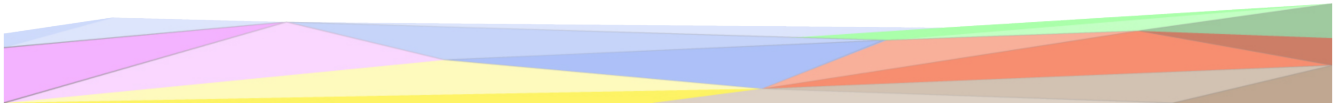
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KDI 2021 - Project Events in Trentino

Document Data:	Reference Persons:
January 3, 2022	Anna Maria Fetz, Lucia Hrovatin



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	20.10.2021	Anna Maria Fetz, Lucia Hrovatin	Domain and Personas definition
0.2	27.10.2021	Anna Maria Fetz, Lucia Hrovatin	Core, common and contextual CQ fragmentation
0.3	09.11.2021	Anna Maria Fetz, Lucia Hrovatin	Inception phase Evaluation
0.4	10.11.2021	Anna Maria Fetz	Data cleaning and parsing
0.5	12.11.2021	Anna Maria Fetz, Lucia Hrovatin	Informal Modeling sheet
0.6	14.11.2021	Lucia Hrovatin	ER model formalization
0.7	16.11.2021	Anna Maria Fetz, Lucia Hrovatin	Alignment between ER and data
0.8	19.11.2021	Anna Maria Fetz, Lucia Hrovatin	Metadata updating
0.9	20.11.2021	Anna Maria Fetz, Lucia Hrovatin	Informal Modeling Evaluation
0.10	25.11.2021	Anna Maria Fetz, Lucia Hrovatin	ER model adjustment
0.11	02.12.2021	Lucia Hrovatin	ETG generation
0.12	06.12.2021	Anna Maria Fetz, Lucia Hrovatin	Formal Modeling Evaluation
0.13	10.12.2021	Anna Maria Fetz, Lucia Hrovatin	Re-alignment data and ETG
0.14	19.12.2021	Anna Maria Fetz	Data Integration on Karmalinker
0.15	21.12.2021	Anna Maria Fetz, Lucia Hrovatin	KGs exploitation on GraphDB
0.16	22.12.2021	Anna Maria Fetz, Lucia Hrovatin	Final presentation and deployment

0 Introduction

Reusability is included in the European FAIR principles [11, 4], aiming at boosting the potential of (digital) data to make it both discoverable and (re)usable either by a human or a machine. When dealing with Data Integration (DI), the concept of reusability becomes even more important and concrete. In this framework, the iTelos methodology can be exploited. Thus, the data integration project documentation plays an important role in enhancing the reusability of the resources handled during the methodology and those produced by the data integration process. A clear description of the resources and the process that manages them provide a deeper understanding of the information involved in the DI project.

Moreover, they allow external readers to exploit the same resources in different projects. The current document provides a detailed report of the DI project developed following the iTelos methodology. The report is structured, on top, to describe:

- Section 1: the project's purpose and the resources involved (both schema and data resources) in the integration process.
- Section 2, 3, 4, 5: the integration process along the iTelos phases.
- Section 6: the open issues
- Section 7 how the result of the integration process (KGs) can be exploited.

1 Purpose and project's resources

This section reports and describes:

- The project's domain and purpose, by reporting the purpose itself and the definition of the project's domain, personas, and scenarios.
- Knowledge resources: The collected reference teleologies that satisfy the purpose along the integration process.
- Data resources: The data collected to satisfy the purpose along the integration process.
- Metadata: The metadata defined for the knowledge and data resources.

1.1 Project Domain and Purpose

The project's scope, meaning its Domain of Interest, intersects the Event domain with the University Student Life within Trentino's area. Due to the placement of University of Trento's departments, the spatial coverage includes only the municipalities of Trento and Rovereto. Temporally, the project targets the period from 2019 to 2022. This wide time-span aims at detecting and modeling the changes occurred due to the Covid-19 pandemic in event organization and participation [7, 8].

The final deliverable should correspond to a website acting as a collective memory of a specific administrative area. The target users are university students, who interact with the service standing as both short-term memory (i.e., working memory) and long-term memory, specifically episodic [1]. While the former function filters the future events by location, duration, target age, personal interests, event category (e.g., music, shows, festivals), and secondary attributes (e.g., language, price), the latter retrieves relevant information about past events. This option focuses mainly on the event's description, tags, ranking, comments, and pictures.

1.2 Scenario and personas descriptions

Nowadays, people are living busy routines cadenced by many events. The Covid-19 pandemic paused people's lives and impacts on the formation of episodic memories [9]. University students are not excluded from this framework, since they are still struggling to find activities fitting their interests and needs. Potential personas and scenarios are listed below.

1. **Giovanni**
   He is 19 years old and is a freshman at the University of Trento. He wants to make friends and get involved in different activities, but with a limited budget due to the flat rent. Since high school, he has been a student representative and has demonstrated an interest in politics and enterprises.
   SCENARIO A: *Giovanni moved to Trento at the beginning of October to start the Bachelor in Economics. Everything is new, and he knows the importance of making friends at the university and, if possible, also outside. One evening, he decides to search for upcoming musical events that gather many people and do not cost too much. This may be a practical solution for the immediate loneliness, but he wants to be involved in more extended activities too. Thus, he wants to find out which student associations are involved in politics and, possibly, also in economics.*

2. **Martina**
   She is 19 years old and law freshman at Trento's University. She looks for events happening during the week to be able to get back home on weekends. Even though she would like to meet many people at once at popular events, she is afraid of the ongoing pandemic.
   SCENARIO B: *Even if freshman, Martina shows excellent social skills. She has already organized an Aperitif with her new colleagues but the location is undefined yet. She plans to find the best bar in the city center by ranking them based on the last two events and the extra benefits offered during those occasions. She also wants to convince her colleagues to go to the Poplar festival, a famous music festival in Trento. Price, online booking, and Covid restrictions are the relevant information she should give them.*

3. **Arianna**
   She is an undergraduate in Biology, is 20 years old, lives in a small town close to Trento,

and commutes every day between her hometown and the university. She does not have a driving license; therefore, her routine is based on the public transport timetable. Most of the time, she uses buses. During the week the last bus is scheduled at 10 p.m., while on Saturdays and Sundays it stops at 9 p.m. She constantly seeks new seminars and meetings about scientific topics, mostly biology and medical engineering.

SCENARIO C: *Arianna finds her courses extremely theoretical and is deeply unsure about her choice. She loves biology but would like to see the practical projects and their results in the medical engineering field. Even though her professors are well referenced, she is interested in something more. Interviews, seminars, and lectures held by (international) guests are her passion. Sometimes she also tries to contact the lecturers or the event organizers to ask further information about the topics.*

4. **Simon**
   They are 24 years old and attend a Master in Economics, after the Bachelor in Management. They have lived in Trento for almost five years and are well settled in the city but still looking for new opportunities. Indeed, they were a member of the CUS, the University Sports Group (i.e., Centro Universitario Sportivo), before Covid, but did not exploit their courses and offers. Now they want to participate in the CUS open events and learn new skills and sports (such as Hokey). They also know that Trento's basketball and volley teams play in the first league (Serie-A).

   SCENARIO D: *Simon is a sports fan. Independently from the sport, they try to watch as many matches as possible. Before the Covid pandemic, they went to some matches of the city's basketball team: L'Aquila Trento. This year they want to discover new (Paralympic) sports, such as Hokey and Volleyball, hoping for affordable prices, starts at reasonable hours and are in places reachable by bus.*

5. **Clara**
   She is 26 years old and attends a Master in Cognitive Science at CIMeC in Rovereto. When she applied for this Master's degree, she decided to live in Trento and commute using regional trains. Luckily, this term she has lecture only until 3 p.m. Therefore, she is constantly looking for art workshops, better if near to Rovereto's train station. When she finds evening courses, she must consider the last train's departure. She is also catholic, and tries to find an active and young community to volunteer and meet new people.

   SCENARIO E: *Clara has a busy routine. Due to the constraints posed by lectures and train timetables, she struggles to find activities parallel to the university. Specifically, she looks for art workshops providing the necessary material, preferably free and not far from the train station. Knowing that the probability of finding such an event in Rovereto is extremely low, she also looks for spiritual associations or groups organizing volunteering activities and meetings.*

6. **Timon**
   Timon is a 23 years old Erasmus student coming from Germany. He knows that Trento's

ESN (i.e., Erasmus Student Network) section is active and organizes different events for international students. As he is interested in meeting other internationals rather than locals, he planned to join as many ESN events as possible.
SCENARIO F: *Most of his colleagues are Italian-only speakers, knowing little English. This makes him feel as an outcast. Therefore, days pass by slowly without having a real chance of integrating and getting to know different cultures better. He decides to turn things upside down and start living the pure Erasmus experience by finding next ESN events, and related activities.*

7. **Paola** She is a Master's student in Sociology, 25 years old, and with an impaired mobility. She does not accept the concept of *limit*. She is deeply interested in cultural events, mainly art exhibitions, and tries to visit them. She is constantly looking for information about the location of those events, the degree of accessibility (i.e., architectural barriers) and the possibility of having reserved parking lots close to the entrance.
   SCENARIO G: *Although having lived in Trento for many years, Paola has never had really the chance of sightseeing. Therefore, one day she spontaneously decides to go visit the city and a new photography gallery which just opened in the center. However, she discovers that to get there steps are the main way in, and nobody seems to know whether there are impaired-suited entrances.After this experience, she decided to map the city events to find out wheter they are accessible, offer specific offers for people with disabilities and can offer reserved parking lots.*

8. **Davide**
   28 years old Ph.D. student in the Department of Engineering. Every weekend, he escapes from the stress caused by the workload of academic duties and goes hiking (or skiing) with the SAT (i.e., Società Alpinisti Trentini), the local trekking group. He is an experienced walker but avoids safeguarded climbs and very binding stretches. He has a higher budget than an average student and can also reach farther places by car.
   SCENARIO H: *Due to organizational issues, Davide's friends decide to cancel the hiking excursion which had been planned for weeks. However, Davide already prepared everything for a nice weekend in the mountains and now wants to check whether any trekking group in the city is organizing similar events in the weekend. By tag, he would like to retrieve hiking experiences that take place during the weekend and require medium skills. If the search will not return results, he would also be open to taking advantage of the free weekend by attending seminars which could boost his career.*

9. **Daniel**
   He is attending the last year of his high school and is choosing where to attend the University. The choice is between Padova and Trento, cities offering top-ranked universities but with different social lives. He is no social butterfly and very picky, therefore he has few friends that he deeply trusts. He prefers culturally challenging events rather than going to the disco, and thus he is most likely going to choose the university offering the most intriguing and interesting events, with companions he might like. Given that

he has a strong passion for chess, he would also like to subscribe to the local chess club to bond with people similar to him.

SCENARIO I: *Daniel has thought out the choice of the university, but he is still skeptical. One last parameter is left out: social life. He focuses on two aspects: musical events and chess. Firstly, he retrieves musical events and festivals based on friends' advice. He wants to judge them by reading comments, looking at pictures of past editions, and evaluating the overall offer. Secondly, he searches for recent achievements and competitions of the university´s chess club. Alternatively, if there is none, similar activities organized by private clubs or student associations.*

10. **Andrea**
    They are 26 years old and attend the third year of (music) conservatory for pop music. Besides music, they are keen on classical dance and theatre shows. They would like to watch as many shows as possible and buy the best seats at the lowest price.

    SCENARIO J: *Besides the pure passion for music, Andrea fell in love with someone and would like to make a great impression at their first date. Therefore, after having carefully scrutinised their Social Media profile, the decision lands over assisting at the concert of their crush' favourite musician. They would like to pay for the best available seats, in the best location possible, according to the price range.*

1.3 Knowledge resources

The backend apparatus of the project relies on both linguistic and knowledge resources. In particular, the preferred linguistic resource is WordNet, one of the most famous lexical repository for English [10]. The chosen reference websites publishing documentation and guidelines to use mark-up on web pages are Schema.org and DBpedia.

1.4 Data resources

The informational core of the project relies on the following listed websites. The **general** purpose pursued here was to to retrieve data shaping the concepts of *time-span, recurrency, student life, appreciation, event categorization*.

- Crushsite. Online culture&arts agenda in Trento and Trentino area. The website categorises events according to the type of event. Each item presents Date, Place/Location, Time and some further Notes or Information. It has been found to be useful especially with regards to start time and location definitions, as well as concept and events fragmentation into categories. Another main feature regards the amount of listed events.
- StayHappening. List of events and activities in the selected city (namely Trento). Each occurrence reports its Title, Start and End Date, Organizer and Description. Moreover, classification into event categories is well organized, but differs from CrushSite's classification.

- ESN Trento. Main resource to retrieve and gather events for international students. One of the main downsides is the bad structure of the posted events: Start and End Date are not always shown and the standard location is set to Italy without being further specified. However, the website is rich of various events, having the main advantage of presenting different editions of the same occurrence over time.
- TripAdvisor. Very useful to get ratings and reviews about a location. Events' reviews are difficult to access, especially the ones regarding students, as they have been posted on Social Media such as Facebook and Instagram, which are not accessible by Non-Organizers. On the other hand, landmarks reviews are of easy access and thus they have been used as main referral for defining a liked or less-liked event by *general* audience.
- MyMovies. As for ESN Association, this website has been used to retrieve further information for international students and to cross-match information from different events, in order to widen the audience as prospected in our purpose.
- JeTn, L'Universitario and UDU events from Eventbrite. They have been retrieved through the student association itself in order to get a closer look to a typical student-organized occasion. This data is useful for different check-ins fragmentation, but the raw format will not be published accordingly to the European Union's data protection laws, in particular to the 26^{th} Recital of the General Data Protection Regulation (GDPR)[5].
- OpenDataTrentino for the events organized in the municipalities of Rovereto and Trento. The datasets are publicly available and (partially) georeferenced. They gather wide-scope and significant events usually promoted and sponsored by the municipalities themselves.
- OpenStreetMap data regarding Trento and Rovereto territory. This data contains tags to mark places' characteristics. Examples of relevant keys are wheelchair, train station and student dorms.

1.5 Metadata

Due to generic and variegate collection of data and the lack of parsing for some scraped data, metadata has been collected within a unique `.ttl` file.
For clarity purposes, all data publishers not directly connected to a physical or legally recognized organization have been linked to an *Agent* node rather than an *Organization* node. Moreover, all restricted data gathered through Event organizers posting on Eventbrite had been structured in the following way. As interpretation may be very wide, Eventbrite has been considered as the main organizer, while student associations as *Agents* of the same Organizer. All data sources have been linked to a specific *Location*, Publisher, and *ContactPoint*. On the other hand, version notes, dataset description and format may vary according to the source.
Other kind of resources, namely the `.json` files extracted via web scraping and API calls present a similar structure, namely:

- **Name** (str): the name of the file built in the following way: category-year-name
- **Location** (str): postal address or *online*

- **Lat_lon** (float): location's geocoordinates in reference system WGS84
- **Description** (str): event description containing relevant information, such as ticket prices, contacts, office hours, event content
- **Info** (str): additional event information whose content may vary according to the website
- **Links** (str): URL, reporting all link references encountered during parsing
- **Schema** (str): all links referring to schema.org and defining the `<div>` object class
- **Recurrency** (str): it identifies events that repeat over time (i.e., weekly, monthly, or daily)

Additionally, data gathered from Open Data Trentino are structured differently. Whereas the variables in Rovereto's dataset were available with an exhaustive description (Open Data Rovereto), Trento's data does not offer further specifications. On the one hand, to exploit the information embedded in the former dataset, the keys `weekday`, `address`, `duration_hour`, `duration_days` and `repeated` (a Boolean variable returning whether an event is repeated over time or not) have been introduced. On the other hand, Trento's data displays each item (i.e., event) as a dictionary having as main keys `metadata` and `data` that are then further sub-specified. In particular, the `data` key presents a (language - dependant) description per event via several attributes.

