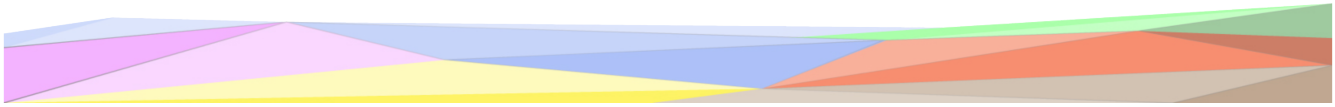
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Tourism And Waste Management in Trentino

Document Data:	Reference Persons:
November 13, 2024	Maria Amalia Pelle, Gaudenzia Genoni, Yishak Tadele Nigatu



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 30, 202	Maria Amalia Pelle	Document created

1 Information Gathering

In this second phase of the project, we searched for and gathered the data resources needed for building the graph. The datasets identified, in particular, are:

1. Waste Types
2. Waste Baskets
3. Municipalities
4. Waste Production
5. Tourist Attraction Places

For each dataset, we dedicated three subsections to source identification, dataset collection, and cleaning/standardization procedures. Furthermore, all datasets and scripts are available on the project's GitHub page.

1.1 Dataset "Waste Types"

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

1.1.1 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.

1.1.2 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.

1.1.3 Data Preprocessing

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_type* column into two: *waste_type* and *disposal_type_1*, separating the main waste category from disposal method information following a hyphen. If *disposal_type_1* contained multiple disposal methods separated by a semicolon, the data was split further into a new column, *disposal_type_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_type**: The main waste category.
- **waste_notes**: Additional information extracted from the *waste_type column*, if any.
- **disposal_type_1**: The primary disposal type.
- **notes_1**: Additional notes extracted from the *disposal_type_1* column, if any.
- **disposal_type_2**: The secondary disposal type.
- **notes_2**: Additional notes extracted from the *disposal_type_2* column, if any.

1.2 Dataset "Waste Baskets"

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

1.2.1 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.

1.2.2 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).

1.2.3 Data Preprocessing

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.

As a final note, 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.

1.3 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.

1.3.1 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.

1.3.2 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;
```

1.3.3 Data Cleaning and Merging

Once collected, both datasets underwent preprocessing to ensure uniformity and compatibility for merging. The ISTAT dataset was formatted with consistent column headers, and 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
- **postal_code**: The postal code (CAP) of the municipality
- **ref**: The ISTAT code of the municipality
- **Totale**: 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.

1.4 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.

1.4.1 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.

1.4.2 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.

1.4.3 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. During this process, unnecessary columns with missing or irrelevant data were removed to improve the quality and relevance of the dataset. The cleaned dataset was then saved as a single CSV file presenting the following attributes:

- **Comune**: The name of the municipality
- **Istat**: The ISTAT code of the municipality
- **Frazione organica (t)**: Organic waste produced in tons
- **Ing. misti a recupero(t)**: Organic waste produced in tons
- **Carta e cartone (t)**: Paper and cardboard waste produced in tons
- **Altro RD (t)**: General waste produced in tons
- **Legno (t)**: Wooden waste produced in tons
- **Metallo (t)**: Metal waste produced in tons
- **Plastica (t)**: Plastic waste produced in tons
- **RAEE (t)**: Electronic waste produced in tons
- **Selettiva (t)**: Selective demolition waste produced in tons
- **Tessili (t)**: Fabric waste produced in tons
- **Vetro (t)**: Glass waste produced in tons
- **Rifiuto da costruzione e demolizione (t)**: Construction and demolition waste produced in tons
- **Anno**: Year in which data was collected

Some of the rows present missing values especially for those waste that are less common and require specific processing like clothes, construction material and electronic waste.

1.5 Dataset ”Tourist Attraction Places”

Tourist attraction places 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.

1.5.1 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.

1.5.2 Data Collection

The following is a list of tourist attractions places which 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 center;
```

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 center;
```

1.5.3 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 1 provides comprehensive information on each processed attraction type.

Table 1: 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	-	
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	

Generated File Name	Columns	Unique Values	NaN Values	Table Size
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	
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		

Generated File Name	Columns	Unique Values	NaN Values	Table Size
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		
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	

1.6 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 places based on our purpose definition. Once this decision was made collec-

tively, 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 Type” and ”Waste Basket” datasets, and Yishak on the ”Tourist Attraction Places” 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.