Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2024 - Student life in Trento

Document Data:	Reference Persons:
January 12, 2025	Davide Cavicchini, Yesun-Erdene Jargalsaikhan

Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 30, 2024	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase1 - Purpose defiition
0.2	November 13, 2024	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase2 - Information Gathering
0.3	November 25, 2024	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase3 - Language Definition
0.4	December 4, 2024	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase4 - Knowledge Definition
0.5	December 28, 2024	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase5 - Entity definition
0.6	January 12, 2025	Davide Cavicchini, Yesun-Erdene Jargalsaikhan	Completed Phase6 - Evaluation & Exloitation - Metadata - readability pass - Completed project

1 Introduction

In this project, we will develop a knowledge graph by integrating the following data sources: the iLog app collects information from students studying at the University of Trento; Open Street Map is an open data resource for geographic data; and Trentino Trasporti provides public transportation data in Trento. The resulting knowledge graph will be utilized to facilitate informed decision-making.

2 Purpose Definition

This section introduces the purpose, domain of interest, scenarios and personas, competency questions, and concept identification for the project.

2.1 Informal Purpose

The objective of this project is to build a knowledge graph that assists students in planning their trips from one location to another using public transportation efficiently and comfortably. This tool aims to facilitate informed decision-making and enhance students' overall university experience. This will be achieved by integrating historical data on student commutes and activities, public transportation information, and points of interest.

2.2 Domain of Interest

In this section, we outline the main domains we are interested in to ground our representations. The project is focused on modeling students' lives in Trento, particularly their daily commutes, activities, and interactions. This involves integrating static and dynamic data sources to create a knowledge graph that captures meaningful patterns and supports decision-making.

The spatial dimension centers around Trento, focusing on locations where students might stay. In particular, as we are mostly interested in the commutes and daily activities of students around the city, we identify two main spatial domains of interest:

- **Points of Interest (PoIs)**: Popular locations such as bars, restaurants, libraries, and university facilities.
- **Public Transportation Infrastructure**: Bus stops and routes in Trento.

The choice of the temporal domain is driven by both the need to model the bus routes and to capture the student activities over time. This means we need to both take into account:

- **Time-Specific Patterns**: Modeling bus schedules, routes, and temporal variations (e.g., weekdays, weekends, festive days).
- **Temporal Events**: Tracking the duration and timing of student activities and visits to various PoIs.

One interesting domain that we are interested in exploring is the emotional state of students and their activities. While this information can be grounded in the world by the location and the time at which it occurs, we also need to define their own contained domains. Following the iLog data format, for the mood values we use a discrete scale whose values range from 0 to 5. The activities also come from the SU2 dataset and thus use the same list, which will be formalized during the language definition phase.

2.3 Personas & Scenarios definition

In this section, we introduce the personas and scenarios to ground the purpose on possible use cases of actual users of our knowledge graph.

2.3.1 Personas

To formalize the purpose of the project, we provide personas that cover various lifestyles among students which are useful to define diverse interactions with the knowledge graph.

Person 1 Alessia, a new international student, has recently started studying at the university.

Person 2 Paolo, a second-year master's student.

Person 3 Houda, an Erasmus student who wants to save up money

Person 4 Lucia, a student habit of dining in restaurants quite frequently

Person 5 Emanuele, a student who lives in San Bartolomeo student residence

2.3.2 Scenarios

For the persons we defined, we described some scenarios students could encounter during their university lifestyle in which our Knowledge Graph can assist for making decisions on planning.

1. **Social Interaction** - Alessia has recently moved to Italy and is excited to spend time with her new friends, exploring the city center of Trento, as she is eager to get to know the city.
2. **University Facilities** - Paolo is a second-year master's student at the University of Trento, currently working on his thesis. He wants to study in a quiet, uncrowded place, so he needs to choose one of the university's facilities.
3. **Daily life** - As an exchange student, Houda has started living in the city center and is planning to go grocery shopping. Since the atmosphere in supermarkets varies, he wants to choose the one that best suits his preferences.
4. **Dinner Place** - Regarding her dining habits, Lucia is looking for decent places to have dinner with her flatmate. While exploring restaurants she had both good and bad experiences with plates, so she doesn't want to choose a bad one.
5. **Personal Activity -** Emanuele is a professional athlete looking to have permanent training at the nearest sports facility to his student residence. A regular commute to the facility is an important part of his daily routine, so he needs to choose the one that will save him time.

2.4 Competency Questions

Following the paper on Big-Thick Data generation via reference and personal context unification, what we want to be able to answer are about personal-reference (PR) and reference-personal (RP) context questions. The following is a list of relevant questions for the scenarios and personas we defined which align with the purpose of our Knowledge Graph:

1. **P1-S1**. Is public transport available to reach the destination?
2. **P1-S1**. How many people are currently present on the chosen bus?
3. **P1-S1**. How many people are the social interaction locations in the city?
4. **P2-S2**. Which university facility best fits the student's needs or has the least impact on their mood?
5. **P2-S2**. How crowded is BUC?
6. **P3-S3**. Which supermarket best meets the student's needs?
7. **P3-S3**. What was the student's mood when they were at the Coop supermarket?
8. **P3-S3**. What is the best route to the Coop supermarket?
9. **P3-S3**. How did I feel about the trip to the Coop supermarket?
10. **P4-S4**. Which restaurant served a meal that met the student's expectations?
11. **P5-S5**. Which sports facility is closest to the student?
12. **P5-S5**. What is the best bus route to the sports facility?
13. **P5-S5**. What is the closest bus stop to reach the facility?

2.5 Concepts Identification

In this section, we try to come up with a mostly complete list and description of what type of concepts we are interested in modeling and which properties are fundamental for the Knowledge Graph. In the next section, we will use this information to guide the modeling of the ER diagram.

Table 1 reports the identified concepts and relates each to the specific competency question in which their use is required. The most important concept we identified for this part is the **Emotional State** which will be used by our Knowledge Graph to address the majority of the reference-to-personal and personal-to-reference queries we aim to tackle.

2.6 ER modeling

Having defined the specific concept instances relevant to our identified competency questions, we now aim to formalize the newly acquired insights. To do so, we use the Entity-Relationship (ER) modeling to identify and define the entity types that we will need to manage and use. The resulting model is depicted in the image 1.

Scenarios	Personas	CQs	Entities	Properties	Focus
1-5	1-5	1-11	Student	student_id, name, current_position	Contextual
1, 5	1, 5	1, 13	Bus Stop	name, direction, time_table, location	Contextual
3, 5	2, 5	2, 8, 12	Bus route	number, time_domain, start_time	Contextual
4	4	10	Restaurant	civic_number, name, location	Common
2	2	3-5	University Facility	civic_number, name, location	Common
5	5	11	Sport Facility	civic_number, name, location	Common
1	1	3	Bar	civic_number, name, location	Common
5	5	11	Residence	intercom_name, civic_number, location	Contextual
3	3	6-9	Supermarket	civic_number, name, location	Common
2-4	2-4	4,7,9-10	Emotional state	time, duration, location, user, mood	Core

Table 1: Concepts Identification Table

The final model we settled on comes from wanting to ground all of the entities and information that will be contained in our KG in the pair (`Student, Time`). This choice can be easily justified by how our competency questions are formulated and the process we follow to answer each one of them. Initially, it might seem strange to not ground the information also on the location of each event, emotional state, etc... However, we argue that the information about where a particular event or emotional state occurred can be obtained using a simple walk on the KG nodes, by using the path `Stream_1 - Student - Stream_2` and aligning the two using the time coordinate. For example, to answer competency question 7, we can start by collecting all the times Houda went to the Coop supermarket using the location data stream and extracting the time intervals she spent in them. Later I can use this information to intersect all of the Emotional states of Houda which fall into the retrieved intervals we are interested in.

Additionally, in all the identified questions we are only interested in a subset of places that satisfy specific requirements, such as being suitable for hanging out with friends, having dinner, or working out. Therefore, it was unnecessary to model each place (e.g., restaurant, bar, sports facility) as separate entities. Instead, we collapsed them under a single **PoI** (Point of Interest) entity, differentiated by a *type* property, whose values and definitions we will go over in the next sections.

Using this formulation, we can draw a parallel between how we are storing the information in our entities and the streams of sensor data that might come in from a mobile device. Furthermore, populating the KG with a continuous stream of data from a particular sensor is easy, as we have access to both the `Person` and `Time` coordinate For this project, we are only focusing our attention on modeling the location data stream (GPS), activities, and emotional states, which are more than enough to demonstrate the ability of our KG to answer both PR and RP queries easily. However, it would be relatively trivial to add a new stream of data, as it only needs to be grounded on the Persons involved and the time interval it takes place in to be able to answer queries about it and jointly retrieve other streams.

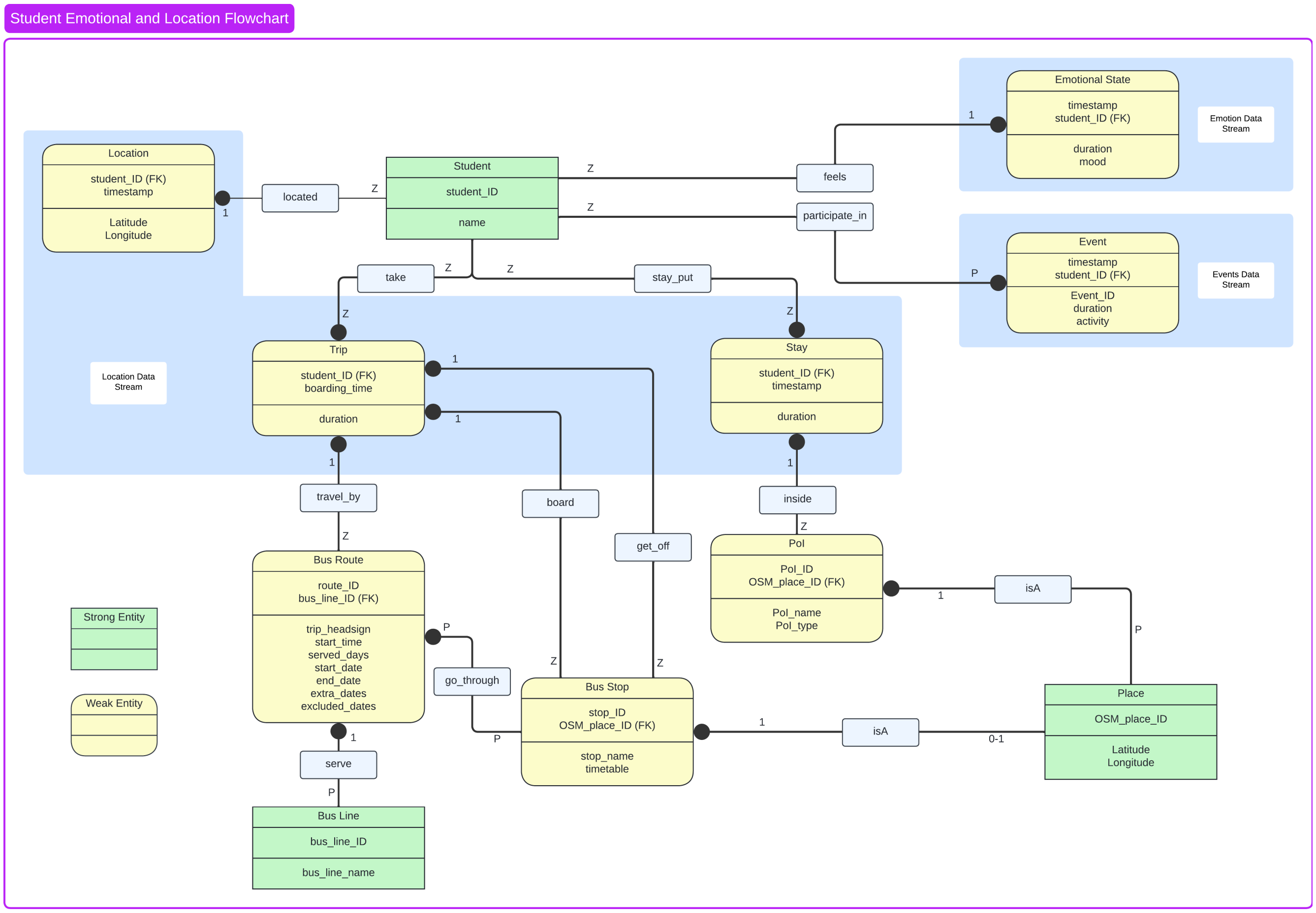

Figure 1: ER model

3 Information Gathering

In this section the second main input for the project is described, namely the data sources. For each resource, we report:

- The name, and the description of the information the resource is carrying.
- Type of resource. If it is a language, schema or data value dataset.
- The source from which such resource can be collected.