The spatial reference of the project relies on data extracted from OpenStreetMap and, specifically, from the keys `tourism`, `leisure`, `building` and `amenity`. The verbose output shows a common structure, characterised by:

- Node's attributes
- API references such as changeset
- key:addr used to provide address information for buildings and facilities
- key-dependant attributes

Lastly, remaining data sources (i.e., Meetup) have not been listed because neither additional parsing nor further changes were applied. Their variables' basic description is reported in the metadata file.

2 Inception

This section reports the sub-activities performed during the Inception phase and describes them both in schema and data layer.
Inception sub-activities:

- Purpose formalization (inception part) and Inception sheet description
- Data and knowledge resource collection

- Resource classification
- Inception phase evaluation

The Inception phase report includes an explanation of the different choices along with their strength and weaknesses.

2.1 Purpose Formalization and Inception Sheet

”A service which helps the citizens to find events of interest in Trentino.”

Following the wide-scope research question stated above, the project’s purpose constraints it by focusing on the University Life domain. In particular, the final service should respond to specific event requests performed by a (future) student of the University of Trento, within the age range 18-30.
The concept of *Event* itself has been retrieved by Wordnet, as

something that happens at a given place and time.

Therefore, spatial and temporal information plays a fundamental role. The supported requests and sources will refer to a specific period: 2019 - 2022. This time-span has been chosen in order to describe in the best way possible different frameworks, namely the activities before, during and (almost) after the Covid-19 pandemic. The geographical coverage includes Trento’s and Rovereto’s surroundings and circumscriptions.
Queries from the personas’ side can vary according to personal interests (such as music, sports, or politics), field of study, hobbies, and impairments. Further supported searches filter events by occurrence and event organizer (i.e., associations or/and stakeholders).
The features presented above refer to short-term memory activities. However, the project guarantees an episodic memory function for querying past events via tags and name and retrieving descriptions, comments, and photos.
All privately-organized events are out-of-scope and therefore not included.

After the purpose formalisation, at least five Competency Questions for persona and scenario have been identified. Their natural language formulations and the extrapolated and categorised concepts (i.e., common, core, and contextual) are reported in the table below.

Persona	CQ	Common K.C.	Core K.C.	Contextual K.C.
1. Giovanni	1.1 Return all university student associations in Trento that organize events.	Administrative Area, Audience, Event, Location, Organisation, Person, Student	JoinAction	Student Organization, University
	1.2 Given the name of a student association, list all the titles and type of event it will organize in the next month.	Duration, Event, Organization, Person, Student	Student Organization	Time-span (month), Title, TypeOfEvent
	1.3 List yearly musical Festivals in Trento gathering more than 2000 people.	Administrative Area, Audience, Event, Facility, Location, Number, Person	Festival, ListenAction, Schedule	CreativeWork, MusicGenre, Time-span
	1.4 Find concerts in Trento sorted by ticket price.	Administrative Area, Audience, Event, Number, Person	MusicEvent, Ticket	Concert, Price Specification
	1.5 Given the name of the Artist, retrieve all their future and past concerts in Trento.	Administrative Area, Duration, Event, Location, Person	Artist(s), MusicEvent, playAction, Schedule, singAction	Concert, isFinished, Name, Time-span
2. Martina	2.1 Rank Trento city center's bars based on the reputation of the last 2 events.	Administrative Area, Audience, Facility, Food Facility, Duration, Event, Location, Number, Person	Bar, City, SocialEvent	isFinished, Ranking
	2.2 Given that there exists an event in a specific location (e.g., Doss of Trento) and on a specific day, return the ticket price and if it can be booked online.	Administrative Area, Audience, Civic Structure, Duration, Facility, Event, LocalBusiness, Location, Person	Address, payAction, Ticket	Landmark, Online booking, Price Specification

	2.3 Get all the events targeting university students and weekly offered in the bars of the city center.	Administrative Area, Duration, Event, Facility, FoodEstablish-ment, LocalBusiness, Location, Person, Student	Bar, City, Schedule, SocialEvent	City center, Time-span, University Student
	2.4 For closed and crowded places, get all the regulations and/or requirements related to the Covid pandemic.	Administrative Area, Duration, Event, Facility, LocalBusiness, Location, Medical Condition, Person	Special An-nouncements, closed Place	Health issues
	2.5 For an event organized in bars, return all the extra benefits offered to the audience (i.e., free drink).	Audience, Duration, Event, Facility, Food-Establishment, Location, Person	Bar, DanceEvent, drinkAction, payAction, SocialEvent, Ticket	listOfBenefit, Price Specification
3. Arianna	3.1 When and where will be the next open seminar about biology organized during the week and finishing before 22 p.m.?	Audience, Duration, Event, Facility, Location, Person	Address, Edu-cationEvent, endTime, Topic/Category	Lecturer, Price Specification, Science, Seminar
	3.2 List the topics of all the meetings/interviews included in a scientific Festival and organized in blended mode.	Audience, Duration, Event, Facility, Location, Person, URL	EducationEvent, Festival, Topic/Category	BlendedMode, Forum, Science
	3.3 List all the online seminars sponsored by the Department of Engineering.	Duration, Event, Organization, Person, URL	Department, Education-Event, VirtualLocation	Faculty, Online, Seminar, University

	3.4 Find the address of the next seminar having as lecturer the professor X, expert in Y, and held in English.	Administrative Area, Civic Structure, Duration, Event, Facility, Location, Person	Address, EducationEvent, Employee	Language, Name, Professor, Research Field, Seminar, Time-span
	3.5 Retrieve the email contact of the organizer of a past seminar.	Duration, Event, Location, Organization, Person	Contact, EducationEvent, findAction, Organiser	Email, Seminar, Time-span
4. Simon	4.1 Get the starting time of the basketball match taking place in the evening.	Civic Structure, Duration, Event, Facility, Location, Person	playAction, Schedule, SportsEvent, startAction, Team, watchAction	Basketball, startDate, startTime
	4.2 Get, if any, discount options for university students for watching basketball matches.	Civic Structure, Duration, Event, Facility, Location, Person, Student	SportsEvent, Team, Ticket, watchAction	Basketball, Match, Price Specification, Seat
	4.3 Get if the Uni Sports promotes events of a uncommon discipline.	Event, Facility, Organization, Person, Student	exerciseAction, SportsEvent	Ranking, Sport type, University
	4.4 Get the location of the main Trento's Football Stadium.	Administrative Area, Civic Structure, Facility, Location, Organization	Address, Latitude, Longitude, SportsFacility	Football Stadium, Ranking
	4.5 List all the Hokey matches in Trento.	Administrative Area, Duration, Event, Organization, Person	playAction, SportsEvent, Team	Hokey, listOf, Match, Ranking
	4.6 Are there Paralympic events in Trento's circumscription?	Administrative Area, Audience, Event, Location, Medical Condition, Organization, Person	playAction, SportsEvent, Team	Circumsciption, Paralympic discipline

5. Clara	5.1 Find the description of the activities performed during an open artistic workshop.	Duration, Event, Location, Organization, Person	typeOfActivity, CreativeWork, createAction, Workshop	Artwork, listOf, Price Specification
	5.2 List all the social events near (within 2km) Rovereto's train station that end before the last train to Trento.	Administrative Area, Civic Structure, Duration, Event, Facility, LocalBusiness, Location, Person, Transport	moveAction, SocialEvent, Station	isDistant, Schedule, Time-span, Train, TrainStation
	5.3 Given an artistic workshop organized in Rovereto, list all the materials she should bring.	Administrative Area, Duration, Event, Location, Organization, Person	CreativeWork, organizeAction, Workshop	Artwork, listOf
	5.4 List all the events promoted by spiritual associations in Rovereto and Trento.	Administrative Area, Civic Structure, Duration, Event, Location, Organization, Person	joinAction, placeOfWor-ship, SocialEvent	Church, listOf, NoProfit
	5.5 Retrieve the information about Tutors' names of an artistic workshop.	Administrative Area, Duration, Event, Location, Organization, Person, Product	Contact, CreativeWork, organizeAction, Workshop	Artwork, listOf, Name, Tutor
6. Timon	6.1 List all the cinemas offering movies in English with subtitles.	Audience, Civic Structure, Duration, Event, Facility, LocalBusiness, Location, Person	Cinema, CreativeWork, CulturalEvents, Entertainment-Facility, Language, watchActivity	listOf, Movie, Subtitles

	6.2 Get places and days of Student Association's (ESN) movie nights.	Audience, Duration, Event, Facility, Location, Organization, Person, Student	Address, CreativeWork, joinAction, SocialEvent, startDate, watchActivity	Cineforum, listOf, Movie, Student Organization, UniversityStudentType
	6.3 Are there cross-cultural challenges (e.g. languages Cafè/game nights) organized by ESN?	Audience, Duration, Event, Facility, Location, Organization, Person, Student	joinAction, playAction, SocialEvent	Challenge, Language, Student Organization, UniversityStudentType
	6.4 Get the *All you can drink* bars close (within 2 km) to the student dorms of Opera Universitaria.	Accommodation, Administrative Area, Event, Facility, FoodEstablishment, LocalBusiness, Location, Organization, Person, Student	Bar, consumeAction, Hostel, studentDorm	drinkAction, isDistant, University
	6.5 List all the languages offered during guided tours in the Buonconsiglio Castle.	Administrative Area, Audience, Civic Structure, Duration, Event, Facility, Location, Person, Tourist Attraction	CulturalEvent, Guided Tour, Landmark, Tourist, Language	listOf, Name
7. Paola	7.1 Return if an assistant has to pay an extra ticket, when visiting the event X.	Administrative Area, Civic Structure, Event, Facility, Location, Medical Condition, Person, Touristic Attraction	Assistant, CulturalEvent, Landmark, Ticket	Exhibition, Museum, Price Specification, Title
	7.2 Get the level of accessibility (architectural barriers) of the well-known landmarks organizing events.	Administrative Area, Event, Facility, Location, Medical Condition, Person, Touristic Attraction	Architectural Barriers, Landmark	Accessibility Level, Ranking

	7.3 Get if special support is offered to guarantee a good experience to the visitor.	Administrative Area, Civic Structure, Event, Facility, Location, Medical Condition, Person, Touristic Attraction, Transport	Employee, Guided Tour	Ranking, Review
	7.4 Are there reserved parking lots near the event's venue?	Administrative Area, Civic Structure, Event, Facility, Location, Person, Medical Condition, Touristic Attraction, Transport	Car, Landmars, Parking	isDistant, isReservedParking
8. Davide	8.1 Given Trentino's territory, are there any après-ski?	Accommodation, Administrative Area, Duration, Event, Facility, FoodEstablishment, LocalBusiness, Location, Person	drinkAction, Landscape, Resort, SocialEvent	Happy hour, Malga, Mountain
	8.2 List the traditional and folkloric food events in Trento's circumscription where participants can cook too.	Administrative Area, Duration, Event, Facility, Food Establishment, LocalBusiness, Location, Person	cookAction, CreativeWork, Workshop	(typical) Food, Recipe
	8.3 Return the description of a path of an excursion organized by the trekking group of the city (e.g., SAT).	Administrative Area, Duration, Event, Location, Organization, Person	City, Landscape, travelAction, Tour	Excursion, Map, Mountain, trekking Organization, tripItinerary

	8.4 Given that a trekking excursion is organized, list all the extra information.	Administrative Area, Audience, Duration, Event, Location, Organization, Person	Landscape, travelAction, Tour	listOf, Mountain, required Skills, targetAge, trekkingOrganization
	8.5 Get the name of the buildings of all open debates organized by University Professors in Trento.	Administrative Area, Civic Structure, Duration, Event, Facility, Location, Organization, Person	Address, Auditorium, EducationEvent, Employee, listenAction	Conference, isDistant, Price Specification, University professor
	8.6 Get the target age of an excursion.	Audience, Duration, Event, Location, Organization, Person	Landscape, travelAction, Tour	Excursion, Mountain, trekkingOrganization, targetAge
	8.7 Returns the specific coordinates of the starting point of an excursion.	Duration, Event, Location, Organization, Person	Altitude, Landscape, Latitude, Longitude, Tour	Excursion, Mountain, trekkingExcursion, trekkingOrganization, tripDestination
9. Daniel	9.1 Knowing that a friend participated in the X music event, retrieve it via tag.	Administrative Area, Duration, Event, Facility, Location, Person	CulturalEVent, listenAction, watchAction	Concert, Interaction Counter, isFinished, musicEvent, Post, Tag
	9.2 Are there clubs to which students can subscribe based on their interests? e.g. book clubs, poetry clubs, chess clubs	Facility, Location, Organization, Person, Student	Club, subscribeAction	culturalClub, Hobby
	9.3 Knowing the name of a Festival, retrieve pictures of past editions.	Administrative Area, Audience, Duration, Event, Facility, Location, Person	Festival, Image, Schedule, watchAction	Edition, isFinished, Time-span

	9.4 Knowing the title of an event, retrieve its evaluation and reviews.	Administrative Area, Duration, Event, Facility Location, Person	Schedule	Evaluation, Name, Review
	9.5 List all the Event contributors (stakeholders) given the event title.	Administrative Area, Duration, Event, Facility, Location, Person	Stakeholder(s)	ListOf, Name, Title
	9.6 List the free entry monuments due to special occasions (e.g., F.A.I. days) in Trento and Rovereto area.	Administrative Area, Civic Structure, Event, Facility, Location, Organization, Person, Touristic Attraction	Address, endDate, Landmark, Schedule, startDate	isFree, Monument, Name, NoProfit, Price Specification
10. Andrea	10.1 List all the musicals offered by a theater in Trento.	Administrative Area, Audience, Civic Structure, CreativeWork, Duration, Event, Facility, Location, Person	CreativeWork, culturalEvent, Show, Theater, watchAction	ListOf, Musical, Title, TheaterEvent
	10.2 Given that a drama show exists in at least one theater, are there student discounts?	Audience, Civic Structure, Duration, Facility, Location, Person, Student	CreativeWork, culturalEvent, PayAction, Show, Theater, Ticket	Drama, Price Specification, TheaterEvent
	10.3 Given that a drama show exists in at least one theater, return all the ticket prices depending on the seat position.	Audience, Civic Structure, Event, Duration, Facility, Person, Location, Student	CreativeWork, CulturalEvent, PayAction, Show, Theater, Ticket	Drama, Seat, TheaterEvent
	10.4 Given a specific band, list all the concerts offered within Trento's municipality and their locations.	Administrative Area, Audience, Duration, Event, Facility, Location, Person, Student	Address, Artist, culturalEvent, playAction	Band, Concert, Municipality, musicEvent
	10.5 Return the link of live stream musical performances.	Audience, Duration, Event, Location, Person, URL	CreativeWork, culturalEvent, VirtualLocation	BeingLive, musicEvent

	10.6 List the addresses and contacts of theaters present in Trento.	Administrative Area, Civic Structure, Facility, Event, Location, Organization, Person, URL	Address, ContactPoint, Theater	Email, ListOf, Telephone
	10.7 Return if it is possible to book a specific seat for the future X theater show.	Civic Structure, Duration, Event, Facility, Location, Number, Person	CreativeWork, culturalEvent, Theater, Ticket	Name, reserveAction, Seat, SeatNumber, TheaterEvent, Time-span

2.2 Data and Knowledge resource collection

2.2.1 Data resource collection

For data collection, a twofold approach was adopted, according to data availability and regulations. Some information had to be retrieved through direct scraping of the web page it was found in, while basic REST API calls were used in other cases. Inevitably, collected data was heterogeneous and data cleansing had to be performed before data deployment.

