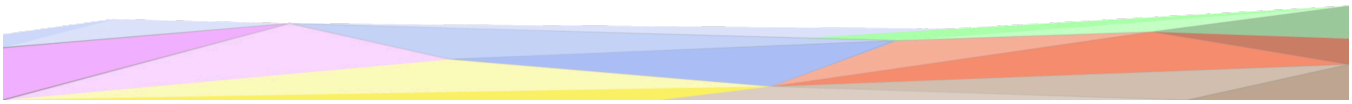
Dipartimento di Ingegneria e Scienza dell'Informazione

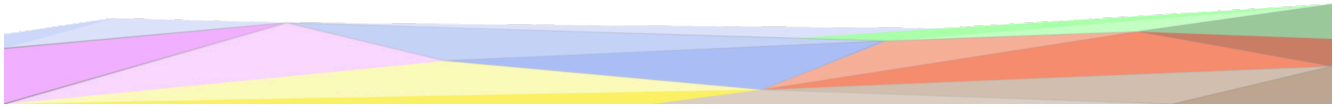
– KnowDive Group –

KGE 2024 - Ecological relations between members of the microbiome

Document Data:	Reference Persons:
November 5, 2024	Eleonora Giuliani, Virginia Leombruni, Marc Shebaby



Index:



1 Introduction

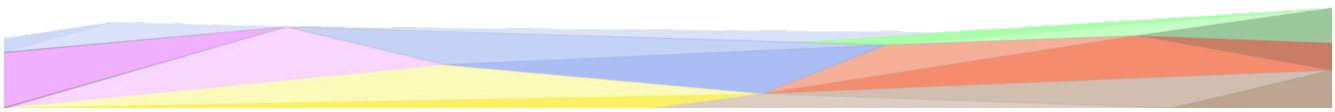
Microbial communities, or microbiomes, represent complex networks of microorganisms that coexist and interact in various environments, from soil and water ecosystems to the human body. These microorganisms, including bacteria, archaea, fungi, and viruses, interact within a complex web of ecological relationships that shape community dynamics and impact the survival of specific species. Through interactions such as competition, cooperation, and even predation, microbes respond to internal and external stimuli—such as changes in nutrient availability, pH, or the introduction of new microbial species. These relationships can significantly affect the microbiome's structure, function, and stability, shaping the presence and abundance of specific species. For instance, some microbes compete for limited nutrients, potentially inhibiting each other's growth, while others participate in syntrophic relationships, where one organism benefits from the by-products of another. This intricate balance of interactions underpins the resilience and adaptability of microbial ecosystems, impacting broader ecological and host-related processes.

To obtain species relative abundance of microbial species several techniques exist such as flow cytometry via microarrays to ribosomal RNA and metagenomic sequencing. Additionally, significant effort is required to cluster the sequences with reference databases to obtain the overall abundances and corresponding taxonomic classifications.

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays a crucial role in enhancing the reusability of the resources handled and produced during the process. A description of the resources and the process developed, provides an understanding of the project, thus serving such information to external readers for the future exploitation of the project's outcomes.

The current document aims to provide a detailed report of the project developed following the iTelos methodology. The report is structured as follows:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: High-level description of the project, based on the Produce role's objectives.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities.
- Section 9: The description of the evaluation criteria and metrics applied to the project.
- Section 10: The description of the metadata produced for all (and all kinds of) the resources handled and generated by the iTelos process, while executing the project.
- Section 11: Conclusions and open issues summary.



2 Purpose Definition

Informal Purpose

The main objective of this project is to construct a Knowledge Graph that reveals new insights into how different microorganisms interact and relate to various diseases across global cohorts, aiming to enhance our understanding of the microbiome's role in human health.

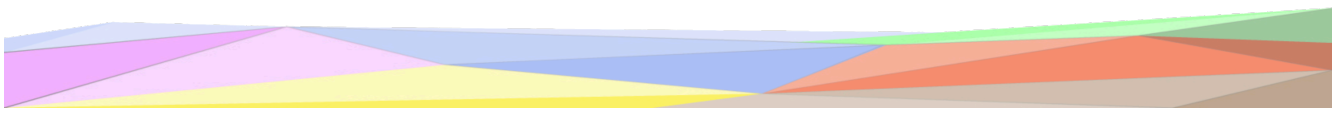
Domain of Interest (DoI)

The main goal of this project is to examine the relative abundances of microbial species in healthy Persons compared to those with diseases, specifically focusing on tumours and cancers such as colorectal cancer. This is achieved by exploiting the curatedMetagenomicData (cMD) package in R, which contains curated datasets from several different independent studies conducted across different periods.

