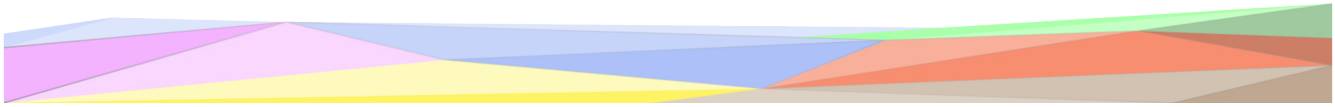
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2024 Project
Sport Facilities & Events in Trentino

Document Data:	Reference Persons:
November 25, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 16, 2024	Christian Sassi, Pietro Bologna	Document created
1.1	October 25, 2024	Christian Sassi, Pietro Bologna	Purpose definition phase
1.2	October 29, 2024	Christian Sassi, Pietro Bologna	Revision of purpose definition phase
1.3	November 6, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Correction to ER model and PF Sheet
2.1	November 7, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Information Gathering phase
2.2	November 9, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Dataset collection
2.3	November 13, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Revision of information gathering phase
3.1	November 20, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Language Definition phase
3.2	November 21, 2024	Christian Sassi, Pietro Bologna, Mouez Khelifi	Revision of language definition phase

1 Introduction

Access to sports facilities and related events is essential for enhancing the quality of life in modern cities and regions. In areas like Trentino Province, providing convenient access to a range of sports infrastructure and events is increasingly important for both residents and visitors. Promoting an active lifestyle and supporting community-driven initiatives are key to strengthening the livability of the region.

To address these needs, we present a project for the Knowledge Graph Engineering course, integrating data on sports facilities and events in Trentino. This Knowledge Graph will offer citizens, tourists, and local authorities a comprehensive, interconnected view of available sports facilities, and related sport events. By combining data on facilities and events, the Knowledge Graph will empower users to make informed decisions that enhance community participation, public health, and the region's sporting culture.

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important role to enhance the reusability of the resources handled and produced during the process. A clear description of the resources as well as of the process (and single activities) developed, provides a clear understanding of the project, thus serving such an information to external readers for the future exploitation 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 as follows:

- 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 Produce role's objectives.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities.
- 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.

You can access the GitHub repository, which contains all the materials used during the project's development, via this link.

2 Purpose Definition

Access to sports facilities and events has a significant impact on the quality of life in regions like Trentino Province. Reliable information on the availability and accessibility of these resources can inspire a more active lifestyle and enhance community engagement. The **purpose** of this project is to develop a Knowledge Graph that consolidates information on sports facilities and events in Trentino, creating a unified resource for residents, tourists, and local authorities. By providing a comprehensive overview, the Knowledge Graph supports informed decision-making, promotes community involvement, enhances public health, and cultivates a vibrant sports culture in the region.

2.1 Informal Purpose

We want to build a Knowledge Graph that brings together all the information on sports facilities and events across Trentino. The goal is to make it easy for people to find sports venues, check out upcoming events, and get involved in physical activities. By offering details on what facilities are available, when events are happening, and what kinds of sports are offered, this project aims to promote an active lifestyle. Ultimately, we want this Knowledge Graph to be a go-to resource for anyone looking to make informed decisions about sports and activities in the region.

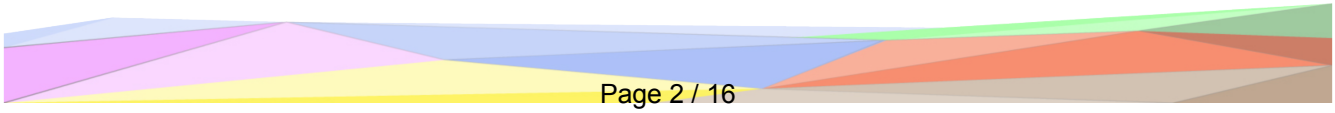
2.2 Domain of Interest (DoI)

After analyzing the purpose, the next step is to define the Domain of Interest (DoI). The DoI provides details about the geographical area and time frame relevant to the project purpose. The Domain of Interest for this project is as follows:

- The **geographical space** for this project is defined by the administrative boundaries of the Trentino Province. We ensure that only sports facilities and centers located within this region are included in our dataset, encompassing both urban and rural locations where sports facilities, such as soccer fields, tennis courts, basketball courts, and other venues, are situated. Additionally, this geographical space applies to the sports events data, ensuring that all events included are those taking place within the Trentino Province. Many of these events are based on those organized for the Festival dello Sport 2024, which brings a concentrated focus on sports culture and activities within one of the main city of the region.
- The **temporal domain** for this project is focused on the year 2024. The data on sports facilities and events reflects the information available for this specific period, capturing the current landscape of sports resources and scheduled events within Trentino Province throughout the year.