1. Information retrieved through API calls was returned in `.JSON` format. However, keys differed consistently among files, especially concerning the event description and location. Namely, the *location* concept embeds different interpretations and the importance of retrieving address, latitude and longitude coordinates was crucial. Similarly, parsing both `description` and `info` keys was essential, due to different understandings and usages of the *additional information* concept.
2. Data gathered through Web Scraping was severely spurious. Thus, the collection followed four different steps:
   (a) Firstly, a request was sent to establish a connection to create a Soup object with Beautiful Soup library.
   (b) The Web Site had to be visited looking for the `<div>` classes of interest, namely, events information.
   (c) Only `<div>` containers belonging to the class of interest were scraped and saved to a local `.CSV` file.
   (d) All `.CSV`-saved `.html`-like files were partially parsed by removing `.html` containers and by creating a basic `.JSON` file with duration in hours, date, event's title, description and further links.

- In Crushsite.it and StayHappening.com, each event was already partially organized within the main `<body><div>` block. Therefore, the scraping extracted the main column from the

web page and scanned for keywords. In particular, data was organized into **arbitrary macro** keys and saved into a `.JSON` file. This choice aimed at aligning different data sources to the same data type.

- ESN Trento website is differently organized. Firstly, there was a lack of a proper event declaration (i.e., no exact address or start/end date). This absence may be justified by the presence of different editions of the same event. Secondly, the scraping has been less straightforward than the previously-mentioned sites due to changing URLs depending on the page number. Thus, programmatic access was almost impossible, and the code was manually modified for each event page and relative scraping procedure. Moreover, a general parsing to remove `.html` code was applied, but most of the information contained in the event description has been kept for further parsing, according to keywords such as *Edition*, *MonthName* and *DateTime*. Data is stored in a `.JSON` file having the same keys as those applied for CrushSite and StayHappening.
- Trip Advisor has been employed as both the project purpose and the competency questions underlined the necessity of reviewing the places where the events occur. The most meaningful choice would be to scrape comments on social networks or, in this case, on similar services. This approach involved three Landmarks of Trento and Rovereto and an additional student-suited location: the Buonconsiglio Castle, MART, Birreria Pedavena, and Bookique. The scraping was applied only on the most popular and shorter comments due to the difficulty of accessing longer reviews via secondary links. The parsed outcome, saved in a `.CSV` file, will be further analysed by applying sentiment analysis.
- MyMovies has been partially scraped in order to retrieve movies screened in original language (with subtitles). As for CrushSite and StayHappening, the web page was adequately divided by the `<div>` container and the respective classes. The downloaded file was firstly saved into a `.CSV` and then formatted into `.JSON` to easily match by key `name` the already-downloaded and parsed movie events.
- The data gathered from events published on Eventbrite by some student associations aimed at detecting student-suited occasions and checking whether the number of followers influenced the number of participants per event. This data has been directly requested to the event organizers (`xlsx` format) and contained personal data. Therefore it was anonymized before upload [5] and further parsed by choosing keys similar to the other `.JSON` files.
- OPENDataRovereto and OPENDataTrento. Data was retrieved through REST API calls getting `.JSON` files with georeferences (i.e., latitude and longitude in WGS84 reference system). Even if metadata is provided along with informative keys, parsing the event description was necessary to extract information about the time, date, and address (mostly in Rovereto's dataset). Again, separate files have been locally stored and aligned to CrushSite and StayHappening data structure for future integration.
- OpenStreetMap data can be used in many ways. Even though the limits of an open community are well known (e.g., the instability and volatility of data), the degree of data precision and richness is extremely high thanks to local contributors. In particular, a place can be

described by different tags encoding its features. Examples are the information about building accessibility for people with impairments (`key:wheelchair`), the presence of students' dorms (`tag:building=dormitory`) or train stations (`tag:railway=station`).
Data has been requested via URL to the service made by Wikimedia Italia, saved as Protocol buffer Binary Format and read with the python library Pyrosm. Data will be saved locally in `geojson` format.

- MeetUP data has been retrieved through the provided REST API that returns data in `.JSON` format. However, only one event has been considered due to the lack of variety and absence of proper events organized in the territory of interest. The motivation behind this choice can be justified by the interest reserved to the structure and nomenclature of the `.JSON` keys.

Some hurdles were faced during the data collection process, ranging from the lack of quantity and specificity to difficulties in accessing websites, exploiting API resources, and generalising parsing and scraping steps. The first problem we dealt with was regarding data access. Thus, it had to be reshaped entirely due to Social Network's Privacy Terms and Eventbrites' REST API restrictions. An attempt was made with Twitter hashtags too (i.e., `#visittrentino`), but the a-specificity and hollowness of tweets, not directly referring to an event, led to meaningless results. Secondly, different data sources had to be exploited in order to get both *producers'* and *consumers'* perspectives and roles within an event. Thus, in addition to Web-Agendas for Events, a connection with TripAdvisor Reviews has been established.
Lastly, another main issue faced during the Inception phase regards the data format. Namely, collected data was often returned as `.CSV` files. Further parsing and reshaping steps have enriched data and allowed an understanding of the available information. The last note regards StayHappening and the scraping procedure that needed different access settings to avoid being signaled as bot.

Besides the `.html` code residuals and some not-standardised data (i.e., time, date, and georeferences), the collection and parsing phase, with cleansing and normalisation procedures included, results satisfactory but still not sufficient for the project's scope. Further work will be applied to obtain cleansed data, fragmented according to the correct reference keys (i.e., Location, Duration, Time), and to compensate for chaotic descriptions, different website structures, and missing or misplaced information.

2.2.2 Knowledge resource collection

The knowledge and linguistic resources exploit different but parallel roles. While the lexical repository WordNet has been used to identify and uniquely define the keywords included in the project's assignment (e.g., *event*), the vocabularies schema.org and DBpedia covered entities, their relationships and actions.

2.3 Resource classification

As already applied to the Competency Questions, data collection resources can also be assigned to different levels.

Most of the data retrieved via web scraping, namely from CrushSite, StayHAppening, and ESN, belongs to the core resources. The main reason behind this statement lies in how they categorise events, which eases the common and core components detection. Thus, a variety of actions, occurrences, event types, and locations were detectable. Additionally, some event-specific details were retrieved, such as as entrance fees and contacts, which turned out useful jointly with data retrieved via Open Trentino as context resources. On the other hand, data retrieved from OpenStreetMap can be assigned to two different levels: it belongs to the common resources category when employed as a geographical reference, but it becomes contextual if the potential of tags is exploited as well. Even if the tags do not follow a standard, vary, and are unstable over time, they greatly detect and describe places' attributes and peculiarities (e.g., place usages, building characteristics).

The knowledge resources cannot be sharply divided into categories. The project's baseline equally relies on schema.org and DBpedia. However, for specific categories within the impairment field, we considered ada.gov as well.

2.4 Inception phase Evaluation

2.4.1 Schema Level

Considering the schema level of the inception phase, the *Coverage* shown by the formula 1 and the *Extensiveness*, reported in formula 2, will be computed between the Competency Queries (CQ) and the aligned ontology (Ont).

$$Cov(CQ) = \frac{|CQ \cap Ont|}{|CQ|} \tag{1}$$

$$Ext(CQ) = \frac{|Ont - CQ|}{|Ont \cup CQ|} \tag{2}$$

At the schema level, the inception phase presents ideal results. Thus, both measures settle down to values around $0.5 - 0.7$, as the table below shows.

2.4.2 Data Level

At the data level, the Coverage measure still offers an overview of the goodness of the Inception phase, but it is now paired with the *Sparsity* (3). It is also worth noticing as the CQ will be compared with the Datasets (D).

$$Spr(CQ) = 1 - \frac{|CQ \cap D|}{|CQ \cup D|} \tag{3}$$

Schema evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 130 \rvert}{\lvert 46 \rvert} = \frac{\lvert 36 \rvert}{\lvert 46 \rvert} =$ **0.78**	$\frac{\lvert 121 \cap 212 \rvert}{\lvert 121 \rvert} = \frac{\lvert 87 \rvert}{\lvert 121 \rvert} =$ **0.71**
Extensiveness	$\frac{\lvert 130 - 46 \rvert}{\lvert 130 \cup 46 \rvert} = \frac{\lvert 94 \rvert}{\lvert 140 \rvert} =$ **0.67**	$\frac{\lvert 212 - 121 \rvert}{\lvert 212 \cup 121 \rvert} = \frac{\lvert 127 \rvert}{\lvert 248 \rvert} =$ **0.51**

Table 2: Summary of inception's schema evaluation

Data evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 17 \rvert}{\lvert 46 \rvert} = \frac{\lvert 15 \rvert}{\lvert 46 \rvert} = 0.326 \approx$ **0.33**	$\frac{\lvert 121 \cap 98 \rvert}{\lvert 121 \rvert} = \frac{\lvert 90 \rvert}{\lvert 121 \rvert} =$ **0.74**
Sparsity	$1 - \frac{\lvert 46 \cap 17 \rvert}{\lvert 46 \cup 17 \rvert} = 1 - \frac{\lvert 15 \rvert}{\lvert 48 \rvert} = 0.687 \approx$ **0.69**	$1 - \frac{\lvert 121 \cap 98 \rvert}{\lvert 121 \cup 98 \rvert} = 1 - \frac{\lvert 90 \rvert}{\lvert 129 \rvert} =$ **0.30**

Table 3: Summary of inception's data evaluation

Differently from the schema level, the data level displays a certain degree of misalignment concerning the CQs and, specifically, the Etypes. However, this impairment was expected due to the different nature of the raw data acquired, namely web scraping and tabular data (i.e., OpenData Trentino). They will be further parsed and aligned in the following phases, but as the high value of Etype Sparsity shows ($Spr(CQ_e) = 0.69$), the amount and variety of data seem to satisfy and be consistent with the main scope of the project. An opposite scenario is pointed out when observing the Properties. Whereas the Datasets greatly covers the CQs, the low Sparsity highlights the necessity of extracting more information via parsing.

3 Informal Modeling

This section describes the informal modeling phase. Like in the previous section, the current one aims to describe the different sub activities performed, as well as the phase outcomes. More in details, this section provides a description of the following activities:

- Purpose formalization (informal modeling part) and Modeling sheet description
- ER model description

- Informal Modeling evaluation

Like the previous phase, the decisions made, along with their weak and strong points, will be reported. Any difficulty and/or open issue are also highlighted below.

3.1 Purpose formalization and Modeling sheet description

3.1.1 Purpose formalization

The purpose presented in the Inception phase (Section 2) corresponds to a website designed for enrolled and future students (of age 18-30) at the University of Trento. The queries are based on specific *filters* that select and return a set of events. In light of the positive evaluation and the data availability, the project's scope has been pondered and no further changes were applied.

3.1.2 Modeling Sheet description

The building process of the Informal Modeling Sheet was intended to resemble a *real query* on the final service. Thus, this hierarchy has been followed:

- **Common objects** were mainly considered to be existence-independent as they are uniquely identified by their attributes, and do not depend on other entity. Thanks to their generic representation and vast usage, they could be used for a (future) broader DoI.
- **Core objects** represent purpose-specific entities, such as the event `type`, main actors (e.g., Student) or specific facilities. According to the intended output, core objects mirror the infrastructure of the filtering system of the website. Specifically, these entities should direct the initial query and skim all the events not matching preferences. For instance, if a specific place is selected, the output will list only events taking place there.
- **Contextual objects** were thought of being a further specification provided by the user to obtain a narrower list of events matching all criteria. For instance, the user would use location-specific keywords when searching for a specific set of events (e.g., Concerts in Trento's Doss).

With concerns to actions, they mostly were intentionally not defined as entities. Thus, they have been integrated as either properties or bidirectional relationships between main entities (i.e., *superclasses*). This choice is justified by focusing more on the event *types* and, consequently, rejecting of a dense ER model. On the one hand, this leads to a model that fully matches the project's purpose. On the other hand, it implies a more complex matching between the ER model and an already-existing Ontology. Functions, representing a role or a static nature of an Etype, were mostly integrated as properties. However, in modeling certain events, this representation does not hold and the creation of sub-entities is preferred.

In the Informal Modeling sheet, the properties are presented as follows. On the one hand, *Data* properties are displayed conceptually, (e.g., `title`), rather than with the corresponding data-type (e.g. `title`: *string*). On the other hand, *Object* properties are represented as `concept(domain,range)`, aiming to describe the concept underlying the considered property (e.g.,

holder(Event, Person)). Moreover, the understanding of the relationship being mono-directional or bi-directional depends on whether two transposed *(Domain, Range)* tuples are present (i.e., if two tuples with the same elements are present, but with switched positions).

3.1.3 Difficulties and open issues

During the filling process some hurdles were faced:

- The definition of some common entity types represented a significant difficulty. While the creation of many common objects, such as Facility or Location has been straightforward, other cases gave rise to problems. Specifically, some core objects, such as Accommodation and FoodEstablishment, were transformed into common objects but then placed under the Facilitymacro entity. Even though this choice might be debatable, it is justified by the specificity of the project's purpose and the objects' frequency in the CQs.
- Secondly, the definition of sub-entities falling under Person ended up in settling for Student and Employee, while event-specific entries such as Artist, or University Student Type are kept as actions.
- Another issue faces the definition of appreciation and rating, which can be assessed via quantitative or qualitative scales. One suitable qualitative approach is the sentiment analysis of review's content. Due to the specificity of this resource, a Review entity has been established and linked to the more general Ranking.
- A constant difficulty was evident when dealing with the type of event and its specification either as a sub-event category (for instance, Concerts in Music Events) or as a property (i.e., isConcert: Bool). Therefore, sub-entities have been created only when the sub-event type had different and more specific features from its parent entity.
- As data introduces information in English and Italian, some terms were avoided due to a semantic misalignment among languages. German was also taken into account, considering the target area of the service (Trento and Rovereto). A concrete example comes from the term *Trip* which has been changed into *Tour* thanks to its more homogeneous translation: Escursione, Tour, Ausflug.
- The duality of Duration for smaller events falling within the main one, such as a Festival, arose. The final decision considers each event falling within a major one as a single unit. Accordingly, also the concept of Schedule had been fixed. Indeed, the original solution confused Schedule, reflecting the pace at which an event is repeated, with Calendar, the actual dates of collateral events or next appointments.
- Lastly, the concepts of Location and VirtualLocation introduced some consistent interference, especially when defining blended mode events. The choice of defining three separate entities (OfflineEvent, OnlineEvent, BlendedEvent) might be debatable, but it mirrors our purpose and maintains the definition of *location* as a point within the space.

