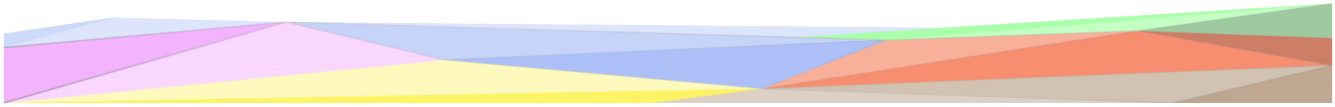
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2023 - Trentino Territory and Transportation

Document Data:	Reference Persons:
October 25, 2023	Rubens Rissi Onzi, Ferrari Eugenio



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 16, 2023	Rubens Rissi Onzi, Ferrari Eugenio	Phase 1 - Purpose Definition
0.2	October 25, 2023	Rubens Rissi Onzi, Ferrari Eugenio	Phase 2 - Information Gathering

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 sub processes) developed, provides a clear understanding of the project, thus serving such an information to external readers for the future exploitations of the project's outcomes.

This project aims at producing useful data for applications that intend to tell it's users about possible delays using the data we provide. The data will feature details mainly about bus stops locations using it's territory data. With that benefiting Trento inhabitants of urban areas.
The current document has the objective provide a detailed report of the project developed following the iTelos methodology. The report is structured, to describe:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: Description of the project development, based on the two main sub process considered by iTelos: producer and consumer.
- Section 4: Purpose formalization step of the iTelos methodology. This step aims to formalize the purpose, by extracting the functional requirement.
- Section 5: Information gathering step of the iTelos methodology. This step aims to formalize the sources, Resourced and data.

2 Purpose and Domain of Interest (DoI)

2.1 Project Purpose

The goal of this project is to provide data to applications and services that need information suitable for predicting bus delays. To do so, we want to incorporate data that could affect transportation delays in urban areas, more in detail. We will consider supermarkets, catering stores, education facilities, tourism destinations, population density, number of right and left turns and number of traffic signs [1]. The data shall be integrated in knowledge graphs (KGs), using as base bus transportation and territorial data available. This release is intended to enable applications to use our KGs, created by the available data, to predict possible transportation delays in Trento, city of Italy.

2.2 Domain of Interest

The project will focus on bus transportation and will utilize data from the following time frames:

[1] Relevant features to predict delays: `https://journals.sagepub.com/doi/abs/10.3141/1666-12`

- Bus transportation data in Trento urban areas, covering the period from 1 September 2022, to June 2023.
- Territorial and facilities data from *Trentino* OSM Places, collected up to 28 February 2023.
- Population data for the year 2018, sourced from the most recent survey conducted by the *Comune di Trento*.

3 Project Development

3.1 Data Production

To fulfill our purpose, we need to produce some resources. These resources either do not currently exist or are of insufficient quality.

The first resource the delay out group could not find any data of it as the *trentino* transport doesn't release any data on the subject. For further works as data becomes available can be added to this dataset. For its properties were chosen the predicted delay and actual delay, as its a good way to predict actual delays and compare them to what was chosen.

The second resource we require regards the Trento city areas, we want data about their names, boundaries, and population density. The Municipality of Trento is divided into 12 official areas, the population density can be used as one of the factors that cause bus delays. The density of population can be a factor in the delays of within these subdivisions busses, as it can bring too many people in one station making the bus wait more for it.

The third resource we need is about the number right and left turns, traffic signs, and the length of specific bus lines. We will integrate this new information on the bus lines, adding information regarding it's name in a matter to identify the lines path. With the path gathered we will construct the entities properties: length, left_turns, right_turns and traffic_signals. As mentioned in the paper of 2.2, the length of a line can be a factor in the delays of buses as the left turns case the right turn case was added as it could be used in countries where people use the left-hand traffic system. We plan to create this data by using the coordinates of the lines and checking changes on its axis that do not correspond to roundabouts.

