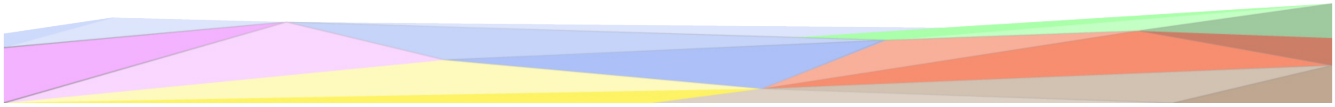
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Tourism And Waste Management in Trentino

Document Data:	Reference Persons:
November 25, 2024	Maria Amalia Pelle, Gaudenzia Genoni, Yishak Tadele Nigatu



Index:

1 Purpose Definition **1**
- 1.1 Informal Purpose . . . 1
- 1.2 Domain of Interest (DoI) . . . 1
- 1.3 Scenarios definition . . . 1
- 1.4 Personas . . . 2
- 1.5 Competency Questions . . . 3
- 1.6 Concepts identification . . . 4
- 1.7 ER model definition . . . 5
- 1.8 Overview of the first iTelos phase . . . 6

2 Information Gathering **8**
- 2.1 Dataset ”Waste Baskets” . . . 9
  - 2.1.1 Source Identification . . . 9
  - 2.1.2 Data Collection . . . 9
  - 2.1.3 Data Cleaning and Standardization . . . 10
- 2.2 Dataset ”Waste” . . . 11
  - 2.2.1 Source Identification . . . 11
  - 2.2.2 Data Collection . . . 12
  - 2.2.3 Data Cleaning and Standardization . . . 12
- 2.3 Dataset ”Municipality” . . . 13
  - 2.3.1 Source Identification . . . 13
  - 2.3.2 Data Collection . . . 13
  - 2.3.3 Data Cleaning and Merging . . . 13
- 2.4 Dataset ”Waste_production” . . . 14
  - 2.4.1 Source Identification . . . 14
  - 2.4.2 Dataset Collection . . . 14
  - 2.4.3 Dataset Cleaning and Merging . . . 14
- 2.5 Dataset ”Tourist Attraction” . . . 15
  - 2.5.1 Source Identification . . . 15
  - 2.5.2 Data Collection . . . 16
  - 2.5.3 Data Preprocessing . . . 17
  - 2.5.4 Merging Dataframes . . . 20
- 2.6 Overview of the second iTelos phase . . . 20

3 Language Definition **21**
- 3.1 Concept identification and Dataset filtering . . . 21
- 3.2 Overview of the third iTelos phase . . . 23

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 24, 2024	Maria Amalia Pelle	Document created
1.0	October 31, 2024	Amalia, Gaudenzia, and Yishak	Phase I Compilation
2.0	November 13, 2024	Amalia, Gaudenzia, and Yishak	Phase II Compilation
3.0	November 25, 2024	Amalia, Gaudenzia, and Yishak	Phase III Compilation

1 Purpose Definition

The initial phase of our Knowledge Graph project focuses on clearly defining the purpose, establishing the domain of interest, and identifying personas and scenarios. These elements inform the formulation of competency questions, which address the needs and goals of the identified personas and scenarios: the competency questions will later be answered by querying the completed knowledge graph. This phase also includes the completion of the Purpose Definition Sheet, where all relevant entities and their associated properties are outlined, and the Entity-Relationship (ER) Modeling.

1.1 Informal Purpose

The purpose of our project is to offer comprehensive data regarding waste management and its relationship with tourism in the Province of Trento. The final Knowledge Graph (KG) will serve as a valuable tool for various stakeholders, including tourists, facility owners, and waste management authorities, by providing accessible information on waste disposal locations, recycling options, and the environmental impact of tourism on local waste management systems. Additionally, the KG will support researchers and experts by enabling in-depth analysis of the interactions between tourism activities and waste generation, promoting evidence-based decision-making in administration and policy development.

1.2 Domain of Interest (DoI)

The focus of this project is on waste production data and tourist infrastructure in the Province of Trento (Trento coordinates are 46°04′00″N, 11°07′00″E). The waste production data, which pertains to individual cities, covers a yearly timeline from 2010 to 2022.

1.3 Scenarios definition

S1 General waste disposal: It's a sunny winter morning at Monte Bondone, a popular skiing destination in Trentino. As visitors enjoy their skiing activities, they generate waste such as snack wrappers and drink containers, which need to be properly disposed of to maintain the area's natural beauty.

S2 Following policies guidance: Facility owners in Trentino must enforce regulations related to waste disposal to comply with local environmental laws. To achieve this, they should regularly review the waste management guidelines provided by the municipality and stay updated on any changes in legislation.

S3 Analysis on tourism waste impact: Trentino requires a thorough analysis of the impact of tourism on waste management to propose new regulations. In response to growing concerns about the increasing waste generated by tourist activities, a comprehensive study is needed to examine current waste management practices in popular tourist areas.

S4 Time Series Analysis: During the COVID-19 pandemic, Trentino faced an unprecedented tourist season with marked fluctuations in visitor numbers due to shifting travel restrictions and health protocols. These changes might have impacted waste production at popular tourist destinations.

S5 Special waste disposal: Sport tourism in Trentino attracts various travelers who bring along sporting equipment such as bicycles, climbing harnesses, and skis. However, accidents can happen even on vacation, leading to damaged gear. Unfortunately, this type of waste cannot always be disposed of easily and often requires transportation to specialized facilities for proper handling.