2.3 Scenarios definition

Here we define a set of usage scenarios, describing the multiple aspects considered by the project purpose. Four key scenarios are considered:



1. **Weekday**: Across the Trentino, on a typical weekday.
2. **Weekend**: In the province of Trento, on a weekend, when sports facilities may see an higher activity as locals and tourists alike participate in and spectate at community sports events.
3. **Holidays**: In Trentino, during the holiday periods (e.g. Christmas, Summer, etc.), when a high influx of tourists is expected.
4. **Festival dello Sport:** In Trento, during the *Festival dello Sport 2024*. During this time, a variety of sports events and exhibitions are scheduled, attracting both residents and visitors.

2.4 Personas

Now, we define a set of real users acting within the scenarios defined above. Each Persona is defined over a specific features included in the main purpose.

1. **Luca** is a 35-year-old engineer working in Trento. He is passionate about outdoor sports, especially padel.
2. **Anna** is a 21-year-old university student from Verona, studying in Trento. She enjoys playing volleyball with her friends.
3. **Matteo** is a 42-year-old tourist visiting Trento during the Christmas holidays. He has a passion for winter sports.
4. **Camilla** is a 24-years-old a volunteer student at the annual ”Festival dello Sport” in Trento and assist visitors discovering the wide range of activities available throughout the city.

2.5 Competency Questions (CQs)

From the above information, here we have the list of Competency Questions (CQs) created considering the personas in the defined scenarios.

1. **CQ1**: Luca inquires about available padel courts in Trento after 7 PM.
2. **CQ2**: Luca also asks if there are any padel events during the weekend of the Festival dello Sport.
3. **CQ3**: Luca asks if there will be events in Trentino that have Sara Errani as a guest.
4. **CQ4**: Anna wants to know what sports can be practiced in Trentino.
5. **CQ5**: Spending the weekend with friends in Folgaria, Anna asks if there are any lighted volleyball or beach volleyball courts available throughout the day.
6. **CQ6**: As a volleyball enthusiast, Anna would like to know if there will be any volleyball events during the summer holidays of 2024.
7. **CQ7**: Matteo also wants to know if there are any skiing events held in Stelvio National Park during the winter season.

8. **CQ8**: Not finding what he's looking for, Matteo asks if there are any sport events during his vacation in Trentino.
9. **CQ9**: A visitor asks Camilla about the events happening today, October 10, at the Festival dello Sport.
10. **CQ10**: While volunteering at the Festival dello Sport, Camilla becomes interested in tennis and wants to know if there are any tennis-related events and if tennis courts are available when she returns to Molveno for the weekend.

2.6 Concepts identification

From the scenarios, personas and CQs we extract the following entities with their properties:

Scenarios	Personas	Competency Questions	Entities	Properties	Focus
1	1	CQ1	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			SportFacility	id, legalName, openingHours, telephone, email, url, sports, location	Core
			Location	id, address, municiplaity	Core
			PadelCourt	id, name, covered	Core
2, 4	1	CQ2	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
			Organization	id, name	Contextual
1,2,3,4	1	CQ3	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
			Guest	id, givenName, familyName	Contextual
1,2,3,4	2	CQ4	EndUser	id, name, surname, birth_date, occupation	Contextual
			Sport	id, name	Core
2	2	CQ5	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			SportFacility	id, legalName, openingHours, telephone, email, url, sports, location	Core
			Location	id, address, municiplaity	Core
			VolleyballCourt	id, name, surface, lit	Core
3	2	CQ6	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
1,2,3	3	CQ7	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
			Location	id, address, municiplaity	Core
1,2,3	3	CQ8	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
4	4	CQ9	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Event	id, name, startDate, endDate, sports, location, organization, guests	Core
			Organization	id, name	Contextual
2, 4	4	CQ10	EndUser	id, givenName, familyName, birthDate, occupation	Contextual
			Sport	id, name	Core
			Event	id, name, start_date, end_date, sports, location, organization, guests	Core
			SportFacility	id, legalName, openingHours, telephone, email, url, sports, location	Core
			Location	id, address, municiplaity	Core
			TennisCourt	id, name, surface, lit, covered	Core

Figure 1: PF Sheet

2.7 ER model definition

The ER diagram, which graphically represents the knowledge gathered in the prior stages. This ER diagram provides detailed information for a technician to delve deeper into the project. The diagram is based on the entity types (ETypes) and attributes identified before.

