provided but also expand to identify additional resources when the initial sources are insufficient to meet the outlined Purpose (see Chapter 2). A core principle here is data reuse, where the user actively searches for pre-existing datasets that can fulfill the informational needs of the project, reducing redundant data generation.

Given the diversity of data types required, the sources identified vary in structure and quality, ranging from high-quality, interoperable catalogs to less formal or heterogeneous sources. The high-quality sources, often catalogs that distribute well-documented and interoperable datasets, provide data with enhanced metadata and adherence to established standards, and because of that are preferred. For example, repositories such as LiveData catalogs host resources like Knowledge Graphs (see Table 6, Table 5), produced by processes like iTelos, prioritizing reusability and standardization.

In contrast, low-quality sources may include more diverse data providers, such as general-purpose websites or purpose-specific databases. These sources often have limited metadata and may lack standardization, resulting in lower interoperability and reusability. Examples include basic web pages (see Table 2), local databases, or repositories that serve a narrow, application-specific purpose.

The tables at the end of this section (see Table 2 to Table 8) summarize the various data sources identified in this phase, detailing their domain, keywords, providers, data formats, and descriptions. These resources, once identified, will be curated and improved through processes like data cleaning and the adoption of standardized formats to ensure they contribute effectively to the knowledge graph's final structure and usability.

Resource name	**IlMeteo.it**
Domain	Trento
Keyword	Weather data
Language	English
Provider	ilMeteo.it
Data URL	ilmeteo.it
Data Format	html pages
Data Description	Archive of weather data from 1973 up to the present.
Knowledge URL	N/A
Knowledge description	N/A

Table 2: IlMeteo.it

Resource name	Meteo Trentino
Domain	Trento
Keyword	Weather data
Language	Italian
Provider	meteotrentino.it
Data URL	meteotrentino.it
Data Format	xlsx
Data Description	Historical weather data from different stations in the territory of Trentino
Knowledge URL	N/A
Knowledge description	N/A

Table 3: MeteoTrentino.it

Resource name	Air Quality data
Domain	Trentino
Keyword	Air quality
Language	Italian
Provider	OpenData Trentino
Data URL	OpenData
Data Format	.csv file
Data Description	Numero di superamenti del valore limite giornaliero previsto per le PM10
Knowledge URL	N/A
Knowledge description	N/A

Table 4: Air Quality Resource

Resource name	KGE23 - Trentino Climate Change
Domain	Trentino
Keyword	Weather Forecast, Weather History, Climate Change
Language	English
Provider	University Of Trento
Data URL	Datasets
Data Format	.ttl files
Data Description	Knowledge graph about Weather Data
Knowledge URL	PDF report
Knowledge description	PDF report

Table 5: KGE23 - Trentino Climate Change

Resource name	**KGE22 - Trentino Climate Change**
Domain	Trentino
Keyword	Weather Forecast, Weather History, Climate Change
Language	English
Provider	University Of Trento
Data URL	Datasets
Data Format	.ttl files
Data Description	Knowledge graph about Weather Data
Knowledge URL	PDF report
Knowledge description	PDF report

Table 6: KGE22 - Trentino Climate Change

Resource name	**Pollution**
Domain	Trentino
Keyword	Pollution, Air Quality
Language	Italian
Provider	Agenzia Provincia per la Protezione Ambiente
Data URL	OpenData
Data Format	.csv file
Data Description	Pollution for various measures (Biossido di Zolfo, Ozono, Monossido di Carbonio, Biossido di Azoto, PM10, PM2,5)
Knowledge URL	N/A
Knowledge description	N/A

Table 7: Pollution resource

Resource name	**Waste**
Domain	Trentino
Keyword	waste, differentiated, undifferentiated
Language	Italian
Provider	Provincia Autonoma di Trento
Data URL	Undifferentiated, differentiated
Data Format	.csv files
Data Description	Contains the waste produced by various zones in Trento
Knowledge URL	N/A
Knowledge description	N/A

Table 8: Waste resource

3.2 Data Collection

To initiate the data collection process, we sourced datasets from several providers, primarily through downloadable CSV or Json files, as listed in the previous section (see Table 2 to Table 8). However, accessing certain open datasets presented notable challenges:

- Weather Data:
  - Data from ilMeteo.it required month-by-month scraping, which posed significant time and resource costs.
  - Data from MeteoTrentino was explored as an alternative source for historical data.
- Pollution Data: Data from the Agenzia Provinciale per la Protezione Ambiente was downloadable as a CSV file through bollettino.appa.tn.it.
- Consumption Data: Open APE provided energy consumption data as an easily accessible CSV file.
- Waste, Air Quality, and Demographic Data: Intended resources from Trentino's Open Data portal (e.g., rifiuti urbani, demographic data) were unfortunately inaccessible due to server issues or page errors. It was then possible to retrieve some of the data by tweaking the APIs calls, while other still are not accessible.