S6 Waste management practices:The waste management provider aims to conduct a detailed comparative analysis of waste production levels across various municipalities within the region. The primary objective is to identify and understand the most effective waste management practices employed by cities with consistently lower waste outputs.

1.4 Personas

P1 Maximilian is a 35-year-old Austrian citizen living in Innsbruck. An outdoor enthusiast, he enjoys winter sports, particularly skiing, and frequently travels to Trentino during the winter season. Passionate about environmental sustainability, Maximilian actively seeks ways to reduce his ecological footprint wherever he goes.

P2 Luciana is a 45-year-old facility owner in Trento, operating a charming bed-and-breakfast that attracts tourists year-round. She actively seeks information on current waste management regulations and best practices to ensure her establishment complies with legal requirements and promotes eco-friendly tourism.

P3 Diego is a 52-year-old policy maker in Trentino Province. Among his responsibilities, he reviews existing waste management policies to evaluate their effectiveness, identifies gaps, and proposes improvements. Diego collaborates with various stakeholders, including facility owners and environmental organizations, to ensure that the regulations he advocates are practical and beneficial for both the community and the environment.

P4 Chiara is a 28-year-old researcher at FBK, currently participating in a collaborative research project. The goal of her study is to assess how tourism in various cities of Trentino con-

tributes to waste production. She is particularly interested in comparing fluctuations in waste generation during the COVID-19 pandemic. Chiara is passionate about sustainable tourism and aims to identify strategies for reducing waste in the region.

P5 Filippo is a 24-year-old sports enthusiast living in Verona. He thrives on outdoor activities and finds it hard to relax, even on vacation. For this reason, he has chosen Caldonazzo for a weekend getaway with his girlfriend, Giulia. Eager for adventure, he has brought along his own canoe and is excited to make the most of their time together. However, upon arriving at their destination, he realizes that he has broken his paddle during the journey and is unsure of where to dispose of it properly.

P6 Giovanni is a 41-year-old waste management analyst working for Dolomiti Ambiente. With a background in environmental science, he is passionate about finding innovative solutions to improve waste management practices. Giovanni currently aims to compare waste production across various cities in Trentino to identify successful strategies employed by those with lower waste outputs.

1.5 Competency Questions

CQ-1 (P3-S3, P4-S4, P6-S6): As a policy maker/researcher evaluating waste management, what is the estimated total waste generated in specific tourist areas during peak seasons, and how is this waste categorized (e.g., plastic, organic)?

CQ-2 (P1-S1, P5-S5): While enjoying winter sports, where can I find the nearest recycling bins, and what types of waste can I dispose of in these bins?

CQ-3 (P2-S2): As a facility owner, what waste disposal options are available to me in Trentino, and which types of waste do these facilities accept?

CQ-4 (P2-S2): In order to maintain compliance with local regulations, what are the waste disposal requirements specific to my bed-and-breakfast? What are the acceptable disposal methods for different waste types?

CQ-5 (P3-S3, P6-S6): What types of waste are generated by various tourism activities in Trentino, and how do these types impact overall waste management practices?

CQ-6 (P3-S3, P5-S5, P6-S6): What special waste disposal facilities are available in Trentino, including their locations, capacities, and the types of waste they manage?

CQ-7 (P3-S3, P4-S4, P6-S6): Which areas in Trentino generate the most waste from tourism activities?

CQ-8 (P3-S3, P4-S4, P6-S6): During the COVID-19 pandemic, how did visitor fluctuations impact waste production at popular tourist destinations in Trentino? Can we analyze the yearly waste production trends for specific municipalities using the available data?

1.6 Concepts identification

Table 1 below lists the entities - along with their respective properties - related to the competency questions, taking into consideration both the purpose (knowledge layer) and the available data sources (data layer).

Table 1: Scenarios, Personas, and related Competency questions

Scenarios	Personas	Competency Questions	Entities	Properties	Focus
S3, S4, S6	P3, P4, P6	CQ-1, CQ-7, CQ-8	Municipality	Municipality_ID (PK) Name Location_ID Coordinates Population_size	Common
-	P1, P2, P3, P4, P5, P6	-	Person	Person_ID (PK) Role Name Surname Date_of_birth Country Is_a_Tourist Location_ID	Common
S1, S2, S3, S4, S5, S6	P1, P2, P3, P4, P5, P6	CQ-1, CQ-2, CQ-3, CQ-4, CQ-5, CQ-6, CQ-7, CQ-8	Waste Type	Waste_Type_ID (PK) Category Recyclability Disposal_Method	Core
S1, S4	P1, P3, P4, P5, P6	CQ-1, CQ-2, CQ-5, CQ-7	Location	Location_ID (PK) Name Category Latitude Longitude Municipality_ID (FK)	Core

S2	P2	CQ-4	Tourist Facility	Facility_ID (PK) Name Type Location_ID (FK)	Core
S1, S3	P1, P3	CQ-1, CQ-5	Tourist Activity	Activity_ID (PK) Type Seasonality Location_ID (FK)	Core
S1, S2, S3, S5	P1, P2, P3, P5	CQ-2, CQ-3, CQ-6	Waste Management Facility	Waste_Facility_ID (PK) Name Location_ID (FK)	Contextual
S2, S3	P2, P3	CQ-2, CQ-4	Waste Regulations	Policy_ID (PK) Name Effective_Date Municipality_ID (FK)	Contextual
S3, S4, S6	P3, P4, P6	CQ-1, CQ-5, CQ-7, CQ-8	Waste Production	Waste_Production_ID Quantity Waste_Type_ID Municipality_ID	Contextual

1.7 ER model definition

The ER model in Figure 1 (below) was designed using the previously considered concepts of entities and properties. It represents the initial graphical version of the final structure of the Knowledge Graph.