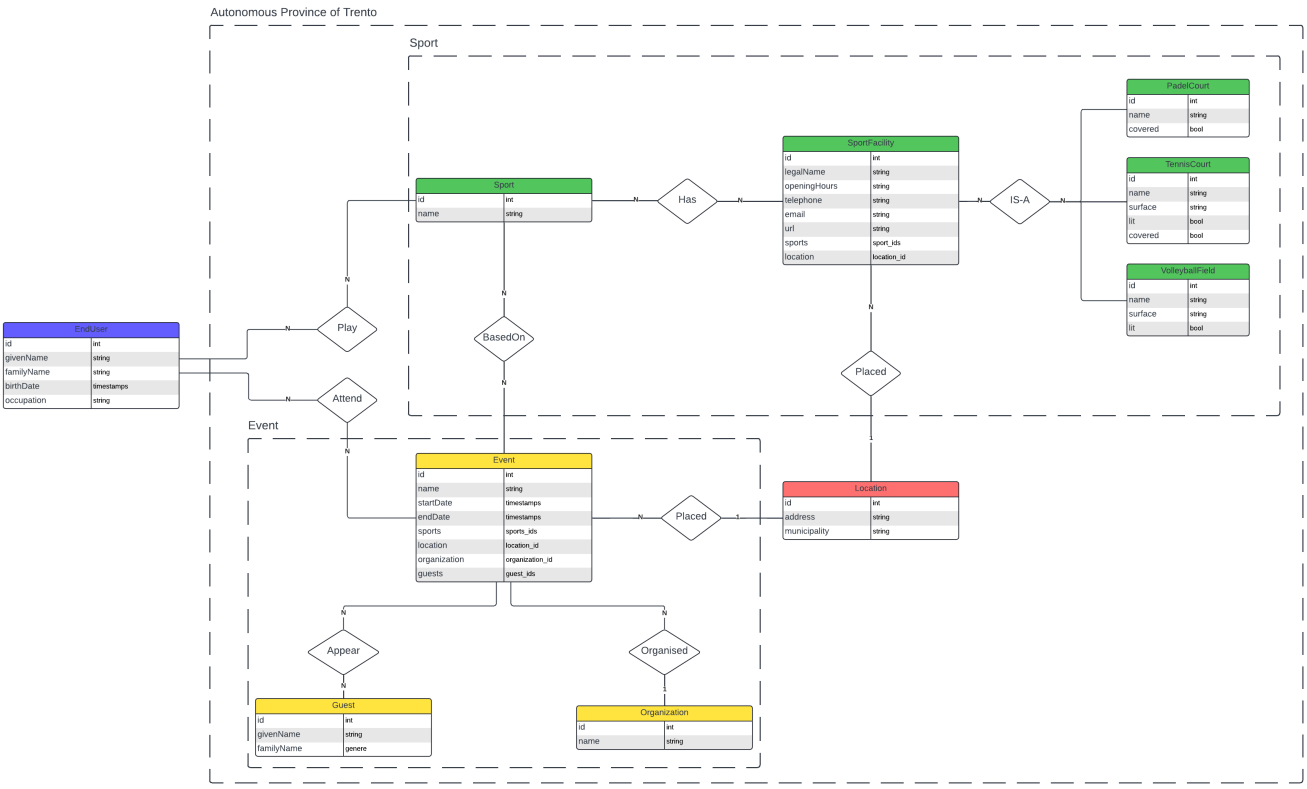


Figure 2: ER model

The following ETypes shown in Figure 2 have been identified to illustrate how the different entities interact within the model, providing a clear and coherent structure for representing information about sports facilities and events in Trentino.

1. **SportFacility**: Represents a facility dedicated to sports, where individuals can participate in various physical activities.
    - `id`: Unique identifier for each sports facility.
    - `legalName`: Name of the sports facility.
    - `openingHours`: Operating hours for the facility.
    - `telephone`: Contact phone number for the facility.
    - `email`: Contact email address for the facility.
    - `url`: Website link for more information.
    - `sports`: List of sports available at the facility, linked to sport IDs.
    - `location`: Address or general location information.
2. **PadelCourt**: A court dedicated to padel, a racquet sport.
    - `id`: Unique identifier for the padel court.
    - `name`: Name of the padel court.

- `covered`: Indicates if the court is covered.

