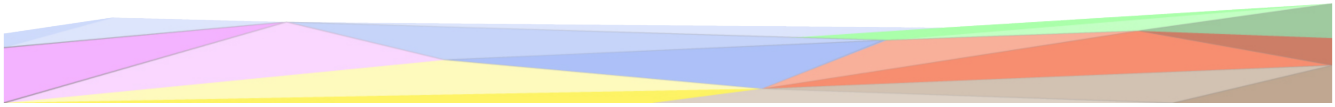
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2022 - Trentino Sports

Document Data:	Reference Persons:
December 20, 2022	Erik Nielsen, Shandy Darma



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.0	2.11.2022	Erik Nielsen, Shandy Darma	Document created
0.1	10.11.2022	Erik Nielsen, Shandy Darma	Purpose Formalization steps formalized
0.2	10.11.2022	Shandy Darma	Introduction, Purpose and Domain of Interest, Data Sources written
1.0	12.11.2022	Erik Nielsen	Inception phase written
2.1	22.11.2022	Shandy Darma	Added descripion Informal Modelling
2.2	22.11.2022	Erik Nielsen	Added ER description and ER Diagram
2.3	23.11.2022	Erik Nielsen	Added Datasets Filtering and Alignment description
2.3	23.11.2022	Erik Nielsen	Added Some Scrape Info in the Inception
2.4	23.11.2022	Shandy Darma	Added Teleology Building
3.0	5.12.2022	Erik Nielsen	Added Introduction Formal Modelling
3.1	5.12.2022	Shandy Darma, Erik Nielsen	Added Schema Alignment
3.2	7.12.2022	Shandy Darma, Erik Nielsen	Fixed Data Management and Open Issues
4.0	15.12.2022	Erik Nielsen	Added KGC, data mapping, entity matching, KG
5.0	17.12.2022	Shandy Darma, Erik Nielsen	Added Exploitation chapter
5.1	17.12.2022	Shandy Darma	Fixed KG evaluation
6.0	17.12.2022	Erik Nielsen	Conclusions
7.0	19.12.2022	Shandy Darma, Erik Nielsen	General Revision

1 Introduction

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important role in order to enhance the reusability of the resources handled and produced during the process. A clear description of the resources and the process developed, provides a clear understanding of the KGE project, thus serving such an information to external readers in order to exploit that in new projects.

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

- Section 2: The project's purpose and the domain of interest and the resources involved (both schema and data resources) in the integration process.
- Section 3: The input resources considered by the KGE project.
- Section 4, 5, 6, 7: The integration process along the different iTelos phases, respectively.
- Section 8: How the result of the KGE process (the KG) can be exploited.
- Section 9: Conclusions and open issues summary.

2 Purpose and Domain of Interest (DoI)

In this section, we are going to describe the purpose and the Domain of Interest (DoI) of this project.

2.1 Purpose

To build a knowledge graph, we first need to define a purpose of the knowledge graph. A purpose is an informal specification of the problem which a user aims to solve. A purpose is necessary to define a boundary for the knowledge graph project. With this in mind, the purpose of this project is as follows:

A service which help the users to find information about sport facilities in Trentino.

2.2 Domain of Interest

After confirming the project's purpose, we need to define the project's Domain of Interest (DoI). The DoI of this project consists of two boundaries: space boundary and time boundary. The space boundary of this project is the Autonomous Province of Trento, while the time boundary of this project is from June 2022 to June 2023.

3 Data Sources

In this section, we are going to describe the knowledge sources and the data sources that we consider for this project.

3.1 Knowledge Sources

The reference schemata that we use for this project are as follows:

- BBC Sport Ontology. This is a simple ontology, provided by BBC, for publishing data about competitive sports events.
- General Transit Feed Specification (GTFS) Static.
- Internet Calendaring and Scheduling Core Object Specification (iCalendar).
- INSPIRE.
- Data Catalog Vocabulary (DCAT) Version 2.

Moreover, the metadata of these reference schemata are as follow:

Title	Description	URL
BBC Sport Ontology	A simple ontology for representing competitive sports events.	`https://lov.linkeddata.es/dataset/lov/vocabs/sport`
GTFS Static	GTFS Static defines a common format for public transportation schedules and associated geographic information.	`https://developers.google.com/transit/gtfs`
iCalendar	iCalendar is a data format for representing and exchanging calendaring and scheduling information such as events, to-dos, journal entries, and free/busy information, independent of any particular calendar service or protocol.	`https://www.rfc-editor.org/rfc/rfc5545`
INSPIRE	A schema for modelling space.	`https://drive.google.com/file/d/1oFYjzx6uuV0p7ZZrXElAygOV-ODZv211/view`
DCAT Version 2	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web.	`https://www.w3.org/TR/vocab-dcat-2`

3.2 Data Sources

The data we use for this project originates from these sources:

- Localizzazione Impianti Sportivi, provided by Comune di Trento, hosted in Open Data Trentino.
- Sport Magazine Trentino.
- Trentino Transporti Opendata.

In addition, the metadata of these data sources are as follow:

Title	Description	URL
Localizzazione Impianti Sportivi	A list of sport facilities in the municipality of Trento.	https://dati.trentino.it/dataset/localizzazione-impianti-sportivi-open-data
Sport Magazine Trentino	Sport magazine that covers the region of Trentino.	https://sportmagazinetrentino.it
Trentino Transporti Opendata	Urban and suburban public transport data in GTFS format. Main information available: stop records (georeferenced), list of lines, routes and arrival and departure times.	https://www.trentinotrasporti.it/open-data
Impianti Sportivi	A list of sport facilities in the municipality of Trento.	https://www.comune.trento.it/Aree-tematiche/Sport/Impianti-sportivi
Centri Sportivi	A list of sport facilities in the municipality of Arco.	https://www.visitvaldisole.it/it/guida-a-z/centri-sportivi

4 Purpose Formalization

In this section, we define the purpose formalization into different sections:

- **Scenarios**: a set of usage scenarios, describing examples of usage of the Purpose.
- **Personas**: a set of real users examples acting within the scenarios defined above. Each Persona is defined over a specific features included in the main Purpose.
- **Competency Questions (CQs)**: the list of CQs created considering the personas in the scenarios defined.
- **Entities identified**: the terms representing the entities to consider in the KGE project, classified using the popularity categories.

4.1 Scenarios

We define scenarios as, within the constraint of the Domain of Interest, description of the possible use cases of the knowledge graph. As such, we created the scenarios as follow:

- Scenario 1: In the municipality of Trento, between 08:00 - 22:00, standard weekday.
- Scenario 2: In the municipality of Trento, between 08:00 - 22:00, one week before the Trento Half Marathon event.
- Scenario 3: In the municipality of Trento, during the evening, Trentino Valley home game day.

4.2 Personas

We define personas as a set of actors which have specific needs within the constraints of this knowledge graph. As such, we created the personas as follow:

- Paolo is 21 years old and a new student in the University of Trento. He is passionate about climbing. In his hometown, he usually practices three times a week.
- Sally is 17 years old and a long distance athletic runner who lives in Villazzano. She mainly focuses on 1500 m and 5000 m races. She goes to the athletic pitch twice a week at 18 o'clock with her club.
- Mario is a 35 years old father. He has a son called Guilio, who is 12 years old and loves to do Judo. Both Mario and Giulio live in Trento center. Giulio has judo lessons four times a week after school at 15:00 o'clock
- Marta is 57 years old and a Trento Tennis Open organizer. The race takes place on the 2nd of October, and Marta is in charge of delivering information from the parking spots to the Tennis Court.
- Pietro is 48 years old. He is a huge fan of Trentino Volley team and always wants to see their home matches.
- Elena is 14 years old. She just transferred to Trento. She loves football and she used to play for a team in her previous hometown.