The cMD package provides microbiome data, and for each collected sample, it includes information about gene families, marker abundance, marker presence, pathway abundance, pathway coverage, and relative abundance.

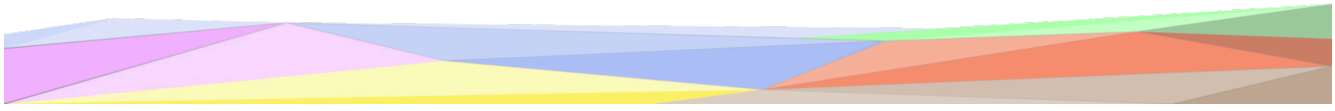
Moreover, this analysis considers several critical factors, including individual characteristics (e.g., smoking and alcohol consumption levels), microbial taxonomy, and disease location. This comprehensive approach allows for a clearer depiction of microbial relative abundances, facilitating insights into both inter-species relationships (species-species interactions) and intra-species connections (species-disease associations or species-epidemiological factor correlations).

Due to the study’s purpose we only use the relative abundance information from the package.



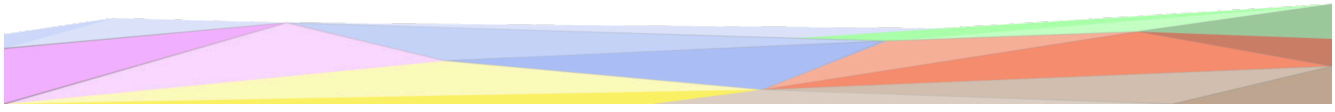
Scenarios definition: a set of usage scenarios, describing the multiple aspects considered by the project purpose.

- **Species Interactions (S1)**: Understanding the interactions among microbial species remains a challenging aspect of research. This Knowledge Graph facilitates the process, by mapping the complex relationships between microorganisms, shedding light on how they influence each other's presence, abundance, and roles. The graph enhances our ability to detect underlying ecological dynamics, cooperation, competition, and other interspecies interactions that may impact microbial community stability.
- **Microbiome epidemiology (S2)**: Microbiome epidemiology examines how population-level trends, such as geography, and socioeconomic factors, correlate with microbial profiles and disease outcomes. In this context, the Knowledge Graph is crucial as it allows scientists to analyze and integrate diverse datasets, revealing patterns of microbial variation across populations and uncovering potential links between specific microbial compositions and health conditions. Thus, researchers can identify broader trends that suggest how certain microbes or microbial communities might be protective or detrimental to health at a population level.
- **Microbiome risk factor analysis (S3)**: This field focuses on person-specific risk factors—such as smoking, alcohol use, diet, and medication—that may influence the microbiome and impact disease susceptibility. The Knowledge Graph plays a vital role by linking microbial data with these specific risk factors, helping researchers observe how shifts in the microbiome might correlate with lifestyle or environmental exposures. It is known that observing such shifts concerning risk factors could act as early indicators of cancer or other diseases and be beneficial to the development of targeted approaches.



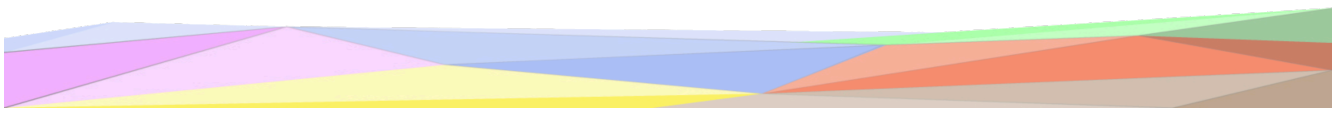
Personas: a set of real users acting within the scenarios defined above. Each Persona is defined over a specific feature included in the main Purpose.

