UNIVERSITY
OF TRENTO - Italy

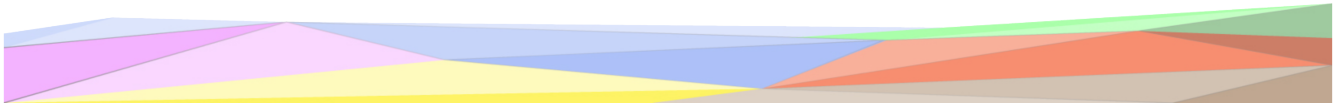
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2024 - HealthRoute Trentino

Project Report

Document Data:	Reference Persons:
November 30, 2024	Alice Massacci



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 16, 2024	Alice Massacci	Document created
1.0	October 19, 2024	Alice Massacci	Informal Purpose & DoI
1.1	October 26, 2024	Alice Massacci	Activity 1: Personas & Scenarios definition
1.2	October 27, 2024	Alice Massacci	Activity 2: Competency Questions definition
1.3	October 27, 2024	Alice Massacci	Activity 3: Concepts Identification
1.4	October 30, 2024	Alice Massacci	Purpose Definition completed
2.0	November 1, 2024	Alice Massacci	Information Gathering - Data Sources
2.1	November 2, 2024	Alice Massacci	Information Gathering - Data Sources and Resources Collection
2.2	November 3, 2024	Alice Massacci	Information Gathering - Data Sources and Resources Collection
2.3	November 10, 2024	Alice Massacci	Information Gathering - Resource Cleaning
2.4	November 11, 2024	Alice Massacci	Information Gathering - Resource Cleaning
2.5	November 13, 2024	Alice Massacci	Information Gathering - Resource Cleaning
2.6	November 14, 2024	Alice Massacci	Information Gathering - Resource Cleaning
2.7	November 15, 2024	Alice Massacci	Information Gathering completed
3.0	November 17, 2024	Alice Massacci	Language Definition - Concept Identification
3.0	November 17, 2024	Alice Massacci	Language Definition - Concept Identification
3.1	November 23, 2024	Alice Massacci	Language Definition - Concept Identification
3.2	November 24, 2024	Alice Massacci	Language Definition - Concept Identification
3.3	November 30, 2024	Alice Massacci	Language Definition - Dataset Filtering

1 Introduction

The current document aims to provide a detailed report of the project developed following the iTelos methodology. The report is structured as follows:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: Definition of the Information Gathering iTelos phase.
- Section 4: Definition of the Language Definition iTelos phase.

2 Purpose Definition

2.1 Informal Purpose

The **HealthRoute Trentino** project aims to develop a Knowledge Graph (KG) that provides seamless access to up-to-date information about Healthcare Facilities across the Trentino Province, along with their integration into the public transportation network. The KG will empower users—whether healthcare professionals or patients—to easily locate medical services (e.g., hospitals, pharmacies, medical centers) and discover the most efficient transit routes to access them. The system aims to streamline healthcare access and improve overall coordination between health services and transportation options.

The informal purpose can be stated as: *"A KG that helps users quickly find healthcare facilities/services in Trentino and determine how to reach them efficiently using public transportation, based on their location and transport data."*

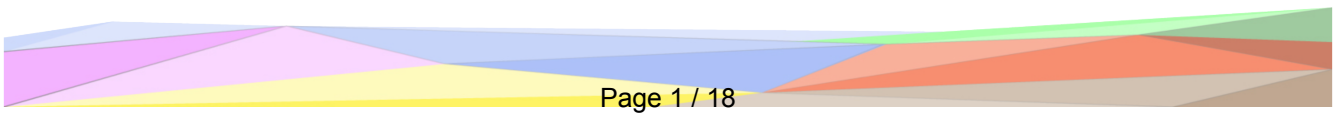
The KG should be capable of answering complex queries like "Which hospital can I reach by bus within 30 minutes from my current location?" or "Which pharmacies are accessible via public transport in the evening?" The KG will integrate static transportation data from different sources with healthcare facilities and synthetic user data, and will serve as a foundation for web or smartphone applications that assist users in planning trips to healthcare facilities based on real-time schedules, transportation options, and user location.

2.2 Domain of Interest (DoI)

The DoI for this project is bounded in space and time:

- Space: autonomous province of Trentino.
- Time: To ensure practical applicability, the project will utilize current public transportation data from the year 2024.

These boundaries ensure that the project operates within a realistic and usable framework for public transportation information.



2.3 Purpose Definition – Activities

Once the informal purpose and DoI are defined, the iTelos methodology moves to formalization of the initial purpose statement into a purpose-specific Entity-Relationship (ER) model. Purpose formulation involves four key Activities:

- Personas & Scenarios definition
- Competency Questions definition
- Concepts Identification
- ER modeling

2.3.1 Activity 1: Personas & Scenarios definition

To guide the design and development of the KG, a set of personas—fictional representations of different end-user types with specific needs and objectives—and scenarios have been defined. These scenarios cover a variety of transportation and healthcare-related challenges, ensuring coverage of rush hour demands, working days and holidays. By framing the project through these user-centered examples, the practical applications of the HealthRoute Trentino KG are demonstrated.

Scenarios

Working Day (Monday to Friday) – Morning (S1)
On a typical weekday morning in Trento, between 7 a.m. and 12 p.m., the city is bustling with activity as residents commute to work and students head to school. The public transport system runs at full capacity. Buses, trams, and trains operate frequently to accommodate the surge in passengers, with most routes servicing both urban and suburban areas. Healthcare workers–doctors, nurses, and administrative staff–rely heavily on public transport to reach hospitals, clinics, and medical centers on time to start their shifts. Patients, especially those with scheduled appointments, also need reliable transportation.

Nighttime Travel (S2)
It's late at night in Trento, between 12 a.m. and 6 a.m. The city is quiet, with most residents at home and businesses closed, except for emergency services like hospitals. Public transportation services are limited, with only a few night buses running. Passengers needing to travel during these hours, such as night-shift workers, may experience longer waiting times and fewer route options.