4.3 Competency Questions

From the scenarios and personas we have created, we are going to define the Competency Questions as follow:

- CQ1: Paolo discovered that Sanbàpolis is a climbing gym with a boulder area, which is good to meet new friends. He is currently looking at the information about entrance price, opening hours and which bus to take arrive there after having class in Povo. (Scenario 1)

- CQ2: Today, Sally does not have a ride to practice, so she needs to find the best way to get to the athletic pitch in north of Trento for 18 o'clock to get in time to the scheduled training. (Scenario 1)
- CQ3: Mario is unable to bring Giulio from the school to the gym, so he needs to find the buses with minimum transit to the Judo gym so Giulio won't get lost on his way. (Scenario 1)
- CQ4: As a Trento citizen, Marta knows that the biggest free parking lot in Trento is Zuffo parking place, so she needs to check the best bus schedule for the athletes to get to the Trento Tennis Open Arena. (Scenario 2)
- CQ5: On the 16th of November, there is a Champions League home match. As a fan of Trentino Volley, Pietro really looks forward to this match. The match starts at 21:00 and the supporters are allowed to get in at 20:30. As Pietro wants to get top spots, he wants to find the best transport to be there on time. Pietro will take the bus/train from Gardolo. (Scenario 3)
- CQ6: Elena is currently looking for a new club to continue playing football and make some new friends. She can't drive, so she wants to find a club which is within bike riding distance from her new house which is in Cognola. She also wants to get the best trade-off between pricing, how far the clubs train and time schedules. (Scenario 1)

The 6 Personas and CQs above described are not covering all the events that the Scenario and the Purpose can generate, but at the same time provide a clear vision of what the Purpose can cover while taking into consideration a wide range of possibilities.

4.4 Entities Identified

We have identified three types of entities: common entities, core entities, and contextual entities.

Competency Query	Common Entities	Core Entities	Contextual Entities
CQ1	Location, Trento Public Transportation, Facility	Sport	Facilities Pricing, Transport Pricing, Availability, Opening Hours
CQ2	Location, Trento Public Transportation, Facility	Sport, Association Groups	Transport Pricing
CQ3	Location, Trento Public Transportation, Facility	Sport	Transport Pricing
CQ4	Location, Event, Trento Public Transportation, Trento Parking Areas, Facility	Sport	Transport Pricing, Availability
CQ5	Location, Event, Trento Public Transportation, Facility	Sport, Association Groups	Transport Pricing, Availability, Opening Hours
CQ6	Location, Trento Public Transportation, Facility	Sport, Association Groups	Facilities Pricing, Transport Pricing, Opening Hours

5 Inception

This section aims to report the KGE sub process performed during the inception phase, by describing each activities both in schema and data layer.

Inception sub activities:

- Resources collection/scraping
- Resources filtering and classification over common, core and contextual
- Resources knowledge definition/extraction
- Resources formatting (semi-formal transformation)

5.1 Inception Idea

This phase of the project consist in scrape, clean, format, and model all the data and resources collected and found in the previous phases as inputs for the inception part. More deeply, in this part of the project the input data, identified in phase-0 considering the Purpose, are firstly

scraped, to gather more information and data, then cleaned and formatted in the way needed. After that, it is important to proceed with modeling the Schema of the collected data, by using Protègè and formally defining them in RDF-OWL. Then the final step of the Knowledge Graph Construction activity is initialized, which gets as inputs the dataset collected and the schemas produced. The outputs of the inception phase are a set of semi-formal resources. The phase is summarized in the following picture 1.

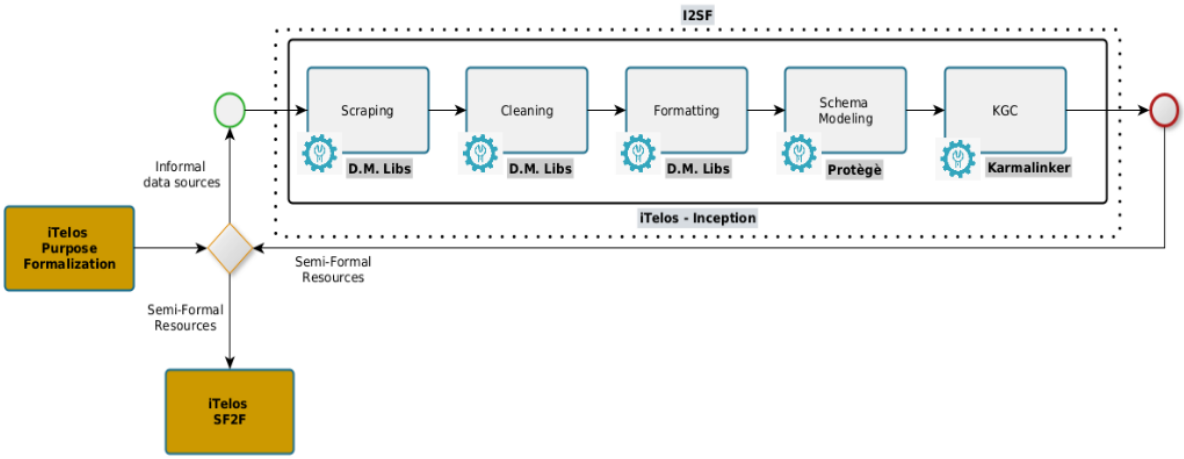


Figure 1: Inception Phase Schema

5.2 Scraping, Cleaning, Formatting

The Purpose Formalization part of the project, gather as input a list of website which provided some information about Sport Facilities in Trentino, which became relevant for the Purpose and much needed after the formalization of the CQs. For this part of the project the scraped websites needed where the ”comune.trento.it” and ”comune.arco.tn.it”. Both of them provide a section for sport facilities in the municipality of Trento and Arco. To gather the data it has been done the following steps:

- Wrote a script in python using the Scrapy libraries in a Poetry environment
- The data to collect where in a table inside each facility page, so the script would first gather all the links by checking s in the page, then find the containing the rows of the table in each link. The main data gathered, in both municipalities, were Address, Type of Facility, Telephone number of the managers, E-mail and facility website.
- Each data were ”cleaned” and formatted during the scraping by using strip() function, further cleaning wasn’t required in this phase.
- The scraped data were formatted as a .json output file, ready for the KGC part of the inception phase

On the need of collecting more data, it was found as useful resource the ”Pagine Gialle” website, which provide many information of different type of facilities in Trentino. This website was a bit more challenging than the previous ones, but by following similar steps the main data collected were: Address, Telephone, Openings Hourrs and Type of Facilities. There are two main reason why this website became vital:

- many facilities had ”easy” to scrape openings
- the collected facilities are located all around Trentino and not only in Trento and Arco, providing more knowledge

During this part, more website were found as possible resources to the purpose, but the ones described above where the only ones providing a dedicated page for sport facilities, to have a wider range of data able to provide knowledge, further scraping on other websites with more different structure should be done.