The fourth resource we need is bus_ride, but this information can not be found in any of the datasets currently available, as it uses the seats and bus id as its proprieties.

3.2 Data Composition

For our project we are using for the territorial data, *Trentino* OSM places, that contain data for urban transportation that will be used for both buses entities, as the name of bus stops and traffic signal proprieties. This dataset will provide the information about the facilities in the city as well, such as, name types and coordinates. Within the information gathered we have chosen some sub classes of facilities as more important than others. Then important facilities were chosen as probable cause on many people trying to access the transport service at same time or in case that it causes business in the area they are located. The chosen facilities were: catering, supermarkets, tourism locations and education facilities. With the KGE22 -

Trentino Urban Transportation we will gather bus line schedule, their coordinates, names and lines. These selected resources will be composed together with the ones created in the data production step. As mentioned before, we want to integrate the produced data regarding bus line characteristics with the KGE 22 dataset such as: number of signs, turns and length. This integration will provide valuable information for predicting delays. We also want to link each bus stop within the KGE22 dataset with its respective city area, allow us to relate the bus stops with the population density in that area which may contribute to delays. Last step is to connect each bus stop with the facilities in that area (that we get from Trentino OSM Places dataset), services and facilities in an area can serve as an indicator of the volume of people visiting, potentially leading to delays.

4 Purpose Formalization

Our Purpose can be formalized by the scenarios, personas and competency questions. For this Project they are as it follows:

4.1 Scenarios

1. A day in Trento on a weekday.
2. A day in Trento on a weekend.
3. A day in Trento on rush hours.
4. A day in Trento on nighttime.

4.2 Personas

1. Giovanni, 19, is a college student that lives in the city centre, even though he studies far, in Mesiano.
2. Isabella, 83, is a senior citizen and lives in the outskirts of Trento, Cassoti di Povo, she often goes with her husband for groceries in weekdays.
3. Lily, 21, is a waiting staff worker in a hotel in the city center, she lives with roommates in a flat in Madonna Bianca. She also likes to go to parties and events in the city.
4. Giosepina, 45, is a wildlife biologist that lives in Mattarello, she has to spend time both in the gathering of samples and behaviours in the wild. She also need to record reports about the samples gathered to the biology department in Povo.

4.3 Competency Questions

1. Isabella, after lunch, wants to reach the city center, where she can find lot of shops to buy groceries to prepare sweets to her daughter. She wants to know how much time it is gonna take to reach the center, and arrive home for dinner.

2. On Tuesday, Giovanni ends his lectures at 19:30 in Mesiano university department. He's curious about his arrival time at the city center, considering that 19:30 falls during rush hour when many people are heading home from work.
3. Isabella wants to visit her daughter in the weekend. As her daughter is available only on Sunday mornings, she have to be aware at witch time she should take the bus, as in the weekend there aren't many available in her part of the city. Her daughter also lives in the other side of the city, Gardolo. She also needs to change bus lines in between.
4. Lily don't want to go too early to work, but as she works sometimes in the night shift starting at 18:30, or in the day shift starting at 7:00. She usually have to go early to not get late. The bus that she rides in the afternoon is always full, by the time that she needs to leave. The hotel she works is well placed, having many amenities like museum and attractions nearby.
5. After enjoying a dinner with his friends, Giovanni decide to head to one of his friend's houses in Martignano. They are fortunate to be right on schedule for the last bus. The buses to Martignano are usually punctual, as the area had fewer residents compared to the city center, resulting in less traffic. They want to confirm if the bus will run on time so that they can arrive at their destination on time.
6. Giosepina needs to present the data gathered in the province of Trento, but as it is raining she prefer not to go by motorcycle, her preferred method, as it would drench her clothes and risk getting sick. Then she decides to take a bus from her house in Mattarello to go to Povo, but she has to take multiple lines for it, she then needs to check the delay of the second line in order to arrive in time in Povo, as her bus stop is the first in the bus ride.