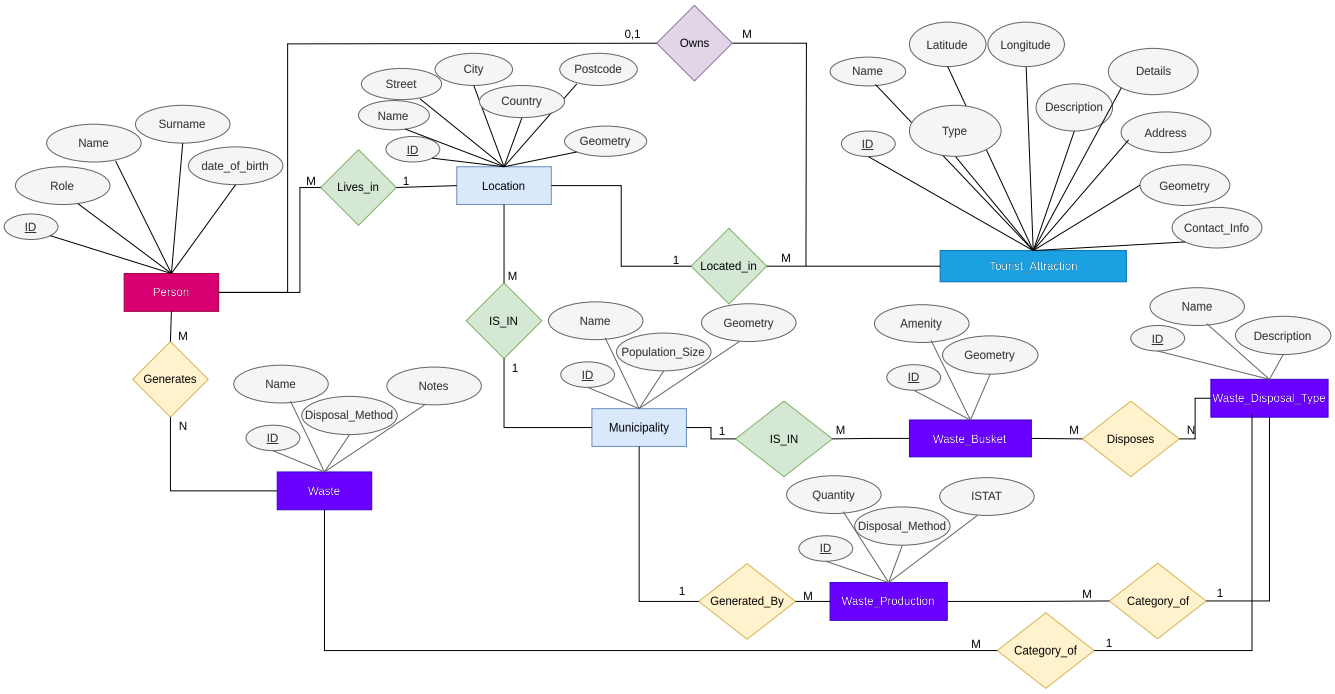


Figure 1: ER Diagram

1.8 Overview of the first iTelos phase

In the initial phase of this project, our focus was to clearly define the project's purpose and identify key personas and scenarios likely to interact with the Knowledge Graph (KG). These personas include tourists visiting Trentino for leisure, analysts from the Province of Trento, and researchers from external institutions (such as FBK) who may use the KG to access information related to waste management and its connection to tourism. Guided by the competency questions, we then identified common, core, and contextual entities - along with their relevant properties - to structure the KG. This preliminary framework remains flexible, allowing us to adjust it based on data availability and ensuring the project's feasibility throughout later stages.
In this phase, we adopted the so-called Middle-out approach, considering both the knowledge layer and the data layer to facilitate easy data adaptation and knowledge modeling.
As a preliminary step, we collaboratively reviewed the data resources provided by our instructor:

- A comprehensive Knowledge Graph on tourism data
- The ISPRA website, which provided aggregated data on waste production per municipality and the locations of facilities for handling specific types of waste (e.g., special waste). Waste production data spans from 2010 to 2022, adding a valuable temporal dimension which helped to frame our Domain of Interest.

After analyzing these sources, we brainstormed additional datasets that could strengthen the KG's depth and utility. Key suggestions included

- Data on local waste disposal policies: Useful for managers of tourist facilities who require up-to-date disposal guidelines (Proposed by Gaudenzia).
- Data on public waste bin locations: Important for tourists who may need guidance on waste disposal options in unfamiliar areas (Proposed by Amalia).
- Real-time, geolocated data on tourist presence: Valuable for facility managers to optimize waste bin placement and collection logistics (Proposed by Yishak).

However, acquiring this additional data may present challenges. No centralized collection of waste disposal policies exists, so manual entry may be necessary. For public waste bin data, official sources are unavailable, so we considered relying on crowdsourced platforms like Overpass Turbo.

