UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

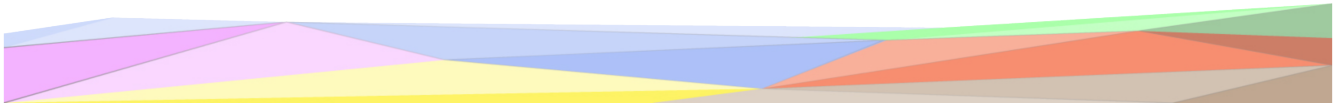
– KnowDive Group –

KGE 2023 - Project Report

Document Data:	Reference Persons:
February 18, 2024	Patrick Nanys, Abdelhakim Rabia

© 2024 University of Trento
Trento, Italy

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



Index:

1 Introduction **1**

2 Purpose and Domain of Interest (DoI) **1**
2.1 Domain of Interest . . . 1
2.2 Project's Purpose . . . 2

3 Project Development **2**
3.1 Data Production . . . 2
3.2 Data Composition . . . 3

4 Purpose Formalization **3**
4.1 Scenarios . . . 3
4.2 Personas . . . 3
4.3 Competency Questions (CQs) . . . 3
4.4 Concept Identification . . . 4
4.4.1 Design process . . . 4
4.5 ER model definition . . . 6
4.5.1 Design process . . . 6

5 Information Gathering **7**
5.1 Consumer Activities . . . 7
5.1.1 Knowledge Layer . . . 8
5.1.2 Data layer . . . 8
5.2 Producer Activities . . . 8
5.2.1 Knowledge Layer . . . 9
5.2.2 Data layer . . . 9
5.3 Schema generation . . . 11
5.4 Formal resource generation . . . 12

6 Language Definition **13**

7 Knowledge Definition **15**
7.1 Ontology . . . 16
7.1.1 Schema.org . . . 16
7.1.2 DataScientia LiveKnowledge . . . 17
7.2 Teleology . . . 17
7.3 Teleontology . . . 18

8 Data Definition **19**
8.1 Entity Matching . . . 19
8.2 Entity Identification . . . 20
8.3 Entity Mapping . . . 20

9 Evaluation **22**
9.1 Knowledge Layer Evaluations . . . 22
9.1.1 Teleontology . . . 22
9.1.2 Teleontology vs CQs . . . 22
9.1.3 Teleontology vs. Reference Ontologies . . . 23
9.2 Evaluating the Data Layer . . . 23
9.3 Queries and Their Relevance to Competency Questions . . . 25
9.3.1 Competency question 1 . . . 25
9.3.2 Competency question 2 . . . 25
9.3.3 Competency question 3 . . . 26
9.3.4 Competency question 4 . . . 28
9.3.5 Competency question 5 . . . 29

10 Metadata Definition **31**
10.1 Dataset Metadata . . . 31
10.2 Knowledge Metadata . . . 31
10.3 Language Metadata . . . 31
10.4 Project Metadata . . . 32
10.5 People Metadata . . . 32

11 Open Issues **32**

Revision History:

Revision	Date	Author	Description of Changes
0.1	February 18, 2024	Patrick Nanys, Abdelhakim Rabia	Document created
0.2	February 18, 2024	Patrick Nanys, Abdelhakim Rabia	Project Development and Purpose Formalization sections written
0.3	February 18, 2024	Patrick Nanys	Gathered datasets
0.4	February 18, 2024	Patrick Nanys	Cleaned the datasets
0.5	February 18, 2024	Patrick Nanys	Created base ontologies
0.6	February 18, 2024	Patrick Nanys	Linked the cleaned datasets to the base ontologies with Karmalinker
0.7	February 18, 2024	Patrick Nanys	Wrote the Information Gathering section in the documentation
0.8	February 18, 2024	Patrick Nanys	Identified and described language concepts and gathered them in the Language spreadsheet
0.9	February 18, 2024	Patrick Nanys	Created the ontology for phase 4
0.10	February 18, 2024	Patrick Nanys	Created the teleology for phase 4
0.11	February 18, 2024	Patrick Nanys	Created the teleontology for phase 4
0.12	February 18, 2024	Abdelhakim Rabia	Updated the ER diagram
0.13	February 18, 2024	Patrick Nanys	Mapped the teleontology onto the cleaned datasets for phase 5
0.14	February 18, 2024	Patrick Nanys	Created the and ran the evaluation sparql queries
0.15	February 18, 2024	Patrick Nanys	Collected and identified the metadata that was saved in the project
0.16	February 18, 2024	Patrick Nanys	Wrote the Language Definition section in the document
0.17	February 18, 2024	Patrick Nanys	Wrote the Knowledge Definition section in the document
0.18	February 18, 2024	Patrick Nanys	Wrote the Data Definition section in the document
0.19	February 18, 2024	Patrick Nanys	Wrote the Evaluation section in the document
0.20	February 18, 2024	Patrick Nanys	Wrote the Metadata Definition section in the document
0.21	February 18, 2024	Patrick Nanys	Wrote the Open Issues section in the document
0.22	February 18, 2024	Patrick Nanys	Created the github webpage for the project

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 to enhance the reusabiltiy of the resources handled and produced during the process. A clear description of the resources as well as of the process (and sub processes) developed, provides a clear understanding of the project, thus serving such an information to external readers for the future exploitations of the project's outcomes.

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

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: High level description of the project development, based on the two main sub process considered by iTelos, producer and consumer, respectively.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities, as well as considered from the point of view of the producer first, and the consumer later.
- Section 9: The description of the evaluation criteria and metrics applied to the project final outcome.
- Section 10: The description of the metadata produced for all (and all kind of) the resources handled and generated by the iTelos process, while executing the project.
- Section 11: Conclusions and open issues summary.

Here is a link to the GitHub repository that contains all the material used during the development of this project: https://github.com/patricknanys/SportsFacilitiesAndTransportationInTrentino

2 Purpose and Domain of Interest (DoI)

2.1 Domain of Interest

The Domain of Interest (DoI) for this project encompasses the Trentino Province, a region located in Northern Italy. This geographical space is renowned for its rich cultural heritage, picturesque landscapes, and a plethora of sports facilities. The time frame considered for this project is contemporary, focusing on the state of sports facilities in December of 2023 (OpenStreetMaps) and transportation in Trentino from February to April of 2023 and December of 2023. This domain has been chosen to provide a comprehensive understanding of how sports facilities are distributed across the region and how they are connected through various transportation means.

2.2 Project's Purpose

The primary purpose of this project is to engineer a Knowledge Graph (KG) that can support applications and services offering detailed information about sports facilities and their interconnection with transportation in the Trentino Province. This KG aims to provide insights into the availability, accessibility, and distribution of sports facilities, and how they are linked with different modes of transportation. This information is crucial for both residents and tourists who wish to leverage these facilities and understand the transportation options available to reach them. The description provided here is informal, based on natural language, and serves as a foundation upon which formal elements and resources will be built and integrated.

3 Project Development

This section describes, at top level, how the project's purpose will be satisfied. More in details the current section is divided into two main subsections, defined as follows.

3.1 Data Production

The primary focus of this phase is to produce quality datasets which revolve around the sports facilities and their interconnection with transportation in the Trentino Province. Given the significance of this domain, it's imperative to ensure that the data is of high quality and accurately represents the state of sports facilities and transportation in the region in the given time ranges.

To achieve this, the data producer will:

- Extract and formalize data about sports facilities in Trentino. This data should include point geometry features, represented by pairs of longitude and latitude coordinates, indicating the exact location of each facility.
- Gather data about transportation means in Trentino, including bus stops, train stations, and other relevant transportation hubs. This data should also have point geometry features.
- Collect information about transportation routes, which should have line geometry features, represented by sets of pairs of longitude and latitude coordinates. These routes will provide insights into how different sports facilities are interconnected and accessible.

The following data sources will be used to create the quality datasets:

- OpenStreetMap - openstreetmap.org
- Overpass Turbo - overpass-turbo.eu
- Overpass API - overpass-api.de

These resources will be pivotal in this phase, ensuring that the data is comprehensive and accurate. In any case when we only have access to low-quality data for a dataset, it's the producer's job to make sure we get a better, high-quality version of that data.

3.2 Data Composition

Once the data production phase is complete, the data consumer's role becomes central. The objective in this phase is to create a well-defined dataset in a standardized format which can possibly be reused in other scenarios and can be provided to the end user.

Based on the above description of this phase, the consumer has the following high level tasks to perform:

1. Download transportation route data and extract the useful information from them.
2. Download information about facilities that provide a place to do specific sports.
3. Merge these gathered sport activity place data based on the facilities they can be performed at.

The ultimate goal is to create a Knowledge Graph that offers a comprehensive view of sports facilities in Trentino and their connection with transportation. Some of the steps the work plan also involves, to get to that goal in this project, after performing the above tasks, are cleaning the data, checking the correctness of it and creating additional synthetic data which we then perform entity mapping on and merge the datasets to form the final KG.

4 Purpose Formalization

Our project purpose is to integrate sports facilities and transportation ways in Trentino to provide a comprehensive understanding of how sports facilities are distributed across the region and how they are connected through various transportation means. To describe multiple aspects considered by the project purpose, we list a set of usage scenarios as follows:

4.1 Scenarios

1. Train trip
2. Bus trip
3. Morning (7-12AM)
4. Afternoon (1PM-5PM)
5. Evening (after 5PM)
6. Urban trip

4.2 Personas

4.3 Competency Questions (CQs)

The CQs have been created in a way to cover different scenarios and persona needs that may arise regarding transportation and sports facility needs.

