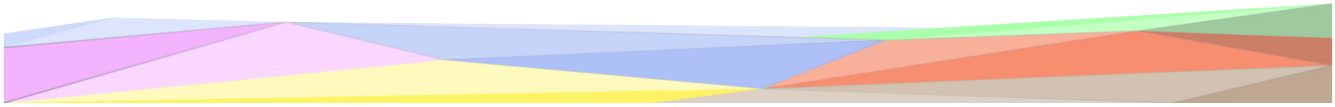
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Trentino Territory & Tourism Facilities

Document Data:	Reference Persons:
October 25, 2023	Marina Bueno García, Lucía Trillo Carreras



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 10, 2023	Marina Bueno García, Lucía Trillo Carrera	Document created
0.2	October 10, 2023	Marina Bueno García, Lucía Trillo Carrera	Purpose and Domain of Interest
0.3	October 10, 2023	Marina Bueno García, Lucía Trillo Carrera	Project Development
0.4	October 10, 2023	Marina Bueno García, Lucía Trillo Carrera	Purpose formalization

1 Introduction

2 Purpose and Domain of Interest (DoI)

2.1 Project purpose

The purpose of this project is to provide in a single place all the information that is available on nature tourism in the Trentino territory. To provide that service we will build a knowledge graph (KG) that provides all the information about the natural tourism attractions such as lakes, natural parks, camp sites, waterfalls... and how to get there via bus/train/bike/taxi.

2.2 Project Domain of Interest

The domain of interest for this project is natural tourism in the 176 municipalities of the Trentino province 1. There is no temporal boundary as the region receives tourism throughout the cold (ski tourism) and hot seasons.

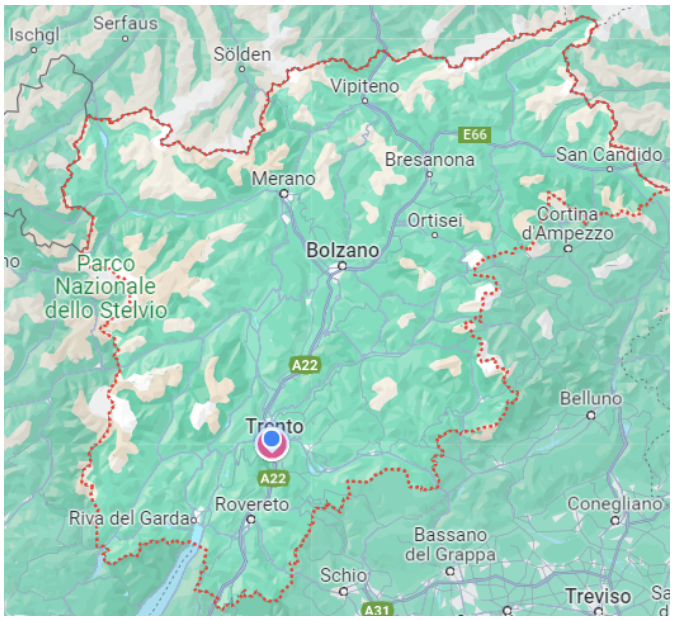


Figure 1: Trentino Autonomous Province

3 Project Development

To achieve the project's purpose described before, we divide the project development into two main subsections:

3.1 Data Production

In the data production phase, we focus on the role of the data producer and how to acquire the necessary data to later build the KG given the purpose.

As data producers, our mission is to make this data as available and reusable as possible, (compliant with the quality and reusability guidelines defined by iTelos. 6*, or at least 5*). We will document and publish the project result on Github and on the KnowDive group so that future researchers or tourists can easily access it.

3.2 Data Composition

As data consumers, our mission is to compose the high quality formal resources produced into the final Knowledge Graph. We will perform this task using the iTelos methodology, after thoroughly defining our Language, Knowledge and Data. The quality of this Knowledge graph will be evaluated based on the Competency questions that will be developed in the next section.

4 Purpose Formalization

In this section, with the goal of identifying the possible use cases of our resources, we first divide the purpose formalization into 3 different subsections: scenarios definition, personas and competency questions (CQs). Once we have listed down the competency questions, we fill a Purpose Formalization sheet (PFsheet) to extract the concepts identifying the information entities and their properties. Finally, the last part of the purpose formalization and the first phase of the iTelos is to shape a ER model to representate formally the initial purpose.

