Trentino Transportation and Education Facilities

University of Trento

Thomas Pasquali, Riccardo Peiretti

Department of Computer Science

January, 2024

Resources

The project resources are publicly and freely available at:

- **GitHub repository:** `https://github.com/ThomasPasquali/TransportationAndEducationFacilities_KGE2023`;
- **Drive:** `https://drive.google.com/drive/folders/1PUmhXYnM5fd-TqePUDAumem9luqa6qcv?usp=sharing`
- **Website:** `https://thomaspasquali.github.io/TransportationAndEducationFacilities_KGE2023/`

Domain of Interest

The *Domain of Interest* of this project consists of two boundaries: **space** and **time**.

- **Space:** Region of Trentino Alto Adige and Main Italian Cities (e.g. Milano, Bologna, Torino etc.);
- **Time:** The (main) temporal domain boundary is given by when Trentino Trasporti changes from winter to summer timetables.

Note: the final knowledge graph considers only the main public transportation itineraries that should never change.

Informal purpose

This KG purpose is to provide data to a web/smartphone application (e.g. UniTrentoApp) for, given the user weekly schedule, the current date and time: easily and conveniently gathering information about public transportation in order to organize the most convenient trips.
This purpose includes, both out-of-door and Trentinos students, teachers that rarely have to reach Trentino and commuters.

Informal Purpose

Build a Knowledge Graph (KG) which satisfies the many different need of students, professors and educational staff to reach their residence/domicile or middle/high school and university by public transportation within the region of Trentino or from the main Italian cities, depending on the specific person engagements, domicile, residence and date-time.

Personas and Scenarios

Personas

- Middle/High school students
- University students, both out-of-doors and locals
- Educational facilities staff e.g. professors, janitors etc.

Scenarios

- Weekdays
- Holidays e.g. Sunday, Christmas etc.
- Different time of the day e.g. 7AM, 4:30PM, 9PM etc.

Competency Questions

FTM - Middle school student

Mario lives in Mollaro and from Monday to Friday needs to go to the middle-school in Mezzolombardo from 8AM to 1PM, except on Thursday he needs to stay in school until 4PM. He only has a train subscription and he must be independent because his parents work all day.

Trento Urban (periphery) - High school student

Jessica lives in Povo and from Monday to Saturday she has to go to the high-school in the center of Trento. Her hours are from 8AM to 1PM.

Trento Urban (city center) - High school student (with wheelchair)

Luca lives in Trento and he goes to high-school from Monday to Saturday in the center of Trento from 8AM to 1PM, but he is in a wheelchair, and he must be independent using suitable and accessible buses.

Competency Questions cont.

Trento Extraurban (city center) - University student

Gian is a Rovereto university student who lives in Trento and he needs to reach the university from Monday to Wednesday from 10:30AM, and he needs to arrive at home before 4:30PM.

Multiple schedules and shifts - University student

Gaia lives at Sanbapolis. From Monday to Friday she needs to reach the University in Povo from 7:15AM to 1PM, but on Wednesday she goes to uni library in University of Sociology in Trento from 9:30AM to 10:30AM. However, on 2024-03-07 she will have a special lecture in University of Sociology in Trento at 12:30AM.

Competency Questions cont.

Flixbus Torino and Bologna - University student

Carla is an out-door-students and she lives in Lungadige in Trento. She return at home in Turin from Trento every Friday to stay at home in the weekend.

Trenitalia (with one transfer) - University professor (one time shift)

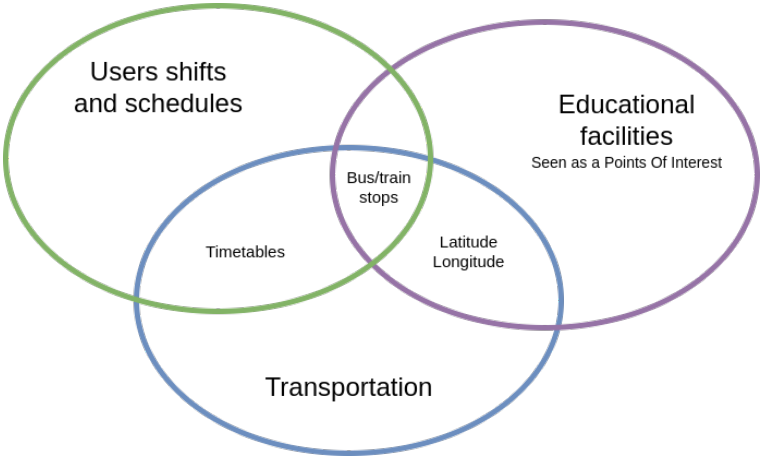
Fausto is a Professor at University of Milan, and he needs to book a ticket train from Milano Centrale to reach the University of Sociology on the 13st, Nov 2023 (Monday) before 10:30AM to have an important seminary.