After gaining a general understanding of the available data sources, we shifted our focus to envisioning the final product and identifying potential personas and scenarios, as well as competency questions. For personas development, we leveraged our local experience and knowledge of the area, including our understanding of the types of tourists who frequent Trentino—primarily outdoor sports enthusiasts. We considered their possible needs and the types of waste-related information that would interest them, concluding that their primary concern would likely be access to disposal facilities. This approach also applied to tourist facility owners, who share similar concerns around waste disposal options for their guests. On the other hand, data such as annual waste production by municipality and ISPRA's other aggregated metrics, which are less relevant to the general public, are highly useful for policymakers, waste management authorities, and city administrators interested in broader environmental management. Lastly, we recognized that researchers could also benefit significantly from the KG, particularly given the temporal data dimensions, which support more in-depth analysis. While developing the Purpose Formalization Sheet and the Entity-Relationship Model, we took into consideration several reference context schemas, and particularly Schema.org.

Throughout this phase, each team member contributed to defining the project's purpose and domain of interest, as well as discussing the available data sources. Amalia and Gaudenzia identified personas, scenarios, and competency questions, creating the concept identification table. Gaudenzia and Yishak developed the Entity-Relationship (ER) schema using the IDEF1X notation, as recommended. Yishak managed all layout and formatting in LaTeX, while Amalia updated the repository on GitHub.

2 Information Gathering

In this second phase of the project, we identified and gathered all the resources needed to build the graph. These resources include knowledge ontologies, language vocabularies, and data value datasets.

- Regarding the **knowledge ontologies**, we consulted schema.org for the ETypes in our ER model that directly correspond to it. However, since schema.org is a high-quality, general-domain schema, some of our ETypes—particularly those related to waste types and waste disposal methods, which are highly domain-specific and mainly defined by available datasets—were not represented: a future ambitious project could involve surveying all relevant resources in the Trentino region, with the goal of standardizing, harmonizing, and making them reusable and interoperable. Table 2 pairs our ETypes with their corresponding schema.org categories.

Table 2: Four ETypes with corresponding schema.org categories

Concept	Schema	Route	Properties
Person	Person	Thing > Person	identifier, givenName, familyName, birthDate
Location	Place	Thing > Place	identifier, name, geo
Municipality	City	Thing > Place > AdministrativeArea > City	identifier, name, geo
Tourist_Attraction	TouristAttraction	Thing > Place > TouristAttraction	identifier, name, latitude, longitude, description, address, telephone

- Regarding the **language vocabularies**, we primarily referred to the Universal Knowledge Core (UKC) and the OpenStreetMap glossary, which will be particularly useful in Phase Three of the project. Both of them are high quality sources.
- With respect to the **data value resources**, we identified seven datasets:
  1. Waste
  2. Waste Baskets
  3. Waste Disposal Types

4. Waste Baskets Disposal Types
5. Municipalities
6. Waste Production
7. Tourist Attraction

For each dataset, we dedicated three subsections to source identification, dataset collection, and cleaning/standardization procedures, as detailed in the following paragraphs.

2.1 Dataset ”Waste Baskets”

This first dataset provides geospatial information on the distribution of waste baskets, organic bins, and recycling points for various materials across the Province of Trento.

2.1.1 Source Identification

In conducting research on waste basket distribution within the Province of Trento, we found no pre-existing datasets suitable for reuse. Consequently, data acquisition was undertaken using OpenStreetMap (OSM), a widely recognized, community-contributed mapping platform. Given the open-source nature of OSM, it provides freely accessible and editable geographical data, which serves as a valuable resource for mapping facilities and amenities. The community-driven aspect of OSM allows for dynamic updates but also introduces potential limitations in terms of data completeness and accuracy.

2.1.2 Data Collection

The data retrieval was facilitated by Overpass Turbo, a tool that allows for specific and customizable querying of the OSM database. The Overpass Turbo query used to collect data is provided below.

```
[out:json];
area["name"="Provincia di Trento"]->.a;
(
  node["amenity"="waste_basket"](area.a);
  node["amenity"="recycling"](area.a);
  node["amenity"="waste_basket"]["waste:organic"="yes"](area.a);
  node["amenity"="recycling"]["recycling:glass"="yes"](area.a);
  node["amenity"="recycling"]["recycling:plastic"="yes"](area.a);
  node["amenity"="recycling"]["recycling:paper"="yes"](area.a);
  node["amenity"="waste_disposal"](area.a);
  node["amenity"="waste_collection_point"](area.a);
);
out body;
>;
out skel qt;
```

This query yielded data on a variety of waste management facilities in the Province of Trento, including general-purpose waste baskets, bins for organic waste, and recycling points for specific materials such as glass, plastic, and paper. These data were collected in a GeoJSON file containing 3,883 points of interest and as many as 84 attributes.
Each point's precise location within the Province of Trento can be visualized on an accompanying map (please,allow a few seconds for it to load).
Given the community-driven nature of the OSM platform, we acknowledge the potential limitations regarding data completeness and accuracy; it is plausible that some areas or types of bins may be underrepresented or inconsistently documented. Nevertheless, in the absence of a more comprehensive source, OSM serves as a viable data source that provides sufficient granularity and geographical coverage for our research objectives.

