



UNIVERSITY
OF TRENTO - Italy



DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

— KNOWDIVE GROUP —

Facilities activity in lockdown

Document Data:

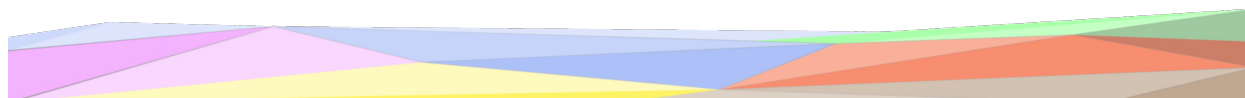
October 11, 2021

Reference Persons:

Manh Tuan NGUYEN and Nicola Giuseppe MARCHIORO

© 2021 University of Trento Trento,
Italy

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

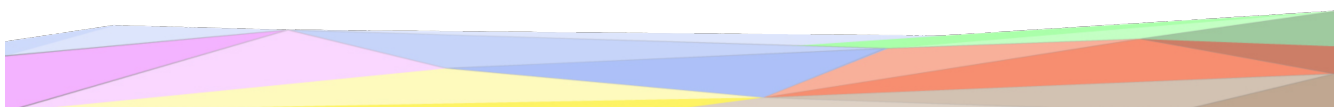


Contents

1	Introduction	1
2	Purpose and project's resources	1
3	Inception	1
4	Informal Modeling	2
5	Formal Modeling	2
6	Data Integration	2
7	Outcome exploitation	3

Revision History:

Revision	Date	Author	Description of Changes
0.1	02.11.2021	Manh Tuan NGUYEN	Purpose and project's resources section, Inception section written
0.2	03.11.2021	Nicola Giuseppe MACHIORO	Fill-up necessary information
0.3	20.11.2021	Manh Tuan NGUYEN	Updated inception phase - dataset metadata section
0.4	20.11.2021	Manh Tuan NGUYEN	updated knowledge sources description
0.5	21.11.2021	Nicola Giuseppe MACHIORO	Updated analyzed CQs
0.6	22.11.2021	Manh Tuan NGUYEN	Informal modeling's backbone written and finished the data part in this section
0.7	03.12.2021	Manh Tuan NGUYEN	Update inception phase kernel concept along with knowledge resources. Update informal modeling's phase.
0.8	20.12.2021	Nicola Giuseppe MACHIORO	Re-update EG description, edit evaluation phase
0.9	05.01.2022	Manh Tuan NGUYEN	Fill in Data Itergation and Data Exploitation part
1.0	10.01.2022	Nicola Giuseppe MACHIORO	Final review.



1 Introduction

Reusability is one of the main principles in the Data Integration (DI) process defined by iTelos. The data integration project documentation plays an important role in order to enhance the reusability of the resources handled during the methodology, as well as for the resources produced by the data integration process. A clear description of the resources and the process that has to manage them, provides a clear understanding of the information handled in the DI project, allowing external readers to exploit the same resources in different projects.

The current document aims to provide 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: How the result of the integration process (KGs) can be exploited.

2 Purpose and project's resources

a) Project's purpose:

Our project goal is to study and analyze students' behavior, the main focus is on the location and movement of students in order to generate a heatmap and to create an average person profile for specific Points of Interest.

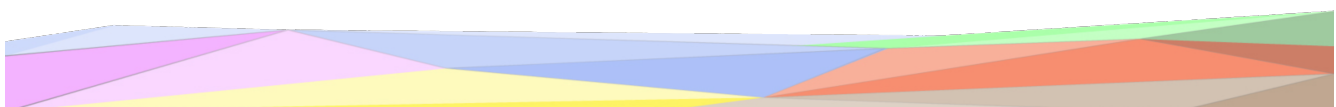
b) Personas:

First, it's important to analyze all the personas that could be relevant to the project's purpose. From that we can actively know the DoI and begin to determine the reusability and shareability of the dataset and knowledge we'll collect for this project following the Itelos Methodology.

- **Marco (39 years old):** a shop owner in the city of Trento. He wants to profile his customers to have a clear picture of the pandemic environment and how it is affecting his business. From that, he can find a way to make his shop more appealing to the public.
- **Sara (28 years old):** a sociology researcher in Trento. She is interested in analyzing the behavior of students throughout the year to discover interesting and unusual patterns.
- **Luca (30 years old):** a public administration employee. He wants to monitor the movement of students to find out the most crowded places during the day and evening in order to apply necessary measures if needed to prevent the outbreak of Covid-19.
- **Angela (25 years old):** an employee of a public transport company in Trento. She wants to understand which are the most popular facilities and areas of the city, at different times throughout the day. Depending on what she finds out she may request the addition of bus/train rides.
- **Chi (35 years old):** An employee at R&D department of a network provider company in Italy. The company currently wants to expand more in the Trento area and Chi is assigned to do some research about the signal around the city. She wants to know the quality of networks around this city, from which she may suggest to the headquarter to improve the network station at that place in order to please the customer.

c) Scenarios:

As this is a research project and our point of view in terms of facilities is quite big compared to others, we need to narrow it down by building different scenarios. We use different personas above to make a hypothesis scenario which our system could be used for.



- **Scenario 1:** Mr. Marco's store was really crowded before with all the international students around the campus, but since the pandemic, his sales have dramatically dropped. He wants to understand how his store is doing compared to others in the city center. He also wants to study a plan to revive his business after the covid period.
- **Scenario 2:** Mrs. Sara is really interested in human behavior, especially international students during the pandemic in Italy since she is a sociology researcher. She wants to gather and analyze all the information about what place students could use and the frequency of that. From that information, she could make assumptions about the average behavior during the Covid-19 crisis.
- **Scenario 3:** Mr. Luca is a concerned government staff member about the Covid-19 disease. After reducing significantly the covid cases, Trento city plans to open up about the restriction of the lock down. But by the experiences from the last 3 waves, Mr. Luca's plan is not to rush adopting a slow approach. For that, he wants to know where the crowded places which may be a good environment for the virus in order to apply specific restriction rules at these locations.
- **Scenario 4:** Ms. Angela is assigned to plan the re-opening of public transport. From the lockdown, a lot of places have to close so the number of customers using public transport has dropped dramatically, so the public transport company has reduced the number of busses/trains. But now, since the situation is getting better and there is a rumor about the re-open strategy from the government, the company wants to increase some necessary routes. Ms. Angela oversees the project, and she wants to know which places could be used a lot by people in Trento and from that, she can decide which parts of the city need better transport coverage.
- **Scenario 5:** Mrs. Chi is head of R&D department of a new network provider company. The company plans to expand their market to Trento. In order to do that, Mrs. Chi needs to know the quality of the network provider in Trento, so that her company can improve and have a certain advantage over other providers in the terms of customer satisfaction.

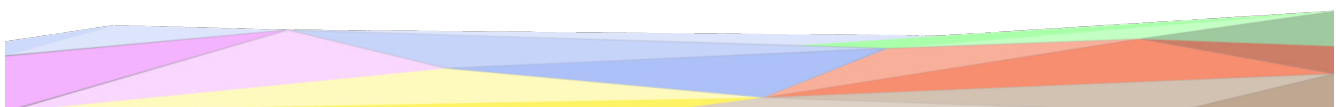
d) Knowledge resources:

To fulfill the purpose of the project, we need to find and decide the knowledge resource where we get the ontology information. After looking at the internet and also consulting the suggestion from the professor and his teaching assistant team, we decided to pick schema.org as our knowledge resource. The reason is that schema.org is a big collaborative, community activity that is easy to find on the internet with a mission to create, maintain and promote schemas for structured data. It also provides a website that contains all the vocabularies (ontologies term) in a hierarchical form. Also we can easily download the vocabularies set in a form of specific type we want from .csv to specialize ontology document .owl/.

e) Data resources:

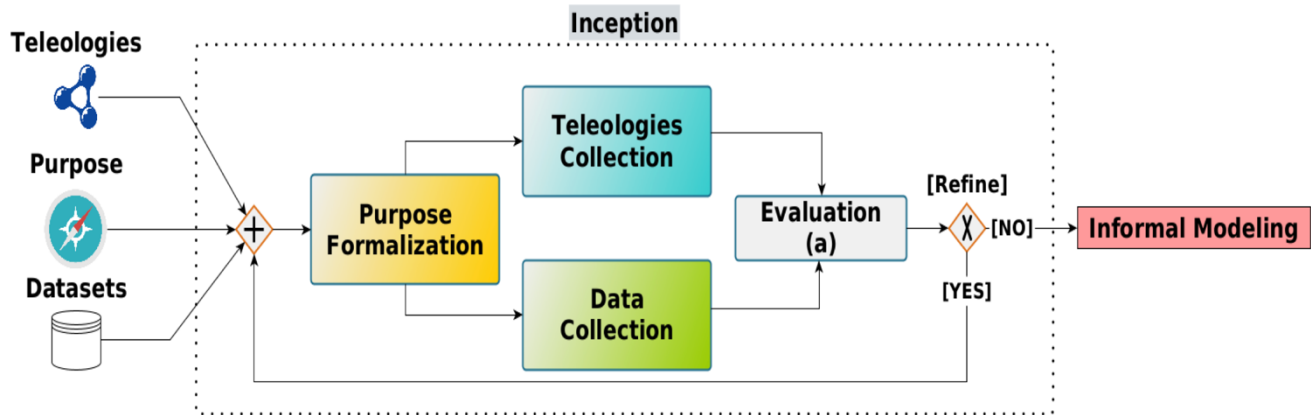
WeNet Diversity pre-pilot data

This project is part of a WeNet European big study. So the dataset has been provided by the Knowdive research group in Trento university. The dataset was collected last year thanks to the participation of 254 students in Trento University. They answered the necessary question and shared anonymized information about their general location to the Knowdive group. Also, the dataset provides the Point of Interest based on the OpenStreetMap dataset. All the detailed information will be described in the Inception section.