- **Haruto (P1)**: is a 68-year-old man living in Japan. He has been smoking since he was a teenager. His long history of smoking may expose him to health risks that could impact his microbiome.
- **Yuki (P2)** is 66 and lives in Japan. He has been overweight since a young age. He has recently noticed that he feels more fatigued than usual and has started losing weight without changing his diet.
- **Emiko (P3)**: is a 65-year-old woman living in the Japanese countryside. She is thin and very active, indeed, she enjoys running. Nevertheless, she is quite hypochondriacal and anxious about her health. She has decided to see a doctor for a diagnosis.
- **Riku (P4):** is a 72-year-old man living in Japan. A month ago, during some routine tests, he was diagnosed with stage IV cancer.
- **Kaito (P5)**: is a 61-year-old man. Yesterday, he noticed some blood in his stool. Feeling scared, he immediately scheduled an appointment for a stool exam.
- **Aiko (P6)**: is a 38-year-old single woman who used to go out and drink at parties. In recent weeks, she has noticed that she has lost some weight.
- **Hana (P7):** is a 54-year-old woman in good health. She enjoys an active lifestyle, often spending time outdoors and engaging in various hobbies. With a strong appreciation for nature, she loves exploring the local flora and fauna.
- **Ren (P8)**: is 56 years old. From a stool test, he was diagnosed with a high abundance of Helicobacter pylori. In the past, he had only heard about this bacteria. He is wondering if it can affect his health and predispose him to CRC.
- **Daiki (P9)**: is 81 years old. He was diagnosed with stage III colorectal cancer (CRC) and has been in remission for 5 years. To reduce the probability of a recurrence, he is committed to maintaining a healthy diet and a balanced microbiome.
- **Sakura (P10):** is 74 years old. She has a diagnosis of stage II colorectal cancer (CRC). In her last lab analysis, a high abundance of E. coli was also detected.



Competency Questions (CQs): the list of CQs created considering the personas in the scenarios defined.

- **CQ-1 (P1-S1)**: Haruto is curious about how different microbes in his body interact and whether these interactions could impact his health, especially given his recent stage I cancer diagnosis and his long history of smoking. What are the main microbial interaction patterns observed in smokers? Are there high or low-level abundances of certain species that correlate with smoking habits?
- **CQ-2 (P2-S2)**: Yuki doesn't understand why he is finally losing weight. Could he be facing an imbalance of the microbiome that impacted his weight? Are there specific correlations between low or high levels of certain microbiome species and weight loss? Does this increase or decrease his probability of developing CRC?
- **CQ-3 (P3-S3)**: Given that Emiko is a healthy woman, what can be observed in the relative abundances of her microbiome? Can a particular pattern from the latter contribute to her risk of developing CRC?
- **CQ-4 (P4-S1)**: Given Riku's stage IV cancer diagnosis, which microbial species are characteristic of advanced CRC patients? Are there patterns associated with this stage?
- **CQ-5 (P5-S3)**: After noticing blood in his stool, Kaito is anxious about his health. What is the relative abundance of the microbiome of Kaito considering his age, gender, and the assumption that he is healthy? Can we infer if he has a high chance of developing gastrointestinal problems, specifically CRC?
- **CQ-6 (P6-S2)**: How does Aiko's alcohol consumption influence the composition and abundance of bacteria in her microbiome? What effect might this have on her risk of CRC?
- **CQ-7 (P7-S1):** Hana, in good health, wonders about positive co-occurrence relationships in her microbiome. Which species demonstrate co-occurrences with high relative abundances greater than 0.621215?
- **CQ-8 (P8-S2)**: Ren is curious about his high levels of *Helicobacter pylori*. How does its abundance correlate with other microbial species, and could it affect his gut health and CRC risk?
- **CQ-9 (P9-S3)**: Daiki wants to compare his microbiome composition with healthy individuals. Which changes in his microbiome should he focus on to reduce his cancer recurrence risk?
- **CQ-10 (P10-S2)**: How does Sakura's high *E. coli* abundance correlate with other bacterial species in CRC patients? Which species show an anti-correlation with *E. coli*?



