UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

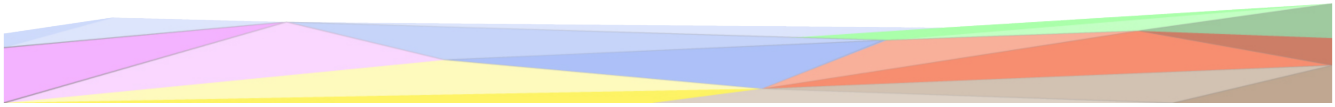
– KnowDive Group –

KGE 2024 - Trentino Territory & Transportation

Document Data:	Reference Persons:
December 4, 2024	Mores Nicola, Roccon Marco

© 2024 University of Trento
Trento, Italy

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



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 21, 2024	Mores Nicola, Roccon Marco	Document created
0.2	October 30, 2024	Mores Nicola, Roccon Marco	Phase 1 - Purpose Definition
0.3	November 13, 2024	Mores Nicola, Roccon Marco	Phase 2 - Information Gathering
0.4	November 26, 2024	Mores Nicola, Roccon Marco	Phase 3 - Language Definition
0.5	December 04, 2024	Mores Nicola, Roccon Marco	Phase 4 - Knowledge Definition

1 Introduction

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important role to enhance the reusability of the resources handled and produced during the process. A clear description of the resources as well as of the process (and single activities) developed, provides a clear understanding of the project, thus serving such an information to external readers for the future exploitation of the project's outcomes.

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: High level description of the project development, based on the Produce role's objectives.
- 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

In this section we will cover the first phase defined by the iTelos methodology: The Purpose Definition. In this phase we aim to concretely define in a formal way the user's Purpose and what will be the information requirements that our Entity Graph will be able to satisfy. In order to do so, we will start from an Informal purpose, defining our Domain of Interest, and proceed with the creation of Personas, Scenarios. Using these we will define a set of Competency Questions (CQs), later used to identify the concepts (entities and properties) that we will work on and that are used to create an ER Model, the first purpose-specific version of the knowledge layer. Thus, at the end of this first step we will have a set of CQs, a set of identified concepts and an ER model that, all together, define our formal Purpose.

2.1 Informal Purpose

The first step to create an Entity Graph is the definition of a starting informal Purpose, stating, through a natural language sentence, the objective that drives us to the usage of the iTelos methodology.
In our case we want to create an Entity Graph containing information about transportation in the Trentino Region, focussed mainly on the city of Trento. In particular we want to extract not only data about Busses and Trains, but also regarding shared mobility alternatives (bike, scooter and car sharing), taxis, parking facilities and bike racks. In a more concise way, the informal purpose is:

> ”A person wants to move in an easy and efficient way through the Trentino region using public transports and other transport services available”

2.2 Domain of Interest

Having finalized our starting Purpose, we can also define the domain of interest in which our project will work and reason.
Our domain of interest will be the one of transportation services and, as stated in the informal purpose, from a spacial point of view, we will focus on the Italian region of Trentino, with special focus on it’s capital city: Trento.
Having delineated a first constraint on the space, we can also define one for the timespan we will consider: the project will focus on the currently available data about Trentino’s public transportation services, that covers a period of time around 10 months (from September 2024 to the end of June 2025).

2.3 Scenarios definition

In this section we define the set of Scenarios that will be taken into account during the project, showing the context in which our final users will act.
Every Scenario has been described in terms of a general description of the context and some possible needs that it may give rise to.

1. **Weekday:**
    - **Description:** It’s a weekday in Trento, with residents primarily commuting for work or study purposes. Buses and trains follow regular schedules, with commuters checking schedules to plan their movements.
    - **Needs emerged:**

  - Access to updated public transport schedules.
  - Travel planning to avoid peak hours.

2. **Weekend Excursion:**
   - **Description:** It's a weekend, and many residents take advantage of their free time to go on bike excursions around Trento. Public transport offers special services to carry bicycles.
   - **Needs emerged:**
     - Information on cycling racks available in Trento.
     - Schedules and regulations for bike transport on public transit.
3. **Holiday (Christmas):**
   - **Description:** During the Christmas season, celebrations lead to changes in public transport schedules. People use buses and trains to visit family and friends or participate in festive events in the city.
   - **Needs emerged:**
     - Information on special holiday public transport schedules.
4. **Rainy Day:**
   - **Description:** On a rainy day in Trento, residents prefer using taxis or car-sharing services to avoid walking in the rain. The demand for private transport increases significantly.
   - **Needs emerged:**
     - Access to information on the availability of taxis and car-sharing services.
5. **Cultural Event in the City Center:**
   - **Description:** A cultural event, such as a fair or festival, is taking place in Trento, attracting a large crowd to the city center. People seek to reach the event quickly and conveniently, so many decide to use electric scooters or bike sharing to avoid traffic and find bike parking easily.
   - **Needs emerged:**
     - Information on available bike-sharing racks and stations near the event.
     - Details on designated areas for rental scooters.
6. **Start of the Academic Year:**

- **Description:** At the start of the new academic year, new university students move to Trento and search for apartments. In their search, they consider not only price and availability but also how well-connected the area is to university departments and daily activities such as supermarkets and gyms.
- **Needs emerged:**
  - Access to information on the proximity of public transport stops in a certain location.
  - Details on public transport routes between two areas in the city.

7. **Graduation Day:**

   - **Description:** It's graduation day at the University of Trento. Family and friends of graduates come to the city to attend the ceremony, causing a significant increase in traffic and high demand for parking. Drivers are looking for information on available parking near the university and alternative parking options.
   - **Needs emerged:**
     - Access to information on available parking near the university.
     - Directions on how to reach the ceremony location by public transport from identified parking areas.