3 Inception

This project strictly follows the Itelos methodology so this section is dedicated to the Inception phase – the first phase of our methodology. Inside this part, we discuss the initial definitions for Competency Questions (CQs), initial dataset collected and relevant metadata.



a) Purpose formalization:

Covid-19 – the disease that changed the way humans will live forever. We could have never imagined a world where everything is closed, where we have to do everything from a distance, or a world where we only have anti-social interactions. Throughout history, mankind has never had to face this kind of crisis. We have never been prepared for this situation. At this time of crisis, mankind feels confused and doesn't know how to deal with the pandemic while improving the economy. So, this project's been created to analyze the pandemic's life, specially in Trento. With this study we hope we can explore different aspects of our pandemic's life so from that, mankind could be well prepared if we ever face this situation once again. For the scope of the project, we consider the Trento city's facilities as an environment where we can integrate heterogeneous data about the university students' behavior during the lockdown period caused by the Covid-19 pandemic. This project focuses on the facilities' point of view in order to analyze all the relevant Domain of Interest (DOI).

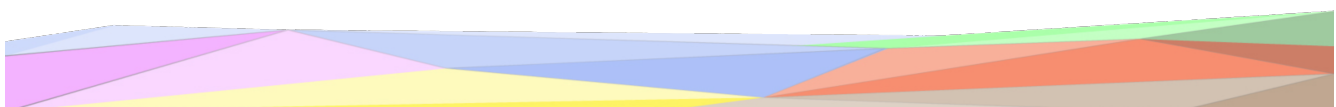
During this section, we must first formalize the purpose of the project. In order to do so, we have to define the personas, the scenarios and also the Domain of interests (DOI). Since we already defined Personas and Scenarios in the previous section, in this part, we will first define the DOI and report the inception sheet.

Following the scenarios and personas, we can first define the Dols list that we need and from that, we could classify the type of CQ we have.

- *Public transport routes in Trento*
- *Facilities frequency in Trento*
- *Store customer profiling in Trento*
- *Network quality in Trento*

From the 4 Dols above, now we present the produced inception sheet. Inside the sheet, we will go more in detail on CQs and the type of each, based on the DOI, Personas and Scenarios.

i. Raw CQs:



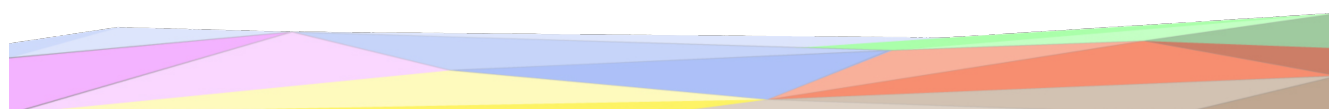
This part we introduced our first informal list of CQs which are elicited from the the purpose of this project.

Personas	Scenario	Competency Question
Marco	1	What is the average profile of my shop's customer?
Marco	1	What is the type of food my customer bought? (Organic,...)
Marco	1	How often do people come to my shop?
Sara	2	What is the most crowded part of the city depending on time and date?
Sara	2	What is the most popular kind of facility visited by students?
Sara	2	How much time do students spend inside/outside each day?
Luca	3	What are the most visited places (type)?
Luca	3	Are there differences between people of different age/gender/nationality regarding where they prefer to spend their time?
Angela	4	What is the most populated part of Trento?
Angela	4	What are the time slots that require more public transport rides?
Chi	5	What is the area with the worst cell phone signal quality?
Chi	5	What are the usual places where students use the network?

ii. **Kernel CQs:**

In this part, we will continue to expand the CQs list by remove all the auxiliary/apparatus words, which lead to a result of each term will denote a concept

Personas	Scenario	Competency Question	Kernel CQs
Marco	1	What is the average profile of my shop's customer?	Profile, Customer, Shop
Marco	1	What is the type of food my customer bought? (Organic,...)	Type , Food, Customer, Bought
Marco	1	How often do people come to my shop?	Often, People, Shop, Come
Sara	2	What is the most crowded part of the city depending on time and date?	City, Crowded, Time, Date
Sara	2	What is the most popular kind of facility visited by students?	Popular, Facility, Student, Visit
Sara	2	How much time do	Student, inside/outside,

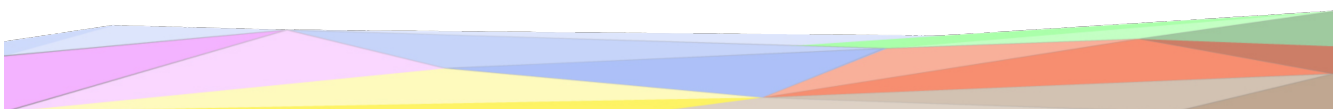


		students spend inside/outside each day?	Time
Luca	3	What are the most visited places (type)?	Most visited, Place
Luca	3	Are there differences between people of different age/gender/nationality regarding where they prefer to spend their time?	People, Age, Gender, Nationality, Time, Where
Angela	4	What is the most populated part of Trento?	Populated
Angela	4	What are the time slots that require more public transport rides?	Time Slots, Public Transport
Chi	5	What is the area with the worst cell phone signal quality?	Area, Cellphone signal quality
Chi	5	What are the usual places where students use the network?	Places, Students, Use, Network

iii. Analyzed CQs:

After achieved the Kernel CQs table, normally we should classify all the kernel concepts as common, core and contextual. But because of the complexity of the dataset, after we sit back and analyze again the Kernel CQs table and unified the kernel concept before going to the Analyzed CQs section. This indicate to minimized the number of Concepts created and bring better result.

Personas	Scenario	Competency Question	Common Kernel Concepts	Core Kernel Concepts	Contextual Kernel Concepts
Marco	1	What is the average profile of my shop's customer?	Person	Person Information	POI
Marco	1	What is the type of food my customer bought? (organic,...)	Person	Person Information	
Marco	1	How often do people come to my shop?	Person, Establishment	VisitPlace	POI
Sara	2	What is the most crowded part of the city depending on time and date?	Establishment	VisitPlace	
Sara	2	What is the most popular kind of facility visited by students?	Person, Establishment	VisitPlace	
Sara	2	How much time do students spend inside/outside each day?	Person, Establishment	Position	



Luca	3	What are the most visited places (type)?	Establishment	Position, VisitPlace	POI
Luca	3	Are there differences between people of different age/gender/nationality regarding where they prefer to spend their time?	Person	Person Information	
Angela	4	What is the most populated part of Trento?	Establishment	Position, VisitPlace	Region
Angela	4	What are the time slots that require more public transport rides?	Establishment	VisitPlace, Position	POI
Chi	5	What is the area with the worst cell phone signal quality?	Establishment	NetworkConnect, Position	POI
Chi	5	What are the usual places where students use the network?	Person, Establishment	Position, NetworkConnect	POI

b) Data and knowledge resource collection:

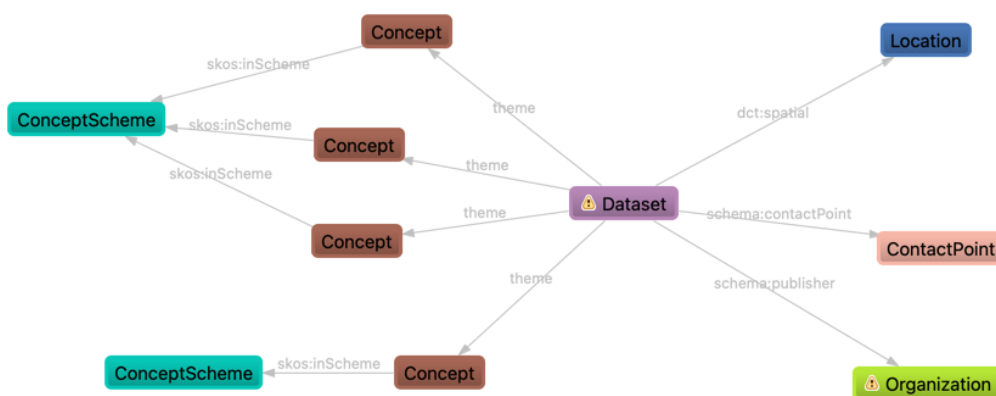
This dataset is named **WeNet Diversity pre-pilot data** which is provided by **Knowdive Research Group** in **Trento University**. In this part, we will present the metadata of this dataset by 3 levels : **Data sources**, **Data collected** and **Data attribute**.

i. Knowledge resource description

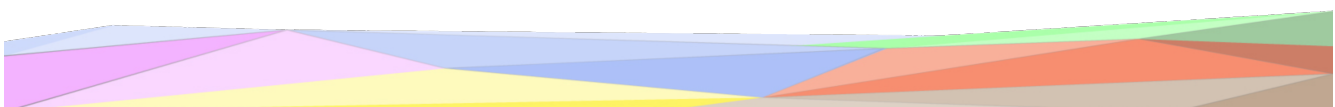
As we indicated in the second section, the knowledge resource we collected is from schema.org. To easily inspect and analyze, we use the .owl format for the knowledge resources. This resource is created by 2 groups : **Steering Group** and a larger **Community Group**. All the information about metadata of this resource can be found at this [link](#) .

ii. Data resources description

In this section, we report about the metadata at data sources level. That's to say in this part, we will provide the metadata regarding the sources, the authors, version, and all other necessary information that need to describe the dataset.



Above is the RDF graph which is used to describe the metadata information we will provide. First let's talk about the dataset itself. This dataset contains 4 csv files which is provided by **Knowdive Research Group** in **Trento city** where our **Contact Point** is **Mr.Matteo Busso**. 4 csv files is represent 4 different concept which are :



- POI location
- Approximate Location with Time series information
- GPS location (latitude ,longitude and attitude)
- Background information of candidate in experiment

To bring these concepts together, we have grouped them into 2 different Concept Schemes : Location and background information scheme. The Location scheme contains the first 3 concepts above and the rest belong to the Background information scheme. For more detail, please refer to the turtle file which will be provided along with this report at the end of KDI course.

iii. Dataset collected description

This project is part of a WeNet European big study. So the dataset has been provided by the Knowdive research group in Trento university. The dataset was collected last year thanks to the participation of 254 students in Trento University. They answered the necessary question and shared anonymized information about their general location to the Knowdive group. Also, the dataset provides the Point of Interest based on the OpenStreetMap dataset followed with approximate location and GPS location. This dataset was created during the lockdown period.