3.2 ER model description

3.2.1 Structural description

The designed Entity-Relationship (*ER*) model, shown below 1, exploits the Crow's foot notation by displaying entities as boxes and relationships as lines between them [6]. The relationship's cardinality is defined by the shape at the end of the lines: a *ring* for zero, a *dash* for one and *crow's foot* for many.
Moreover, the model structures the entities following a bottom-up approach, called generalization. If a set of entities has some common properties, they are extracted and generalized to a higher-level entity (i.e., `Student` and `Employee` are generalized to `Person`). A sub-class inherits all the attributes and relationships from the superclass(es) and, additionally, defines its specific attributes. A non filled arrow is placed with the head touching the superclass object, while all sub-classes are linked to this arrow by a black line.

3.2.2 Conceptual description

The ER model has been structured following the future website's filters and its focus on searching for events, as other websites already do (Figure 2 and 3). The main event's type (i.e., *categories*) identified are the following:

- **Social Events**: lead to a socialization process (i.e., people forming groups) and present repeated actions, such as drinking or dancing. A prototypical example is a Happy Hour.
- **Education Events**: (academic) prearranged meetings for consultation, exchange of information, or discussion. The audience might be not actively involved.
- **Workshop**: umbrella term for any laboratory event that actively involves the participants and delivers a tangible result (i.e., product), or breaks new skills.
- **Cultural Events**: knowing that *cultural* has a broad meaning, within this project, an event is defined as cultural if it is designed for entertainment and enjoyment and is related to some branch of art, culture, or local values (i.e., performing arts, musicals, photography, and literature). The definition here adopted partially resembles the DutchCulture Database structure as it is consistent with the project's purpose.
- **Tour**: traveling means going from one place to another. Whereas the event's types above may be represented as happening at a specific time and in a static space, a tour event involves both a movement and some personal interests. For example, if a hiking tour is considered, two decoupled levels can be identified. Firstly, the static one, meaning the event description, follows the standard place-time definition. Then, the movement and its volatility can be introduced. Thus, creating a specific entity in the ER model combines the two levels and tries to solve a corner case.
- **Sports Events**: organized occasions where a sports or exercise activity is performed at a specific location in a temporal interval. Most sports events are also part of bigger meetings and competitions. They can be periodically organised, such as the Facoltiadi or any

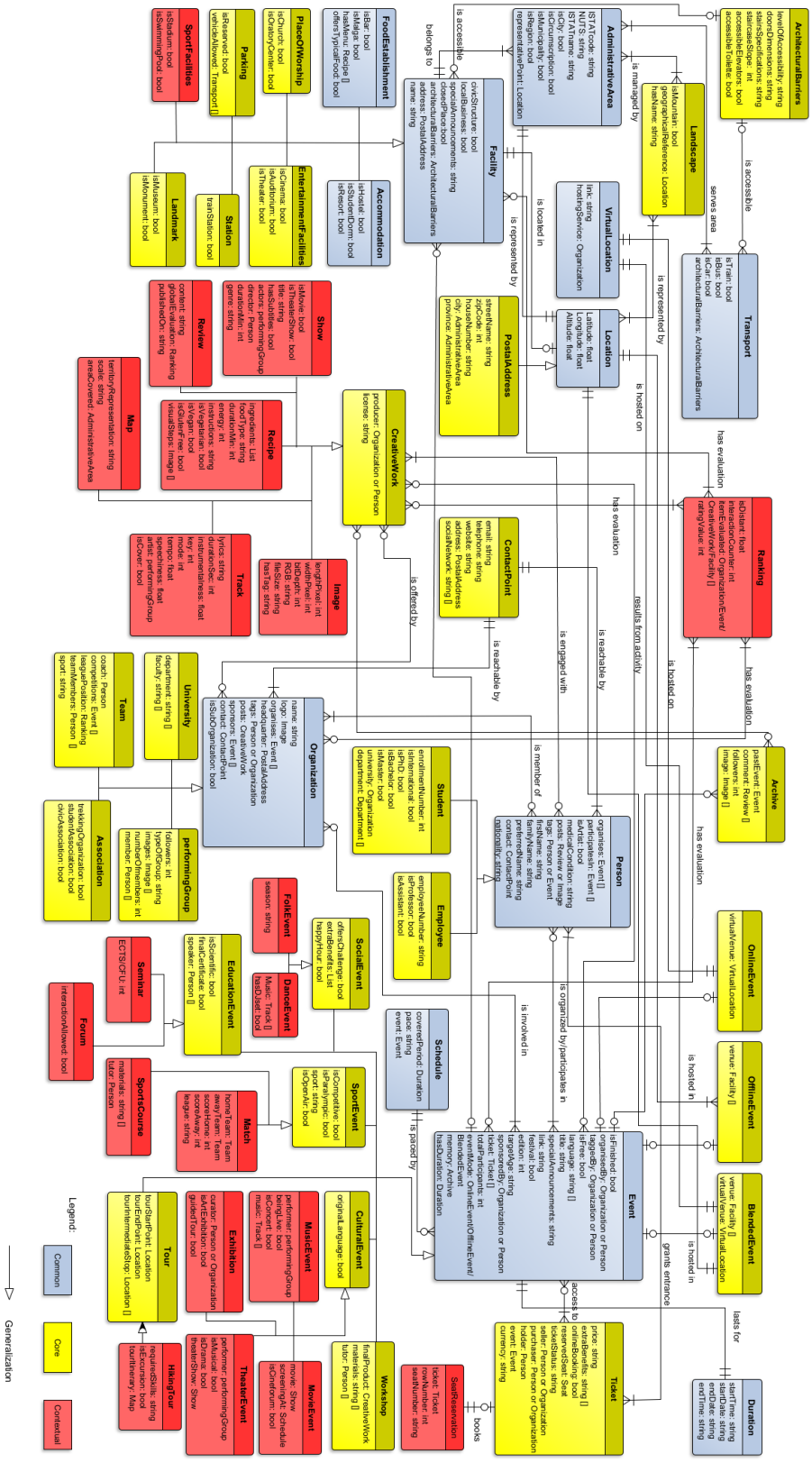


Figure 1: Entity Relationship model

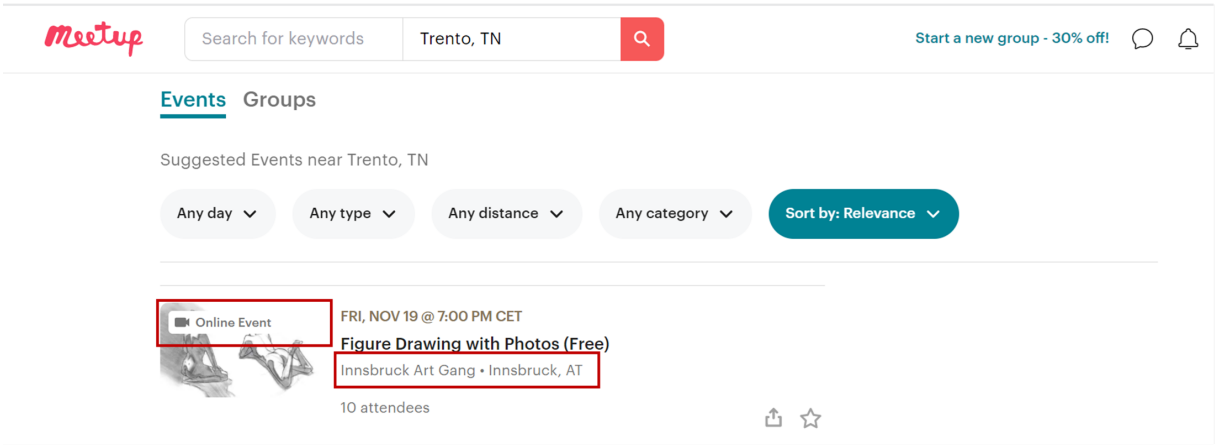


Figure 2: Example of online event promoted on Meetup

sports course (i.e., dance course). Depending on the disciplines, schedule, competitors, and scope (e.g., tournaments, leagues, fund-raising), the events may embed different descriptors. However, accordingly to the project's purpose only the general ones have been chosen. Specifically, they resemble the SportsCompetition model developed by the W3C Community Group.

While the Covid-19 pandemic did not particularly impact the events' categories, it constraints and modifies the concept of *location*. Thus, three scenarios are possible:

- *Offline events* have a location traditionally defined as an identifiable point in the two (or three) dimensional space
- *Online events*, meaning occasions happening only in a virtual environment hosted on video communication services, such as Zoom and Google Meet.
- *Blended or Hybrid events* offer an offline or online attendance depending on the participant's needs and possibilities.

Differences among websites can be detected when modeling this trichotomy. Whereas Meetup always spatially defines the events and, consequently, their nature as *online* (figure 1), websites as Eventbrite simply categorize the event as *online* and do not further specify spatial coordinates (figure 3). The latter implementation has been preferred due to its consistency. Indeed, a double label may confuse the user and results redundant when dealing with events offered exclusively online. The third case, blended events, is the most problematic as it combines an offline location with a virtual one. The chosen compromise defines the event based on the offline `Location` (spatial relation) but constraints it with an additional `VirtualLocation`. A significant disadvantage of this approach is the event's disassociation from the *online* category. Nevertheless, it satisfies the project's purpose and maintains a high degree of coherence.

Even though an event is defined as an entity temporally and spatially constrained, many exceptions can arise. The proposed ER model tries to overcome three exceptions, namely

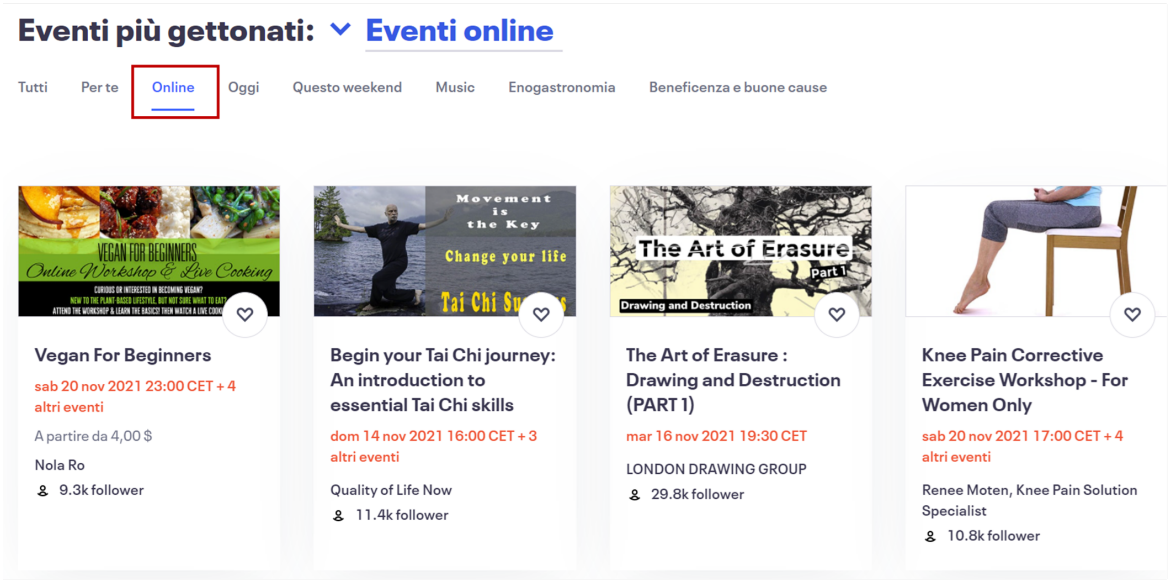


Figure 3: Example of online events promoted on Eventbrite

events belonging to Festivals, Editions, and periodically repeated. The first exception is encoded as a boolean property, `festival`, of the `Event` entity. If true, the event will be linked to another `Event` reporting the information of the festival itself. In this way, the event will be considered as a sub-event of the festival but will maintain its independence. The second exception, the edition issue, is unraveled by placing an `edition` property to the `Event`. This property is filled with numeric references (i.e., integer). Lastly, repeated events can be considered as scheduled. The `Schedule` is a distinct entity stating the temporal difference between consecutive events, such as on a daily, weekly, monthly, or yearly basis. If a `Schedule` is initialized, only one `Event` will be posted, but with the possibility of filtering the preferred time-span (i.e., hour or date), as in Eventbrite (figure 4).

The entity `CreativeWork` is another choice that may be debatable. This concept gathers all human artifacts produced during an activity or employed in an event. The properties for each sub-class have been extracted by benchmarking other services, for example:

- **Track** follows the audio features returned by the SpotifyAPI.
- **Recipe** is mainly modeled on Giallo Zafferano, a well-known Italian food website and magazine that greatly satisfies user's needs but does not offer tabular data. Further attributes were extracted from the RecipeDB food database, which encapsulates the information about recipes, ingredients and nutrition profiles interlinked with flavor profiles and health associations [2].
- **Map** should comply with some standards and minimum requirements depending on the type of product, as the Copernicus program states. In this framework, the maps are defined as a geographical depiction of the territory and routing instrument.