ID	Name	Age	Interests	Usage	Residence	Special needs
1	Paolo	25	All sports	Practice	Trento	None
2	Patrick	20	Gym	Practice	Povo	None
3	José	42	Football	Dad	Trento	Wheelchair (for son)
4	Lucia	50	Volleyball	Fan	Rovereto	None
5	Ginevra	13	Athletics	Practice	Villazzano	None

Table 1: Personas

1. Paolo lives in Trento and is in love with sports. Which facilities can Paolo use in Trento that is in walking distance to him?
2. Patrick lives in Povo and would like to go to a gym in the morning to work out. Which is the closest gym to him that is open at a specific time?
3. José has a son and who is in a wheelchair and they live in Trento. He wants his son to enjoy playing football. Which transportation route options do they have that is wheelchair accessible?
4. Lucia is a Volleyball fan in Rovereto and loves to watch volleyball matches, but she is not that into long walks and she prefers buses. Which urban bus route should she take to the next match that involves the least amount of walking?
5. Ginevra lives in Villazzano and she is a professional athlete. Which track has a tartan surface and where she needs to walk the least before and after practices while taking public transport?

4.4 Concept Identification

From the CQs, referring to Personas and Scenarios, we extract Entities with properties. These entities are categorized as either Common, Core, or Contextual entities by considering Focus and Popularity classification. *Focus* defining the importance of an entity given the main purpose and *Popularity* the reuse of the entities in the already existing input informations sources.

4.4.1 Design process

1. Entities identification
   - Bus Stop and Train Stop: These entities represent the physical locations where passengers can board or get off from buses and trains, respectively. Essential attributes like `id`, `name`, `location`, and `timetable` provide detailed information about each stop.
   - *Bus Trip* and *Train Trip*: These entities capture the essence of a journey between two points, detailing the route taken by a bus or train. Attributes such as `id`, `from`, `to`, `wheelchairAccessible`, and `importance` not only offer route-specific information but also cater to the inclusivity and prioritization of different trips based on their significance within the network.
   - *EndUser*: This entity focuses on the individuals utilizing the transportation network, incorporating attributes like `id`, `name`, `location`, `address`, and `specialNeeds`. This allows for personalized routing.

- *Region*: Identified to contextualize the geographic scope of the knowledge graph, the `Region` entity, with its `id` and `name` attributes, serves as a spatial delimiter, grouping together stops, trips, and sports facilities based on their location.
- *Fitness Station, Fitness Centre, Sports Centre, Stadium, Track*, and *Pitch*: These entities represent various types of sports facilities, each with a unique set of attributes tailored to their specific characteristics. Common attributes include `id`, `name`, `location`, and `sport type`, while some, like Fitness Centre and Sports Centre, also detail `opening hours`. These entities are crucial in linking the transportation network directly to sports venues, facilitating easy access for end users.
- *Sport*: This entity abstracts the concept of sports into a manageable form, with attributes such as `id` and `typeOfSport` simplifying the categorization and discovery of related sports facilities.

2. Attributes definition

   These attributes are meticulously chosen to provide comprehensive details about each entity.

   - Bus Stop & Train Stop:
     - `id`: A unique identifier for each stop.
     - `name`: The name of the stop, providing easy identification.
     - `location`: Geographic coordinates or address, enabling precise mapping.
     - `timetable`: A schedule of arrival and departure times, aiding in planning.
   - Bus Trip & Train Trip:
     - `id`: A unique identifier for each trip.
     - `from`: The starting point or origin of the trip.
     - `to`: The destination point of the trip.
     - `wheelchairAccessible`: An attribute indicating if the trip accommodates wheelchair users.
     - `importance`: A measure of the trip's significance in the network.
   - EndUser:
     - `id`: A unique identifier for each user.
     - `name`: The name of the user.
     - `location`: The user's current geographic location.
     - `address`: The permanent address of the user.
     - `specialNeeds`: Information regarding any specific requirements the user may have. (eg. wheelchair access)
   - Region:
     - `id`: A unique identifier for the region.
     - `name`: The name of the region.
   - Fitness Station, Fitness Centre, Sports Centre, Stadium, Track, and Pitch:
     - `id`: A unique identifier.

Scenarios	Persona	CQs	Entity	Properties	Focus	Popularity
2	1-5	4	Bus Stop	id, name, location, timetable	Core	Common
1	1-5	4	Train Stop	id, name, location, timetable	Core	Common
1-2	1-5	4	Bus Trip	id, from, to, wheelchairAccessible, importance	Core	Core
1-2	1,2,3,5	4	Train Trip	id, from, to, wheelchairAccessible, importance	Core	Core
1-6	1-5	2	EndUser	id, name, location, address, specialNeeds	Contextual	Contextual
1-6	1-5	1-5	Region	id, name	Core	Contextual
1-6	1	1	Fitness Station	id, name, location, sport type	Core	Core
1-6	1-2	1-2	Fitness Centre	id, name, location, sport type, opening hours	Core	Core
1-6	1-5	1-5	Sports Centre	id, name, location, sport type, opening hours	Core	Core
1,2,6	1,3,4,5	1,3,4,5	Stadium	id, name, location, sport type	Core	Core
1-6	5	5	Track	id, name, location, sport type, surface	Core	Core
1-6	4,5	4,5	Pitch	id, name, location, sport type	Core	Core
1-6	1-5	1-5	Sport	id, typeOfSport	Core	Core

Table 2: Entities extraction and classification

  - `name`: The facility's name.
  - `location`: Geographic coordinates or address.
  - `sport type`: The type of sport(s) that the facility supports.
  - `opening hours` (for Fitness Centre, Sports Centre): The hours during which the facility is open to the public.
- Sport:
  - `id`: A unique identifier for each sport.
  - `typeOfSport`: A categorization attribute, specifying the sport type.

4.5 ER model definition

At the very end we use the outputs of the concept identification with the extracted entities to create an initial ER model.

4.5.1 Design process

1. Relationships identification

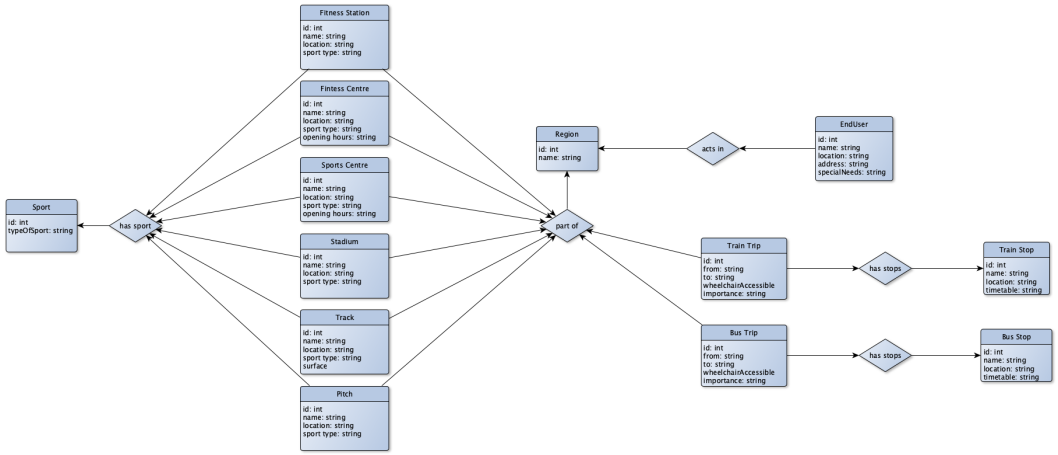


Figure 1: ER model

The following relationships have been identified to determine how the entities relate to each other.

- *part of*: Identified to link each entity (except *EndUser*) to a contextualized geographic scope.
- *acts in*: Identified to link the *EndUser* to the scope.
- *has sport*: Identified to connect each sport facility to a specific sport that can be performed there.
- *has stops*: Identified to link the bus and train stops to specific routes to form a network.

2. Relationship cardinality

   For each relationship, the specification should detail the number of instances of the involved entities. For example the *has stops* relation has *1* to *N* cardinality specifying that to each trip there are multiple stops.

3. Create the ER Diagram.

5 Information Gathering

This section aims at reporting the execution of the activities involved in the Information Gathering iTelos phase. The report, starting from the current section, is organized along two main dimensions. The first one considers the parallel execution of the producer and consumer processes, while the second dimension takes into account the activities operating over data and knowledge layers.

5.1 Consumer Activities

These tasks aim to gather resources that are of "high quality and formal" nature, which have been identified as appropriate for the project.

5.1.1 Knowledge Layer

In the fields of sports facilities and transportation, projects from the previous year had developed a specialized knowledge layer designed specifically for their objectives, in harmony with the data they had. Yet, these existing knowledge layers are not directly transferable to the current project because its scope is either too broad or too narrow. As a result, this task is approached as an activity of the producer. OpenStreetMap is identified as a pivotal source of geographical data, offering an extensive, up-to-date map of the world, contributed to by a diverse global community. Its comprehensiveness makes it particularly suitable for the project's needs. In the scope of the project the focus did not go into the timetables of the specific routes, more about how the sports facilities with different properties suiting different needs are connected. For this reason adopting the GTFS schema did not make much sense and mainly looking at the schema.org schemas in conjuction with the Open Street Map - Trentino Territory Ligthweight Ontology which were readily available. The final schema came together using both the mentioned schemas while also adding some specifics related to this project. For more details refer to the producer part of the knowledge layer.

5.1.2 Data layer

The following have been the main sources for the data which contained high quality resources:

- Overpass Turbo - overpass-turbo.eu
- Overpass API - overpass-api.de

Initially Overpass Turbo has been used to manually run queries on the site to extract resources in geojson format. This has been done for all the sports that was planned to be covered. Example query for the *sport=athletics* can be seen here:

```
[out:json][timeout:25];
area[name="Provincia di Trento"][admin_level=6]->.searchArea;
nwr[sport=athletics](area.searchArea);
out geom;
```

To gather data for the transportation stops and routes this method was not sufficient. The amount of data that had to be extracted was too large and had to opt for using Overpass API for them. To query the data *query.py* was used with the *train_routes.query* and *bus_routes.query* input files to perform the queries. Later on query files have also been created for the sport that have been previously extracted with manual querying to automate the process of up-to-date data gathering.

Running the queries for the transportation routes resulted in an 11Mb file for the train routes and 117Mb file for the bus routes creating a huge dataset.

5.2 Producer Activities

These tasks focus on acquiring "informal" resources from sources characterized by a higher degree of diversity. The materials gathered through the producer process do not adhere to the iTelos

standards for quality and reusability. These are the resources that the producer process ultimately converts into high-quality materials by the end of the procedure.