Given these accessibility challenges, we adapted by embracing data reuse from high-quality knowledge graph datasets developed in prior years (KGE22 and KGE23). These knowledge graphs already contained weather and air quality data with high metadata quality, which allowed us to bypass scraping and source exploration for weather data, further enhancing data reliability and consistency. By prioritizing the use of these KGs, we aligned with the reuse philosophy central to our project, focusing instead on areas such as demographic, energy consumption, waste and pollution data, expanding those KGs with the data we were interested into.

3.3 Data Cleaning

To ensure data usability, Python scripts were employed to parse and clean each CSV dataset.

3.3.1 Demographic and Energy Consumption Data

Cleaning and enriching this dataset required the use of an auxiliary CSV containing Italian municipality names and their corresponding ISTAT codes. By joining this utility file with the primary energy consumption data, we were able to associate each building's consumption data with a specific municipality. For privacy reasons, and due to the lack of precise geolocation data that could be mapped onto OpenStreetMap, the project did not use precise building coordinates. City-level aggregation was deemed sufficient for project objectives.

3.3.2 Waste Data

The waste data came divided in 2 main files, differentiated waste and undifferentiated waste. Cleaning and enriching this dataset would have required the use of an auxiliary CSV containing the Waste Company Codes in order to link those data to a location, similarly to what we did for the Energetic Consumption data. In this case such data was not accessible, neither with the APIs tweaking performed for other sources, therefore the cleaning consisted in the merging of the two files, adding a column to identify the value type (Differentiated or Undifferentiated). We hope to gain access to a way to map those codes to the territory in the future, to gain real value of these data.

3.3.3 Data Column Pruning

Each dataset underwent a column-filtering process, retaining only the fields pertinent to our analysis. This step was essential for removing unnecessary data and optimizing storage and processing efficiency.

3.3.4 Cleaning Output

This process lead to the creation of new, processed csv files, now cleaned and ready for the next iterations. In particular the file obtained were:

- `open_ape_clean.csv`: The energetic consumption cleaned data.
- `waste_clean.csv`: The waste cleaned data.

3.4 Data Standardization

For standardization, we adopted a unified CSV format across all datasets, using a comma delimiter to ensure compatibility and readability. This approach allowed for seamless integration of diverse data sources and supported efficient data manipulation and analysis in subsequent phases. To accomplish this, other python scripts were used to transform the data retrieved in other formats (e.g. Json) to the accorded csv standard.

4 Language Definition

This section describes the process of selecting concepts essential for representing data in our Knowledge Graph.

We leveraged the Universal Knowledge Core (UKC) for predefined concepts and introduced new ones where necessary to ensure comprehensive coverage.

In the first stage, we analyzed the concepts derived from the ER model in Figure 1 and matched them against the UKC. Any concepts absent from the UKC were documented in a spreadsheet and defined accordingly.

The second stage involved Dataset Filtering, where we refined the data to align with the project's objectives and ensure relevance.

The finalized concepts are illustrated in Figure 3 and Figure 2.

4.1 Concepts Identification

For the identification of the concepts, we reviewed the ER model, extracting relevant keywords.

The concepts present in the UKC include:

- **Weather station**: A facility that collects meteorological data, such as temperature, humidity, and rainfall.
- **Waste**: Refuse or materials discarded by human activity, including recyclable and non-recyclable waste.
- **Pollutant**: Substances introduced into the environment that cause harm or contamination.
- **Location**: A specific place where data is collected or an event occurs.
- **Latitude**: The geographical coordinate indicating a location's distance north or south of the equator.
- **Longitude**: The geographical coordinate indicating a location's distance east or west of the Prime Meridian.
- **Altitude**: The height of a location above sea level.
- **Rainfall**: The total amount of precipitation measured over a period of time.
- **Snowfall**: The amount of snow that falls in a given period.
- **Humidity**: The concentration of water vapor present in the air.