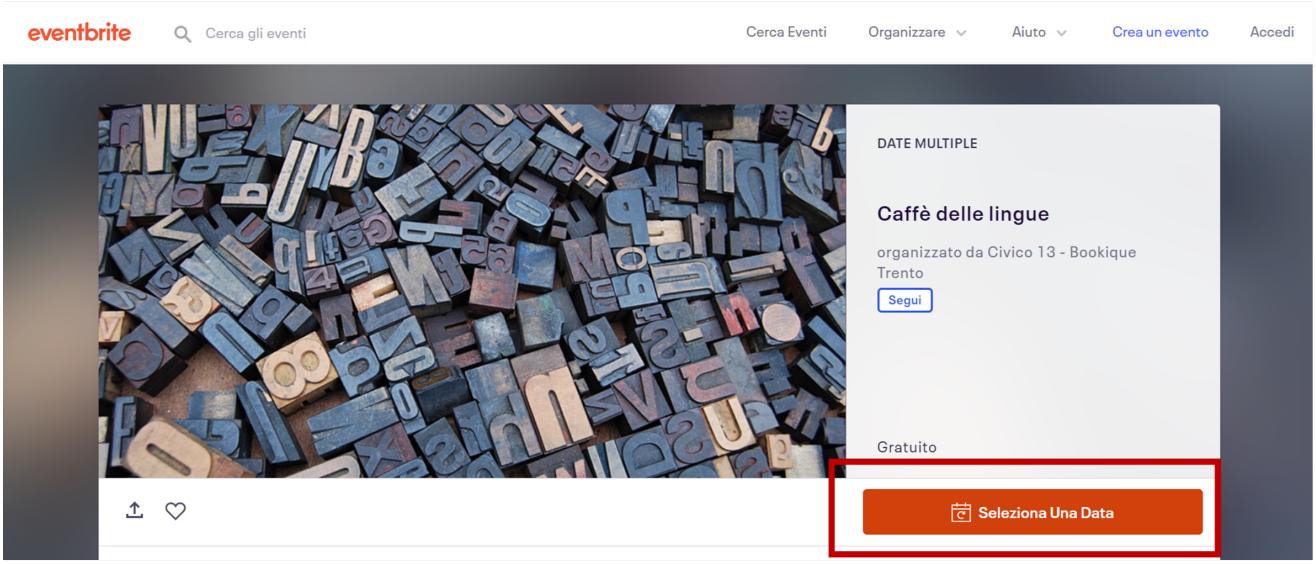


Figure 4: Example of an event scheduled on weekly basis and posted on Eventbrite

- **Image** entity preserves the standard characteristics of a (digital) picture but also mimics Pinterest's functionality via (optional) tags, allowing an easy retrieval.

Regarding the `Organization` sub-classes, two specifications are necessary. Firstly, `PerformingGroup` is an umbrella term for describing single artists, groups, bands, or orchestras but also theater or artistic groups. Secondly, the `Team` sub-entity is based on the tabular structure of TheSportsDB by leaving out the attributes referring to streaming data (e.g., live scores).

To satisfy the necessity of a *memory*, the ER model integrates an `Archive` entity too, aiming at keeping track of events' past history.

A final remark regards the creation of the `ArchitecturalBarriers` entity that enlarges the plethora of users and increases the quality of the final service. The information delivered focuses on the event's venue (and transport) and its accessibility. Due to the lack of standardization in this field, these properties comply with the Italian legislation DM 236/89.

3.3 Informal Modeling evaluation

3.3.1 Schema Level

As in the inception phase, also the schema level of the Informal Modeling phase considers Coverage and Extensiveness as measures of goodness. A relevant difference regards the term to which the CQs are paired: the ER model instead of the Ontologies.

As the table above shows, the Coverage measure frames a solid situation, assessing extremely high values for both Etypes and properties, respectively, $Cov_e = 0.96$ and $Cov_p = 0.95$. However, the goodness of this result is debatable as too high Coverage might imply an ER model

Schema evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 59 \rvert}{\lvert 46 \rvert} = \frac{\lvert 44 \rvert}{\lvert 46 \rvert} = 0.956 \approx$ **0.96**	$\frac{\lvert 121 \cap 217 \rvert}{\lvert 121 \rvert} = \frac{\lvert 115 \rvert}{\lvert 121 \rvert} =$ **0.95**
Extensiveness	$\frac{\lvert 59 - 46 \rvert}{\lvert 59 \cup 46 \rvert} = \frac{\lvert 15 \rvert}{\lvert 61 \rvert} = 0.245 \approx$ **0.25**	$\frac{\lvert 217 - 121 \rvert}{\lvert 217 \cup 121 \rvert} = \frac{\lvert 102 \rvert}{\lvert 224 \rvert} = 0.455 \approx$ **0.46**

Table 4: Summary of the evaluation on the informal modeling phase at the schema level

excessively designed on the CQs. This specificity may prevent its usage in other contexts and fail in delivering a reusable product. On the other hand, the ER model slightly extends the CQs properties' coverage by reaching almost $Ext_p \approx 0.5$, thus safeguarding purpose-specific information. Since the ER model has been structured following a certain degree of generalization and consistency with the purpose, the relative low extensiveness of Etypes ($Ext_e = 0.25$) shows an alignment between this willingness and the obtained result.

3.3.2 Data Level

At the data level, the Coverage measure is paired again with the Sparsity. Comparing the Inception phase with the Informal Modeling one, the two measures were not subject to significant alterations. This condition is justified by the ER model formulation's effort, which postponed other processes, such as data parsing and alignment with the CQs.

Data evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 21 \rvert}{\lvert 46 \rvert} = \frac{\lvert 20 \rvert}{\lvert 46 \rvert} =$ **0.43**	$\frac{\lvert 121 \cap 99 \rvert}{\lvert 121 \rvert} = \frac{\lvert 90 \rvert}{\lvert 121 \rvert} =$ **0.74**
Sparsity	$1 - \frac{\lvert 46 \cap 21 \rvert}{\lvert 46 \cup 21 \rvert} = 1 - \frac{\lvert 20 \rvert}{\lvert 47 \rvert} =$ **0.57**	$1 - \frac{\lvert 121 \cap 99 \rvert}{\lvert 121 \cup 99 \rvert} = 1 - \frac{\lvert 90 \rvert}{\lvert 130 \rvert} = 0.307 \approx$ **0.31**

Table 5: Summary of the evaluation on the informal modeling phase at the data level

4 Formal Modeling

This section is dedicate to the description of the formal modeling phase. Like in the previous section, the current one aims to describe the different sub activities performed and the relative outcomes.
More in detail, it provides a description of the following activities:

- ETG generation
- Data management (syntactic heterogeneity)
- Formal Modeling evaluation

Like the previous phase, also the current one has to report the decision made during the phase activities, along with their weaknesses and strengths. Any difficulty and/or open issue has been reported as well.

4.1 ETG generation

4.1.1 Ontology selection and ER adjustment

Previous to the ETG generation on Protégé, some changes were introduced in the ER model presented in the Section 3. Hence, adjustments were carried out not only with regards to *Object* properties, but with *Data* properties as well. A short explanation justifying the choices taken and embedded in the new ER model (figure 5) is presented below.

- `Event`.
  The CQs requiring to navigate back in time through an event's collective memory (i.e., `Archive`) and those demanding the distance to a certain venue were not satisfied by the previous version of the ER model. While the latter has been solved via a `boolean` Data property `distance` that filters the events based on a ray of $2km$ from a specific venue, the former issue induced more structural changes. Hence the adjustments focused on the consistency of information and the storage costs of *sub-events* belonging to a *super-event*, such as a Festival or a Fair. To simplify the model, the Object property `superEvent` has been introduced, allowing a one-to-one relationship with the main `Event` and guaranteeing a single storage in `Archive`. Moreover, to follow the trail of a collective memory, and to ease future project's expansions, the Object property `post` was introduced, linking to `CreativeWork`, and modeling the act of posting pictures and reviews.

  Within this framework, also the Object property `EventType` and the relative sub-categories have undergone several modifications. Ideally, the ER event-classification model should have been mirrored by the final ETG, as it perfectly satisfied the granularity required by the CQs. However, due to the fact that the majority of our reference files required pondered parsing, and several `Event`'s leaf nodes contained `boolean` properties, a less-detailed but more performing option has been preferred. Thus, the ETG reflects the final ER model, that merges the contextual and core event-related Etypes and keeps the six main sub-categories of `Event`. Hence, the `Event` child `SocialEvent` displays all the Data properties

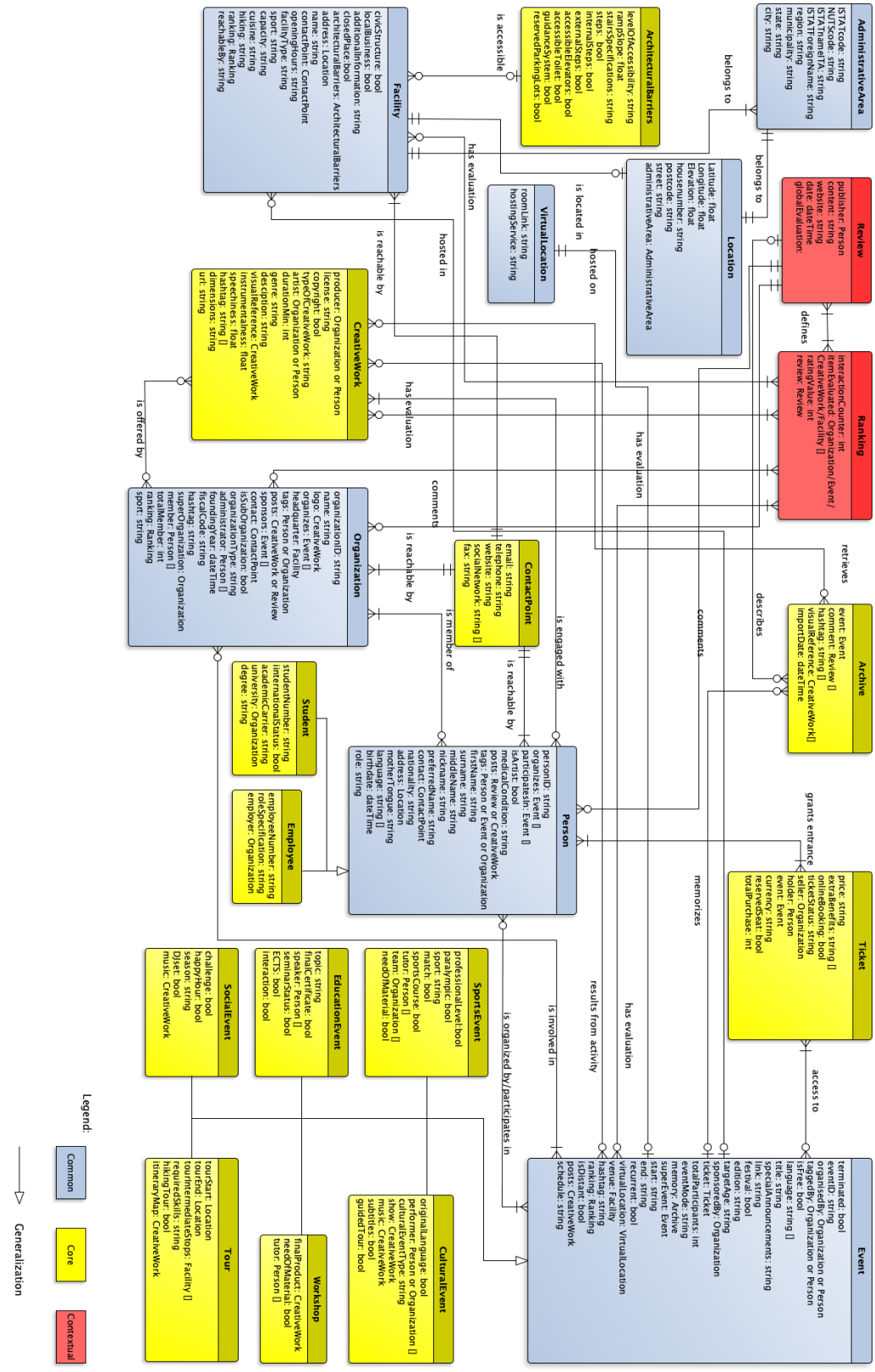


Figure 5: Entity Relationship model with adjustments

contained in DanceEvent (e.g., happyHour) and in FolkEvent. This choice might be highly debatable, as it does not represent the *state-of-the-art* solution, but it is an alternative and feasible way to correctly satisfy the Competency Queries and stick to the Purpose.

Lastly, instead of keeping one Entity per event mode (namely, offline, online, and blended), a new attribute eventMode was introduced. The values allowed should identify (as string) the three event statuses: *online, offline, hybrid*. This simplification skims some relationships between event and location, and retains the most relevant information.

- Duration.
  The Object property duration, contained in Event, was also subject to changes. Thus, the concept of Duration has been simplified and reduced to the Data Properties start and end, which define the Event's range of action in time.

- Schedule.
  Due to the overlapping of Schedule with Event, it has been replaced by two Data properties: recurrency and schedule, respectively defined as *boolean* datatype and *string*. The latter should specify the pace of the Event, guaranteeing both an appropriate output for CQs requiring weekly or yearly-scheduled events, and for those having an irregular pace.

- Facility.
  A trade-off was faced: on the one hand, facilities could be categorized in sub-classes, such as FoodEstablishment or Accommodation, keeping a high level of detail. On the other hand, this structure was not feasible due to the time required in data preparation and management. Therefore, all boolean Data properties of the sub-classes have been merged within facilityType property. Additionally, considering the CQs' requirement of listing specific food (e.g., typicalFood) or activity (e.g., hiking), the contextual attributes have been kept in separate Data properties. Hence, Facility presents properties such as cuisine and hiking. This latter example also solved the issue regarding the erasure of Landscape, that was requiring a granularity out of project's scope.

- Ranking.
  In Ranking, the isDistant property was discarded in light of what discussed for Event and data availability. Thus, instead of keeping both dichotomies distance-ranking and enjoyment-ranking, only the latter has been kept. Moreover, the ER model was lacking the connection between the concept of ranking and the criterion followed, namely the reviews. To fill this gap, the Object property review was introduced in Ranking and linked to Review, a contextual and independent entity.

- CreativeWork.
  The changes regarding Review addressed a further issue: several children of CreativeWork were sharing properties and the available data would not have reached such granularity. Thus, all the common properties of CreativeWork's children have been collapsed and the children (i.e., Image, Show, Recipe, Track and Map) discarded. The super-entity CreativeWork is therefore defined as a "manifestation of creative effort including fine artwork (sculpture, paintings, drawing, sketching, performance art), dance, writing (literature),

film-making, and composition"[1].

- `Archive`.
  Due to the concept of collective memory pursued by the Purpose, some changes were applied on the Etype `Archive` as it was missing some information. Thus, the hyperlinks created by `hashtags`, the `importDate` and the `visualReferences` were added as Data properties.
- Further entities, such as `SeatReservation`, `Transport` and `PostalAddress` have been deleted as such granularity was not expected. Moreover, the information required by some CQs has been either collapsed in the super-entities (e.g., `PostalAddress` moved in `Location`), or added as data properties (i.e., `trasportationMode`) or integrated within already existing properties, like `additionalInformation` or `description`.

When selecting the appropriate concepts to reshape the ER model, some general purpose ontologies such as schema.org and DBPedia have been exploited, along with Wordnet database. Moreover, the concepts related to the spatial information have been treated according to OpenStreetMap's ontologies OSMonto and LikedGeoData. A last open issue regards the definition of architectural barriers, which depends on the country and its relative legislation. Having as spatial coverage the area of Trento and Rovereto, the project reports the guidelines proposed by the cooperative HandiCREA in the mobile application TrentinoAccessibile.