2.4 Personas definition

In this subsection we will define a set of Personas: Fictional actors involved in the project domain, characterizing user's needs and perception, and that will act in the previously defined Scenarios.
Below we describe our Personas, stating for each of them a brief description of their lives and what their needs and goals are:

1. **Sara:**

   - **Occupation:** University student
   - **Age:** 23 years
   - **Description:** Sara lives downtown near the station and regularly attends classes in the Department of Economics. Without a car, she uses buses and trains for her travels, and occasionally bike sharing.
   - **Needs/Objective:**
     - Find fast and direct routes to the campus.
     - Check for any unavailability of transportation.

2. **Marco**:
    - **Occupation:** University student and amateur cyclist
    - **Age:** 22 years
    - **Description:** Marco studies law in the city center and prefers to cycle when the weather is nice. He lives just a few minutes from the department, but on rainy days he prefers public transport.
    - **Needs/Objective:**
        - Discover the locations of public bike racks.
        - Find public transport alternatives in case of rain.
3. **Luisa**:
    - **Occupation:** Commuter
    - **Age:** 32 years
    - **Description:** Luisa works in the center of Trento but lives in a nearby town. Luisa suffers from motion sickness whenever she has to work on the computer during her travels. For this reason, and to have more space, she prefers the train for her daily commutes.
    - **Needs/Objective:**
        - Check train schedules and verify availability, even on holidays.
        - Plan trips that minimize wait times between trains.
4. **Giovanni**:
    - **Occupation:** Business executive
    - **Age:** 48 years
    - **Description:** Giovanni has frequent appointments in various parts of the city. He needs to move quickly and efficiently, often working while on the go, and for this reason, he prefers taxis.
    - **Needs/Objective:**
        - Know the nearest taxi parking in useful areas.
5. **Andrea**:
    - **Occupation:** Out-of-town student
    - **Age:** 24 years

- **Description:** Andrea and his fellow out-of-town students regularly organize weekend trips. They do not own a car, so they rent car-sharing vehicles for longer trips.
- **Needs/Objective:**
  - Find designated car-sharing zones and check the availability of vehicles for day trips.

6. **Helmut**:
   - **Occupation:** Tourist
   - **Age:** 36 years
   - **Description:** Coming from a nearby country, Helmut arrives by bike and wants to explore the city center without bringing it with him.
   - **Needs/Objective:**
     - Identify public bike racks and secure bike parking.
     - Check if the trains connecting his country to the city center have appropriate racks for transporting bikes.
7. **Francesca**:
   - **Occupation:** Employee
   - **Age:** 56 years
   - **Description:** To attend her son's graduation in the city center, Francesca drives there, but since she is not from the area, she doesn't know the locations of nearby parking.
   - **Needs/Objective:**
     - Find public and paid parking available for extended stays.
     - Know if the parking areas are accessible and close to the ceremony location.
8. **Davide**:
   - **Occupation:** University student
   - **Age:** 21 years
   - **Description:** Davide uses electric scooters to get around in the evening when public transport is less frequent. He is also a frequent user of bike sharing, having recently lost his own bike.
   - **Needs/Objective:**
     - Know the location of scooter and bike-sharing racks.
     - Find quick and flexible transport solutions during the evening hours.

9. **Anna**:
    - **Occupation:** University student
    - **Age:** 19 years
    - **Description:** Anna is a student with reduced mobility who moves in a wheelchair. She lives in the suburbs and attends university in the city center, so she relies on public transport for her daily travels. She often needs to check bus arrival and departure times and ensure they are accessible.
    - **Needs/Objective:**
        - Verify the accessibility of buses on urban routes and if the service is active during a specific time frame.
        - Know in advance if a stop is accessible in a wheelchair and if there are detours or route changes that could affect her mobility.

2.5 Competency Questions (CQs)

Now that we have defined Personas and Scenarios we can proceed extracting the KG functional requirements, defining ours Competency Questions. Each of them will refer to one of the Personas acting in one of the Scenarios previously enumerated, and will be used to identify the questions that our EG, once completed, will be able to answer.
Below are the CQs identified, grouped by the Scenario they are referring to:

1. **Weekday:**
    1.1 **Sara:** Which buses and trains can I take to go from Trento Station to the Department of Economics between 8:00 and 9:00?
    1.2 **Marco:** What is the arrival time of the next bus from the ”Povo Valoni” stop heading downtown?
    1.3 **Anna:** Which stops on bus line 5 are wheelchair accessible?
2. **Weekend Excursion:**
    2.1 **Marco:** Given a point on a map with its coordinates, what is the nearest stop to it?
    2.2 **Helmut:** Which train routes allow bicycle transportation during the weekend?
    2.3 **Helmut:** How many bikes can be parked in the rack closest to my current location?
    2.4 **Giovanni:** How many parking spots are available in the car-sharing station near Piazza di Fiera?
3. **Holiday (Christmas):**

3.1 **Luisa:** How do holiday schedules for busses and trains change on Christmas Day?

3.2 **Sara:** If I get on bus line 5 at “Povo Salé” stop at 12:05, when can I expect to arrive at the “S.francesco Porta Nuova” stop?

3.3 **Anna:** Are there any routes on the “P.Dante Rosmini S.Rocco Povo Polo Soc.” line that go directly to “Povo Polo Sociale”?

4. **Rainy Day:**

4.1 **Andrea:** Which car-sharing stations are closest to the San Bartolomeo area?