4.1 Scenarios definition

In this section, we outline three distinct scenarios designed to cater to diverse preferences and interests in exploring the natural wonders of Trentino. Each scenario offers a unique perspective, guiding individuals through different levels of adventure and nature immersion:

1. Challenging Adventure: a plan to enjoy adventurous hike in Trentino's mountains, looking for challenging trails and stunning views.
2. Moderate Nature Experience: an exploration of Trentino's hills, combining nature with cultural experiences, seeking easy to moderate trails.
3. Relaxing Scenic Strolls: gentle walks around picturesque locations in Trentino, focusing on easy trails and accessible spots.

4.2 Personas

In this section, we introduce six diverse personas, each bringing a unique perspective and set of preferences to the exploration of Trentino's natural treasures.These personas provide a diverse lens through which we can tailor scenarios and recommendations, ensuring a personalized and enriching experience for each individual exploring Trentino.

Id	Origin	Nature expertise	Age	Description
1	Italy,Lombardy	High	55	**Antonio** is a local outdoor enthusiast, familiar with Trentino's trails, seeks challenging hikes and hidden gems. He enjoys skiing after picking it up from his business trips to the Alps. He has a car.
2	Italy,Toscana	Medium	28	**Giulia** is a young professional traveling with her boyfriend, Matio. Giulia enjoys art, history and good coffee .She is visiting Trentino for the first time, seeks diverse trails and wants detailed information. Matio loves biking
3	Italy,Toscana	Low	27	**Mario** is a young professional traveling with her wife, Julia, and children by car. A tourist with minimal nature experience, wants picturesque spots, child-friendly and needs guidance on transportation options.
4	Czech Republic, Brno	High	23	**Veronica** is doing her Erasmus for 6 months in Trento and wants to thoroughly explore everything the region has to offer. Time is not a limitation on her travels, but money is. She's into climbing and doing challeging hikes.
5	Germany, Munich	Low	32	**Dalim** is a novice hiker on a weekend vacation, interested in scenic spots and gentle walks, prefers easily accessible locations.
6	Poland, Krakow	Medium	26	**Tymoteusz** is a nature lover, keen on exploring Trentino's beauty at a relaxed pace, enjoys both hills and mountains. He loves swimming and beautiful rocks

4.3 Competency questions

In this section, we articulate a set of key competency questions that serve as the foundation for unlocking the information within the Trentino Knowledge Graph. These questions are strategically designed to cater to the diverse needs and interests of individuals exploring Trentino's natural wonders.

1. What are the top three challenging mountain trails in Trentino?
2. Can you suggest a moderate hill trail with historical points of interest?
3. Are there any beginner-friendly walks with scenic views near Trento?
4. How can I reach the Dolomites?
5. Can you provide a list of must-visit lakes in the Trentino region?
6. Who can I ask for general touristic info in each comune?
7. Can you recommend a nature guide/tours or group?

8. What public transportation options are available from Trento to the starting point of a beginner-friendly trail?
9. Any recommendations for family-friendly nature activities in Trentino?
10. Where can I buy sleeping bags and a basic tent?
11. What points of Trentino's natural landscape can be visited in a day without strenuous physical activity?
12. What transportation is available from Trento to popular hiking destinations?
13. Where I can rent a bike?
14. How can I access the different skis stations in Trento?
15. Can you suggest a budget-friendly transportation option?
16. What are the top three panoramic viewpoints accessible by car?
17. Which caves or grottoes are worth exploring for those interested in geological formations?
18. Can you provide a list of mountain shelter of camping sites along their chosen mountain trails?
19. What is the closest gas station to every ski station?
20. Which natural attraction is close to this train station?

4.3.1 PF Sheet

In this table, we present a comprehensive overview of the relationships between scenarios, personas, competency questions (CQs), entities, properties, focus, and popularity within the context of our Trentino Knowledge Graph project. The table encapsulates the intricate web of connections, guiding the design and implementation of our knowledge graph.