Formal purpose

The informal version of the purpose can be formalized as the answer to questions like:

- "On day d at time t, which bus or train shall the person p take in order to reach its school or university in time?"
- "On day d at time t, which bus or train shall the person p take once finished his/hers duties to get back home?"
- "On day d at time t, which bus or train shall the person p take in order get to his/her real home?" (out-of-doors students)

Project Domains Intuition



Consumer activities

The following websites provided the high-quality data used for the project development:

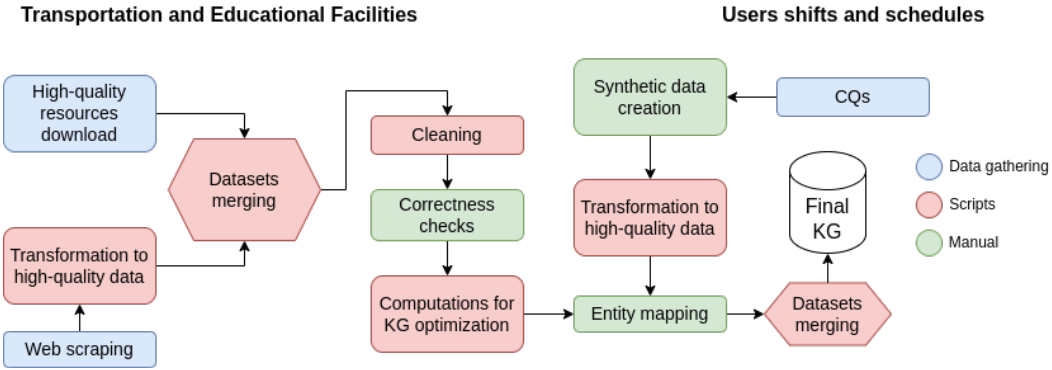
- GTFS datasets:
  - https://www.trentinotrasporti.it/open-data
  - https://www.dati.lombardia.it/Mobilit-e-trasporti/Orario-Ferroviario...
  - https://www.transit.land/feeds?search=trenitalia
  - https://transitfeeds.com/p/actv/630
  - https://www.transit.land/feeds/f-u-flixbus
- Educational facilities datasets:
  - https://www.tuttitalia.it/trentino-alto-adige/19-scuole/
  - https://webapps.unitn.it/du/it/StrutturaAccademica

Produces activities

High-quality data that was not available has been obtained via web scraping supported by manual correctness checks:

- E656 has been scraped to obtain missing GTFS data: Bolzano-Verona, Ferrovia Trento-Malè and Valsugana trains
- Wikipedia helped to fetch missing data, for example FTM stops and exact locations
- PDF timetables allowed to integrate missing data, for example the FTM trains *weekly schedules*
- Trenitalia has been used to check if E656 data was up-to-date

Process Intuition



Goal

Refine and formalize the language (concepts and words) employed for representing the information needed to fulfill the project purpose.
This activity has been divided in:

1. Language concepts for e-types;
2. Language concepts for e-types attributes i.e. data properties;
3. Language concepts for e-types relations i.e. object properties.

Methodology

Here is the list of steps that have been followed to formalize each concept/word that has been used during the project

1. **IF** reference ontologies already define the concept **THEN** adopt it **OTHERWISE**
2. **IF** KGE annotator has an existing concept that fits well **THEN** adopt it **OTHERWISE**
3. Use KGE annotator to create a new concept

Top-Down phase - Reference ontologies

This project uses GTFS as the main inspiration for:

- Transportation
- User schedules

They share *weekly schedule* and *schedule exception* e-types

The reference ontology has been downloaded from Datascientia LiveKnowledge.