Weekend Reduced Service (S3)
On weekends, especially Sundays, Trento slows down as businesses close. Public transportation runs on reduced schedules, with fewer buses and trains in operation. Hospitals and essential services remain open.

Heavy Rain (S4)
It's raining heavily, and streets are slick with water, making outdoor travel less appealing. Fewer people are

walking outside, opting instead for the warmth and shelter of public transportation. Buses are more crowded than usual. The weather conditions also slow down traffic. Passengers face longer waiting times and a slower commute.

Strike Day (S5)
On strike days, both buses and trains can be delayed or cancelled altogether, creating significant disruptions for commuters. People who rely solely on public transportation face difficulties, as they might be unable to get to work, miss important meetings, or face long delays.

Traffic Disruption Due to Construction Work (S6)
Ongoing construction work on a major road in Trento has caused significant delays throughout the city. Public buses are forced to take alternative routes, resulting in longer travel times for passengers who might miss connections or arrive late at their destinations. The detours also add extra pressure on already busy streets.

Personas

Stefania (P1): Stefania is a 36-year-old social worker living in Borgo Valsugana. She frequently visits nursing homes and hospitals as part of her job. Stefania also has to manage her own health, picking up prescriptions for thyroid medication and visiting her endocrinologist in Trento.

Riccardo (P2): Riccardo is a 45-year-old IT consultant from Pergine Valsugana. After suffering a minor heart attack a year ago, he became committed to improving his health through regular cardiovascular check-ups and a strict fitness routine. He prefers using public transportation to reduce stress and improve his health further, often combining his trips with walks in the local parks. He also suffers from seasonal allergies and regularly visits the local pharmacy during spring and summer to buy antihistamines and nasal sprays. Riccardo often schedules his trips to the pharmacy for early morning to avoid the intense heat in the summer.

Elena (P3): Elena is a 38-year-old teacher living in Rovereto. She has asthma and needs regular check-ups with a pulmonologist. When the weather is favorable, Elena uses her scooter to travel to her appointments. However, during colder months or when her asthma flares up, she relies on public transportation for her commutes. Her appointments are often scheduled for the early morning, during the busy weekday rush hours.

Chiara (P4): Chiara is a 29-year-old nurse working at a hospital in Trento. She is often on rotating shifts, including nights and weekends. Her day shifts run from 7 am to 3 pm, while night shifts run from 11 pm to 7 am. In addition to her regular duties, Chiara is part of a professional development program that requires her to travel to different healthcare facilities across the region. She mentors new nurses and attends training sessions and seminars to stay updated on the latest medical practices. Chiara often relies on public transport for her commutes.

Edoardo (P5): Edoardo is a 50-year-old man who lives in Caldonazzo and works in Trento. Due to his job as a restaurant manager, Edoardo often works late into the night. He has chronic insomnia and sometimes needs to visit the pharmacy after work to get over-the-counter sleep aids. Since he doesn't drive, Edoardo depends on late-night bus services and 24-hour pharmacies to manage his condition without having to take time off work.

2.3.2 Activity 2: Competency Questions (CQs) definition

- **CQ-1 (P1-S5)**: Stefania has a busy Monday morning with back-to-back appointments at two nursing homes, but a public transport strike has been announced.
  a) Which train/bus lines does she rely on for her appointments?
  b) Are they affected by the strike?

- **CQ-2 (P1-S1)**: Stefania has a morning visit scheduled at a nursing home in Trento at 11:00 a.m.
  a) What is the most efficient public transportation route for her to arrive at the nursing home on time?
  b) Is there a pharmacy within 500 meters of the final bus stop?
  c) If she leaves home at 9:30 a.m., will she have sufficient time to stop by the pharmacy to buy her thyroid medications before proceeding to her appointment?

- **CQ-3 (P2-S1)**: I'ts Wednesday, Riccardo has a cardiovascular check-up scheduled for 9:00 a.m. at Ospedale Santa Chiara and plans to pick up some antihistamines beforehand.
  a) How many pharmacies are located within a 5-10 minute walk from each bus stop on his route to the hospital, and which one would be the most convenient to stop at?
  b) If he leaves home at 7:30 a.m., will he arrive on time for his appointment?

- **CQ-4 (P2-S3)**: After enjoying a walk in Parco delle Albere on a sunny Sunday afternoon, Riccardo begins experiencing allergy symptoms and needs to buy antihistamines.
  a) Is there a nearby pharmacy open on Sundays?
  b) Given the reduced weekend public transport service, what is the fastest route he can take to reach the pharmacy and then continue home?

- **CQ-5 (P3-S6)**: Elena has a morning appointment with her pulmonologist at Ospedale Santa Chiara in Trento. Due to ongoing construction on the train line between Rovereto and Trento, she must rely on replacement bus services.
  a) What is the best route for her to take?
  b) Which is the travel time using the replacement bus?
  c) If Elena's appointment is scheduled for 10:00 a.m., what time should she leave her home to catch the replacement bus and ensure she arrives on time?

- **CQ-6 (P4-S6)**: Chiara finishes her shift at Ospedale S. Camillo at 3:00 p.m. and needs to travel to Centro Medico di Rovereto for a training session at 4:30 p.m., but there is ongoing construction causing delays on the usual bus route. What alternative bus routes can she use to reach the training center on time despite the disruption?

- **CQ-7 (P4-S4)**: Chiara is finishing her night shift at the hospital at 7:00 a.m. on a rainy weekday.
  a) What public transportation options are available for her to get home after her shift?
  b) If the bus arrives but is overcrowded due to the heavy rain, how long will she have to wait for the next available bus?

- **CQ-8 (P5-S2)**: Edoardo finishes his late shift at the restaurant at 1:00 a.m. and needs to stop by a 24-hour pharmacy to pick up sleep aids. Given the limited late-night bus service, what is the quickest way for him to reach the nearest open pharmacy and then head home to Caldonazzo?