4.4 Concepts Identification and Categorization

From the scenarios, personas and CQs we extract the following entities with their properties:

Table 1: CQs Extraction

Scenarios	Personas	Competency Questions	Entities	Properties	Focus	Popularity
1 - 2 - 3	2	1	bus_stop	(id: int, name: string, coordinates: location, arrival_times: schedule)	Contextual	Common
1 - 2 - 3	2	1	delay	(id: int, predicted_delay_time: string, actual_delay_time: string)	Contextual	Contextual
1 - 2 - 3	2	1	supermarket	(id: int, name: string, coordinates: string,)	Common	Common
1 - 3 - 4	1	2	education_facilities	(id: int, name: string, coordinates: string,)	Common	Common
1 - 3 - 4	2	3	bus_line	(id: int, name: string, length: float, left_turns: int, right_turns: int, traffic_signals: int)	Contextual	Contextual
1 - 3 - 4	2	3	city_area	(id: int, name: string, boundary: polygon, population_density: float)	Core	Contextual
1 - 2 - 3 - 4	3	3	tourism_destinations	(id: int, name: string, coordinates: string,)	Common	Common
1 - 2 - 4	1	4	catering	(id: int, name: string, coordinates: string,)	Common	Common
1 - 3	4	5	bus_ride	(id: int, seats: int, bus_id: int)	Contextual	Core

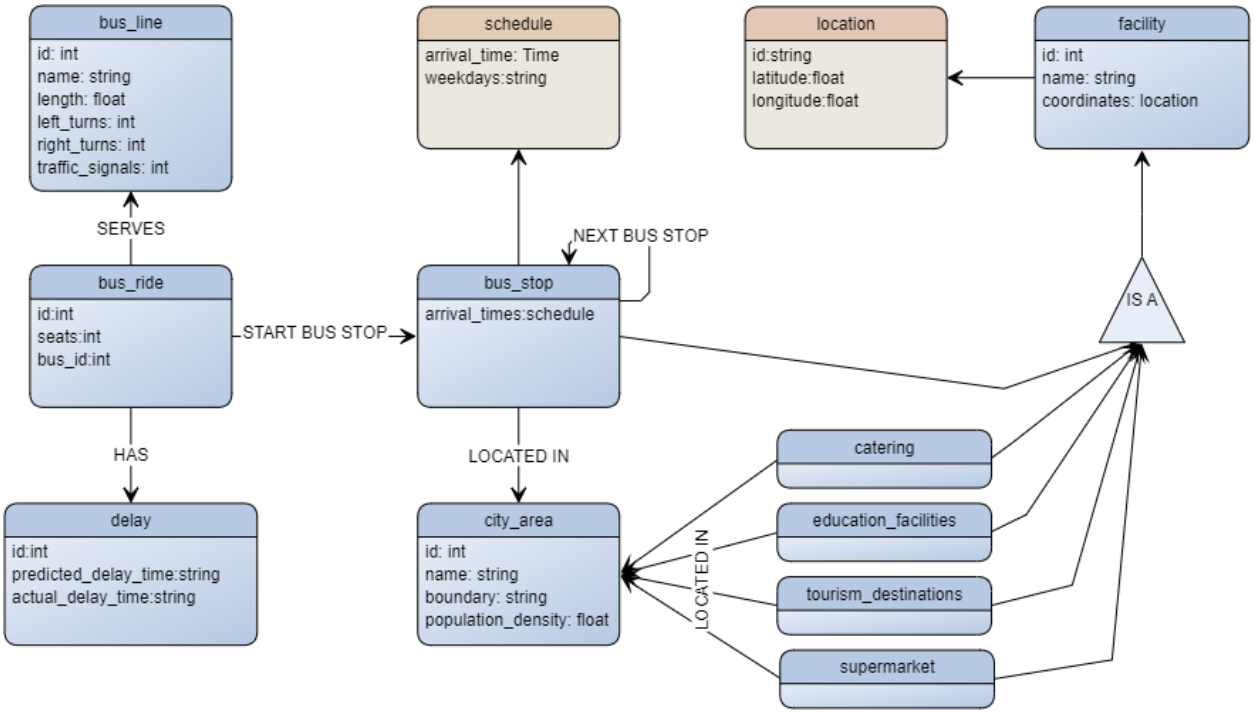


Figure 1: ER model

4.5 ER Modeling

Given the entities and property identified in the step above, we can design the purpose ER model as in Figure 1.