General CQs:

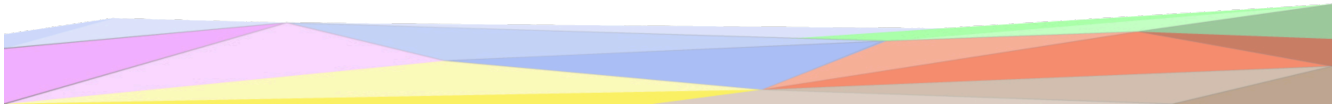
- Which species are associated with the CRC?
- How is a specific species' high or low abundance in the microbiome associated with CRC risk?
- Which species co-occur in Persons with CRC?
- How does the microbiome composition differ between healthy Persons and diseased those affected by CRC?
- What is the relationship between the abundance of bacteria and gender, age, BMI smoking, and alcohol consumption habits?
- Which species show a positive relationship and, as a consequence, a similar abundance pattern in Persons with CRC? Which species exhibit an anti-correlation?

Concepts identification:

Starting from the formulated CQs that combined each persona with specific scenarios, it is now possible to identify the entities of our ER model. These entities are categorized into common contextual and core entities.

Entities:

- Persons: This entity describes the specific characteristics of each participant from whom the samples are collected.
- Microbiome composition: This entity describes the unique and diverse microbial composition of each Person.
- Stage 1: first stage of CRC. It is the initial stage, limited to a specific tissue.
- Stage 2: second stage of CRC. Cancer has increased in size and may involve nearby tissue.
- Stage 3: third stage of CRC. Cancer cells spread to lymph nodes.
- Stage 4: fourth stage of CRC. Advanced stage with metastasis in different parts of the body.
- Healthy: Represents individuals without cancer.
- Risk factors: There are certain habits and personal characteristics that can influence cancer development.
- Correlation: The entity “correlation” is fundamental for identifying a link between species and a person's health condition. It can be classified into three levels: High, Medium and Low level.



Attributes

- Person:
  - *Subject id (SRR)*: identifier code unique for each Person
  - *Age*: specific age of the Person
  - *Age category*
  - *Gender*: Can be male or female
  - *Country*: Ethnicity of the Person
  - Study: Refers to the specific research study that has been considered
- Stage 1, 2, 3, 4:
  - *Disease location*: The site of the body where the cancer is developed
- Risk factors:
  - *BMI*: Body Mass Index
  - *Smoking habits (Brinkman index)*: correlated to smoking exposure. It is calculated as the product of the number of cigarettes per day and years of smoking
  - *Alcohol habits (alcohol numeric)*: measures the Person's alcohol intake. It is computed by multiplying the weekly number of drinks by the units of alcohol per drink
- Microbiome composition:
  - *Taxonomy*:

    each species is defined based on the following hierarchical structure:
    Kingdom (k__), Phylum (p__), Class (c__), Order (o__), Family (f__), Genus (g__), Species (s__)
- Correlation:
  - *Relative abundance*: expresses the proportion of each species within each Person.
  - *Presence/absence*: a binary variable indicating if a species is present or absent in a sample.