2.3.3 Activity 3: Concepts Identification

Considering the scenarios and personas involved in the CQs, the following entities and their properties can be identified:

Scenarios	Personas	CQs	Entities	Properties	Focus
1–6	1–5	1–8	End_User	id, name, gender, age, type	Contextual
1–6	1–5	1–8	Trip	id, start_time, end_time, start_point, end_point, route	Contextual
1–6	1–5	1–8	Route	id, type, provider, length, schedule	Core
1–6	1–5	1–8	Position	id, address, latitude, longitude	Contextual
1–6	1,2	2b,3a	Stop	id, name, coordinates, arrival_time, departure_time	Core
1–4	1,2,4,5	1a,2,3,4,7,8	Weekly_Schedule	id, monday, tuesday, wednesday, thursday, friday, saturday, sunday	Core
5,6	1,3,4	1b,5,6	Schedule_Exception	id, date, exception_type, affected_routes	Core
5,6	1,3,4	1b,5,6	Event	id, type, date	Core
1–6	1–5	1–8	Health_Facility	id, name, coordinates, type, access_schedule	Core
1,5,6	1–7	1,2,3,5c,6	Appointment	id, user, when, where	Contextual

Table 1: Purpose Formalization sheet

2.3.4 Activity 4: ER modeling

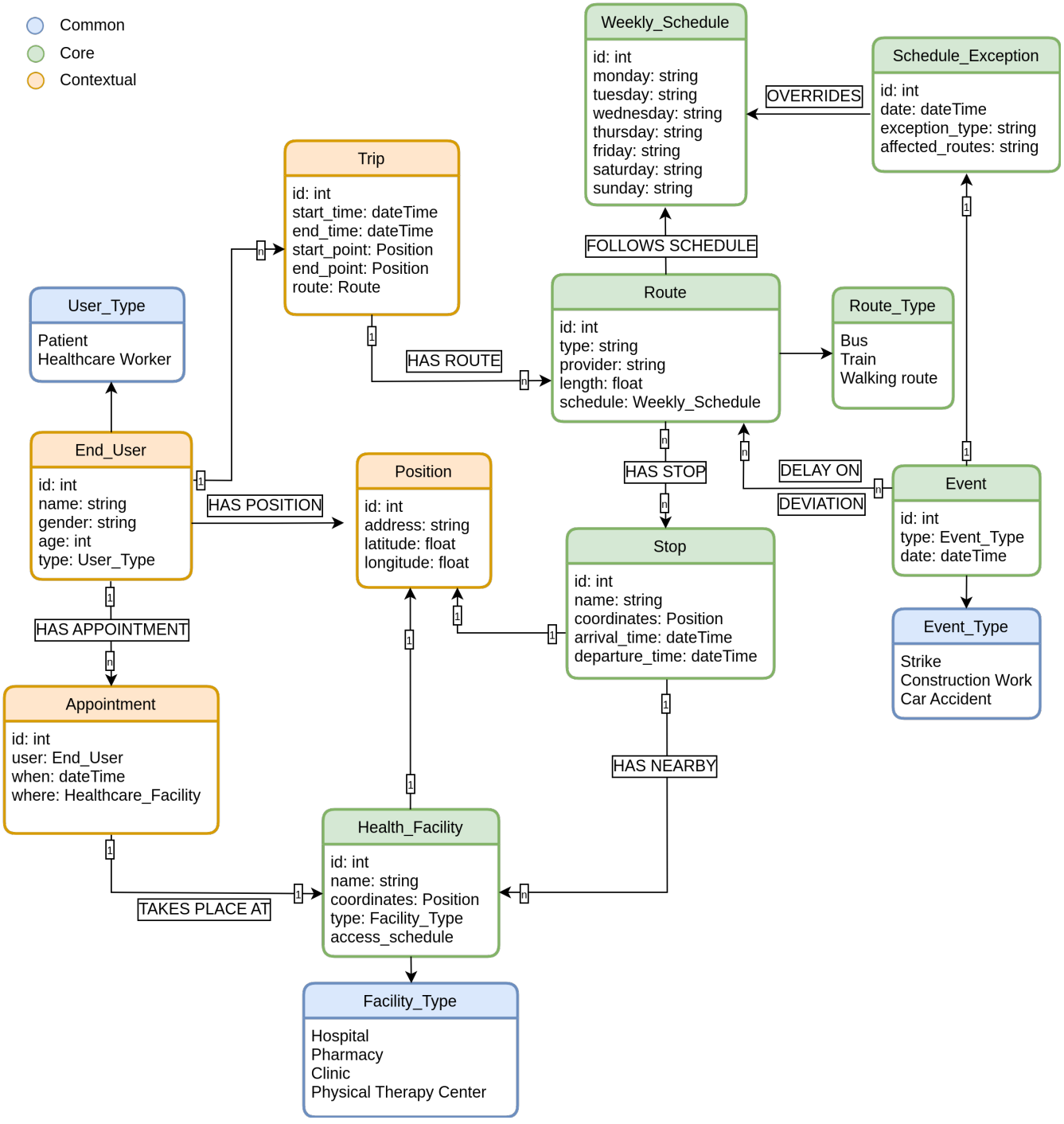


Figure 1: ER model

Each EType in the ER model has specific attributes that provide information relevant to answering CQs and fulfilling the KG's purpose.

The **End_User** EType captures individual user data and categorizes users into distinct **User_Type** groups—patients and healthcare professionals. **End_Users** are linked to the **Appointment** EType, which is key to aligning travel routes with users' healthcare schedules, supporting the KG's goal of matching users with timely travel options to and from their appointments.
Transportation-related ETypes—**Trip**, **Route**, and **Stop**—model the transportation network. A **Trip** defines a single journey taken by an **End_User** from a start point to a destination, and is served by one or more **Routes** that define the specific transit paths and schedules. This design supports queries like finding the nearest healthcare facility within a set travel time. The **Stop** EType defines precise locations along a route where users can board or exit. Each **Stop** has coordinates, scheduled arrival and departure times, and supports queries about where users can access transport, travel times, and nearby healthcare facilities.
The **Weekly_Schedule** EType defines the recurring weekly timetable for transportation services. By associating each route with its **Weekly_Schedule**, the KG can provide users with detailed timing and frequency information. Real-world schedules are subject to disruptions, and **Schedule_Exception** captures temporary changes to a regular schedule, such as delays or cancellations, caused by external occurrences (**Event** EType) such as traffic accidents, road closures, strikes. This structure allows the model to answer queries about potential disruptions.
Finally, healthcare service locations are managed through the **Health_Facility** EType, which catalogs healthcare access points—hospitals, clinics, and pharmacies—with data properties (attributes) for coordinates, facility type, and open hours.