5.2.1 Knowledge Layer

With the knowledge acquisition taken into consideration from the consumer step, below can be seen the final structure of each GEOJSON and TSV files for the transportation routes and the sports facilities summarized.

Transportation (Core focus)

GEOJSON file	fields
bus_routes	id, name, from, to, min_latitude, max_latitude, min_longitude, max_longitude, type_of_transport, stops, wheelchair, importance
train_routes	id, name, from, to, min_latitude, max_latitude, min_longitude, max_longitude, type_of_transport, stops, wheelchair
bus_stop	id, lat, lon
train_stop	id, lat, lon

Sports facilities (Core focus)

TSV file	Columns
fitness_centre	id, name, lat, lon, sport, leisure, opening_hours
fitness_station	id, name, lat, lon, sport, leisure
pitch	id, name, lat, lon, sport, leisure, surface
sports_centre	id, name, lat, lon, sport, leisure, opening_hours
stadium	id, name, lat, lon, sport, leisure
track	id, name, lat, lon, sport, leisure, surface

End user (Contextual focus)

TSV file	Columns
end_user	id, name, age, interest, usage, lat, lon, wheelchair

5.2.2 Data layer

The data obtained from the consumer contained elements that were not specifically needed for the project, so cleaning had to be performed on the dataset, even though it was of high quality, to specialize it for our use case.

Transportation

The extra information that we need to remove from the data that does not serve any purpose for our KG are the "ways" in the routes. The data is made up from "relations" that consist of "ways" and "nodes". Each node is a stop/station/platform and each way is the exact coordinate

sequence that the train/bus follows during its way from one stop to another. This additional data which is not needed for our KG greatly inflates the dataset. The *1_filter_transportation_routes.py* script loads the geojson traverses the structure to find all the ways in the dataset and remove them.

After the above filtering a cleaning step was also performed on the transportation routes data. In this cleaning step only those tags were retained which were identified as needed in the previous sections for the entities with small exceptions. It was realized that the name and route properties had to be retained as well. The name property contained the train/bus number/reference that identifies the line. This is needed for end users to identify which bus/train they need to take. Also the route tag was retained, because it contained the information that a given route was a "bus" or a "train" route giving us control for future queries to query for specific types based on user preferences outlined in the CQs. The exact steps taken can be seen in the 2_clean_transportation_routes.py script.

The final list of retained tags for trips are: from, importance, name, route, to, wheelchair
For stops the retained fields are: ref, lat, lon

Sports Facilities

For the sports facilities the following steps have been performed:

1. First of all the coordinates had to converted to a simple "Point" format representing each location with latitude and longitude values instead of Polygon or LineString shapes. This is done with the *1_normalize_coordinates_facilities.py* script.
2. All the separate "normalized" geojson files were merged into one json. (*2_merge_facilities.py*)
3. The one merged json file was then cleaned in multiple iterations. First the properties were extracted from the json format and loaded up into a python DataFrame for better a cleaning process. Most of those data points were dropped next where the leisure field was not specified for some specific sports since the leisure field was crucial for sports facility identification which is a main focus in the goal dataset. Some other intermediate steps were also carried out to filter the number of fields that were available to enable us to search for the needed fields faster that take us closer to our needed dataset. Ultimately the DataFrame was split into separate csv files based on the leisure field and only retaining the columns that contain the data we have identified earlier as needed for our KG. For more details about all the steps taken refer to the *3_enrich_and_split_facilities.py*.

The final list of retained fields (combining all the fields of the separate files) are: leisure, name, lon, lat, id, surface, opening_hours.

Other

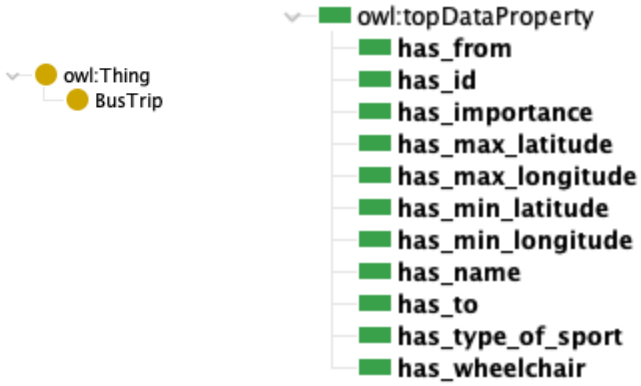
The data for the end user and the region has been created manually based on the previous definitions. Below in the table can be seen the final set of properties created for the datasets.

Manually generated dataset	Final set of properties
end_user.tsv	id, name, age, interest, usage, lat, lon, wheelchair, acts_in
region.tsv	id, name

5.3 Schema generation

Initially, to facilitate the integration of knowledge and data, a range of tools is utilized. The first step involves constructing an ontology with the help of the Protege development tool. This endeavor is guided by the key entities identified from the data's structure and intended use. Following this, examples of the tasks carried out in Protege include developing schemas for each distinct entity, which are stored in separate files. For additional information refer to the projects Github resources in the second phase. Afterward, properties of the data are set according to the structure of the data. The naming is as follows: has_[AttributeName].

Example how this is formulated in Protege can be seen below about the BusTrip EType and its data properties.



In addition, the table 3 lists the data properties, the domains they refer to, and the range.

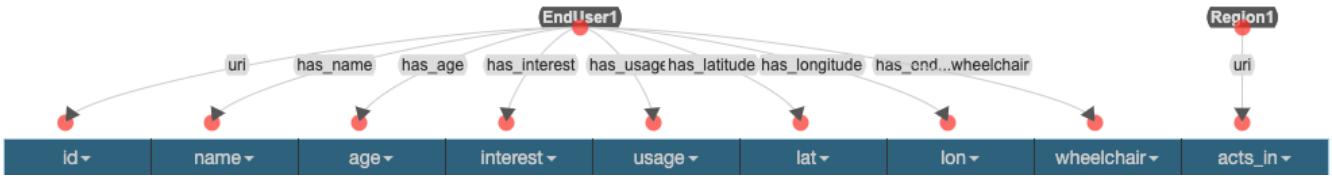
Property name	Domain(s)	Range
has_id	BusStop, BusTrip, FitnessCentre, FitnessStation, Pitch, Region, Sport, SportsCentre, Stadium, Track, TrainStation, TrainTrip, EndUser	xsd:string
has_latitude	BusStop, FitnessCentre, FitnessStation, Pitch, SportsCentre, Stadium, Track, TrainStation, EndUser	xsd:float
has_longitude	BusStop, FitnessCentre, FitnessStation, Pitch, SportsCentre, Stadium, Track, TrainStation, EndUser	xsd:float
has_name	BusTrip, FitnessCentre, FitnessStation, Pitch, Region, SportsCentre, Stadium, Track, TrainTrip, EndUser	xsd:string
has_from	BusTrip, TrainTrip	xsd:string
has_to	BusTrip, TrainTrip	xsd:string
has_max_latitude	BusTrip, TrainTrip	xsd:float
has_max_longitude	BusTrip, TrainTrip	xsd:float
has_min_latitude	BusTrip, TrainTrip	xsd:float
has_min_longitude	BusTrip, TrainTrip	xsd:float
has_type_of_transport	BusTrip, TrainTrip	xsd:string
has_type_of_sport	Sport	xsd:string
has_wheelchair	BusTrip, TrainTrip	xsd:string
has_importance	BusTrip	xsd:string
has_leisure	FitnessCentre, FitnessStation, Pitch, SportsCentre, Stadium, Track	xsd:string
has_opening_hours	FitnessCentre, SportsCentre	xsd:string
has_surface	Pitch, Track	xsd:string
has_age	EndUser	xsd:int
has_interest	EndUser	xsd:string
has_usage	EndUser	xsd:string
has_enduser_wheelchair	EndUser	xsd:string

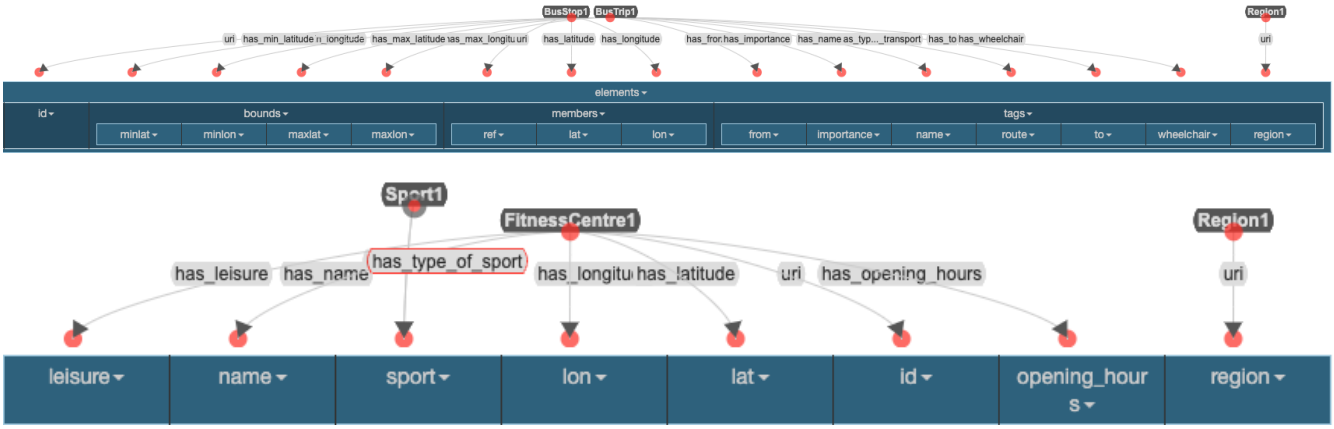
Table 3: Properties, their domains, and ranges

Note: All the examples are designed to give a sense of the process and might vary in their ultimate form.