This section reports the execution of the steps and it is structured in the following subsections:

- Sources identification - where we identify our needs
- Datasets collection - where we identify the data sources we use
- Datasets cleaning - we explain in detail the structure of the data and how we processed it
- Datasets standardization - we show the resulting tables from our processing and the information they contain

3.1 Sources identification

In this section, we want to briefly reason on what type of data we need to have access to fulfill the purpose, answer the competency questions, and populate the entities we identified in the previous section. To do so, we have to find and use the following sources of data:

- Trentino Trasporti: we need to have access to information about the bus routes, the stops around the city, and the timetables for each of them.
- PoI in Trento: we need to collect all the places a student from the University of Trento might be interested in, comprehensive of university facilities, bars, restaurants, etc...
- Student Information: we need to have detailed information about the students, comprehensive of their mood state, activities, and positions.

3.2 Data sources

We now list and explain the sources we identified and decided to use, which should adequately cover the needs we listed in the previous section. For each dataset, we are interested in explaining what it contains, what data it collects, how to get a copy of such data, and what needs it covers. This will be important to define how we will process this data in the next steps.

3.2.1 SmartUnitn2

SmartUnitn2 is a data-value dataset about the everyday life of one hundred fifty-eight university students over four weeks between May and June 2018. The dataset is collected from thirty-four sensors, including location data and responses from questionnaires.
For the purpose of the project, this dataset is used for extracting information about the student commutes and activities including emotional states.
SmartUnitn2 dataset can be downloaded only through a request form from the authors of the original work, which is ”2018-Smart Unitn 2-Trento”, from this web portal.

3.2.2 Trentino Trasporti

Trentino Trasporti is a dataset for public transportation in the Trentino region of Italy. It carries information about public transport stops, lines, schedules, and fare details through the years.
The dataset is crucial for providing bus route, bus stop, and their schedules during the time SmartUnitn2 dataset was collected.
It is available for download via Dati Trentino portal, a website of the public transport agency of the Autonomous Province of Trento.

3.2.3 Open Street Map

Open street map is an open-source data-value dataset that offers a wide range of geographic data such as geospatial information, map features, and metadata for the features, boundaries, and transportation network.

We use this dataset to ground both points of interest and bus stops in a renowned data source, which should facilitate the reuse of the data resulting from this project.
The dataset can be downloaded from Open Street Map (OSM).

3.2.4 Points of Interest in Trentino

PoI in Trentino is a data-value dataset that provides information on points of interest (PoIs) in the Trentino region of Italy, such as university facilities, bars and restaurants and so on.
We are interested in this dataset to be able to cover all the PoI around Trento that a student might be interested in. The datasets can be downloaded from Open Data Trentino.
However, this data lacks information about university facilities, arguably the most important ones for our project. To get around the limitation we scrape university facility data from OSM using Overpass-Turbo, a web-based data mining tool.

3.3 Datasets cleaning

Having identified the sources from which we will gather the data, we now explain its structure and how we processed it to filter out the information we need. To do so, for each dataset we will start by explaining its structure and the data it contains in details. And finally, explain how we extracted the data we need from them.

3.3.1 PoI & OSM Data

The Points of Interest in Trentino dataset contains information about individual places in Trentino. We use the OpenStreetMap (OSM) dataset as a ground for localizing both points of interest and student locations. By integrating the PoI dataset with OSM, we match places based on their latitude and longitude coordinates. For our purpose, we download the following source:

- **wu2013poi.json**: `latitude,longitude,social_address,contact,category,name, ..`
  We keep only the name, type, latitude, and longitude of the places in the cleaned dataset.

We scrape the OSM dataset for the points of interest related to university facilities, and obtained following data source:

- **osm_universities.json**: `latitude,longitude,name,amenity/category,building, ...`
  We keep the name, type, latitude, and longitude of the places in the cleaned dataset.

3.3.2 Trentino Trasporti Data

The data offered by the Trentino Trasporti agency contains information in the gtfs format for the existing bus routes, the individual trips, the days they run in, the exceptional dates, and the bus stops. For our processing, the files we are interested in are the following:

- **routes.txt**: `route_id,agency_id,route_short_name,route_long_name,...`
  Which we use to get information about the route names and collect the existing `route_ids`

- **stops.txt**: `stop_id,stop_name,stop_lat,stop_lon,...`
  Useful to collect all the bus stops in the city of Trento and their position with which we query the OSM data source to get the `osm_id` for each.
- **trips.txt**: `route_id,service_id,trip_id,trip_headsign,...`
  Which reports all of the bus trips around the city and allows us to connect them to the bus routes using the `route_id` property.
- **stop_times.txt**: `trip_id,arrival_time,departure_time,stop_id,stop_sequence`
  This data allows us to connect all the bus trips to both their path and create the timetables for each one of the bus stops.
- **calendar.txt**: `service_id,monday,tuesday,...,saturday,sunday,start_date,end_date`
  This table is useful to populate the `served_days` property for a particular bus trip by joining it with the trips table over the `service_id` column.
- **calendar_dates.txt**: `service_id,date,exception_type`
  Gives information about the extra dates (`exception_type1`) and the excluded ones (2), by merging it with the trip table over the `service_id` column as before.

Note that we focused our attention only on the urban subset of the data, expanding the code to include the extra-urban split of the data should not be expensive.

3.3.3 SmartUnitn2 Data

3.3.3.1 Time-Diaries (Questionnaires for emotions and activities) The SmartUnitn2 questionnaire dataset contains information about students' responses to a couple of questions.

For our purpose, we are only interested in responses on what the students were doing, how they were feeling, when they were doing it, and for how long. Therefore, we create separate datasets for student activity history and emotion history. For the emotion data, we represent the emotional character responses as a numerical scale from 0 to 4, and textual definition depending on how down or excited the student was.

3.3.3.2 GPS Data The SmartUnitn2 also contains a table with the collected GPS data. This data was collected from the devices with a frequency of one minute, and the related tables have the following relevant columns `userid, latitude, longitude`.

This data is then cross-referenced with the position of the PoIs and the time and position of the busses to construct two tables with all the matches that were found within some margin of error. The data produced uses 50 meters of radius to catch the near PoIs and bus stops, and a time buffer of 8 minutes for the bus arrival times.

The resulting data is further processed to filter out as much noise as possible using some consistency constraints on the routes and stays. In particular, we require the user to have followed a bus trip for more than three stops to be deemed a valid trip and to have stayed at least 10 minutes in the same place to consider it a stay.

3.4 Datasets standardization

In this section, we want to elaborate on the resulting tables and data from this phase. As identified in the previous phase, we have two separate types of information:

Static Entities All the entities used to index the information our KG has about the students fall into this category. These correspond to the entities in the ER about buses, locations, and students.

Data Streams Collect the information that needs to be indexed using the time. Such streams are used to model the following entities:

- Emotions
- Activities (Events)
- Locations, which is further processed to have the information about the bus trips a student took and their stays at the PoI around Trento.

3.4.1 Static Entities

3.4.1.1 Students The data we use for the students is anonymized. For this reason, we do not have access to the names of each student, and the only data we collect to identify them is the unique ids used in the SmartUnitn2 dataset in the **user_ids.json**.

3.4.1.2 Locations (PoI) The data obtained from PoI in Trentino and the scraped data from OSM are merged into poi_and_osm.json. Here, all the individual places in Trento are provided, including the OSM ID, as well as their identifying name and corresponding category. A sample of the dataset is shown in Figure 2.

	osm_id	latitude	longitude	name	type
0	292004245	46.076974	11.141749	Biblioteca Argentario	library
1	428313995	46.051743	11.127649	Biblioteca Clarina	library
2	1559090545	46.066395	11.138857	BUM - Biblioteca Universitaria Mesiano	library
3	1607714595	46.006060	11.127804	Biblioteca Mattarello	library
4	1721783679	46.062734	11.124417	Biblioteca del Polo Umanistico	library

Figure 2: Point of Interest in Trento

3.4.1.3 Busses The resulting datasets from Trentino Trasporti are shown in Figures 3, 4, 5 and cover all the information we need to populate our entities and match our users positions to their trips.

In **bus_routes.json** we collected all the routes served by Trentino Trasporti agency with their unique name, which will be used to give more human-readable responses to the user of our Knowledge Graph.

	route_id	route_name
0	396	3 - Cortesano Gardolo P.Dante Villazzano 3
1	400	5 - Piazza Dante P.Fiera Povo Oltrecastello
2	402	7 - Canova Melta Piazza Dante Gocciadoro
3	404	8 - Centochiavi Piazza Dante Mattarello
4	406	9 - P.Dante S.Donà Cognola Villamontagna

Figure 3: Bus routes dataset

	stop_id	stop_name	stop_lat	stop_lon	osm_id	route_id	trip_id	arrival_time
0	1	Baselga Del Bondone	46.078325	11.047358	1246012567	[536, 536, 536, 536, 536, 536, 536, 536, 536, ...	['0002896002017091120180607', '000289888201709...	['07:04:00', '07:14:00', '07:46:00', '07:51:00...
1	2	Baselga Del Bondone	46.078581	11.047541	1363981295	[536, 536, 536, 536, 536, 541, 404, 536, 536, ...	['0002898952017091120180607', '000289300201709...	['05:13:00', '06:08:00', '06:20:00', '06:52:00...
2	3	Belvedere	46.044406	11.105342	1365224515	[411, 411, 415, 411, 415, 411, 411, 404, 415, ...	['0002875242017091120180607', '000288988201709...	['06:07:00', '06:28:00', '06:45:00', '06:55:00...
3	4	Lamar Ponte Avisio	46.134620	11.110914	307367318	[543, 543, 543, 543, 543, 543, 543, 543, 543, ...	['0002871262017091120180607', '000287127201709...	['06:27:00', '06:47:00', '06:57:00', '07:04:00...
4	5	Sp 85 Bivio Sopramonte	46.085226	11.069313	229059978	[536, 536, 536, 536, 536, 536, 536, 536, 536, ...	['0002892982017091120180607', '000289315201709...	['05:55:00', '06:25:00', '06:58:00', '07:18:00...

Figure 4: Bus stops dataset

bus_stops.json contains all the bus `stop_id`s and positions contained in the dataset and merges the data from the bus trips to populate the time tables of each stop. In addition, we also search up the `osm_id` associated to each stop to be able to ground them using the same `Place` entity as the PoIs.

Finally, **bus_trips.json** collects all the bus routes, merged with the information about the sequence of stops and arrival times, days of the week the route is served (`served_days`), and the extra and excluded dates.

3.4.2 Data Streams

3.4.2.1 Emotions The emotional state stream dataset shows the temporal changes in students' moods over time, specifying the duration of each mood. A sample of the dataset is shown in Figure 6

3.4.2.2 Activities The activity stream dataset shows the activities of students over time, specifying how long the activity has took place. A sample of the dataset is shown in Figure 7

3.4.2.3 Locations The Location stream collects the GPS sensor data of each user. Additionally, it includes higher-level abstractions such as bus trips and stays.

	route_id	trip_id	trip_headsign	stops	times	start_time	start_date	end_date	served_days	extra_dates	excluded_dates
0	496	0002864212017091120180607	Corso Bettini Liceo	[2822, 2823, 2504, 2302, 2301, 2303, 2824, 224...	['06:50:00', '06:50:00', '06:52:00', '06:54:00...	06:50:00	20170911	20180607	[1, 1, 1, 1, 1, 0, 0]	NaN	[20180602]
1	490	0002864222017091120180607	Viale Dei Colli Sc. Alberghiera	[1311, 1285, 1283, 1373]	['07:40:00', '07:41:00', '07:43:00', '07:46:00']	07:40:00	20170911	20180607	[1, 1, 1, 1, 1, 0, 0]	NaN	[20180602]
2	490	0002864232017091120180607	Piazzale Orsi Ist. Veronesi	[1339, 2900, 2343, 2296, 2508, 2478, 1372, 128...	['08:45:00', '08:46:00', '08:48:00', '08:49:00...	08:45:00	20170911	20180607	[1, 1, 1, 1, 1, 0, 0]	NaN	[20180602]
3	572	0002864242017091120180607	S.Lucia Chiesa	[1284, 1282, 1312, 1431, 1363, 1419, 1418, 275...	['15:48:00', '15:49:00', '15:50:00', '15:51:00...	15:48:00	20170911	20180607	[1, 1, 1, 1, 1, 0, 0]	NaN	[20180602]

