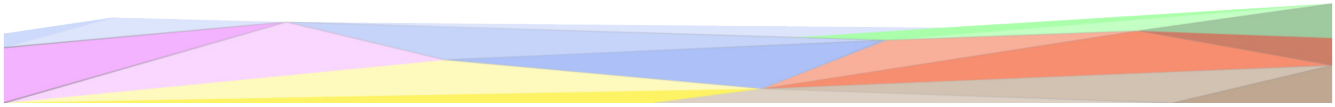
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2023 - Project Report

Document Data:	Reference Persons:
January 23, 2024	Patrick Nanys, Abdelhakim Rabia



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	January 23, 2024	Author1	Document created

1 Purpose and Domain of Interest (DoI)

1.1 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.

1.2 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.

2 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.

2.1 Data Production

The primary focus of this phase is to produce datasets that align with our project's purpose, which revolves 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 resources like Trentino OSM places and OpenStreetMaps will be pivotal in this phase, ensuring that the data is comprehensive and accurate.

2.2 Data Composition

Once the data production phase is complete, the data consumer's role becomes central. The objective in this phase is to integrate the newly produced data with existing high-quality resources to create a cohesive Knowledge Graph.

The steps involved in this phase are:

- **Vertical Composition:** This involves identifying redundant instances of sports facilities or transportation hubs/routes from different datasets and ensuring that only one unique instance is retained in the final Knowledge Graph. For instance, if a particular bus stop is mentioned in both the Trentino OSM places and OpenStreetMaps datasets, it's essential to identify these repetitions and keep only one instance.
- **Horizontal Composition:** This step focuses on establishing relationships between sports facilities and transportation routes/hubs. For example, determining which bus stop or train station is closest to a particular sports facility and how they are interconnected. The Haversine distance can be employed to calculate the distance between two points based on their longitude and latitude, ensuring that the connections made are geographically accurate.

The integration will heavily rely on resources like the Trentino OSM lightweight Ontology, SCHEMA.ORG. The ultimate goal is to create a Knowledge Graph that offers a comprehensive view of sports facilities in Trentino and their connection with transportation, catering to both residents and tourists.

3 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:

- **Scenario 1.** A sports enthusiast is looking for specific sports facilities in Trentino where they can practice their favorite sport.
- **Scenario 2.** A resident wants to find a gym near their residence that aligns with their daily routine.
- **Scenario 3.** A parent is searching for sports facilities in Trentino where their child can practice.

Name	Age	Interests	Usage	Residence	Description
Paolo	25	Climbing	Practice	Trento center	Paolo loves climbing
Patrick	20	Gym	Practice	Povo	He goes to the gym everyday at 6a.m. before work
José	42	Football	Dad	Rovereto	He's the dad of Bernardo who plays football
Lucia	50	Volleyball	Fan	Trento center	Lucia is a fan of Trentino Volley and does not want to miss a single home match
Ginevra	13	Athletics	Practice	Villazzano	She mainly focuses on 5000 m races and goes to the pitch at 10p.m.

Table 1: Table of personas

Scenarios	Persona	CQs	Entity	Properties	Classification
4		4	Bus Stop	id, name, location, timetable	Common
4		4	Train Stop	id, name, location, timetable	Common
4		4	Tram Stop	id, name, location, timetable	Common
1,2,3,4,5		1,2,3,4,5	Sports Facility	id, name, location, sport type, opening hours	Core
1,3,4		4	Sports Venue	id, name, location, events, capacity	Core
2		2	Paolo's Home	id, address, coordinates	Contextual
2		2	Patrick's Home	id, address, coordinates	Contextual
2		2	José's Home	id, address, coordinates	Contextual
2		2	Lucia's Home	id, address, coordinates	Contextual
2		2	Ginevra's Home	id, address, coordinates	Contextual

Table 2: Entities extraction and classification

- **Scenario 4.** A sports fan wants to know the transportation options available to reach a specific sports venue in Trentino.
- **Scenario 5.** An athlete is looking for sports facilities in Trentino that cater to their specific training schedule.

The personas can be seen in the Table of personas.