4.1.2 Evaluation on final ER model

As already stated above, *Extensiveness*' null value was consistent with the choice of reducing the number of Etypes in the final ER, while the remainder still reflects the assumptions and considerations that have been defined in Section 3. Moreover, the most relevant loss regards

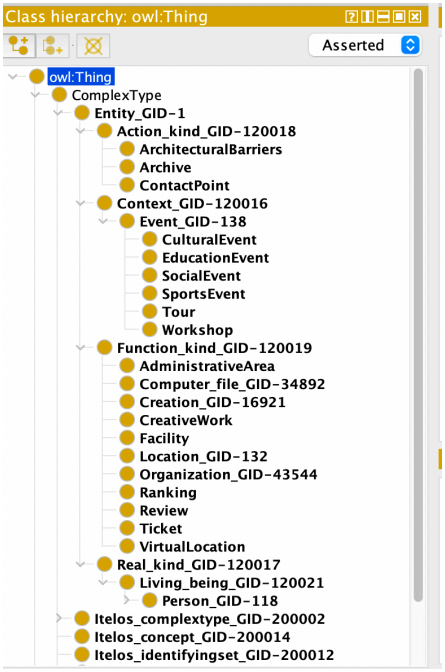
Schema evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 22 \rvert}{\lvert 46 \rvert} = \frac{\lvert 22 \rvert}{\lvert 46 \rvert} = 0.478 \approx$ **0.48**	$\frac{\lvert 121 \cap 210 \rvert}{\lvert 121 \rvert} = \frac{\lvert 115 \rvert}{\lvert 121 \rvert} =$ **0.95**
Extensiveness	$\frac{\lvert 22 - 46 \rvert}{\lvert 22 \cup 46 \rvert} = \frac{\lvert 0 \rvert}{\lvert 46 \rvert} =$ **0**	$\frac{\lvert 210 - 121 \rvert}{\lvert 210 \cup 121 \rvert} = \frac{\lvert 95 \rvert}{\lvert 216 \rvert} = 0.439 \approx$ **0.44**

Table 6: Summary of the evaluation on the informal modeling phase at the schema level

the distinction between events (i.e., the `Event` children) whose description should be the main project's objective but suffered from a low granularity of data and a consequent erasure.

[1] Definition from Wikipedia - Creative Work

4.1.3 Language Alignment via Pròtegè and KOS



As the Data Layer has to be adapted to the Knowledge one, before the ETG implementation an overall alignment between the two layers was performed and implied trimming the ER by re-arranging properties. The adjustments described above were followed by a further pairwise association between each entity and its own `ID`, which is either embedded in data or will be established via an *Identifying Set* in Karmalinker. This is worth noticing, as some choices may represent an ideal solution and will not be feasible on the collection of data here managed.
The ETG generation was based on a two-steps procedure. The first step, performed in Protégé, had three main objectives, namely defining the classes, the object properties and the data properties. While the last two goals were performed straightforwardly thanks to the support of the ER model and required just the addition of the mandatory *has*, the former led to further debatable decisions. The resulting ontology is structured as the caption on the right shows. Specifically, this structure has been modeled starting from the given baseline ontology, having `Event` as pivot class. The classification of the other Etypes followed a deductive reasoning. Specifically, they were assigned based on their dynamic nature (i.e., `Action`) or on their specific role (i.e., `Function`). A distinct class was dedicated to the concept of `Real kind` and the definition of `Person`.
After this first step, the concepts in the ETG are still conceptually lacking a unique and formal definition, and linguistically diverse. The language impairment is clearly detectable due to the lack of several unique identifiers, called **Global IDentifier** (GID) in the UKC (Universal Knowledge Core) framework. A language alignment is performed via the application KOS. Any conflict was faced by either choosing an appropriate definition or establishing a new one. The knowledge sources exploited to fill those gaps are reported below:

- **AdministrativeArea**'s definition is based on the OECD glossary.
- **ECTS** credits report the official definition provided by the European Commission.
- **ArchitecturalBarriers**, as already stated in the Section 1, are defined following the ADA guidelines and conformed to the Italian legislation.
- Subcategories of event have been defined following schema.org and LawInsider dictionary.
- **Review** was already present in schema.org - Review, but the definition provided by IGI-global fits better as it specifies the required nature of review, namely an *online review*.
- **NUTS** code are an European standard and their description reflects the relative Ontology on DBPedia. It is worth noticing that the usual granularity of NUTS codes(i.e., 1, 2 or 3) is hereby not specified and may depend on the user's needs.

- Some properties related to address, such as **housenumber**, was retrieved from GLEIF ontology, meaning the Global Legal Entity Identifier Foundation Base Ontology, while the concept of **region** resembles the ISTAT nomenclature.
- Concepts as **Hashtag** and **Social Network** could not be found in existing Ontologies. Whereas hashtag has been defined as a specification of the **Tag** Ontology SCOT, social network concept rephrases the definition given by Boyd & Ellison [3].
- Another specificity of the Italian framework regards the fiscal code for no-profit organizations. This unique identifier's definition follows the directives of the Agenzia delle Entrate and the AA5/6 model.
- The International Paralympic Committee was the main knowledge source for the concept of **paralympic**.
- The properties referring to **speechiness** and **instrumentalness** come from the Spotify.API and, therefore, are described as on Spotify.com

4.2 Data management (syntactic heterogeneity)

4.2.1 Data types misalignment

Data sources used as reference are quite variegate and present consistently different data types. One may naturally think of aligning syntactic heterogeneity according to a *state of the art* structure, such as the `JSON` keys contained in files from MeetUp or OpenData Trentino. However, this approach might lead to three main issues:

1. The level of granularity provided by those structures may not be reachable when dealing with scraped events' web pages. Furthermore, the pre-defined data are tabular, while the scraped information are unstructured, dense, wordy and does not fit the same set of predefined attributes.
2. The attributes of those (semi) structured data also suffer from intrinsic language misalignment, as the presence of completely empty attributes highlights. This risk may enlarge when generalizing those structures to other data.
3. The data types chosen by those sources might not fully fit the project's purpose when facing a lack of information. For instance, some scraped events do not report the exact starting time or day. Thus, a general `dateTime` data type might imply discarding some relevant information provided in alternative ways (e.g.; every November's Thursday at 4 pm becomes `11/2021 16:00:00` rather than `2021-11-DD 16:00:00`). Therefore, the ideal data type `dateTime` has been imposed in artificial datasets, meaning manually filled. Examples are `CreativeWork` or `Person`, whose dates should be accessible and easily retrievable.

For the above-mentioned reasons, the syntactic alignment has been performed on the basis of the content from scraped websites. Indeed, the largest portion of available data originates from Crushsite, StayHappening, and ESN. Given that the concept of event `Duration` is represented by its start and end date and/or time, there was no need to represent date and time neither in

ISOformat, as no computations have to be performed, nor in dateTime type.
A similar reasoning has brought to the definition of Ticket as class without further distinctions or levels. Therefore, two main reference keys have been introduced in the datasets: has_extraBenefits and has_price. In both cases, the variables are encoded as a string data type. Most of the collected datasets present information in Italian but with English attributes. To allow shareability, the English keywords were used as attributes names even though the information contained is reported in Italian. Additionally, has_language or has_motherTongue attributes follows the ISO 639-1 format, which is in line with the goal of aligning dataset's attributes both to the ER and the existing standards.
The example below aims at clarifying the dataset's structure.

```
event = {"has_eventID": string,
        "has_title": string,
        "has_type":string,
        "has_mode": string, # offline, online or blended
        "has_cost": {
            "has_ticketID":string,
            "has_event": Event,
            "has_freeEntrance": bool,
            "has_onlineBooking": bool,
            "has_extraBenefits": string, # all extras included in the ticket
            "has_totalPurchase": int,
            "has_price": string, #either price or url to the purchasing web service
            "has_currency": string,
            "has_purchaser": string
                    },
        "has_link": string,  # link to the main page of the event
        "has_targetAge": string,
        "has_edition": int,
        "has_festivalStatus": bool,
        "has_language": [],
        "has_start": dateTime, # starting date of the event
        "has_end": dateTime,# ending date of the event
        "has_recurrency": bool, # recurrent event or not
        "has_schedule": string, # define the pace (weekly, daily, monthly, yearly)
        "has_organizer": string,
        "has_specialAnnouncements": string,
        "has_description" :string,
        "has_terminated": bool,
        "has_hashtag": [],
        "has_distance": bool,
        "has_transportMode": [],
        "has_venue": string,
        "has_virtualLocation":string,
        "has_superEvent":string
        }

```

Another structure is presented in the OpenStreetMap dataset, that needed some tweaks to reflect the conceptual changes introduced in the ER model and its related ETG. Specifically,

- Due to the removal of has_menu object property and Recipe entity, the values referring to the diet:type key (i.e., isVegan) have been merged together within the unique concept has_cuisine. Consequently, the concept of *menù* (which did not exist in the original dataset) has been created within has_additionalInformation key of string type.
- A similar reasoning was applied for Facility. Whereas in the first ER model the entity was the parent of many subcategories formatted as boolean data type, the latest version gath-

ers all the different types of facility under the umbrella term `has_facilityType`. This data property describes the type of facility via a `string` (i.e., `has_facilityType = restaurant`).

Even though these choices would result in a less-precise data collection in comparison with the forecast, the available data and the multiple data sources still guarantee a reasonable degree of precision and satisfy most of the CQs listed above (Section 1).

4.3 Formal Modeling evaluation

Since many Etypes were discarded and the ER model adjusted to better fit the data and to frame the project's scope, severe changes in the overall matching between reference Ontologies and the ETG were expected. Therefore, the evaluation was performed on both the reference Ontologies (wide scope) and the CQs (project's scope).

As a matter of fact, already with *Coverage*, controversial values are assessed. While the

Schema evaluation		
	Etype	**Property**
Coverage of Ont	$\frac{\lvert 130 \cap 22\rvert}{\lvert 130\rvert} = \frac{\lvert 18\rvert}{\lvert 130\rvert} =$ **0.14**	$\frac{\lvert 212 \cap 210\rvert}{\lvert 212\rvert} = \frac{\lvert 141\rvert}{\lvert 212\rvert} = 0.665 \approx$ **0.67**
Coverage of CQ	$\frac{\lvert 46 \cap 22\rvert}{\lvert 46\rvert} = \frac{\lvert 21\rvert}{\lvert 46\rvert} = 0.456 \approx$ **0.46**	$\frac{\lvert 121 \cap 210\rvert}{\lvert 121\rvert} = \frac{\lvert 119\rvert}{\lvert 121\rvert} = 0.983 \approx$ **0.98**
Sparsity on Ont	$1 - \frac{\lvert 130 \cap 22\rvert}{\lvert 130 \cup 22\rvert} = 1 - \frac{\lvert 18\rvert}{\lvert 134\rvert} =$ **0.87**	$1 - \frac{\lvert 212 \cap 210\rvert}{\lvert 212 \cup 210\rvert} = 1 - \frac{\lvert 141\rvert}{\lvert 281\rvert} =$ **0.50**
Sparsity on CQ	$1 - \frac{\lvert 46 \cap 22\rvert}{\lvert 46 \cup 22\rvert} = 1 - \frac{\lvert 21\rvert}{\lvert 47\rvert} = 0.554 \approx$ **0.55**	$1 - \frac{\lvert 121 \cap 210\rvert}{\lvert 121 \cup 210\rvert} = 1 - \frac{\lvert 119\rvert}{\lvert 212\rvert} =$ **0.43**
Extensiveness on Ont	$\frac{\lvert 22 - 130\rvert}{\lvert 130 \cup 22\rvert} = \frac{\lvert 3\rvert}{\lvert 134\rvert} = 0.022 \approx$ **0.02**	$\frac{\lvert 210 - 212\rvert}{\lvert 212 \cup 210\rvert} = \frac{\lvert 69\rvert}{\lvert 281\rvert} = 0.245 \approx$ **0.25**
Extensiveness on CQ	$\frac{\lvert 22 - 46\rvert}{\lvert 46 \cup 22\rvert} = \frac{\lvert 1\rvert}{\lvert 47\rvert} =$ **0.02**	$\frac{\lvert 210 - 121\rvert}{\lvert 121 \cup 210\rvert} = \frac{\lvert 91\rvert}{\lvert 212\rvert} = 0.429 \approx$ **0.43**

Table 7: Summary of the evaluation on the formal modeling phase at the schema level

$Cov_e(Ont)$ scores an extremely low performance, the $Cov_p(Ont)$ reaches a high value, indicat-

ing worse adhesion to the Ontologies, but still high granularity. This unbalance may be easily justified by the difference in the cardinalities of the Ontologies ($|Ont_e| = 130$) and the CQs ($|CQ_e| = 46$). Thus, if the latter is considered, the $Cov_e(CQ)$ reaches a value around 0.5, and an even higher value for $Cov_p(CQ)$, indicating an almost excessive adhesion to the purpose.
Another unbalanced outcome is displayed by the *Sparsity* measure. On the one hand, the Etypes in the ETG highly differ from the Ontologies ($Spr_e(Ont) = 0.87$) due to an overall realignment. Thus, the ETG seems to better fit the CQs, namely the result of the realignment, as the $Spr_e(CQ) = 0.55$ shows. On the other hand, the properties register a slight and not significant difference: $Spr_p(Ont) = 0.50$ and $Spr_p(CQ) = 0.43$.
Since the ER model has been restructured, and several context-specific entities with respective properties were erased, low values in the *Extensiveness* measure were expected. Even though the extremely low results of Ext_e show the limited contribution of the created knowledge graph, two further conclusions could also arise. Firstly, a general inconsistency with respect to the Ontologies may be caused by the main knowledge resource employed: *schema.org*. Given that it is a wide-scope resource and aims at generalizing concepts, it loses precision and descriptive power in corner cases. Secondly, the ETG is mainly modeled on the ER, which already had a low *Extensiveness* value with respect to CQs ($Ext_e(CQ) = 0.25$), as the evaluation in Section 3 shows. Therefore, a high $Ext_e(CQ)$ value would have highlighted an extension of the CQs. Within this framework, the extension is neither expected nor wanted due to the risk of modeling an ETG excessively generalized or with entities out of scope.
In conclusion, the consequences of shrinking most of *contextual* Etypes became evident, resulting in a more generic ETG. This outcome does not reflect the ideal solution, as `EventType` categories should have been kept to be more cohesive with the overall Purpose. Although the previous ER would better fit the reality, these outcomes are still fairly good and aligned to the available data as well.

4.4 Open issues

During the language alignment in KOS, one major issue arose and prevented the closure of the formal modeling phase. Namely, when trying to align the data property `has_email`, belonging to the Etype `ContacPoint`, to its meant respective concept, other properties were discarded. Example of properties deleted due to this phenomenon were: `has_onlineBooking`, `has_cp:website` or `has_addr:city`. This unexpected behaviour has been temporarily solved by linking `has_cp:email` to a different and more specific concept, avoiding the previously-mentioned issues. Therefore, although the chosen meaning would have been:

E-Mail: *Communicate electronically on the computer.* GID: 105296

The final association applied corresponds to:

Electronic E-Mail: *(Computer Science) a system of world-wide electronic communication in which a computer user can compose a message at one terminal that is generated at the recipient's terminal when he logs in.* GID: 33745

A further issue arose with the introduction of object property `has_ratingValue`, which has been related to `Rating` domain, but disconnected to `Review`, as oppositely presented in the newly-defined ER.

Lastly, for a broader and more comprehensive evaluation of the richness of the ETG, *Cue validity* computation would have been of use. However, due to a down of the liveschema.eu server, the measurement will be incorporated in the Data Integration phase.