4.2 **Giovanni:** Where can I catch a taxi near Piazza Duomo?

4.3 **Marco:** On line 7, how many stops are there from “Gocciadoro Arcate” to “Adamello Gorizia”?

4.4 **Francesca:** Having just washed my car, where can I find an underground parking garage to avoid the rain?

5. **Cultural Event in City Center:**

5.1 **Davide:** Where can I find electric scooter stations near Piazza Fiera?

5.2 **Davide:** How many rental bikes can be parked in the station near the city center?

5.3 **Marco:** Where can I find a bike rack with frame locks near Piazza Duomo?

5.4 **Anna:** Which runs of bus line 5 heading downtown are wheelchair accessible?

6. **Start of the Academic Year:**

6.1 **Anna:** Which bus and train stops are available within a 500-meter radius of my apartment in the Santa Chiara area?

6.2 **Andrea:** How many stops separate the area of Piazza Dante from the Department of Economics on public transport lines?

6.3 **Luisa:** Which organization manages public transportation services in the city of Trento, and how can I contact it?

7. **Graduation Day:**

7.1 **Francesca:** What is the average maximum capacity of public parking spaces within a 1 km radius of the Department of Medicine at the University of Trento?

7.2 **Francesca:** What are the opening hours of the “Piazza di Fiera” parking lot?

7.3 **Andrea:** How many free public parking spots are there in Povo?

2.6 Concepts Identification

Having defined the CQs, we can proceed with the following step: Concept Identification. During this step we will extract the concepts identifying Entity Types (ETypes) and properties that will be modelled in our KG. To do so we will take into account the previously defined purpose and also the data layer, in terms of data sources availability. The final result of this step will be a Purpose Formalization Sheet (PFsheet), a dedicated spreadsheet combining Knowledge and Data Layer.

The following table shows the PFsheet we can generate from the Personas and Scenarios described in the previous sections. Each row contains one Entity and its corresponding properties, stating from which Personas, Scenarios and CQs these concepts have been extracted from.

In order to enhance the reusability, flexibility and quality of our future EG we will also consider well-known schema providers, such as SCHEMA.org, trying to find a proper mapping between their resources and our concept vocabulary whenever possible.

Finally we classified each entity with respect to its Focus, a parameter used to represent how much a concept is relevant to one's purpose, more concretely it can assume one of these values: Common, universal and commonly used concepts; Core, essential concepts for our domain; Contextual, highly specific concepts of our context and thus, usually, less reusable.

Scenarios	Personas	Competency Questions	Entities	Properties	Focus
6	3	6.3	City		Common
6	3	6.3	Organization	name, telephone	Common
1, 2, 4, 5, 6, 7	1, 2, 4, 5, 6, 7, 8, 9	1.1, 2.1, 2.3, 2.4, 4.1, 4.2, 5.1, 5.2, 5.3, 6.1, 6.2, 7.1, 7.3	Point	latitude, longitude	Common
1, 2, 3, 4, 5, 6	1, 2, 3, 6, 9	1.1, 1.2, 1.3, 2.2 3.1, 3.2, 3.3, 4.3, 5.3, 5.4, 6.2	Line (Bus/Train)	shortName, longName, type	Contextual
2, 3, 5, 6	1, 2, 6, 9	1.1, 1.2, 2.2, 3.1, 3.3, 5.4, 6.2	Trip	direction, headsign, bikeSlots, wheelchair	Contextual
1, 3, 4, 5, 6	1, 2, 8, 9	1.1, 1.2, 1.3, 3.2, 3.3, 4.3, 5.2, 6.1, 6.2	Stop (Bus/Train)	name, wheelchair, pos	Contextual
1, 3, 4, 6	1, 2, 5	1.1, 1.2, 3.2, 4.3, 6.2	Stop Event	arrivalTime, departureTime, stopSequence	Contextual
2, 5	2, 6	2.3, 5.3	Bike Rack	pos, capacity, type	Core
1, 2	1, 2, 6	1.1, 2.2	Schedule	byDay, validity(start end date)	Contextual
3	3	3.1	Special Schedule	date, type(variazione del servizio)	Contextual
2, 4	4, 5	2.4, 4.1	Car Sharing Station	pos, capacity	Core
4	4	4.2	Taxi Station	pos	Core
5	8	5.1	Scooter Sharing Station	pos	Core
5	8	5.2	Bike Sharing Stations	pos, capacity	Core
4, 7	5, 7	4.4, 7.1, 7.2, 7.3	Parking Lot	name, capacity, pos, type, openingHours isAccessibleForFree	Core

Figure 1: Purpose Formalization sheet

2.7 ER model definition

The last step of this initial phase, that will lead us to a complete formalization of the purpose, is to design an Entity Relation (ER) model using the concepts previously obtained. This ER model will be our first version of the final knowledge layer.
In order to create the model we will use the IDEF1X Notation ERD, a notation that will allow us to define entities and their attributes more precisely, compared to the traditional ERD, thus obtaining a clearer representation. To illustrate more clearly the focus level previously assigned to each entity, in the following diagram we will also use various colours to highlight the different levels: Red for Common ETypes; Green for Core ETypes; Blue for Contextual Etypes.