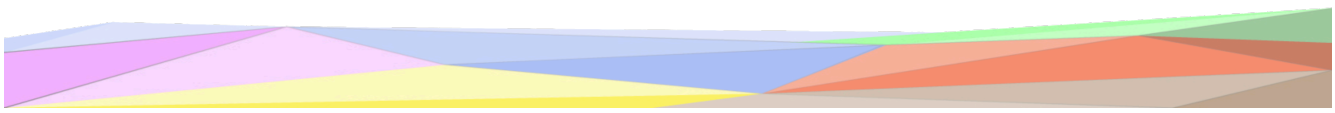
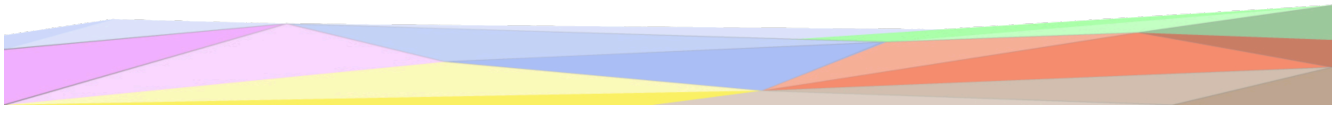
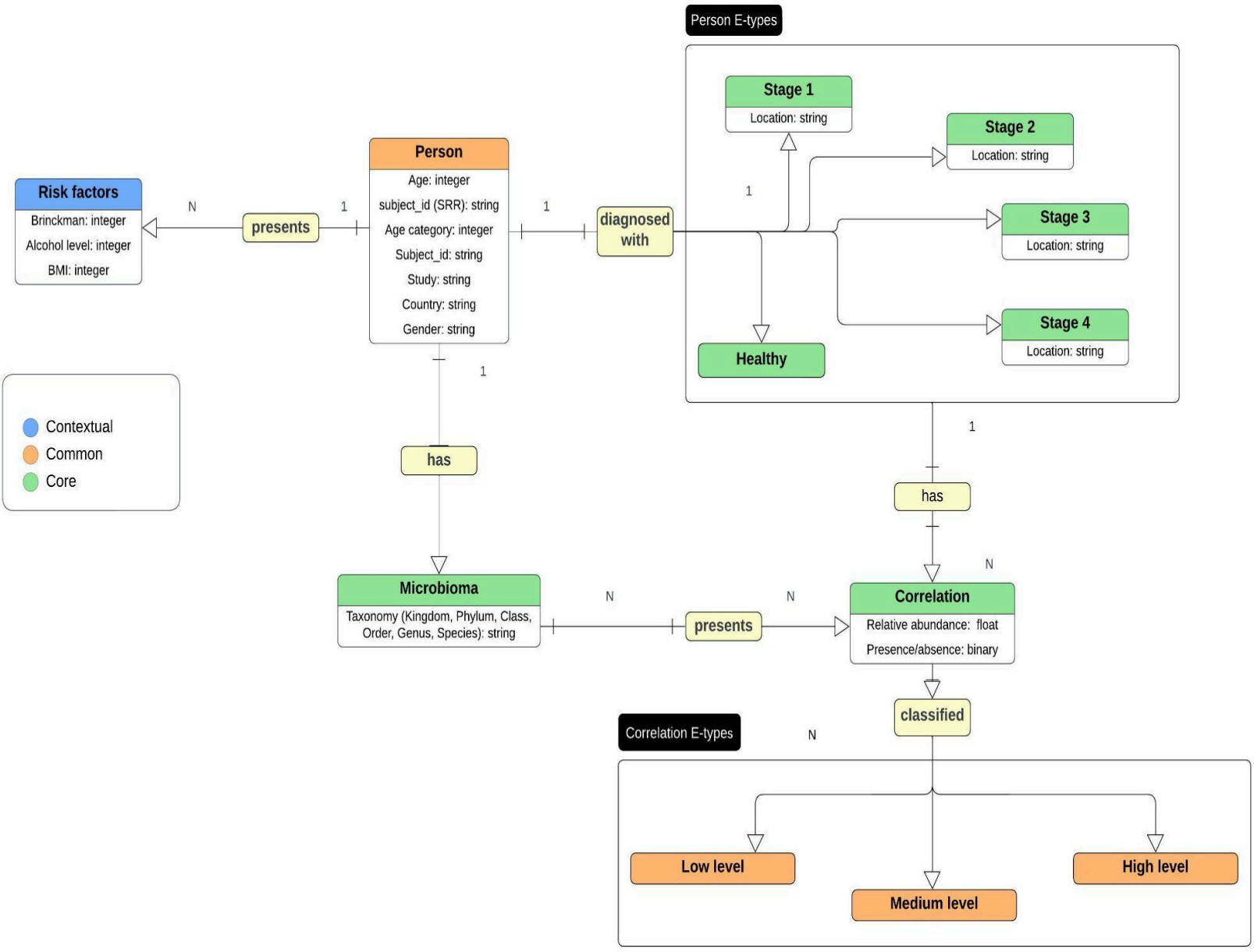


Table: Extraction of Entities Based on CQs and Focus/Popularity Classification

Scenario	Persona	CQs	Entities	Properties	Focus
S1, S2, S3	P1 - P9	1 - 9	Person	ID, age, age category, country, gender, study	Common
S1, S2, S3	P1 - P9	1 - 9	Microbiome	Taxonomy	Core
S1, S2, S3	P1, P2, P3	1, 2, 3	Risk Factors	Smoking habits, BMI, Alcohol habits	Contextual
S1, S2, S3	P1, P2, P8	1, 2, 8	Correlation	Relative abundance, presence/absence	Core
S1, S3	P3, P5, P7	3, 5, 7	Health		Core
S1	P1	1	Stage 1	Location	Core
S2	P10	10	Stage 2	Location	Core
S3	P9	9	Stage 3	Location	Core
S1	P4	4	Stage 4	Location	Core