5 Data Integration

This section is dedicated to the description of the data integration phase. Like in the previous section, it aims to describe the different sub-activities and the phase outcomes produced. More in details, this section provides a description of the following activities:

- Data management (semantic heterogeneity)
- Entity matching
- Data integration phase evaluation

Like the previous phase, the current one reports the decision made and their weak and strong points. Any difficulty and/or open issue has been reported below.

5.1 Data management (semantic heterogeneity)

5.1.1 Entity alignment (DTA - 2.2)

This sub-phase aims at mapping multiple entity representations (among different datasets) to the purpose-specific schema, ETG, generated in Section 4. Due to the multiplicity of data sources, the data management and alignment followed a two-steps procedure. While a first step was reserved for parsing scraped data and creating files aligned with the ETG, other data required a different approach. Specifically, if data was already structured, such as those from OpenData Trentino, their attributes have been straightforwardly aligned. An example is the `Organization` file, whose columns have been merged or adjusted to maximise the information expressed. The only dataset requiring a radical change was about the `ArchitecturalBarriers`. Due to a lack of

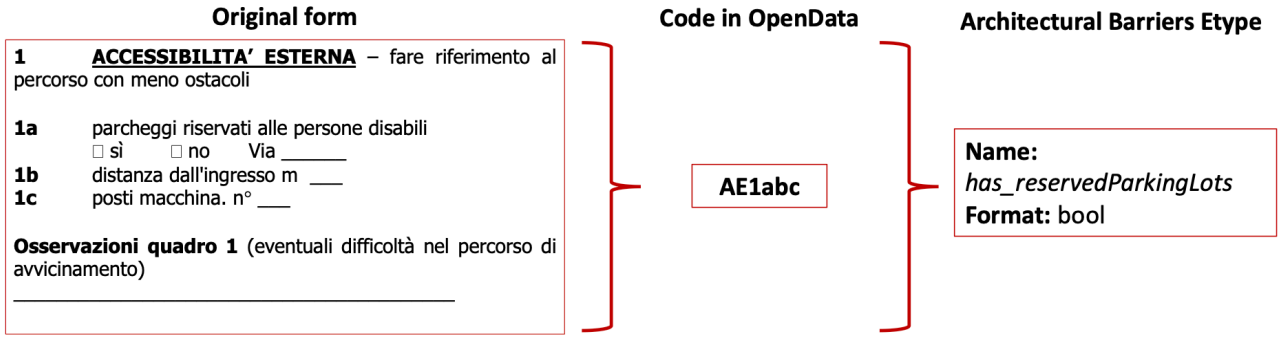


Figure 6: Cross referencing of the OpenDATA Trentino dataset and the original form provided by HandyCREA. NOTE: The example displayed is not implemented in the deployed model.

metadata, a meeting with the cooperative responsible HandiCREA was necessary, along with a crossing procedure to match codes in the dataset with the form used to assess the observations[2].
For OpenStreetMap data, retrieved by the REST API as GEOJson, a (re)structuralization was necessary and followed the guidelines of the ETG. Therefore, as showed below, four main changes were necessary.

- Tags have been merged together, such as `cuisine`, `diet:vegan`, `diet:gluten_free` and `diet:vegetarian`, to create independent properties (i.e., `menu`)
- the attribute's data type was aligned with the ETG
- places retrieved as `MULTIPOLYGON` were reduced to `POINT` geometry (i.e., Python's library shapely) based on their address. This approximation loses granularity but was consistent with the project's purpose and dataset's structure.
- data was split in two distinct `Location` and `Facility` files.

Example of the `GEOJson` retrieved from OpenStreetMap (before the alignment):

```
OSM_record_before = {
        "lat": float,
        "id": int,
        "tags": dict(),
        "lon": float,
        "addr:city": string,
        "addr:country": string,
        "addr:housenumber": string,
        "addr:housename": string,
        "addr:postcode": string,
        "addr:street": string,
        "email": string,
        "name": string,
        "opening_hours": string,
        "operator": string,
        "phone": string,
        "website": string,
        "information": string,
        "tourism": string, # definition of tourism facility (e.g., museum)
        "leisure": string, # definition of leisure facility (e.g., theater)
        "outdoor_seating":string, # wither yes or null
        "amenity": string, # definition of amenity facility (e.g., restaurant)
        "building": null,
        "building:levels": string,
        "internet_access": string, # either yes, no or null
        "parking": string, # either yes or null
        "wikipedia": string,
        "shop": string
    }
```

Example from `Facility` dataset.

```
OSM_record_after = {
  "has_latitude": float,
  "has_longitude": float,
  "opening_hours": string,
  "facility_type": string,
  "architecturalBarriers": string, # code linking to the relative architectural barriers
```

[2]Copy of reference form and summary table will be provided as additional material in the Git repository

```
  "name": string,
  "sport": string, # type of sport played
  "has_capacity": int,
  "menu": string, # type of menu offered
  "hiking": string, # hiking use or not
  "has_osmID": int,
  "has_contactPoint": string, # code linking to the contact point of the facility
  "has_ranking": list # ranking of the facility based on reviews
}
```

Lastly, when dealing with the `Review` dataset, sentiment analysis was performed to extract the feeling thereby expressed. The analysis exploited the Python library nltk, whose categorization was standardized within the range $[0, 10]$. On this scale, 0 corresponds to a strongly negative comment and 10 to a positive one.
Several choices performed in the choices of the identifiers might be debatable. Even though better choices exist, they were not doable within this framework and data. The table 7 displays the identifiers, their format and an example taken from the actual data.

5.1.2 Entity matching (DTA - 2.3)

The problem related to the match between different entities in the datasets pointing at the same real-world entity arose mainly in the case of buildings and their geocoordinates. Indeed, due to a misalignment between the reference systems of the datasets, some places were reported twice. Therefore, a two-step procedure has been employed. Firstly, all the entities in `Location` and `ArchitecturalBarriers` were georeferenced via the open-source geocoder Nominatim in order to obtain latitude and longitude in the reference system WGS84/UTM zone 32N. Secondly, the entities were compared, and the duplicates within the relative datasets erased.
Few cases required an approximation and lacked granularity, as in the case of MART (Museo di arte moderna e contemporanea di Trento e Rovereto), which hosts a museum, a library, conference halls, and art galleries. These facilities shared the address and, therefore, the `osm_id`:

`address`: Corso Bettini, 43, 38068 Rovereto TN
`osm_id`: 7968922

While the main facility (i.e., MART museum) was retained, the others were merged and their records erased. This choice avoided duplicates and seemed the most coherent within the project's purpose.

5.2 Entity matching

The entity matching activity was responsible for the merge of the two distinct layers, knowledge and data. During this phase, KarmaLinker has been exploited. This tool maps the datasets values (representing entities and entity properties) on the Etypes and their properties defined in the ETG. By matching them, the semantic misalignment between the two layers is solved.
Whereas most of the Etypes already included a unique identifier, as the table 7 shows, an *Identifying Set* was established to link each `Facility` to the relative `Location`. An effortless solution was employed, namely the use of existing geocoordinates latitude and longitude. Even though this choice may be debatable, it might be supported by two main reasons. Firstly, all the

addresses and their geocoordinates have been retrieved from OpenStreetMap or georeferenced via the open-source geocoder Nominatim. Moreover, the two sources exploit the same reference system (i.e., WGS 84/UTM zone 32N) to avoid any possible discrepancy. Secondly, the osm_id identifier, retrieving a specific `node` in OpenStreetMap, was already assigned to `Facility`. The Identifying set representing a `Location` within the EG is:

$$IS_{Location} = has\_latitude, has\_longitude$$

It is worth noticing how the matching procedure between the elements of the ETG and the datasets' variables was facilitated by denominations' similarity. Indeed, all datasets have been preprocessed by standardizing and, sometimes, translating the variables' names. Even though a high similarity between the ETG and the data layer was present, some exceptions arose. For instance, the Excel spreadsheet comuni_italiani.xlsx about the Italian municipalities was not modified beforehand and, therefore, reported Italian names and poorly formatted variables (i.e., with spaces and wordy definitions).

5.2.1 Problems of entity matching procedure

During the matching, some problems arose with `bool` variables in `JSON` files that were transformed in `string`. The misrepresentation of data was solved by repeatedly applying the following function:

```
return not getValue("<property_name>") == 'false'.
```

Moreover, the unique identifiers used for `VirtualLocation` were crashing because `URL`s were interfering with `URI` creation. Even though each identifier was formatted as `string` having `VL_+ link event`, it was automatically converted in `URL` by Karmalinker and did not allow a generation of the `RDF` file. Even if not optimal, the solution satisfies the requirements, is feasible and intuitive, but source dependent. Thus, each identifier is defined depending on its data resource:
- **OpenData Trentino:** `VL_ + oggetto_rovereto + url number`. An example is given by the instance:

```
VL_https://www.comune.rovereto.tn.it/openpa/object/129034
    --to--> VL_oggetto_rovereto129034
```

- Web scraped sites are saved in a different way, namely: `VL_+website name+event's title`. A concrete example, having as website ESN, follows:

```
VL_esn-speed-friending-w-esn
```

5.3 Data integration phase evaluation

5.3.1 Quantitative evaluation

Given the formal ETG, the Dataset Schema's sparsity is extremely low. As the table below displays, both values are slightly below 0.2. These results may be interpreted differently. On the

one hand, they emphasize the model's specificity and represent the outcome of a purpose-driven process. On the other hand, they may suggest a biased model and a lack of generalization due to an excessive purpose coherence. Consequently, the model overfits this data and be only partially exploitable in other contexts.

5.3.2 Non quantitative evaluation

The qualitative evaluation of the deployed EG considers three different dimensions: consistency, accuracy and completeness.
Consistency Dimension:

- **Cycles in a class Hierarchy**. Given that only two classes are hierarchically structured, namely `Event` and `Person`, they do not present circles. Even though some data properties may seem redundant, like `studentNumber` and `employeeNumber`, they are not considered sub-identifiers. Indeed, they assess the existence of a `Student` or `Employee` within the databases of a company or university. The person *per se* is not detached from its identifier, which follows the definition above (table 7).
- **Polysemous Elements.** There is one polysemous element due to a oversight:

  `has_record_GID-35630_Type-138` and `has_memory_GID-300040_Type-138`

  As the figure 7 shows, the latter has been preferred and used as object property in the entity alignment.
- **Multiple Domains/Ranges** per property have been avoided. Even though this approach reduces possible conflicts between properties, it also leads to wordy definitions. A clear example is represented by the concept of Item rated in a `Ranking`. It has been fragmented into four distinct object properties instead of a unique `has_itemEvaluated`:
  - `has_evaluationOnCreativeWork_GID-300045_Type-31330`
  - `has_evaluationOnEvent_GID-300122_Type-31330`
  - `has_evaluationOnFacility_GID-300044_Type-31330`
  - `has_evaluationOnOrganization_GID-30012_Type-31330`

  The concepts of identifier, tag and organizer were subject to a similar fragmentation. However, due to an oversight, the definitions of `Review` and `Ranking` identifiers were missing and both entities have been linked to the Data Property:

  `has_identifier_GID-39085_Type-1`

  This property belongs to the Etype `Entity` and should fill the ontological gap satisfactorily, but it introduces an overlap between definitions that should be avoided.
- **Semantically Identical Classes** were strongly avoided. No classes have similar semantics thanks to the adjustments performed on the ER model at the beginning of the formal modeling phase, Section 4.

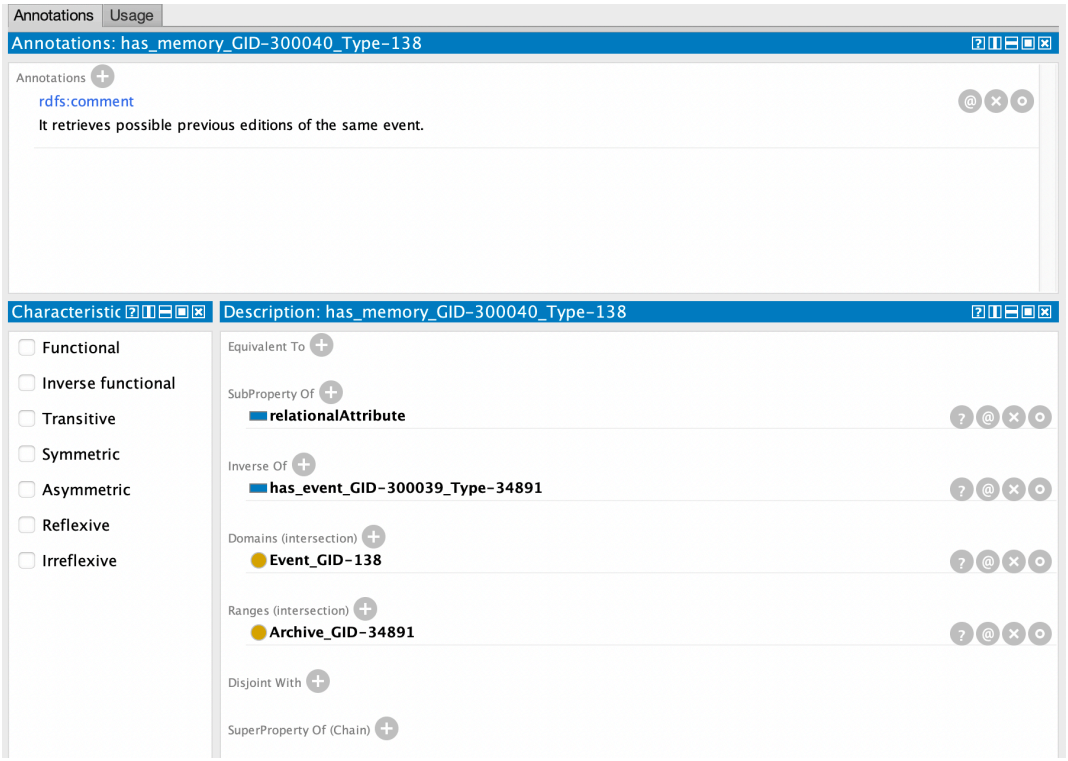


Figure 7: has_memory object property

Accuracy Dimension

The accuracy of the deployed ETG should meet the purpose's needs. Indeed, the relationships are always defined as `has_` followed by a substantive or declined verb. Moreover, relationships such as `sameAs` and `isA` were not used due to their generic meaning. Wordy but more intuitive definitions have been preferred. Specifically, they assemble `has_typeOf` with the Etype denomination (e.g., `has_typeOfCreativeWork`) or the Etype with `type`, like `has_facilityType`. As stated above, the model fits available data and, therefore, does not present a hierarchy over-specification. Indeed, any leaf class without instances was discarded. Thanks to a similar approach, all miscellaneous classes have been pruned already in the final ER model.

Completeness Dimension

The ontology is deployed as complete, meaning without missing domain or range in properties and absent isolated elements. Thus, all the elements declared have been used with the only exception mentioned above (i.e., `has_record`). However, a typo is present in the definition of the `has_beenProducted` property, which should be replaced with `has_beenProduced`. A further typo, `has_headquarter`, should be replaced with `has_headquarters`

6 Open Issues

This section describes any issues/problems remained open along the DI process. The description of open issues provides a clear explanation about the problems, the approaches adopted while trying to solve them, and eventually, any proposed solution that has not been applied.