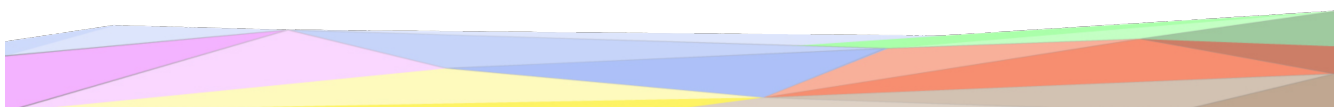
iv. Data's attribute description

During the inception phase, as the Data Scientist, Tuan is fully focused on analyzing the codebook provided by the Knowdive team. From that, he will first have a grasp on the quality of the data in the dataset and which features we should use. Since we don't have to collect any dataset because it's already provided by the Knowdive team, the hardest job is analyzing and understanding the features. Below is the metadata description of each file inside the WeNet Diversity pre-pilot data

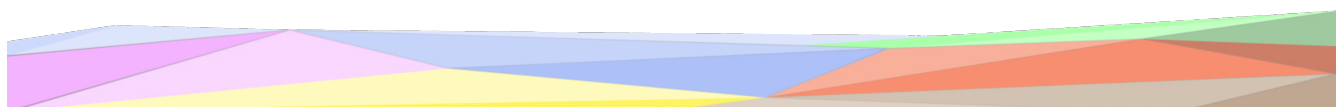
questionnaires_data.csv:

Dataset containing general info about the participants to the experiment, since the data is anonymized every person is identifiable only by a progressive ID. It contains basic personal data such as Nationality, Age and Gender, and more details about specific queries.

Variable	Class	Description
Token	String	The survey unique ID generated
userID	Int	The iLog unique ID of participant
Complete 2 nd survey	Int	If the participant completed the 2 nd survey
Complete 3 rd survey	Int	If the participant completed the 3 rd survey
Installed iLog	Int	If the participant installed the iLog
W1_A01	Int	Gender information
W1_A02	Int	Ages
W1_A03	Int	Nationality
Extraversion	Int	Personality scale (Base on Big 5 Inventory scale)
Agreeableness	Int	Personality scale (Base on Big 5 Inventory scale)
Conscientiousness	Int	Personality scale (Base on Big 5 Inventory scale)
Neuroticism	Int	Personality scale (Base on Big 5 Inventory scale)
Openness	Int	Personality scale (Base on Big 5 Inventory scale)
Linguistic	Int	Multiple Intelligences scale
Logicmath	Int	Multiple Intelligences scale
Spatial	Int	Multiple Intelligences scale



Bodykines	Int	Multiple Intelligences scale
Musical	Int	Multiple Intelligences scale
Interpersonal	Int	Multiple Intelligences scale
Intrapersonal	Int	Multiple Intelligences scale
Environmental	Int	Multiple Intelligences scale
Spiritual	Int	Multiple Intelligences scale
conformity	Int	Basic Human Value scale
tradition	Int	Basic Human Value scale
benov	Int	Basic Human Value scale
univers	Int	Basic Human Value scale
self	Int	Basic Human Value scale
stim	Int	Basic Human Value scale
hedon	Int	Basic Human Value scale
achieve	Int	Basic Human Value scale
power	Int	Basic Human Value scale
security	Int	Basic Human Value scale
open	Int	Basic Human Value scale
selfenh	Int	Basic Human Value scale
selftran	Int	Basic Human Value scale
conserv	Int	Basic Human Value scale
w2_B01_1	Int	Do you have a car driver 's license?
w2_B01_2	Int	Do you have a motorbike driver license?
w2_B01_3	Int	Do you have a bicycle?
w2_B01_4	Int	Do you have a car?
w2_B01_5	Int	Do you have a motorbike?
w2_B01_6	Int	Access to a car?
w2_B01_7	Int	Access to a motorbike?
w2_B02_1	Int	Transport method: Waking
w2_B02_2	Int	Transport method: Cycling
w2_B02_3	Int	Transport method: Car
w2_B02_4	Int	Transport method: Car sharing
w2_B02_5	Int	Transport method: Motorbike
w2_B02_6	Int	Transport method: Bus
w2_B02_7	Int	Transport method: Train
w2_B02_8	Int	Transport method: Electric scooter
w2_B03	Int	Frequent of using public transport per day
w2_C10_1	Int	Type of food last month: Organic
w2_C10_2	Int	Type of food last month: Zero-mile
w2_C10_3	Int	Type of food last month: Weight-loss product
w2_C10_4	Int	Type of food last month: Dietary supplements
w2_C10_5	Int	Type of food last month: Frozen Items
w2_C10_6	Int	Type of food last month: Allergen-free products
w2_C10_7	Int	Type of food last month: Ready to eat meals



w2_D02	Int	Have you been physically active on a regular basis in the last year or so?
w2_D04	Int	Frequent of physical exercise

timediaries_data.csv:

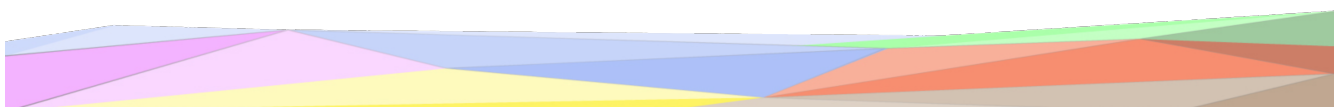
Dataset containing answers to a position query, for each entry it stores many information regarding the query instance time and the answer time.

Variable	Class	Description
userID	Int	The iLog unique ID of participant
Date	Date	Date when the notification was to be sent
Time	Time	Time when the notification was to be sent
Dt	DateTime	Date and time the notification was to be sent
Instance_date	Int	Date when the notification was sent
Instance_time	Int	Time when the notification was sent
Notification_date	Int	Date when the notification was received
Notification_time	Int	Time when the notification was received
Answer_date	Int	Date when the notification was answered
Answer_time	Int	Time when the notification was answered
Delta	Int	Time taken to respond the notification
Where	Int	Location (Base on timediaries-value_labels.xlsx)
Week	String	Week of the experiment

POI_data.csv:

Dataset containing info about a position automatically acquired by a person's smartphone network or GPS signal. It doesn't contain data about the location of the person, but about the surrounding area in which someone is. An entry is generated when a participant stands in a place for a certain period of time.

Variable	Class	Description
Timestamp	Int	Date and time of POI (form Month, day, hour, minute, second, decimals)
ExperimentID	Int	Experiment id
userID	Int	The iLog unique ID of participant
Bearing	Int	Compass direction
Speed	Int	Speed of device (m/s)
Network_provider	Boolean	If using network/Wifi
Gps_provider	Boolean	If using GPS
Suburb	String	The neighbor where the POI is located
City	String	the city where the POI is located
Region	String	The region where the POI is located
Moving	Boolean	If the user was moving
Coden	Int	4-digit code defining the feature class of each POI near the user
Fclassn	String	Class name of this location
Namen	String	Name of this feature (for example street or place name)



gps_data.csv:

The dataset contains information about a position automatically acquired by a participant's smartphone using GPS or network. This dataset is described in detail about the coordinate of the participant by latitude and longitude also with altitude.

Variable	Class	Description
Timetamp	Int	Date and time of POI (form Month, day, hour, minute, second, decimals)
ExperimentID	Int	Experiment id
userID	Int	The iLog unique ID of participant
Bearing	Int	Compass direction
Speed	Int	Speed of device (m/s)
Network_provider	Boolean	If using network/Wifi
Gps_provider	Boolean	If using GPS
Accuracy	Int	The GPS accuracy in meters
Latitude	Int	Latitude
Longitude	Int	Longitude
Altitude	Int	Altitude

c) Inception Evaluation:

Following the diagram above, after formalizing the purpose of the project, we have divided our task separately, Nicola oversees Knowledge and Tuan oversees Data. But we still crossover, for example Tuan may need to think about some example CQ in order to analyze the dataset and vice versa. After defining the list of CQs it will be evaluated by Tuan (Data Scientist) to determine if the dataset we have is suitable to solve them. If not, we must roll back again and filter the Competency Questions. This section is like a loop until we meet an agreement. During this progress, the CQs list will be strictly co-operated with data. So we worked closely together, and at each evaluation loop, whenever Tuan found a problem occurred he brought up the reasons and made examples of doable CQs in order to reach the goal and close the evaluation phase.

Beside that, iTelos methodology also provides the metrics in order to calculate the “fitness for use” of our work. With this evaluation, we can guarantee to obtain qualified entity graph in the final phase.

Below is where we calculate the necessary component for the metrics. Each component is calculated base on our kernel concepts and the ontologies provides by schema.org

- Kernel Classes (CQs) :

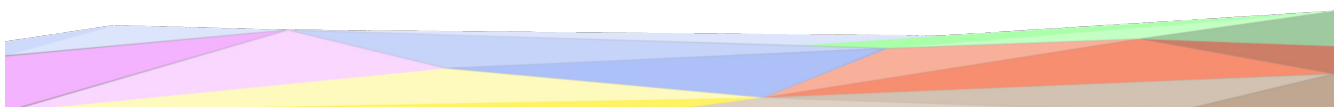
$$C_c = \{Customer, Shop, Food, Bought, People, Come, City, Facility, Student, Visit, Public Transport, Area, Places, Use, Network\} = 15$$

- Kernel Properties (CQs) :

$$C_p = \{about, Profile, Type, Often, Crowded, Time, Date, Popular, Most visisted, Age, Gender, Inside Outside, Time Slots, Populated, Cellphone signal Quality\} = 14$$

- Ontology Classes :

$$O_c = \{Person, Product, LocalBusiness, ShoppingCenter, City, Vehicle, Place, UseAction, ArriveAction, TravelAction, BuyAction\} = 11$$



- Ontology Properties :

$$O_p = \{BirthDate, Nationality, about, TypeofGood, acitvityFrequency, repeatFrequency, location, timeofday, strenghtUnit, strenghtValue \} = 10$$

Metrics	Classes	Properties
Coverage	6/15 = 0.4	6/14 = 0.42
Extensiveness	5/20 = 0.25	4/18 = 0.22
Sparsity	14/20 = 0.7	12/18 = 0.67

4 Informal Modeling

This section is dedicated to the description of the informal modeling phase. Like in the previous section, the current one aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced.