The ER Model


Person E-types
Stage 1
Location: string
Stage 2
Location: string
Stage 3
Location: string
Stage 4
Location: string
Healthy
Risk factors
Brinckman: integer
Alcohol level: integer
BMI: integer
presents
Person
Age: integer
subject_id (SRR): string
Age category: integer
Subject_id: string
Study: string
Country: string
Gender: string
diagnosed with
Contextual
Common
Core
has
has
Microbioma
Taxonomy (Kingdom, Phylum, Class, Order, Genus, Species): string
presents
Correlation
Relative abundance: float
Presence/absence: binary
classified
Correlation E-types
Low level
Medium level
High level
1
N

The ER diagram represents two core entities: **Person** and **Microbiome**. A relationship labelled **“has”** exists between them, directed from **Person** to **Microbiome**.

The **Person** entity is divided into several entity types (e-types) based on cancer diagnosis stages for carcinoma (Stages 1, 2, 3, and 4). While this project utilizes only one study, this structure can accommodate multiple studies by associating each person with a specific disease to form a unique e-type.

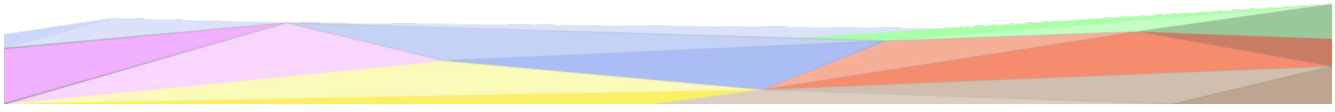
Each **Person** is also associated with a **Risk Factor** entity, which includes attributes such as **BMI, Brinkman’s Index,** and **Alcohol Level**. The **Risk Factor** entity provides additional context, potentially revealing patterns in the data that could address some of the competency questions outlined in the project.

The **Microbiome** entity represents the various species present within a person’s microbiome. Each species has a **relative abundance** value, allowing correlation with the **Person** entity. Correlations between species and persons are categorised into **low, medium,** and **high** levels, depending on the relative abundance of a species in a given individual. Specifically:

- **High level** is designated if the species' abundance for a person is at or above the third quartile (Q3).
- **Low level** occurs if the abundance is at or below the first quartile (Q1).
- **Medium level** represents values between Q1 and Q3.

To determine these correlation categories, the relative abundance values are discretized based on the median, Q1, and Q3 thresholds across all species (where, for example, median = 0.09878, Q1 = 0.01567, and Q3 = 0.621215).

This correlation classification supports the analysis of species interactions. Species with high level of correlations within a particular e-type of **Person** are assumed to exhibit **mutualistic interactions**, while those with lower level of correlations are assumed to have **competitive interactions**. Ultimately, machine learning models can be constructed to infer these interactions further and assess the completeness of this knowledge graph.



3 Information Gathering

In this section, the second main input for the project is described, namely the data source list (if available). The resources (language, schema, and data values) available as input for projects, have to be properly described. More details for each resource have to be reported:

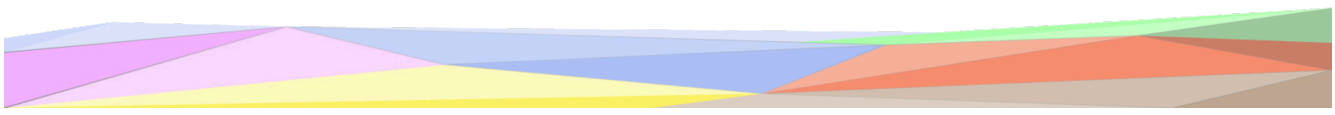
- The name, and the description of the information the resource is carrying.
- Type of resource. If it is a language, schema, or data value dataset.
- The source from which such resources can be collected.
- If the resource is diversity-aware (thus already produced by iTelos) or needs to be improved in terms of diversity (i.e., data coming from low-quality sources).

Moreover, this section aims at reporting the execution of the activities involved in the Information Gathering iTelos phase.

Information Gathering sub-activities:

- Sources identification
- Datasets collection
- Datasets cleaning
- Datasets standardization

The report of the work done during the first phase of the methodology has to include also a description of the different choices made, with their strong and weak points. In other words, the report should provide the reader, with a clear description of the reasoning conducted by all the different team members.



