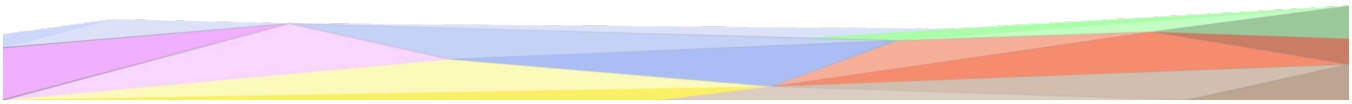
Dipartimento di Ingegneria e Scienza dell'Informazione

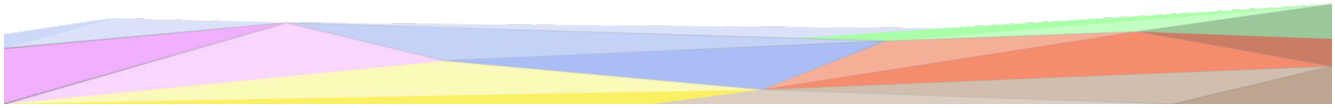
– KnowDive Group –

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

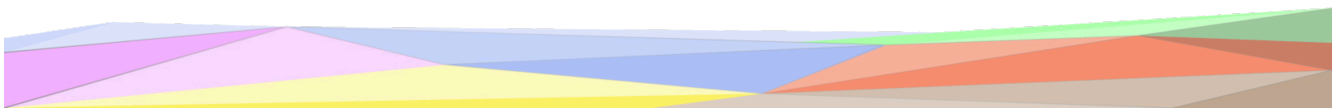
Microbial communities, or microbiomes, represent complex networks of microorganisms that coexist and interact in various environments, from soil and water ecosystems to the human body. These microorganisms, including bacteria, archaea, fungi, and viruses, interact within a complex web of ecological relationships 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.

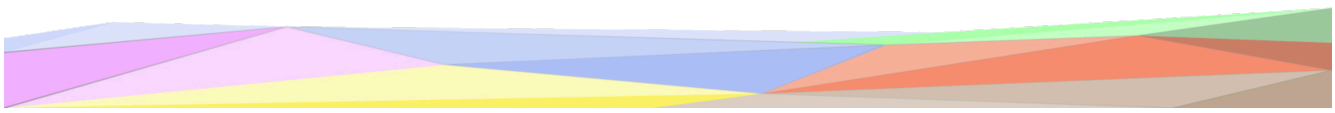
Domain of Interest (DoI)

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.

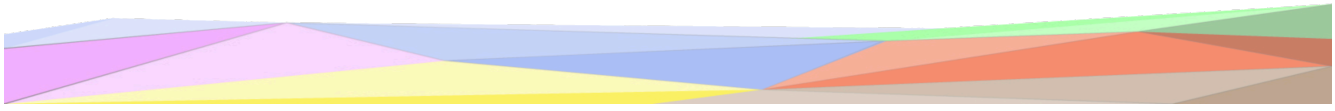
Moreover, this analysis considers several critical factors, including individual characteristics (e.g., smoking and alcohol consumption levels), microbial taxonomy, and disease location. This comprehensive approach allows for a clearer depiction of microbial relative abundances, facilitating insights 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.



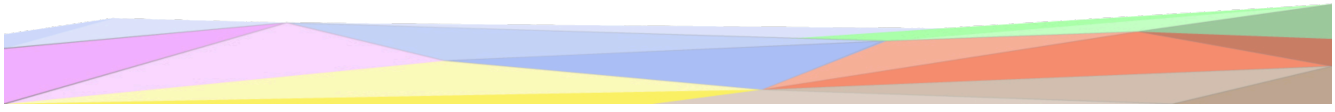
Scenarios definition: a set of usage scenarios, describing the multiple aspects considered by the project purpose.

- **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.



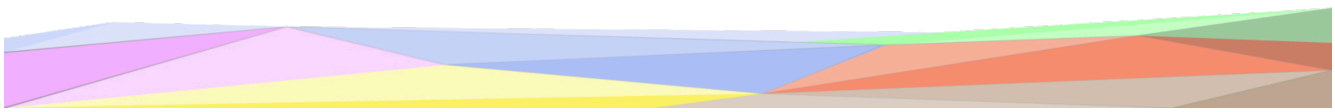
Personas: a set of real users acting within the scenarios defined above. Each Persona is defined over a specific feature included in the main Purpose.

- **Haruto (P1)**: is a 68-year-old man. 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 Cancer.
- **Daiki (P9)**: is 81 years old. He was diagnosed with stage III colorectal cancer 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. 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 cancer?
- **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 cancer?
- **CQ-4 (P4-S1)**: Given Riku's stage IV cancer diagnosis, which microbial species are characteristic of advanced patients with cancer? 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 cancer?
- **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 cancer?
- **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 cancer 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 cancer patients? Which species show an anti-correlation with *E. coli*?