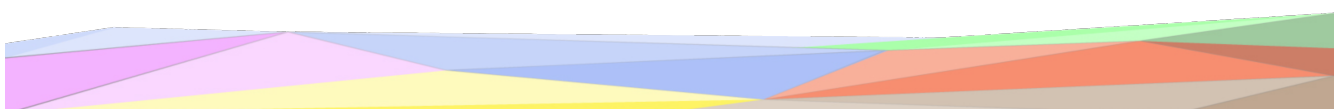
More in details, this section provides a description of the following activities:

a) Purpose formalization (informal modeling part) and Modeling sheet description

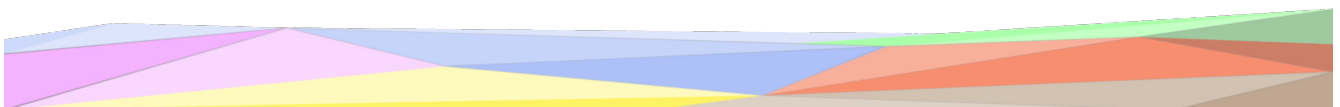
Inside this section, we continue to formalize our purpose from the last phase. In Inception phase, we already retrieved necessary Analyzed CQs which has been agreed by both Knowledge engineer and Data Scientist. From that, now we will devote it into 2 steps: Classified CQs and Atribute CQs. Please refer to this link for original table of this section.

i. Classified CQs:

From all the concepts that we analyze in the last phase, now we will go deeper by classify it as Object, Function or Action in order to make the ETypes in ER diagram. The table below will describe the classified CQs.



Competency Questions	Common Kernal Concepts			Core Kernal Concepts			Contextual Kernal Concepts		
	Object	Function	Action	Object	Function	Action	Object	Function	Action
1.1 What is the average profile of my customer?	Person				Person Information			POI	
1.2 What is the general idea about specific types of food that customer bought? (organic,...)	Person				Person Information				
1.3 How often do people come to my shop?	Person, Establishment					VisitPlace		POI	
2.1 What is the most crowd part of the city depending on time and date?	Establishment					VisitPlace			
2.2 What is the most popular kind of facility visited by students?	Person, Establishment					VisitPlace			
2.3 How much time do students spend inside/outside each day?	Person, Establishment			Position					
3.1 What are the most visited places (type)?	Establishment			Position		VisitPlace			
3.2 Are there differences between people of different age/gender/nationality regarding where they prefer to spend their time?	Person				Person Information				
4.1 What is the most populated part of Trento (by students)?	Position			Position		VisitPlace		Region	
4.2 What are the time slots that require more public transport rides?	Position					VisitPlace			
5.1 What is the area with the worst cell phone signal quality?	Position					NetworkConnect			
5.2 What are the usual places where students use network?	Position			Position				POI	



ii. Attribute CQs:

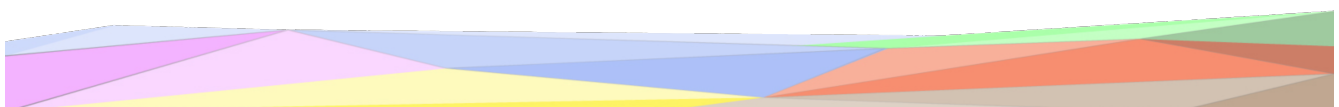
From the above Classified CQs table, now we need to assign the correspond attribute from the dataset we have collected at the beginning. Inside the table I describe all the attribute by the meaning of features of dataset (easier for everyone to get the idea of ER diagram) So for the correct features name term inside dataset, please refer to attribute metadata above.

EType	Description	Relation	Data Properties
Person	A common Etype describe a human living	It's a PartOf Things and also a superclass of Person Information (HasFunction) and VisitPlace with NetworkConnect (HasAction)	userID: String Gender: Int Age: Int Nationality: String
Establishment	A common Etype describe a location	It's a PartOf Things and also a superclass of Position (HasFunction)	TypeOfLocation: Int
Position	A Core Etype describe detail about a location by GPS coordinate	It's a Function of Establishment since every place will have a coordinate , therefore it inherit all the attributes of Establishment It's also a superclass of Region and POI (hasFunction)	ExperimentID: Int Latitude: Long Longitude: Long
Person Information	A Core Etype describe "about" properties of a human	It's a function of Person since all person will have their information.	HaveCar: Int HaveBike: Int AccessToCar: Int AccessToBike: Int OrganicFood: Int ZeroMileFood: Int WeightLossProduct: Int DietarySupplements: Int FrozenItems: Int AllergenFreeProduct: Int ReadyToEat: Int ExerciseFrequency: Int
VisitPlace	A Core Etype describe an action to go to specific place	It's an Action of Person Etype and also it's an admissible action to a position since Person can visit any place	Timestamp: Long Answer_time: Int Answer_date: Int
NetworkConnect	A Core Etype describe an action to connect to specific type of network	It's an Action of Person Etype and also it's an admissible action to a position since Person can perform connection network at any place	NetworkProvider: Boolean GPSProvider: Boolean Accuracy: Int
POI	A Contextual Etype describe the specific name of the location by code of OpenStreetMap	It's an Function of a Position since we will know the name of a street, store,... when we know their coordinate.	POIClass: String POICode: Int POIName: String
Region	A Contextual Etype describe the neighbor of the location	It's an Function of a Position since we can know in which neighbor (city, region, suburb) we are in if we know the GPS coordinate.	Suburb: String City: String Region: String
Things	The most generic object of all item	The parent Etype of everything	N/A

b) ER Model description

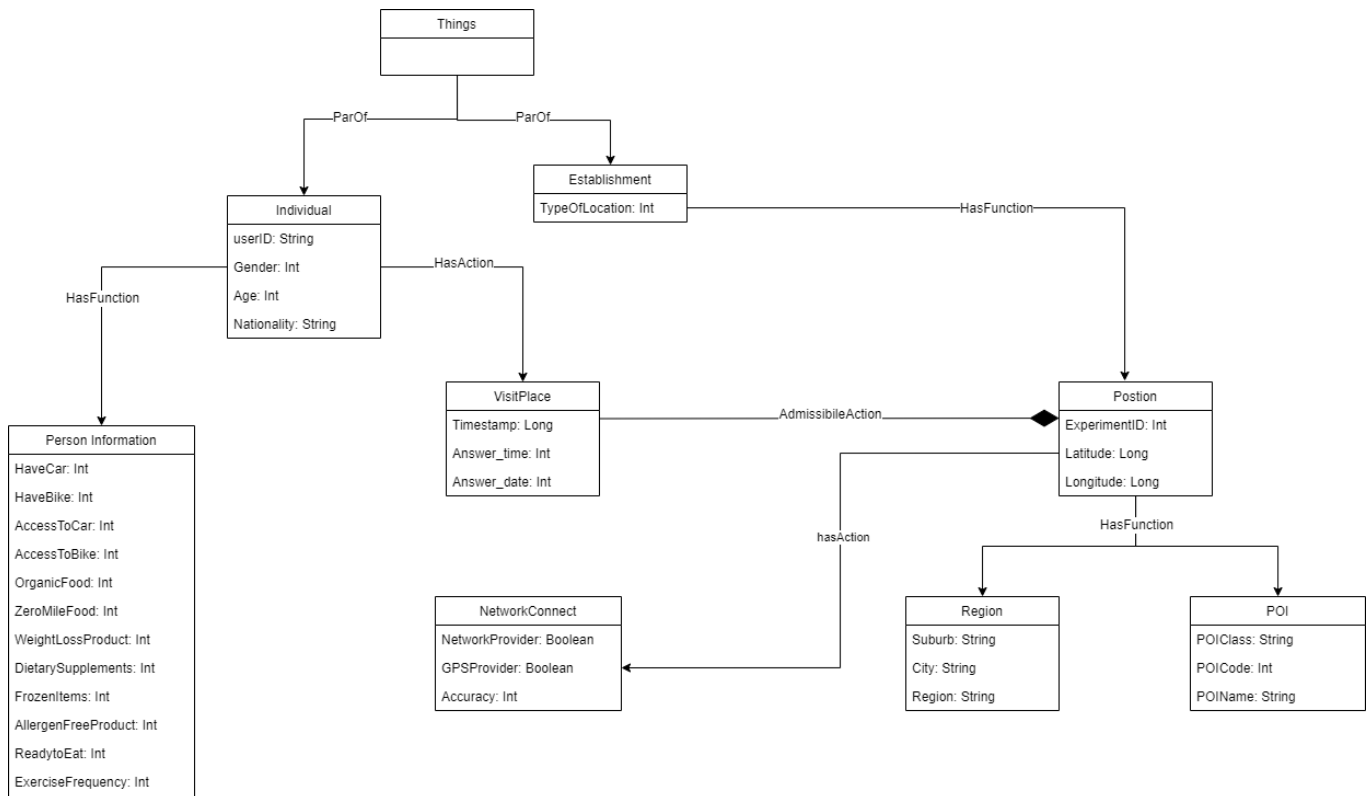
After archived the list of ETypes along with the correspond Attribute, we created the ER diagram follow these steps:

- First, we defined all Etypes by Object, Function and Action.
- We created all Objects which is all the subclass of Things.
- Then we matched all correspond Function to each Object.
- After obtain all the function, we begin to matched all relevant Action.