Taking into account the personas in the scenarios defined, we create Competency Questions (CQs):

- **CQ 1.** Which sports facilities are available in Trentino for a specific sport?
- **CQ 2.** Where can a resident find a gym near their residence that matches their routine?
- **CQ 3.** Can you list the sports facilities in Trentino suitable for children?
- **CQ 4.** What are the transportation options to reach a specific sports venue in Trentino?
- **CQ 5.** Where can an athlete find sports facilities in Trentino that are open during specific hours?

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 classification and Popularity classification.

At the very end we use the extracted entities to create an initial EMR model.

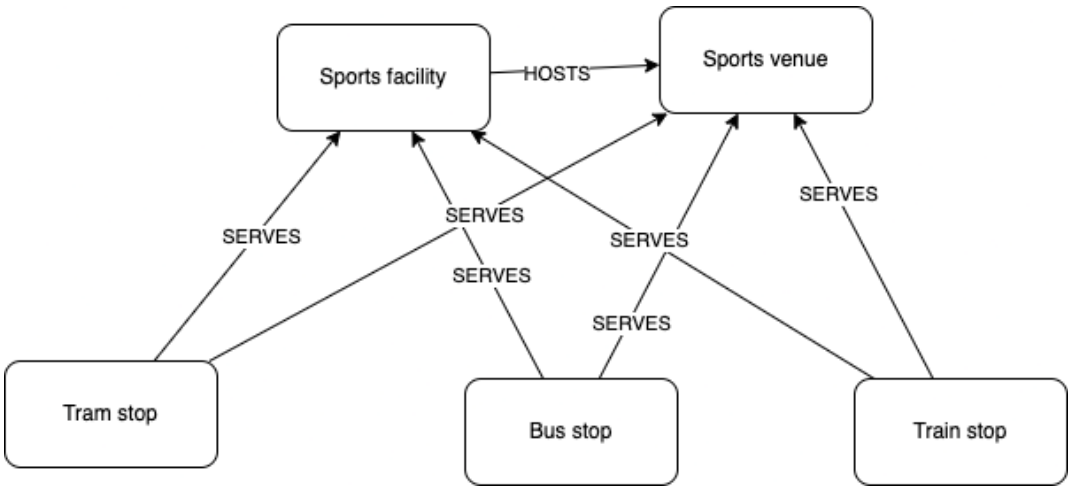


Figure 1: EMR model

4 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.

4.1 Consumer Activities

4.1.1 Knowledge Layer:

- **Sources Description**: 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.
- **Formal Resources Collection**: Accessing OSM's database through its API or data export service was critical for acquiring current and relevant geographic information, specifically about sports facilities and transportation networks in the Trentino Province.
- **Formal Resources Classification**: The OSM data was meticulously classified into 'common', 'core', and 'contextual' categories, aligning the data with the project's goal of developing a Knowledge Graph focused on sports facilities and transportation links.

4.1.2 Data Layer:

The approach in the data layer mirrors that of the knowledge layer, emphasizing the comprehensive nature of the OpenStreetMap data in meeting the project's requirements.

4.1.3 Reflection on Choices Made:

- **Strengths**: Utilizing OSM was instrumental due to its encompassing range of data points, covering all necessary aspects to fulfill the original goal of the project. This comprehensiveness eliminated the need for supplementary data sources, ensuring efficiency in data integration.
- **Weaknesses**: While reliance on an external data source like OSM typically presents limitations in control over specific data nuances, OSM's dynamic and community-driven updates significantly mitigate this issue.

4.1.4 Reasoning Behind the Approach:

The decision to rely exclusively on OSM data was driven by its unparalleled alignment with the project's needs. OSM's extensive coverage of sports facilities and transportation networks in the Trentino Province provided a solid foundation for the Knowledge Graph. This strategic choice reflects the team's commitment to creating a comprehensive, efficient, and data-driven solution to support applications and services aimed at enhancing the accessibility and enjoyment of sports facilities in the region.

5 Language Definition

This section is dedicated to the description of the Language Definition phase. During this phase, our team focused on developing a consistent and precise language framework that serves as the foundation for our project. This phase was crucial for ensuring that the information was accurately represented and understood across all team members and stakeholders.