Object properties (relationships) between ETypes have also been established. Key relationships include **HAS ROUTE** (linking a **Trip** to its **Route/s**), **FOLLOWS SCHEDULE** (associating **Route** with a **Weekly_Schedule**), and **HAS STOP** (connecting a **Route** to its **Stop** locations). Each relationship has defined cardinalities to specify how many instances of the two entities are involved. For instance, the **HAS STOP** relationship between **Route** and **Stop** typically has a many-to-many cardinality, as each route has multiple stops, and each stop may serve multiple routes. The **HAS APPOINTMENT** relationship, instead, exhibits a one-to-many cardinality, whereby an **End_User** can have multiple appointments, while each **Appointment** is associated with a single user. Defining these cardinalities ensures that the model accurately represents the complexity of real-world entities and can effectively processes user queries.

3 Information Gathering

3.1 Data Sources and Resources Collection

This section contextualizes each data source's role in the project.

The primary source of public transportation data for this project was the Trentino Transporti website. Through its Open Data section, Trentino Trasporti provides curated datasets covering both urban and extra-urban transportation following the General Transit Feed Specification (GTFS) standards. This adherence to GTFS ensures that data on bus stops, routes, and timetables is consistent and well-structured, supporting easy processing and meeting the iTelos methodology's requirements for interoperability. The latest update to these datasets was on September 09, 2024, covering the period from September 09, 2024 to June 12, 2025. This ensures that the KG is built upon the most current information available.
Public transportation datasets collected by project teams from previous academic years and stored within the LiveData catalogs were carefully reviewed to assess their relevance and potential for reuse. However, data was limited to the years 2022-2023, and could not fully meet the current project's needs for up-to-date transit information.

Data on Trentino region's pharmacies and para-pharmacies was initially sought from the 2022 Trentino Healthcare Connectivity Project. The datasets, sourced from OPENdata Trentino (dataset identifiers: *apss_farmacie-pat* and *apss_parafarmacie-pat*), covered the periods from January 1, 2021, to December 31, 2021, and from January 1, 2017, to December 31, 2017, respectively. Given the outdated nature of this data, it was deemed unsuitable for the current project's objectives. OPENdata Trentino references the Ministero della Salute's Open Data system as the original source of these datasets. Therefore, an attempt was made to access the Ministero della Salute's Open Data directly in the hope to acquire more up-to-date information. A couple of datasets were identified that met the project's requirements:

- FRM_FARMA provides a list of pharmacies across the entire national territory, covering branches, dispensaries, and seasonal dispensaries. Specifically, it provides detailed administrative data for each pharmacy (e.g. Ministerial Identification Code, name, VAT number, local health authority (ASL) pharmacy code) and precise location details, including the full address (street, postal code, district, municipality, province, and region). For further details, a dictionary of the dataset's fields is available for reference. The dataset is provided in JSON format and is updated weekly. At the time of retrieval for this project, the latest update was on November 6, 2024.
- FRM_PFARMA provides a complete list of businesses, other than pharmacies, authorized to sell medications to the public, along with administrative and location data. For more information about the dataset, a dictionary is available for reference. The dataset is provided in JSON format and is updated daily. At the time of retrieval for this project, the latest update was on November 6, 2024.

The FRM_FARMA and FRM_PFARMA datasets were both deemed suitable for the project, as they provide not only precise addresses and geographical coordinates (latitude and longitude), but also include the key fields "data_inizio_validità" (start date of validity) and "data_fine_validità" (end date of validity). These fields enable filtering to include only currently active businesses, ensuring that the KG can support users in identifying nearby

pharmacies and para-pharmacies based on up-to-date operational status.

OPENdata Trentino provides a dataset of hospitals within the municipality of Trento, available in both CSV and GeoJSON formats. The dataset was last updated on October 18, 2023, and contains precise location coordinates, institution names, facility types, and full addresses. Given that hospitals are typically stable institutions with minimal turnover, it is reasonable to assume that the listed facilities remain operational and relevant for the current project's objectives.

The 2022 Trentino Healthcare Connectivity Project has collected and processed a dataset on Trentino's public and accredited healthcare facilities, sourced from OPENdata Trentino under the identifier *apss_strutture-sanitarie-dell-azienda-sanitaria-e-convenzionate*. The dataset covers a timeframe from December 1, 2017, to December 31, 2017. Given the age of this data, its relevance and accuracy for the current project requires careful consideration. Although healthcare facilities tend to have low turnover and often remain operational over long periods, significant updates or structural changes may have occurred since 2017. Therefore, while this dataset remains available for reference, efforts were made to identify up-to-date resources.
An alternative, more suitable source was identified in the *Aziende Ospedaliere, Aziende Ospedaliere Universitarie e IRCCS pubblici* dataset available through the Ministero della Salute's Open Data system. This dataset provides a list of healthcare facilities—including hospital trusts, university hospital trusts, and public IRCCS (Istituti di Ricovero e Cura a Carattere Scientifico)—actively operating across Italy as of January 1, 2024. For detailed information on the dataset, please refer to the accompanying data dictionary.