5.4 Formal resource generation

The derived ontologies were uploaded to the Karma Data Integration Tool, together with the concluding datasets. Utilizing the accessible information, relevant schema, and the Entity-Relationship diagram, connections between the schema and the columns of the CSV datasets concerning data properties are determined. Subsequently, the end product, referred to as Karma models, is generated in Turtle (TTL) format and is available in the Github repository in the Phase 2 directory. In Fig **??** some images can be seen of the process.





6 Language Definition

In the language definition phase, the focus is on tailoring the language (concepts and words) used to represent the necessary information for achieving the project's goals. This involves utilizing the outputs from previous phases, with the consumer examining the broader project objectives and the involved concepts in data composition, including the relationships between entity types.

On the producer side, the efforts are dedicated to developing and formalizing the concepts within each dataset. This is achieved through the "kge-annotator", which aids in identifying or crafting the appropriate concepts by looking for synonyms or similar terms for each word or term previously utilized. When an existing concept in the UKC precisely fits, it is directly adopted. Alternatively, when concepts are deemed too general or are missing entirely, new ones are created to ensure precise representation.

The final set of concept labels can be seen in Table 4 for the EType Concepts, in Table 6 for the Data Properties Concepts and in Table 5 for the Object Properties Concepts. In all the tables the first column shows the initial concept words that have been used up until now in the project and the Concept Label column the attached label and the Description column a short explanation what it means. The Concept Labels where the GID starts with 13XXX have been newly created and defined, the rest have been reused.

In conclusion, after formalizing the concepts relevant to the project, all mentions of e-types, relationships and properties will adhere to the Concept Label column from this point forward.

Initial concept word	Concept Label	Description
FitnessCentre	fitness_centre_GID-13003	A place equipped with machines and facilities for physical exercise
FitnessStation	fitness_station_GID-13004	A spot designated for outdoor physical workouts
Pitch	pitch_GID-13006	a designated outdoor area, often grassy, used for playing sports
Track	track_GID-22259	a course over witch races are run
SportsCentre	sports_centre_GID-13007	a facility that offers a variety of sports and physical activities
Stadium	stadium_GID-23800	a large structure for open-air sports or entertainment
BusStop	bus_stop_GID-45937	a. place on a bus route where buses stop to disembark and take on passengers
TrainStation	train_station_GID-22321	terminal where trains load or unload passengers or goods
Sport	sport_GID-2681	an active diversion requiring physical exertion and competition
BusTrip	bus_trip_GID-13008	a journey taken on a bus, typically over a long distance
TrainTrip	train_trip_GID-13009	a journey made by traveling on a train, often for long distances
Region	region_GID-46452	the extended spatial location of something
EndUser	end_user_GID-53816	the ultimate user for which something is intended

Table 4: EType concept labels and descriptions

Initial concept word	Concept Label	Description
part_of	part_of_GID-13034	denotes that something is a segment, member or component within a larger group, structure, or entity
has_stops	has_stops_GID-13035	refers to the characteristic of a transportation route indicating the presence of designated locations where vehicles pause to pick up or drop off passengers
has_sport	has_sport_GID-13036	indicates the presence or availability of sports activities or facilities at a given location or within an organization
acts_in	acts_in_GID-13037	referring to the actions or behaviors of person within a specific geographical or virtual area

Table 5: Object properties concept labels and descriptions

.

Initial concept word	Concept Label	Description
has_from	from_GID-13015	indicating the starting point of a movement, action, or activity
has_id	id_GID-13016	abbreviation for "identification," referring to a unique identifier
has_importance	importance_GID-76752	a prominent status
has_latitude	latitude_GID-46263	the angular distance between an imaginary line around a heavenly body parallel to its equator and the equator itself
has_leisure	leisure_GID-13017	refers to the time spent in non-work activities
has_longitude	longitude_GID-46270	the angular distance between a point on any meridian and the prime meridian at Greenwich
has_max_latitude	max_latitude_GID-13018	refers to the highest or northernmost latitude point in a specific area
has_max_longitude	max_longitude_GID-13019	refers to the easternmost longitude point in a specific area
has_min_latitude	min_latitude_GID-13020	refers to the lowest or southernmost latitude point in a specific area
has_min_longitude	min_longitude_GID-13021	refers to the westernmost longitude point in a specific area
has_name	name_GID-2	a language unit by which a person or thing is known
has_opening_hours	opening_hours_GID-13023	specific times during which a business or public building is open to the public
has_surface	surface_GID-24186	the outer boundary of an artifact or a material layer constituting or resembling such a boundary
has_to	to_GID-13026	a preposition indicating direction towards a destination
has_type_of_sport	type_of_sport_GID-13028	a specific category or kind of sport, characterized by a set of rules or customs
has_type_of_transport	type_of_transport_GID-13030	specific category or mode of transportation
has_wheelchair	wheelchair_GID-13033	indicating whether a transportation option is accessible to wheelchair users
has_age	age_GID-27200	how long something has existed
has_interest	interest_GID-2206	a diversion that occupies one's time and thoughts (usually pleasantly)
has_usage	usage_GID-4887	the act of using
ha_enduser_wheelchair	wheelchair_GID-25446	a movable chair mounted on large wheels; for invalids or those who cannot walk; frequently propelled by the occupant

Table 6: Data properties concept labels and descriptions

7 Knowledge Definition

This section focuses on detailing the kTelos phase. Beginning with the project's gathered resources, the formalized objective (as partially depicted by the ER model), and the acquired data, the aim is to develop the final KG's teleontology. Specifically, the knowledge resources created during this phase are intended to standardize the information representation, enhancing the interoperability and reusability of the final KG. This is achieved by constructing knowledge resources that maximize

the use of established standard domain ontologies and data schemas.

The Teleontology enables the reuse of project data. In this phase, as with earlier ones, tasks are divided into two categories: producer and consumer. On the producer side, the goal is to develop interoperable ontologies for each dataset that will be created, resulting in multiple ontology files, one for each Knowledge Graph (KG) the Producer will generate. On the consumer side, the aim is to design a single, unified interoperable ontology for the composite final KG, leading to the creation of a single ontology output file.

7.1 Ontology

This section outlines the top-down knowledge definition stage within the kTelos process. The objective is to employ Lightweight Ontologies, already harmonized with the UKC, to establish a high-level perspective of the entities participating in the project.

The following two sources have been used when possible for reference ontologies:

- DataScientia LiveKnowledge
- Schema.org

Unfortunately, due to the unique characteristics of the EndUser and Sport ETypes in this project, no existing ontology meets the requirements. As a result, a very basic ontology has been developed which can be seen below.

7.1.1 Schema.org

The primary ontology that was taken into consideration was Schema.org. This decision was made based on the fact that this is the one of the most widely used/known schema knowledge bases. Multiple types could be found on the site that was fitting for our project. The types that have been chosen to be included into our reference ontologies are the following:

- Thing > Intangible
- Thing > Intangible > Trip
- Thing > Intangible > Trip > BusTrip
- Thing > Intangible > Trip > TrainTrip
- Thing > Place
- Thing > Place > CivicStructure
- Thing > Place > CivicStructure > BusStop
- Thing > Place > CivicStructure > TrainStation

7.1.2 DataScientia LiveKnowledge

From the OSM Lightweight Ontology only the *Region* type was needed to be taken into the reference for completion.

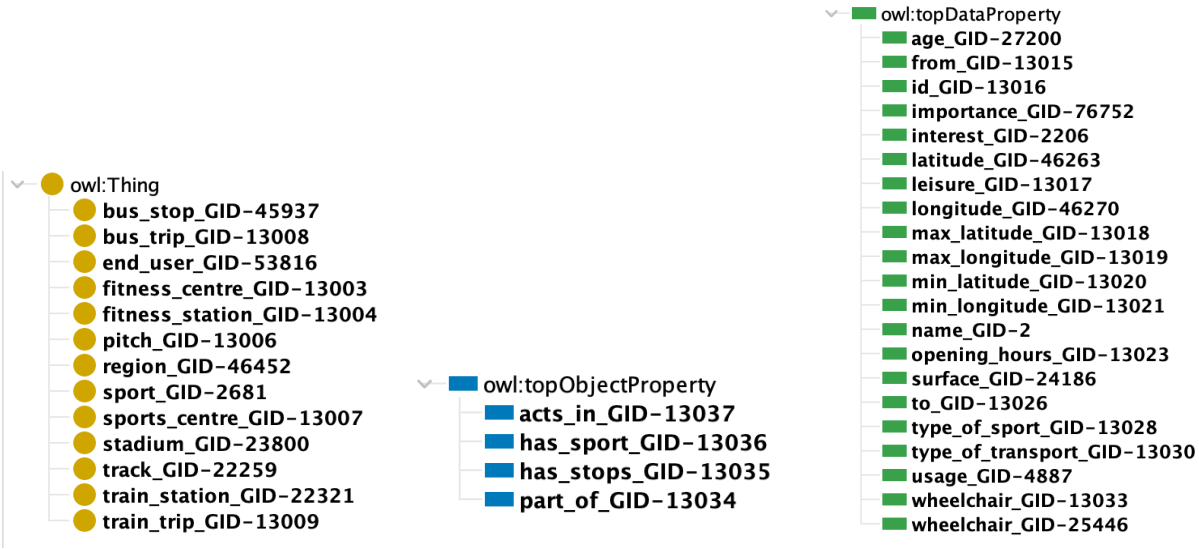
7.2 Teleology

This section outlines the bottom-up Knowledge Definition phase of the kTelos process. Its objective is to construct a teleology that matches the project's purpose and data. This means establishing a teleology that aligns with the requirements formulated as Competency Questions (CQs).

Six main areas have been identified that had to be connected:

- Transportation routes
- Transportation stops
- Facilities
- Sports
- EndUser
- Region

The types in these main areas have then been defined based on the Concept Labels from the previous section, so aligned with the UKC, and connected with the use of object properties in Protege while retaining the data properties that we have defined before, just with the new Concept Labels. The results can be seen in the following three pictures in Protege (yellow dots: Protege classes, blue rectangles: Protege object properties, green rectangles: Protege data properties).