6.1 Identifiers' definition

The table 7 summarizes all the identifiers, except the *Identifying Set* of `Location`. The objectives of an identifier are to differentiate between entities and grant their existence. Whereas the former goal is achieved, the latter may be disregarded. Thus, several identifiers, as creative work or ticket, imply a dependence on an event even though it does not hold in the real world. A clear example is given by a creative work, which exists *a priori* and should be linked to the event and not defined by it.
The deployed solution is acceptable in a data-driven procedure with questionable goodness but should radically change in a more realistic framework.

6.2 Language alignment

As stated in the Open Issues section of the Formal Modeling phase, problems arose during the language alignment in KOS. Along with the issue reported in that context (section 4) and referring to the definition of `has_cp:email` property, another unexpected behaviour was registered. Thus, the properties whose denomination contain a colon, namely all properties with `has_cp:` and `has_addr:`, were subject to a forgetting after refreshing the web page.
This behaviour slowed down the process, but the language alignment has been carried out and successfully closed without consequences on the data integration phase.

6.3 Approximation of some Etypes

Several Etypes were characterized following a purpose-driven approach. Therefore, they may not be generalizable to other contexts. Two main issues might be pointed out. Firstly, a specification due to the spatial boundaries of the project. Several properties belonging to administrative area and architectural barriers classes apply only to the Italian or, even specific, regional scenario. Thus, they rely on territorial (national, regional or provincial) legislation or depend on Italian institutions, such as ISTAT (i.e., `has_ISTATcode`) or Agenzia delle Entrate (e.g., `has_fiscalCode`). Secondly, some classes are overgeneralized due to a lack of data. An emblematic case is the facility class, whose property `has_facilityType` is just a buffer for places' multiplicity of usages and/or services.

6.4 Events' duration and location

Since the Inception phase (Section 2), a major problem arose. Before the Covid-19 pandemic, an event was commonly addressed as a point in space happening at a given time. Whereas

the concept of *duration* was already subject to problems, the past definition of *location* started to be problematic as well.

6.4.1 Duration

The concept of duration, namely a starting and ending point bounding a temporal interval, seems straightforward, but it is often disregarded in real-world applications. Events' such as festivals, art expositions or courses last more days (or months) and suggest the existence of decoupled levels. The first level refers to the duration of the main event (e.g., festival), while the second level covers the time a subevent (i.e., a conference, a visit or a lesson) lasts. Aiming to implement a model encoding both levels, three different solutions were exploited:

- **Duration**. It was introduced as Etype and object property of `Event` in the ER model of section 3. That implementation was conceptually correct because it guaranteed a precise definition of a temporal interval by setting both date and time. However, two main limitations were underlying. Firstly, it still implies duplicating each event, namely recording a sub-event as contained in a super one. Secondly, with the available data, this solution would not be feasible.
- **Schedule**. Along with `Duration`, the first ER model presented the scheduling concept. As schema.org suggests, the `Schedule` entity would simplify courses' encoding and paced appointments. Once again, this optimal solution was not feasible due to the quality of scraped data. Moreover, as schema.org reports, there is still pending feedback about the term's implementation in applications and websites.
- **Calendar**. The introduction of a temporal entity `Calendar` could be another feasible solution, also thanks to the already existing W3C Time Ontology. This structure solves the issue similarly to `Duration`. Thus, it decouples the super- and sub-events into distinct levels by applying different time-spans to them.

Even though the chosen solution is not the best one, it simplifies the concept of Duration by reducing it to two Data Properties setting the boundaries: `has_start` and `has_end`. Moreover, the pace of an event is established using the properties `has_recurrency` and `has_recurrencySpecification` along with a decoupling between super- and sub-events. The sub-event calls the parent via the Object Property `has_superEvent`, which guarantees a *per-se* existence but also an underlying degree of dependence.

6.4.2 Venue and location

As stated in Section 3, the concept of `Location` leads to problems related to the coherence with the real world. Two main issues can be addressed. Firstly, the introduction of a `VirtualLocation` when speaking about online or blended events (e.g., webinar). Secondly, the difference between `Location` as anonymous point in the space displayable via geocoordinates, and the place where an event takes place, hereby addressed via `Facility`. The solutions proposed in this project are debatable but the most suitable ones given the purpose and data.

- **Location vs VirtualLocation**
  A trade-off was faced: on the one hand, each `Event` entity can state its `mode`, namely its online, offline or blended nature, as a Data Property. On the other hand, `VirtualLocation` and `Location` (specifically `has_venue`) are defined and filled depending on the event's nature. Thus, if an event is blended, they will be both filled; otherwise, only one will be linked to other Etypes. In this way, events organized exclusively online will be categorized as *on line* and will not be subject to any spatial boundary.
- **Location vs Facility**
  On most websites, an event is usually retrieved or categorized by calling a facility's name or showing its position on a map. The geospatial coordinates are retrieved to localize it when neither the facility name nor the address is available. Because the final service of this project should embrace high usability, the preferred choice followed the user's perspective. Hence, each `Event` is linked to a `Facility`, meaning an Etype representing a place with a specific purpose or offering a service. Each `Facility`, in turn, is connected to a `Location` that localizes it within the space and labels it with an address (whenever possible).

7 Outcome exploitation

The final section of the current document aims to provide a description of the data integration process outcome. Here have to be reported the final Knowledge Graph (KB) information (like, number of etypes and properties, number of entities for each etype, and so on). Moreover this section has to provide a description for the KG possible exploitation.

7.1 KG general information

As the figure 8 shows, GraphDB's Knowledge Graph is fully exploitable through `CONSTRUCT WHERE` queries. Each instance is clearly defined and the user is able to reach the desired output. Furthermore, `SPARQL` queries can be manually applied but only partial results have been achieved due to the team's lack of expertise.

7.2 KG exploitation

On the one hand, the manual exploitation of the graph displayed its fully potential. On the other hand, three `SPARQL` queries were chosen, reflecting the willingness of displaying user searches (i.e., CQs) respectively falling within `common`, `core`, and `contextual` concepts. Those queries corresponds to the following CQs:

- *Giovanni* - **common** : Return all university student associations in Trento that organize events.
- *Martina* - **core** : Get all the events targeting university students and weekly offered in the bars of the city center.
- *Daniel* - **contextual** : Knowing the title of the Event, retrieve its evaluation and reviews.

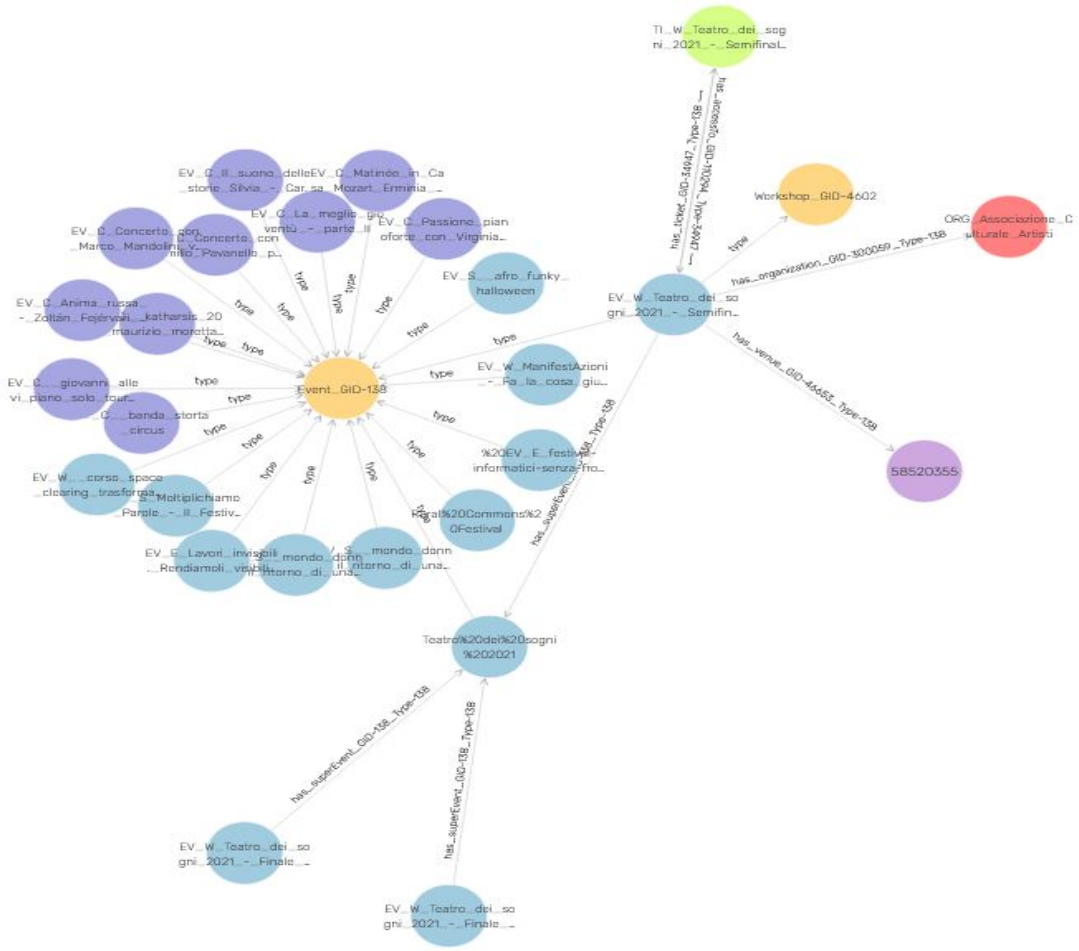


Figure 8: KG partially exploited

An example of the (partial) query is reported below:

```
# Get all the events targeting university students and weekly offered in the bars of the city center

PREFIX foaf:<http://knowdive.disi.unitn.it/etype#>
SELECT ?target ?title ?pace ?name
WHERE {
    ?s a foaf: SocialEvent GID-300005;
    foaf: has title it GID-300033 Type-138 ?title
    foaf:has_targetAge_GID-300054_Type-138 ?target
    foaf: has recurrencySpecification GID-300123 Type-138 ?pace
    foaf: has venue GID-46653 Type-138 ?venue ;
    FILTER (?target='18-30' && ?pace= "weekly")

    ?venue foaf: has officialName GID-34017 Type-17982 ?name.
    }
```

As stated above, the manual searches are still possible by running the following snippet:

```
PREFIX foaf: <http://knowdive.disi.unitn.it/etype#>

```

```
CONSTRUCT WHERE{
    ?s a foaf:Event_GID-138
}LIMIT 100
```

7.3 Issues along iTelos procedure

Three main issues should be pointed out because they persisted during the Itelos procedure and (partially) affected the final deliverable.
At the schema level, the KG was structured following a purpose-driven approach and, therefore, the outcome focuses on a model representing events. Even though the event representation is consistent and feasible, it should extend its scope and encompass other existing Ontologies. Thus, a deeper understanding of Etypes as `Organization`, `Duration`, and `Person` might be necessary for a real-world scenario.
Further issues regarded the data level. Firstly, the GDPR sets specific limitations to the disclosure of personal information [5]. Accordingly, all the information related to individuals and not publicly disclosed were discarded. Due to this radical forgetting process, the `Person` and `Review` datasets report only a few records and with partial information. A real-world scenario would deal with higher quality and granularity of data, resulting in a more performing KG exploitation. Secondly, the overall data management procedure required demanding data cleansing, parsing, and adjustment. This procedure followed the choices taken at the schema level, and each step of the data pipeline is reported in the public git repository used by the team to organize and parallelize the work. It is worth noticing, as the last adjustments are not reported as Python scripts because they were either manually performed, or directly applied in Karmalinker.

7.4 Material and repositories

All the phases described above and the relative (meta)data can be found at the following links:

- Events in Trentino Gitpage, a website generated from the main repository containing all the information needed to understand the project
- Events in Trentino, repository populated along with the iTelos procedure
- KDI-project, public GitHub repository used by the team members to organize and parallelize the work

References

[1] A. Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[2] D. Batra, N. Diwan, et al. Recipedb: a resource for exploring recipes. *Database*, 2020, 2020.

[3] D.M. Boyd and N.B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[4] S. Collins, F. Genova, et al. Turning fair into reality: Final report and action plan from the european commission expert group on fair data, 2018.

[5] European Commission. *2018 reform of EU data protection rules*, 2018.

[6] G. Everest. Basic data structure models explained with a common example. *Computing Systems*, 1976.

[7] S. Hanaei, A. Takian, et al. Emerging standards and the hybrid model for organizing scientific events during and after the covid-19 pandemic. *Disaster medicine and public health preparedness*, pages 1–6, 2020.

[8] J. A. L. Ludvigsen and J. W. Hayton. Toward covid-19 secure events: Considerations for organizing the safe resumption of major sporting events. *Managing Sport and Leisure*, pages 1–11, 2020.

[9] K. Ritchie and D. Chan. The emergence of cognitive covid. *World Psychiatry*, 20(1):52, 2021.

[10] Princeton University. *Princeton University "About WordNet."*, 2010.

[11] M. D. Wilkinson, M. Dumontier, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Identifiers			
Etype	**Identifier**	**Data type**	**Example**
Administrative Area	ISTAT code for municipality (i.e., Codice Comune formato alfanumerico)	`int`	*001001*
Architectural Barriers	AB_ + either identifier assigned by OpenData Trentino or random integer in range $[1000, 100000]$	`string`	*AB_3813*
Archive	AR_ + event's identifier	`string`	*AR_EV_S_Processione_annuale*
Creative Work	CR_ + event's identifier	`string`	*CW_EV_C_NO_TIME_TO_DIE*
Contact Point	CP_ + random integer	`string`	*CP_9194104150*
Event	EV_ + one letter category (i.e., *C* for cultural) + event's title	`string`	*EV_E_Ivona_Pablo_Girolami_*
Facility	osm identifier	`int`	*269440263*
Organization	ORG_ + either identifier assigned by OpenData Trentino or random integer in range $[1000, 100000]$	`string`	*ORG_2394*
Person	PR_ + random integer in range $[1000, 100000]$	`string`	*PR_42531*
Ranking	RA_ + name of the item evaluated	`string`	*RA_bookique1*
Review	RE_ + name of the item evaluated	`string`	*RE_bookique1*
Ticket	TI_ + event's identifier	`string`	*TI_E_Abitare_la_speranza*

Table 8: Summary of the unique identifiers used at data level

Data evaluation		
	Etype	**Property**
Sparsity	$1 - \frac{\lvert 22 \cap 18 \rvert}{\lvert 22 \cup 18 \rvert} = 1 - \frac{\lvert 18 \rvert}{\lvert 22 \rvert} =$ **0.18**	$1 - \frac{\lvert 210 \cap 204 \rvert}{\lvert 210 \cup 204 \rvert} = 1 - \frac{\lvert 186 \rvert}{\lvert 228 \rvert} = 0.185 \approx$ **0.19**

Table 9: Summary of the evaluation in the Data Integration phase at data level