Data on healthcare facilities were ultimately sourced from the Humanitarian Data Exchange (HDX) platform, specifically from the Italy Healthsites dataset within the Geodata Datasets section. This dataset is part of the Healthsites.io initiative, which aims to build an open data commons of health facility data with OpenStreetMap. The dataset is available in GeoJSON format and includes records on hospitals, clinics, pharmacies, medical offices, and dental practices throughout Italy. It provides key attributes such as facility type, location, and opening hours. Designed to support humanitarian efforts and improve accessibility to healthcare infrastructure data, this resource is regularly updated and validated through contributions from local and international health organizations. At the time of retrieval for this project, the latest update was on February 8, 2024.

GeoJSON data was download from the openpolis/geojson-italy GitHub repository which contains geo-referenced limits for all municipalities in Italy. Files are upgraded periodically, and refer to the latest administrative subdivisions, as published by ISTAT. Each municipality's boundary is represented as a polygon or multipolygon, defining the precise shape and size of its geographic area.

3.2 Resource Cleaning and Formatting

Once the relevant datasets were identified, the next step was to ensure their suitability for the construction of the KG by cleaning and formatting the data. This process involved several key steps, such as removing duplicate records, outdated data points or irrelevant attributes for the purpose, standardizing field names and values, and ensuring consistent formatting across all datasets. Data anonymization was not required, as all the data used was sourced from publicly accessible platforms, and it is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).
Each dataset was prepared to address the CQs defined during Purpose Formalization, ensuring they serve the information needs of scenarios. For reusability and compliance with iTelos standards, all resources have been formatted according to the CSV standard.
In accordance with the FAIR principles, access to both the original and processed datasets, as well as the Python scripts used for data processing, is provided through the project GitHub repository.

- ***Public Transportation Data:*** Urban and extra-urban transportation data from Trentino Trasporti Open Data was well-organized and fully compliant with the General Transit Feed Specification (GTFS) standard. The data was already aligned with best practices for transit data interoperability. As a result, no preprocessing or data manipulation was necessary. This RMarkdown notebook offers an overview of the files included in the dataset, providing insight into their structure and content.
  Train service data was unstructured (i.e. timetables in PDF format). Manual extraction was necessary to convert the data into a usable and machine-readable format for analysis. This process involved reviewing the timetables, identifying relevant information such as train schedules and station stops, and manually entering the data into a structured format suitable for further processing. Google Maps service was used to geolocate the train lines of interest. Due to the time-intensive nature of the manual extraction process, the final train service dataset has not yet been completed as of the writing of this report. Upon completion, it will be made available in the GitHub repository, along with the accompanying processing code.

- ***FRM_FARMA and FRM_PFARMA:*** The data cleaning process involved filtering the records to retain only those related to the 'PROV. AUTON. TRENTO' region. After isolating these entries, irrelevant fields such as 'cod_farmacia_asl' and 'p_iva' were removed to streamline the dataset. Address information was standardized by splitting the 'indirizzo' field at the first comma, separating the street name and house number and assigning each to distinct fields. If no house number was provided, 'None' was used to indicate the absence of data. The 'indirizzo' field displayed inconsistent casing, with some entries in uppercase and others in lowercase. For consistency and ease of use, this field was standardized to title case. Coordinates were standardized by replacing commas with periods.
  Each pharmacy and para-pharmacy in the dataset is uniquely identified by an ID, yet it is common to find multiple records with the same ID, sometimes repeated dozens of times. These duplicates only vary in their validity period, defined by the 'data_inizio_validità' and 'data_fine_validità' fields. It appears that, rather than updating an existing record when a pharmacy renews its validity period, a new record is added to the dataset. Outdated entries were filtered out, leaving a single valid record for each ID.
  'None' was used for missing data in the dataset.

- ***Aziende Ospedaliere, Aziende Ospedaliere Universitarie e IRCCS pubblici:*** Upon processing the dataset, it became clear that no entries corresponded to the 'Prov. Auton. Trento' region. The dataset dictionary does include a region code (i.e., 042) specifically assigned to the Province of Trento, suggesting that data for this region was intended to be included but ultimately missing. The complete absence of relevant records for the targeted region makes the dataset unusable for the project's objectives.

- ***Italy Healthsites:*** To narrow down the GeoJSON data to healthcare facilities specifically located within Trentino, a spatial filter was applied. The filter relied on a separate GeoJSON boundary file of Trentino-Alto Adige, and retained only the entries that fell within the specified polygonal geographic boundaries. Before applying the filter, both the healthcare facilities data and boundary geometry were aligned to the same coordinate reference system (CRS). Further filtering was applied to isolate municipalities within the 'Provincia Autonoma di Trento'. A list of Trentino municipalities, sourced from Wikipedia, guided this process, allowing the dataset to retain only the relevant municipalities for targeted analysis. To streamline the dataset, several non-essential fields were dropped.

  The dataset includes records for hospitals, pharmacies, clinics, and dental offices. Many pharmacy entries overlap with those found in the Ministero della Salute datasets, FRM_FARMA and FRM_PFARMA. Since these datasets are updated weekly and feature a 'data_fine_validità' field, they are reliable for maintaining an up-to-date list of active businesses, making them preferable as primary sources. However, the geographic coordinates (latitude and longitude) in FRM_FARMA and FRM_PFARMA are sometimes imprecise.
  For instance, in the FRM_FARMA dataset, the coordinates for *Farmacia Grandi* at *Via Alessandro Manzoni 7A, Trento*, are given as 46.0748132409383, 11.1251346567747 (marked by the red pin in Figure 2). Meanwhile, the coordinates from the Healthsites dataset are 46.073819833990825, 11.125191578971851 (marked by the green pin). As shown in Figure 2, the green pin closely matches the pharmacy's actual location, whereas the red pin is a relatively rough approximation. This pattern holds for many pharmacies in the dataset, with similar discrepancies observed across many entries. Given the recurring issue, it was decided to prioritize the Healthsites dataset's location data for spatial accuracy, while still relying on FRM_FARMA and FRM_PFARMA for information on the active status and operational details of each pharmacy. This combined approach aimed to combine the precise geographical positioning from Healthsites with the regularly updated and reliable business data from the Ministero della Salute sources, ultimately improving the dataset's overall quality for downstream applications.