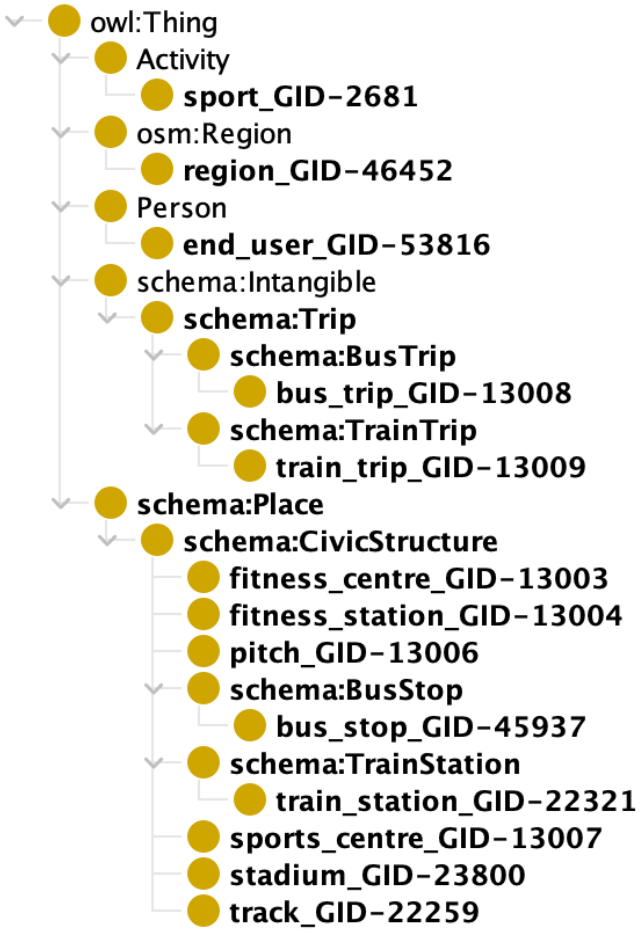
How each of the object properties were used to connect the types can bee seen in the following table.

Domain	ObjectProperty	Range
end_user_GID-53816	acts_in_GID-13037	region_GID-46452
fitness_centre_GID-13003, fitness_station_GID-13004, pitch_GID-13006, sports_centre_GID-13007, stadium_GID-23800, track_GID-22259	has_sport_GID-13036	sport_GID-2681
bus_trip_GID-13008, train_trip_GID-13009	has_stops_GID-13035	bus_stop_GID-45937, train_station_GID-22321
bus_stop_GID-45937, bus_trip_GID-13008, fitness_centre_GID-13003, fitness_station_GID-13004, pitch_GID-13006, sports_centre_GID-13007, stadium_GID-23800, track_GID-22259, train_station_GID-22321, train_trip_GID-13009	part_of_GID-13034	region_GID-46452

7.3 Teleontology

This section details the middle-out Knowledge Definition phase of the kTelos process. The aim is to integrate the project-specific teleology with the general-purpose Lightweight Ontology to create a Teleontology.

How the two were connected can be seen in the following Protege structure.



8 Data Definition

This section focuses on the final phase of the methodology, known as Data Definition. In contrast to the previous phase, this section is organized along a singular dimension that differentiates between producer and consumer processes. Unlike earlier stages, there is no distinct separation between knowledge and data activities here, as this phase integrates both into a unified data structure. This structure combines the knowledge frameworks established in the preceding section with the aligned dataset, starting from cleaned and aligned data resources and the teleontology. The objective is to create a comprehensive Knowledge Graph that encapsulates both layers.

Addressing the diversity in meaning is crucial for generating a Knowledge Graph that meets the initial project goals. Therefore, the last phase of the iTelos methodology is divided into three key activities: Entity Matching, Entity Identification, and Entity Mapping.

8.1 Entity Matching

When managing and integrating data from various datasets, entities in the real world, denoted by their values, can be depicted through a variety of properties and their corresponding values. This scenario gives rise to what is commonly referred to as the entity matching problem. This

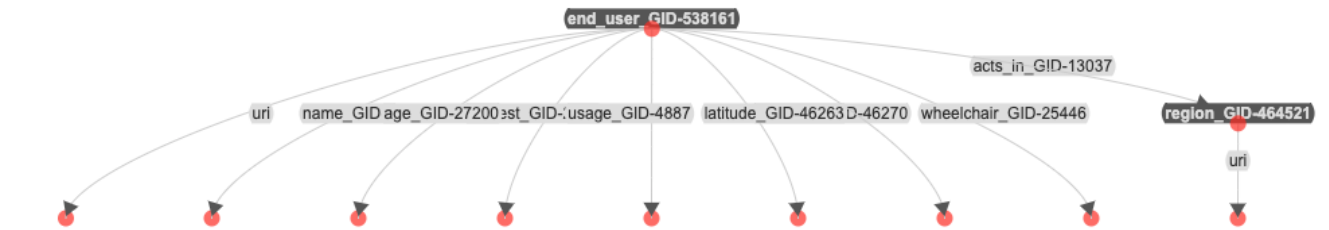
problem manifests in two primary areas of concern. Firstly, at the schema layer, there arises the challenge of identifying the appropriate set of properties across different datasets where multiple representations of the same real-world entity may exist. This necessitates a careful comparison and selection process to ensure that the representations are consistent and accurately reflect the entity in question. Secondly, at the data layer, there is the critical task of assigning the correct property values when multiple representations share the same properties but present different values. This requires a meticulous approach to reconcile these differences to achieve a unified and accurate representation of each entity. During the course of this project, it is noteworthy that neither of these challenges were encountered while evaluating and utilizing the available datasets. This absence of issues related to entity matching at both the schema and data layers indicates a level of consistency and coherence in the datasets that simplifies the processes of data integration and analysis.

8.2 Entity Identification

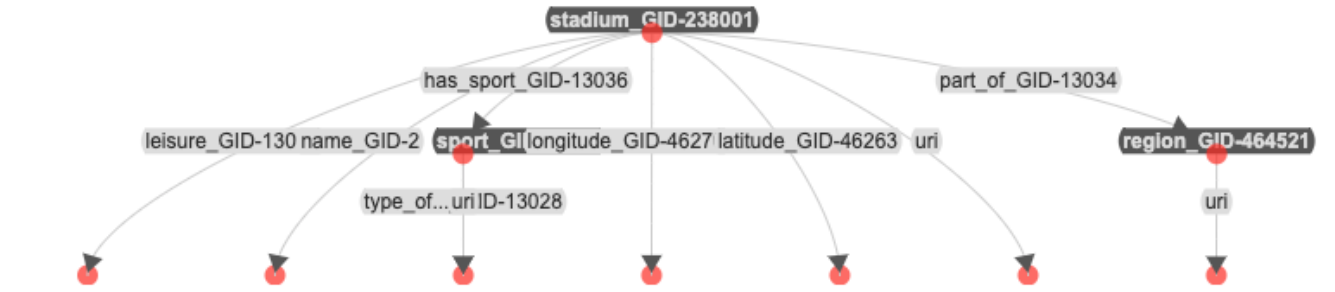
The following step involves recognizing an entity in a single dataset and then applying the same identification method if the same entity appears in two or more distinct forms across various datasets. In the data collection phase, careful attention was given to assigning unique identifiers (IDs) to each data entry, both when sourcing data from online platforms and when creating datasets manually. For instance, a region ID was assigned to all entities with a spatial component within the Trentino region. Similarly, IDs were allocated to end users to facilitate their identification. For data acquired from the internet, fields that acted as unique identifiers were already present, eliminating the need for any further identification efforts.

8.3 Entity Mapping

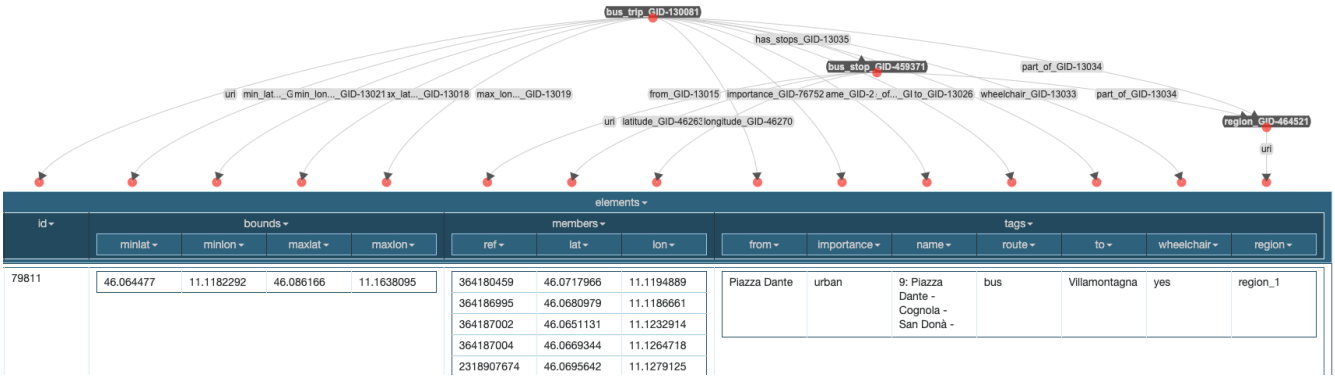
The final activity is focused on seamlessly integrating the information framework outlined in the teleontology with the corresponding data values found in the datasets. This process involves numerous mapping operations that effectively address the entity matching challenge. The task of entity mapping is carried out utilizing the Karma Linker tool. Below are some examples from the process.



id	name	age	interest	usage	lat	lon	wheelchair	acts_in
1	Paolo	25	all	practice	46.0748	11.1217	no	region_1
2	Patrick	20	fitness	practice	46.0659	11.1526	no	region_1



leisure	name	sport	lon	lat	id	region
stadium		soccer	10.9466532	45.7383564...	relation/669...	region_1
stadium	Stadio Briamasco	soccer	11.1144498...	46.0642568...	way/24860...	region_1
stadium	Campo da Calcio AC Nago (chiuso)	soccer	10.9029658...	45.8730441...	way/28591...	region_1



On the producer's end, files are kept separate to facilitate their potential use in various applications. The Knowledge Graph (KG) that is generated is structured both linguistically, aligning with UKC concepts, and conceptually, organized using a teleontology. This approach results in the creation of multiple KG files, one for each KG that needs to be developed.