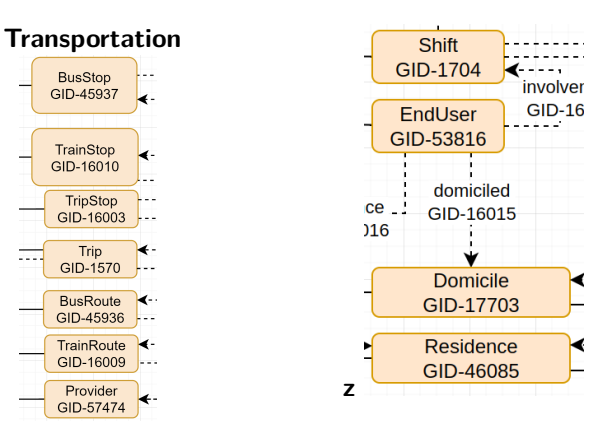
The rest of the users knowledge is really specific, therefore, an ad-hoc simple ontology has been created for the users.

Top-Down phase - Reference ontologies cont.

Since educational facilities are seen as Points Of Interest, only the purpose related slice of knowledge has adopted Schema.org:

- Thing > Place > CivicStructure > EducationalOrganization
- Thing > Intangible > StructuredValue > GeoCoordinates
- Thing > Place > Residence

Bottom-up phase - Teleology



Bottom-up phase - Teleology cont.

A summary of the final project's teleology entities and relations between them

Entities
BusStop
TrainStop
BusRoute
TrainRoute
TripStop
Trip
Provider
ScheduleException
WeeklySchedule
Shift

Entities	Relation	Entities
Bus/TrainRoute	operated	Provider
Trip	characterized	Bus/TrainRoute
TripStop	of	Trip
Bus/TrainStop	at	TripStop
Bus/TrainStop	localized	Position
Trip	avaiability schedule	WeeklySchedule
Trip	avaiability schedule exception	ScheduleException
Shift	occurence schedule	WeeklySchedule

(continues in the next slide)

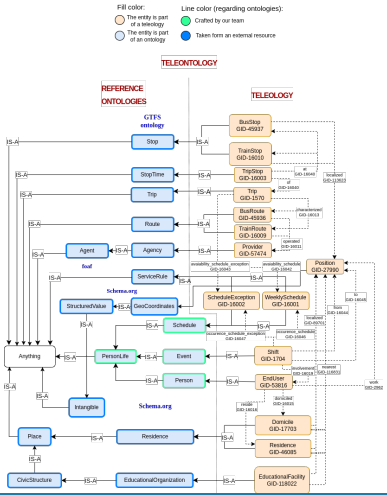
Bottom-up phase - Teleology cont.

Entities
EndUser
Domicile
Residence
EducationalFacility
Position

Entities	Relation	Entities
Shift	occurence schedule	ScheduleException
Shift	involvement	EndUser
Shift	from	Position
Shift	to	Position
EndUser	domiciled	Domicile
EndUser	reside	Residence
EndUser	work	Position
Domicile, Residence	localized	Position
Domicile, Residence	nearest	Position
EducationalFacility	localized	Position
EducationalFacility	nearest	Position

Middle-out phase - Teleontology

GitHub link



Entity Matching - First Phase

The entities, represented by their values, can be represented through different properties, and properties values. This is known as the entity matching problem with two main consequences:

1. Schema layer: find the right set of properties between the different datasets where multiple representations of the same entity can be present;
2. Data layer: the need to set the correct property values, if multiples representations share the same properties, but having different values.

In our case: **web scraping** + **ad-hoc parsing to GTFS**

Entity Identification - Second Phase

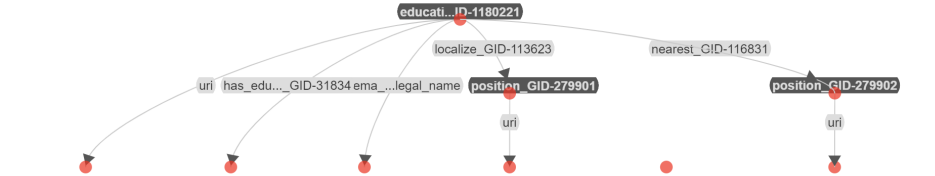
It is then necessary to identify an entity within a single dataset and adopt the same type of identification, if the same entity represented in two (or more) different ways, within different dataset. For instance:

1. The creation of an universal dataset Position containing address, latitude, and longitude for Educational Facilities, End User, and Stops;
2. The creation of unique column 'id' with the corresponding entity with conventions name like *pos_edu_fac*, *pos_user_dom*, *pos_user_res*, *pos_stop_bus*, *pos_stop_train*

Entity Mapping - Third Phase

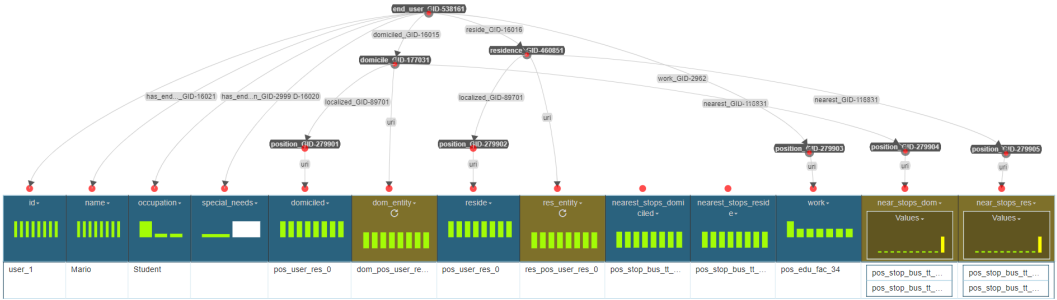
- Concretely merging the relative information values in the datasets;
- A specific type of mapping operation is performed to define the identifies for the entities to be considered the final KG.

Entity Mapping (example) - Education Facilities



id	type	legal_name	localize	nearest_stops	nearest_stops_split
edu_fac_TNMM84…	middleschool	Ala Cesare Battisti	pos_edu_fac_0	pos_stop_bus_tt_…	pos_stop_bus_tt_… pos_stop_bus_tt_… pos_stop_bus_tt_… pos_stop_bus_tt_… pos_stop_bus_tt_… pos_stop_bus_tt_…
edu_fac_TNMM81…	middleschool	Albiano	pos_edu_fac_1	pos_stop_bus_tt_…	pos_stop_bus_tt_…

Entity Mapping (example) - End User



Teleontology Size

To give an idea about the "size" of the final teleontology, a summary of the number of e-types, data and object properties:

	Instances Count
Etypes	33
Object Properties	16
Data Properties	33

Teleontology vs CQs

The following formula aims to evaluate how much the teleontology covers the Competency Questions:

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

	Etypes	Cov_E	Object Properties	Cov_{OP}
Total identified from CQs	13		15	
Total defined for the project	33	100%	16	100%

	Data Properties	Cov_{DP}
Total identified from CQs	28	
Total defined for the project	33	100%

Teleontology vs Reference Ontologies

The following formula aims to evaluate how much the teleontology reuses the reference ontologies:

$$Cov_E(RO_E) = \frac{|RO_E \cap T_E|}{|RO_E|}$$

GTFS ontology	Etypes	Cov_E	Object Properties	Cov_{OP}
Total in the ontology	30		19	
Total reused in the project	9	30%	0	0%

GTFS ontology	Data Properties	Cov_{DP}
Total in the ontology	22	
Total reused in the project	10	45%

From *Schema.org* have been reused 7 e-types and 4 data properties.

Connectivity Matrix

Since the connectivity matrix quite wide, here are link to the GtHub repository:

Connectivity Matrix:
https://github.com/ThomasPasquali/TransportationAndEducationFacilities_KGE2023/blob/main/Evaluation/connectivity_matrix.csv

Legend:
https://github.com/ThomasPasquali/TransportationAndEducationFacilities_KGE2023/blob/main/Evaluation/connectivity_matrix_legend.csv

Entities Count by E-types

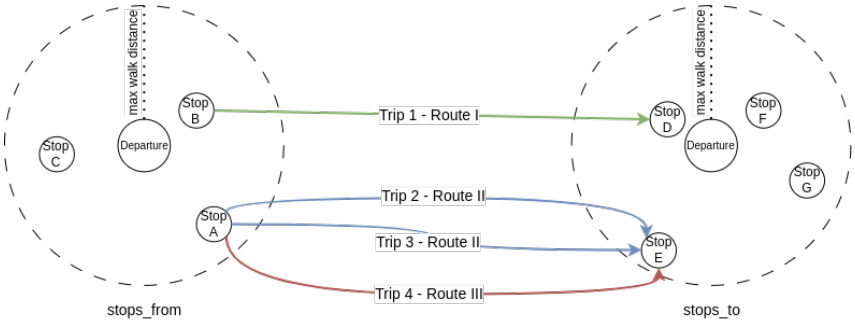
To conclude the KG evaluation, some information about the number of entities for each e-type:

Etype	Total number of entities
TripStop	509,214
ScheduleException	246,192
Trip	47,885
WeeklySchedule	41,571
Stop	9,273
Position	9,413
Route	1,347
EducationalFacility	139
Shift	19
EndUser	7
Residence	7
Domicile	7
Provider	3

CQs as SPARQL queries

All CQs share the same template which is available on GitHub (*template.sparql*).

Transportation query slice intuition:



CQ2 - Jessica - Trento suburbs

Jessica lives in Povo and from Monday to Saturday she has to go to the high-school in the center of Trento. Her hours are from 8AM to 1PM.

```
BIND ("false"^^xsd:boolean as ?wheelchair_need)
BIND ("45"^^xsd:long as ?max_wait_time)
# BIND ("0.5"^^xsd:float as ?max_stop_dist)
BIND ("_urban" as ?stops_filter)

BIND ("1" as ?direction) # For outward while "0" for return

BIND ("2024-01-29T06:00:00"^^xsd:dateTime as ?curr_datetime)
# For outward while "2024-01-29T12:00:00" for return

BIND ("Jessica" as ?username)
```

CQ2 - Jessica - Trento suburbs - Results outward

	shift	stop_from_name	stop_to_name	route_name	departure_time	arrival_time	wait_time	walk_from_m	walk_to_m	trip_duration	score	accessibility
1	data:shift_5	"Povo Piazza Manci"	"Venezia Cave"	"5"	"07:37:00"	"07:44:00"	"16"^^xsd:long	"12"^^xsd:int	"24"^^xsd:int	"7"^^xsd:long	"41.0"^^xsd:float	"1"
2	data:shift_5	"Povo "Centro Civico""	"Venezia Cave"	"5"	"07:38:00"	"07:44:00"	"16"^^xsd:long	"30"^^xsd:int	"24"^^xsd:int	"6"^^xsd:long	"49.0"^^xsd:float	"1"
3	data:shift_5	"Povo "Polo Sociale""	"Venezia Cave"	"5"	"07:36:00"	"07:44:00"	"16"^^xsd:long	"37"^^xsd:int	"24"^^xsd:int	"8"^^xsd:long	"54.5"^^xsd:float	"1"
4	data:shift_5	"Povo Piazza Manci"	"Venezia Cave"	"5"	"07:17:00"	"07:24:00"	"36"^^xsd:long	"12"^^xsd:int	"24"^^xsd:int	"7"^^xsd:long	"61.0"^^xsd:float	"1"
5	data:shift_5	"Povo Piazza Manci"	"Venezia Cave"	"5"	"07:14:00"	"07:20:00"	"40"^^xsd:long	"12"^^xsd:int	"24"^^xsd:int	"6"^^xsd:long	"64.0"^^xsd:float	"1"
6	data:shift_5	"Povo "Centro Civico""	"Venezia Cave"	"5"	"07:18:00"	"07:24:00"	"36"^^xsd:long	"30"^^xsd:int	"24"^^xsd:int	"6"^^xsd:long	"69.0"^^xsd:float	"1"
7	data:shift_5	"Povo PantÃ "	"Venezia Cave"	"5"	"07:15:00"	"07:20:00"	"40"^^xsd:long	"34"^^xsd:int	"24"^^xsd:int	"5"^^xsd:long	"74.0"^^xsd:float	"1"
8	data:shift_5	"Povo "Polo Sociale""	"Venezia Cave"	"5"	"07:16:00"	"07:24:00"	"36"^^xsd:long	"37"^^xsd:int	"24"^^xsd:int	"8"^^xsd:long	"74.5"^^xsd:float	"1"
9	data:shift_5	"Povo "Polo Sociale""	"Venezia Cave"	"5"	"07:13:00"	"07:20:00"	"40"^^xsd:long	"37"^^xsd:int	"24"^^xsd:int	"7"^^xsd:long	"77.5"^^xsd:float	"1"