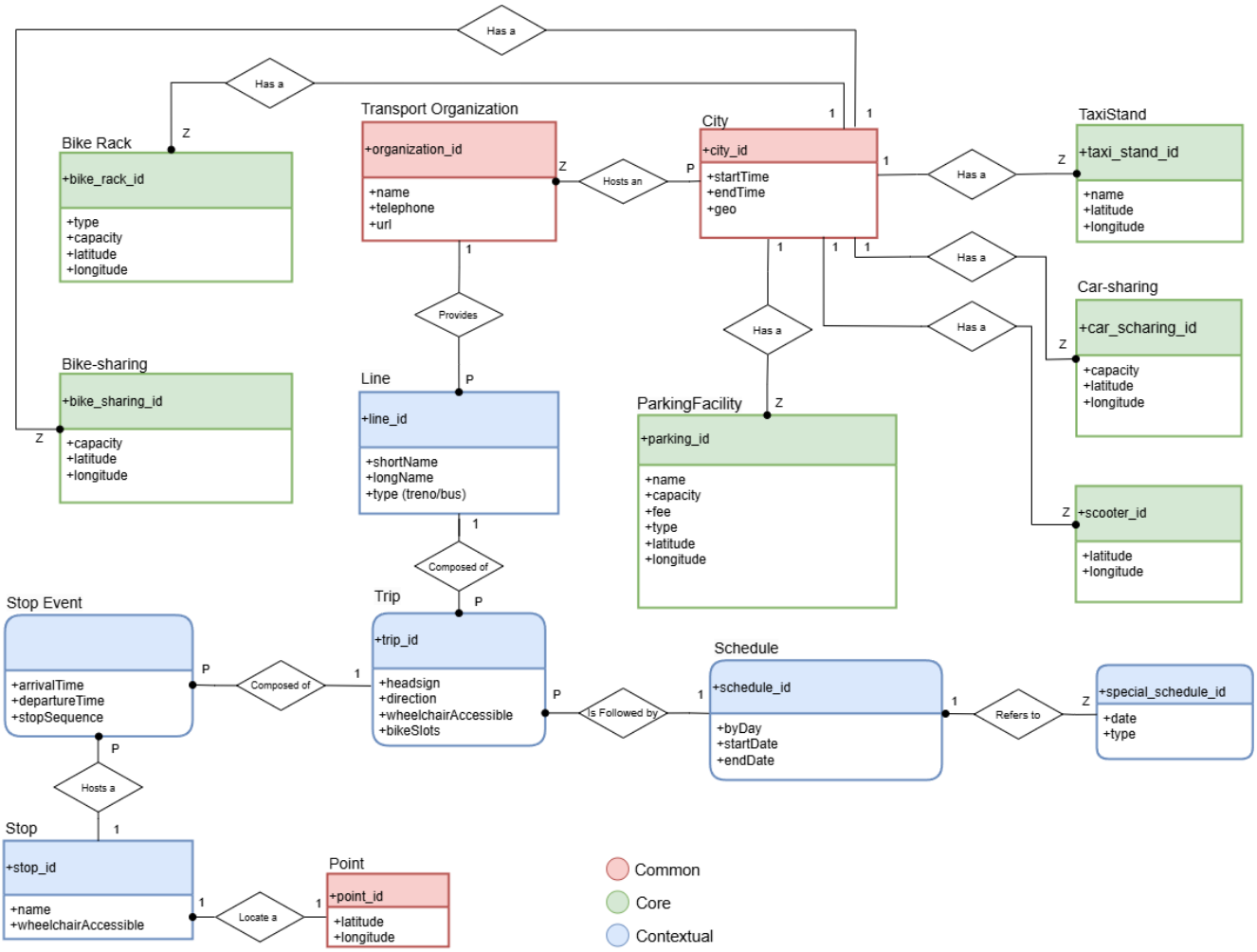


Figure 2: The ER model in IDEF1X Notation

3 Information Gathering

In this section we will cover the second phase of the iTelos methodology. Having formally defined the purpose of our project, we can focus on the collection on resources valuable for the creation of a EG able to satisfy the purpose. These resources includes dataset about every information's layers: Data, Knowledge and Language datasets. Moreover, during this chapter, we will also enhance their quality and reusability, by means of cleaning from unnecessary noise and standardizing them.

3.1 Source Identification

The first step of this phase covers the identification of the sources taken into account while gathering information. Among the many possible sources we can classify them in two main groups, based on their quality: High quality data sources contains distributed dataset characterized by high interoperability and reusability; On the other hand, Low quality sources' dataset are less standardized and with poor metadata, thus being less interoperable and understandable.

During our research we have founded these information resources:

- **Data value Datasets**
  - OPENdata Trentino: Open data Trentino is a platform, managed by the autonomous province of Trento, that offers an unique catalogue of reusable data and allows the search, access and download of open data about Trentino and services in its territory. We will use this platform to gather many of our datasets about the public transportation services and more. This source's dataset have really high variability in terms of quality: some of them are easily accessible (through APIs or direct download) and uses high quality, open license and non-proprietary data formats that support machine-readability, such as JSON and CSV; on the other hand many of their resources are occasionally accessible or not at all, while other are provided in formats such as PDF, making really hard to reuse them.
  - OpenStreetMap: OpenStreetMap is a collaborative project that provides free and editable maps of the world, also maintaining geographic data about roads, trails, railway stations, and so on. As an open-source and widely recognized resource, we will leverage OpenStreetMap to obtain additional information about parking lots, trainline and cycle path in the Trentino region, enriching our project and allowing us to cover also these part of the previously defined domain.
  - Trentino Trasporti Website: Trentino Trasporti is the main public transportation service provider in the city of Trento. In their website are provided information about the train

and bus lines available.