The examples provided above illustrate the method of associating data with the relevant Knowledge and Data schemas. Within each dataset, specific data properties corresponding to the teleontology have been identified. Furthermore, these examples highlight the process of linking columns across different datasets that contain related values. For instance, the method by which the region type is connected to identifiers in other datasets, and how the actual value, such as the region name, is found in yet another dataset, yet all are unified through the use of identifiers.

On the consumer's side, the separate files are integrated into a single file to fulfill a specific purpose. The final KG meets the requirements derived from the user's objectives (Competency Questions), but this will be proved during the next evaluation section. In practical terms, a file for each dataset model has been created using Karma, and these files have been concatenated to produce the final KG.

9 Evaluation

This section outlines the evaluation conducted at the conclusion of the entire iTelos methodology process, encompassing both the producer and consumer stages, and assesses the final outcome. The evaluation criteria, detailed below, address both the Knowledge Layer and the Data Layer.

9.1 Knowledge Layer Evaluations

The iTelos methodology employs a series of metrics for these evaluations, among which Coverage stands out as particularly significant. Coverage measures the extent to which a Knowledge Graph (KG) encompasses a given body of knowledge. It is applied in evaluating the Knowledge Layer with two distinct aims:

1. Primary Objective (Teleontology vs. Competency Questions): This focuses on the degree to which the final KG fulfills the Competency Queries. It assesses the extent of the teleontology's coverage of entities and properties identified from the Competency Questions.
2. Secondary Objective (Teleontology vs. Reference Ontologies): This examines how thoroughly the teleontology encompasses the entity types and properties derived from the reference ontologies.

9.1.1 Teleontology

Below is a summary table that reflects the total count of entity types, object properties, and data properties, which are instrumental in calculating coverage.

	Counts
Etypes	24
Object Properties	4
Data Properties	21

9.1.2 Teleontology vs CQs

Taking into account Table 2 and the Competency Questions outlined in the Purpose Formalization section, in relation to the final figures from the above table of the Teleontology, the coverage for entity types, object properties, and data properties is determined as follows. For instance, for the entity type, with a defined set of Competency Question Entities (CQ_E), the coverage of entity types (Cov_E) by the Teleontology (T) is calculated in this manner:

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

where CQ_E represents the count of entity types derived from the Competency Questions, and T_E signifies the total number of entity types present in the Teleontology.

Following, a table is provided that presents the final evaluation metrics, based on the coverage of entity types, object properties, and data properties.

	Etypes	Cov_E	Object Properties	Cov_{OP}	Data Properties	Cov_{DP}
Total identified from CQs	13		4		14	
Total defined for the project	24	100%	4	100%	21	100%

The table indicates that, for each evaluated criterion, the final Teleontology encompasses a greater or equal number of entity types, data properties, and object properties. This increase is attributed to the evolution of the project's specific knowledge design choices and requirements, which were initially not fully determined. Throughout the project's development, these aspects were refined, leading to the establishment of a more comprehensive and effective knowledge structure to achieve the project's goals.

9.1.3 Teleontology vs. Reference Ontologies

In the context of the ontologies used and comparing the final figures of the Teleontology, the coverage for entity types (Cov_E), object properties (Cov_{OP}), and data properties (Cov_{DP}) is determined in the following manner. For instance, regarding entity types, with a given set of Reference Ontologies (RO), the coverage of entity types (Cov_E) by the Teleontology (T) is calculated as follows:

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

Here, RO_E represents the count of entity types identified from the Reference Ontologies, and T_E indicates the total count of entity types in the Teleontology.

A table is provided below that shows the final evaluation metrics, highlighting the coverage of entity types, object properties, and data properties.

	Etypes	Cov_E	Object Properties	Cov_{OP}	Data Properties	Cov_{DP}
Schema.org ontology						
Total in the ontology	806		-		1474	
Total reused in the project	8	~1%	0	0%	4	<1%
OSM ontology						
Total in the ontology	791		0		0	
Total reused in the project	1	<1%	0	0%	0	0%

For Schema.org, both the entity types and data properties coverage are not optimal. This is primarily because Schema.org encompasses a vast ontology, while our project is highly specific, resulting in a relatively low coverage percentage. Regarding the OSM ontology, only one entity type was utilized since it served as a supplementary ontology for our project, and given that there are no data properties within the OSM ontology, none were employed.

9.2 Evaluating the Data Layer

The evaluation of the Data Layer is specifically focused on the main goal, which is to assess its effectiveness in meeting the intended purpose. This involves examining the density and intercon-

nectedness of the Knowledge Graph (KG). The connectivity of the KG is assessed at two critical points:

- At the end of the development process, to gauge the overall interconnectedness of the KG;
- Throughout the construction phase, to determine how the inclusion of each dataset contributes to the overall connectivity of the KG.

Connectivity within a KG is analyzed through two key dimensions:

1. Entity connectivity, which measures the level of linkage among the various entities within the KG.
2. Property connectivity, which assesses the level of linkage between the entities of the KG and their associated property values.

The connectivity between entities and properties is quantified using a connectivity matrix, as depicted in the following table.

	sport	region	end_user	...
sport	0|0 tot: 0			
region		0|0 tot: 0		
end_user		0 tot:0	0|..|0 tot: 0	
bus_trip		0 tot:0		
train_trip		0 tot:0		
...				

Each cell within the connectivity matrix illustrates the extent to which entities, identified by their etype, are interconnected. The matrix's main diagonal highlights the proportion of null values within the data properties of the entities corresponding to each etype. Here, each value corresponds to a specific property, with the final value aggregating all properties. The table below (legend) serves as a guide to match each value with its respective data property, as per their sequence.

The other cells of the matrix depict the level of connectivity between entities, specifically showing the percentage of object properties that are null. The rows represent entities that reference other entities. Similar to data properties, the table below provides detailed insights into which object property values are being referenced.

	sport	region	end_user	...
sport	id|type_of_sport			
region		id|name		
end_user		act_in	age|id|interest|latitude|longitude|name|usage|visits	
bus_trip		part_of		
train_trip		part_of		
...				

This document presents just a sample of the matrix. For a comprehensive view, please consult the complete matrix and its accompanying legend in the GitHub repository files in the Evaluation directory.

9.3 Queries and Their Relevance to Competency Questions

To finalize the assessment phase, competency questions were converted into SPARQL queries to evaluate the knowledge graph's ability to meet the project's objectives.

9.3.1 Competency question 1

Competency Question: "Paolo lives in Trento and is in love with sports. Which facilities can Paolo use in Trento that is in walking distance to him?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>

SELECT ?name ?leisure ?sportName
WHERE {
    BIND("Paolo" AS ?queryName)
    BIND(2 AS ?walkingDistance)

    ?user etype:name_GID-2 ?userName .
    ?user etype:latitude_GID-46263 ?userLat .
    ?user etype:longitude_GID-46270 ?userLon .

    ?facility etype:leisure_GID-13017 ?leisure .
    ?facility etype:name_GID-2 ?name .
    ?facility etype:latitude_GID-46263 ?facilityLat .
    ?facility etype:longitude_GID-46270 ?facilityLon .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .

    FILTER(?userName = ?queryName)
    BIND(omgeo:distance(?userLat, ?userLon, ?facilityLat, ?facilityLon) AS ?distance)
    FILTER(?distance < ?walkingDistance)
}
```

Results:

	name	leisure	sportName
1	"Juta Wellness Center"	"fitness_centre"	"fitness"
2	"Palestra Defant's Club"	"fitness_centre"	"fitness"
3	"Mrs. Sporty"	"fitness_centre"	"fitness"
4	"Campo "Carlo Prada""	"pitch"	"soccer"
5	"Campo atletica CONI "Covi-Postal""	"sports_centre"	"athletics"
6	"Stadio Briamasco"	"stadium"	"soccer"

9.3.2 Competency question 2

Competency Question: "Patrick lives in Povo and would like to go to a gym in the morning to work out. Which is the closest gym to him that is open at a specific time?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>

SELECT ?name ?leisure ?sportName ?distance ?openingHours WHERE {
    BIND("Patrick" AS ?queryName)

    ?user etype:name_GID-2 ?userName .
    ?user etype:latitude_GID-46263 ?userLat .
    ?user etype:longitude_GID-46270 ?userLon .
    ?user etype:interest_GID-2206 ?userSportInterest .

    ?facility etype:leisure_GID-13017 ?leisure .
    ?facility etype:name_GID-2 ?name .
    ?facility etype:latitude_GID-46263 ?facilityLat .
    ?facility etype:longitude_GID-46270 ?facilityLon .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?facility etype:opening_hours_GID-13023 ?openingHours .
    ?sport etype:type_of_sport_GID-13028 ?sportName .

FILTER(?userName = ?queryName)
    FILTER(?sportName = ?userSportInterest)

    BIND(omgeo:distance(?userLat, ?userLon, ?facilityLat, ?facilityLon) AS ?distance)
}
ORDER BY ASC(?distance)
LIMIT 2
```

Results:

	name	leisure	sportName	distance	openingHours
1	"Mrs. Sporty"	"fitness_centre"	"fitness"	"2.8194347149479513"^^xsd:float	"Mo-Fr 07:00-21:00; Sa 07:00-18:00"
2	"Star Club"	"fitness_centre"	"fitness"	"4.5701257091358 63"^^xsd:float	"Tu-Fr 09:30-22:00;Sa 09:00-12:00"

9.3.3 Competency question 3

Competency Question: "José has a son and who is in a wheelchair and they live in Trento. He wants his son to enjoy playing football. Which transportation route options do they have that is wheelchair accessible?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?name ?leisure ?sportName ?tripName ?userDistance ?facilityDistance ?totalI
    BIND("JosÃ©" AS ?queryName)
```

```
    ?user etype:name_GID-2 ?userName .
    ?user etype:latitude_GID-46263 ?userLat .
    ?user etype:longitude_GID-46270 ?userLon .
    ?user etype:interest_GID-2206 ?userSportInterest .
    ?user etype:wheelchair_GID-25446 ?wheelchairNeed .