General CQs:

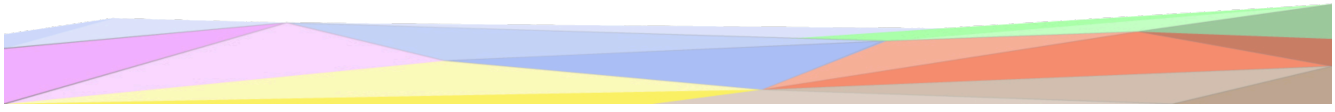
- Which species are associated with cancer?
- How is a specific species' high or low abundance in the microbiome associated with cancer risk?
- Which species co-occur in Persons with cancer?
- How does the microbiome composition differ between healthy Persons and diseased those affected by cancer?
- 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 cancer? 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 cancer. It is the initial stage, limited to a specific tissue.
- Stage 2: second stage of cancer. Cancer has increased in size and may involve nearby tissue.
- Stage 3: third stage of cancer. Cancer cells spread to lymph nodes.
- Stage 4: fourth stage of cancer. 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.

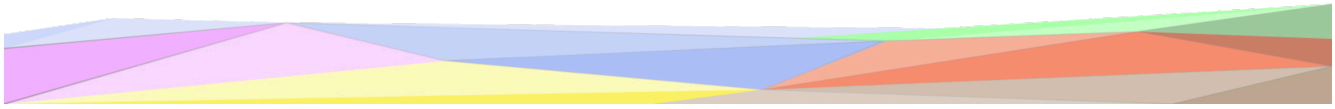
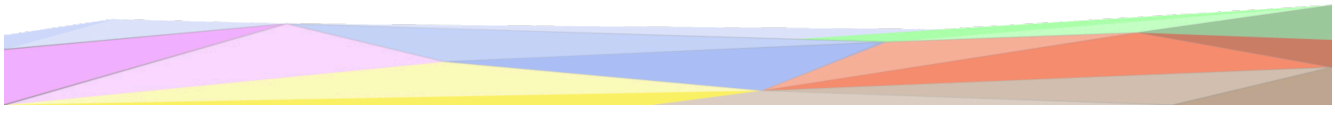
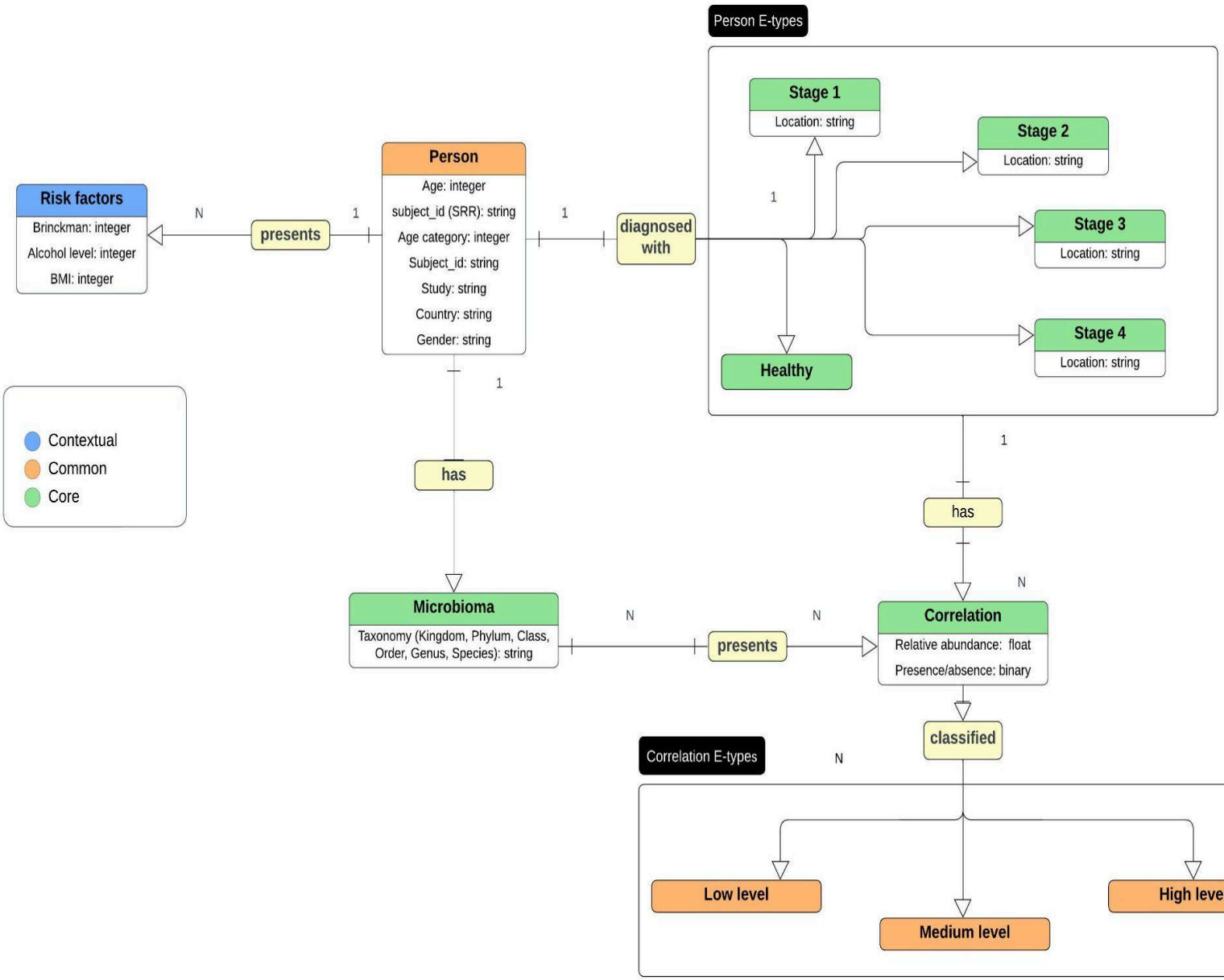