5.3 Schema modelling

The next phase of the inception part is schema modeling. in this part, the entities found in the Purpose Formalization phase by the CQs definition were modeled and formally defined in RDF-OWL by Protègè. Each found Entity was modeled by referring to the schemas provided by Schema.org regarding Facilities, Sport related schemas and Public Transports to follow one of the main objectives of iTelos methodology which is reusability. In the following picture there is an example of how protegè was exploit in this pahse 2

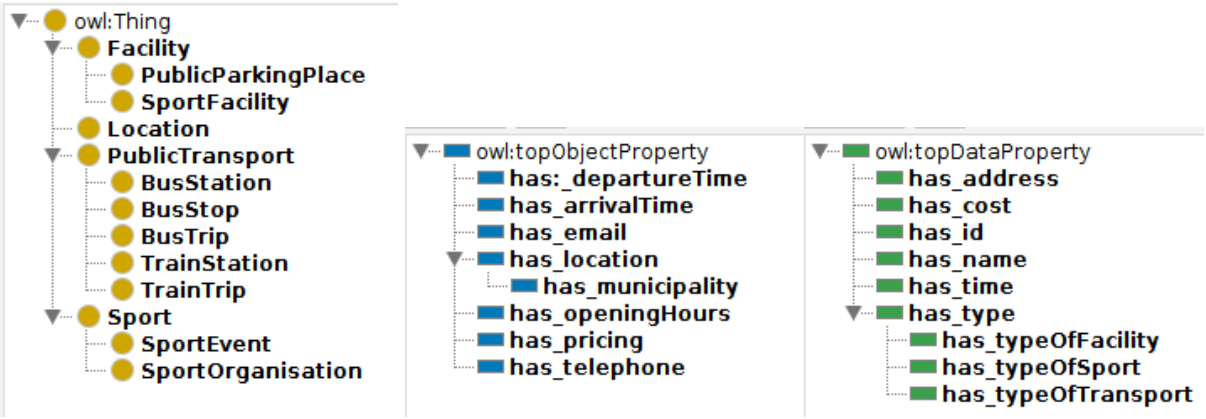


Figure 2: Schema modelling via Protege

5.4 KGC

To conclude the phase, the gathered datasets and the Schemas modelled in the previous part were processed in Karma application, which return as output a single object representing a semi-formal information model.

6 Informal Modeling

This section is dedicated to the description of the informal modeling phase. In this phase, we built an entity-relationship (ER) model that describes the dataset that we worked on. On top of it, we build the teleology by importing the structure within the ER model into Protégé. Lastly, we updated our datasets by filtering them and aligning them with the teleology we created. In addition, we discussed the issues that we encountered while working with each of the sub phases.

6.1 ER Model Description

After Phase-0 and the inception part of the project, iTelos continues by creating an ER-Diagram which is the graphical representation of the acquired knowledge of the previous phases. The ER provides enough information for a technician to get deeper into the project, and at the same time is easy to understand for the domain expert.

The diagram in question is the results of the ETypes and Attributes found during the Phase-0 and the Schema Modelling. In particular, from the previous phases, each found entity was divided into Objects and Functions and then linked to their Actions as described in table 3.

Objects	Functions	Actions
Everything		SpatialPartOf
Region		SpatialPartOf, LocatedIn
Facility		LocatedIn, WorkIn, Near, IsTypeOf
	ParkingPlace	Near, IsTypeOf
	SportFacility	Near, IsTypeOf
Sport		HasTypeOf, PracticeTypeOf
SportOrganization		LocatedIn, PracticeTypeOf
SportEvent		LocatedIn, HasTypeOf, Near
BusStation		LocatedIn, Near, SpatialPartOf
BusStop		LocatedIn, Near, SpatialPartOf
	BusTrip	SpatialPartOf, Operate
TrainStation		LocatedIn, Near, SpatialPartOf
	TrainTrip	SpatialPartOf, Operate
PublicTransport		LocatedIn, Operate

Figure 3: Object, Function and Action table

The diagram was created by:

- Defining all the ETypes by Object, Function and Action as written in table 3

- Defining all objects as subclass of Everything
- Defining all objects as ”Located in” Trentino Region as it is the spatial area of the Purpose
- Linking all objects with their corresponding functions
- Adding and filtering all actions
- Adding colors to the entities based on their categories: blue color for Common Entities, green color for Core Entities, and red color for Contextual Entities

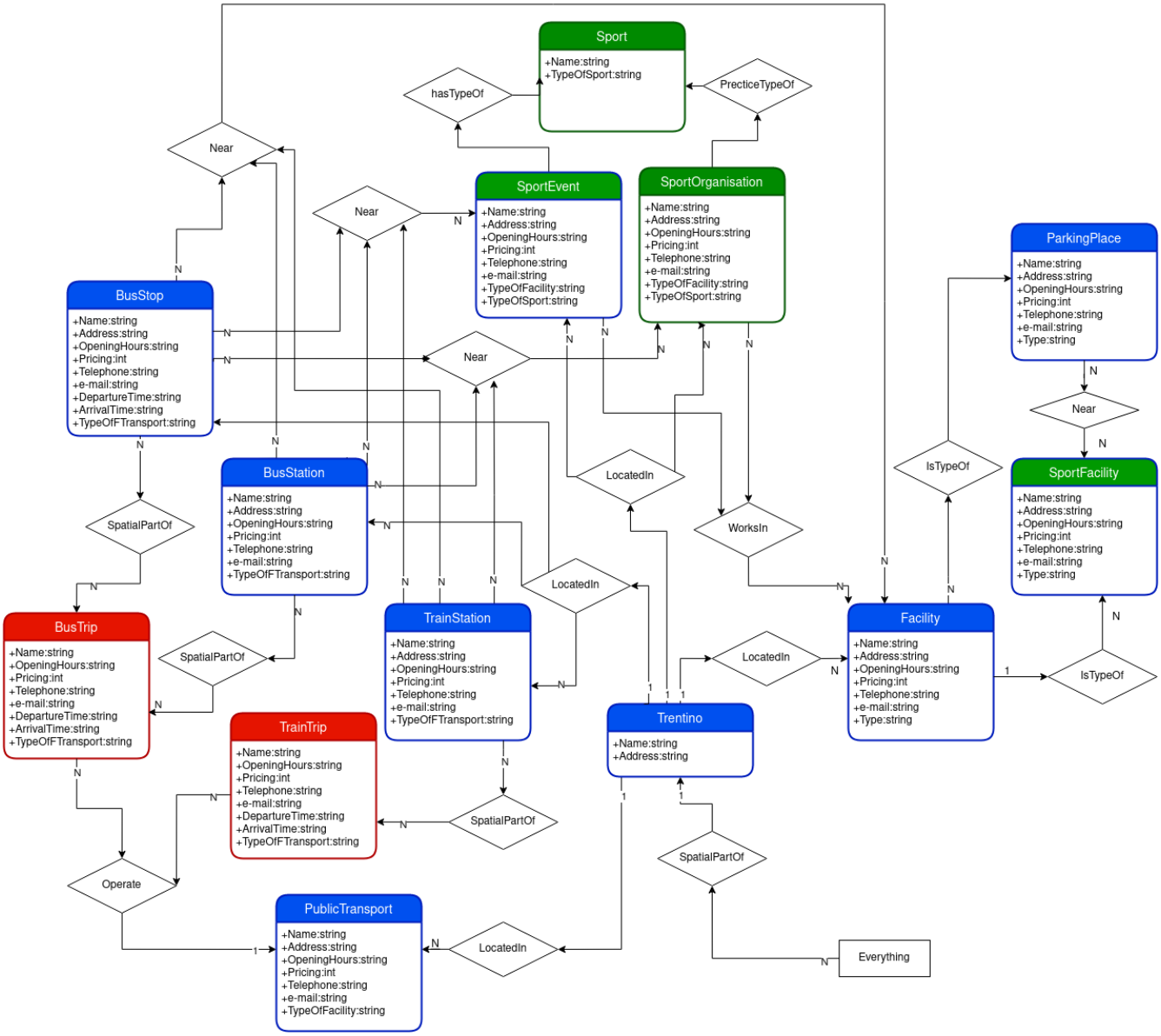


Figure 4: ER Diagram of Trentino Sport

In this section, there is one main open issue that is visible to the reader. The diagram itself became too complex due to the impossibility to create some ”Superclass” to generalize

some entities such as BusStop, BusStation, and TrainStation. All of these objects should have dedicated ”Near” and ”location” actions, which didn’t fit in the diagram. Then for sure, a better organization of the diagram is much needed.

6.2 Teleology Building

During this sub phase, we used the ontology we built during the inception phase to build a new teleology. We primarily worked using the Protege tool. First, we created the new Region class. In our project, the region is Trentino, so we defined the Region as Trentino. Next, we imported all of the entity types that we have defined during the previous phases. Furthermore, we created object properties according to the ER diagram. For example, the BusTrip object has SpatialPartOf relationship with BusStation. We implemented the relationship in Protege by adding it as the object properties of BusTrip.