Figure 5: Bus trips dataset

	userid	timestamp	mood	mood_text	duration
0	0	2018-05-09 19:33:48.395	3	happy	180.0 min 25.0 sec
1	0	2018-05-09 22:34:13.508	4	excited	88.0 min 7.0 sec
3	0	2018-05-10 12:38:14.671	3	happy	73.0 min 13.0 sec
4	0	2018-05-10 13:51:27.778	4	excited	12.0 min 41.0 sec
5	0	2018-05-10 14:04:08.978	3	happy	138.0 min 46.0 sec

Figure 6: Emotional state dataset

	userid	timestamp	activity	duration
0	0	2018-05-09 19:33:48.395	Guardo Youtube, Serie-Tv, ecc.	96 min 3 sec
1	0	2018-05-09 21:09:51.600	Cinema, Teatro, Concerto, Mostra, ...	22 min 54 sec
2	0	2018-05-09 21:32:45.845	Guardo Youtube, Serie-Tv, ecc.	120 min 22 sec
3	0	2018-05-09 23:33:07.865	Leggo un libro; ascolto musica	29 min 12 sec
4	0	2018-05-10 00:02:20.655	Dormire	755 min 54 sec

Figure 7: Activity stream dataset

The **user_poi_stays.json** file specifies the duration of a student's stay at a location over time.

	user_id	osm_id	timestamp	duration
0	0	2607250269	2018-05-17 21:18:55	12.866667
1	1	215037366	2018-05-22 23:17:48	505.483333
2	1	275318445	2018-05-16 08:24:41	29.666667
3	1	275318445	2018-05-16 09:37:21	44.966667
4	1	275318445	2018-05-23 08:32:04	90.666667

Figure 8: Student stays in the location dataset

In **user_poi_matches.json** is collected the data for student locations matching with corresponding PoIs. We also developed a visualization for the extracted data illustrated in Figure 11.

	userid	osm_id	user_timestamp
0	0	2607250269	2018-05-13 00:34:53.189
1	0	2607250269	2018-05-13 21:37:05.588
2	0	2607250269	2018-05-13 21:37:06.871
3	0	2607250269	2018-05-13 21:37:07.557
4	0	2607250269	2018-05-13 21:37:08.979

Figure 9: Student location matches with PoI dataset

The result from cross-referencing the students' locations with the bus trips is collected in **user_likely_trips.json** and it contains information about where each user took the bus, at what time, and where and when they got off. We also developed a visualization for the extracted data which is illustrated in Figure 12.

	user_id	date	trip_id	boarding_stop_id	alighting_stop_id	boarding_time	alighting_time	duration_minutes	number_of_stops	total_tir
0	2	2018-05-18	0002869582017091120180607	1318	1421	2018-05-18 12:36:54.539	2018-05-18 12:43:21.746	6.453450	4	
1	6	2018-05-29	0002875752017091120180607	407	348	2018-05-29 12:45:13.865	2018-05-29 12:50:45.182	5.521950	3	
2	6	2018-05-30	0002892692017091120180607	2426	2780	2018-05-30 17:22:22.525	2018-05-30 17:32:30.575	10.134167	3	
3	6	2018-06-01	0002894352017091120180607	349	406	2018-06-01 17:01:00.655	2018-06-01 17:11:59.210	10.975917	3	
4	6	2018-06-04	0002891992017091120180607	172	348	2018-06-04 09:43:12.040	2018-06-04 09:57:49.920	14.631333	4	

Figure 10: Student likely trips dataset

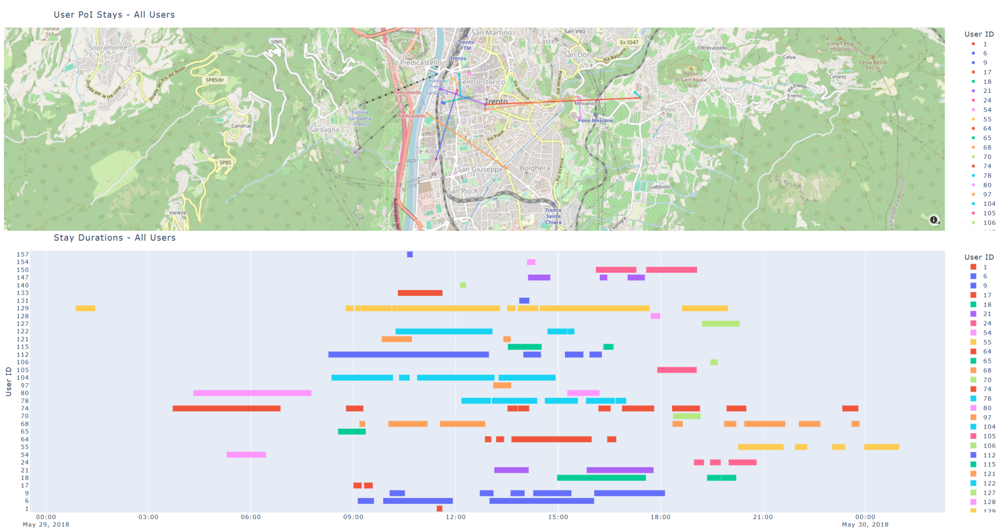

Figure 11: Visualization of the extracted stays for a day

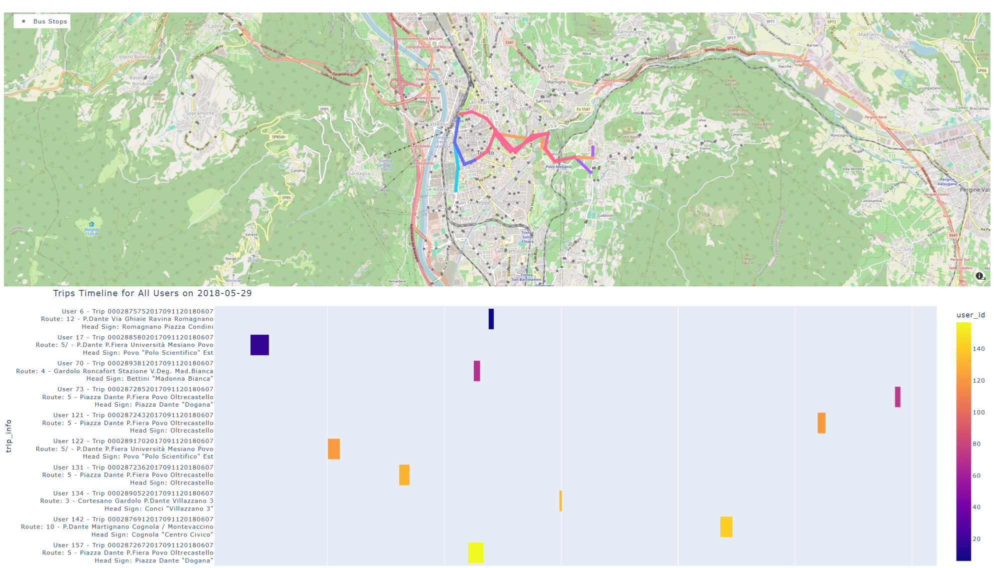

Figure 12: Visualizaton of the extracted trips for a day

4 Language Definition

In the Language definition phase, we formally define the concepts used to represent the information included in the final knowledge graph.

4.1 Concept identification

In this activity, we define the purpose-specific language resources by selecting all the concepts from the information in the dataset and formalizing them using Universal Knowledge Core (UKC) by aligning. Then we generated language resources representing the formalized concepts and their description.

4.1.1 Entity types

In the Language Definition phase, entity types are identified to structure the fundamental building blocks of the knowledge graph. Each entity type corresponds to a core concept derived from the datasets and competency questions contained in the ER model. Table 2 contains the definition of the entity types.
These entity types form the backbone of the knowledge graph, allowing us to represent the data we need.

Label	Description
bus_stop_GID-45118	A place on a bus route where buses stop to discharge and take on passengers
bus_route_GID-45117	The route regularly followed by a passenger bus
bus_line_KGE24-0A-1	The public transport line offered by the transportation agency
trip_journey_GID-1474	The act of traveling from one place to another
stay_GID-5335	Continuing or remaining in a place or state
location_GID-695	A determination of the place where something is
student_GID-53021	A learner who is enrolled in an educational institution
mood_GID-39957	A characteristic (habitual or relatively temporary) state of feeling
place_GID-45095	Any area set aside for a particular purpose
point_of_interest_KGE24-0A-20	Any place in which a person stays for a continuous period of time
event_GID-56	Something that happens at a given place and time

Table 2: Etypes definition

4.1.2 Data properties

Data properties describe the specific attributes of our entities relevant to our purpose identified in the ER modeling and data-gathering phase. The data properties identified for the knowledge graph are shown in Table 3,

Label	Description
duration_GID-72859	The period of time during which something continues
timestamp, timing_GID-27373	The time when something happens
latitude_GID-45424	An imaginary line around the Earth parallel to the equator
longitude_GID-45429	The angular distance between a point on any meridian and the prime meridian at Greenwich
activity_GID-2006	Any specific behavior
mood_GID-39957	A characteristic (habitual or relatively temporary) state of feeling
timetable_GID-34211	A schedule of times of arrivals and departures
stop_GID-5348	A brief stay in the course of a journey
start_time, starting_time_GID-73577	The time at which something is supposed to begin
boarding_time_KGE24-0A-2	The time at which someone enters a vehicle
start_date_KGE24-0A-3	The date at which something begins
end_date_KGE24-0A-4	The date at which something ends
extra_dates_KGE24-0A-5	Dates where something wasn't supposed to happen but exceptionally it does
excluded_dates_KGE24-0A-6	Dates where something was supposed to happen but it was removed
PoI_type_KGE24-0A-7	Class of appartenance of the Point of Interest (Restaurant, supermarket, ...)
PoI_name_KGE24-0A-8	Common name used to reference the Point of Interest
bus_line_name_KGE24-0A-9	Name of the bus line (5, 6, NP, ...)
stop_name_KGE24-0A-10	Name of the bus stop
trip_headsign_?_KGE24-0A-11	Text displayed on the head sign of the bus
served_days_monday_KGE24-0A-12	The considered service is being provided each week on monday
served_days_tuesday_KGE24-0A-29	The considered servide is being provided each week on tuesday
served_days_wednesday_KGE24-0A-30	The considered servide is being provided each week on wednesday
served_days_thursday_KGE24-0A-31	The considered servide is being provided each week on thursday
served_days_friday_KGE24-0A-32	The considered servide is being provided each week on friday
served_days_saturday_KGE24-0A-33	The considered servide is being provided each week on saturday
served_days_sunday_KGE24-0A-34	The considered servide is being provided each week on sunday
student_ID_KGE24-0A-13	Unique Identifier for the student
stop_ID_KGE24-0A-14	Unique Identifier for the bus stop
PoI_ID_KGE24-0A-15	Unique Identifier for the Point of Interest
route_ID_KGE24-0A-16	Unique Identifier for the bus route
bus_line_ID_KGE24-0A-17	Unique Identifier for the bus line
Event_ID_KGE24-0A-18	Identifier for the event
OSM_place_ID_KGE24-0A-19	Unique Identifier from the OSM resource for a point on the map

Table 3: Data properties definition

4.1.3 Object properties

In Table 4, we identified the concepts of object properties. Object property describes how the source entity relates to the target entity, where each entity corresponds to independent concepts.

Properties	**Description**	Etypes
particpate_in, participate_GID-97811	become a participant; be involved in	*Source etype*: Student, *Target etype*: Event
feel_GID-101373	undergo an emotional sensation	*Source etype*: Student, *Target etype*: Mood
is_a_GID-105454	have the quality of being; (copula, used with an adjective or a predicate noun)	*Source etype*: PoI, BusStop , *Target etype*: Place
located_GID-85982	situated in a particular spot or position	*Source etype*: Location, *Target etype*: Student
stay_put_GID-101763	stay put (in a certain place)	*Source etype*: Student, *Target etype*: Stay
take_GID-101699	travel or go by means of a certain kind of transportation, or a certain route	*Source etype*: Student, *Target etype*: Trip
get_off_GID-102589	leave a vehicle, aircraft, etc.	*Source etype*: Trip, *Target etype*: Event
board_GID-102601	get on board of (trains, buses, ships, aircraft, etc.)	*Source etype*: Trip, *Target etype*: BusStop
using, travel_by_KGE-0A-50	use a certain mean of transportation to travel	
go_through_GID-102767	go across or through	*Source etype*: BusRoute, *Target etype*: BusStop
serve_GID-97875	do duty or hold offices; serve in a specific function	*Source etype*: BusRoute, *Target etype*: BusLine
inside_GID-106969	within a building	*Source etype*: Stay, *Target etype*: PoI

Table 4: Formalized concepts of object properties

4.1.4 Data properties values

In Table 5, we identified and formalized the specific concepts appears in the data data property values in the dataset.

4.2 Dataset Cleaning

To formalize and disambiguate our knowledge representation we need to change some of the names used by the datasets we collected.