Table: Extraction of Entities Based on CQs and Focus/Popularity Classification

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

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.

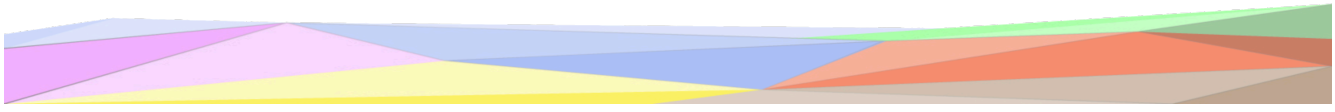
Each **Person** is also associated with a **Risk Factor** entity, which includes attributes such as **BMI, Brinkman’s Index,** and **Alcohol Level**. The **Risk Factor** entity provides additional context, potentially revealing patterns in the data that could address some of the competency questions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

- A **high level** is designated if the species' abundance for a person is at or above the third quartile (Q3).
- A **low level** occurs if the abundance is at or below the first quartile (Q1).
- A **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 levels of correlations within a particular e-type of **Person** are assumed to exhibit **mutualistic interactions**, while those with lower levels 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

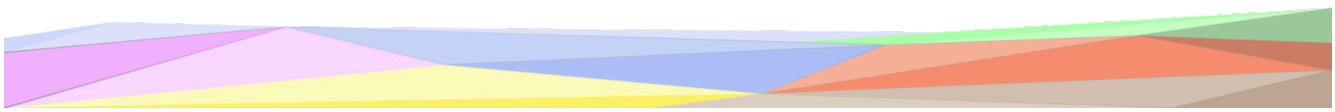
After completing the first phase of the project, which involved defining the objectives and building the ER model, we now move on to the information-gathering phase. In this second phase, we identify and accurately describe the sources and structure of the available data while performing a thorough cleaning and filtering process. The goal is to produce standardized datasets ready for the subsequent stages of analysis.

Sources Identification

For this project, the data source used is the R package *CuratedMetagenomicData* (CMD). This package provides curated and standardized human microbiome data, useful for innovative analyses. It also includes a range of tools that provide information on gene families, marker abundance, marker presence, pathway abundance, pathway coverage, and relative abundance for samples collected from various body sites. The taxonomic abundances of bacteria, fungi, and archaea for each sample were calculated using *MetaPhlAn3*, while the metabolic functional potential was determined with *HUMAnN3*. The manually curated sample metadata and standardized metagenomic data are made available as (Tree)*SummarizedExperiment* objects. Additionally, specific metadata for each sample is collected based on the study's objectives. In total, the package includes over 26 studies comprising 5716 samples and 34 diseases. *Table 1* shows the different types of studies available in the package, along with their respective properties.

Dataset Name	Body Site	Disease	# Total Samples	# Case Samples	Average Reads per Sample (std) (M)	Size (Tb)	# Reads (G)	Reference
AsnicarF_2017	Stool, milk	None	26	-	21.4 (19.8)	0.2	0.5	7
BritoIL_2016	Stool, oral	Other condition	312	-	67.4 (51.8)	5.6	21.0	8
Castro-NallarE_2015	Oral	Schizophrenia	32	16	61.0 (25.2)	0.5	2.0	9
ChngKR_2016	Skin	Atopic dermatitis	78	38	15.8 (7.5)	0.3	1.2	10
FengQ_2015	Stool	Colorectal cancer	154	93	53.8 (8.5)	2.3	8.3	11
Heitz-BuschartA_2016	Stool	Type 1 diabetes	53	27	44.5 (0.9)	0.5	2.4	12
HMP_2012	Several	None	749	-	51.5 (44.8)	9.4	38.6	4
KarlssonFH_2013	Stool	Type 2 diabetes	145	53	31.0 (17.6)	1.4	4.5	13
LeChatelierE_2013	Stool	Obesity	292	169	69.0 (23.2)	4.0	20.1	14
LiuW_2016	Stool	Other condition	110	-	58.3 (26.8)	1.8	6.4	15
LomanNJ_2013	Stool	Shiga-toxigenic *E. coli*	43	43	9.2 (12.1)	0.1	0.4	16
NielsenHB_2014	Stool	Inflammatory bowel diseases	396	148	53.9 (20.2)	3.5	21.4	17
Obregon-TitoAJ_2015	Stool	Other condition	58	-	47.1 (20.9)	0.6	2.7	18
OhJ_2014	Skin	None	291	-	24.7 (38.1)	2.2	7.2	19