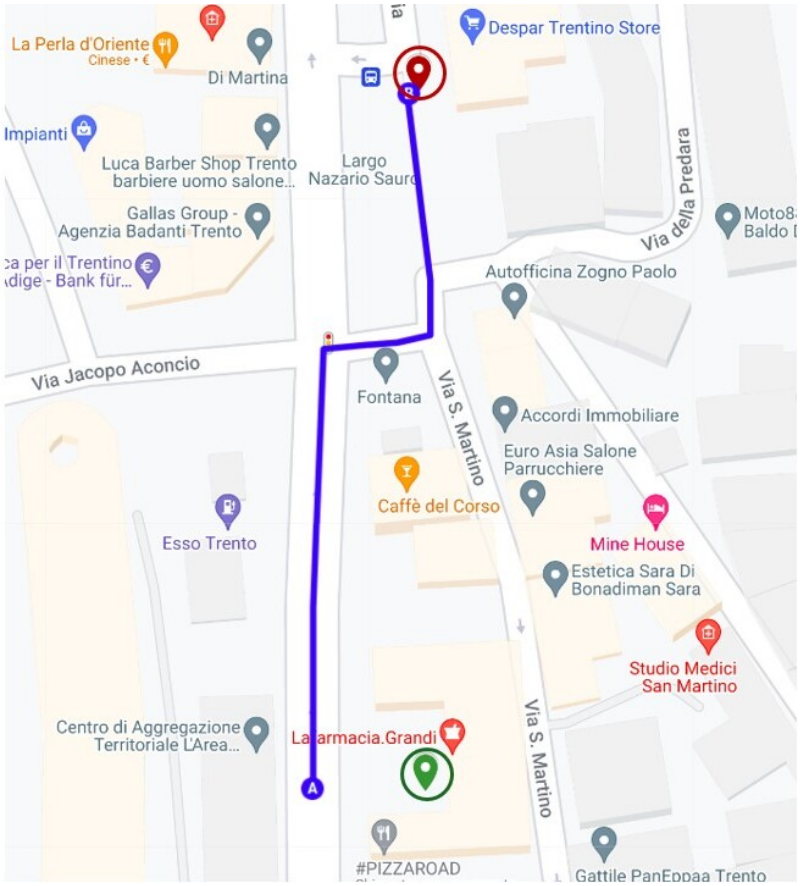


Figure 2: Discrepancies in spatial coordinates between Healthsites and Ministero della Salute datasets

The challenge, however, lies in integrating the two sources. Matching by pharmacy name proves unreliable due to inconsistencies: in the Ministero della Salute datasets, the pharmacy name is theoretically derivable from the 'descrizione_farmacia' field, intended to include the *ragione sociale* according to the data dictionary. However, in practice, this field varies widely—sometimes it contains only the pharmacy name, sometimes only the owner's name, and occasionally both. Merging by address is also impractical, as the Healthsites dataset often lacks address data, and address formats differ across the datasets (e.g., *Piazza Dante Alighieri* and *Piazza D. Alighieri*).

To address the inconsistencies, a geospatial matching approach was implemented, based on proximity rather than exact matches on name or address. Coordinate columns or WKT geometry were transformed into spatial point data (any non-Point geometries like polygons were converted to centroids), and a 50-meter buffer zone was defined around each point in the Healthsites dataset. Any point from the Ministero della Salute datasets falling within the search radius was then treated as a potential match, suggesting likely correspondence to the same pharmacy. A KD-Tree Nearest Neighbor Search was used to find points in close proximity (*cKDTree* from *scipy.spatial*).

Following the initial identification of potential matches based on spatial proximity, additional criteria were applied to refine these matches. When pharmacy names were available in both datasets, a fuzzy matching algorithm (using the *FuzzyWuzzy* Python library) assessed name similarity. Where address information was present, fuzzy matching was also applied to the address fields. Finally, a combined metric incorpo-

rating spatial distance, name similarity, and address similarity, was used to evaluate the quality of each match.

During processing, it became clear that the spatial coordinates in the Ministero della Salute datasets were more imprecise than anticipated. As a result, the buffer radius had to be increased from the initial 50 meters to 10 kilometers to ensure potential matches were captured. Ultimately, the maximum distance observed between two pharmacies identified as a match reached 6.54 km! (Figure 3).