- **Knowledge Datasets:**
  - SCHEMA.org: ”Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond”. Being this a well-known and standardized resource, we will extract some reference schemas from it, in order to help us during the modelling of the needed ETypes. Proceeding in this way we can make our project more understandable and easily interoperable.
  - GTFS.org: ”The General Transit Feed Specification, also known as GTFS, is a standardized data format that provides a structure for public transit agencies to describe the details of their services such as schedules, stops, fares, etc”. As a widely adopted standard for transit data, GTFS will be integral in defining and modelling transportation-related entities, routes, stops, schedules, and calendars within our project.
- **Language Datasets**
  - Universal Knowledge Core (UKC): ”The Universal Knowledge Core (UKC) is a multilingual, high quality, large-scale, machine-readable, and diversity-aware lexical resource”. This will be our main resource at the language level, providing a formalization of many of the purpose-specific concepts founded during the project.
  - SCHEMA.org: In addition to the previously cited schemas and ontologies, Schema.org provides also informations about the language layer. As a matter of facts, each schema is provided also with a brief explanation of the meaning of the concepts introduced by itself, together with a description of each of its properties.
  - OpenStreetMap Wiki: OpenStreetMap, through their wiki page, provides also high quality dataset at language level, offering many formalized definitions of concepts about transportation services and streets.

3.2 Dataset Collection

In this subsection we will list the various dataset extracted from the previously cited data sources, showing for each of them their content and giving a brief note about their quality whenever needed.

- **OPENdata Trentino:**
  - Trasporti pubblici del Trentino: This dataset contains data about Trentino Trasporti’s urban and suburban lines. The data provided for the two types of lines follows

the GTFS standard and is similar, except for the missing ”Shapes.txt” file and the ”wheelchair_accessible” property in the ”Trips.txt” and ”Stops.txt” files for the suburban lines.. In particular, inside of this dataset we can find various files:

  * Agency.txt - This file contains data about Trentino’s Trasporti, the main public transportation provider in the city of Trento and in its region.
  * Calendar.txt - Contains data about the frequencies of a particular service, showing whether it is provided in each day of the week or not and also the timespan validity of each entry.
  * Calendar_dates.txt - Provides information about exceptional dates, showing how the service may change during these days.
  * Routes.txt - Contains information regarding the various line available. In particular, we can find their names (in a brief or more precise manner), their type (busses or trains), identifying colour and about which agency provides them.
  * Shapes.txt - Provide data showing the physical route followed by the vehicle during their trips, by means of sequences of points, characterized by their latitude and longitude.
  * Stops.txt - This file contains data about public transportation’s stops, including their names, descriptions, positions, zones and information regarding wheelchair accessibility.
  * Stop_times.txt - Contains information about the arrival/departure time of each trip at every stop station, also showing the sequence in which these stops will occur.
  * Transfers.txt - Provide data regarding the intersection point between different lines, showing if it is possible for the user to leave their current transportation vehicle to get on another one.
  * Trips.txt - Shows information about the various lines’ trips that occurs during a specific service’s timespan. For each of them data about the head-sign of the trip, its direction, its physical shape and wheelchair accessibility are provided

- Taxi: In this dataset are contained information regarding the various taxi stands located in the city of Trento, such as their position (both in WKT coordinate system standard, lat/long and street name) and the stand’s name.
- Car sharing: Through this dataset we can extract data about the parking lot used for the car sharing service in Trento. For each lot is provided its position (WKT and street name), its capacity and its decree number.
- Stazioni Bikesharing Trentino: This dataset provides information about bike racks used for bike-sharing services. For each rack we have its id, a brief description, location, capacity and type.

- C'entro in bici: Similarly to the previous dataset, this one provides details of bike racks designed for the "C'entro in bici" bike sharing public project. For each of them its provided the position (WKT), a brief description and the capacity.
- Rastrelliere per biciclette: In this dataset we can find details regarding public bike racks available in Trento, in particular we have their position (WKT, zone in the city, street name, street number and nearby palaces), type, capacity, year of installation, number of modules, photo, and category. However, this information aren't always available for every bike rack and many of them have one or more missing values.
- Parcheggio protetto per biciclette: This dataset focuses on bike racks that are protected, usually within a covered area or warehouse. For every rack we have its position (WKT, park and street name), a brief description and its capacity.
- Punti sosta monopattini condivisi a tariffa agevolata: In this dataset are stored data about scooter sharing stations. In particular there are information regarding location (WKT and street name), id, decree, reference code and additional notes.
- Itinerari ciclabili esistenti: This dataset provides information about the existing bike paths, showing for each of them their shape, type, year of construction, itinerary and decree.

- **OpenStreetMap:** In order to extract data from OpenStreetMap, we will use Overpass turbo, a tool that allow to easily query OpenStreetMap, through its APIs, and to show the obtained results directly on a interactive map. The query used to gather the following dataset can be found in our Github Repository.
  - Trento's Parking Lot: Using OSM, we can get details about parking lot in the city of Trento. These lots can be characterized by many different properties, such as: id, position (latitude, longitude), accessibility, capacity, information of eventual owners, fees, name, opening hours and information about the floor they are located on; However, most lots included only subsets of these properties, with smaller ones having often having just a few of them.
  - Bycicle path in Trentino Alto Adige - Sudtirol: From a specifically crafted query, we can gather information about the bicycle paths in the Trentino region. For each path, details on shape, walkability, surface type, and lighting are provided.
  - Trentino's Train lines' shapes: With a query we can gather information about the railways in Trentino, with details on shape, maximum allowed speed, name, lines and whether they are electrified.