Table 1 *- Datasets available in the package*



Datasets Collection

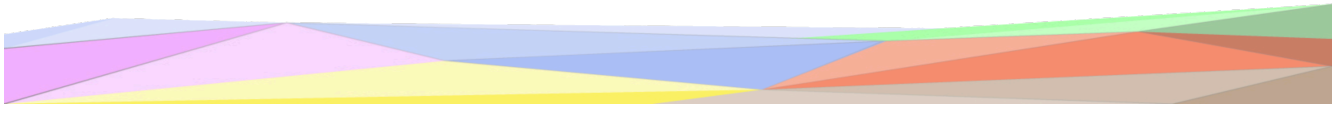
For the purposes of this study and to simplify the analysis, we selected the dataset **2021-10-14.YachidaS_2019**, which focuses on the disease 'Carcinoma Cancer' and contains data on 712 species and 616 samples. We extracted the relative abundance data, which indicates the quantity of each species found in each sample while excluding unrelated data on genetic markers. Additionally, we retrieved the sample metadata and converted the datasets into CSV files to ensure the data's integrity and quality. The conversion process is illustrated in the code below, which generates two CSV files. Moreover, we also created a function in R to map the species names with their taxonomy and added the latter as a new field in the CSV file.

Code

```
library(curatedMetagenomicData)
#data_check_all=curatedMetagenomicData(".*.relative_abundance",dryrun = FALSE, rownames = "short")
data <- curatedMetagenomicData("YachidaS_2019.relative_abundance",dryrun = FALSE, rownames = "short")

relative_abundance <- as.data.frame(data[[1]]@assays@data@listData[["relative_abundance"]])


person_data <- as.data.frame(data[[1]]@colData@listData)


relative_abundance$taxonomy <- matched_species

#write relative abundance to csv
write.csv(relative_abundance, "relative_abundance.csv")
#write person data to csv
write.csv(person_data, "person_data.csv")
```

The first file contains all the relevant metadata associated with each sample, providing detailed information about their characteristics. This metadata includes key variables such as the study name, subject ID, body site, study condition, age, gender, BMI, and disease location, among others. These attributes offer important contextual insights that allow us to better understand the conditions under which the samples were collected and the individual features of each sample.

The second file, on the other hand, outlines the relative abundance of each species across the samples, offering a clear representation of the species distribution within the dataset. This file provides data on the specific quantities of various microbial species found in each sample, helping to reveal patterns and variations in microbial composition across different conditions or patient groups.



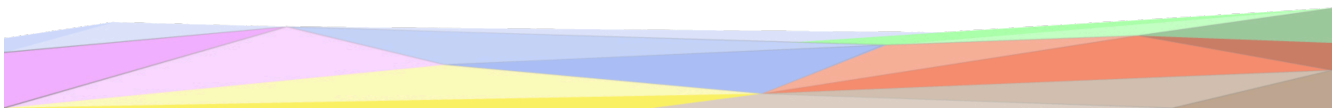
Together, these files form the essential data foundation that will support and guide the subsequent stages of our analysis. By combining detailed metadata with species abundance data, we ensure that the analysis is both comprehensive and contextually grounded, allowing for meaningful interpretations of the microbial data concerning the study's objectives. *Table 2* summarizes the two key files generated in the information-gathering phase of the project.

CSV file	Columns
Metadata	study_name, subject_Id, body_site, study_condition, diseases, age, age_category, gender, country, non_westernized, sequence_platform, PMID, number_bases, minimum_read_length, median_read_length, curator, BMI, disease_location, ajcc, brinckman_index, alcohol_numeric
Relative abundance	samples

Table 2 *- Metadata and Relative Abundance Files*

Data cleaning and standardization

Once the data were collected and displayed, the next step involved filtering and retaining only the information that was relevant to the specific purpose of this study. This process ensured that the dataset was focused and manageable, eliminating any unnecessary details that could introduce noise into the analysis. As part of the cleaning procedure, certain columns from the metadata file were removed, as they were not essential for the current analysis. Additionally, instead of discarding the NA (missing) values for the metadata attributes, we decided to preserve them in case the data was reused and updated in subsequent studies. By retaining these values, we maintain the flexibility to incorporate additional data or refine the metadata as new information becomes available. *Table 3* summarizes the metadata columns retained after cleaning, highlighting key attributes relevant to this study while preserving NA values for potential future data integration.