After following these steps for all entity types, we exported the ontology to an OWL format file, so that it can be accessed again using the Protege tool. In the following pictures,5 and 6, is shown how protegè was exploited.

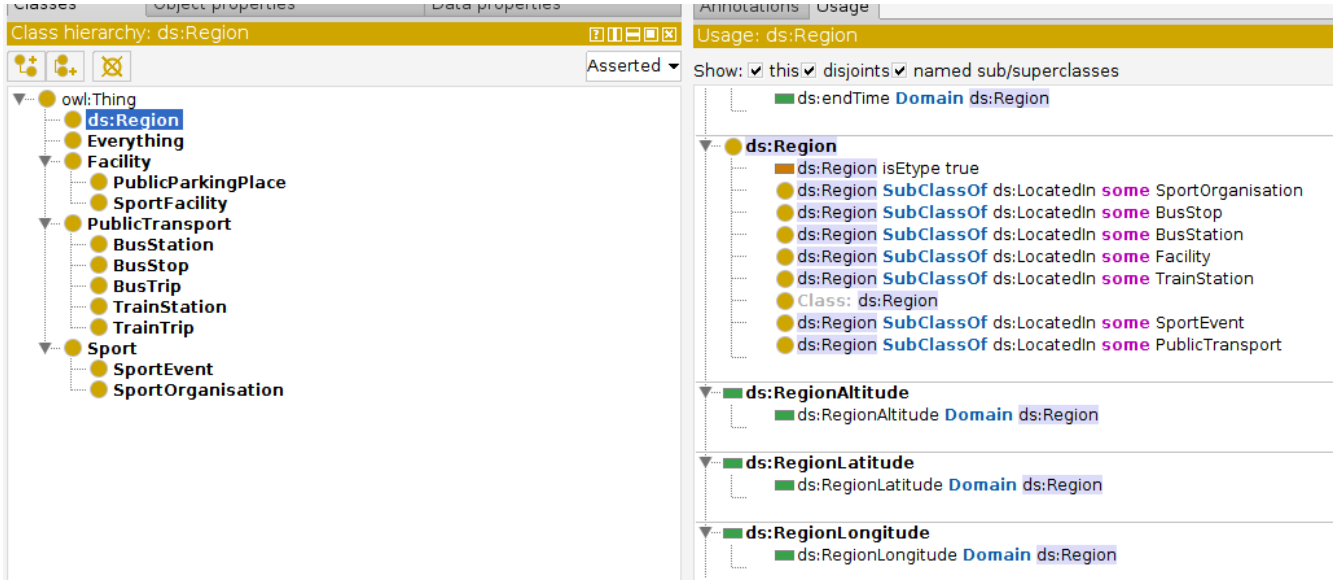


Figure 5: Teleology via Protegè 1

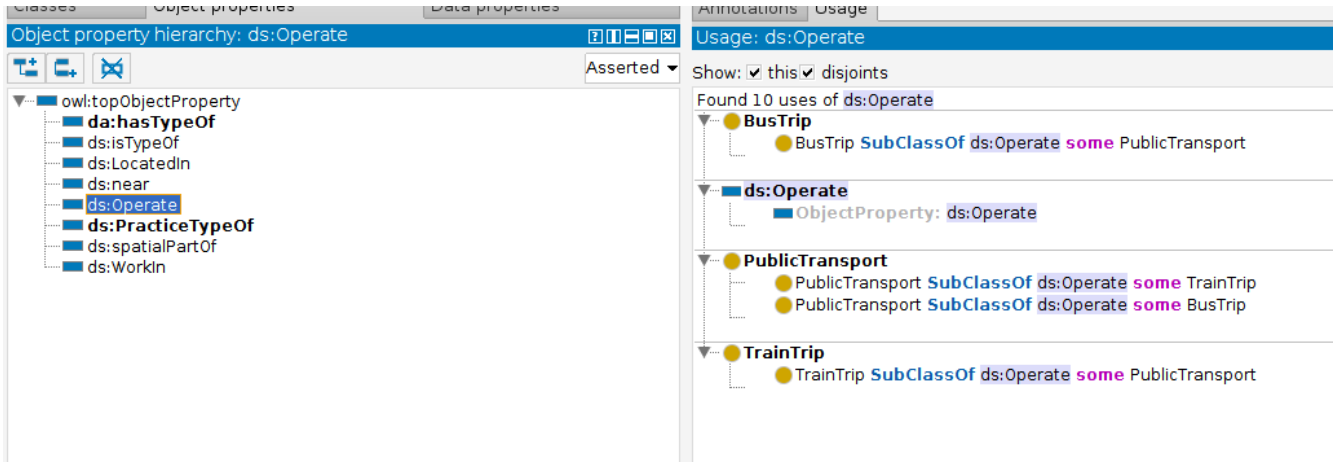


Figure 6: Teleology via Protegè 2

6.3 Datasets Filtering and Alignment

After having created the ER-Diagram and built the Teleology, some data must be aligned to these models. In order to do so, the datasets used as resources went under filtering and alignment process. While doing so, some features of the datasets were filtered out. The following table is a summary of this part of the Informal Modelling Phase:

Dataset	ImpiantiSportivi.csv	Arco_Web_Facilities.json	Trento_Web_Facilities.json	PagineGialle_Web_Facilities.json

Remained Attributes	WKT;
	Impianto;
	Ubicazione;
	Comcat;
	Gestione;
	Disciplina;
	Discipline;
	Strutture;
	Tiplogia;
	Management;
	Typology
		Title;
		Infos:
		-Indirizzo;
		-Telefono;
		-E-Mail;
		-WebSite
		}
			Title;
			Infos{
			-Indirizzo;
			-E-mail;
			-Telefono;
			-Indirizzo Web;
			-Impianto gestito da;
			-Tipologia di luogo
			}
				Title;
				Address;
				Openings;
				Telephone;
				Type_Of_Facility

Figure 7: Data alignment table

A necessary consideration must be done. As it is an ongoing project, some more data were scraped outside the main inception phase, which led to some open issues with the filtering and alignment part due to the continued discovery of new useful resources.

7 Formal Modeling

At this point of the project, the generated outputs from the previous phase are the Teleology and the datasets filtered with the useful features. These outputs are necessary to the next phase of

the methodology, the formal modelling, which has the aim to create an Ontology, link it to the previously mentioned Teleology in order to generate a Teleontology. This phase request are:

- Create an ontology
- Align it to the Teleology to generate the Teleontology, which on the following steps it will be call *ETG* (Entity Type Graph)
- Continue with the lenguage alignment
- Conclude the phase by align the datasets with the ETG

7.1 ETG Generation

7.1.1 Schema Alignment

One of the main idea of building a knowledge graph is reusability. One such method to achieve that is by reusing an existing teleontology, which allows the user to reuse and improve our knowledge graph based on the existing teleontology. Which is why in this phase, ideally it is necessary to find a teleontology that matches the current purpose in LiveSchema. Nevertheless, currently, not every purpose has its teleontology in LiveSchema yet. Thus, it is necessary to build the teleontology fit for the current purpose.

While building a new teleontology, it is important to explore existing ontologies to find any that matches the existing entities. For this puspose, the chosen schemata are from General Transit Feed Specification (GTFS) Ontology and schema.org Ontology. Specifically, the following ontologies were used:

- **GTFS**: Station, Stop, Trip
- **Schema.org**: Event, Place, CivicStructure, Intangible, Organisation

Each of the ontology entities became a ”superclass” of the entities found in the previous phases, emphasizing generalization and aligning the various heterogeneity created from the earlier phases. The following diagram 8 summarizes this step.