- At the end, we will look back and see if any EType will accept these action as “Admissible Action”

The relation has been already described above since it’s better to read it in the table. Below is the ER diagram which we created by draw.io

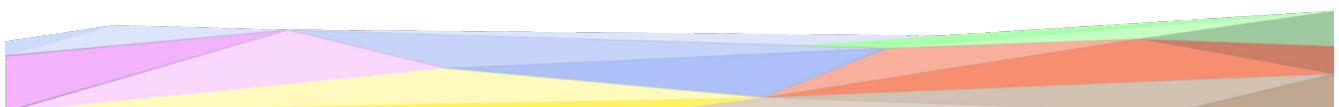


c) Data level

In this section, we report about the evolution of the dataset that we have collected in the previous phase. So during the inception phase, we have been provided the metadata of **WeNet Diversity pre-pilot** dataset. Following the iTelos methodology , during the informal modeling phase, we have to research carefully about each attribute of the datasets and from that provide a final dataset which only contains vital information.

After finalizing the ER models, we decided to remove the ones that would not be any relevant to our scope of project and cannot integrate with the initial set of entities defined above. All the detailed information about each attribute can be found in the metadata attributes level above in the Inception phase section. Below is the table detail about the necessary attributes will be keep in each of our csv files

	<i>questionnaires_data</i>	<i>timediaries_data</i>	<i>POI_data</i>	<i>gps_data</i>
Attributes remain	UserID; W1_A01; W1_A02; W1_A03; w2_B01_4; w2_B01_3;	userID; Instance_date; Instance_time; Notification_date;	userID; Timestamp; Network_provider; Gps_provider; Coden; Fclassn; Namen	Timestamp; userID; Network_provider; Gps_provider; Accuracy; Latitude;



	w2_B01_6; w2_B01_7; w2_C10_1; w2_C10_3; w2_C10_4; w2_C10_5; w2_C10_6; w2_C10_7; w2_D04	Notification_time; Answer_date; Answer_time; Where		Longitude
--	---	--	--	-----------

Beside analyzing the attributes, the data scientist role also did the pre-processing action to clean the dataset. We did clear all missing value data points, also only choose the data which consider the region: **Trentino-Alto Adige/Sudtirolo** which will be suitable for all scenarios and personas defined above.

d) Informal Modeling evaluation:

During this phase, following the iTelos methodology, we continue to work in parallel and meet each other at the end of the phase. After a few iterations, we came to an agreement of how the ERs model should be and also all the necessary attributes will be used to cover up the CQs list. Due to the iteration action, currently our CQs list is 100% covered by our dataset.

- **Coverage:**

Class Coverage= $Nc(CQ) | Nc(ER)=1$

Property Coverage = $Np(CQ) | Np(ER)=1$

- **Extensiveness:**

Class Coverage= $Nc(CQ) | Nc(ER)=1$

Property Coverage = $Np(CQ) | Np(ER)=1$

(Nc: Number of classes, Np: Number of properties)

5 Formal Modeling

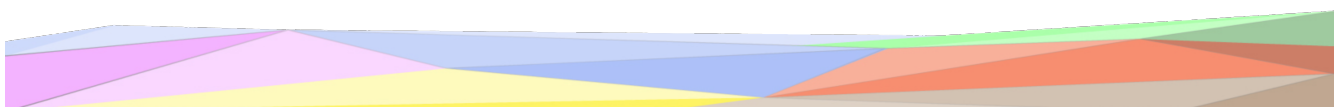
a) ETG Generation

During this phase we used the ER previously generated to build the ETG model, doing so we mapped our Etypes to already-existing concepts when possible. First we created our ontology using Protégé, after defining Classes, Object and Data properties we started doing the language alignment through the kos application. In this way we were able to map our Etypes to already-existing concepts when possible.

Language alignment: The goal of this phase is to remove the intrinsic ambiguity that lies in natural language, to do so we need to assign a precise identifier (and thus a definition) to each concept in our ETG. We used the KOS application to browse the Universal Knowledge Core to find the most appropriate concept to each term and, when no description would fit our needs, we created new entries.

Schema alignment: To carry out this task we made use of Protégé. We added all the concepts we needed following the given hierarchy, and then we uploaded the file to the KOS application. At first, we had some problems with one of our main Etypes (Individual), but after a discussion with S. Bocca we figured out what the problem was and decided to change it to Participant in order to fix the issue.

These steps are fundamental to close the gap between the ETG model and the Foundational Teleology



producing a sharable and reusable formal ETG made of objects, functions, and actions.

Here are listed all the Etypes used in the ETG with their associated GID:

Individual (GID:118)

"A human being"

UserID: String (Even if the IDs in the dataset are integer numbers, we decided to put it as a string for future implementation that might use a different id system to identify individuals, for example the SSN)

Gender: Int (Not Boolean as we have 3 different possible values in the dataset)

Age: Int

Nationality: String

Establishment (GID:17902)

"Any area set aside for a particular purpose"

TypeOfLocation: Int (This refers to a <Int, String> map given by the Knowdive Group to categorize places)

Person_information (GID:45661)

"A collection of facts from which conclusions may be drawn"

HaveCar: Int (Not Boolean as we have more than 2 different possible values in the dataset)

HaveBike: Int (Not Boolean as we have more than 2 different possible values in the dataset)

AccessToCar: Int (Not Boolean as we have more than 2 different possible values in the dataset)

AccessToBike: Int (Not Boolean as we have more than 2 different possible values in the dataset)

BoughtOrganicFood: Int (Describes frequency over the last month)

BoughtZeroMileFood: Int (Describes frequency over the last month)

BoughtWeightLossProduct: Int (Describes frequency over the last month)

BoughtDietarySupplements: Int (Describes frequency over the last month)

BoughtFrozenItems: Int (Describes frequency over the last month)

BoughtAllergenFreeProduct: Int (Describes frequency over the last month)

BoughtReadytoEat: Int (Describes frequency over the last month)

ExerciseFrequency: Int

Position (GID:27990)

"The spatial property of a place where or way in which something is situated"

ExperimentID: Int

Latitude: Long

Longitude: Long

Region (GID:300002)

"A geographic area delimited by political border"

Suburb: String

City: String

Region: String

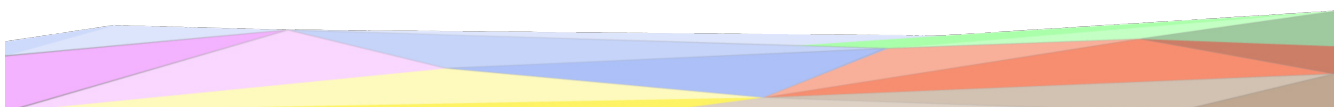
POI (GID:300000)

"A specific point location that someone may find useful or interesting"

POIClass: String (Class generated by OpenStreetMap API, it defines the general purpose assigned to a location)

POICode: Int (Code generated by OpenStreetMap API)

POIName: String (Specific name of a POI)



Network_connect (GID:300001)

"The ability to connect to a universal network of communication"

NetworkProvider: Boolean

GPSProvider: Boolean

Accuracy: Int

VisitPlace (GID:300003)

"The action of being in a place at a given time"

Timestamp: Long

Answer_time: Int (Given as the integer representation of the string "HHmmss")

Answer_date: Int (Given as the integer representation of the string "yyyyMMdd")

Other than these classes and Data properties we have defined the following Object properties:

has_additional_information (Domains: Individual, Ranges: Person_information)

has_data_acquisition_info (Domains: Position, Ranges: Network_connect)

has_establishment_location (Domains: Establishment, Ranges: Position)

has_moved_to_place (Domains: Individual, Ranges: VisitPlace)

has_point_of_interest (Domains: Position, Ranges: POI)

has_suburb_to_which_it_belongs (Domains: Position, Ranges: Region)

has_visited_position (Domains: VisitPlace, Ranges: Position)

b) Data Management (syntactic heterogeneity)

During this phase, in the aspect of data scientist, we have to look deeper. In order to archive better alignment between dataset and the schema, we performed various analysis such as : unify format, unify value information, remove meaning-less datapoint. About the data value misalignment (format and language), our result is the dataset is well format and has the same Italian language using econde utf_8 format. This result could be easy to undersand since the dataset is come from the same source which is knowdive group and also we don't have any other dataset so the data itself has follow the rule when they collected it.

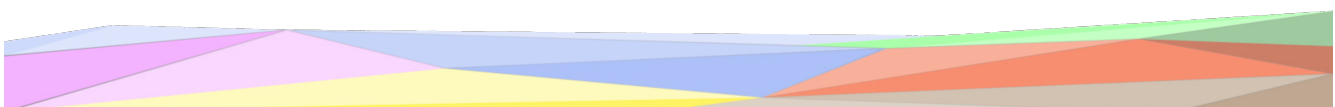
To be detail, I will describe each of data file change :

Questionnaire_data.csv:

- Format: CSV
- Total datapoint: 294
- Final total datapoint: 254
- With this file, we analyzed and found out that inside this, they also include a lot of "fake" users which is not install necessary device which relevant to the rest of dataset. So we decided to remove all "fake" users. Also, we checked and make sure all the data value of each features is correct to the description of Knowdive group.
- With all the missing value, we decided to keep it for reusable reason – if any other researchers want to do the research on how student interact with task for example, these missing value could be helpful.
- Remove all duplicate information

Timediaries_data.csv:

- Format: CSV



- Total datapoint: 264136
- Final total datapoint: 180869
- Not like questionnaire data, with this file, we removed all the datapoint has missing information in the feature “where” which indicate the position due to if we don’t know the position, that datapoint will be nonsense to the project purpose
- Remove all duplicate information

Poi_data.csv

- Format: CSV
- Total datapoint: 1919933
- Final total datapoint: 1029901
- Like timediaries data, we will drop all the datapoint has missing value on all the column relevant to POI position by OpenStreetMap (fclass, fcode and name – please refer to attribute metadata in Inception phase for more detail) .
- We also check all the string value to be consist and no duplicate. We found that all the name is follow the English noun with Italian name. So there is no conflict here since it’s all the latin word and all the term strictly follow English language.
- Set utf_8 decode in order to keep Italian special character.
- Remove all duplicate information

Gps_data.sv

- Format: CSV
- Total datapoint: 80574668
- Final datapoint: 80574668
- The file is too large to be exploit deeply like others but lucky that the file only contains information about the gps coordinate which we only need to guarantee the data value format is “long” type.
- Remove all duplicate information

Data pre-processing tools:

- Language: Python
- Framework: Pandas

c) Formal Modeling Evaluation

This phase we continue to develop the ETG from ER diagram from Informal Modeling phase and also align the dataset in the right format in both term of file and data value. We have managed to achieve both the ETG and dataset in time and right form.