CSV file	Columns
Metadata	study_name, sudject_id, body_site, diseases, age, age_category, gender, country, non_westernized, BMI, disease_location, ajcc, brinckman_index, alcohol_numeric

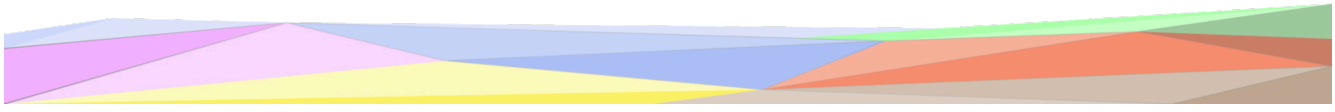
Table 3 *- Metadata Columns Post-Cleaning and Standardization*

Schema Generation

In the Schema Generation phase, we define a conceptual and logical structure to organize collected data for effective analysis. The objective is to create a schema that represents the relationships between various variables, ensuring both data and metadata align with the project's analytical goals. The objects and their properties for the schema are already reported previously in the ER description, so we expect the data layer of the knowledge graph to be composed of nodes representing the instantiation of the objects (person and species) and the edges that model the relationship between them. Additionally, our schema includes an event type, *Correlation*, represented as a node that links a *Person* with *Species*. This *Correlation* event node holds properties that classify whether a species has a high association with a disease, providing a basis for inferring inter-species relationships through methods like Pearson correlation. Therefore, representing our knowledge graph in this way facilitates interoperability and enables data reuse for future studies.

Formal resource generation

During the Formal Resource Generation phase of the project, we focused on creating and organizing the formal resources necessary for the subsequent stages of the analysis. These resources include the finalized datasets, data processing and analysis scripts, and detailed documentation that ensures consistency and reproducibility throughout the entire analysis process.

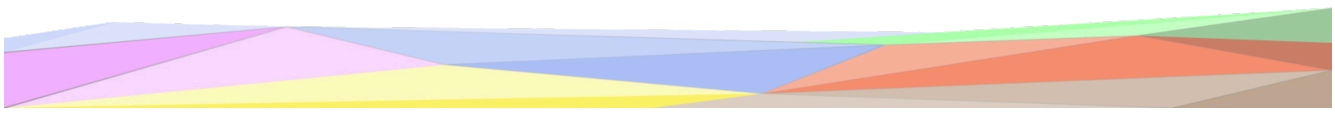


4 Language Definition

Until now, our purpose formulation has been carried out without a clearly defined language resource. This presents a significant challenge, as the words or concept structures used to represent the Etypes and their properties can be interpreted differently by users. Furthermore, some words are polysemous, meaning they have multiple meanings, which makes them ambiguous. Therefore, it is crucial to address this linguistic diversity by associating each concept with a formal definition. This can be effectively achieved using the Universal Knowledge Core (UKC), a high-quality, diversity-aware database.
Following the iTelos methodology, the concepts to be identified should initially represent the Etypes, object properties, and data properties. These elements were already established during Phase 1, resulting in the creation of a CSV file that serves as a language dataset. This dataset contains formally defined concepts tailored to our purpose-specific domain.

Three different tables are presented below: the first contains the entity types, the second the properties, and the third the relationships. For each concept, an ID has been assigned. Most of these concepts are found in the UKC ontologies. At the same time, for more biological terms, such as *Microbiome* or *Relative Species Abundance*, it was possible to locate them in other ontologies available on BioPortal, with the specific link provided. However, some concepts could not be found in any external ontology and have been identified using the code *KGE-QCB1-number*.

For the term 'Brinkman Index,' while some similar concepts related to smoking habits were found in biological ontologies (such as the number of cigarettes smoked per person), we decided to create a new ID. This is because the Brinkman Index is calculated specifically as the product of the number of cigarettes smoked per day and the number of years of smoking, which makes it distinct from other related concepts.

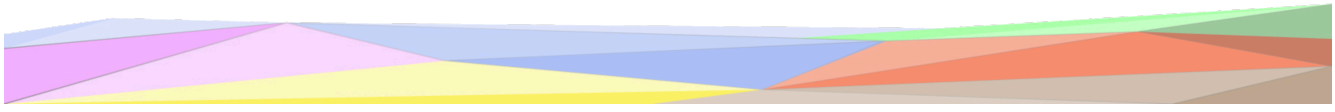


Language concepts for e-types

ConcetID	Word-en	Gloss-en
UKC-36	Person	A human being
KGE_QCB1-1	Risk factor	Something that makes a person more likely to get a particular disease or condition
http://purl.obolibrary.org/obo/OHMI_0000003	Microbiome	A biome that consists of a collection of microorganisms (i.e., microbiota) and the surrounding environment where the microorganisms reside.
UKC-43176	Species	A taxonomic group whose members can interbreed (biology)
UKC-65892	Correlation	A reciprocal relation between two or more things
UKC-67961	Cancer	Any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream
UKC-27611	Stage	A position on a scale of intensity, amount or quality
UKC-80514	Healthy	Having or indicating good health in body or mind; free from infirmity or disease

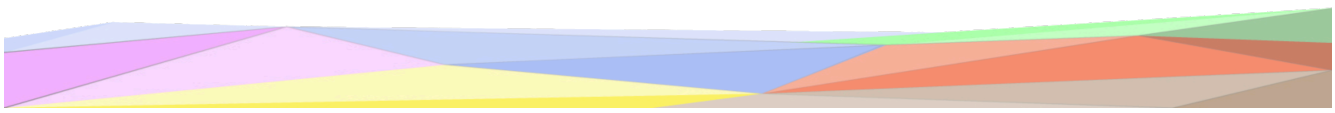
Language concepts for e-types relations (object properties)

ConcetID	Word-en	Gloss-en
UKC-681	has_Medical_diagnosis	Identification of a disease from its symptoms
KGE-QCB1-2	has_Species	A person has a taxonomic group whose members can interbreed.
KGE-QCB1-3	has_Interaction	A species correlates with a particular person
KGE-QCB1-4	has_Risk_Factor	A person is associated with risk factors.