3. **TennisCourt**: A court dedicated to tennis.
   - `id`: Unique identifier for the tennis court.
   - `name`: Name of the tennis court.
   - `surface`: Type of surface on the court (e.g., clay, grass).
   - `lit`: Indicates if the court has lighting.
   - `covered`: Indicates if the court is covered.
4. **VolleyballField**: A field dedicated to volleyball.
   - `id`: Unique identifier for the volleyball field.
   - `name`: Name of the volleyball field.
   - `surface`: Type of surface on the field (e.g., sand, grass).
   - `lit`: Indicates if the field has lighting.
5. **Sport**: Represents a specific type of sport.
   - `id`: Unique identifier for each sport.
   - `name`: Name of the sport.
6. **Event**: Represents an organized sports event, including details on participation and scheduling.
   - `id`: Unique identifier for each event.
   - `name`: Name of the event.
   - `startDate`: Start date of the event.
   - `endDate`: End date of the event.
   - `sports`: Sports included in the event, linked to sport IDs.
   - `location`: Location where the event is held, linked to location ID.
   - `organization`: Organization responsible for the event, linked to organization ID.
   - `guests`: List of guest participants, linked to guest IDs.
7. **Guest**: Represents a guest or participant involved in a sports event.
   - `id`: Unique identifier for each guest.
   - `givenName`: The first name of the guest.
   - `familyName`: The last name of the guest.
8. **Organization**: Represents an organization responsible for hosting or coordinating sports events.
   - `id`: Unique identifier for each organization.
   - `name`: Name of the organization.

9. **Location**: Represents a physical location where sports facilities and events are held.
    - `id`: Unique identifier for each location.
    - `address`: Full address of the location.
    - `municipality`: City in which the location is situated.
10. **EndUser**: Refers to the end user who will utilize the service..
    - `id`: Unique identifier for each user.
    - `givenName`: The first name of the end user.
    - `familyName`: The last name of the end user.
    - `birthDate`: Birth date of the user.
    - `occupation`: Occupation of the user (e.g. student, worker, etc.).

This ER model is also structured around several relationships:

1. **Placed**: It contextualizes where the facilities and events are physically situated, allowing users to identify precise addresses or areas for sports facilities and event venues.
2. **Has**: It defines the kinds of sports activities or facilities available within each SportFacility, supporting users in finding facilities based on specific sports or equipment.
3. **IS-A**: It allows the model to recognize these entities as specific kinds of SportFacility, enabling queries about both general facilities and specific facility types, based on shared attributes.
4. **Organise**: It identifies the entity responsible for organizing an event, allowing users to find events hosted by specific organizations or understand the event's affiliation.
5. **Appear**: It enables user queries about the guest lists.
6. **BasedOn**: Identifies the sport associated with an event.
7. **Play**: Indicates the sport played by the user.
8. **Attend**: Indicates the event joined by the user.

Due to the broad accessibility of the sports world, we decided not to focus solely on the university domain; instead, we envisioned this service as accessible to everyone in Trentino. Building on this idea, our goal became developing a service capable of helping citizens find relevant sports events and facilities throughout Trentino.
From a temporal perspective, we chose to support requests specific to the year 2024, while from a geographical perspective, we based our service on the entire Trentino region.
Both these temporal and geographical choices have their strengths and weaknesses. Regarding the temporal choice, the main limitation is that we are restricted to the year 2024, as sports events data from previous years is often incomplete or inconsistent, with some years missing data entirely. Looking forward, another limitation is that our service remains fixed to 2024, as we cannot guarantee that future data providers will address these critical gaps in coverage. By

focusing on 2024, however, we ensure that our service operates reliably, leveraging a dataset we know to be comprehensive—this is a major strength of our approach.
As for the geographical choice, as outlined in our purpose definition, our service is designed to function across the entire Trentino region, making it one of the service's greatest strengths. However, it is important to note that for some entities, especially sports facilities, certain fields may be missing, such as contact information—particularly for less documented facilities. This also applies to specific types of facilities, such as volleyball fields, where some sports facilities may have incomplete data. Despite these gaps, we chose to model these properties because they align with the project's purpose: providing a complete and unified view of sports facilities and events in the region. By including these attributes in the model, we establish a structure that can accommodate future data as it becomes accessible. This approach ensures that the Knowledge Graph can be expanded and refined over time, allowing it to better serve the needs of residents, tourists, and local authorities seeking information about sports opportunities in Trentino.

3 Information Gathering

Information gathering is the second phase in the iTelos methodology for knowledge graph engineering. This phase involves collecting and processing the resources needed to build the final Entity Graph (EG), following the purpose defined in the first phase. The process works with three types of datasets: data value datasets, knowledge datasets (ontologies), and language datasets. Beyond just collecting data, this phase aims to improve the quality and reusability of the gathered information through cleaning and standardization steps. This organized approach ensures the knowledge graph is built using high-quality, compatible data that effectively meets its intended purpose.