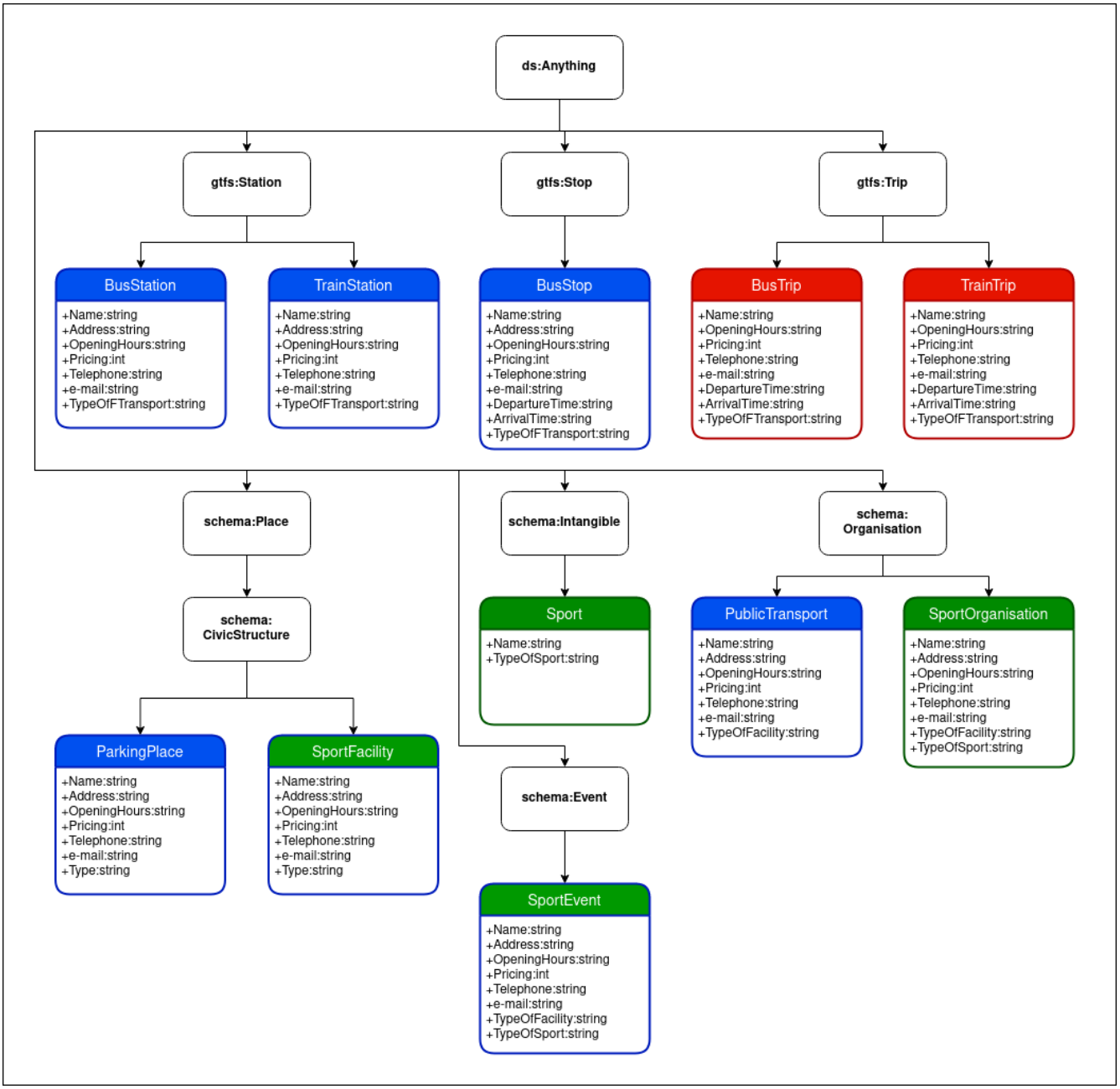


Figure 8: Ontology Alignment Idea Diagram

Furthermore, Figure 9 explained the alignment of the existing teleology with the used ontology by using the Protegee application.

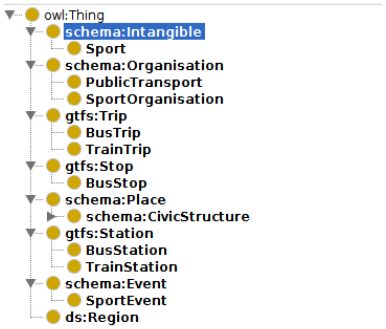


Figure 9: Teleontology Alignment

7.1.2 Language Alignment

In this phase, a crucial point is to align the different languages used during the previous phase, to provide a more general integration in the next phases. As natural language present an intrinsic ambiguity between the different idioms, the goal is to assign a specific identifier to each concept used in the ETG, to provide more general knowledge in a heterogeneous entity like language. In the following pictures 10 are reported the output ontology alignments from KOS application via Protege.

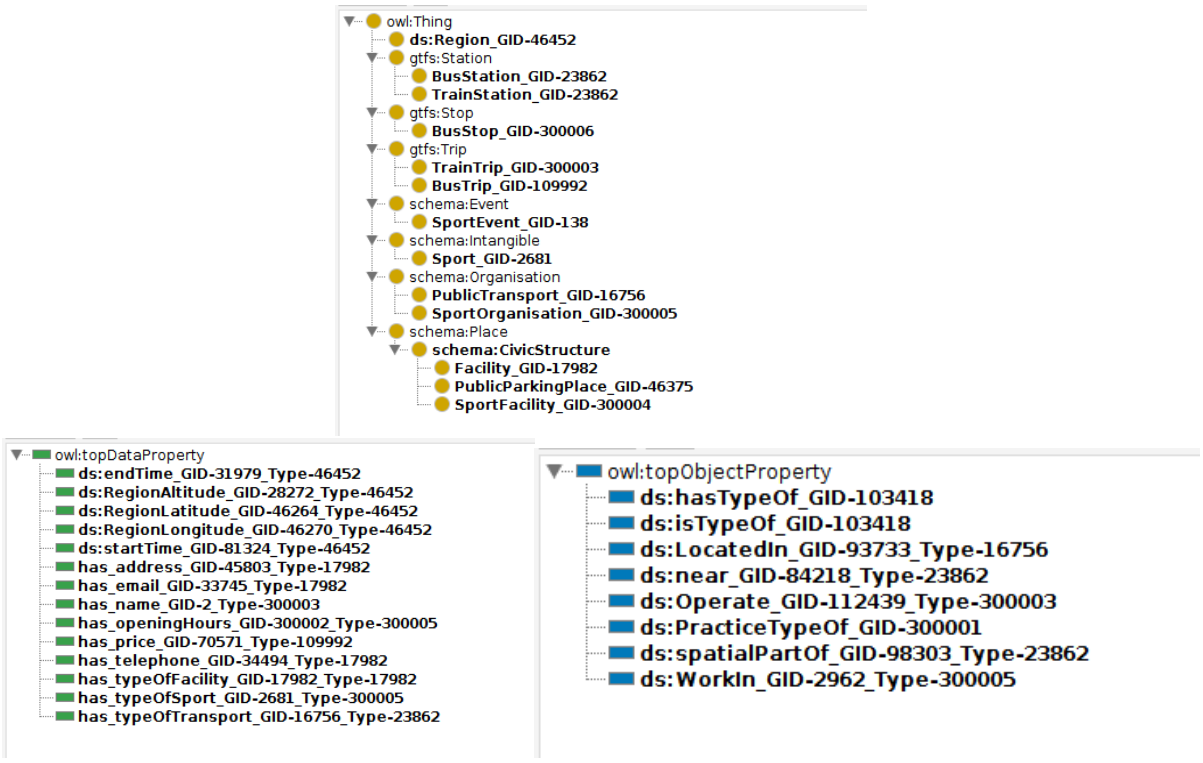


Figure 10: Language alignment via Protege

7.2 Data Management and Alignment

During this step, it is important to align the dataset retrieved from the previous phases to the newly-formed ETG. The collected data have many heterogeneous properties, nevertheless managing the integration of the datasets in this part becomes trivial. Particularly, the data found during the Phase-0 are in CSV format, while the ones collected in the inception phase are JSON files. Thus, a Python script was used to adapt the JSON formats into a CSV, in order to synchronise all the datasets in the same file format. Additionally, further filtering and management has been done to the scraped data in order to have them align with the ETG.
For example, in the ”PagineGialle_Web_Facilities.json” file, the opening hours were divided with different features for each day, while during the ETG creation this attribute was identified as a single feature named *”openingHours”*. To align them, each day was concatenated into a single string attribute called *”openingHours”*. Another important fact is that all the missing values in the JSONs were set as blank during the conversion.