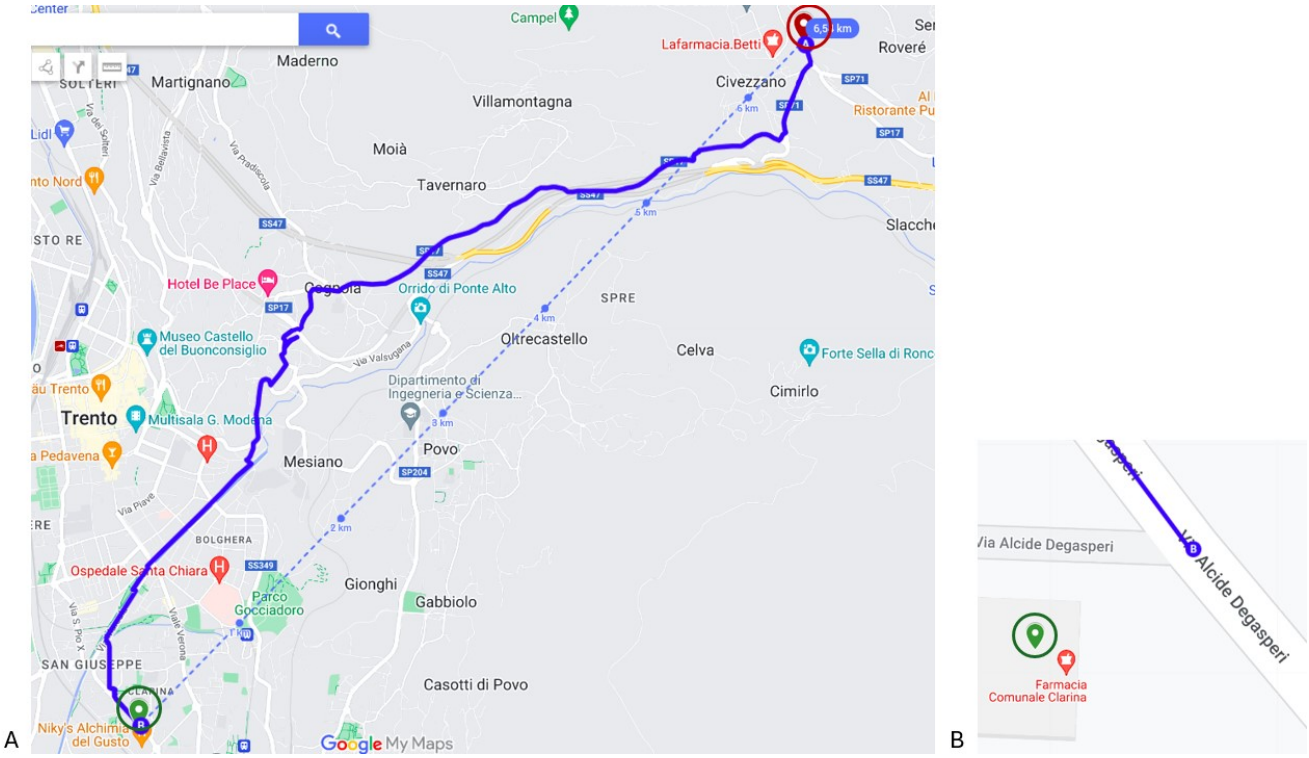


Figure 3: A) The red pin marks the Ministero della Salute coordinates for ”Comunale N.4 Clarina” at Via De Gasperi 112, 38123, Trento, while the green pin marks the Healthsites coordinates for ”Farmacia Clarina” at Via Alcide De Gasperi 112, 38123, Trento. Despite the differences in name and address, there is high confidence that these entries refer to the same pharmacy. B) The green pin aligns with the pharmacy’s location on Google Maps, whereas the red pin, positioned 6.54 km away in a different postal code (38045), reflects the high imprecision in the Ministero della Salute dataset.

Out of 94 Healthsites records that included a full address, 74 matched entries in the Ministero della Salute datasets. All matches were manually reviewed.
In conclusion, the final integrated dataset comprises entries from FRM_FARMA and FRM_PFARMA, with Healthsites’ more accurate spatial data replacing the original coordinates wherever possible. Additionally, opening hours from the Healthsites dataset were included in the entries when available.

Following initial processing, records for hospitals, clinics, and dental offices from the Healthsites dataset required no further refinement.

4 Language Definition

The Language Definition phase of the iTelos methodology addresses the challenges of language heterogeneity by establishing the formal concepts required to represent the information in the Knowledge Graph. This phase is pivotal in creating a consistent and unambiguous domain language tailored to the KG's purpose, facilitating integration of heterogeneous data sources and supporting robust query capabilities.
The Language Definition phase comprised two key activities:

- Concept Identification
- Dataset Filtering

4.1 Activity 1: Concept Identification

This activity involved identifying all relevant concepts to be used to represent the information in the final KG. These concepts included ETypes, data properties, and object properties, which had been outlined during the Purpose Formalization phase and documented in the Purpose Formalization Sheet and ER model.
The identified concepts were then aligned with the Universal Knowledge Core (UKC). If a concept already existed in the UKC, it was directly adopted for its broad reusability and established formalization. If the concept was absent from the UKC, it was searched for in other ontologies or vocabularies (e.g., Schema.org, General Transit Feed Specification). When existing concepts were either too general or absent, new concepts were defined to address the project's specific context. For instance, *Weekly Schedule* and *Schedule Exception* were introduced to address the periodicity of public transit schedules and deviations from them, respectively, filling gaps not covered by existing definitions in the UKC.
The formalized concept definitions resulting from the UKC alignment process are presented in Tables 2, 3, and 4, with each entry uniquely identified and accompanied by a precise label and gloss.

Note to the Reader: In Table 3, references to specific concepts from the General Transit Feed Specification (GTFS) are provided in the *Language_Resources.html* file available on the Project's GitHub repository. In this document, only a general link to the GTFS Reference is provided, as LaTeX does not support URLs with text highlights.

Language concepts for ETypes

Concept ID	Language Concept Word	Gloss
UKC-49456	end user	The ultimate user for which something is intended.
UKC-20121	health facility	Building where medicine is practiced.
KGE2024-XXXXX*	health appointment	A scheduled meeting with a healthcare provider for consultation, treatment, or check-up.
UKC-27518	position	The spatial property of a place where or way in which something is situated.
UKC-45117	bus route	The route regularly followed by a passenger bus.
KGE2024-XXXXX*	train route	The route regularly followed by a passenger train.
UKC-1484	trip	A journey for some purpose (usually including the return).
UKC-45118	bus stop	A place on a bus route where buses stop to discharge and take on passengers.
KGE2024-XXXXX*	train stop	A place on a train route where trains stop to discharge and take on passengers.
KGE2024-XXXXX*	weekly schedule	An ordered timetable outlining activities or operations for each day of the week.
KGE2024-XXXXX*	schedule exception	An instance that does not conform to a schedule.

Table 2: In this table is presented how ETypes have been formalized and assigned to a UKC concept. Concept IDs marked with a ” * ” are new concepts specifically created for this project.