The following concepts were added to capture additional details not present in the UKC:

- **pm10**: Particulate matter with a diameter of 10 micrometers or less.
- **pm25**: Particulate matter with a diameter of 2.5 micrometers or less.
- **so2**: Sulfur dioxide, a gas produced by volcanic activity and industrial processes.
- **o3**: Ozone, a gas found in the Earth's atmosphere, important for blocking ultraviolet radiation but harmful at ground level.
- **air_quality**: A measure of the cleanliness or pollution level of the air.
- **temperature_avg**: The average temperature recorded over a specific period.
- **temperature_min**: The minimum temperature recorded over a specific period.
- **temperature_max**: The maximum temperature recorded over a specific period.
- **rain_mm**: The total rainfall measured in millimeters.

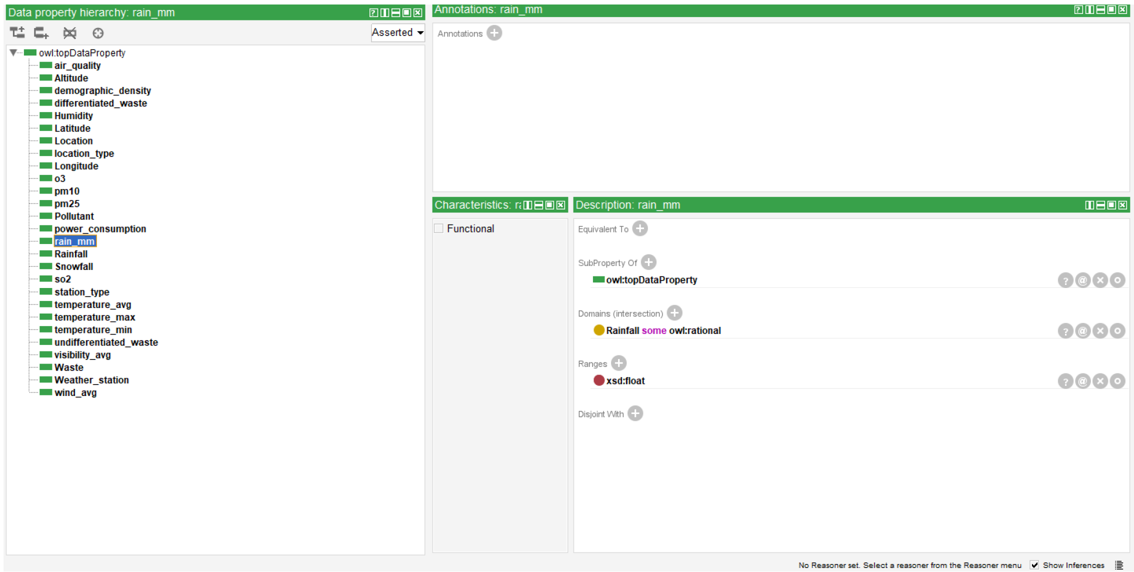

Figure 2: Protege

- **visibility_avg**: The average distance over which objects can be seen, indicating atmospheric clarity.
- **wind_avg**: The average wind speed measured over a specific period.
- **differentiated_waste**: Waste that has been sorted for recycling or specific disposal processes.
- **undifferentiated_waste**: Waste that has not been sorted and is disposed of as general refuse.
- **power_consumption**: The total amount of electricity used by a specific location or population.
- **demographic_density**: The number of people living per unit area of a region.
- **location_type**: The classification of a location (e.g., urban, rural, industrial).
- **station_type**: The type of weather or monitoring station, categorized by its purpose or data collection capabilities.

4.2 Dataset filtering

Dataset filtering was performed in the previous phase to retain only the columns relevant to the construction of the knowledge graph.

In particular, several columns were removed from the 'open_ape" dataset, such as 'anno_costruzione', 'qualita_involucro', and 'qualita_impianti'.

Furthermore, the two waste data sets (differentiated and undifferentiated) were merged by introducing a 'waste_type' field to distinguish between the two.