7.3 Open Issues

An important aspect of this phase is the language. While the creation of teleontology seemed to follow a good workflow, the Language Alignment didn’t return a complete alignment of each entity, meaning that either some language concept has been misunderstood or there was a lack of optimum Protege exploitation during the alignment, further work should be done to provide a better language alignment.
Data management and alignment is instead an important step for the data integration part. In this case, the alignment seems to be good enough to continue the work, but yet as it is an ongoing project, this step could require some major changes.

8 KGC

The final step of the methodology is to finally bring all together the collected resources and information by link the ETG created in the previous step. This part is trivial as the project objective to build a Knowledge Graph (KG) becomes reality, as the data gathered in the previous phases are reshaped into the final KG. The important steps in this section are:

- find a identifier of the entities to do a proper Entity Matching
- map the ETG to the data and retrieve the EG

8.1 Entity Matching

In this phase, it is important to generate the data integration idea inside the EG. The final result will not be a single dataset where all the information and features are collected together, but instead, the same entity belonging to a different dataset will have links to the knowledge provided to other data. In order to have a proper link to each entity, it becomes trivial to find a smart identifier that is unique to each entity of the eType.

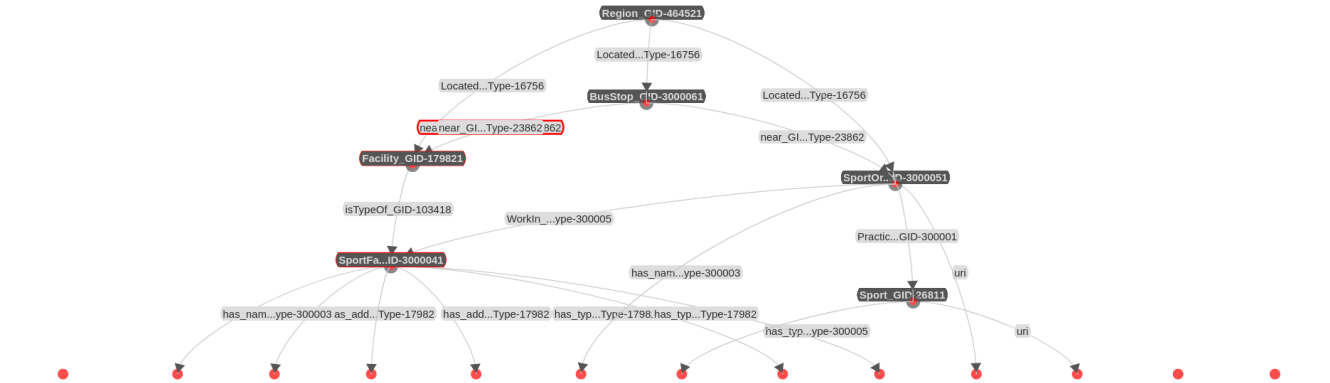
This step can be divided into two categories, one compose of the dataset provided by ”Trentino Trasporti” and the other that had the facilities datasets and the scraped ones. The first category was already formatted in gtfs standards and provided a unique ID for each entity inside each dataset, which made the entity matching task fairly easy. On the other hand, the facilities linkage required some hands-on work to create a good identifier. The dataset ”ImpiantiSportivi.csv” had already the IDs, so it became the key dataset to start with the integration to the scraped data collected from different web resources. The scraped dataset of course didn't have ID features. The problem then become to find a good identifier that was similar for each facility in the different datasets.

In order to do so, a quite unique identifier to the facilities is the ”Name” and the ”Address”. As both of them can have some differences between resources, to find the ones similar to each other, it was decided to code a small script in Python that apply the similarity evaluation via Edit Distance between names and addresses. If the Edit Distance was below a certain threshold, the compared entities would be assigned to the same ID, otherwise, a unique new ID would have been assigned.

Due to the presence of strong semantic heterogeneity, this idea worked for 20% of the entities. For example, the Track and Field pitch in Trento was found in 3 different datasets (”ImpiantiSportivi.csv”, ”Trento_scraped_web_infos.csv”, ”PagineGialle_scraped.csv”), in all of them the name was slightly different, but the address collected didn't match at all, which made clear that some of the work had to be done by hand.

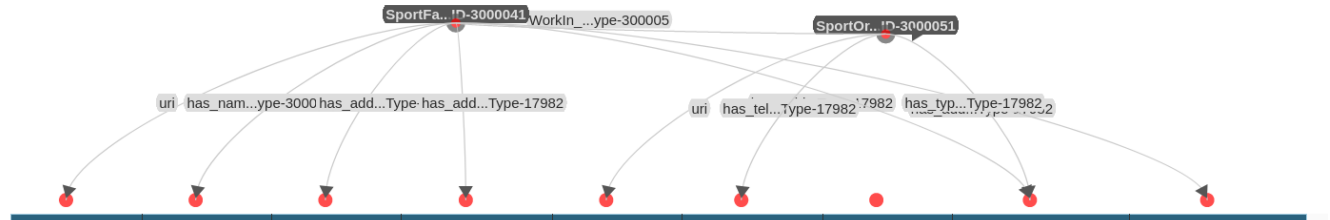
8.2 Data Mapping

After have found a feasible identifier for each entity, now it becomes trivial to link all the datasets to the ETG and finally build the EG. For this phase of the prject it was used the KarmaLinker tool, the opensource tool that provides a good solution to data linking and mapping. The following screenshots shows some examples of the KarmaLinker Tool.



WKT	id	Name	Address	Municipality	gestione	disciplina	strutture	tipologia	management	discipline	facilities	typology
POINT (664758.2828944... 5101773.5804799...	2	BOCCIODR... COMUNALE	VIA FERMI, 35	TRENTO	UFFICIO ISTRUZIONE E SPORT	BOCCE		BOCCIODR... O CAMPO BOCCE	MUNICIPAL...	BOWLS		BOWLING RINK or BOWLING GREEN
POINT (664194.2690442... 5101119.4528796...	21	BRIDGE CLUB	VIA FERSINA	TRENTO	ASSOCIAZI... SPORTIVE	BRIDGE		BRIDGE CLUB	SPORTS CLUB	BRIDGE		BRIDGE CLUB

Figure 11: Example of KarmaLinker on ImpiImpiantiSportivi.csv



id	Name	Address	Municipality	Gestione	Telephone	Email	Web	Type
100	Centro tennis di via Fersina	via Fersina, 21	Trento	ATA Battisti - Associazione Trentina Atletica	0461 921515		http://tennis.atatre... /portali.asp	Campo da tennis
30	Blm Group Arena - Palestre	Via Fersina 11, 38123	Trento	A.S.I.S.	0461 959812		www.asis.trento.it	Impianto sportivo multidisciplinare
27	Centro sci da fondo Viote Monte Bondone	Strada delle Caserme 3-5-7-9 38123 Monte	Trento	A.S.I.S.	0461 948105		http://www.asis.tre... /it/impianti/centro-sci-da-fondo-viote-monte-	Impianto sportivo multidisciplinare
103	Impianto	via Bettini 7,	Trento	A.S.I.S.	0461 959812,		http://www.asis.tre...	Piscina