5.1 Language Definition Sub Activities

In this particular section, as opposed to the other segments of our work, we concentrated exclusively on the knowledge layer. This involved a thorough examination of our ontology, during which we meticulously identified every term that required clear definitions. The purpose of this exercise was to ensure that users of our system have a comprehensive understanding of each entity, data, and object property within our framework. To achieve this, we compiled a detailed list of these identified terms, culminating in the creation of 55 distinct entries within our language spreadsheet. For each entry, we provided an in-depth description, thereby facilitating a deeper understanding of the concepts. Additionally, we introduced concept labels for each term, offering a concise and recognizable identifier. These labels, coupled with their attached descriptions, are designed to enhance user comprehension and navigation through our system, ensuring clarity and precision in the understanding of all the concepts involved.

6 Knowledge Definition

In this particular section of our project, we dedicated our efforts towards the development and structuring of three pivotal files: the ontology, teleology, and teleontology. Our approach in

constructing the ontology was predominantly top-down. This method involved an initial focus on modeling the overarching problem by defining the primary entities, outlining their characteristics, and elucidating the interrelations between them. The top-level entities identified in our ontology framework included:

- Place
  - Facilities
  - Stop
- Sport
- Trip
- Region

This hierarchical structure was intentionally chosen to model the ontology as it effectively encapsulates those components which inherently lack direct data association. Therefore, by integrating them into the ontology, we ensure a comprehensive representation of all elements, including those that are not directly linked to tangible data.

Moving on to the teleologies, we adopted a bottom-up approach. This perspective focuses on incorporating entities that are directly associated with raw, empirical data. The entities integrated into the teleologies include:

- BusStop
- BusTrip
- FitnessCentre
- FitnessStation
- OutdoorPlace
  - Pitch
  - Track
- SportsCentre
- Stadium
- TrainStation
- TranTrip

This bottom-up methodology in teleology allows for a granular focus on data-rich entities, ensuring that all aspects with substantial data backing are duly represented and accounted for.

Finally, for the teleontology, we employed a middle-out approach. This strategy serves as a bridge, intertwining the top-down perspective of the ontology with the bottom-up approach of the teleologies. The result of this integration is a well-rounded teleontology that effectively merges both the theoretical framework and the empirical data elements. This comprehensive teleontology is represented through the detailed depiction of Entities, Object properties, and Data properties in our corresponding figures. This holistic approach ensures that our model is both structurally sound and data-informed, thereby enhancing the overall robustness and utility of our project.