ConcetID	Word-en	Gloss-en	Word-Other language
KGE2024-1-0001	pm10	Particulate matter with a diameter of 10 micrometers or less.	particolato_pm10
KGE2024-1-0002	pm25	Particulate matter with a diameter of 2.5 micrometers or less.	particolato_pm25
KGE2024-1-0003	so2	Sulfur dioxide, a gas produced by volcanic activity and industrial processes.	biossido_di_zolfo
KGE2024-1-0004	o3	Ozone, a gas found in the Earth's atmosphere, important for blocking ultraviolet r	ozono
KGE2024-1-0005	air_quality	A measure of the cleanliness or pollution level of the air.	qualità_dell'aria
KGE2024-1-0006	temperature_avg	The average temperature recorded over a specific period.	temperatura_media
KGE2024-1-0007	temperature_min	The minimum temperature recorded over a specific period.	temperatura_minima
KGE2024-1-0008	temperature_max	The maximum temperature recorded over a specific period.	temperatura_massima
KGE2024-1-0009	rain_mm	The total rainfall measured in millimeters.	pioggia_mm
KGE2024-1-0010	visibility_avg	The average distance over which objects can be seen, indicating atmospheric cla	visibilità_media
KGE2024-1-0011	wind_avg	The average wind speed measured over a specific period.	vento_medio
KGE2024-1-0012	differentiated_waste	Waste that has been sorted for recycling or specific disposal processes.	rifiuti_differenziati
KGE2024-1-0013	undifferentiated_waste	Waste that has not been sorted and is disposed of as general refuse.	rifiuti_indifferenziati
KGE2024-1-0014	power_consumption	The total amount of electricity used by a specific location or population.	consumo_energetico
KGE2024-1-0015	demographic_density	The number of people living per unit area of a region.	densità_demografica
KGE2024-1-0016	location_type	The classification of a location (e.g., urban, rural, industrial).	tipo_di_località
KGE2024-1-0017	station_type	The type of weather or monitoring station, categorized by its purpose or data coll	tipo_di_stazione

Figure 3: Concepts added

5 Knowledge Definition

This section is dedicated to the description of the Knowledge Definition phase. Like in the previous section, it aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced.

Knowledge Definition sub activities:

- KTelos
  - Teleology definition
  - Teleontology definition
- Dataset cleaning and formatting

The report of the work done during this phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

6 Entity Definition

This section is dedicated to the description of the Entity Definition phase. Like in the previous section, it aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced.
Entity Definition sub activities:

- Entity matching
- Entity identification
- Data mapping

The report of the work done during this phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

7 Evaluation

This section aims at describing the evaluation performed at the end of the whole process over the final outcome of the iTelos methodology. More in details, this section as to report:

- the final Knowledge Graph information statistics (like, number of etypes and properties, number of entities for each etype, and so on).
- Knowledge layer evaluation: the results of the application of the evaluation metrics applied over the knowledge layer of the final KG.
- Data layer evaluation: the results of the application of the evaluation metrics applied over the data layer of the final KG.
- Query execution: the description of the competency queries executed over the final KG in order to test the suitability of the KG to satisfy the project purpose.

8 Metadata Definition

In this section the report collects the definitions of all the metadata defined for the different resources produced along the whole process. The metadata defined in this phase describes both the final outcome of the project, and the intermediate outcome of each phase (language, schema, and data source standardised values).

The definition of the metadata, is crucial to enable the distribution (sharing) of the resource produced, through the data catalogs. For this reason it is important to describe also where such metadata will be published to distribute the resources it describes (for example the DataScientia catalogs).

In particular the structure of this section is organized as follows, with the objective to describe the metadata relative to all the type of resources produced by the project.

- Project metadata description
- Language resources metadata description
- Knowledge resources metadata description
- Data resources metadata description

9 Open Issues

This section concludes the current document with final conclusions regarding the quality of the process and final outcome, and the description of the issues that (for lack of time or any other cause) remained open.

- Did the project respect the scheduling expected in the beginning ?
- Are the final results able to satisfy the initial Purpose ?
  - If no, or not entirely, why ? which parts of the Purpose have not been covered ?

Moreover, this section aims to summarize the most relevant issues/problems remained open along the iTelos process. The description of open issues has to provide a clear explanation about the problems, the approaches adopted while trying to solve them and, eventually, any proposed solution that has not been applied.

- which are the issues remained open at the end of the project ?