In our ER model we have 10 ETypes.

1. delay: as the main importance for the project, the delay will have the following proprieties:
   - id: this propriety is the identification of all the entities, as it was used in all the other entities it will not be explained.
   - predicted_delay_time: chosen to store possible tries to predict the delay in contrast with the actual_delay_time. Can also be used in future predictions.
   - actual_delay_time: chosen as it will help the future predictions of models in contrast with the predicted_delay_time. It has the problem that we currently have no free database available with it.

2. facility: is used as a superclass for the bus_stop, catering, education_facilities, tourism_destinations and supermarket subclasses. It is used to represent a location with a purpose that could be used to predict delays. We decided to model these specific subclasses as we think that these are the most important for predicting delays.
    - name: Naming used for the facility, as a matter of identification for people.
    - coordinates: A location based propriety that has its own proprieties that will be explained below. Is used to locate a facility on a map that uses longitude and latitude.
3. catering: is a subclass of facility, it represent all those places, like restaurants, bars and pubs, as defined by the OpenStreetMap Data in Layered GIS Format [2] used for Trentino OSM Places in the catering layer. Catering facilities are prone to have rush hours in specific times, lunch and dinner, this could contribute to delays in nearby bus stops.
4. education_facilities: is a subclass of facility, it represent all the education related amenities: universities, schools, kindergartens, colleges and public buildings, again following the standard cited above. They could also inflict rush in the beginning of lessons, but not only that as they bring many people to use this services at one time, leading to delays.
5. tourism_destinations: is a subclass of facility, it represents all the facilities that could be a tourist destination, for example: museums, castles and monuments. Touristic places can bring many people in one place and also not only tourists. This movement of in or out of the busses at some bus stops could bring a delay effect after the tourism spot station.
6. supermarket: is a subclass of facility and it represents the supermarkets. This facilities can be very busy at end of shifts or at lunches, as many people gather around them, or sometimes use the public service to buy groceries.
7. city_area: as previously said in the report, we consider the twelve official division of the urban area of the municipality of Trento, this allow us to model population density for the Trento urban areas. It will be used as way to find how many people live in the area, as many people use the transport, it will bring delays. Each city area is modeled by this entity with the following properties:
    - name: Name of the *circoscrizione*, as a matter of identification for people.
    - boundaries: the geographical boundaries of the area, used to know where the greater population densities are.
    - population_density: the density of the population living in that area of the city of Trento, used to try to infer delays by people using the public transportation.
8. bus_line: this entity represents a bus line with some of the information needed for our purpose that aims at predicting delays, as some of it's proprieties are used in this endeavor, these are the properties:
    - name: the name of the bus line, as a matter of identification for people.

[2] OpenStreetMap Data in Layered GIS Format used in Trentino OSM Places: `https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf`

- length: the length of the bus line path, as delays times can cause ripple effect on it's path.
- left_turns: the number of left turns in the path, as this maneuver can sometimes conflict with the ongoing traffic, making it slower to follow the predicted path.
- right_turns: the number of right turns in the path, as the left_turns but in right side driving countries.
- traffic_signals: the number of traffic signals on the path, can cause small delays if the bus is unlucky to get all of them red.