2.1.3 Data Cleaning and Standardization

In this case, data cleaning involved the removal of 45 columns containing mostly NaN values or information not relevant to our project, while retaining the 39 most meaningful columns. Furthermore, the GeoJSON file was transformed into a CSV format: although the dataset is generally sparse, with many rows missing values for most columns, the available data is very useful.
The most important attributes in the dataset are:

- **id**: a unique identifier for each data point.
- **amenity**: it describes the type of facility or service (e.g., waste basket, recycling point, or waste collection point).
- **recycling types**: it indicates the types of materials accepted for recycling, such as glass, plastics, metals, paper, organic waste, and special items.
- **geometry**: it provides spatial coordinates (latitude and longitude) essential for mapping and spatial analysis.
- **municipality**: it lists the municipality in which the waste basket is located (this attribute was derived from a spatial join with the "Municipalities" dataset, presented below.).

The recycling columns were initially filled with values such as 'yes', 'no', and NaN. To standardize the data, we converted 'yes' values to True, 'no' values to False, and retained NaN values as False. Additionally, we observed that a single waste basket could be associated with multiple recycling types in a so-called "many-to-many" relationship. As a result, we decided to create a new dataset, namely **"Waste Disposal Types"**, containing 10 waste disposal categories, as follows:

1. Organic: Includes food waste, organic waste, garden waste, and green waste.
2. Paper/Cardboard: Includes paper, cardboard, paper packaging, books, magazines, and newspapers.
3. Glass: Includes general glass waste and glass bottles.
4. Metal: Includes aluminium, cans, scrap metal, and other metal items.
5. Plastic: Includes PET, plastic, plastic bottles, plastic packaging, and beverage cartons.
6. Textiles: Includes clothes and shoes.
7. Electronic Waste: Includes batteries, electrical appliances, electrical items, small appliances, and lamps.
8. Wood: Includes wood items and garden pots.
9. Construction Waste: Includes construction-related waste.
10. Miscellaneous: Includes items like tetrapak and other non-categorized materials.

Following this categorization, which proved to be very important also in the subsequent analysis of waste types and waste production, we mapped the individual recycling types of each waste basket to these higher-level disposal categories, resulting in the creation of a new dataset, **“Waste Baskets Disposal Type”**: the dataset lists baskets alongside the types of waste they accept, with separate rows for each applicable category.

2.2 Dataset ”Waste”

This dataset provides structured information on waste types and disposal methods in the Province of Trento, supporting analysis of local waste management practices.

2.2.1 Source Identification

To obtain data regarding waste production, management, and disposal in the Province of Trento, we referred to Dolomiti Ambiente, a subsidiary of the Dolomiti Energia Group responsible for environmental hygiene services and waste collection. Although Dolomiti Ambiente operates exclusively in the municipalities of Trento and Rovereto, it was selected as a representative case study, under the assumption that waste collection practices in other municipalities within the province are managed in a similar and comparable manner. Future research could expand this study to include an analysis of waste collection regulations in municipalities beyond Trento and Rovereto.

2.2.2 Data Collection

Although Dolomiti Ambiente's website does not provide a dedicated dataset on waste types and their proper disposal methods, the company offers a downloadable PDF document titled "Riciclabolario", which serves as a guide for navigating waste sorting for domestic waste. This document became the primary source for the data extraction process. From this source, an initial CSV file was created, listing all waste types alongside their corresponding disposal categories.

2.2.3 Data Cleaning and Standardization

The initial CSV file was subjected to extensive cleaning to extract relevant data into appropriate columns and standardize the structure. The cleaning process began by splitting the *waste* column into two: *waste* and *disposal_method_1*, separating the main waste category from disposal method information following a hyphen. If *disposal_method_1* contained multiple disposal methods separated by a semicolon, the data was split further into a new column, *disposal_method_2*, to ensure each disposal method was clearly identified. Parenthetical information, such as special handling instructions or waste characteristics, was extracted and moved into separate "notes" columns (*waste_notes*, *notes_1*, *notes_2*), ensuring that relevant details were retained without cluttering the primary columns. Finally, an index column was added to provide a sequential identifier for each row, facilitating easier referencing and further indexing operations as needed.
After performing the above cleaning steps, the final dataset consists of the following columns:

- **index**: A sequential identifier for each row.
- **waste**: The main waste category.
- **waste_notes**: Additional information extracted from the *waste column*, if any.
- **disposal_method_1**: The primary disposal type.
- **notes_1**: Additional notes extracted from the *disposal_method_1* column, if any.
- **disposal_method_2**: The secondary disposal type.
- **notes_2**: Additional notes extracted from the *disposal_method_2* column, if any.
- **category**: The ID of the waste category to which the disposal type belongs, based on the dataset "Waste Disposal Types".

2.3 Dataset ”Municipality”

This dataset contains the geographical boundaries of municipalities in the Province of Trento, along with the corresponding ISTAT code and population data. It can be useful for determining which municipality a user is located in.

2.3.1 Source Identification

The ”Municipality” dataset was compiled using two primary sources: population data provided by ISTAT (Italian National Institute of Statistics) and geographic coordinates obtained from OpenStreetMap (OSM). ISTAT is a reliable source for demographic data in Italy, regularly publishing population statistics, including municipal-level data. The Overpass Turbo API was used to retrieve the spatial boundaries of each municipality in the specified region.