3.1 Sources identification

The first activity within the Information Gathering phase involves identifying and accessing relevant sources of information. This step includes examining the input sources provided and potentially exploring additional sources if the initial inputs prove insufficient.

The goal is to enable data reuse by locating pre-existing data sources that can provide the required information, by actively searching for datasets aligned with the project's objectives, ensuring the efficient use of available resources
Given the heterogeneous nature of data, multiple types of sources must be considered to effectively cover the diverse aspects of the project. To ensure high-quality and reliable information, the emphasis is placed on using ”high-quality” sources, such as Pagine Gialle and catalogs of interoperable, reusable datasets. These sources are often maintained in repositories or live data catalogs (e.g., OpenData Trentino), where data and other relevant resources are published and made accessible.

a. Data Layer

The Data Layer encompasses the diverse data sources used to populate and structure information about sports facilities and events in the Trentino region. The goal is to gather comprehensive, high-quality datasets that provide detailed and relevant information. The selected sources include publicly accessible datasets and online directories, ensuring broad coverage of the various aspects related to sports and recreation.

Below is an overview of the primary data value sources utilized in this project:

1. **Overpass Turbo**
    - *Description*: A web-based tool for querying OpenStreetMap (OSM) data, allowing detailed searches for features like sports facilities, parks, and roads using the Overpass API. Ideal for extracting geospatial data.
    - *Access Link*: overpass-turbo.eu
2. **OpenData Trentino**
    - *Description*: The official Open Data Portal of Trentino, offering access to datasets across various sectors like transportation, tourism, and public services. Supports data reuse and innovation.
    - *Access Link*: dati.trentino.it
3. **Pagine Gialle**
    - *Description*: An online directory of businesses in Italy, categorized by industry. Provides contact details, addresses, and user reviews for services like restaurants and sports facilities.
    - *Access Link*: paginegialle.it
4. **Festival dello Sport**
    - *Description*: An annual sports event in Trento featuring panels, workshops, and live sports. It gathers athletes and sports enthusiasts, providing insights into the sports industry.
    - *Access Link*: ilfestivaldellosport.it
5. **Comune di Trento**
    - *Description*: The official website of Trento's municipal administration, offering information on public services, events, and datasets related to urban planning and local services.
    - *Access Link*: www.comune.trento.it
6. **List of Municipalities**
    - *Description*: List of municipalities of Trentino province.
    - *Access Link*: github.com/alihamzaunitn/kdi-educationtrentino

For this project, it wasn't possible to determine if one data source was better than another. This is because, for the type of data we handle—especially event-related data—available resources were limited. Choosing between them would have left us with very little data or, in the worst case, none at all.

b. Knowledge Layer
In designing the Knowledge Layer for Sports Facilities and Events in Trentino, we initially considered using the **General Transit Feed Specification (GTFS)**, a standard for public transit data. However, as the project's focus is not on public transit schedules or routes, the GTFS schema was deemed unsuitable. Instead, we opted to use **Schema.org**, which better aligns with the project's requirements and is more readily available for use.

Additionally, some property names in our model do not strictly follow the Schema.org vocabulary. This deviation was necessary to better address the specific needs and purposes of our project. This custom adaptation ensures the Knowledge Graph accurately represents the context and requirements of project's purpose.

3.2 Datasets collection

The Data Collection phase focuses on acquiring relevant datasets to populate the project's knowledge base. This phase involves several key steps, including the selection of data sources, gathering the data, and ensuring its quality and completeness.

The data collection process employed three main approaches:

- **Direct Downloads**: Some datasets, such as those available from *OpenData Trentino*, were collected directly in formats like CSV or JSON. These datasets are freely downloadable, ensuring that they are immediately usable without requiring additional extraction steps. However, it should be noted that in some circumstances, especially for *OpenData Trentino*, web scraping was necessary to gather these resources more efficiently and quickly.
- **API**: For sources like *Overpass Turbo*, an automated approach was used to gather data. The Overpass API, specifically designed for querying OpenStreetMap data, was utilized to collect information on sports facilities across the Autonomous Province of Trento. This method facilitated large-scale and precise data collection.

```
[out:json][timeout:60];
// Define the Trentino administrative area
{{geocodeArea:Trentino}}->.searchArea;
// Fetch data for sports centers within Trentino
(
  nwr["leisure"~"sports_centre|pitch"](area.searchArea);
);
out body;
>;
out skel qt;
```

- **Web Scraping**: In cases where structured data was not readily accessible, custom Python scripts were developed to extract information from HTML-based directories. Web scraping was used to collect data from sources such as *Festival dello Sport*, *Pagine Gialle*, *OpenData Trentino* and *Comune di Trento*. This approach allowed for the extraction of relevant data that was otherwise embedded in unstructured web pages.