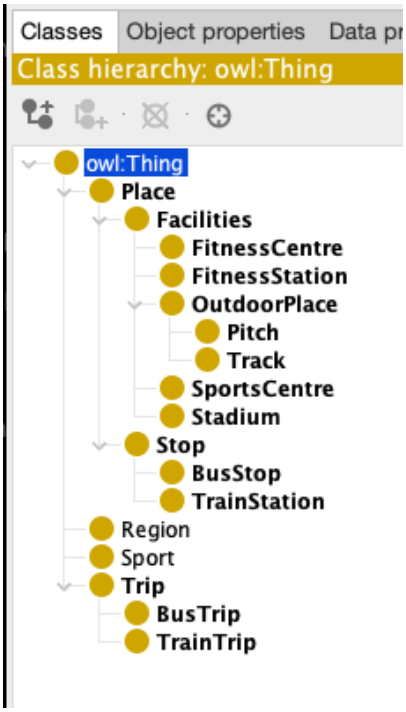


Figure 2: Entities

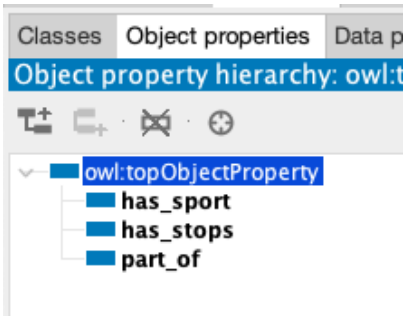


Figure 3: Object properties

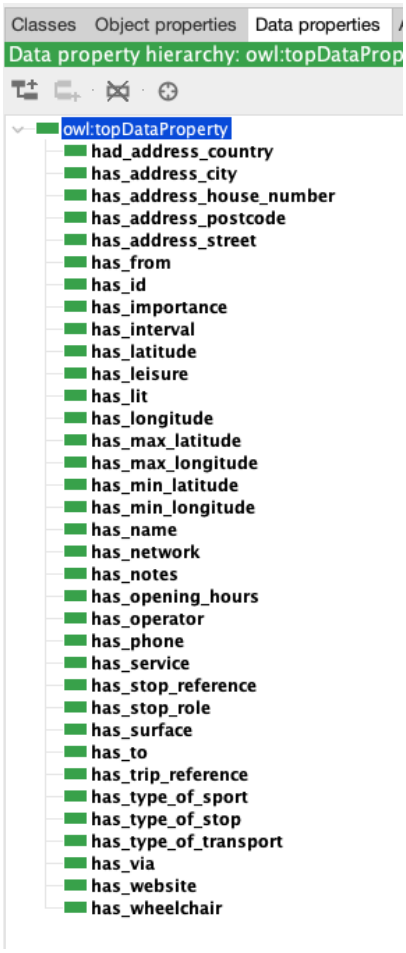


Figure 4: Data properties

7 Data Definition

In the current stage of our project, we embarked on the crucial task of synthesizing the data that has been meticulously cleaned and prepared in the preceding phases. This integration was accomplished by aligning the prepared data with the teleontology framework that was recently developed. The primary objective of this endeavor was to construct a comprehensive knowledge graph. This graph is not just a mere collection of data points; it is a well-structured representation of information that is poised to be instrumental in addressing the initial Competency Queries Constructs (CQs) that were identified at the onset of our project.

The process of linking the cleaned data with the teleontology was executed with remarkable ease, a testament to the thorough and strategic planning undertaken in the earlier stages of our project. This meticulous planning played a pivotal role in simplifying what could have otherwise been a complex and daunting task. As a result of this process, we successfully generated eight Resource Description Framework (RDF) files in the Turtle (TTL) format. These files are not just standalone entities; they are interconnected components that collectively form the backbone of our knowledge graph.

In addition to these RDF files, we have also produced eight corresponding models. These models are more than just technical documents; they are detailed narratives that elucidate the various steps and decisions involved in the data linking process. By articulating these steps, the models provide a clear and comprehensive account of how the individual pieces of data were transformed and integrated to form the cohesive structure of the knowledge graph. Recognizing the importance of reproducibility in scientific and technical endeavors, we have taken the conscientious step of saving these models. The preservation of these models serves multiple purposes. Firstly, it ensures transparency in our methodology, allowing others to understand and evaluate the processes we employed. Secondly, it facilitates reproducibility, enabling other researchers or project teams to replicate our methods in their own endeavors. This aspect is particularly crucial in the realm of data science and knowledge management, where the ability to reproduce results is a cornerstone of validity and reliability.

Furthermore, the saved models act as a valuable resource for future projects. They can be referred to as blueprints or guidelines for similar tasks, thereby streamlining the workflow and reducing the learning curve for future teams. This archival of our methodologies and results not only contributes to the body of knowledge in our field but also stands as a testament to our commitment to best practices and scientific rigor. The current phase of our project has been a significant milestone. We have successfully linked the cleaned data with the teleontology to create a dynamic and informative knowledge graph. This graph is set to play a crucial role in addressing our initial queries. The processes we followed were systematically planned and executed, resulting in a set of RDF files and comprehensive models. The preservation of these models ensures that our work is transparent, reproducible, and serves as a valuable resource for future endeavors. This phase not only marks the culmination of our meticulous planning and preparation but also sets the stage for the practical application of our knowledge graph in addressing the initial questions that guided our research.

8 Evaluation

In this comprehensive analysis, we scrutinize the effectiveness of our Knowledge Graph, focusing on its ability to answer Competency Questions (CQs) that were established at the project's outset. This evaluation is achieved through a series of tailored queries. Notably, our efforts have borne fruit, with four out of five initial CQs being successfully answered by these queries.