    ?facility etype:leisure_GID-13017 ?leisure .
    ?facility etype:name_GID-2 ?name .
    ?facility etype:latitude_GID-46263 ?facilityLat .
    ?facility etype:longitude_GID-46270 ?facilityLon .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    ?trip etype:type_of_transport_GID-13030 ?typeOfTransport .
    ?trip etype:wheelchair_GID-13033 ?wheelchairAccessible .
    ?trip etype:name_GID-2 ?tripName .
    ?trip etype:has_stops_GID-13035 ?startStop .
    ?startStop etype:latitude_GID-46263 ?startSlat .
    ?startStop etype:longitude_GID-46270 ?startSlon .

FILTER(?userName = ?queryName)
    FILTER(?sportName = ?userSportInterest)
    FILTER(?wheelchairAccessible = ?wheelchairNeed)

    BIND(omgeo:distance(?userLat, ?userLon, ?startSlat, ?startSlon) AS ?userDistance)
    FILTER(?userDistance < 1)

    ?trip etype:has_stops_GID-13035 ?endStop .
    ?endStop etype:latitude_GID-46263 ?endSlat .
    ?endStop etype:longitude_GID-46270 ?endSlon .
    BIND(omgeo:distance(?facilityLat, ?facilityLon, ?endSlat, ?endSlon) AS ?facilityDistanc
    BIND((?userDistance + ?facilityDistance) AS ?totalDistance)
    FILTER(?totalDistance < 1)
}
ORDER BY ASC(?totalDistance)
LIMIT 10
```

Results:

	name	leisure	sportName	tripName	userDistance	facilityDistance	totalDistance	wheelchairAccessible	startSlat	startSlon	endSlat	endSlon
1	"Campo "Mirko Bonetti""	"pitch"	"soccer"	"8: Mattarello - Piazza Dante - Centochiavi"	"0.16654076010129956"^^xsd:float	"0.0570600203222465"^^xsd:float	"0.22360078"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.0927358"	"11.1185106"
2	"Campo "Mirko Bonetti""	"pitch"	"soccer"	"3: Villazzano 3 - P. Dante - Gardolo - Cortesano"	"0.16654076010129956"^^xsd:float	"0.0708975127145485"^^xsd:float	"0.23743826"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.0926912"	"11.1200641"
3	"Campo "Mirko Bonetti""	"pitch"	"soccer"	"8/: Centochiavi - Piazza Dante - Mattarello"	"0.1977164353293706"^^xsd:float	"0.0570600203222465"^^xsd:float	"0.25477645"^^xsd:float	"yes"	"46.0647351"	"11.1222519"	"46.0927358"	"11.1185106"
4	"Campo "Mirko Bonetti""	"pitch"	"soccer"	"8: Centochiavi - Piazza Dante - Mattarello"	"0.1977164353293706"^^xsd:float	"0.0570600203222465"^^xsd:float	"0.25477645"^^xsd:float	"yes"	"46.0647351"	"11.1222519"	"46.0927358"	"11.1185106"
5	"Campo sportivo Meano"	"pitch"	"soccer"	"3/: Villazzano 3 - P. Dante - Gardolo - Cortesano"	"0.16654076010129956"^^xsd:float	"0.09876411161246162"^^xsd:float	"0.26530486"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.1282784"	"11.1202984"
6	"Campo sportivo Meano"	"pitch"	"soccer"	"3: Villazzano 3 - P. Dante - Gardolo - Camparta - Cortesano"	"0.16654076010129956"^^xsd:float	"0.09876411161246162"^^xsd:float	"0.26530486"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.1282784"	"11.1202984"
7	"Campo sportivo Meano"	"pitch"	"soccer"	"3: Villazzano 3 - P. Dante - Gardolo - Cortesano"	"0.16654076010129956"^^xsd:float	"0.09876411161246162"^^xsd:float	"0.26530486"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.1282784"	"11.1202984"
8	"Campo "Mirko Bonetti""	"pitch"	"soccer"	"3: Villazzano 3 - P. Dante - Gardolo - Cortesano"	"0.16654076010129956"^^xsd:float	"0.10178936333707596"^^xsd:float	"0.26833013"^^xsd:float	"yes"	"46.0651131"	"11.1232914"	"46.0920065"	"11.119035"

9.3.4 Competency question 4

Competency Question: "Lucia is a Volleyball fan in Rovereto and loves to watch volleyball matches, but she is not that into long walks and she prefers buses. Which urban bus route should she take to the next match that involves the least amount of walking?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>

SELECT DISTINCT ?name ?leisure ?sportName ?tripName ?typeOfTransport ?userDistance ?facilit
    BIND("Lucia" AS ?queryName)

    ?user etype:name_GID-2 ?userName .
    ?user etype:latitude_GID-46263 ?userLat .
    ?user etype:longitude_GID-46270 ?userLon .
    ?user etype:interest_GID-2206 ?userSportInterest .

    ?facility etype:leisure_GID-13017 ?leisure .
    ?facility etype:name_GID-2 ?name .
    ?facility etype:latitude_GID-46263 ?facilityLat .
    ?facility etype:longitude_GID-46270 ?facilityLon .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    ?trip etype:type_of_transport_GID-13030 ?typeOfTransport .
    ?trip etype:name_GID-2 ?tripName .
    ?trip etype:has_stops_GID-13035 ?startStop .
    ?startStop etype:latitude_GID-46263 ?startSlat .
    ?startStop etype:longitude_GID-46270 ?startSlon .

FILTER(?userName = ?queryName)
    FILTER(?sportName = ?userSportInterest)
    FILTER(?typeOfTransport = "bus")
```

```
    BIND(omgeo:distance(?userLat, ?userLon, ?startSlat, ?startSlon) AS ?userDistance)
    FILTER(?userDistance < 5)

    ?trip etype:has_stops_GID-13035 ?endStop .
    ?endStop etype:latitude_GID-46263 ?endSlat .
    ?endStop etype:longitude_GID-46270 ?endSlon .
    BIND(omgeo:distance(?facilityLat, ?facilityLon, ?endSlat, ?endSlon) AS ?facilityDistanc
    BIND((?userDistance + ?facilityDistance) AS ?totalDistance)
    FILTER(?totalDistance < 10)
}
ORDER BY ASC(?totalDistance)
LIMIT 10
```

Results:

	name	leisure	sportName	tripName	typeOfTransport	userDistance	facilityDistance	totalDistance	startSlat	startSlon	endSlat	endSlon
1	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.45609987598485646"^^xsd:float	"4.934798782044648"^^xsd:float	"5.390898"^^xsd:float	"45.8903772"	"11.0342751"	"45.8785467"	"10.8906829"
2	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.47322614072762753"^^xsd:float	"4.934798782044648"^^xsd:float	"5.408025"^^xsd:float	"45.8905981"	"11.0340124"	"45.8785467"	"10.8906829"
3	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.45609987598485646"^^xsd:float	"4.954039213616975"^^xsd:float	"5.410139"^^xsd:float	"45.8903772"	"11.0342751"	"45.8783575"	"10.8906618"
4	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.47322614072762753"^^xsd:float	"4.954039213616975"^^xsd:float	"5.427265"^^xsd:float	"45.8905981"	"11.0340124"	"45.8783575"	"10.8906618"
5	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.45609987598485646"^^xsd:float	"5.413342199895871"^^xsd:float	"5.869442"^^xsd:float	"45.8903772"	"11.0342751"	"45.8715904"	"10.8752846"
6	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.45609987598485646"^^xsd:float	"5.424140783615866"^^xsd:float	"5.880241"^^xsd:float	"45.8903772"	"11.0342751"	"45.8714749"	"10.8750524"
7	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.47322614072762753"^^xsd:float	"5.413342199895871"^^xsd:float	"5.886568"^^xsd:float	"45.8905981"	"11.0340124"	"45.8715904"	"10.8752846"
8	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"0.47322614072762753"^^xsd:float	"5.424140783615866"^^xsd:float	"5.897367"^^xsd:float	"45.8905981"	"11.0340124"	"45.8714749"	"10.8750524"
9	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"3.1789700605148967"^^xsd:float	"4.934798782044648"^^xsd:float	"8.113769"^^xsd:float	"45.8687988"	"11.0142265"	"45.8785467"	"10.8906829"
10	"Palestra Romarzollo"	"sports_centre"	"volleyball"	"Bus B332: Rovereto <=> Bolognano"	"bus"	"3.1789700605148967"^^xsd:float	"4.954039213616975"^^xsd:float	"8.133009"^^xsd:float	"45.8687988"	"11.0142265"	"45.8783575"	"10.8906618"

9.3.5 Competency question 5

Competency Question: "Ginevra lives in Villazzano and she is a professional athlete. Which track has a tartan surface and where she needs to walk the least before and after practices while taking public transport?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?leisure ?sportName ?surface ?tripName ?userDistance ?facilityDistance ?totalDistanc
    BIND("Ginevra" AS ?queryName)

    ?user etype:name_GID-2 ?userName .
    ?user etype:latitude_GID-46263 ?userLat .
    ?user etype:longitude_GID-46270 ?userLon .
    ?user etype:interest_GID-2206 ?userSportInterest .

    ?facility etype:leisure_GID-13017 ?leisure .
```

```
    ?facility etype:latitude_GID-46263 ?facilityLat .
    ?facility etype:longitude_GID-46270 ?facilityLon .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?facility etype:surface_GID-24186 ?surface .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    ?trip etype:type_of_transport_GID-13030 ?typeOfTransport .
    ?trip etype:name_GID-2 ?tripName .
    ?trip etype:has_stops_GID-13035 ?startStop .
    ?startStop etype:latitude_GID-46263 ?startSlat .
    ?startStop etype:longitude_GID-46270 ?startSlon .