Fortunately, these changes are subtle and only cover the Trentino Trasporti dataset. In particular, we change the name of ”bus trips” to ”bus routes”, since it contains information about the specific route a bus takes from departure to its destination, and the original ”bus routes” now is referred to as ”bus line” which is the broader concept of, for example, the bus line 5, 5/, ...

Concept	Description
Concepts for point of interests in student PoI matches dataset	
restaurant_GID-22077	a building where people go to eat
bar_GID-14628	a room or establishment where alcoholic drinks are served over a counter
university_GID-24619	establishment where a seat of higher learning is housed, including administrative and living quarters as well as facilities for research and teaching
supermarket_GID-23735	a large self-service grocery store selling groceries and dairy products and household goods
sport_facility_GID-18566	athletic facility equipped for sports or physical training
library_GID-19663	a depository built to contain books and other materials for reading and study
student_accommodation, dorm_GID-17139	a college or university building containing living quarters for students
hotel_GID-18979	a building where travelers can pay for lodging and meals and other services
pizzeria_GID-21345	a shop where pizzas are made and sold
sports_field_GID-45304	a piece of land prepared for playing a game
pharmacy_GID-17277	a retail shop where medicine and other articles are sold
bank_GID-14574	a building in which the business of banking transacted
dentist, dental_office_KGE24-0A-28	professional healthcare facility where dentists, dental hygienists, and other dental professionals provide oral healthcare services
train_station_GID-21898	terminal where trains load or unload passengers or goods
health_service_GID-21898	a medical institution where sick or injured people are given medical or surgical care
hairdresser_GID-22377	a shop where hairdressers and beauticians work
jewelry_store_GID-42955	a firm that sells and buys jewelry
monument_GID-45446	an important site that is marked and preserved as public property
museum_GID-20454	a depository for collecting and displaying objects having scientific or historical or artistic value
shop_GID-22782	a mercantile establishment for the retail sale of goods or services
ice_cream_shop_pastry_shop, milk_bar_GID-20240	snack bar that sells milk drinks and light refreshments (such as ice cream)
post_office_GID-110273	a public building in which mail is received, sorted and distributed
beer_shop_GID-23949	a building with a bar that is licensed to sell alcoholic drinks
stationery_bookstore_GID-15080	a shop where books are sold
bus_station_GID-15406	a terminal that serves bus passengers
parking_space_GID-45482	a space where an automobile can be parked
tobacco_shop_GID-24229	a shop that sells pipes and pipe tobacco and cigars and cigarettes
market_GID-44731	place designated for commercial exchanges, generally outdoors
hospital_GID-18969	a health facility where patients receive treatment
cabelcar_GID-24016	a transportation system in which cars (telphers) are suspended from cables and operated on electricity
grocery_store_GID-18499	a marketplace where groceries are sold
bedbreakfast_GID-14768	an overnight boardinghouse with breakfast
fast_food_GID-39991	inexpensive food (hamburgers or chicken or milkshakes) prepared and served quickly
appartamento	a suite of rooms usually on one floor of an apartment house
internet_point_KGE24-0A-35	a place in which customers pay to use computers to access the internet
souvenir_shop_KGE24-0A-36	a store primarily selling souvenirs, memorabilia, and other items relating to a particular topic or theme
clothin_store_KGE24-0A-37	a store that sells ready-made clothing

Concept	Description
bank_branch_KGE24-0A-38	a retail location where a bank, credit union, or other financial institution (including a brokerage firm) offers a wide array of face-to-face and automated services to its customers
typical_product_store_KGE24-0A-39	a store that sells typical products
ski_rental_KGE24-0A-40	a place where ski equipment is rented
electronics_store_KGE24-0A-41	the ideal place to shop for all types of consumer electronics, from televisions and sound systems to computers and gaming devices
perfumery_GID-21134	store where perfumes are sold
health_club_GID-18769	a place of business with equipment and facilities for exercising and improving physical fitness
shopping_mall_GID-21414	mercantile establishment consisting of a carefully landscaped complex of shops representing leading merchandisers; usually includes restaurants and a convenient parking area; a modern version of the traditional marketplace
swimming_pool_GID-23807	pool that provides a facility for swimming
camping_site_GID-45121	a site where people on holiday can pitch a tent
theatre_GID-24081	a building where theatrical performances or motion-picture shows can be presented
nightclub_GID-15442	a spot that is open late at night and that provides entertainment (as singers or dancers) as well as dancing and food and drink
ice_stadium_GID-22152	building that contains a surface for ice skating or roller skating
residence_GID-45262	any address at which you dwell more than temporarily
stationery_shop_KGE24-0A-42	a retail store that sells stationery, such as paper and paper products (including printing and engraving), postcards, and paper novelties
jewellry_store_KGE24-0A-43	a retail business establishment, that specializes in selling (and also buying) jewellery and watches
pet_store_KGE24-0A-44	an essential services retailer which sells animals and pet care resources to the public
farmhouse_GID-17664	house for a farmer and family
dairy_KGE24-0A-45	a farm where dairy products are produced
sports_store_KGE24-0A-46	a shop where sports clothes and equipment are sold
optical_store_KGE24-0A-47	a specialized retail establishment dedicated to providing various products and services related to vision care and eye health
toy_store_KGE24-0A-48	a specialized retail establishment dedicated to the provision of various products and services relating to vision care and eye health
Concepts for emotional states in SU2 dataset	
sad_GID-81576	experiencing or showing sorrow or unhappiness
down_GID-77954	low in spirits
neutral_KGE24-0A-24	feeling calm and measured emotion
happy_GID-80395	enjoying or showing or marked by joy or pleasure or good fortune
excited_GID-79092	in an aroused state

Concept	Description
Concepts for activities in SU2 dataset	
watching_GID-103268	look attentively
reading_GID-95695	interpret something that is written or printed
listening_to_GID-103365	hear with intention
sleeping_GID-73617	be asleep
studying_GID-30956	applying the mind to learning and understanding a subject (especially by reading)
eating_GID-4227	the act of consuming food
working_GID-2874	activity directed toward making or doing something
shopping_GID-303	searching for or buying goods or services
pause_GID-5384	temporary inactivity
social_activity_GID-5475	activity considered appropriate on social occasions
personal_care_KGE24-0A-21	the activity to support and supervision of daily personal living tasks and private hygiene
hobby_GID-2123	an auxiliary activity
housework_KGE24-0A-22	the tasks or activities for taking care of works around the house
coffee_break_GID-40097	a snack taken during a break in the work day
sport_GID-2593	an active diversion requiring physical exertion and competition
other_KGE24-0A-23	activities that are different or distinct from one already mentioned
lesson_GID-4491	a unit of instruction
movie_theater_GID-16033	a theater where films are shown
social_media_KGE24-0A-25	interactive technologies that facilitate the creation, sharing and aggregation of content among virtual communities and networks
en_route_GID-1415	the act of going from one place to another
on_the_phone_KGE24-0A-26	talking with somebody else using the telephone or other devices
in_chat_KGE24-0A-27	chatting or having a conversation with people using technologies for messaging
free_time_GID-72882	time available for hobbies and other activities that you enjoy
cigarette_GID-4210	the act of smoking tobacco or other substances
beer_GID-4255	the act of drinking alcoholic beverages to excess
physical_activity_GID-3158	the activity of exerting your muscles in various ways to keep fit
rest_GID-5391	freedom from activity (work or strain or responsibility)
nap_GID-4331	sleeping for a short period of time (usually not in bed)
fun_GID-2113	activities that are enjoyable or amusing
exhibit_GID-4451	the act of looking or seeing or observing

Table 5: Formalized concepts of data property values

Additionally, we exploded the served_days column into the corresponding seven concepts for a service being active on monday, tuesday, ... This change will help us streamline the next phase, allowing us to have meaningful data properties for the bus_route etype.

To define the concept in data property values for the concepts associated with the Points of Interest, we filtered out the most relevant places based on the locations visited by the students from the 9 user_poi_matches.json dataset.

No other changes are required to clean the collected dataset, and they can be used as is.

5 Knowledge Definition

In this phase, we produce the knowledge teleontology for the final knowledge graph by exploiting the concepts we defined in the Language definition phase.

5.0.1 Teleontology definition

To create our own teleology with entity types aligned to our purpose and competency questions, we take as reference the following three resources, also shown in Fig13;

- Schema.org version 28.1
- GTFS
- OSM teleontology

First of all, we identified from the collected resources the entities that aligned the most for our own, defined during the past phases, which are the following:

- schema:Action is used for stay_GID-5335. Unfortunately, we were unable to find any more specific entity that would map to our intended meaning.
- schema:Place is used for location_GID-695, tho we do not use it as intended, it is only a point in space where a person was on a certain time.
- schema:ReactAction used for mood_GID-39957 since we intend to use it as the emotional state which is produced by interacting with the environment
- schema:TravelAction used for trip_journey_GID-1474
- schema:Event used for event_GID_56
- schema:Person used for student_GID-53021
- gtfs:Route is used for bus_line_KGE24-0A-1
- gtfs:Stop and OSM:OSM place is used for bus_stop_GID-45118, since we are combining the properties from the gtfs files and the OSM id to be able to have a common reference connecting both PoIs and bus stops.
- gtfs:Trip is used for bus_route_GID-45117

Figure 13: Reference Ontologies: GTFS,

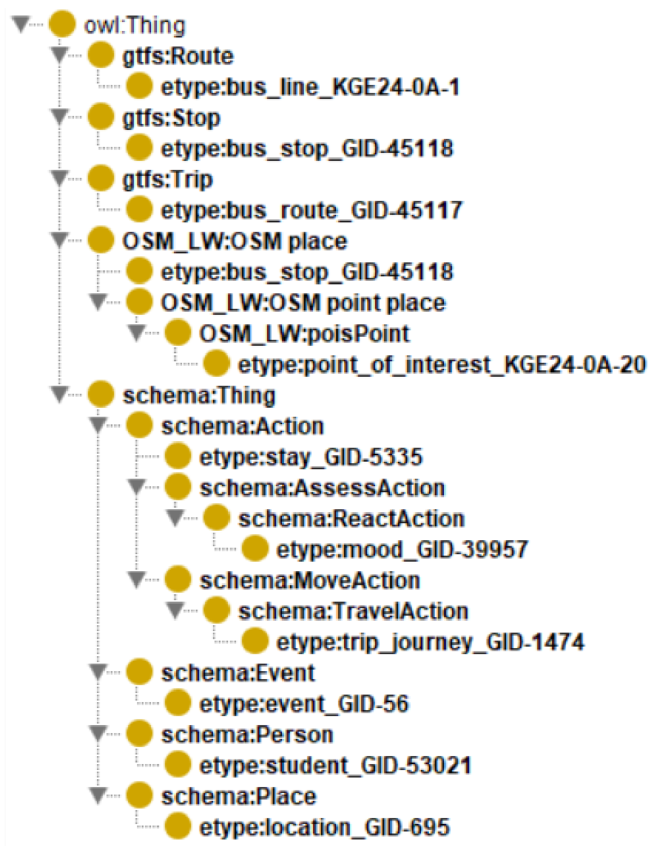

Figure 14: Produced Teleontology

- `OSM:poisPoint` is used for `point_of_interest_KGE24-0A-20`

From what we can see from the entities we aligned to the selected resources, why did we not use any equivalence? In some cases, we wanted our data to allow easy integration with many different sources, so we decided to not be bound to a single source. While others we just did not find the best match for our purpose.

An example of the first case is the bus stops, the main data source to be used to populate them are the gtfs files from the Trentino Trasporti agency we mentioned in the second phase. However, we also integrated into them the OSM identifiers; allowing us to tap into both knowledge bases. Indirectly, aligning all our spatial references to the OSM allows us to also be able to connect all of our streams to the knowledge contained in OSM. Skipping ahead, this will allow us in SPARQL to bridge multiple KGs that use the OSM ID to ground the places.

While for the second, we simply did not find a good match for the public schemes we searched. Especially schema modeling the emotional states of a person in general and not in response to web content, which we think is the intended use of the `ReactAction` class in the schema.org resource.

5.0.2 Teleology definition

To obtain the teleology, we need to choose only the leaf entity types from the teleontology we defined. Afterward, we leave the entity types that are general to their leaf entity types. For this case, we drop the entity type `Place` as it is a general to `Point_of_interest`. The resulting teleology is shown in the Figure15

In the end, we have the following entity types:

- `student_GID-53021`
- `mood_GID-39957`
- `event_GID_56`
- `point_of_interest_KGE24-0A-20`
- `stay_GID-5335`
- `location_GID-695`
- `trip_journey_GID-1474`
- `bus_line_KGE24-0A-1`
- `bus_stop_GID-45118`
- `bus_route_GID-45117`

With the following object properties:

- `has_board_GID-102601`: from `trip_journey_GID-1474` to `bus_stop_GID-45118` indicating the bus stop the student got on the bus.
- `has_feel_GID-101373`: from `student_GID-53021` to `mood_GID-39957` indicating the mood of the student.
- `has_get_off_GID-102589`: from `trip_journey_GID-1474` to `bus_stop_GID-45118` indicating the bus stop the student got off the bus.