Scenarios	Personas	CQs	Entities	Properties	Focus	Popularity
2,3	1,3,5	3,16	tourist_trip	itinerary, tourist_type, trip_origin, arrival_time, departure_time	Core	Core
1,2,3	1,2,3,4,5	1,2,4,9	mountain	address, aggregateRating, geo, hasMape, photo, publicAccess, review, specialOpeningHoursSpecification, tourBookingPage, telephone, description, image, name, url	Core	Contextual
1	1,2,4,6	18	accomodation	accommodationCategory, numberOfBedrooms, occupancy, petsAllowed, tourBookingPage, address, aggregateRating, photo	Contextual	Contextual
1,2,3	3,6	5	lake	address, geo, aggregateRating	Core	Contextual
1,2,3	2,3,4,5,6	6,7	person	name, email, affiliation, birthdate, gender, nationality, owns	Common	Common
1,2,3	2,4,5,6	8,12,15	BusStation	arrivalBusStop, departureBusStop, openingHours	Contextual	Contextual
1,2,3	2,4,5,6	8,12,15	TrainStation	arrivalStation, departureStation, openingHours	Contextual	Contextual
1,2,3	2,4,5	10,13	store	address, brand, department	Common	Common
1,2,3	1,3	19	GasStation	adress, brand	Contextual	Contextual
1,2,3	1	14	SkiResort	address, name	Core	Contextual
1,2,3	2,4,5,6	17,11,20	Landform	adress,geo	Core	Contextual
1,2	1,3	16,19	Car	name	Common	Common

4.3.2 ER model

In the landscape of Trentino's natural tourism, we have designed a robust Entity-Relationship (ER) model to encapsulate the information tpresented . The ER model is the backbone of our knowledge graph, crafted to establish meaningful connections between entities, personas, scenarios, and competency questions. It serves as the architectural blueprint that not only structures our data but also provides a comprehensive framework for users to navigate and explore the diverse facets of Trentino's natural tourism.

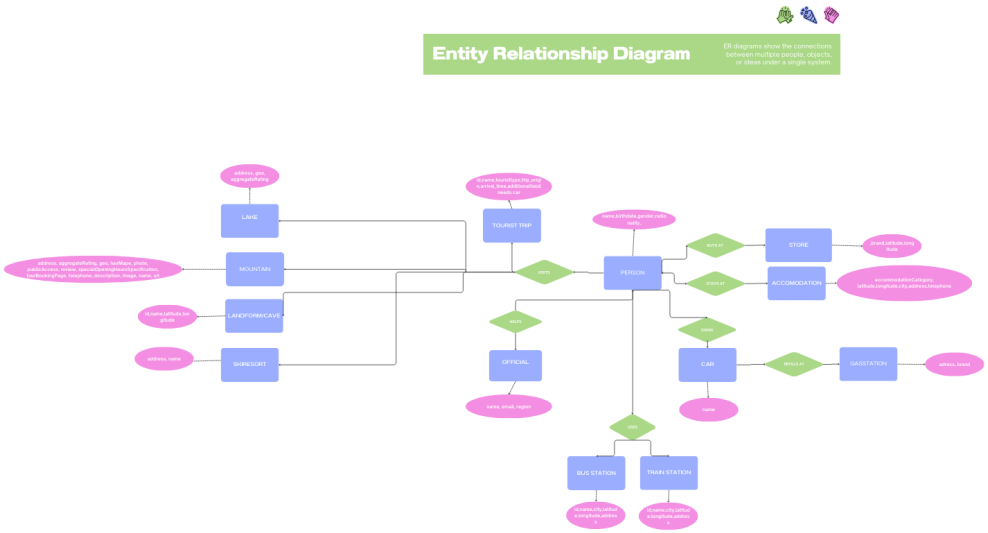


Figure 2: ER model

5 Information Gathering

In this phase of the project both, the Data Producer and the Data Consumer, play an important role. From the producer side, the aim is to collect informal resources from sources with an higher level of heterogeneity and from the consumer side, the aims is to collect formal resources to be composed with the objective of building the final KG.