3.3 Datasets cleaning

The primary goal of data cleaning is to eliminate noise and inconsistencies, enabling more accurate and meaningful analysis. In our case, cleaning the dataset involves addressing various issues to improve its quality. The process begins by identifying and removing duplicate entries within the dataset. Next, irrelevant data (e.g. records from years other than 2024) is filtered out to ensure the dataset remains focused and relevant to the analysis.

The data cleaning process was specifically necessary for events sourced from Trentino's open data portal. While other data sources in our project could be processed automatically, the events dataset required manual processing for two key reasons. First, as our knowledge graph focuses specifically on sports events, we needed to carefully identify and extract these from a dataset that included a wide variety of event types (cultural, social, artistic, etc.). Second, the source data structure was inconsistent, with event information spread across various fields and often embedded within long descriptive text rather than in structured fields. For instance, in the Mori dataset, the "*VII^ Coppa Italia ParaHockey*" event had its timing information, location details, and participant categories scattered throughout a long description field, requiring careful manual extraction.

3.4 Datasets standardization

Standardizing a dataset involves transforming the data into a consistent format or structure. First, we ensure that all measurements use uniform units, such as converting all dates into timestamps. For our dataset, we adopted a consistent datetime format "YYYY-MM-DD HH:MM:SS" (e.g., "2024-02-10 20:30:00"), with events missing specific time information defaulting to "00:00:00". Next, we focus on maintaining consistent naming conventions, making sure terms and locations are written uniformly (e.g., "Trento" vs. "Trento city"), and standardizing categorical data to follow a consistent format (e.g., using "True" and "False" instead of "Yes" and "No"). This process is very important in our case, as it facilitates the seamless integration of data from multiple sources into a cohesive dataset.

a. Data Layer
For the data layer part, we performed a cross-check among the different cleaned datasets to ensure consistency, particularly for attributes that reference IDs from other datasets. For instance, in the *Sport Facility* dataset, attributes like "*sports*" and "*location*" reference the IDs from the Sport and Location datasets, respectively. This ensures that the data is interconnected, allow-

ing accurate representation of which sports are available at a given facility and its geographical context. By validating these cross-references, we establish a coherent data structure.

b. Knowledge Layer

After data acquisition and pre-processing to ensure quality and consistency, the tables below presents the final structure of our datasets for sports facilities and their associated events.

I. Sport

CSV File	Columns
Sport	id, name
Sport Facility	id, legalName, openingHours, telephone, email, url, sports, location
PadelCourt	id, name, covered
TennisCourt	id, name, surface, lit, covered
VolleyballField	id, name, surface, lit

II. Events

CSV File	Columns
Event	id, name, startDate, endDate, sports, location, organization, guests
Guest	id, givenName, familyName
Organization	id, name

III. Location

CSV File	Columns
Location	id, address, municipality

IV. End User

CSV File	Columns
EndUser	id, givenName, familyName, birthDate, occupation

3.5 Data cleaning and standardization explanation

In this section, we provide a detailed explanation, accompanied by snapshots, of the steps we took to transform the raw dataset into a cleaned and subsequently standardized dataset for our project. We began with a collection of unstructured and inconsistent event data, encountering common issues such as missing fields, duplicate entries, and non-uniform formatting. Specifically, when working with data obtained from OpenData Trentino, we faced additional challenges: the datasets often had different structures and included an extensive number of attributes per event (in some cases, more than 40 fields).

As shown in Figure 3, many of these attributes were not relevant to our project goals, containing extraneous or redundant information that needed to be filtered out. Our process involved identifying and selecting only the most pertinent fields to ensure a streamlined and standardized dataset ready for integration into the knowledge graph.