- has_go_through_GID-102767: from bus_route_GID-45117 to bus_stop_GID-45118 indicating all the stops a bus takes in its route.
- has_inside_GID-106969: from stay_GID-5335 to point_of_interest_KGE24-0A-20 to tell where the student stayed at.
- has_located_GID-85982: from student_GID-53021 to location_GID-695 to store all the GPS history of the students.
- has_partecipate_in_GID-97811: from student_GID-53021 to event_GID_56 indicating the events a student attended to.
- has_serve_GID-97875: from bus_route_GID-45117 to bus_line_KGE24-0A-1, allowing us to tell the bus number associated to the route of interest
- has_stay_put_GID-101763: from student_GID-53021 to stay_GID-5335 associating the students to their stays at the PoIs.
- has_take_GID-101699: from student_GID-53021 to trip_journey_GID-1474 associating the students to their recorded bus trips.
- has_travel_by_KGE24-0A-50: from trip_journey_GID-1474 to bus_route_GID-45117 indicating which bus the student took for this trip.

And data properties:

- has_activity_GID-2006 defines the activity associated to event_GID_56.
- has_boarding_time_KGE24-0A-2 the time the student got on the bus, associated to trip_journey_GID-1474.
- has_bus_line_ID_KGE24-0A-17 the id originally contained in the GTFS file for the bus route, now renamed to bus line, associated to bus_line_KGE24-0A-1.
- has_bus_line_name_KGE24-0A-9 the name of the bus_line_KGE24-0A-1
- has_duration_GID-72859 the number of minutes a trip_journey_GID-1474 or stay_GID-5335 or mood_GID-39957 or event_GID_56 lasts for.
- has_end_date_KGE24-0A-4 the last served date of a bus_route_GID-45117.
- has_Event_ID_KGE24-0A-18 an extra ID associated to an event, can be used to match the same event being attended by multiple people.
- has_excluded_dates_KGE24-0A-6 the dates in which a bus_route_GID-45117 is not active.
- has_extra_dates_KGE24-0A-5 the dates in which a bus_route_GID-45117 is being served exceptionally.
- has_latitude_GID-45424 the latitude of a place or of a person at a certain time, associated to bus_stop_GID-45118 point_of_interest_KGE24-0A-20 and location_GID-695.
- has_longitude_GID-45429 the longitude of a place or of a person at a certain time, associated to bus_stop_GID-45118 point_of_interest_KGE24-0A-20 and location_GID-695.
- has_mood_GID-39957 the mood associated to mood_GID-39957.
- has_OSM_place_ID_KGE24-0A-19 the OSM id of a place in Trento, associated to point_of_interest_KGE24-0A-20 or bus_stop_GID-45118.
- has_PoI_ID_KGE24-0A-15 URI of point_of_interest_KGE24-0A-20.
- has_PoI_name_KGE24-0A-8 the common name associated to a point_of_interest_KGE24-0A-20.
- has_PoI_type_KGE24-0A-7 the type of PoI associated to point_of_interest_KGE24-0A-20.
- has_route_ID_KGE24-0A-16 the id originally contained in the GTFS file for the bus trip, now

renamed to bus route, associated to `bus_route_GID-45117`.

- `has_served_days_monday_KGE24-0A-12` tells whether the route is being served on Monday, associated to `bus_route_GID-45117`.
- `has_served_days_tuesday_KGE24-0A-30` tells whether the route is being served on Tuesday, associated to `bus_route_GID-45117`.
- `has_served_days_thursday_KGE24-0A-31` tells whether the route is being served on Thursday, associated to `bus_route_GID-45117`.
- `has_served_days_wednesday_KGE24-0A-15` tells whether the route is being served on Wednesday, associated to `bus_route_GID-45117`.
- `has_served_days_friday_KGE24-0A-32` tells whether the route is being served on Friday, associated to `bus_route_GID-45117`.
- `has_served_days_saturday_KGE24-0A-33` tells whether the route is being served on Saturday, associated to `bus_route_GID-45117`.
- `has_served_days_sunday_KGE24-0A-29` tells whether the route is being served on Sunday, associated to `bus_route_GID-45117`.
- `has_start_date_KGE24-0A-3` the first served date of a `bus_route_GID-45117`.
- `has_start_time_GID-73577`
- `has_stop_ID_KGE24-0A-14` the id originally contained in the GTFS file for the bus stop, associated to `bus_stop_GID-45118`.
- `has_stop_name_KGE24-0A-10` the common name of a `bus_stop_GID-45118`.
- `has_stop_order_KGE24-0A-9` the route and timetable of a `bus_route_GID-45117` through the different bus stops.
- `has_student_ID_KGE24-0A-13` the URI of the `student_GID-53021`.
- `has_timestamp_GID-27373` the timestamp at which something happens, associated to `trip_journey_GID-1474` or `stay_GID-5335` or `mood_GID-39957` or `event_GID_56` or `location_GID-695`.
- `has_timetable_GID-34211` the timetable associate to a `bus_stop_GID-45118`.
- `has_trip_headsign_KGE24-0A-11` The text displayed on top of a bus, associated to `bus_route_GID-45117`.

6 Entity Definition

In the last phase of the methodology, we aim to merge the knowledge and the data layers into a single structure, which is our final KG.

6.1 Entity Matching

In the final datasets that we produced, there was no single entity in the different datasets described by different set of properties because in the information gathering phase, we already matched the entities, as we modeled the teleontology considering the dataset we had and the datasets are already been aligned based on the entity types for the teleontology. For example, point of interests, we matched the places by their `osm_id`.

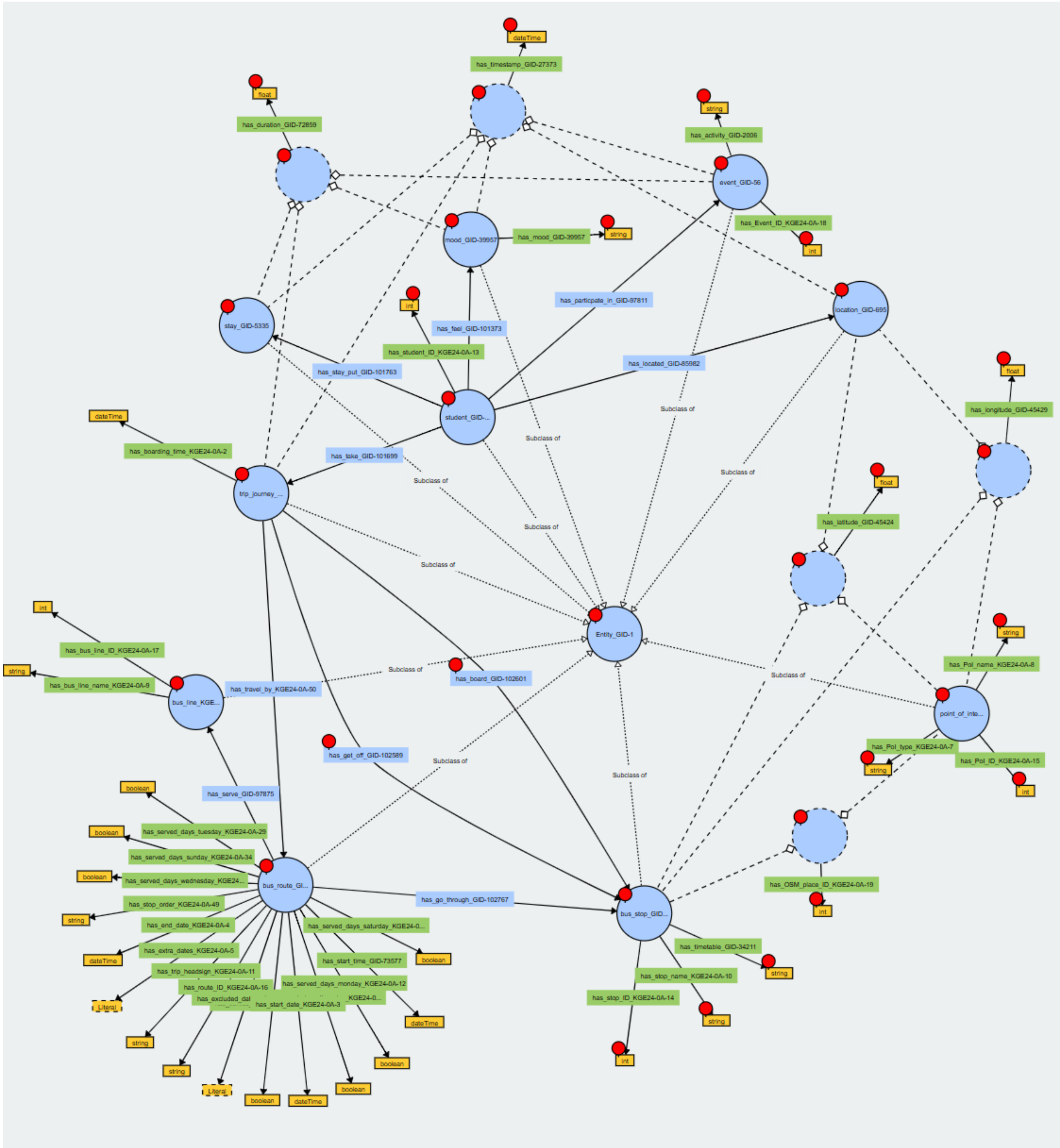

Figure 15: Teleology

6.2 Entity identification

To identify a single entity within the dataset we need to assign a unique identifier to each entity; in our case, we used the Uniform Resource Identifier (URI). All entities, except for the `Bus Stop`, `Bus Line`, `Student`, we used the Identifying Set, which is used to identify the entity using the composition of two or three values of the properties of the entities.

For example: The **Location** entity which represents the student's location stream is identified by the combination of : URI of the EType Location + URI of the EType Student + URI of the student entity, + value of the timestamp, resulting following unique URI: `location_GID-695_student_GID-53021_01526135431101`

6.3 Entity mapping

In entity mapping, we merge the values in the dataset based on teleontology which we produced in the previous phase. For this, we use Karma tool, an information integration tool for integrating data with knowledge schema. Karma maps selected data with the built ontologies or teleontologies. And all of the entity mapped models we produced are shown in Figure 16, 17, 18, 20, 21 22, 23, 24, 19,