5.1 Source identification

The original sources of information for the knowledge and the data layer are the following:

- Knowledge layer
  - Formal: Open Street Map - Trentino Territory Lightweight Ontology
  - Standard references:
    - SCHEMA.ORG describes the hierarchy of different Entity types and data types.
    - SCHEMA.ORG LOV schema of the schema.org vocabulary
    - GTFS STATIC is the General Transit Feed Specification, that allows to get data in the correct format. This link describes the format and structure of the files.
    - GTFS LOV describes the GTFS vocabulary.
    - GTFS UPGRADED (Subashish) describes the GTFS ontology
    - GEOSPATIAL ONTOLOGY (Subashish) is a data model for the geospatial domain.

    * TIME ONTOLOGY is a vocabulary for temporal entities such as time intervals, their properties and relationship.

- Data layer
  - Formal: Trentino OSM places

    This dataset contains information scraped from Open Street Maps about the Trentino region: springs, peaks, cave entrances and other natural wonders are listed here.
  - Semi-Formal:KGE22 - Trentino Tourist Facilities This source contains individual datasets of bus,train and bike stops; taxi companies; bars and restaurants; campsites, hotels and vacation houses;gas stations and souvenir shops; and museums, ski stations and natural attractions.
  - Informal:
    * ISTAT Turismo This source contains contact emails of experts in the ISTAT declaration (economic activity classification) by municipality. For the moment, we see no relevance to the project.
    * OPEN DATA TRENTINO: This source contains 6.556 different datasets of the Trentino province of different data formats and topics.

5.2 Resources collection and scraping

However, as stated in our purpose, we mainly aim to cover everything related to nature tourism (lakes, mountains, hikes, ski and the transportation to these places), so we extracted the data that fits our needs the best in the following files (github):

File name	Source	Topic	Classification
acommodation.csv	Trentino Tourism KG 2022	Accomodation	Common
bus_station.csv	Trentino Tourism KG 2022	Bus transportation	Core
train_station.csv	Trentino Tourism KG 2022	Train transportation	Core
gas_station.csv	Trentino Tourism KG 2022	Gas Station	Core
skiresort.csv	Trentino Tourism KG 2022	Ski resort	Contextual
person_tourism_official.csv	ISTAT Turismo	Local information reference people	Common
mountain.csv	OSM Places	Montains	Contextual
lake.csv	OSM Places	Lakes	Contextual
caves.csv	OSM Places	Landform (caves)	Contextual
store.csv	OSM Places	Mountain equipment stores	Core
tourist_trip.csv	Komoot	Trails and hikes	Contextual
car.csv	Custom	Our personas	Contextual
person_tourist.csv	Custom	Our personas' cars	Contextual

All resources were distributed except for those published online on the Komoot website. We tried several approaches, we asked the owners for the dataset, but they only provide it to companies. Next we tried to scrap the data from the source with BeautifulSoup4, but it was not possible. So we moved on to the next approach, creating our own data using the information provided on the Komoot website.

The progress of the scraping attempts can be seen in the following notebook.

	A	B	C	D	E	F	G	H	I	J
1	Numero	Codice comune ISTAT	Comune	Azienda per il turismo	Referente sta	e-mail				
2	3	22003	Aldeno	Trento, Monte Bondone e Altopianc	Pierangela Pe	pierangela.pedron@trento.info				
3	2	22002	Albiano	Trento, Monte Bondone e Altopianc	Pierangela Pe	pierangela.pedron@trento.info				
4	5	22236	Altopiano della Vigolana	Altipiani cimbri e Vigolana	Verena Pasca	office@alpecimbra.it				
5	8	22006	Arco	Garda Trentino, Valle di Ledro, Term	Marica Cattoi	marica.cattoi@gardatrentino.it	claudia.ventura@gardatrentino.it			
6	33	22042	Carisolo	Madonna di Campiglio, Pinzolo, Val	Claudia Trogg	sales@campigliodolomiti.it				
7	30	22038	Canal San Bovo	San Martino di Castrozza, Primiero e	Wilma Loss	infovanoi@sanmartino.com				
8	26	22034	Caldonazzo	Valsugana, Tesino e Valle dei Mochi	Anna Dorigon	anna.dorigoni@visitvalsugana.it				
9	28	22036	Campitello di Fassa-Ciampedel	Val di Fassa	Nicoletta Artc	infocanazei@fassa.com				
10	27	22035	Calliano	Rovereto, Vallagarina e Monte Bald	Matteo Zocca	zocca@visitrovereto.it				
11	32	22040	Capriana	Val di Fiemme e Val di Cembra	Susanna Bone	susanna.bonelli@visitfiemme.it				
12	31	22039	Canazei-CianacÃ¤i	Val di Fassa	Nicoletta Artc	infocanazei@fassa.com				
13	29	22037	Campodenno	Val di Non	Elisa Widman	info.coredo@visitvaldinon.it				
14	25	22033	Caldes	Val di Sole	Mavie Cattoi	mavie.cattoi@visitvaldisole.it				
15	61	22083	FiavÃ	Garda Trentino, Valle di Ledro, Term	Marica Cattoi	marica.cattoi@gardatrentino.it				
16	68	22093	Giustino	Madonna di Campiglio, Pinzolo, Val	Claudia Trogg	sales@campigliodolomiti.it				
17	65	22090	Frassilongo-Garait	Valsugana, Tesino e Valle dei Mochi	Anna Dorigon	anna.dorigoni@visitvalsugana.it				