Language concepts for e-types attributes (data properties)

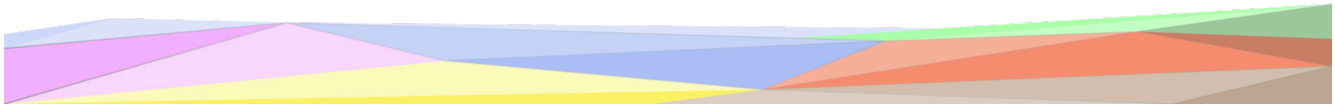
ConcetID	Word-en	Gloss-en
UKC-44477	Taxonomy	A classification of organisms into groups based on similarities of structure or origin etc
UKC-42545	Kingdom	The highest taxonomic group into which organisms are grouped; one of five biological categories: Monera or Protoctista or Plantae or Fungi or Animalia
UKC-43156	Phylum	The major taxonomic group of animals and plants; contains classes (biology)
UKC-43160	Class	A taxonomic group containing one or more orders (biology)
UKC-43163	Order	A taxonomic group containing one or more families (biology)
UKC-43171	Genus	A taxonomic group containing one or more species (biology)
UKC-26728	Age	How long something has existed
UKC-27174	Gender	The properties that distinguish organisms based on their reproductive roles
UKC-45187	Country	The territory occupied by a nation
http://purl.bioontology.org/ontology/MESH/D015992	BMI	An indicator of body density as determined by the relationship of BODY WEIGHT to BODY HEIGHT. BMI=weight (kg)/height squared (m2).
KGE-QCB1-5	Brinkman Index	Is calculated from cigarettes per day times smoking years.
KGE-QCB1-6	Alcohol level	Is a measure of alcohol in the blood as a percentage.
UKC-2	Name	A language unit by which a person or thing is known
UKC-66329	Medium	A state that is intermediate between extremes; a middle position
UKC-80756	Low	Less than normal in degree or intensity or amount



UKC-80747	High	Greater than normal in degree or intensity or amount
http://purl.obolibrary.org/obo/OHMI_0000468	Relative Species Abundance	A quality of ecological community that refers to how common or rare a species is relative to other species in a defined location or community

Data filtering

The second part of the language definition focuses on the data filtering process to ensure that the data layer resources match the identified concepts. In this case, no further filtering is needed because the resources are already well-aligned with the data layer.



5 Knowledge Definition

This section details the kTelos phase. The goal is to develop the final Knowledge Graph's teleontology, starting from the resources gathered for the project, the formalized objectives (as partially depicted by the ER model), and the acquired data. The knowledge resources created during this phase aim to standardize information representation, improving the interoperability and reusability of the final Knowledge Graph. This is achieved by leveraging established domain ontologies and data schemas.

The Teleontology facilitates the reuse of project data. As in earlier stages, tasks are divided into two categories: producer and consumer. On the producer side, the goal is to develop interoperable ontologies for each dataset, resulting in multiple ontology files, one for each Knowledge Graph (KG) generated. On the consumer side, the goal is to design a unified interoperable ontology for the final composite KG, leading to a single output ontology file.

Producer activities:

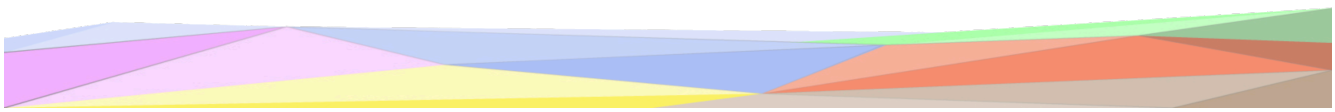
This section describes the top-down knowledge definition stage within the kTelos process. The objective is to use ontologies, harmonized with the UKC, to establish a high-level view of the entities involved in the project.

The following sources have been used for reference ontologies:

- **BioPortal**: Provides ontologies related to health and disease, particularly disease staging and microbiome-related terms.
- Within BioPortal **OHMI:** is a biomedical ontology that represents the entities and relations in the domain of host-microbiome interactions (owl file).
- Within BioPortal **DOID**: is Human disease ontology (owl file).

The first reference schema is the **Ontology for Human-Microbe Interactions** (OHMI), which includes classes relevant to our work. Some of these classes are defined and conceptualized more precisely than how we approached them during the language definition phase. One notable example is the **"Human-Microbiome Interaction"** class, which aligns closely with our previously defined class, **"correlation."** However, since "correlation" is a more generic term and not specific to our dataset, we decided to adopt the OHMI concept instead. This allows us to accurately describe the event that captures the interaction between humans and microbiomes (Figure 1).