Language concepts for ETypes attributes (Data Properties)

Data Property Name	Concept ID	Language Concept Word	Gloss
End User			
has_user_id	UKC-38423	identifier	A symbol that establishes the identity of the one bearing it.
has_user_first_name	UKC-33531	first name, given name, forename	The name that precedes the surname.
has_user_surname	UKC-33528	surname, family name, cognomen, last name	The name used to identify the members of a family (as distinguished from each member's given name).
has_user_type	UKC-31362	type	A subdivision of a particular kind of thing.
Health Facility			
has_health_facility_legal_name	Schema.org	legalName	The official name of the organization, e.g. the registered company name.
has_health_facility_type	UKC-31362	type	A subdivision of a particular kind of thing.
has_health_facility_opening_hours	Schema.org	openingHours	The general opening hours for a business. Opening hours can be specified as a weekly time range, starting with days, then times per day.
Health Appointment			
has_health_appointment_user	UKC-53492	user	A person who makes use of a thing; someone who uses or employs something.
has_health_appointment_place	UKC-66454	place	Proper or appropriate position or location.
has_health_appointment_date	UKC-73013	date	The specified day of the month.
has_health_appointment_time	UKC-38632	time	An instance or single occasion for some event.
Position			
has_position_addr_street	UKC-45008	street address	The address where a person or organization can be found.
has_position_addr_housenumber	KGE2024-XXXXX*	house number	A number identifying a building on a street, used in postal addresses.
has_position_addr_postcode	UKC-33635	zip code, zip, postcode, postal code	A code of letters and digits added to a postal address to aid in the sorting of mail.
has_position_municipality	UKC-45537	municipality	An urban district having corporate status and powers of self-government.
has_position_latitude	UKC-45423	latitude	The angular distance between an imaginary line around a heavenly body parallel to its equator and the equator itself.
has_position_longitude	UKC-45429	longitude	The angular distance between a point on any meridian and the prime meridian at Greenwich.
Route			
has_route_id	General Transit Feed Specification Reference	route_id	Identifies a route.
has_route_type	General Transit Feed Specification Reference	route_type	Indicates the type of transportation used on a route. Valid options are: 0 - Tram, 1 - Subway, etc.

has_route_short_name	General Transit Feed Specification Reference	route_short_name	Short name of a route. Often a short, abstract identifier.
has_route_long_name	General Transit Feed Specification Reference	route_long_name	Full name of a route. This name is generally more descriptive than the route_short_name.
Trip			
has_trip_id	General Transit Feed Specification Reference	trip_id	Identifies a trip.
has_trip_headsign	General Transit Feed Specification Reference	trip_headsign	Text that appears on signage identifying the trip's destination to riders.
Trip Stop			
has_trip_stop_arrival_time	UKC-73119	arrival_time	The time at which a public conveyance is scheduled to arrive at a given destination.
has_trip_stop_departure_time	UKC-73120	departure_time	The time at which a public conveyance is scheduled to depart from a given point of origin.
Stop			
has_stop_id	General Transit Feed Specification Reference	stop_id	Identifies the serviced stop. A stop may be serviced multiple times in the same trip, and multiple trips and routes may service the same stop.
has_stop_name	General Transit Feed Specification Reference	stop_name	Name of the location. The stop_name should match the agency's rider-facing name for the location as printed on a timetable, published online, or represented on signage.
has_stop_lat	General Transit Feed Specification Reference	stop_lat	Latitude of the location.
has_stop_lon	General Transit Feed Specification Reference	stop_lon	Longitude of the location.
Weekly Schedule			
has_weekly_schedule_start_date	General Transit Feed Specification Reference	start_date	Start service day for the service interval.
has_weekly_schedule_end_date	General Transit Feed Specification Reference	end_date	End service day for the service interval. This service day is included in the interval.
has_weekly_schedule	UKC-34204	schedule	An ordered list of times at which things are planned to occur.
Schedule Exception			
has_schedule_exception_type	UKC-31362	type	A subdivision of a particular kind of thing.
has_schedule_exception_date	UKC-73013	date	The specified day of the month.

Table 3: In this table is presented how ETypes properties (first column) have been formalized and assigned to a UKC concept (second and third columns). The last column provides a description of what the concept means. Concept IDs marked with a " * " are new concepts specifically created for this project.

Language concepts for ETypes relations (Object Properties)

Object Property	Relationship	Language Concept Word	Concept ID	Gloss
has_appointment	End User - Appointment	scheduled	UKC-87477	Planned or scheduled for some certain time or times.
has_position	End User - Position; Health Facility - Position; Stop - Position	localized	UKC-81950	Confined or restricted to a particular location.
plans	End User - Trip	plan	UKC-96049	Make plans for something.
takes_place_at	Appointment - Health Facility	happen, hap, go on, pass off, occur, pass, fall out, come about, take place	UKC-94276	Come to pass.
has_route	Trip - Route	follow, travel along	UKC-102467	Travel along a certain course.
has_stop	Route - Stop	include	UKC-105576	Have as a part, be made up out of.
has_nearby	Stop - Health Facility	nearest, nighest, closest	UKC-109080	(superlative of 'near' or 'close') within the shortest distance.
follows_schedule	Route - Weekly Schedule	adopt, follow, espouse	UKC-104222	Choose and follow; as of theories, ideas, policies, strategies or plans.
overrides	Schedule Exception - Weekly Schedule	override	UKC-1088	The act of nullifying; making null and void; counteracting or overriding the effect or force of something.
causes	Event - Schedule Exception	cause, do, make	UKC-100755	Give rise to; cause to happen or occur, not always intentionally.
delays	Event - Route	delay, detain, hold up	UKC-94853	Cause to be slowed down or delayed.
deviates	Event - Route	deviate	UKC-102843	Cause to turn away from a previous or expected course.

Table 4: In this table is presented how "generic" words (first column) regarding ETypes relationships have been formalized and assigned to a UKC concept (third and fourth columns). The last column provides a description of what the concept means.

4.2 Activity 2: Dataset Filtering

In Activity 2, the focus shifted to aligning the collected data resources with the formalized language concepts identified in Activity 1. The objective was to remove any irrelevant, redundant, or misaligned data points that did not meet the established definitions and ensure that all data entries conformed to the newly defined domain language. Any data elements that were not explicitly defined in the language resources were excluded. This step also involved renaming data columns to better align with their corresponding ETypes, improving both clarity and consistency in how the data was represented. This renaming process helped ensure that the data was more intuitively organized and aligned with the formalized domain language, making it easier to integrate and query.