2.3.2 Data Collection

Population data for each municipality were collected directly from the ISTAT website in CSV format. These data represent the official population counts recorded on January 1, 2024. The spatial data from OSM, specifically the municipal boundary coordinates, were retrieved using queries to the Overpass Turbo platform, similar to how the waste basket data was collected. The query was designed to capture the administrative boundaries at the municipality level:

```
[out];
area["name"="Provincia di Trento"]
["boundary"="administrative"] ["admin_level"="6"];
rel(area)["boundary"="administrative"]["admin_level"="8"];
out body;
out skel qt;
```

2.3.3 Data Cleaning and Merging

Once collected, both datasets underwent preprocessing to ensure uniformity and compatibility for merging. The ISTAT dataset had all column names translated and adjusted to ensure consistency and compatibility with the other tables. Most columns were dropped, leaving only the ISTAT code and total population. Similarly, the geographic data from Overpass Turbo were structured into a GeoJSON format to facilitate easy data merging and visualization. To create a comprehensive table that includes both population and geographic data, the ISTAT code was used as a key to merge the two datasets. This approach was preferred over using city names, as it ensures consistency—especially given that some municipalities have undergone name changes in the last five years. The GeoJSON file was enhanced by appending a ”Total

Population" tag to each municipality entry, allowing a single dataset to reflect both spatial and demographic information for each municipality.

The final dataset includes the following attributes for each municipality:

- **Name**: The name of the municipality
- **ref: ISTAT**: The ISTAT code of the municipality
- **population**: The population count as of January 1, 2024
- **geometry**: The coordinates of the polygon defining the municipality's borders

As mentioned in the previous section, we acknowledge potential limitations regarding the completeness and accuracy of the data from the OSM platform.

2.4 Dataset "Waste_production"

This dataset contains a table with the annual waste production in tons for all the cities in the Province of Trento, covering the years from 2014 to 2022. Some data are provided in an aggregated form.

2.4.1 Source Identification

The data was sourced from the Italian Institute for Environmental Protection and Research (ISPRA), which provides comprehensive information on waste production across Italy. The datasets are publicly available on the ISPRA website.

2.4.2 Dataset Collection

These data were collected and compiled by ISPRA, offering insights into waste management trends over the specified period. The dataset were collected from their website.

2.4.3 Dataset Cleaning and Merging

To ensure consistency, the data from all years were combined into a single dataset, with each record labeled by the corresponding year. All column names and attributes were translated, and the names were chosen to be compatible with the other tables to ensure uniformity across the dataset. During this process, unnecessary columns with missing or irrelevant data were removed to improve the quality and relevance of the dataset. Additionally, a new "category" column was added, where a numeric value describes the specific waste type, using the "waste_disposal_type" table as a reference. The dataset was then melted to transform it into a

long-format table, making it easier to analyze over time. The cleaned dataset was saved as a single CSV file presenting the following attributes:

- **Municipality**: The name of the municipality
- **ref: ISTAT**: The ISTAT code of the municipality
- **category**: waste id from the waste_disposal_type table
- **year**: Year in which data was collected

2.5 Dataset ”Tourist Attraction”

tourist attraction encompass a diverse range of categories, including natural locations, cultural landmarks, and facilities tailored to enhance visitor experiences. In this project, we considered a wide array of tourist attractions, including:

- **Natural Attractions:** Such as protected areas, lakes, rivers, beaches, peaks, viewpoints, caves, waterfalls, and springs.
- **Accommodation Facilities:** Including hotels, holiday apartments, and houses.
- **Dining and Hospitality:** Encompassing food and drink establishments such as Restaurants, Cafes and Bars.
- **Cultural Sites:** Including museums, historic sites, and other cultural attractions.
- **Entertainment and Recreation:** Such as amusement parks, recreational facilities, skiing and winter sports, and other adventure locations.

This comprehensive categorization reflects the broad appeal and variety of attractions available to tourists, from immersive natural environments to comfortable accommodations and lively entertainment options.

2.5.1 Source Identification

The data for this project was sourced from OpenStreetMap via the Overpass Turbo API. We executed multiple queries tailored to each attraction type to gather comprehensive information. The resulting data was retrieved in GeoJSON format and subsequently processed to align with our specific project requirements.

2.5.2 Data Collection

The following is a list of tourist attractions that were taken into consideration in this project:

- Protected Areas
- Lakes and Rivers
- Beaches
- Peaks and Viewpoints
- Caves
- Hotel and Accommodations
- Holiday Appartments and Houses
- Food and Drink Establishments
- Cultural Attractions
- Amusement and Recreational Facilities
- Skiing and Winter Sport Facilities
- Waterfall and Springs

The following query is executed in Overpass API but. to make the data manageable, the queries are run separately and the data is downloaded separately. For instance for *Hotels and Accommodation* the following query is executed:

```
[out:json];
area["name"="Provincia di Trento"]->.searchArea;
(
  // Hotels and Accommodations
  node["tourism"="hotel"](area.searchArea);
  node["tourism"="guest_house"](area.searchArea);
  node["tourism"="hostel"](area.searchArea);
  node["tourism"="camp_site"](area.searchArea);
  node["tourism"="caravan_site"](area.searchArea);
  node["tourism"="chalet"](area.searchArea);
  node["tourism"="alpine_hut"](area.searchArea);
  node["building"="hotel"](area.searchArea);
);
out geom;
```