CQ2 - Jessica - Trento suburbs - Results return

	shift	stop_from_name	stop_to_name	route_name	departure_time	arrival_time	wait_time	walk_from_m	walk_to_m	trip_duration	score	accessibility
1	data:shift_6	"Venezia Cave"	"Povo Piazza Manci"	"5"	"13:14:00"	"13:22:00"	"14"^^xsd:long	"18"^^xsd:int	"13"^^xsd:int	"8"^^xsd:long	"37.5"^^xsd:float	"1"
2	data:shift_6	"Venezia Cave"	"Povo Piazza Manci"	"5"	"13:17:00"	"13:26:00"	"17"^^xsd:long	"18"^^xsd:int	"13"^^xsd:int	"9"^^xsd:long	"41.5"^^xsd:float	"1"
3	data:shift_6	"S.Francesco Porta Nuova"	"Povo Piazza Manci"	"5"	"13:13:00"	"13:22:00"	"13"^^xsd:long	"29"^^xsd:int	"13"^^xsd:int	"9"^^xsd:long	"43.0"^^xsd:float	"1"
4	data:shift_6	"Venezia Cave"	"Povo S.Agata"	"5"	"13:14:00"	"13:22:00"	"14"^^xsd:long	"18"^^xsd:int	"28"^^xsd:int	"8"^^xsd:long	"45.0"^^xsd:float	"1"
5	data:shift_6	"Venezia Cave"	"Povo PantÃ™"	"5"	"13:14:00"	"13:21:00"	"14"^^xsd:long	"18"^^xsd:int	"32"^^xsd:int	"7"^^xsd:long	"46.0"^^xsd:float	"1"
6	data:shift_6	"S.Francesco Porta Nuova"	"Povo Piazza Manci"	"5"	"13:16:00"	"13:26:00"	"16"^^xsd:long	"29"^^xsd:int	"13"^^xsd:int	"10"^^xsd:long	"47.0"^^xsd:float	"1"
7	data:shift_6	"Venezia Cave"	"Povo "Centro Civico""	"5"	"13:17:00"	"13:25:00"	"17"^^xsd:long	"18"^^xsd:int	"27"^^xsd:int	"8"^^xsd:long	"47.5"^^xsd:float	"1"

CQ6 - Carla - Flixbus (Torino)

Carla is an out-door-students and she lives in Lungadige in Trento. She return at home in Turin from Trento every Friday to stay at home in the weekend.

```
BIND ("false"^^xsd:boolean as ?wheelchair_need)
# BIND ("60"^^xsd:long as ?max_wait_time)
# BIND ("2"^^xsd:float as ?max_stop_dist)
BIND ("flixbus" as ?stops_filter)
BIND ("1" as ?direction)

BIND ("2023-11-03T14:00:00"^^xsd:dateTime as ?curr_datetime) # A friday

BIND ("Carla" as ?username)
```

	shift	stop_from_name	stop_to_name	route_name	departure_time	arrival_time	wait_time	walk_from_m	walk_to_m	trip_duration	score	accessibility
1	data:shift_17	"Trento (Lungadige Monte Grappa)"	"Turin (Vittorio Emanuele)"	"FlixBus 407"	"15:35:00"	"21:10:00"	"155"^^xsd:long	"136"^^xsd:int	"60"^^xsd:int	"335"^^xsd:long	"588.0"^^xsd:float	"1"

CQ7 - Fausto - Trenitalia (transfer)

Fausto is a Professor at University of Milan, and he needs to book a ticket train from Milano Centrale to reach the University of Sociology on the 13st, Nov 2023 (Monday) before 10:30AM to have an important seminary.

```
BIND ("false"^^xsd:boolean as ?wheelchair_need)
BIND ("90"^^xsd:long as ?max_wait_time)
BIND ("0.2"^^xsd:float as ?max_stop_dist)
BIND ("trenitalia" as ?stops_filter)
BIND ("0" as ?direction)

BIND ("2023-11-13T04:00:00"^^xsd:dateTime as ?curr_datetime) # For the first trip
# "2023-11-13T08:00:00 for the second trip

BIND ("Fausto" as ?username)

...

# For the first trip:
## Using "nearest" COMMENTED
# ?edu_res_dom_to prop:localized_GID-89701 ?shift_pos_to.
# ?edu_res_dom_to prop:nearest_GID-116831 ?stop_pos_to.

## UNCOMMENTED
{ ## OR using geo distance
    {
        { ?stop_to a prop:bus_stop_GID-45937. }
        UNION
        { ?stop_to a prop:train_stop_GID-16010. }
    }
}

# For the second trip the previous has been done for ?stop_from
```