This collection includes all the dataset we gathered during the current phase and can also be found on our Github Repository. To confirm that these resources are sufficient, we must ensure

they cover the list of the Competency Questions defined in the previous phase. Since they do, we can proceed with the next step.

3.3 Dataset Cleaning and Standardization

During the last step of this second phase, we will focus on removing any noise in the dataset collected so far. Specifically, we will remove those datasets, entities and properties that don't align to our formal purpose and modify the remaining ones to fit our goals. Note that the original datasets were primarily in Italian, so we will translate properties and values into English as needed. After that, we will convert every file to the same standardized format: CSV.

Follows a more in detail description of how we concretely modified the datasets:

- Trasporti pubblici del Trentino: For what concerns the Trentino Trasporti dataset, we decided to maintain all the files except for transfers.txt, stopslevel.txt, shapes.txt and feed_info.txt, this apply for both urban and suburban. The remaining files contain mostly useful information in an already well known standardized format (GTFS). For this reason only few small changes has been applied:
  - routes.txt: Removed the fields route_color and route_text_color.
  - stops.txt: Removed the fields stop_code, stop_desc, zone_id and renamed wheelchair_boarding as wheelchair_accessible.
  - trip.txt: Added the field bikeSlots, stating how many bicycle slots are available to the trip passengers. Busses will have a default value of 0; trains of 6. This information has been gathered from this page.
  - *.txt: Removed references to not used files.
- Taxi: For what concern the Taxi Stands we decided to maintain only their names and positions. The initial file provided three different properties for the position: WKT, containing a UTM-32 Point, a particular coordinate system, and two others called x and y, expressing the position using latitude and longitude. Being these properties redundant, we decided to keep only the latter two. Follows the list of changes applied to this dataset:
  - id: Not present in the original file, so a hand-crafted one was introduced.
  - x: Renamed to "latitude" in order to increase the readability.
  - y: Renamed to "longitude" in order to increase the readability.
  - nome: Renamed to "name" in order to increase the readability.

  The final CSV file is composed of: id, name, latitude and longitude.

- Car sharing: For what concern the Car sharing stations, we decided to maintain only some properties: via, auto and WKT, although some changes has been applied as follow:
  - id: Not present in the original file, so a hand-crafted one was introduced.
  - WKT: Converted from a UTM (32) Point to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude”.
  - via: Renamed to description. Originally contained the name of the street the station is located and a brief description of the place. Being the street address redundant, we decided to maintain only the description.
  - auto: Renamed to capacity. Used to indicate the number of cars available in every car sharing station, so it has been renamed in order to increase the readability.

  The final CSV file is composed of: id, description, capacity, latitude and longitude.
- Stazioni Bikesharing Trentino: For what concern the Bike-sharing stations, we decided to maintain only some properties: id, desc, ciclopositeggi and WKT, although some changes has been applied as follow:
  - WKT: Converted from a UTM (32) Point to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude”.
  - ciclopositeggi: Renamed to capacity. Used to indicate the capacity of every bike sharing station, so it has been renamed in order to increase the readability.
  - desc: Renamed to description.

  The final CSV file is composed of: id, description, capacity, latitude and longitude.
- C’entro in Bici: For what concern the second bike sharing dataset, we decided to maintain only some properties: desc, ciclopositeggi and WKT, although some changes has been applied as follow:
  - id: Not present in the original file, so a hand-crafted one was introduced.
  - WKT: Converted from a UTM (32) Point to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude”.
  - ciclopositeggi: Renamed to capacity. Used to indicate the capacity of every bike sharing station, so it has been renamed in order to increase the readability.
  - desc: Renamed to description.

  The final CSV file is composed of: id, description, capacity, latitude and longitude.

- Rastrelliere per biciclette: For what concern the bike racks dataset, we decided to maintain only some properties: id, Tipo_generale, tot_bici and WKT, although some changes has been applied as follow:
  - WKT: Converted from a Linestring, that is a list of UTM (32) Points, to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude” and assigning them the average value of the initial one.
  - Tipo_generale: Renamed to type. Used to indicate the bike rack’s type, whose possible values were “tradizionale” (traditional) and “bloccatelaio” (frame lock), so it has been renamed in order to increase the readability.
  - tot_bici: Renamed to capacity. Used to indicate the capacity of the bike rack, so it has been renamed in order to increase the readability.

  The final CSV file is composed of: id, type, capacity, latitude and longitude.
- Parcheggio protetto per biciclette: For what concern the second bike racks dataset, we decided to maintain only some properties: posti and WKT, although some changes has been applied as follow:
  - id: Not present in the original file, so a hand-crafted one was introduced.
  - WKT: Converted from a UTM (32) Point to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude”.
  - type: Not present in the original file, so a hand-crafted one was introduced, with a fixed value of “guarded”. This field was added in order to allow us to merge the two dataset while having a common set of fields among them.
  - posti: Renamed to capacity. Used to indicate the capacity of every bike sharing station, so it has been renamed in order to increase the readability.

  The final CSV file is composed of: id, type, capacity, latitude and longitude.
- Punti sosta monopattini condivisi a tariffa agevolata: For what concern the scooter sharing service dataset, we decided to maintain only some properties: id, note and WKT, although some changes has been applied as follow:
  - WKT: Converted from a UTM (32) Point to EPSG:4326 format, dividing it in two new properties “latitude” and “longitude”.
  - note: Renamed to description. Used to provide additional information that could help locate the Scooter Sharing locations, so it has been renamed in order to increase the readability.

The final CSV file is composed of: id, description, latitude and longitude.