While for *Food and Drink Establishment* the following query is executed and the result is saved in a separate geoJSON file. The total Overpass query is found in the project's GitHub repository.

```
[out:json];
area["name"="Provincia di Trento"]->.searchArea;
(
  // Food and Drink Establishments
  node["amenity"="restaurant"](area.searchArea);
  node["amenity"="cafe"](area.searchArea);
  node["amenity"="bar"](area.searchArea);
  node["amenity"="pub"](area.searchArea);
  node["amenity"="fast_food"](area.searchArea);
);
out geom;
```

2.5.3 Data Preprocessing

The GeoJSON data provides extensive details about each type of attraction, resulting in a sparse dataset with information that varies significantly across attraction types. For instance, food and drink establishments are typically privately owned and include contact details such as phone numbers and emails. In contrast, natural attractions like lakes and rivers lack ownership information, as they are not associated with individual proprietors. Therefore, we carefully selected specific features for each type of attraction to best represent them, ensuring the dataset effectively serves the purpose of our project. Table 3 provides comprehensive information on each processed attraction type.

Table 3: Summary of Generated Data Files and Their Properties

Generated File Name	Columns	Unique Values	NaN Values	Table Size
caves.csv	ID	489	0	489 x 4
	Name	242		
	Latitude	489		
	Longitude	489		
artworks.csv	ID	419	0	419 x 8
	Name	210	208	
	Latitude	419	0	
	Longitude	417	0	
	Artist Name	98	292	
	Artwork Type	11	0	
	Description	14	90	
	Website	2	-	

Generated File Name	Columns	Unique Values	NaN Values	Table Size
memorials.csv	ID	504	0	504 x 7
	Name	222	244	
	Latitude	504	0	
	Longitude	504	0	
	Memorial Type	14	258	
	Historic Type	3	0	
	Description	80	425	
gallery_and_museum.csv	ID	71	0	71 x 10
	Name	69	2	
	Latitude	71	0	
	Longitude	71	0	
	Website	11	61	
	Type	2	0	
	Street	26	46	
	City	18	47	
	Postcode	16	47	
	House Number	21	47	
food_and_drink_establishments.csv	ID	2293	0	2293 x 13
	Latitude	2293	0	
	Longitude	2292	0	
	Name	1942	192	
	Cuisine	139	1598	
	Operator	280	2007	
	Street	645	1185	
	City	163	1314	
	Postcode	66	1298	
	House Number	199	1217	
	Website	334	0	
	Phone	619	0	
	Email	204	0	
holiday_apartments_and_houses.csv	ID	244	0	244 x 12
	Latitude	244	0	
	Longitude	244	0	
	Name	236	2	
	Tourism Type	1	45	
	Street	128	37	
	City	33	40	
	Postcode	21	51	
	House Number	80	66	
	Website	79	22	
	Phone	89	49	
	Email	82	44	

Generated File Name	Columns	Unique Values	NaN Values	Table Size
hotels_and_accommodations.csv	ID	759	0	759 x 15
	Latitude	759	0	
	Longitude	759	0	
	Name	688	28	
	Tourism Type	7	1	
	Street	303	356	
	City	114	388	
	Postcode	56	379	
	House Number	111	365	
	Cuisine	8	739	
	Description	6	753	
	Operator	37	721	
	Website	365	0	
	Phone	326	0	
	Email	264	0	
lakes.csv	ID	38	0	38 x 4
	Central_Latitude	31		
	Central_Longitude	38		
	Name	38		
rivers.csv	ID	138	0	138 x 4
	Central_Latitude	137		
	Central_Longitude	138		
	Name	25		
peaks_and_viewpoints.csv	ID	2542	0	2542 x 9
	Latitude	2540	0	
	Longitude	2539	0	
	Name	1676	799	
	Description	22	2516	
	Historic	4	2533	
	Amenity	3	2524	
	Height	2	2539	
	Website	3	2052	
protected_areas.csv	ID	3	0	3 x 8
	Latitude	3		
	Longitude	3		
	Name	3		
	Source	3		
	Website	3		
	Protection Title	2		
	Leisure	1		

Generated File Name	Columns	Unique Values	NaN Values	Table Size
skiing_and_winter_sports.csv	ID	1292	0	1292 x 6
	Latitude	1292	0	
	Longitude	1291	0	
	Name	461	644	
	Sport	1	1191	
	Description	1	1291	
waterfall_and_spring.csv	ID	295	0	295 x 6
	Latitude	295	0	
	Longitude	295	0	
	Name	98	196	
	Amenity	3	264	
	Description	5	290	

2.5.4 Merging Dataframes

After generating the data in CSV format for each type of attraction, we refined our ER Diagram to highlight the key features aligned with our CQs. We then consolidated this information into the Tourist Attractions table, summarizing the essential attributes as represented by the column names listed in Table 4.

Table 4: Summary of The final Tourist Attraction Table

Column name	Description
ID	Specific Identifier taken from the Overpass API
Name	Name
Type	Tourist Attraction Type. example: hotel, lake, skiing
Latitude and Longitude	Central Coordinates of the place
Description	General Description
Details	Relevant information depending on the type of Tourist Attraction. For instance, a Cafe might have cuisine information, whereas Artwork may consist of the artist's name.
Address	City, Street, Postcode and House number information
Contact	Phone, Email or Website information
Geometry	Location information. A Node, Polygon or way, according to Open Street Map Schema.