FILTER(?userName = ?queryName)
    FILTER(?sportName = ?userSportInterest)
    FILTER(?surface = "tartan")

    BIND(omgeo:distance(?userLat, ?userLon, ?startSlat, ?startSlon) AS ?userDistance)
    FILTER(?userDistance < 1)

    ?trip etype:has_stops_GID-13035 ?endStop .
    ?endStop etype:latitude_GID-46263 ?endSlat .
    ?endStop etype:longitude_GID-46270 ?endSlon .
    BIND(omgeo:distance(?facilityLat, ?facilityLon, ?endSlat, ?endSlon) AS ?facilityDistanc
    BIND((?userDistance + ?facilityDistance) AS ?totalDistance)
    FILTER(?totalDistance < 2)
}
ORDER BY ASC(?totalDistance)
LIMIT 10
```

Results:

	leisure	sportName	surface	tripName	userDistance	facilityDistance	totalDistance	startSlat	startSlon	endSlat	endSlon
1	"track"	"athletics"	"tartan"	"R25: Trento <=> Bassano del Grappa"	"0.5682063623622785"^^xsd:float	"0.36578670078550946"^^xsd:float	"0.9339931"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
2	"track"	"athletics"	"tartan"	"R26: Trento <=> Borgo Valsugana"	"0.5682063623622785"^^xsd:float	"0.36578670078550946"^^xsd:float	"0.9339931"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
3	"track"	"athletics"	"tartan"	"R25: Trento <=> Bassano del Grappa"	"0.5682063623622785"^^xsd:float	"0.429016910025397"^^xsd:float	"0.99722326"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
4	"track"	"athletics"	"tartan"	"R26: Trento <=> Borgo Valsugana"	"0.5682063623622785"^^xsd:float	"0.429016910025397"^^xsd:float	"0.99722326"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
5	"track"	"athletics"	"tartan"	"R25: Trento <=> Bassano del Grappa"	"0.5682063623622785"^^xsd:float	"0.4596052776409246"^^xsd:float	"1.0278116"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
6	"track"	"athletics"	"tartan"	"R26: Trento <=> Borgo Valsugana"	"0.5682063623622785"^^xsd:float	"0.4596052776409246"^^xsd:float	"1.0278116"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
7	"track"	"athletics"	"tartan"	"R25: Trento <=> Bassano del Grappa"	"0.5682063623622785"^^xsd:float	"0.47471942556558006"^^xsd:float	"1.0429258"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
8	"track"	"athletics"	"tartan"	"R26: Trento <=> Borgo Valsugana"	"0.5682063623622785"^^xsd:float	"0.47471942556558006"^^xsd:float	"1.0429258"^^xsd:float	"46.0462939"	"11.1391285"	"46.0522445"	"11.4640027"
9	"track"	"athletics"	"tartan"	"R25: Trento <=> Bassano del Grappa"	"0.5682063623622785"^^xsd:float	"0.5449371897341002"^^xsd:float	"1.1131436"^^xsd:float	"46.0462939"	"11.1391285"	"46.051142"	"11.4541682"
10	"track"	"athletics"	"tartan"	"R26: Trento <=> Borgo Valsugana"	"0.5682063623622785"^^xsd:float	"0.5449371897341002"^^xsd:float	"1.1131436"^^xsd:float	"46.0462939"	"11.1391285"	"46.051142"	"11.4541682"

10 Metadata Definition

This section of the report compiles definitions of all metadata established for various resources created throughout the entire process, encompassing both the producer and consumer perspectives. It details the metadata that characterizes not only the project's final output but also the outputs at each stage. Defining this metadata is essential for facilitating the sharing of the produced resources. Thus, it's vital to also specify where this metadata is made available for the purpose of distributing the resources it pertains to (e.g., through the DataScientia catalogs). The section is structured into five main categories with the goal of outlining the metadata associated with every type of resource generated by the project.

10.1 Dataset Metadata

Contains information about various datasets, including their licenses, URLs, keywords, publishers, creators, owners, language, knowledge level, size, name, publication date, description, version, domain, and file format.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguage	DatLevel	DatSize	DatName	DatPublication	DatDescription	DatVersion	DatDomain	DatFileFormat
Open Data Com	https://github.com/patrick-r	fitness,centre,sp	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	752B	fitness_centre_data.csv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	fitness,station,sp	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	2KB	fitness_station_data.tsv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	pitch,sport,facilit	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	45KB	pitch_data.tsv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	sports,centre,spo	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	11K	sports_centre_data.tsv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	stadium,sport,fac	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	717B	stadium_data.tsv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	track,sport,facilit	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	3.3KB	track_data.tsv	31/12/2023	Provides informa	3	sport facilities	tsv
Open Data Com	https://github.com/patrick-r	bus,route,transp	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	356KB	bus_routes_filtered.geojson	31/12/2023	Provides informa	4	transport	geojson
Open Data Com	https://github.com/patrick-r	train,route,transp	OpenStreetMap	OpenStreetMap	OpenStreetMap Foundation	english	Knowledge level	22KB	train_routes_filtered.geojson	31/12/2023	Provides informa	4	transport	geojson
Apache License	https://github.com/patrick-r	end user,other,tr	Patrick Nanys	Patrick Nanys	Patrick Nanys	english	Knowledge level	335B	end_user.tsv	18/2/2024	Provides informa	1	society	tsv
Apache License	https://github.com/patrick-r	region,other,tren	Patrick Nanys	Patrick Nanys	Patrick Nanys	english	Knowledge level	25B	region.tsv	18/2/2024	Contains a singl	1	territorial	tsv

10.2 Knowledge Metadata

Similar to the Dataset Metadata, it contains detailed descriptions of the knowledge resources, including their subject, type, creator, owner, language, size, name, publication date, description, version, domain, and file format.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguage	DatName	DatPublicationTi	DatDescription	DatVersion	DatDomain
Creative Commons Attribut	https://schema.org/Place	Place	Schema.org	Schema.org	Schema.org	english	place	02/18/2024	Information relat	3	facility
Creative Commons Attribut	https://schema.org/BusStop	BusStop	Schema.org	Schema.org	Schema.org	english	bus stop	02/18/2024	Information relat	3	transport
Creative Commons Attribut	https://schema.org/TrainStation	TrainStation	Schema.org	Schema.org	Schema.org	english	train station	02/18/2024	Information relat	3	transport
Creative Commons Attribut	https://schema.org/BusTrip	BusTrip	Schema.org	Schema.org	Schema.org	english	bus trip	02/18/2024	Information relat	3	transport
Creative Commons Attribut	https://schema.org/TrainTrip	TrainTrip	Schema.org	Schema.org	Schema.org	english	train trip	02/18/2024	Information relat	3	transport
Creative Commons Attribut	https://schema.org/CivicStructure	CivicStructure	Schema.org	Schema.org	Schema.org	english	civic structure	02/18/2024	Information relat	3	facility
Creative Commons Attribut	https://schema.org/Trip	Trip	Schema.org	Schema.org	Schema.org	english	trip	02/18/2024	Information relat	3	transport
Creative Commons Attribut	https://schema.org/Intangible	Intangible	Schema.org	Schema.org	Schema.org	english	intangible	02/18/2024	Information relat	3	intangible
Creative Commons	https://datascientiafoundation.github	Region	Datascientia Foundation	Datascientia Fou	Datascientia Fou	english	region	02/18/2024	Information relat	3	spatial

10.3 Language Metadata

Focuses on language-related metadata, detailing concepts, creators, owners, languages, sizes, names, publication dates, descriptions, versions, domains, and file formats.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguage	DatSize	DatName	DatPublication	DatDescription	DatVersion	DatDomain	DatFileFormat
Apache-2.0 licer	https://github.cor	sport,facility,tran	Patrick Nanys	Patrick Nanys	Patrick Nanys	english	38 concepts	Language sprea	23/01/2024	Description of all	4	sport,facility,tran	tsv

10.4 Project Metadata

Provides comprehensive details on this project, including titles, URLs, keywords, types, descriptions, start and end dates, funding agencies, inputs, outputs, coordinators, and observations.

prjTitle	prjURL	prjKeywords	prjType	prjDescription	prjStartDate	prjEndDate	prjFundingAgency	prjInput	prjOutput	prjCoordinator	prjObservations
Sports Facilities And Transportation In Trento	https://patrick-nanys.github.io/Sport	sport,facility,transportation,trentino	Knowledge Resource Generation	This Knowledge Graph is design	45208	45340	Datascientia foundation	The project incorporate	A Knowledge Graph was developed,	Simone Bocca	

10.5 People Metadata

Lists information about individuals working on the project, including their identifiers, first names, last names, emails, nationalities, genders, affiliations, and personal webpages.

comIdentifier	firstName	lastName	email	nationality	gender	affiliation	personalWebpage
patrick-nanys	Patrick	Nanys	patrick.nanys@s	hungarian	M	Datascientia, Kn	https://www.linkedin.com/in/patrick-nanys/
hakim-rabia	Abdelhakim	Rabia	abdelhakim.rabi	french	M	Datascientia, Kn	https://www.linkedin.com/in/rabia-hakim/

11 Open Issues

The initiative has successfully achieved the objectives outlined by the competency queries, marking a significant milestone in the project's development. However, a few unresolved issues remain that warrant further attention. Notably, the project team developed a specialized ontology tailored to the specific needs of this endeavor. This raises the question of whether there exist alternative ontologies that could potentially encompass the concepts of Activity and Person Etypes more comprehensively. Despite exhaustive searches across multiple platforms and databases, no such ontologies have been identified.

Additionally, the method employed for querying opening hours presents challenges. While the storage format for opening hours aligns with the standards set forth by Schema.org, this approach is not without its drawbacks, particularly in terms of query complexity. The current methodology is not ideally suited for effective querying, suggesting a need to explore other ontologies that offer more query-friendly schemas for representing opening hours. Furthermore, there is considerable room for improvement in the realm of query optimization. The execution time for queries can be extensive without the implementation of adequate filtering mechanisms throughout the process. One potential strategy to enhance query performance involves the precomputation of distances for stops, which could significantly reduce query execution times. By addressing these issues and exploring the possibilities for optimization, the project could achieve greater efficiency and effectiveness in its operations.