To measure the quality of our ETG we tried to apply the Cue validity, we couldn’t use the liveschema tool so we decided to try with the schema.org ontology, but we still didn’t manage to create the FCA lattice, so we just used the info found at this [link \(https://schema.org/docs/schemas.html\)](https://schema.org/docs/schemas.html) to calculate Coverage and Extensiveness.

- **Coverage:**

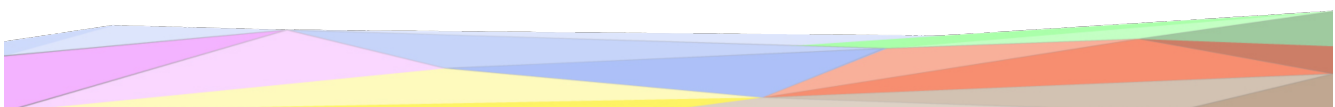
Class Coverage= $Nc(ETG) | Nc(ONT)=0.63$

Property Coverage = $Np(ETG) | Np(ONT)=0.41$

- **Extensiveness:**

Class Coverage= $Nc(ETG) | Nc(ONT)=0.99$

Property Coverage = $Np(ETG) | Np(ONT)=0.98$



(Nc: Number of classes, Np: Number of properties)

6 Data Integration

In this last section, we will talk detail about the data itergration files. This is the last phase of iTelos method where we take all the output of previous phase to process. From the ETG and final data from Formal Modeling phase, after perform *Semantic Heterogeneity* , we can generate the EG out of ETG and data we have.

a) Semantic Heterogeneity

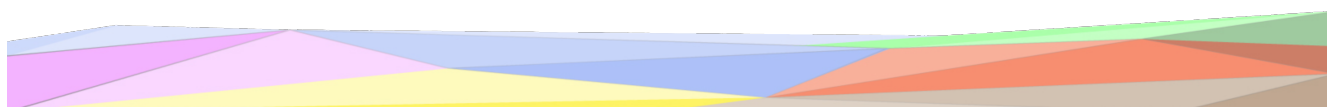
First we need to understand what is *Semantic Heterogeneity*. “Consequence of the more general phenomenon of the diversity of the world and of the world descriptions.” (Giunchiglia , Fumagalli 2020) is a perfect definition of Semantic Heterogeneity. So normally all the data we collect can come from different aspect in life. During our course, a really good example can make it clearer for you. This is an example about a datapoint of the bus. The bus information from manufacturer is different compared to transportation company. Since the factory only care about the type of engine, the fuel type, in the meanwhile that is not what transportation company data show. For them, they care about the capacity of the bus, the licence etc... So when we do any project, the collection of dataset is really important and we have to collect from a lot of sources which can bring this problem. To deal with this, we need to do **Entity Mapping** which aims to different features of each data sources we have represent the same concept in the real world. For this part, it’s not our problem since our dataset is provided by only DISI group. All the features of each data is well present and well defined (please refer to Inception phase for description of each features). So we only left with the **Entity Alignment** where we need to connect our data to the ETG graph.

b) Entity Alignment

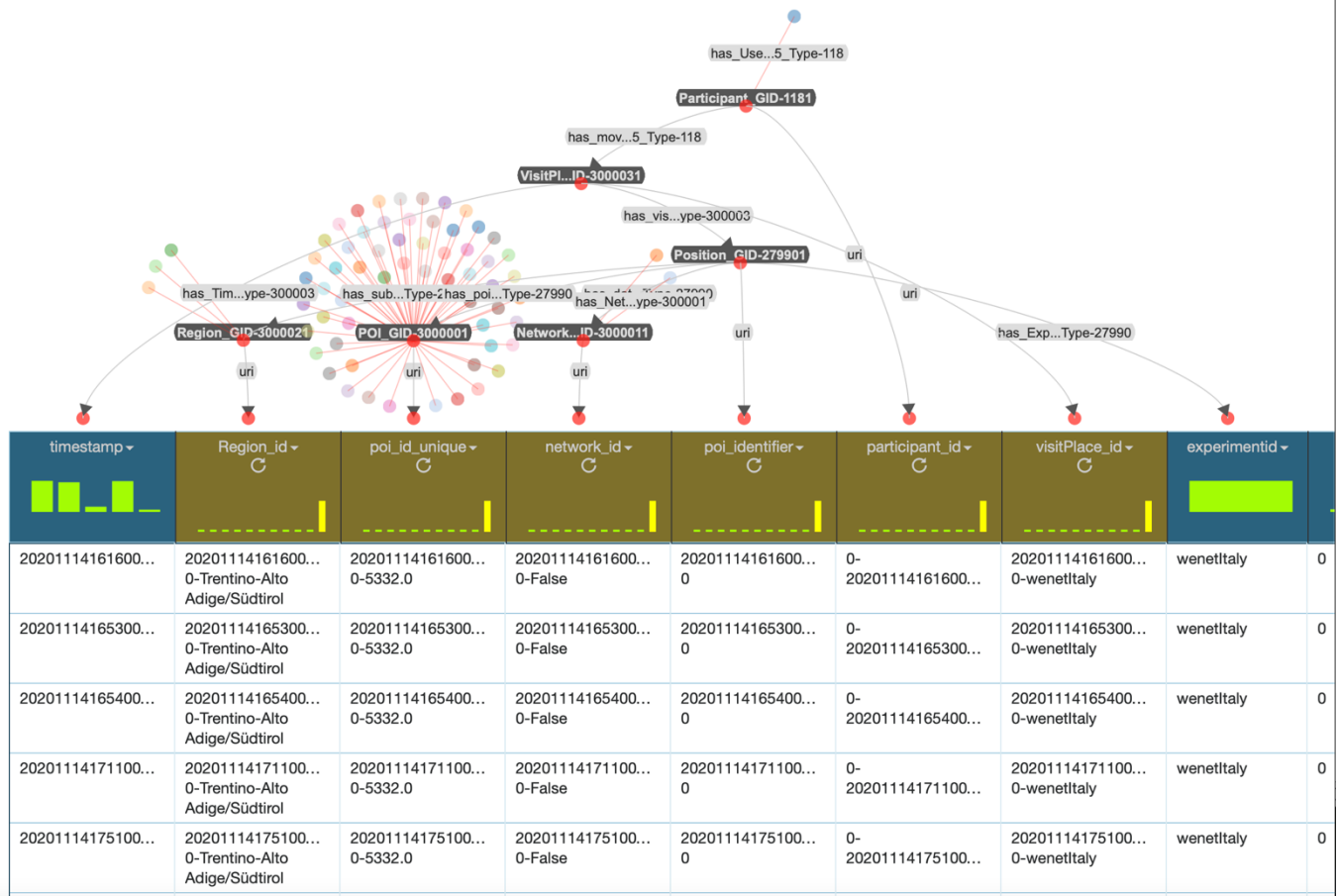
This process aims to map multiple representation of dataset to the single, purpose-specific schema which we already known as ETG from the Formal Modelling phase. To do this, we used the Karmanlinker application to do this. This part is not too heavy compare to other phase of this project since the beginning, we already paid a lot of time to think and analyze our dataset and the ontology we have. So from this, we only map each ontology term we create in ETG with corresponding data features (by description since all the ontology we created based on the description of each data but in the most common way in order of reusable target of iTelos). Below is all the detail about each dataset files. For each of these files, except Questionnaire dataset, all files have to applied PyTransform to create unique column to be uri for each class of EG model.

- Poi.csv

Data Representation	Class	Property
Timestamp	VisitPlace_GID-3000031	has_Timestamp_GID-80563_Type-300003
ExperimentID	Position_GID-279901	has_ExperimentID_GID-39085_Type-27990
userID	Participant_GID-1181	has_UserID_GID-39085_Type-118
Network_provider	Network_connect_GID-3000011	has_Network_provider_GID-300011_Type-300001

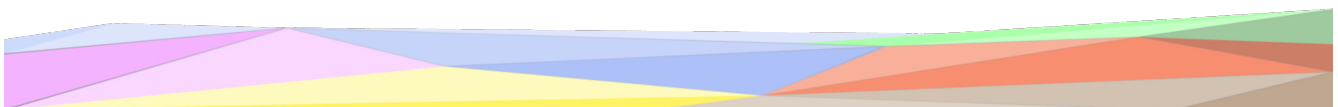


GPS_provider	Network_connect_GID-3000011	has_GPS_provider_GID-300019_Type-300001
Suburb	Region_GID-3000021	has_Suburb_GID-46054_Type-300002
City	Region_GID-3000021	has_City_GID-45969_Type-300002
Region	Region_GID-3000021	has_Region_GID-300002_Type-300002
Fclass#number	POI_GID-3000001	has_POI_class_GID-300010_Type-300000
Code#number	POI_GID-3000001	has_POI_code_GID-300017_Type-300000
Name#number	POI_GID-3000001	has_POI_name_GID-2_Type-300000