	0	1	2	3	4	5	6	7	8	9	10	11
0	remoteId	published	modified	Titolo	Eventuale sottotitolo	Tipo di evento	Identificativo	Date ed orari	Descrizione breve	Argomenti	Descrizione completa	Galleria immagini
1	aa5e02c315	2023-06-21	2023-06-2	CAREZZE SONORE - 03 febbraio 2023	Percorso per ...	Evento culturale			Percorso per ...	Istruzione	Percorso per ...	
2	5e1ec15eb8	2023-06-21	2023-06-2	GIORNATA DELLA MEMORIA 2023 - UN LIBRO DI SANGUE	Atto unico di Renzo ...	Evento culturale			a cura del CLUB ...	Istruzione	In occasione della ...	
3	c50349209b	2023-06-21	2023-06-2	POMERIGGI DEI PICCOLI ALDENERI 21/01/2023	Iniziativa promossa dal	Evento culturale			In collaborazione ...	Demografia	SABATO 21 GENNAIO ...	
4	4510e8694b	2023-06-21	2023-06-2	POMERIGGI DEI PICCOLI ALDENERI 04/02/2023	Iniziativa promossa dal	Evento culturale			In collaborazione ...	Demografia	SABATO 04 FEBBRAIO	
5	0b55cc7386	2023-06-21	2023-06-2	PILATES MAMMA - BIMBO 13/03/2023	Percorso di ginnastica	Evento culturale			Percorso di ...	Istruzione	Percorso di ginnastica	
6	965d63872e	2023-06-21	2023-06-2	POMERIGGI DEI PICCOLI ALDENERI 18/02/2023	Iniziativa promossa dal	Evento culturale			In collaborazione ...	Demografia	SABATO 18 FEBBRAIO	
7	4e2838ed97	2023-06-21	2023-06-2	MINI-CORSO DI TEATRO 17/02/2023		Evento culturale			Organizzato ...	Istruzione	La Filodrammatica ...	

Figure 3: Raw dataset format

Then, during the cleaning process, we resolved inconsistencies by removing duplicate entries and correcting formatting issues. Specifically, we eliminated numerous irrelevant attributes from the raw data, such as *"published"* and *"modified"*, as they were not essential for our project objectives. Additionally, we extracted useful information (such as dates, times, and locations) from the *"descrizione completa"* field, where this data was often embedded. As shown in Figure 4, this meticulous cleaning process allowed us to reduce noise and significantly improve the overall quality and relevance of the dataset.

	0	1	2	3	4	5	6	7
0	name	startData	endDate	location	lat	lon	organization	guests
1	Agor dei Talenti 2024	21/9/24 10.00	21/9/24 19.00	Parco Albere Aldeno				
2	Agor dei Talenti 2024	22/9/24 10.00	22/9/24 19.00	Parco Albere Aldeno				
3	Corso di Difesa Personale - Associazione Judo Zen'Yo Destra Adige	12/11/24 19.00	12/11/24 20.30	Via Martignoni n. 34, Aldeno	45.978.739	11.091.057	Judo Zen'Yo Destra Adige	

Figure 4: Dataset after cleaning process

Finally, we standardized the dataset to ensure consistency and compatibility with our project requirements. This involved restructuring key fields such as event names, dates, and locations to adhere to a uniform format. For instance, we standardized date entries to the format "YYYY-MM-DD HH:MM:SS" and ensured that all textual information was consistent in structure and style.

Additionally, we conducted a "dataset integration" phase, where we linked related datasets to create meaningful connections. For example, we connected the Guest CSV with the Event CSV by associating the *id* of each guest with the corresponding event instance. This step allowed us to establish relationships between datasets, enriching the overall data structure and improving its utility for our knowledge graph. As shown in Figure 5, the event *"ALESSANDRO COLOMBO: TAGLIATO PER VIVERE"* has the guest number *8* which corresponds to *"Alessandro Colombo"* in the Guest CSV in Figure 6 .

	0	1	2	3	4	5	6	7
0	id	name	startDate	endDate	sports	location	organization	guests
1	1	LIBRI DI SPORT: LA VETTA DELLA VITA Libri di sport	2024-10-10 17:00:00		3	1	18	172
2	2	GAZZA CAF	2024-10-11 09:30:00		3	2	18	57
3	3	ANDREA LANFRI: SENZA LIMITI	2024-10-11 15:00:00		3	3	18	23
4	4	CATHERINE DESTIVELLE: UNA VITA IN VERTICALE	2024-10-12 10:00:00		3	4	18	39
5	5	DENIS URUBKO: COLPEVOLE D'ALPINISMO	2024-10-12 12:30:00		3	5	18	61
6	6	VALENTINA CAFOLLA: APNEA GLACIALE	2024-10-12 18:00:00		43	1	18	251
7	7	MATTEO ZURLONI: SPEED(Y) GONZALES	2024-10-10 15:00:00		4	3	18	173
8	8	ALESSANDRO COLOMBO: TAGLIATO PER VIVERE	2024-10-12 15:00:00		4	3	18	8