CQ7 - Fausto - Trenitalia (transfer) - Results

Milano-Verona

	shift	stop_from_name	stop_to_name	route_name	departure_time	arrival_time	wait_time	walk_from_m	walk_to_m	trip_duration	score	accessibility	trip_ex	trip_ws	trip
1	data:shift_18	"MILANO CENTRALE"	"VERONA PORTA NUOVA"	"RE6"	"06:25:00"	"08:17:00"	"25"^^xsd:long	"19"^^xsd:int	"0"^^xsd:int	"112"^^xsd:long	"146.5"^^xsd:float	"NaN"		data:ws_trenitalia/lombardia_249472015856862	data:trip_trenitalia/lombardia_15856862
2	data:shift_18	"MILANO CENTRALE"	"VERONA PORTA NUOVA"	"RE6"	"07:25:00"	"09:17:00"	"85"^^xsd:long	"19"^^xsd:int	"0"^^xsd:int	"112"^^xsd:long	"206.5"^^xsd:float	"NaN"		data:ws_trenitalia/lombardia_249253515856004	data:trip_trenitalia/lombardia_15856004

Verona-Trento

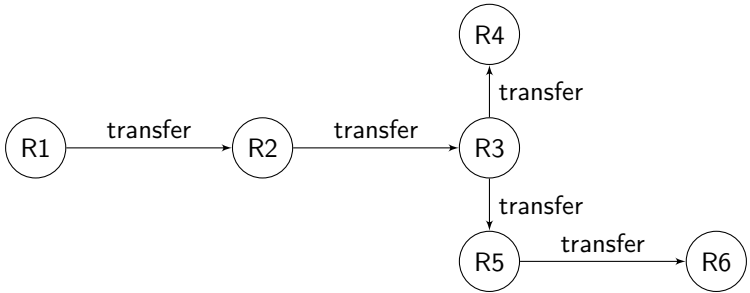
	shift	stop_from	stop_to	route_name	departure_time	arrival_time	wait_time	walk_from_m	walk_to_m	trip_duration	score	accessibility	trip_ex	trip_ws	trip
1	data:shift_19	data:stop_train_scraped_2173	data:stop_train_scraped_2264	"Verona-Bolzano"	"09:06:00"	"10:30:00"	"0"^^xsd:long	"3"^^xsd:int	"49"^^xsd:int	"84"^^xsd:long	"110.0"^^xsd:float	"1"		data:ws_scraped_543210001	data:trip_scraped_16678
2	data:shift_19	data:stop_train_scraped_2173	data:stop_train_scraped_2264	"Verona-Bolzano"	"08:50:00"	"09:50:00"	"40"^^xsd:long	"3"^^xsd:int	"49"^^xsd:int	"60"^^xsd:long	"126.0"^^xsd:float	"1"		data:ws_scraped_543210001	data:trip_scraped_16638

Future Work

- **Query optimization:** some ideas on how to further optimize queries:
  - Divide the KG in sub-graphs that only cover specific regions for transportation;
  - Add indexes to the KG.
- **Raise scraped data quality:** either by creating better scripts or finding available datasets.
- **Update datasets:** for example Trenitalia Lombardia trains.
- **Better data filtering:** for instance, Flixbus non Italian trips and stops.
- **Find or create a better ontology for users.**

Future Work cont.

- **Handle shifts that require more than one trip:**



This sub-graph would allow queries like: ?from prop:transfer{3} ?to

For instance, if ?from has the value R1, ?to will get the values {R4, R5}.

Conclusion

iTelos methodology pros

- The incrementality of the methodology allows to focus on each phase, one at the time;
- Revisiting previous phases is fundamental;
- Formalizing the language really helps.

Criticalities

- During some phases of this project, not knowing what comes next, may lead to wrong or non-optimal choices;
- Some datasets may not be available e.g. trains Bolzano-Verona and Veneto.