Figure 12: Example of KarmaLinker on Trento scraped datset

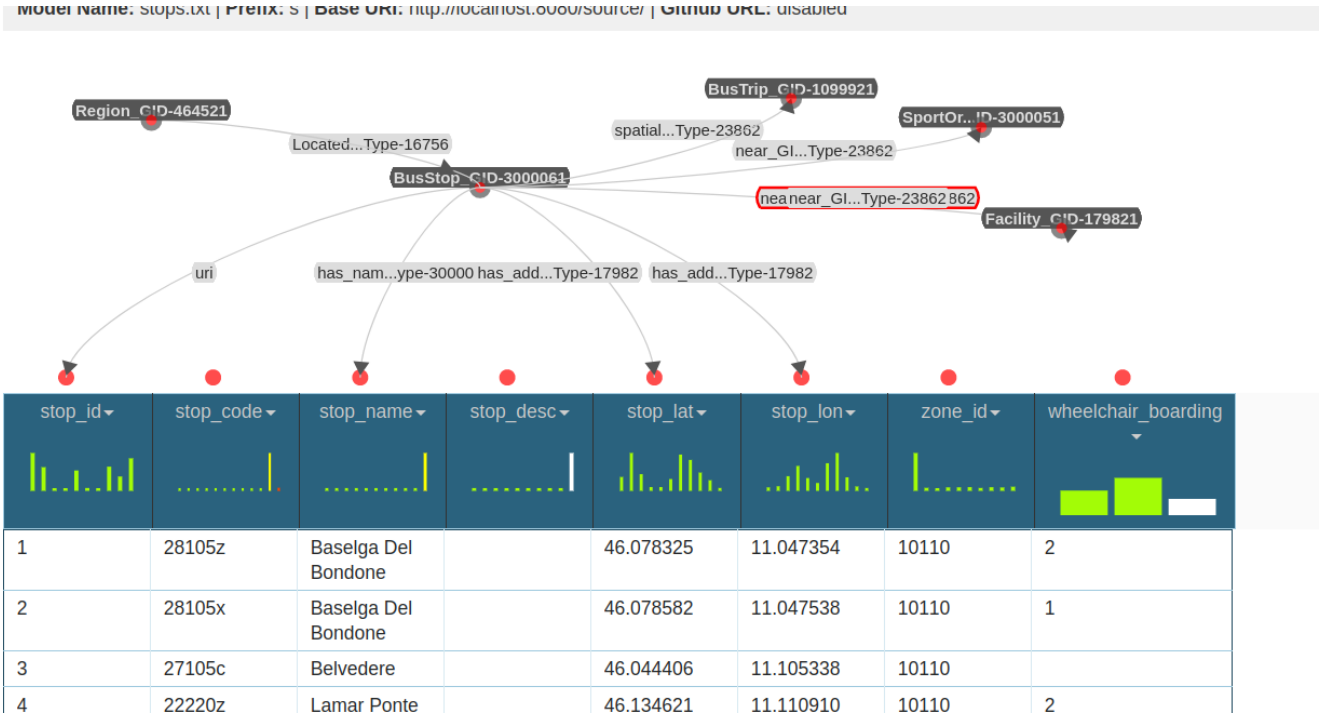


Figure 13: Example of KarmaLinker on BusStop dataset

8.3 KG

After have retrieved the outputs of KarmaLinker, it is possible to upload them on the GrapdDB tool provide by ontotext, which allow to have a visual representation of the KG and to exploit it. Following a visual example at 14.

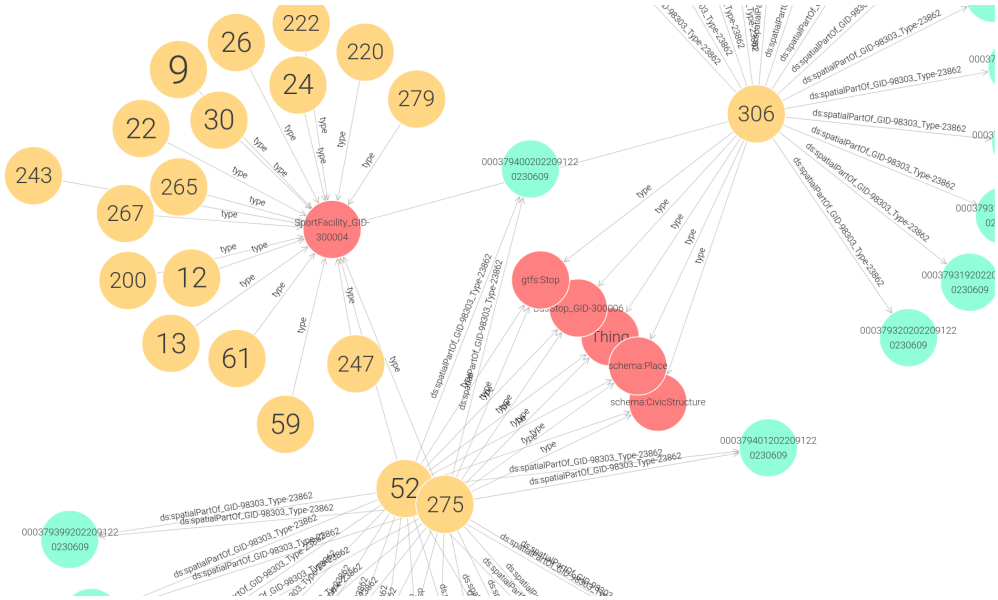


Figure 14: Example of part of the KG

9 Outcome Exploitation

This section aims to provide a description of the KGE process outcome. Here you have to report the final Knowledge Graph information statistics (like, number of etypes and properties, number of entities for each etype, and so on). Moreover this section has to provide a description for the KG possible exploitation, like examples of queries executed, execution time, and so on.

9.1 Knowledge Graph Evaluation

To evaluate the knowledge graph that has been constructed, several metrics can be used. Those metrics are coverage and connectivity.

9.1.1 Coverage

The first metric is coverage, which measures how much knowledge does the knowledge graph represents with entity types and properties. There are two aspects that will be compared to the knowledge graph: the CQs and the Reference Ontologies (ROs). Furthermore, there are two types of coverage that can be measured: etype coverage and property coverage. In total, there are four coverage metrics to measure.

Beforehand, it is necessary to define coverage measurement in this context. For this explanation, assume that the coverage that will be measured is between the knowledge graph and the CQs. Given the knowledge covered by the knowledge graph is represented by α and the knowledge in the CQs is represented by β. First, measure the knowledge covered by both the knowledge and the CQs. This can be done by defining the knowledge intersection between the two, represented by $\gamma = \alpha \cap \beta$. Lastly, measure that out of all knowledge in the CQs, how much is covered by the knowledge graph, which can be calculated with $|\gamma|/|\beta|$. This final value is the coverage. Visually, the coverage can be interpreted as in figure 15.

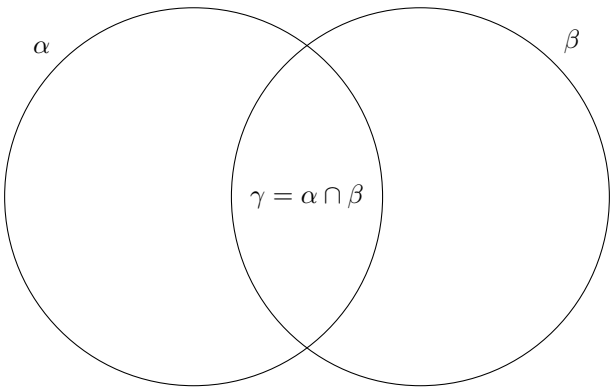


Figure 15: Visual interpretation of the coverage metric

With that in mind, the coverage metrics can be observed in figure 16

	Etype Coverage	Property Coverage
ETG vs CQs	0.64	1.0
ETG vs Reference Ontologies	1.0	0.23

Figure 16: Table of the coverage metrics

9.1.2 Connectivity

The last metric is connectivity. There are three metrics that need to be observed to evaluate this metric: entity connectivity, object property connectivity, and data property connectivity. Entity connectivity can be measured by summing up all the number of entities for each etype in the KG. Object property connectivity can be measured by summing up all the number of object property values that are not empty/null for each etype in the KG. Data property connectivity can be measured by summing up all the number of data property values that are not empty/null for each etype in the KG.