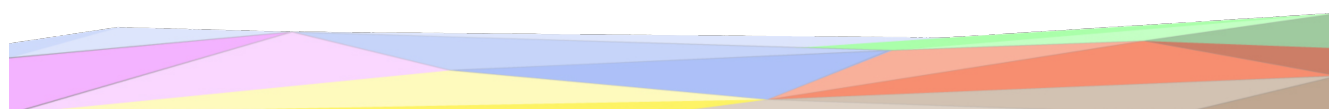


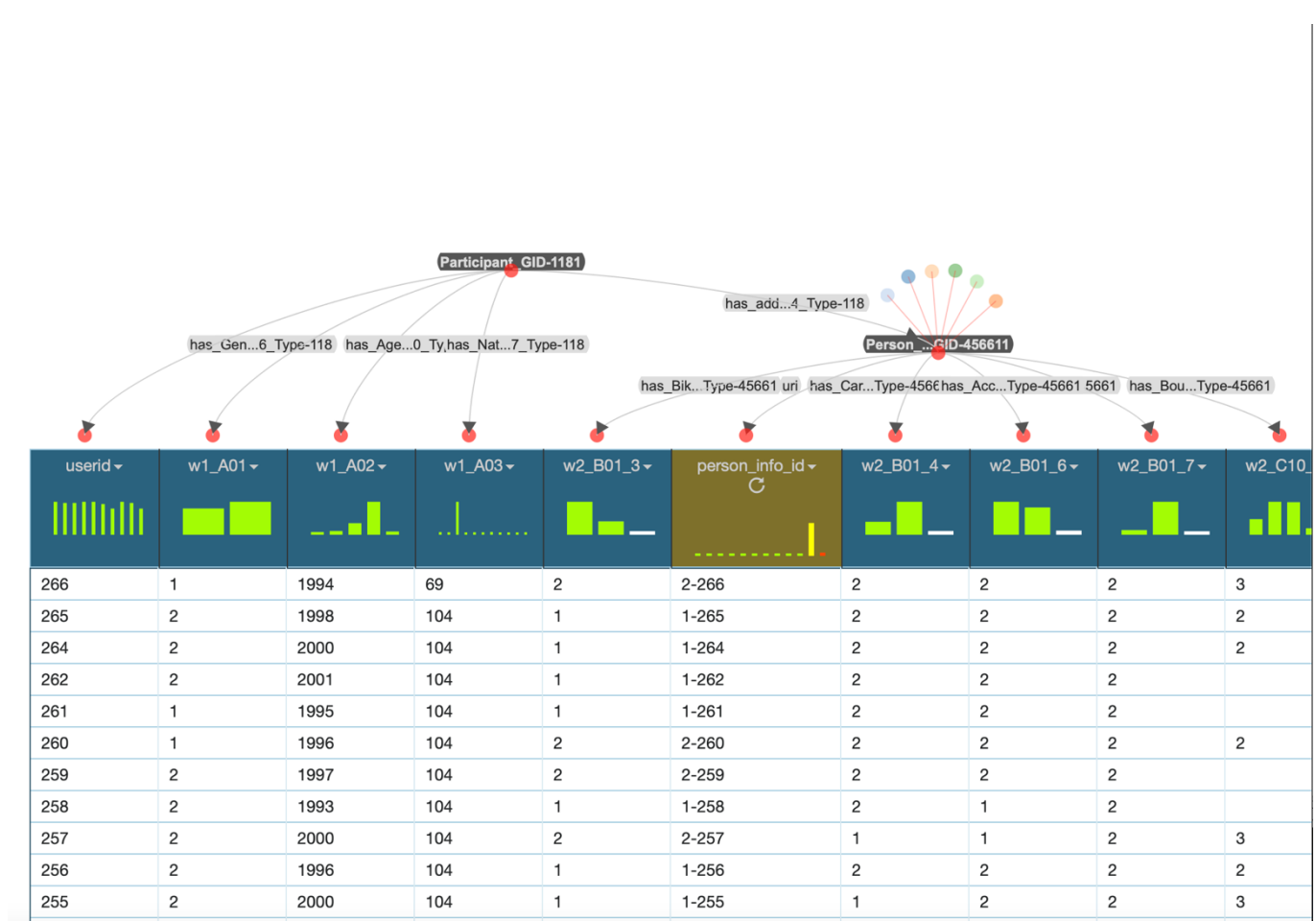
- Questionnaire.csv

Data Representation	Class	Property
userID	Participant_GID-1181	has_UserID_GID-39085_Type-118
W1_A01	Participant_GID-1181	has_Gender_GID-27646_Type-118
W1_A02	Participant_GID-1181	has_Age_GID-27200_Type-118



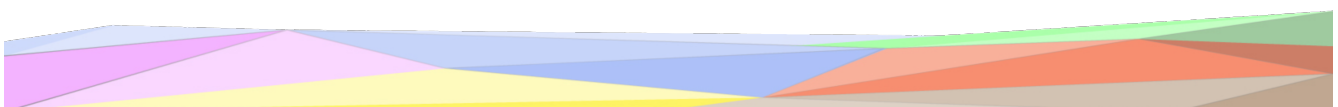
W1_A03	Participant_GID-1181	has_Nationality_GID-74177_Type-118
W2_B01_3	Person_information_GID-456611	has_Bike_GID-15188_Type-45661
W2_B01_4	Person_information_GID-456611	has_Car_GID-15944_Type-45661
W2_B01_6	Person_information_GID-456611	has_Access_to_car_GID-300018_Type-45661
W2_B01_7	Person_information_GID-456611	has_Access_to_bike_GID-110294_Type-45661
W2_C10_1	Person_information_GID-456611	has_Bought_organic_food_GID-300015_Type-45661
W2_C10_3	Person_information_GID-456611	has_Bought_weight_loss_products_GID-300008_Type-45661
W2_C10_4	Person_information_GID-456611	has_Bought_dietary_supplements_GID-300012_Type-45661
W2_C10_5	Person_information_GID-456611	has_Bought_frozen_items_GID-300016_Type-45661
W2_C10_6	Person_information_GID-456611	has_Bought_allergene_free_products_GID-300009_Type-45661
W2_C10_7	Person_information_GID-456611	has_Bought_ready_to_eat_food_GID-300014_Type-45661
W2_D04	Person_information_GID-456611	has_Exercise_frequency_GID-100877_Type-45661

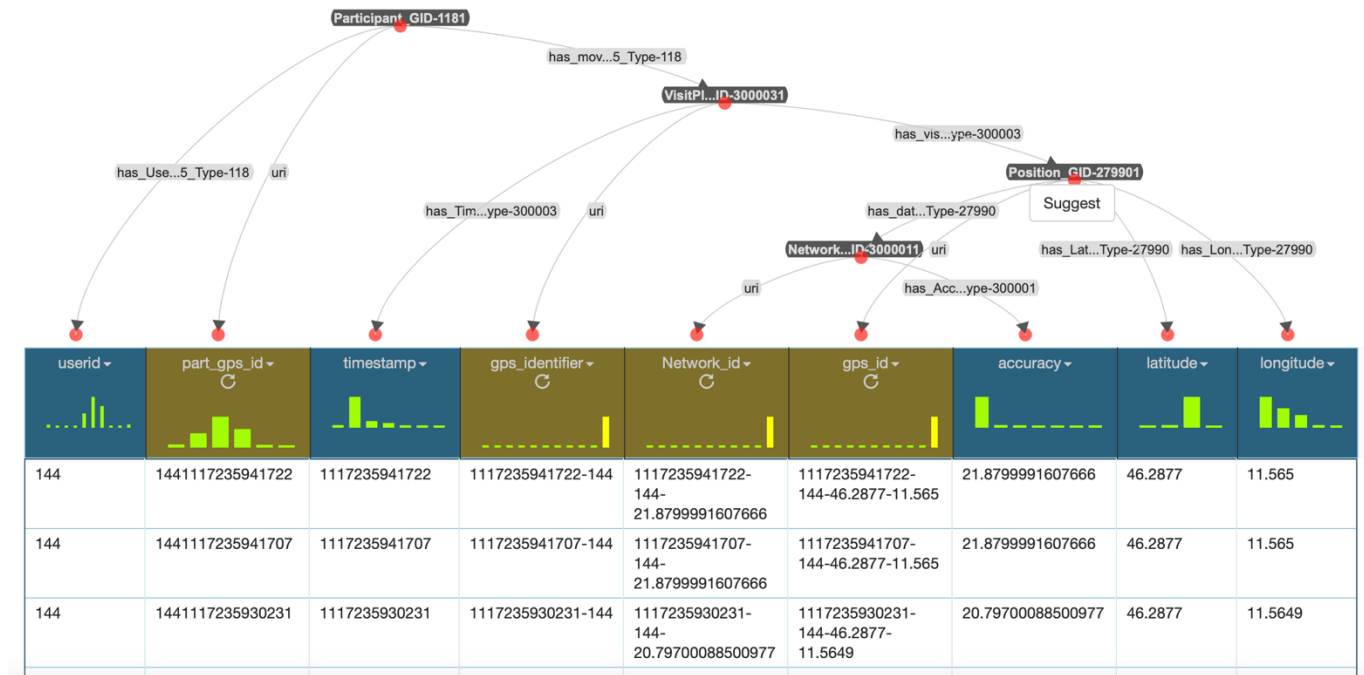




- Gps.csv

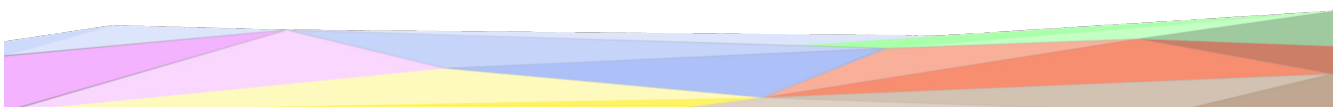
Data Representation	Class	Property
userID	Participant_GID-1181	has_UserID_GID-39085_Type-118
Timestamp	VisitPlace_GID-3000031	has_Timestamp_GID-80563_Type-300003
Accuracy	Network_connect_GID-3000011	has_Accuracy_GID-26585_Type-300001
Latitude	Position_GID-279901	has_Latitude_GID-46263_Type-27990
Longitude	Position_GID-279901	has_Longitude_GID-46270_Type-27990