In the OHMI reference schema, the general subclasses at the top level are divided into two major categories: **continuant** and **occurrent**. Entities whose existence depends on a specific period are grouped under **occurrent**, while entities that persist entirely through time are categorized as **continuant**. In our data, we expect the data properties of a **person** to have mostly fixed values; and to be under "continuant". Similarly, this applies to the **microbiome**, which is classified under the **continuant** category in the reference schema (Figure 2).



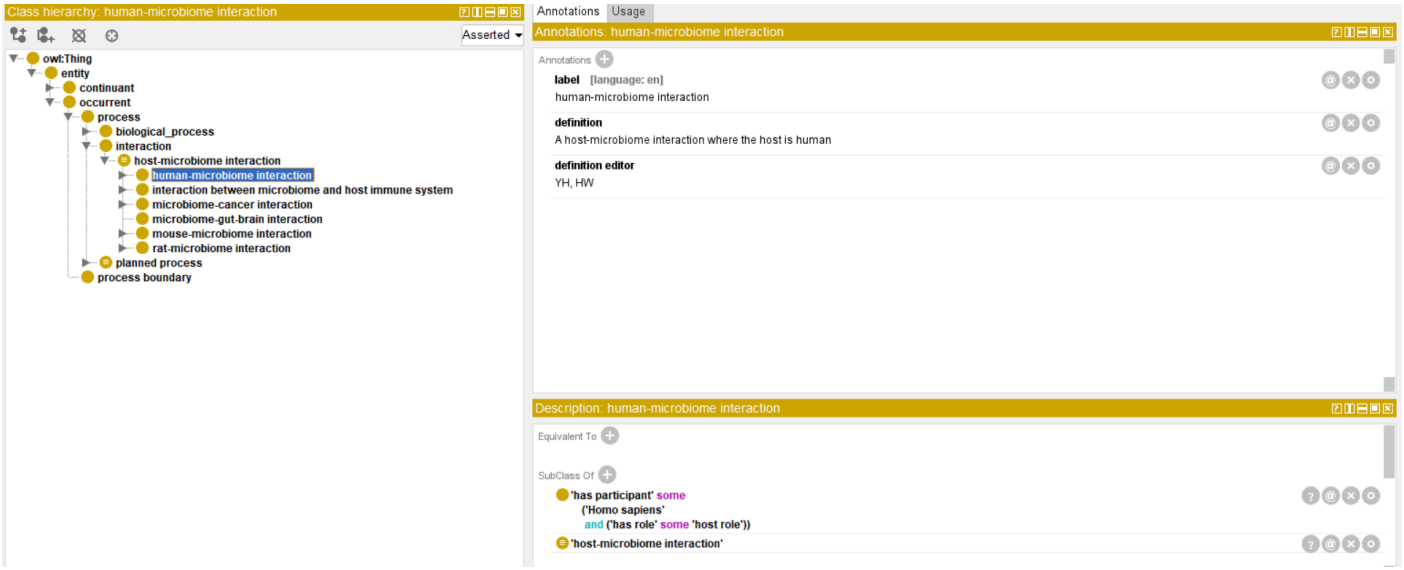


Figure 1. Located class of interest “human-microbiome interaction” in OHMI.owl (reference schema).