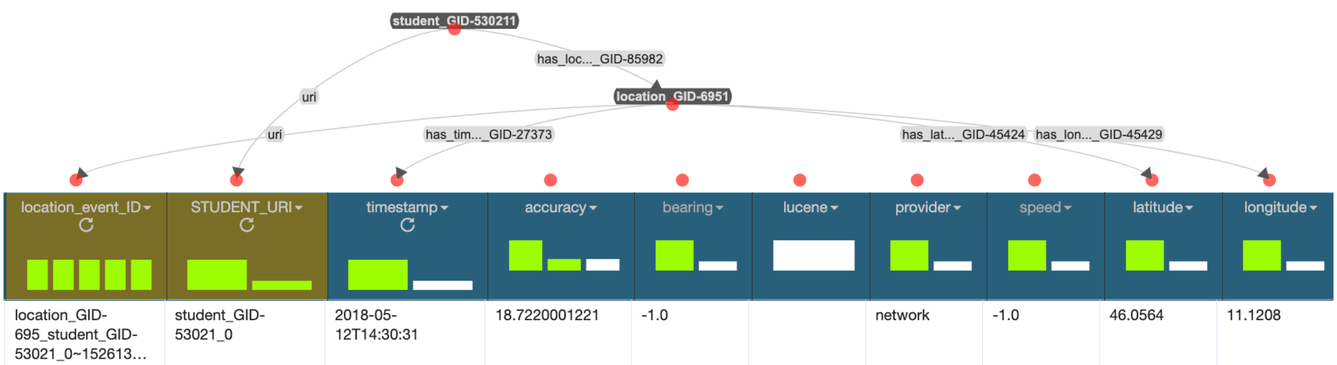

Figure 16: Location Stream Model

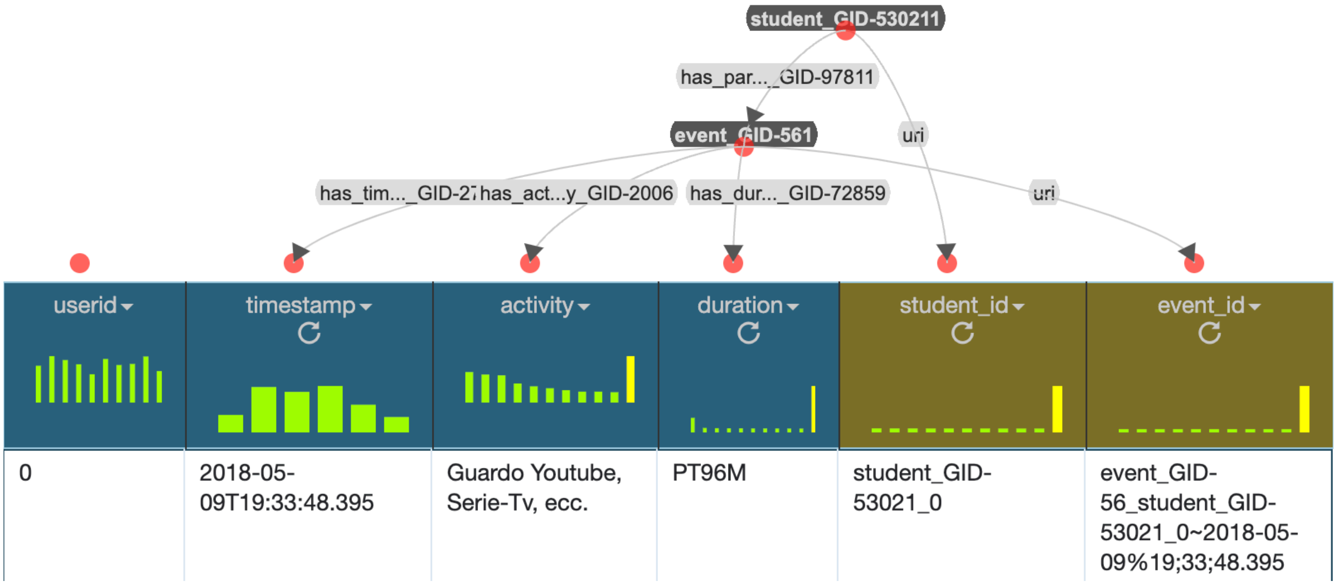

Figure 17: Activity Stream Model

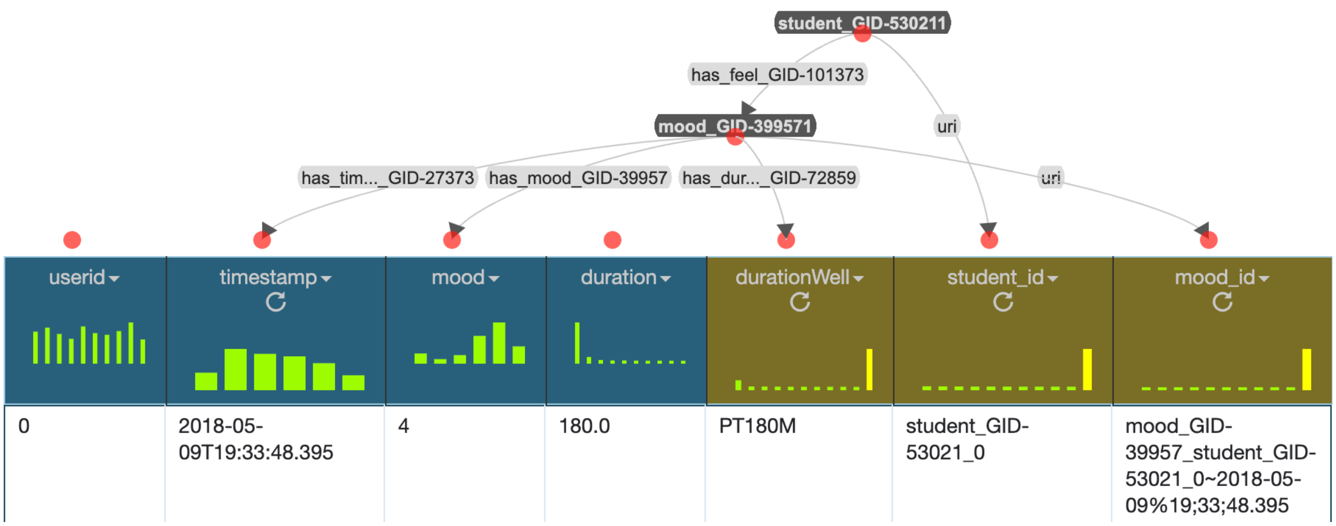

Figure 18: Mood Stream Model

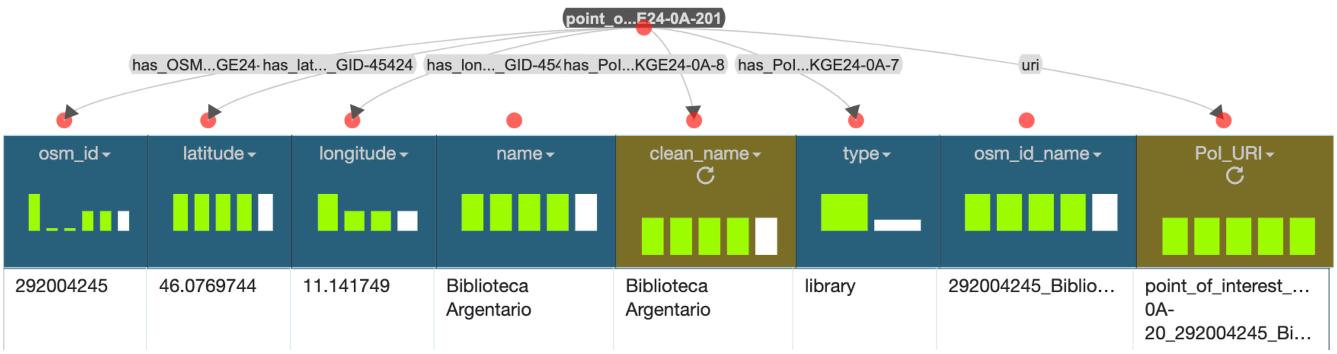

Figure 19: Point of Interest Model

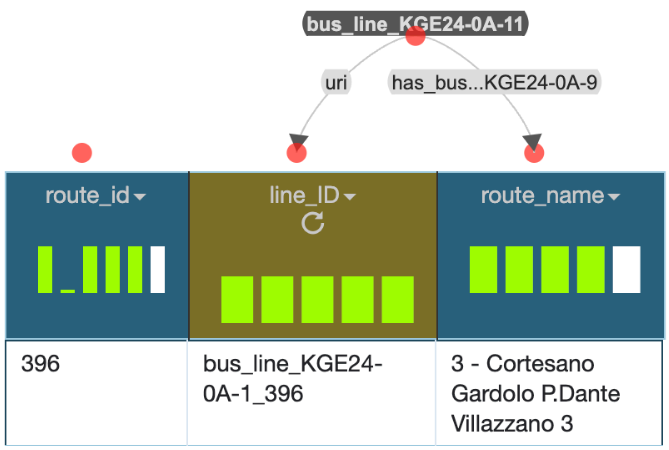

Figure 20: Bus Routes Model

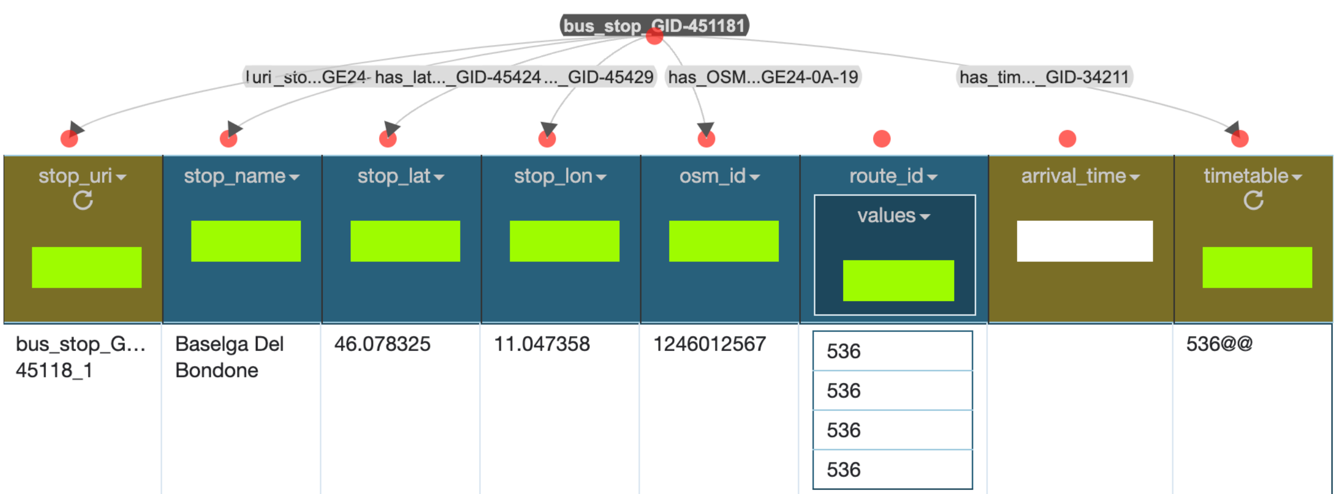

Figure 21: Bus Stops Model

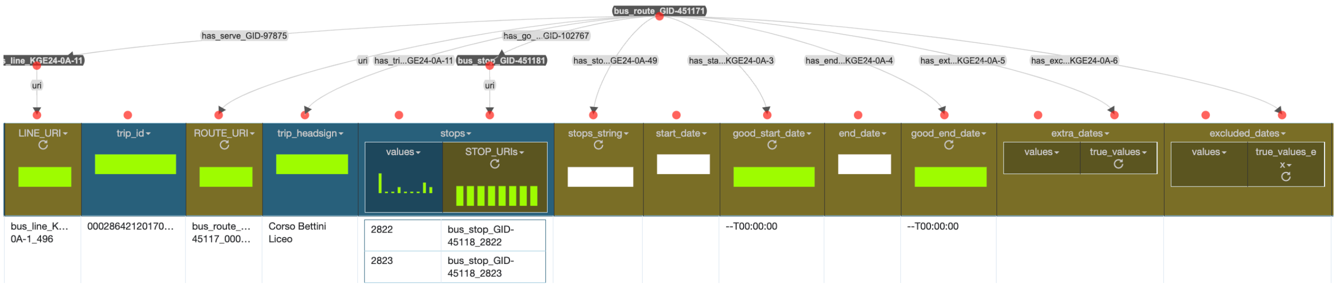

Figure 22: Bus Trips Model

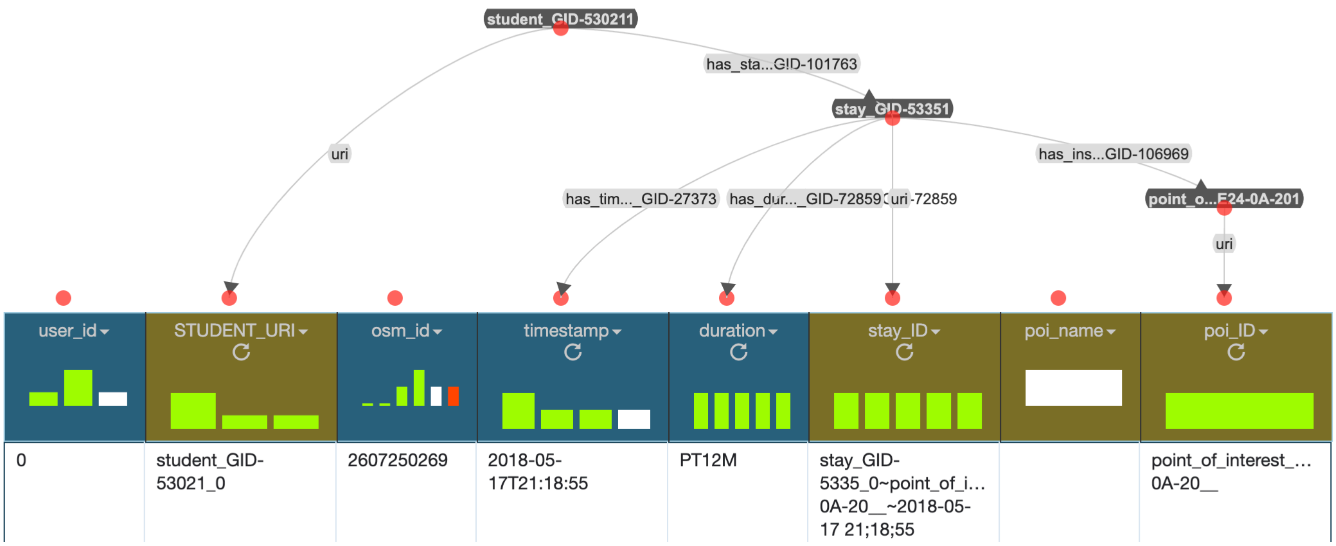

Figure 23: User PoI stays Model

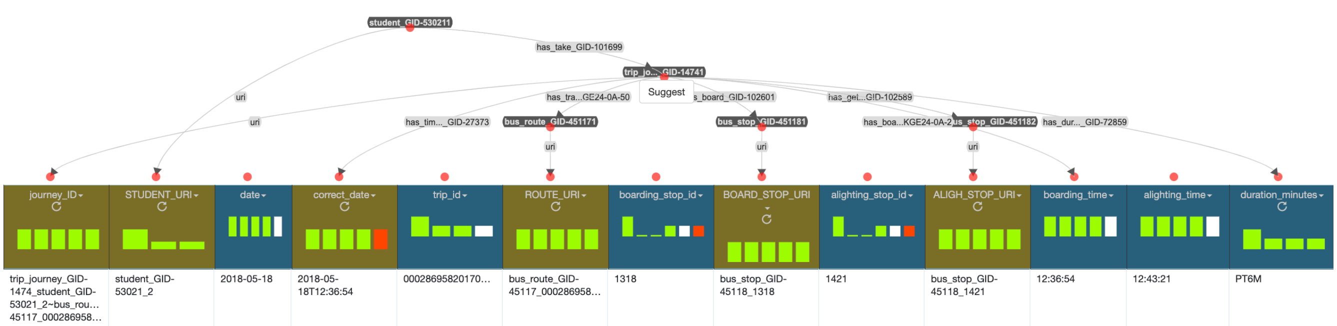

Figure 24: User Trips Model

7 Evaluation

In this section, we describe the evaluation of our final KG by reporting the assessment of the statistical information, such as, evaluation metric over the knowledge and data layer. We describe how our initial purpose can be satisfied by answering the competency questions using the KG we produced as well.

7.1 Knowledge Layer Evaluations

To evaluate the knowledge layer for the primary and secondary objectives which are purpose satisfaction and reusability is evaluated by the coverage metrics in Etype and data property levels.

7.1.1 Teleontology vs CQs - EType level

We extract 10 Etypes from the competency questions as seen in Table1

$$Cov_E(CQ_E) = \frac{|CQ_E \cap T_E|}{CQ_E} = \frac{10}{10} = 1 \tag{1}$$

7.1.2 Teleontology vs CQs - Property level

We extract 33 properties from the competency questions as seen in Table1 . In the teleontology, we defined the properties as according to the properties of Etypes we extracted from the competency questions.

$$Cov_p(CQ_p) = \frac{|CQ_p \cap T_p|}{CQ_p} = \frac{33}{33} = 1 \tag{2}$$

7.1.3 Teleontology vs Reference Ontologies (ROs) - EType level

gtfs 21, OSM 24 Schema +800

$$Cov_E(RO_E) = \frac{|RO_E \cap T_E|}{RO_E} = \frac{10}{843} = 0.01 \tag{3}$$

7.1.4 Teleontology vs Reference Ontologies (ROs) - Property level

$$Cov_p(RO_p) = \frac{|RO_p \cap T_p|}{RO_p} = \frac{33}{\infty} = 0 \tag{4}$$

7.2 Data layer evaluation

To evaluate the data layer for the objectives of the iTelos methodology, we will mostly use two metrics:

- Property Connectivity
- Entity Connectivity

However, the entity distribution was heavenly skewed to the locations stream, as that data is the most numerous. For this reason, we also included the number of entities and the reachability of other entities from our user entities (students) as a metric to make more sense of the results.

7.2.1 Number of Entities by Type

To answer this question we used the following SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?etype (COUNT(?entity) AS ?entityCount)
WHERE {
    ?entity rdf:type ?etype .
}
GROUP BY ?etype
```

Listing 1: query for number of entities by type

Which gave as a result table 6.

etype	entityCount
student_GID-53021	149
event_GID-56	45283
bus_line_KGE24-0A-1	41
bus_stop_GID-45118	1034
bus_route_GID-45117	2934
mood_GID-39957	23117
point_of_interest_KGE24-0A-20	2903
stay_GID-5335	3684
trip_journey_GID-1474	138
location_GID-695	3127734

Table 6: Number of entities by type.

As we can see from the table, our KG is heavily unbalanced toward the *location_GID-695* entity. This is to be expected as it was collected asynchronously from the students' devices. However, we have to consider this when looking at the next information as it might skew the results of the next evaluations regarding the students' connectivity to other entities.

7.2.2 Entity connectivity

The metric for the whole graph obtained is the following:

$$EC(KG) = \sum_{X=1}^{N} EC(X) = 640'781.37$$

However, this does not paint a clear picture of the information and connectivity in our graph. We need to look at the connectivity of each entity using the following formula:

$$EC(X) = \frac{\sum_{Y=1}^{N}(X,Y)}{OP(X)}$$

Where (X, Y) is a cell in the connectivity matrix, and OP(X) is the number of object properties of the ETypes X. To compute this metric we use the following SPARQL query:

```
PREFIX e: <http://knowdive.disi.unitn.it/etype#>

SELECT
 ?etype (SUM(?connectivityValue) / ?objectPropertyCount AS ?EC_X)
WHERE {
  # Calculate connectivity for each entity type X
  {
    SELECT
     ?etype (COUNT(DISTINCT ?reachableEntity) AS ?connectivityValue)
    WHERE {
     ?subject a ?etype .