- Trento's Parking Lot: For what concerns the parking lot JSON dataset, fetched from OpenStreetMap, we decided to maintain only some properties: id, parking, access, capacity, fee, coordinates, although some changes has been applied as follow:
  - access: This property defines the accessibility of the parking, for the purpose of this project we deleted all the private ones. Those were parking spots inside private property that are not of our interests.
  - coordinates: Renamed to latitude and longitude. The coordinates were in the right format but contained in a unique property, so we split it.
  - parking: Renamed to type. Used to indicate the type of every parking lot, so it has been renamed in order to increase the readability.
  - capacity: Used to indicate the capacity of every lot. Since this property wasn't specified for every spot, a default value of -1 has been set whenever needed.
  - fee: Used to indicate the fees required to use a parking lot. Since this property wasn't specified for every spot, a default value of "no" has been set whenever needed.

  The final CSV file is composed of: id, access, fee, capacity, type, latitude and longitude

Note that any dataset mentioned in the Collection phase but missing from the list above has been removed due to discrepancies between the data provided and the data actually required for our purposes.

The results of this cleaning process can be found in our Github Repository, along with the Python script used to alter the various dataset.

4 Language Definition

This section will focus mainly on resources at Language layer: We will identify the purpose-specific concepts meaningful for our project and then formally state their definitions. Through these steps, at the end of this phase, we will produce a purpose-specific language file.

4.1 Concept Identification

The first activity of this Language phase focuses on defining the language resources for our project. To achieve this, we will identity every concept representing ETypes, properties and data properties values that will be used in the final Knowledge Graph to represent information. These concepts will be extracted from the results obtained in the previous iTelos phases, specifically

from the ER model, PFSheet (produced during the first phase), and the resources identified at Language, Data and Knowledge layers (collected during the second phase).

Once identified, we will focus on determining the formal meaning of these concepts, either by finding an existing formal definition or defining one ourself whenever needed. To find these definition, we will leverage already existing resources, with a particular focus on the Universal Knowledge Core (UKC). Specifically, we will explore the UKC to search for the previously identified concepts and use the definitions provided (referred to as gloss) when they align with our objectives. If they do not meet our needs, we will consider other language resources among the ones identified in the Dataset Collection step. If no suitable definition can be found, we will create a formal definition ourselves. This will happen mainly for highly domain-specific concepts, such as the ones related to ETypes classified as Contextual in the PFSheet.

The final step regards the creation of the file representing the language resources meaningful for our purpose. This file will be structured as a table containing the following columns:

- **Concept ID:** Contains the ID identifying each concepts. When mapping to UKC Concepts is available, we will use the corresponding UKCIdentifier; For other resources, such as OpenStreetMaps, we will use the URL of the web page containing the concept and its term; Lastly, when we need to create a definition ourselves, the corresponding concept will use an incremental id in the range assigned to our project, which is KGE24-1. In this way, the IDs will follow the format KGE24-1-<concept_id>, where concept_id is a numerical value obtained incrementing the previous value by 1, starting from 1.
- **Concept label:** Contains the word we are formally defining.
- **Concept description:** Contains the formal definition and meaning of each concept.

During the creation of this file, we considered including details on labels and description in multiple languages: English, as it is the de-facto world-wide standard language; Italian, since Trento is located in Italy and Italian is the main language used in the territory we are working on; German, as it is the third most spoken language in Trentino. German was also taken into account due to the significant number of German tourists who frequently visit Trento, as highlighted during the definition of personas in the first phase of the project. However, for the purpose of this course project, we decided to not include the information in German, mainly because neither of the two team members speak this language. Using automatic translations for such resources, which require exceptional accuracy to avoid misleading users does not seem like an optimal solution. Note that, in most cases, the translation of both words and glosses has been done manually.

Below is an image partially showing our language resource file, while the whole version is available in our GitHub Repository

ConcetID	Word-en	Gloss-en	Word-it
UKC-21898	Train stop, Train station (UKC)	Terminal where trains load or unload passengers or goods	Fermata del treno
UKC-45118	Bus stop	A place on a bus route where buses stop to discharge and take on passengers	Fermata dell'autobus
UKC-24387	Transit line	A line providing public transit	Linea di trasporto
KGE24-1-1	Train line	Part of a transportation system that provides, by means of a train, transportation from a fixed position to another following a predefined railroad	Linea ferroviaria
KGE24-1-2	Bus line	Part of a transportation system that provides, by means of a bus, transportation from a fixed position to another following a predefined path	Linea degli autobus

Figure 3: Section of the Language Resource file

4.2 Dataset Filtering

In this second activity, we will focus on the data layer of our final Knowledge Graph. Specifically, we will align the data-level resources previously collected with the concepts that we have just formalized, filtering out every data element not defined by any of these concepts.

In our case, every EType, attribute and property has a direct mapping to the language resource, so none of them will be removed.

5 Knowledge Definition

Having formalized the language resources, in this section we will focus on the Knowledge layer, aiming to develop a knowledge teleontology for our project. Additionally, we will also align the dataset collected in the phase 2 to ensure that they match the modelling choices defined by the teleontology. This alignment will result in a dataset structure consistent with the teleontology, unifying in this way the representation of the information collected so far.

5.1 Modeling a Knowledge Teleontology using kTelos

The first step of the Knowledge Definition phase aims at modelling the previously collected knowledge in a knowledge teleontology, following the kTelos process. To achieve this, we will leverage the language resources, defined during the third phase, which will help us in modelling both ETypes and properties. Specifically, we will start by creating a hierarchical structure of ETypes using an IS-A hierarchy, selecting terms from the language resource that denotes entity types. Once the ETypes are defined, we will focus on object properties, identifying terms that denotes relationships between ETypes and defining them as object properties, including their

domain and range. A similar approach will be applied to the definition of data properties, ensuring a comprehensive and structured representation of the knowledge, completing in this way our knowledge teleontology file.