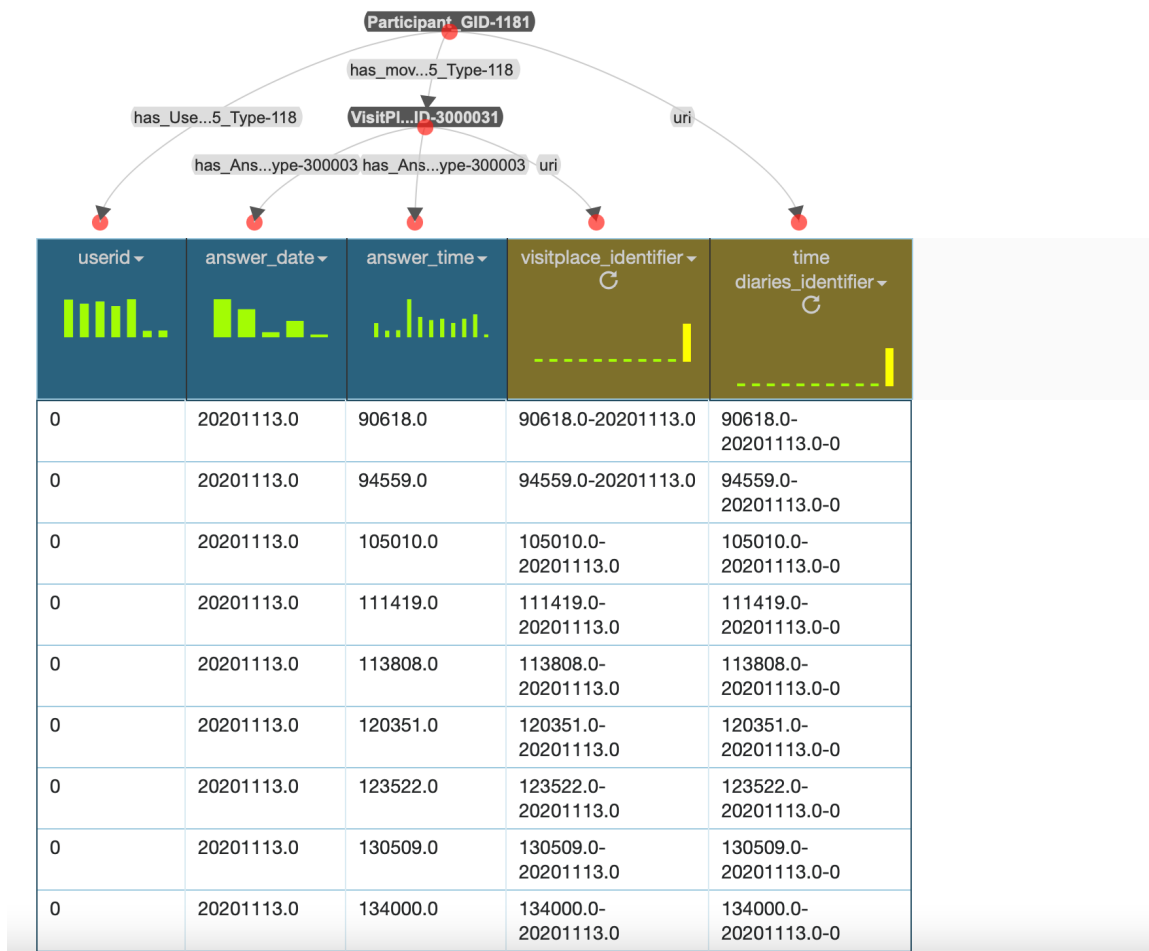




- Timediaries.csv

Data Representation	Class	Property
userID	Participant_GID-1181	has_UserID_GID-39085_Type-118
Answer_date	VisitPlace_GID-3000031	has_Answer_date_GID-80741_Type-300003
Answer_time	VisitPlace_GID-3000031	has_Answer_time_GID-80563_Type-300003





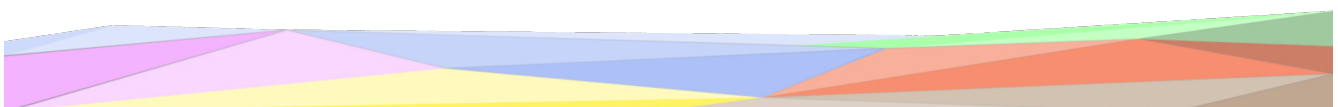
c) Evaluation:

At this phase, we care more about the applicability of this project. So to evaluate this we need to see how many CQs in Inception phase can be answered by the EG. The evaluation will be based on SQL language for example SPASQL. Like we have mentioned in Inception phase, before actual building the CQs, we have a tight connection between Data Scientist and Knowledge Specialist role in order to come up with the best relevant CQs that can be applied to our dataset. So in general, our EG can answer 100% CQs list.

7 Outcome Exploitation

a) Exploitation of KGs

To be able to explore the KGs graph, inside our course we begin to practice and get familiar with GraphDB software with the help of SparQL (Type of SQL language) used to query/visualize graph of KG. For example, if we want to explore all the visit places of a participant inside our dataset, below is the script of SparQL to do that:



Create visual graph config ?

Config name

Place


Starting point

Graph expansion


Node basics

Edge basics


Node extra



Start with a search box
Choose the starting point of your visual graph every time



Start with a fixed node
The visual graph will always start from the chosen node.



Start with graph query results
The results from your query will be the starting point of the visual graph.


1

2

3

```
CONSTRUCT WHERE {  
  ?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
  <http://knowdive.disi.unitn.it/etype#VisitPlace_GID-300003>  
}
```

>>



keyboard shortcuts

Sample queries:

Simple CONSTRUCT
CONSTRUCT or DESCRIBE q...

User queries:

☐ Share visual graph with other users

Save

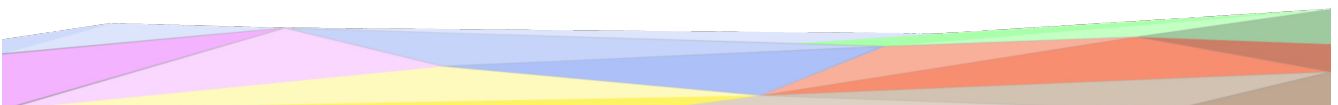
Preview

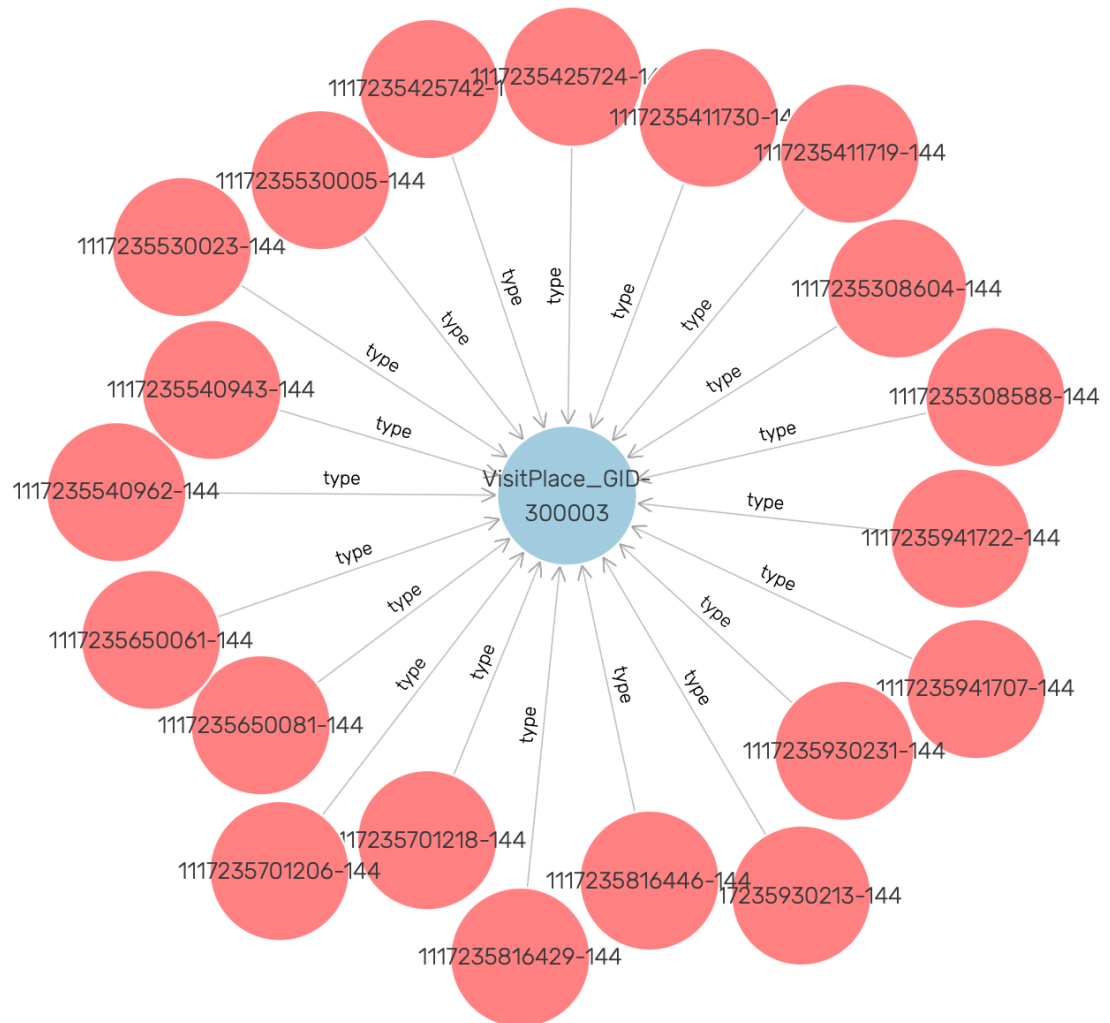
Revert changes

Next

Cancel

After run this query, we will have the graph visualize and from that we can navigate the node by click and drag action (which is provided by GraphDB). Refer to the image below for more detail.





b) Conclusion

The course has brought us an opportunity to actually know and think about the knowledge and data integration part (which we never actually thought about that before). This procedure or more specifically the iTelos method is helpful especially for us who are inside the Computer Science domain to begin to think of preparing data before coding (so that we can reuse the dataset in any project we want). Moreover, we also have a chance to study more about ontology aspects.

About the project, this is a research project so it's really a challenge for us. Despite we finished through all the phases but it's still up to the DISI group to decide whether it's good enough. But in any circumstance, we all hope in some aspect, our KGs are acceptable and can be reusable for further purposes of the DISI group. Despite the time constraint, we managed to finish the project and understand all the theory provided by the professor.