Figure 5: Dataset after standardization process

	0	1	2
0	**id**	**givenName**	**familyName**
1	1	Aaron	March
2	2	Adriano	Galliani
3	3	Agnese	Duranti
4	4	Alba	De Silvestro
5	5	Alberta	Santuccio
6	6	Alessandra	Sensini
7	7	Alessandra	Campedelli
8	8	Alessandro	Colombo

Figure 6: Guest CSV

4 Language Definition

In the Language Definition phase, the focus is on adapting the language, i.e. word and concepts, required to accurately represent the information and relationships within our Knowledge Graph (KG). This phase builds upon the outputs of previous phases, such as the Purpose Formalization Sheet, Entity-Relationship (ER) Model, and the resource set. It aims to create a purpose-specific language that connects the project's goals with the data's conceptual representation.

The objective is to ensure that the concepts used are aligned with the project's goals and effectively represent the data's entity types (ETypes), attributes, and relationships. This is achieved by leveraging existing ontologies and vocabularies to identify standardized definitions. A cascade approach is applied in this process: first, each concept is checked for its presence in the **Universal Knowledge Core (UKC)**. If the concept is not found in the UKC or does not align with the project's requirements, alternative sources such as **schema.org** and **OpenStreetMap wiki** are consulted. If the concept remains undefined, a new ConceptID is created, following a structured format such as `KGE24-SportFacilities&SportEvents-8XXX`. In this phase, it is important to note that it was not possible to create fully automated scripts to extract these language definitions. This is because each word had to be carefully reviewed from both a grammatical and semantic perspective to ensure it aligns with the project's characteristics. For example, if we take the entity "*Guest*", the UKC provides us with the same word but with a meaning that differs from the one given in this project. Alternatively, when looking at "*celebrity*" in UKC, the meaning partially reflects that of "*Guest*" (in the project's context), but the word itself does not align. This is also a particular case, as it is not possible to find a perfect semantic and grammatical match in any resource. The definition of "*Guest*" is a custom language definition uniquely created for this project.

To establish a structured and coherent language definition, we organize the information into three distinct tables:

- Table 1: Dedicated to Entity Types (ETypes).
- Table 2: Focused on relationships.
- Table 3: Covering attributes.

In conclusion, by the end of this phase, all entities, relationships, and attributes in the KG will adhere to the newly formalized vocabulary. This guarantees consistency and clarity in how information is represented.

ConceptID	**Word-en**	**Gloss-en**
UKC-2593	sport	An active diversion requiring physical exertion and competition.
UKC-56	event	Something that happens at a given place and time.
UKC-43416	organization	The persons (or committees or departments etc.) who make up a body for the purpose of administering something.
UKC-695	location	The persons (or committees or departments etc.) who make up a body for the purpose of administering something.
UKC-53492	user	A person who makes use of a thing; someone who uses or employs something.
KGE24-SportFacilities &SportEvents-8001	guest	Represents a guest or participant involved in a sports event.

Table 1: EType concept labels and descriptions.

ConceptID	**Word-en**	**Gloss-en**.
UKC-85982	placed	Situated in a particular spot or position.
UKC-104711	organise	Create (as an entity).
UKC-101132	appear	Character on stage or appear in a play, etc.
UKC-97761	play	Participate in games or sport.
UKC-105477	attend	Be present at (meetings, church services, university), etc.
UKC-92536	basedOn	Being derived from (often followed by 'on' or 'upon').
UKC-103527	have	Have or possess, either in a concrete or an abstract sense.

Table 2: Relationships concept labels and descriptions.

ConceptID	**Word-en**	**Gloss-en**.
UKC-36247	identification	Evidence of identity; something that identifies a person or thing (full form of ”*id*”).
UKC-33531	given_name	The name that precedes the surname.
UKC-33528	family_name	The name used to identify the members of a family (as distinguished from each member’s given name).
schema.org-birthDate	birthDate	Date of birth.
UKC-2910	occupation	The principal activity in your life that you do to earn money.
UKC-2	name	A language unit by which a person or thing is known.
schema.org-startDate	startDate	The start date and time of the item.
schema.org-endDate	endDate	The end date and time of the item.
UKC-45004	address	The place where a person or organization can be found or communicated with.
UKC-45537	municipality	An urban district having corporate status and powers of self-government.
OSM-surface	surface	Describes the surface of a feature.
UKC-75466	lit	Provided with artificial light.
UKC-83504	covered	Overlaid or spread or topped with or enclosed within something; sometimes used as a combining form.

Table 3: Data properties concept labels and descriptions.