9. bus_stop: is a subclass of facility, it represent a bus stop, other than the properties inherited from facility. It also has a self-relation to itself, allowing us to represent the sequence of bus stops that a particular bus ride needs to perform
    - arrival_time: type schedule (the schedule type is defined below), is used as a base for who will use to know when the bus should arrive.
10. bus_ride: this entity represent a specific bus ride for a particular bus line, it as a relation to his first bus stop and contains the following properties:
    - seats: the number of seats for this ride, also this can be considered as a factor that could possibly contribute to delays, as more people can enter and exit it.
    - bus_id: the identifier of the bus ongoing ride, as delays can be chained to other stations delays, we identify the ones that are in delay.

We also defined two datatypes to help us modelling our model:

- location: it represent a position, it is composed by the latitude and the longitude used to know where facilities are located.
- schedule: it has two properties, the first one is the arrival time of the bus, while the other is a string of 7 bits containing information about the weekdays for the schedule as defined by the KGE 2022 - Trento Urban Transportation. The type will be used as a way to know the next bus station of the ride.

5 Information Gathering

5.1 Data and Knowledge Sources Identification

We identified three resources:

- Trentino OSM Places and his relative ontology, this is a formal resource providing a cleaned and classified OSM dataset with boundary of Trentino. It is organised in a folders tree representing categories and subcategories. The dataset is based on the Trentino OSM Lightweight Ontology.
- KGE 2022 - Urban Transportation, this dataset was produced during the KGE course and it models the Urban transportation's in the urban areas of Trento.
- The last one is an informal resource for data about the city areas in Trento and their population, taken from the *Comune* di Trento:
  - City areas: `https://www.comune.trento.it/Aree-tematiche/Cartografia/Download/Circoscrizioni` it contains the official subdivision of Trento with the name and the boundaries.
  - Population: `https://www.comune.trento.it/Aree-tematiche/Statistiche-e-dati-elettorali/Statistiche/Studi-e-analisi/Dati-statistici-nelle-Circoscrizioni-di-Trento` it contains PDF files about the most recent survey on the population in Trento divided by the areas.

5.2 Resource Collection, Processing and Scraping

5.2.1 Data Cleaning and Formatting

This subsection we will explain how we extracted the data or produced the data on our dataset.

1. facility
   - id: Attributed for each instance automatically based at time of insertion on the dataset.
   - name: Will be extracted using the propriety name in the OSM place dataset.
   - coordinates: A location based propriety that has its own extraction methods that will be explained below.
2. catering: is a subclass of facility, it's proprieties will be extracted the same way.
3. education_facilities: is a subclass of facility, it represent all the education related amenities: university, school, kindergarten, college, public_building, again following the standard cited above.
4. tourism_destinations: is a subclass of facility, it's proprieties will be extracted the same way.
5. supermarket: is a subclass of facility, it's proprieties will be extracted the same way.

6. city_area: is a subclass of facility, most it's proprieties will be extracted the same way, the ones that differ are:
    - boundaries: Will be extracted by the polygon type on the XML provided by the *comune* of Trento.
    - population_density: this asset is will be extracted by PDFs in the *comune* of Trento website.
7. bus_line:
    - name: the name of the bus line is gathered via the OSM place dataset.
    - length: the length of the bus line will be reused data from KGE 2022 - Urban Transportation.
    - left_turns: we will need to annotate by hand, as its easier than making a program just for it.
    - right_turns: we will need to annotate by hand, as its easier than making a program just for it.
    - traffic_signals: we will the data gathered on the OSM place dataset using traffic signs entity.
8. bus_stop:
    - arrival_time: will be extracted by using data from KGE 2022 - Urban Transportation where we will use it's schedules to know where each buss will arrive in each station at each time scheduled for the bus station.

Delay and bus_ride we don't have how to extract the data as Trentino *transporti* doesn't make it available for everyone.

5.2.2 Knowledge Modeling

Given the resources identified and the data collected, cleaned and formatted we need to associate a schema to each of them. These are the three schemas created in Protégé for each of the resources identified in Section 5.1.