Figure 3: Example of input dataset

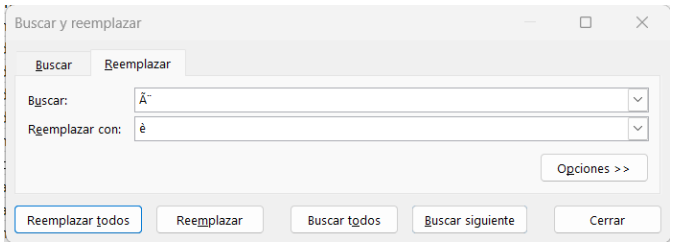


Figure 4: Substitution by ASCII characters

5.3 Data cleaning and formatting

The data cleaning process involved standard tasks such as deleting columns of entities that are not relevant to our purpose. It also comprised modifying word accents, removing repeated entries and null values, transforming data from the form POINT(longitude, latitude) into two different columns to more complex tasks.

A vast majority of our data is location-based, and the formating and information available is very varied amongst entities. In an effort to normalise, we are using the geopy python library to fill in the blanks and enrich current information.

In order to use it, we have modified the coordinates format (in different ways for the OSM and KGE2022 sources) to provide 2 floating points with the correct decimal numbers. We have coded using the spanish version of Excel, which uses commas instead of points to determine decimal points, and uses points to mark the division between the thousands, hundreds of thousands and so on.

Once this transformation was completed, we were able to get the city and address of all the bus stops using the longitude and latitude.

The complete and annotated description of the geographical transformations can be found on

```
#latitude and longitude as floats
#problem -> 2 points on the format, we need to do this transformation 11.118.457 ->  11.118457
#only keeping the first point of each coordinate

for i in df.index:
  longitudeoutput = df.longitude[i].split('.') # we split by points
  latitudeoutput = df.latitude[i].split('.') # we split by points

  a=df.latitude[i]
  a =re.sub('\.','',a)
  df.loc[i,"latitude"] = a[:2] + '.' + a[2:]

  b=df.longitude[i]
  b =re.sub('\.','',b)
  df.loc[i,"longitude"]= b[:2] + '.' + b[2:]

df["longitude"] = df["longitude"].astype(float)
df["latitude"] = df["latitude"].astype(float)
```

Figure 5: Transformation of the geographical data

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="KGETourism")


a=389 #lets take a sample bus station
print(df.loc[a,"name"])
(df.loc[a,"longitude"],df.loc[a,"latitude"])

location = geolocator.reverse((df.loc[i,"latitude"],df.loc[i,"longitude"]))
location.raw.get("address")
```