     ?subject ?property ?reachableEntity .
     ?reachableEntity a ?etype2 .
     FILTER(isUri(?subject) && STRSTARTS(STR(?etype), STR(e:)))
     FILTER(isUri(?reachableEntity) &&
            STRSTARTS(STR(?etype2), STR(e:)))
    }
    GROUP BY ?etype ?property
  }
  # Count the number of object properties for EType X
  {
    SELECT
     ?etype (COUNT(DISTINCT ?property) AS ?objectPropertyCount)
    WHERE {
     ?subject a ?etype .
     ?subject ?property ?reachableEntity .
     ?reachableEntity a ?etype2 .
     FILTER(isUri(?subject) && STRSTARTS(STR(?etype), STR(e:)))
     FILTER(isUri(?reachableEntity) &&
            STRSTARTS(STR(?etype2), STR(e:)))
    }
    GROUP BY ?etype
  }
}
GROUP BY ?etype ?objectPropertyCount
ORDER BY DESC(?EC_X)
```

Listing 2: query for number of entities by type

Which gave us the results in table 7.

etype	Entity Connectivity
student_GID-53021	639991.2
event_GID-56	ND
bus_line_KGE24-0A-1	ND
bus_stop_GID-45118	ND
bus_route_GID-45117	537.5
mood_GID-39957	ND
point_of_interest_KGE24-0A-20	ND
stay_GID-5335	182
trip_journey_GID-1474	70.67
location_GID-695	ND

Table 7: SPARQL query results for entity connectivity.

From the table it seems the students are a highly connected entity. However, this number is heavily skewed thanks to the location information. If we remove it from the picture, our data is not as dense as we would like, especially if we look at the stays and trips captured in our KG. Unfortunately, this is the result of how we computed them. A more robust algorithm would greatly improve the expressivity of our final KG.

Yet, we still lack some information as to how well our most important entity, the students, is connected to the rest of the graph. To compute this metric we resorted to the following query:

```
PREFIX e: <http://knowdive.disi.unitn.it/etype#>

SELECT DISTINCT
 ?reachableEntityType
 (COUNT(DISTINCT ?reachableEntity) AS ?reachabilityCount)
WHERE {
  {
    # Step 1: Direct connections
    ?student a e:student_GID-53021 .
    ?student ?property1 ?reachableEntity .
    ?reachableEntity a ?reachableEntityType .
  }
  UNION
  {
    # Step 2: Indirect connections (2 hops)
    ?student a e:student_GID-53021 .
    ?student ?property1 ?intermediateEntity .
    ?intermediateEntity ?property2 ?reachableEntity .
    ?reachableEntity a ?reachableEntityType .
  }
```

```
}
GROUP BY ?reachableEntityType
ORDER BY DESC(?reachabilityCount)
```

Listing 3: query for number of entities by type

This query computes the 2-hop connectivity, allowing us to compute the connection from the student to all the other relevant distinct entities. Note: to compute the connectivity to the bus lines we included the 3-hops, however, we did not use this result as default as it wrongly increases the number of reachable bus stops to all the ones the bus trips go through. The results are listed in table 8.

etype	Student Connectivity	% of entities reached
event_GID-56	45283	100%
mood_GID-39957	23117	100%
stay_GID-5335	3684	100%
trip_journey_GID-1474	138	100%
location_GID-695	3127734	100%
point_of_interest_KGE24-0A-20	182	6.27%
bus_stop_GID-45118	82	7.93%
bus_route_GID-45117	112	3.81%
bus_line_KGE24-0A-1	20	48.78%

Table 8: SPARQL query results for entity connectivity.

We should focus our attention on the last four entities. And, from the percentage of reachable entities within the entire pool of possible ones, they indicate one of two possibilities: either students tend to lead less exploratory lives, frequently revisiting the same locations, or, our data preprocessing pipeline for computing stays and trips needs refinement. Both could be a plausible explanation, as from our experience students usually go mostly to university, supermarket, apartment, and repeat. However, we cannot exclude the other possibility.

7.2.3 Property connectivity

We finally compute the property connectivity using:

```
PREFIX e: <http://knowdive.disi.unitn.it/etype#>

SELECT
 ?etype (SUM(?dataPropertyValueCount) / ?dataPropertyCount AS ?PC_X)
WHERE {
  # Count the number of non-null data property values for (X, X)
  {
    SELECT
     ?etype ?dataProperty (COUNT(?value) AS ?dataPropertyValueCount)
    WHERE {
      ?entity a ?etype .
      ?entity ?dataProperty ?value .
```

```
      FILTER(!isUri(?value) && STRSTARTS(STR(?dataProperty),STR(e:)))
    }
    GROUP BY ?etype ?dataProperty
  }