2.6 Overview of the second iTelos phase

In the second phase of the project, we began by discussing together the data we needed and the resources we could rely on. The decision regarding which datasets to use was not straightforward and required careful reflection on the purpose outlined in the first phase: for instance, we had to determine whether to create two distinct datasets for tourist facilities and tourist attractions or to combine both domains into a single dataset. We later decided to merge them under tourist attraction based on our purpose definition. Once this decision was made collectively,

each of us took responsibility for individual tasks related to data collection and cleaning. Specifically, Amalia worked on the ”Municipalities” and ”Waste_Production” datasets, Gaudenzia on the ”Waste” and ”Waste Basket” datasets, and Yishak on the ”Tourist Attraction” and ”Location” datasets.
This division of labour guided the structure of the report, with each of us writing the sections corresponding to our respective datasets. We focused on detailing the data source, collection methods, and data cleaning process for each dataset.
One of the main challenges we faced during this phase was finding data related to waste management in the Province of Trento. Waste management regulations vary across municipalities, although the differences are relatively minor. To address this issue, we decided to refer to the Dolomiti Ambiente website, which provides guidelines for the municipalities of Trento and Rovereto, assuming that other municipalities follow similar regulations to those of the provincial capital.
For spatial data, we primarily relied on OpenStreetMap, utilizing Overpass Turbo to execute custom queries. The platform, being voluntary and collaborative, has both advantages and limitations. On the positive side, the available data is extensive, accurate, and reliable, making it a valuable resource for our study. However, due to its collaborative nature, the data is necessarily partial, with potential gaps or inconsistencies in some areas.
If, as we continue with the project, we find that additional data is required, we may consider reusing datasets from previous course projects, particularly those related to tourism in the Province of Trento.
Finally, this phase required us to revise the ER model, as the initial version needed adjustments. It is possible that further refinements will be necessary as the project progresses.
All raw data, pre-processing and querying scripts, and processed datasets are hosted in our GitHub repository and can be found under the ’Phase Two’ section.

3 Language Definition

3.1 Concept identification and Dataset filtering

In this section, the first activity was dedicated to the **concept identification**. We aimed to formalize the concepts and terminologies used in the KG by leveraging two primary data sources for lexical and semantic definitions:

- The Universal Knowledge Core (UKC), a repository of domain-independent concepts.
- OpenStreetMap (OSM), used for domain-specific terminologies, particularly for datasets obtained through the Overpass Turbo tool.

We carefully selected concepts to represent the following elements in the final Language Resource table:

- **Entity Types (ETypes)**, highlighted in red,
- **Data Properties**, highlighted in orange,
- **Object Properties**, highlighted in blue,
- **Data values** (only if used as enumeration categories), highlighted in green.

As input, we primarily relied on the ER model developed in the first phase, which was revised and refined for this purpose based on the real datasets available, ensuring alignment with the actual data.
As mentioned above, the UKC was particularly useful for defining widely-used general terms, such as *name*, *category*, and *description*, while OSM proved effective for domain-specific terminology. For concepts from the UKC, we used the UKC number as the identifier. In the case of OSM, we primarily referenced the link to the relevant web page. However, when a specific web page dedicated to a single concept was unavailable, we used the link to the general web page, preceded by the name of the specific item (for example: aluminium-https://wiki.openstreetmap.org/wiki/Tag:amenity%3Drecycling).
We aimed to maintain consistency across our language resources; however, a few terms had to be newly defined due to their absence in existing vocabularies. Most notably, we defined *id* as ”a unique identifier assigned to each record in a table, ensuring that each entry can be distinctly referenced and retrieved.”, since the UKC vocabulary associated the term’s first concept with a physical object (”a card or badge used to identify the bearer”) and the second concept specifically with psychoanalytic terminology. Another interesting case is the new definition of *geometry* in the GIS context, which was absent in the UKC and was subsequently added by us, as follows: ”In GIS terminology, geometry refers to the spatial representation of geographic features, defining their shape, location, and size using points, lines, and polygons within a coordinate system”. Furthermore, we had to define specific waste categories (such as *organic*, *paper*, *glass*, *metal*, *plastic*) and specific waste type (such as *Selective Collection (t)* and *Street Cleaning Waste for Recovery (t)*), as they belong to a highly contextual vocabulary and were either absent from the UKC or did not align with the meanings relevant to our domain.
The primary focus of the second activity, **dataset filtering**, was to ensure that all elements included in the final KG were clearly defined by formalized concepts. During this process, we discovered that certain terms in the ”Waste Baskets” dataset — specifically, *recycling:bottles*, *recycling:drugs*, *recycling:electrical_items*, *recycling:garden_pots*, *recycling:lamps*, *recycling:organic_waste*, *recycling:small_appliances*, and *recycling:tetrapak* — were no longer maintained by the OpenStreetMap community. As a result, we decided to filter

them out of the final KG.
The final result of this phase is the **Language Resource Table**, which has been uploaded to our GitHub repository under the Phase 3 section.

3.2 Overview of the third iTelos phase

In this phase, each team member focused on defining the concepts primarily associated with the databases curated during the second phase. However, we worked together to resolve any ambiguities and reach a consensus on more complex definitions. This phase also required us to revisit and refine the ER model, a task led by Yishak, in order to clearly identify the key terms to prioritize, based on the data resources and databases gathered in the second phase. Finally, we revised the second phase report, specifically by adding a section on knowledge and language resources, in response to feedback from our tutor.