8.1 Queries and Their Relevance to Competency Questions

8.1.1 Query 1: Identifying Sports Facilities for a Specific Sport

Competency Question: "Which sports facilities are available in Trentino for a specific sport?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX schema: <https://schema.org/>

SELECT ?facility ?sportName WHERE {
    ?facility a etype:facilities_GID-13002 .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    FILTER(?sportName = "SPECIFIC_SPORT_NAME")
}
```

Explanation and Rationale: This query strategically targets facilities in Trentino that are dedicated to a specific sport. By leveraging the predicates 'etype:facilities' and 'etype:has_sport', it filters out facilities based on the chosen sport. This is instrumental for athletes or enthusiasts who seek a facility that caters specifically to their sport of interest, thus making the search more efficient and targeted.

8.1.2 Query 2: Locating a Gym Based on Proximity and Routine

Competency Question: "Where can a resident find a gym near their residence that matches their routine?"

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

SELECT ?gym ?address WHERE {
    ?gym a etype:fitness_centre_GID-13003 .
    ?gym etype:address_street_GID-13014 ?address .
    ?gym etype:latitude_GID-46263 ?latitude .
    ?gym etype:longitude_GID-46270 ?longitude .
    ?gym etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    FILTER(?sportName = "SPECIFIC_SPORT_NAME") # Replace with the sport of interest
```

```
    # Constants for the Haversine formula
    BIND(6371 AS ?earthRadius) # Earth's radius in kilometers
    BIND(PI() AS ?pi)
    BIND(?pi / 180 AS ?deg2rad)
    BIND(<resident_latitude> AS ?resLat)
    BIND(<resident_longitude> AS ?resLong)
    BIND(<near_threshold> AS ?threshold)

    # Convert degrees to radians
    BIND(?latitude * ?deg2rad AS ?latRad)
    BIND(?longitude * ?deg2rad AS ?longRad)
    BIND(?resLat * ?deg2rad AS ?resLatRad)
    BIND(?resLong * ?deg2rad AS ?resLongRad)

    # Haversine formula
    BIND(SIN((?latRad - ?resLatRad) / 2) * SIN((?latRad - ?resLatRad) / 2) + COS(?resLatRa
    BIND(2 * ATAN2(SQRT(?a), SQRT(1 - ?a)) AS ?c)
    BIND(?earthRadius * ?c AS ?distance)

    FILTER(?distance < ?threshold)
}
```

Explanation and Rationale: This query employs geographic coordinates and the Haversine formula to calculate the proximity of gyms to a resident's location. By incorporating specific sports preferences, the query aligns gyms with an individual's routine, offering personalized options. This approach not only simplifies the search for a suitable gym but also ensures that the recommendations are convenient and relevant to the resident's lifestyle.

8.1.3 Query 3: Unanswered Competency Question

Competency Question: "Can you list the sports facilities in Trentino suitable for children?"

Explanation: The inability to answer this CQ highlights the limitations of our current Knowledge Graph, possibly due to gaps in data or the need for more complex query structures. This observation serves as a valuable insight, guiding future enhancements of the graph.

8.1.4 Query 4: Transportation Options to Sports Venues

Competency Question: "What are the transportation options to reach a specific sports venue in Trentino?"

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

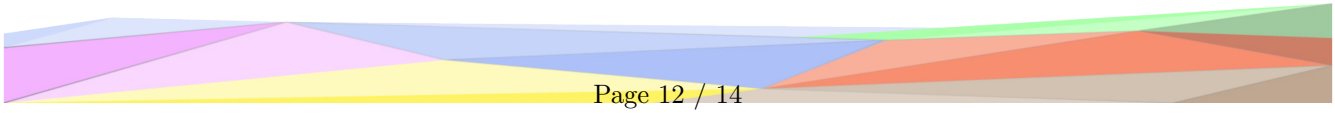
```
SELECT ?facility ?stop ?address WHERE {
    ?facility a etype:facilities_GID-13002 .
    ?facility etype:latitude_GID-46263 ?flat .
    ?facility etype:longitude_GID-46270 ?flon .
    ?facility etype:name_GID-2 ?name .
    ?stop a etype:stop_GID-46571 .
    ?stop etype:latitude_GID-46263 ?slat .
    ?stop etype:longitude_GID-46270 ?slon .

    FILTER(?name = "FACILITY_NAME")

    # Constants for the Haversine formula
    BIND(6371 AS ?earthRadius) # Earth's radius in kilometers
    BIND(PI() AS ?pi)
    BIND(?pi / 180 AS ?deg2rad)
    BIND(<near_threshold> AS ?threshold)

    # Convert degrees to radians
    BIND(?flat * ?deg2rad AS ?latRad)
    BIND(?flon * ?deg2rad AS ?longRad)
    BIND(?slat * ?deg2rad AS ?stopLatRad)
    BIND(?slon * ?deg2rad AS ?stopLongRad)

    # Haversine formula
    BIND(SIN((?latRad - ?stopLatRad) / 2) * SIN((?latRad - ?stopLatRad) / 2) + COS(?stopLat
    BIND(2 * ATAN2(SQRT(?a), SQRT(1 - ?a)) AS ?c)
    BIND(?earthRadius * ?c AS ?distance)

    FILTER(?distance < ?threshold)
}
```