```
Vigo Rendena
{'road': 'Via San Lorenzo',
 'suburb': 'Bolghera',
 'village': 'Vela',
 'city': 'Trento',
 'municipality': "Territorio Val d'Adige",
 'county': 'Provincia di Trento',
 'ISO3166-2-lvl6': 'IT-TN',
 'state': 'Trentino-Alto Adige/Südtirol',
 'ISO3166-2-lvl4': 'IT-32',
 'postcode': '38122',
 'country': 'Italia',
 'country_code': 'it'}
```

Figure 6: Result of the geographical data

this notebook.

For reusability and iTelos compliance reasons, all the resources have been aligned following the CSV format standard, instead of txt or xlsx formats.

5.4 Knowledge modeling and formal resource generation

Once we have all the datasets cleaned and formatted into csv files and in order to generate formal resources first we need to associate our datasets to a schema.

While cleaning the datasets and naming the entities, we had in mind the schemas provided in schema.org, which made the process of modeling quite easy. We have limited the properties shown as information that is both present in the datasets and relatable to an existing schema. We used the following schemas:

Concept	Schema	Route	Properties
Campsites and shelters	Accomodation	Thing > Place > Accommodation	accommodationCategory, latitude, longitude, city, address, telephone
Panoramic viewpoints	TouristTrip	Thing > Intangible > Trip > TouristTrip	id, name, touristType, trip_origin, arrival_time, additional field(needs car)
Private transportation	Car	Thing > Product > Vehicle > Car	brand, model
Our personas, guides and information contacts	Person	Thing > Person	name, birthdate, gender, nationality, description, affiliation, email
Mountains	Mountain	Thing > Place > Landform > Mountain	name, latitude, longitude, city, address
Lakes	LakeBodyOfWate	Thing > Place > Landform > BodyOfWater > LakeBodyOfWater	id, name, longitude, latitude
Mountain equipment stores	Store	Thing > Organization > LocalBusiness > Store / Thing > Place > LocalBusiness > Store	id, name, longitude, latitude
Caves	Landform	Thing > Place > Landform	id, name, longitude, latitude
Public transportation: buses	BusStation	Thing > Place > CivicStructure > BusStation	id, name, latitude, longitude, address
Public transportation: trains	TrainStation	Thing > Place > CivicStructure > TrainStation	id, name, latitude, longitude, address
Gas station	GasStation	Thing > Organization > LocalBusiness > AutomotiveBusiness > GasStation / Thing > Place > LocalBusiness > AutomotiveBusiness > GasStation	id, name, longitude, latitude, address
Ski resort	SkiResort	Thing > Place > LocalBusiness > SportsActivityLocation > SkiResort	address, name

The tool used for modeling the schemas is Protégé and the outcome is a RDF-OWL schema file (see Fig. 7). Once we have the schema, we map the data layer and the knowledge layer using Karma (see Fig. 8), producing RDF files whose content is copied into a ttl file.

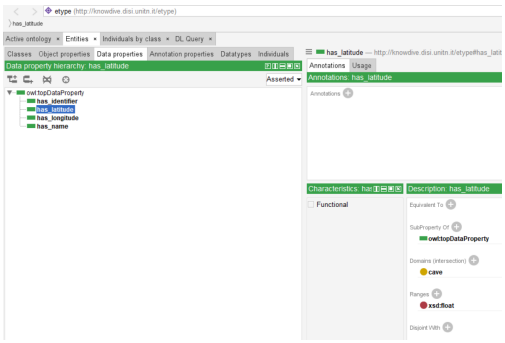


Figure 7: Protegé

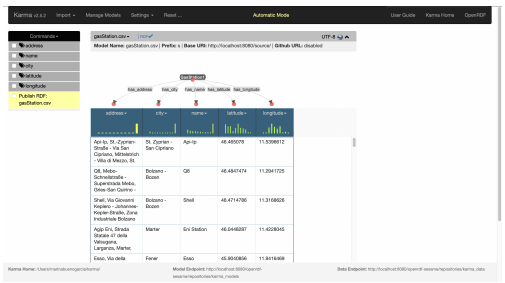


Figure 8: Karma

Throughout this phase, our PFsheet and ER mapping have suffered modifications as we have worked with the data in detail. The hikes are not comprised of natural resources (lakes,mountains,caves) and instead are an alternative element that tourists can visit. We also

realized having the opening hours of the stations was nonsensical in some cases (like bus stops in the street). We are aware that iTelos is an itherative metodology and as such, we are constantly revising and makeing sure every new step is coherent with the previous ones.

6 Language Definition

7 Knowledge Definition

8 Data Definition

9 Evaluation

10 Metadata Definition

11 Open Issues

References