4 Language Definition

This section is dedicated to the description of the Language Definition phase. Like in the previous section, it aims to describe the different sub-activities performed by all the team members, as well as the phase outcomes produced.

Language Definition sub-activities:

- Concept identification
- Dataset filtering

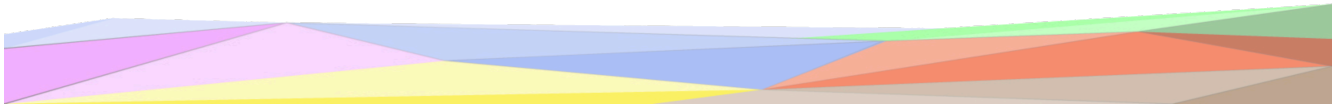
The report of the work done during this phase of the methodology has to include also a description of the different choices made, with their strong and weak points. In other words, the report should provide the reader, with a clear description of the reasoning conducted by all the different team members.

5 Knowledge Definition

This section is dedicated to the description of the Knowledge Definition phase. Like in the previous section, it aims to describe the different sub-activities performed by all the team members, as well as the phase outcomes produced.
Knowledge Definition sub-activities:

- KTelos
  - Teleology definition
  - Teleontology definition
- Dataset cleaning and formatting

The report of the work done during this phase of the methodology has to include also a description of the different choices made, with their strong and weak points. In other words, the report should provide the reader, with a clear description of the reasoning conducted by all the different team members.



6 Entity Definition

This section is dedicated to the description of the Entity Definition phase. Like in the previous section, it aims to describe the different sub-activities performed by all the team members, as well as the phase outcomes produced.
Entity Definition sub-activities:

- Entity matching
- Entity identification
- Data mapping

The report of the work done during this phase of the methodology has to include also a description of the different choices made, with their strong and weak points. In other words, the report should provide the reader, with a description of the reasoning conducted by all the different team members.

7 Evaluation

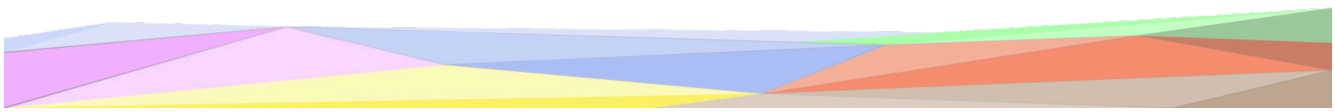
This section aims to describe the evaluation performed at the end of the whole process over the outcome of the iTelos methodology. More in detail, this section is to report:

- the final Knowledge Graph information statistics (like, the number of types and properties, number of entities for each etype, and so on).
- Knowledge layer evaluation: the results of the application of the evaluation metrics applied over the knowledge layer of the final KG.
- Data layer evaluation: the results of the application of the evaluation metrics applied over the data layer of the final KG.
- Query execution: the description of the competency queries executed over the final KG to test the suitability of the KG to satisfy the project purpose.

8 Metadata Definition

In this section, the report collects the definitions of all the metadata defined for the different resources produced along the whole process. The metadata defined in this phase describes both the outcome of the project, and the intermediate outcome of each phase (language, schema, and data source standardised values).

The definition of metadata is crucial to enable the distribution (sharing) of the resource



produced, through the data catalogs. For this reason, it is important to describe also where such metadata will be published to distribute the resources it describes (for example the DataScientia catalogs).

In particular, the structure of this section is organized as follows, to describe the metadata relative to all the types of resources produced by the project.

- Project metadata description
- Language resources metadata description
- Knowledge resources metadata description
- Data resources metadata description

9 Open Issues

This section concludes the current document with conclusions regarding the quality of the process and outcome, and the description of the issues that (for lack of time or any other cause) remained open.

- Did the project respect the scheduling expected in the beginning?
- Are the final results able to satisfy the initial Purpose?
    - If no, or not entirely, why? Which parts of the Purpose have not been covered?

Moreover, this section aims to summarize the most relevant issues/problems remaining open along the iTelos process. The description of open issues has to provide a clear explanation of the problems, the approaches adopted while trying to solve them and, eventually, any proposed solution that has not been applied.

- Which issues remained open at the end of the project?