Explanation and Rationale: This query is key in facilitating access to sports facilities by identifying available transportation options. It integrates transportation data within the Knowledge Graph, using geographic coordinates and the Haversine formula to calculate distances. This feature is crucial for planning journeys to sports venues, especially for those relying on public transport or seeking the most convenient routes.

8.1.5 Query 5: Finding Sports Facilities with Specific Opening Hours

Competency Question: "Where can an athlete find sports facilities in Trentino that are open during specific hours?"

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX schema: <https://schema.org/>
```



```
SELECT ?facility ?sportName ?opening_hours WHERE {
    ?facility a etype:sports_centre_GID-13007 .
    ?facility etype:opening hours_GID-13023 ?opening_hours .
    ?facility etype:has_sport_GID-13036 ?sport .
    ?sport etype:type_of_sport_GID-13028 ?sportName .
    FILTER(?sportName = "SPECIFIC_SPORT_NAME") # Replace with the sport of interest
}
```

Explanation and Rationale: Athletes and sports enthusiasts often have specific schedules, making it crucial to find facilities with compatible opening hours. This query assists in locating such facilities, filtering by opening hours and the type of sport. It thereby enables efficient planning and maximizes the utility of sports facilities to suit diverse schedules.

The development and execution of these queries mark a significant advancement in harnessing the potential of our Knowledge Graph. Successfully addressing four out of five initial competency questions underlines the robustness and practical applicability of our Knowledge Graph. It demonstrates its capability to provide targeted, meaningful information in the realm of sports facilities and activities in Trentino, serving as a valuable tool for athletes, residents, and sports enthusiasts alike.

9 Metadata Definition

The process of gathering the metadata was conducted with exceptional attention to detail and precision. We embarked on an exhaustive and meticulous journey through the entirety of the data and schema resources that have been utilized in the course of our project. This thorough exploration was undertaken with the specific intent of incorporating these elements into the metadata section. Our primary objective in doing so was to ensure the comprehensive acknowledgment and mention of every individual and entity that contributed to the development of our final knowledge graph.

In pursuit of this goal, we employed a methodical approach, carefully examining each dataset and resource to ascertain its relevance and significance in the broader context of our project. We recognized the importance of accurately representing the contributions of all parties involved, understanding that the success of our knowledge graph hinges not just on the data itself, but also on the collective expertise and insights of those who have aided in its compilation and refinement.

As such, we took great pains to cross-reference our resources, verifying the accuracy of our data, and ensuring that no contributor, regardless of the extent of their involvement, was overlooked. This process was not only about paying homage to their contributions but also about maintaining the integrity and reliability of our knowledge graph. By meticulously documenting the sources and contributors in our metadata section, we aimed to create a transparent and trustworthy foundation for our work.

This endeavor, though time-consuming and demanding, was essential for the integrity and credibility of our final product. It involved not just a simple review of the resources but a deep dive into the intricacies of each element that played a role in the formation of our knowledge graph. Our commitment to this process reflects our dedication to excellence and our acknowledgment of the

collaborative nature of this project. It is through this rigorous and detailed approach to metadata collection that we have endeavored to craft a knowledge graph that is not only comprehensive and informative but also respectful and appreciative of every individual's contribution to this collaborative effort.

10 Open Issues

This section concludes the current document with final conclusions regarding the quality of the process and final outcome, and the description of the issues that (for lack of time or any other cause) remained open.

- Did the project respect the scheduling expected in the beginning ?
- Are the final results able to satisfy the initial Purpose ?
  - If no, or not entirely, why ? which parts of the Purpose have not been covered ?

Moreover, this section aims to summarize the most relevant issues/problems remained open along the iTelos process. The description of open issues has to provide a clear explanation about the problems, the approaches adopted while trying to solve them and, eventually, any proposed solution that has not been applied.

- which are the issues remained open at the end of the project ?