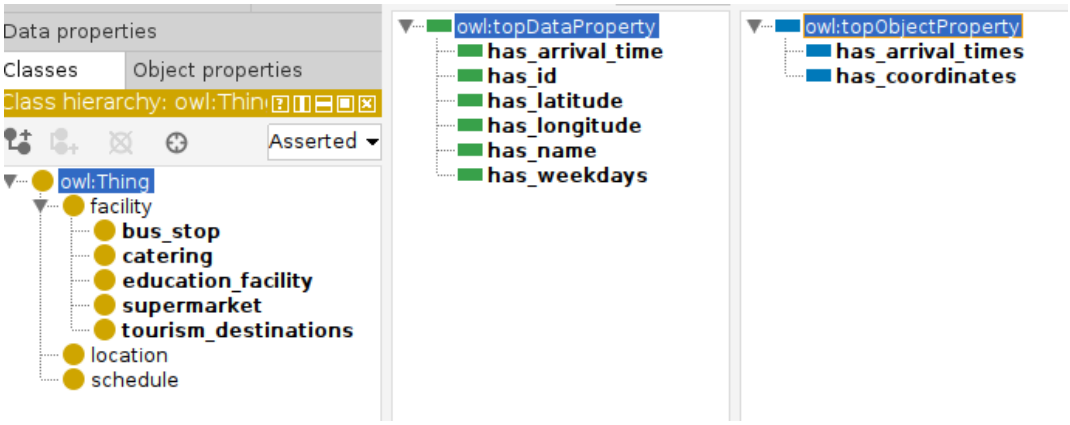


Figure 2: Schema for Trentino OSM Places

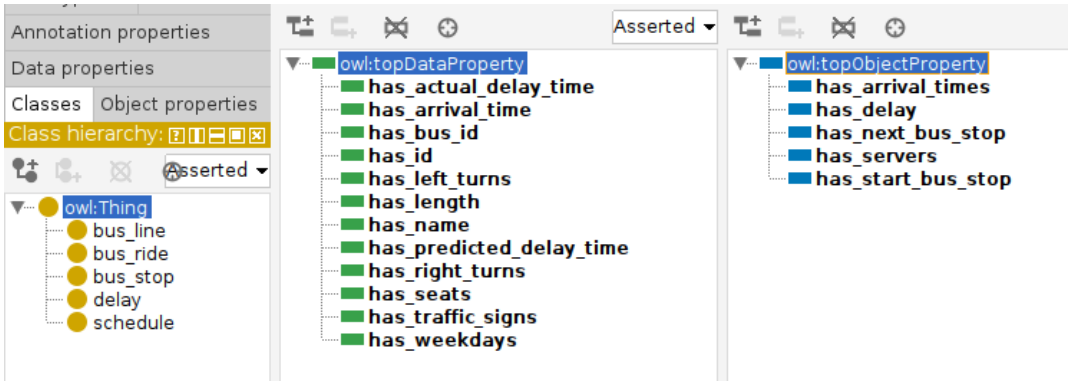


Figure 3: Schema for KGE 2022 - Trentino Urban Transportation

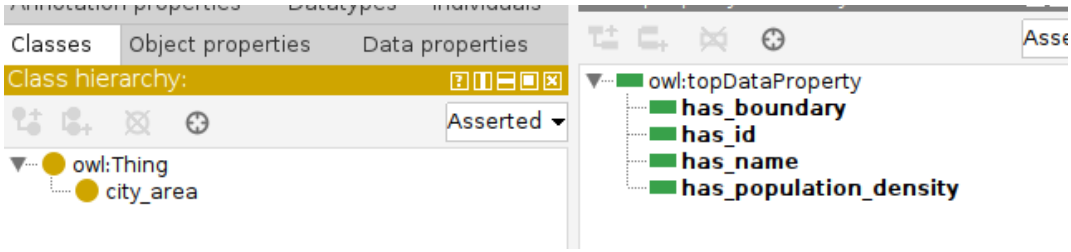


Figure 4: Schema for the data scraped from Comune di Trento