To concretely apply this process, we will use Protégé, ”A free, open-source ontology editor and framework for building intelligent systems”. Specifically, we created the hierarchical structure of ETypes within the Classes section, adding various Annotations for each class:

- **rdfs:label:** A standard annotation where we specify the EType name.
- **rdfs:comment:** A standard annotation where we inserted the Gloss-en defined in the Language Resource File.
- **conceptID:** A custom annotation created by us, representing the ID of the concept. This follows the values and formats specified in the Language Resource File. Since we do not have multiple classes with the same name referring to different concepts (and thus different ETypes), we decided not to include the conceptID directly in the concept names. Instead, we added it as an annotation, ensuring a clearer and more organized final result.
- **isEtype:** Another custom annotation created by us, denoting that the class is indeed an EType.

Note that for the annotations ”rdfs:label” and ”rdfs:comment”, since these were written in English, we also added the language tag, setting it to ”en”.

Then, as specified by kTelos, we proceeded with the definition of our Object properties, specifying for each of them their Domains and Ranges, which represent the ETypes involved in the relationship. During this process, one of our object property, ”has_a”, was used to link multiple pairs of ETypes. This was feasible because each of these relationships connected the City EType to another EType, and the semantic meanings of this link remained consistent across all the cases.

The final step to create our knowledge teleontology involved the Data properties. This process was similar to the one followed for the object properties, as we specified both the domain and range for each property. However, their meaning differed from the previous one: the Domain indicates the ETypes to which the property is relevant; while the Range defines the allowed value types for the property. Specifically, we selected from a limited subset of data types provided by Protégé: xsd:boolean, xsd:int, xsd:float, xsd:string and xsd:dateTime.

This process helped us defining our initial teleontology by formalizing the informal ER model developed during the first phase into an OWL file. This initial version can be found in our Repository.

The next step involved the integration of existing external ontologies, identified during the second phase, with the goal of aligning them with our teleontology. Specifically, we focused on

the ontology provided by Schema.org, leveraging its well-established structure and concepts. During this process, we analyzed the entities and properties in our teleontology to identify:

- **Equivalent concepts:** When a concept in Schema.org fully matched one in our teleontology, we annotated the corresponding class in our teleontology with a custom annotation ”equivalentTo” to explicitly align it with the Schema.org concept.
- **Hierarchical relationships:** When we identified Schema.org concepts representing a more general version of one of our classes, we linked them using the ”SubClass Of” relationship.

Additionally, we reviewed our data properties to improve their alignment with Schema.org’s structure. Specifically, we migrated data properties initially associated with one of our classes to a Schema.org class introduced during this step. This adjustment was particularly relevant when the latter stood in an ”Is-A” relationship with our original class, ensuring better semantic alignment and interoperability.

One of the most significant result of this step was the introduction of a new class, ”Place”, derived from Schema.org, which will be the parent of many previously defined ETypes. Due to this relationship, we also migrated the ”latitude” and ”longitude” data properties shared among these subclasses to their new parent. Another important change introduced with this new class was the removal of ”Point”, since it was equivalent to Place. Consequently, our old EType was no longer relevant and was eliminated.

The final teleontology obtained through this process can be found in our Repository.

5.2 Schema Alignment

The last step of this Knowledge definition phase is Schema Alignment. During this phase we will produce another OWL file representing our final teleology, obtained by aligning the informal ER model to our teleontology. Schema Alignment is a critical step, as it ensures that the teleology reflects the purpose defined for the iTelos iteration. Typically, this phase require a complex process, often involving machine learning models, to ensure a precise alignment between schema components. However, due to the time constraint of the course, we will limit our workflow to the following steps:

1. **Identifying leaf ETypes:** We identified the most specific entity types in the hierarchy (the leaf nodes), for which we have data.
2. **Dropping general ETypes:** General ETypes higher in the hierarchy that are not directly linked to available data will be removed from the schema.

3. **Inheriting purpose-specific properties:** Whenever applicable, object and data properties defined for the removed ETypes will be transferred to the relevant leaf entity types.

Following these steps, we created our reference Teleology, which is accessible in our Repository, while below is an image showing its classes:

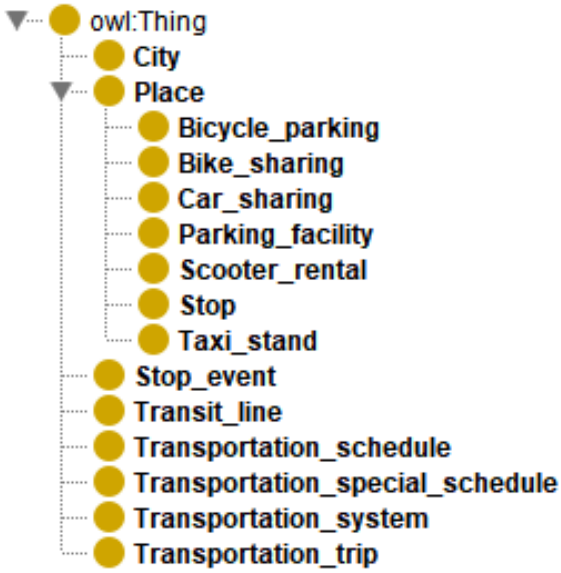


Figure 4: Teleology's classes

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 ?