  # Count the total number of data properties (DP(X))
  {
    SELECT
     ?etype (COUNT(DISTINCT ?dataProperty) AS ?dataPropertyCount)
    WHERE {
     ?entity a ?etype .
     ?entity ?dataProperty ?value .
     FILTER(!isUri(?value) && STRSTARTS(STR(?dataProperty),STR(e:)))
    }
    GROUP BY ?etype
  }
}
GROUP BY ?etype ?dataPropertyCount
ORDER BY DESC(?PC_X)
```

Listing 4: query for number of entities by type

Which resulted in what is reported in table 9.

etype	Property Connectivity
event_GID-56	45283
bus_line_KGE24-0A-1	41
bus_stop_GID-45118	1034
bus_route_GID-45117	2935.36
mood_GID-39957	23117
point_of_interest_KGE24-0A-20	2914.8
stay_GID-5335	3684
trip_journey_GID-1474	138
location_GID-695	3127734
TOTAL	3206881.16

Table 9: SPARQL query results for property connectivity.

As we expected, since we computed most of this data by ourselves, we do not have many null properties. The only ones we do not have data for is for some bus routes where extra or excluded dates were not present in the original GTFS files. We see that for the point of interest we get a number higher than the number of entities present in the KG. Upon further exploration, we discovered that this is the result of our choice in the URI for the PoIs, as we only used the OSM ID and the name, while in the data some of them were duplicated with a different PoI type. Using SPARQL we were able to identify 42/2903 of such entities. Unfortunately, due to time constraints, we were unable to go back and correct this fault. However, given the limited number of such entities, we believe that it does not interfere too much with our evaluation and does not result in any error when exploiting the KG.

7.3 KG Exploitation

From the finalized Entity Graph, we can answer the competency questions using SPARQL queries using GraphDB tool, for testing the exploitation of the final KG and its ability to satisfy the project purpose. The following is one example of a Competency question and how it could be answered using our KG.

1. What is the closest bus stop to reach the facility?

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX e: <http://knowdive.disi.unitn.it/etype#>

SELECT DISTINCT
        ?busStop
        ?poiName
        ?poiType
        ?stopPoiDistance
WHERE {
   # find the PoI
    {
        SELECT
                ?poiName ?poiType ?poiPoint
        WHERE {
            # accepted types
            VALUES ?searchType {
                "camposportivo"
                "piscina"
                "palestra"
                "impiantosciistico"
                "Campo sportivo (calcio, volley, basket, ...)"
                "Stadio del ghiaccio"
                "scuola sportiva"
                "piste da sci"
                "Palestra"
                "Piscina"
                "stadiosalto"
                "impiantosci"
            }
```

```
# 1) Get the user's most recent lat/long before ?departTime
{
    SELECT
        (?locTimestamp AS ?locTime)
        (?lat AS ?uLat)
        (?long AS ?uLong)
    WHERE {
        BIND("2018-05-15T15:00:00"^^xsd:dateTime AS ?deptime)
        BIND(<http://localhost:8080/source/student_GID-53021_1>
            AS ?usr)
        ?usr e:has_located_GID-85982 ?userLoc .
        ?userLoc
            e:has_latitude_GID-45424 ?lat ;
            e:has_longitude_GID-45429 ?long ;
            e:has_timestamp_GID-27373 ?locTimestamp .

        FILTER(?locTimestamp < ?deptime)
    }
    ORDER BY DESC(?locTimestamp)
    LIMIT 1
}

# 2) Get the POI lat/long filtered by type

?poi a e:point_of_interest_KGE24-0A-20;
    e:has_PoI_name_KGE24-0A-8 ?poiName;
    e:has_PoI_type_KGE24-0A-7 ?poiType;
    e:has_latitude_GID-45424 ?poiLat ;
    e:has_longitude_GID-45429 ?poiLong .

FILTER(LCASE(?poiType) = LCASE(?searchType))


# 3) Calculate distance using GeoSPARQL

    BIND(STRDT(CONCAT("POINT(", STR(?uLong), " ",
        STR(?uLat), ")"), geo:wktLiteral) AS ?userPoint)
    BIND(STRDT(CONCAT("POINT(", STR(?poiLong), " ",
        STR(?poiLat), ")"), geo:wktLiteral) AS ?poiPoint)
    BIND(geof:distance(?userPoint, ?poiPoint,
        <http://www.opengis.net/def/uom/OGC/1.0/metre>)
        AS ?poiDistance)
}
ORDER BY ?poiDistance
```

```
    LIMIT 1
    }

    # 4) Get all bus stops
    ?busStop a e:bus_stop_GID-45118 ;
            e:has_latitude_GID-45424 ?stopLat ;
            e:has_longitude_GID-45429 ?stopLong .

    BIND(STRDT(CONCAT("POINT(", STR(?stopLong), " ",
        STR(?stopLat), ")"), geo:wktLiteral) AS ?stopPoint)
    BIND(geof:distance(?stopPoint, ?poiPoint,
        <http://www.opengis.net/def/uom/OGC/1.0/metre>)
        AS ?stopPoiDistance)
}
ORDER BY ?stopPoiDistance
LIMIT 3

% \end{minted}
```

Listing 5: What is the closest bus stop to reach the facility?

The result of the query is shown in figure 25

	busStop	poiName	poiType	stopPoiDistance
1	http://localhost:8080/source/bus_stop_GID-45118_22	"Argentario"	"camposportivo"	"228.54477580305206"^^xsd:double
2	http://localhost:8080/source/bus_stop_GID-45118_23	"Argentario"	"camposportivo"	"241.24936836163323"^^xsd:double
3	http://localhost:8080/source/bus_stop_GID-45118_28	"Argentario"	"camposportivo"	"309.6257993138039"^^xsd:double

Figure 25: ER model

2. Which university facility best fits the student's needs or has the least impact on their mood? The SparQL query exploited the Point of Interest, student stays and mood stream RDFs, first by finding the facility related to universities from PoI, then filtering out the stay data that belong to the specific student and related to the PoI extracted before and extracting the mood data belong to the student stored around the time of the stays. In the result, university facilities that students have visited and expressed their mood are parsed, and the one with the highest mood - in terms of average mood value - is selected. The query is shown in the following:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX e: <http://knowdive.disi.unitn.it/etype#>

SELECT DISTINCT ?poiName ?poiType (AVG(?moodValue) AS ?moodInFacility)
WHERE {
```

```
    {
        # GET THE POI WE WANT
        {
            SELECT ?poi ?poiName ?poiType
            WHERE{
                VALUES ?types { "biblioteca"  "university"
"Biblioteca"  "library" }
                ?poi a e:point_of_interest_KGE24-0A-20;
                    e:has_PoI_name_KGE24-0A-8 ?poiName;
                    e:has_PoI_type_KGE24-0A-7 ?poiType.
                FILTER(LCASE(?poiType) = LCASE(?types))
            }
        }

        # GET THE STAYS FOR THE STUDENT
        {
            SELECT ?stay
            WHERE{
                BIND(<http://localhost:8080/source/student_GID-
                53021_55> AS ?user)
                ?user e:has_stay_put_GID-101763 ?stay.
            }
        }
        ?stay e:has_inside_GID-106969 ?poi;
            e:has_timestamp_GID-27373 ?stayStartTime;
            e:has_duration_GID-72859 ?stayDuration .
        BIND(?stayStartTime + ?stayDuration AS ?stayEndTime)

        # GET THE MOODS OF THE STUDENT AROUND THE
        {
            SELECT ?mood ?moodValue ?moodStartTime ?moodEndTime
            WHERE {
                BIND(<http://localhost:8080/source/student_GID
                -53021_55> AS ?user)

                ?user e:has_feel_GID-101373 ?mood.

                ?mood e:has_mood_GID-39957 ?moodVal ;
                    e:has_timestamp_GID-27373 ?moodStartTime ;
                    e:has_duration_GID-72859 ?moodDuration .
                BIND(xsd:float(?moodVal) AS ?moodValue)
                BIND(?moodStartTime + ?moodDuration AS ?moodEndTime)

                FILTER(?moodValue > 0)
```

```
                }
            }
                    # SELECT THE MOODS ACCORDING TO STAY
            FILTER(
                ((?moodStartTime <= ?stayEndTime)
                && (?moodEndTime >= ?stayStartTime))
            )
        }
}
GROUP BY ?poiName ?poiType
ORDER BY DESC(?moodInFacility)
LIMIT 1
```

Listing 6: Which university facility best fits the student's needs or has the least impact on their mood?

The result of the query is shown in figure 26

	poiName	poiType	moodInFacility
1	"Bup - Biblioteca Universitaria Povo"	"library"	"4.714286"^^xsd:float

Figure 26: ER model

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

8.1 Project Metadata

Table 10: Project Information: KGE 2024 - Student Life in Trento

Field	Description
Project Title	KGE 2024 project - Student Life in Trento
Project URL	NR
Keywords	KGE, knowledge graph, student, Trento, bus, PoI
Project Type	NR
Description	Build a knowledge graph that assists students in planning their trips from one location to another using public transportation in an efficient and comfortable manner. This tool aims to facilitate informed decision-making and enhance students' overall university experience. This will be achieved by integrating historical data on student commutes and activities, public transportation information, and points of interest.
Start Date	20-10-2024
End Date	13-01-2025
Funding Agency	University of Trento
Input Data Sources	SmartUnitn2 - Trentino Trasporti 2018 GTFS data - Points of Interest in Trento - Open Street Map
Outputs	Language, knowledge definitions - elaborated data in both JSON format - data mapping models and RDF formatted data
Coordinator	NR
Observations	The data for student trips and stay resulted sparse, limiting the scope of questions we can tackle.

8.2 Language Metadata

Table 11: Dataset Information: KGE 2024 - Language Resource

Field	Description
Data License	CC-BY-SA-4.0
Data URL	`https://github.com/yesunerdene9/KGE_student_commutes/blob/main/Phase%203%20-%20Language%20Definition/KGE24-UNITN%20-%20Language%20resource.xlsx`
Keywords	Concepts, UKC
Publisher	Davide, Yesun-Erdene
Creator	Davide, Yesun-Erdene
Owner	Davide, Yesun-Erdene
Language	English, Italian
Level	N/A
Size	79 KB
Name	Concepts definition for KGE 2024 project - Student Life in Trento
Publication Timestamp	26-12-2024
Description	Concept definition for data resources using UKC, including point of interest, smartUnitn2, and Trentino Trasporti.
Version	Version 1.0
Domain	N/A
File Format	XLSX

8.3 Knowledge Metadata

Table 12: Dataset Information: Teleontology and Teleology Files

Field	Description
Data License	CC-BY-SA-4.0
Data URL (Teleontology)	`https://github.com/yesunerdene9/KGE_student_commutes/blob/main/Phase%204%20-%20Knowledge%20Definition/Teleonthology.owx`
Keywords (Teleontology)	teleontology
Publisher (Teleontology)	Davide, Yesun-Erdene
Creator (Teleontology)	Davide, Yesun-Erdene
Owner (Teleontology)	Davide, Yesun-Erdene
Language	English
Level	N/A
Size (Teleontology)	1.627 KB
Name (Teleontology)	teleontology
Publication Timestamp (Teleontology)	19-12-2024
Description (Teleontology)	Contains the entities aligned with other publicly available knowledge resources.
Version	Version 1.0
Domain	N/A
File Format (Teleontology)	OWX
Data License	CC-BY-SA-4.0
Data URL (Teleology)	`https://github.com/yesunerdene9/KGE_student_commutes/blob/main/Phase%204%20-%20Knowledge%20Definition/teleology.ttl`
Keywords (Teleology)	teleology
Publisher (Teleology)	Davide, Yesun-Erdene
Creator (Teleology)	Davide, Yesun-Erdene
Owner (Teleology)	Davide, Yesun-Erdene
Size (Teleology)	20 KB
Name (Teleology)	teleology
Publication Timestamp (Teleology)	26-12-2024
Description (Teleology)	The Teleology file contains entity types, data properties, and object properties used in this project.
File Format (Teleology)	TTL

8.4 Data Metadata

Table 13: Dataset Information: KGE 2024 - RDF Models and Data

Field	Description
Data License	CC-BY-SA-4.0
Data URL (Models)	`https://github.com/yesunerdene9/KGE_student_commutes/tree/main/Phase%205%20-%20Entity%20Definition/models`
Keywords (Models)	RDF models, KGE
Publisher (Models)	Davide, Yesun-Erdene
Creator (Models)	Davide, Yesun-Erdene
Owner (Models)	Davide, Yesun-Erdene
Language	English
Level	N/A
Size (Models)	229 KB
Name (Models)	Data mapping for KGE 2024 project - Student Life in Trento
Publication Timestamp (Models)	08-01-2025
Description (Models)	Karma models to map JSON computed datasets into RDFs.
Version	Version 1.0
Domain	N/A
File Format (Models)	TTL
Data License	CC-BY-SA-4.0
Data URL (Data)	`https://github.com/yesunerdene9/KGE_student_commutes/tree/main/Phase%205%20-%20Data%20Definition/rdfs`
Keywords (Data)	RDF data, KGE
Publisher (Data)	Davide, Yesun-Erdene
Creator (Data)	Davide, Yesun-Erdene
Owner (Data)	Davide, Yesun-Erdene
Size (Data)	3.52 GB
Name (Data)	Knowledge graph data for KGE 2024 project - Student Life in Trento
Publication Timestamp (Data)	08-01-2025
Description (Data)	Published RDF data for KGE 2024 project - Student Life in Trento.
File Format (Data)	TTL

9 Open Issues

This section concludes the current document with conclusions regarding the quality of the process and outcome, and the description of the issues that (for lack of time or any other cause) remained open.
Overall, the resulting KG was able to answer our needs and we believe that the data we produced can be effectively re-used by exploiting the time dimension to align all our data streams.

It could however be improved in some areas, especially in its ability to tap into more information sources for what concerns the PoIs around Trento city. Our initial solution included the OSM place ID for each PoI so that more domain-relevant information could be pulled from other sources. Nonetheless, we still believe a more extensive knowledge modeling effort should be undertaken concerning the PoIs information. For example, by cleaning and standardizing the PoI types which cannot be properly used for querying the KG without knowing the full list. However, we did not feel that was the focus of our purpose. We preferred to focus our attention on how to use the students' information to compute useful knowledge, which could then be queried to answer our CQ.

Additionally, we should not have forced ourselves to use the UKC to map the concepts for the GTFS data, as this resulted in confusing and conflicting naming where routes are mapped to lines and trips to routes.

Focusing on the core aspects of our purpose, we think there is room for improvement in how the GPS data is exploited to compute student stays and trips. Better algorithms to gather this information would allow our KG to contain richer information, which could later be exploited to answer the queries relevant to our purpose.

Finally, as already mentioned in section [**?**], one relevant issue not addressed in the final delivered project is the use of an incorrect URI for the PoI entities, which resulted in a minor number of collisions. This only impacted a small number of entities, which we believe did not impact any of our evaluations in any relevant way. Nonetheless, this should be addressed and new RDFs should be published. However, as we already mentioned this part of the work should already receive a bigger modeling effort to increase the expressivity and alignment of the final KG to other knowledge resources.