This metric aims to measure how ”dense” or ”connected” the entities are. Say that there is a KG with one etype with four different object properties. In the ideal case, the object property connectivity should be four times the entity connectivity, which means that all of the object property values are not null/empty. Additionally, the closest the object property connectivity to the maximum value, the better. However, if the object property connectivity is significantly lower than the maximum value, then there will be questions regarding the missing object property values, which remarks the ”sparsity” and ”incompleteness” of the KG. Also, the similar principle is also applicable for data property connectivity.

With this idea in mind, the connectivity of this KG can be observed in figure 17

Entity Connectivity	4,386
Data Property Connectivity	17,706
Object Property Connectivity	16,121

Figure 17: Table of the connectivity metrics

9.2 Graph exploitation

Then another important aspect to evaluate the process is how it can be exploited to answer queries (or how can the graph be questioned). To do so, all the output retrieved from the Karma-Linker application was used inside the open-source tool ”GraphDB”, which is a powerful source that permits to visualization graphically the KG itself, and then running queries to retrieve answers from the graph. GraphDB uses SPARQL programming language as a sort of language, which for some aspects is similar to any SQL-related language, but yet presents some differences which allow to query a graph instead of a relational table.

So GraphDB become trivial to question the identified CQs, and check if the graph is capable to return the informations necessary to answer the CQs. There are two main ways to retireve informations from GraphDB. The fist one is to exploit the visual graph, even by using some tweaks via SPARQL and then find the right answers by playing with the returned grapgh. The other method is to directly query the graph using SPARQL and then check the answers on the

returned tables, just like what would happen on a DBMS database using SQL queries. The first method is represented in the following pictures 18 and 19.

```
# CONSTRUCT or DESCRIBE query. The results will be rendered visually as a graph of triples.
CONSTRUCT WHERE {
    ?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
     <http://knowdive.disi.unitn.it/etype#SportFacility_GID-300004>
} LIMIT 200
```

Figure 18: Query to retrieve visual graph

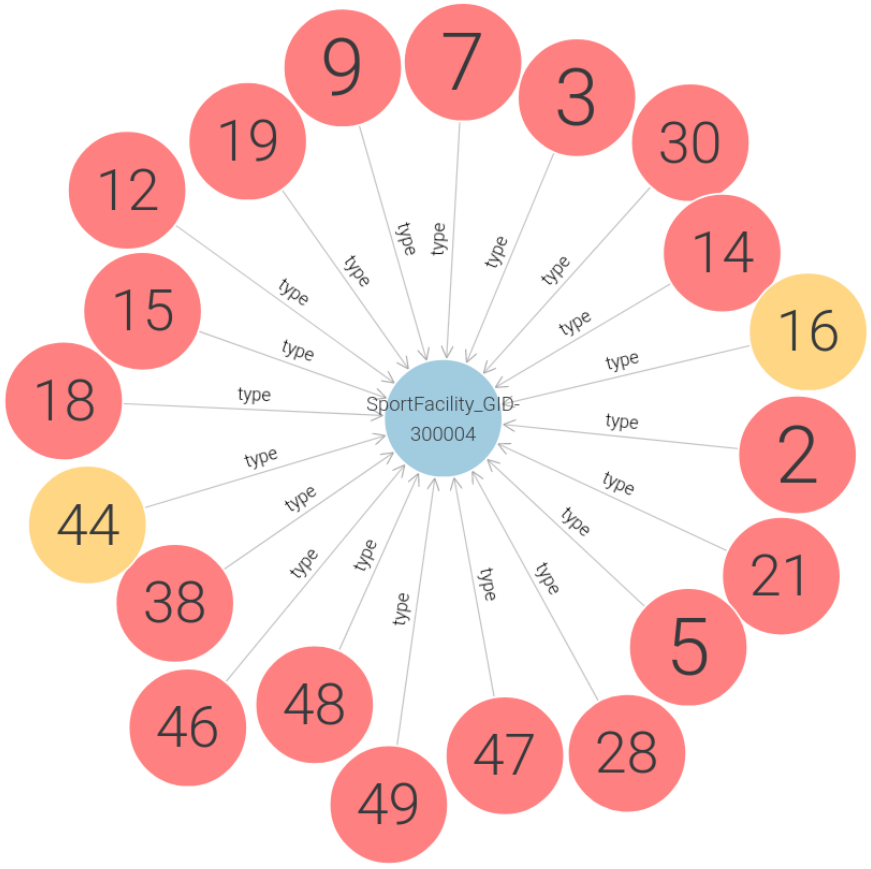


Figure 19: Retrieved Visual Graph

Then the KG can be questioned by SPARQL queries, the following examples (pictures 20 and 21) show how some CQs were answered by using SPARQL.

```
 BASE <http://knowdive.disi.unitn.it/etype#>
PREFIX kge: <http://knowdive.disi.unitn.it/etype#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
select ?facility_name ?openings ?pricing ?no where {
    ?facility kge:has_name_GID-2_Type-300003 "Sambapolis" .
    ?facility kge:has_name_GID-2_Type-300003 ?facility_name .
    filter(?facility_name = "Sambapolis")
    ?facility kge:has_openingHours_GID-300002_Type-300005 ?openings .
    ?facility kge:has_price_GID-70571_Type-109992 ?pricing .
}
```

Table | Raw Response | Pivot Table | Google Chart

Filter query results

	facility_name	openings	pricing
1	"Sambapolis"	"Every day 9:00-23:00"	"https://www.operauni.tn.it/palestradiroccia/listino-prezzi/"

Figure 20: Query for CQ1

```
 BASE <http://knowdive.disi.unitn.it/etype#>
PREFIX kge: <http://knowdive.disi.unitn.it/etype#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
select ?stopName ?busTrip ?TripName where {
    ?busStop kge:has_name_GID-2_Type-300003 ?stopName .
    filter(?stopName = "Maccani Abondi")
    ?busStop kge:ds:spatialPartOf_GID-98303_Type-23862 ?busTrip .

} limit 100
```

ble | Raw Response | Pivot Table | Google Chart

ter query results

stopName	busTrip
"Maccani Abondi"	_:genid-d0e7dfd2b3844e0a95c59548fee33d2f2172-etype_BusTrip_GID-109992 1
"Maccani Abondi"	_:genid-d0e7dfd2b3844e0a95c59548fee33d2f2172-etype_BusTrip_GID-109992 1

Figure 21: Query for CQ2

10 Conclusions & Open Issues

The project is based on the following particularly one important object: the Purpose. The purpose was to retrieve and organize as much information as needed about sports facilities in Trentino. The project respects the scheduling of the purpose, even though it would require much more data to be completed. The main aim of this project remains to use iTelos Methodology to perform and study a KGE and Data Integration inside a university course, with his strict schedule and much of the effort used to learn the methodology instead of only applying it.
Taking that into consideration, the final result satisfies the initial Purpose to return information about sports facilities, but it does not do so extensively in all of the Trentino region. Then during the methodology, some needed datasets weren't collected, due to the impossibility of retrieving them or either because of a lack of time to do a more extensive search. Another open-issue to work on for the future is about location reusability, as it was pointed out how addresses have semantic heterogeneity problems, so it would be more efficient to transform addresses in geographic coordinates to have a better representation of locations. Anyway, the objective of building a KG and following a KGE methodology were fulfilled, in particular, it was shown that the KG created is exploitable to answer the given queries.
To conclude, the project and the course gave us the necessary background and tools to engineer a Data Integration problem and the creation of a Knowledge Graph, which, as shown in the project, can be applied in a great vastity of fields and can be exploited in many future works.