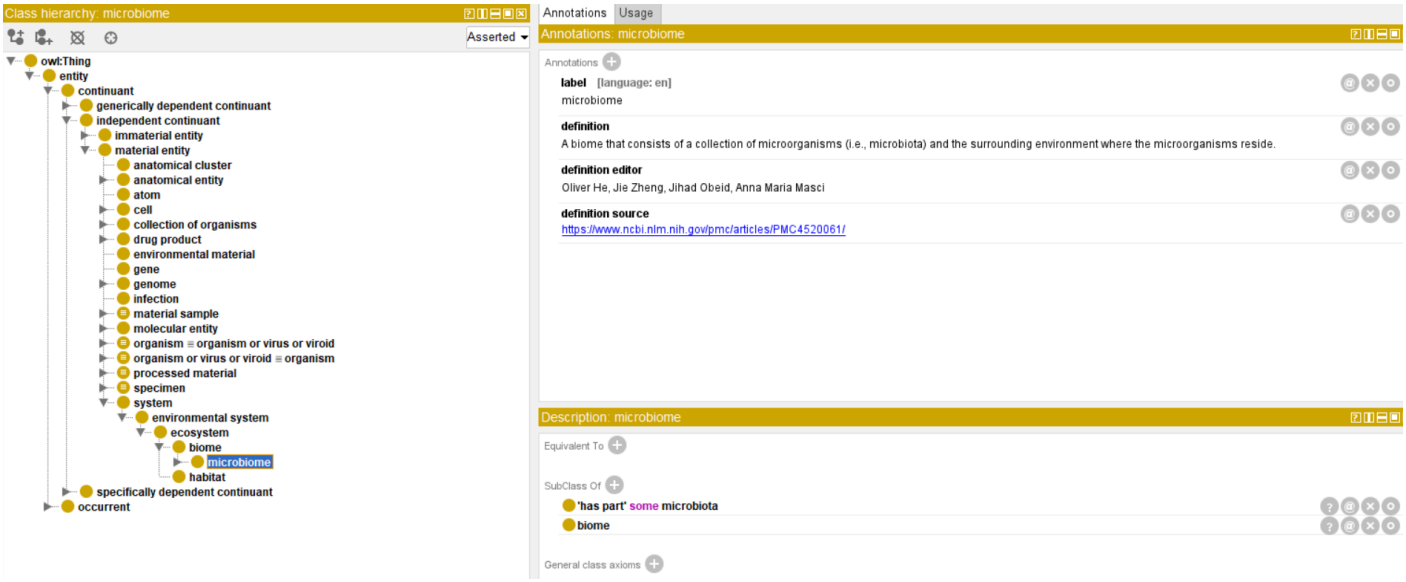


Figure 2. Microbiome Class found in the reference schema under the general class “continuant” in OHMI.owl

The DOID is the other reference ontology used that models the hierarchy of various diseases including our main interest which is cancer. This enormous ontology is composed of around 17,000 classes, where one of the top levels is disease which entails a subclass called “disease of cellular proliferation” and the latter is a direct parent class to cancer which is of interest (Figure 3a and 3b).

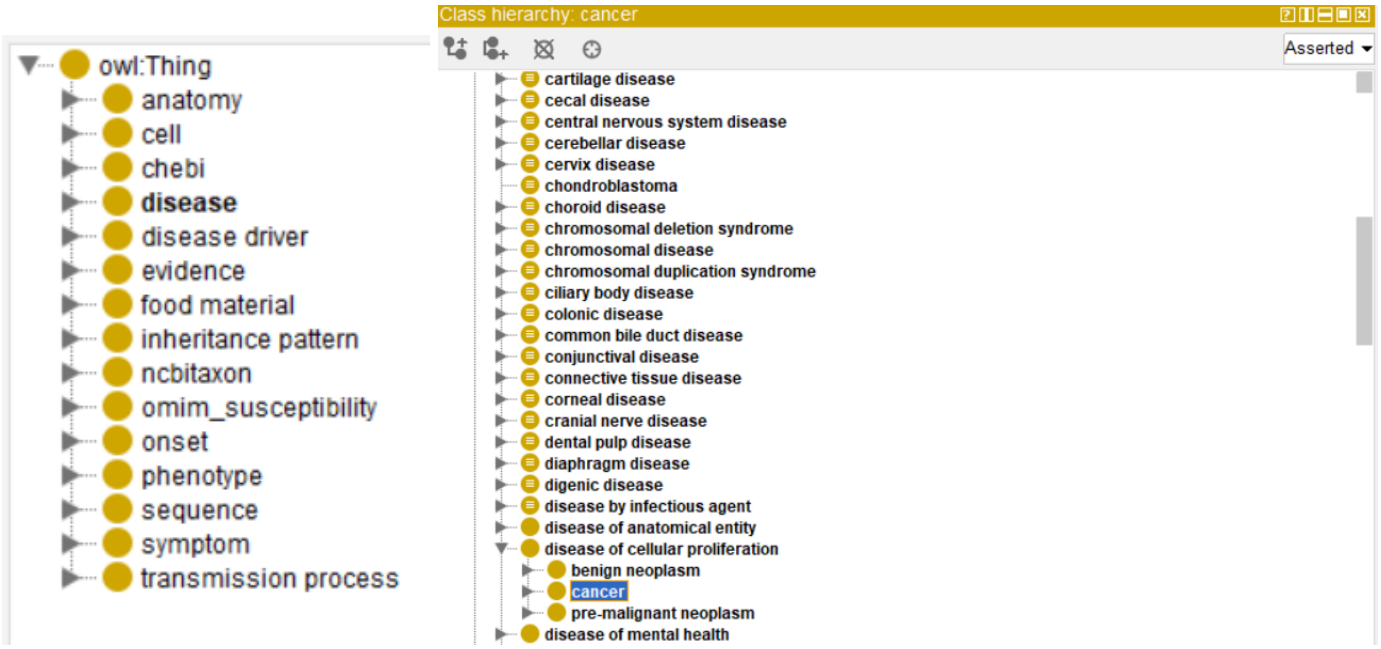


Figure 3

a. “Disease” as a general class in DOID.own (left image)

b. “Cancer” in the DOID.owl as a subclass of “disease of cellular proliferation”

Consumer Activities:

Teleology

This section outlines the bottom-up knowledge definition phase of the kTelos process. The aim is to create a teleology that aligns with the project’s objectives and data, informed by the Competency Questions (CQs), which are detailed at the beginning of the report.

These entities were defined using Concept Labels from the ontologies, aligned with the UKC, and linked using object properties in Protege, a tool used for ontology development. The data properties were maintained, but with the new Concept Labels that are specific to Cancer and microbiome composition..

The result of these connections can be visualized in Protege, with:

- Entities (classes)
- Object properties
- Data properties

Images of the Protege visualizations are provided below:

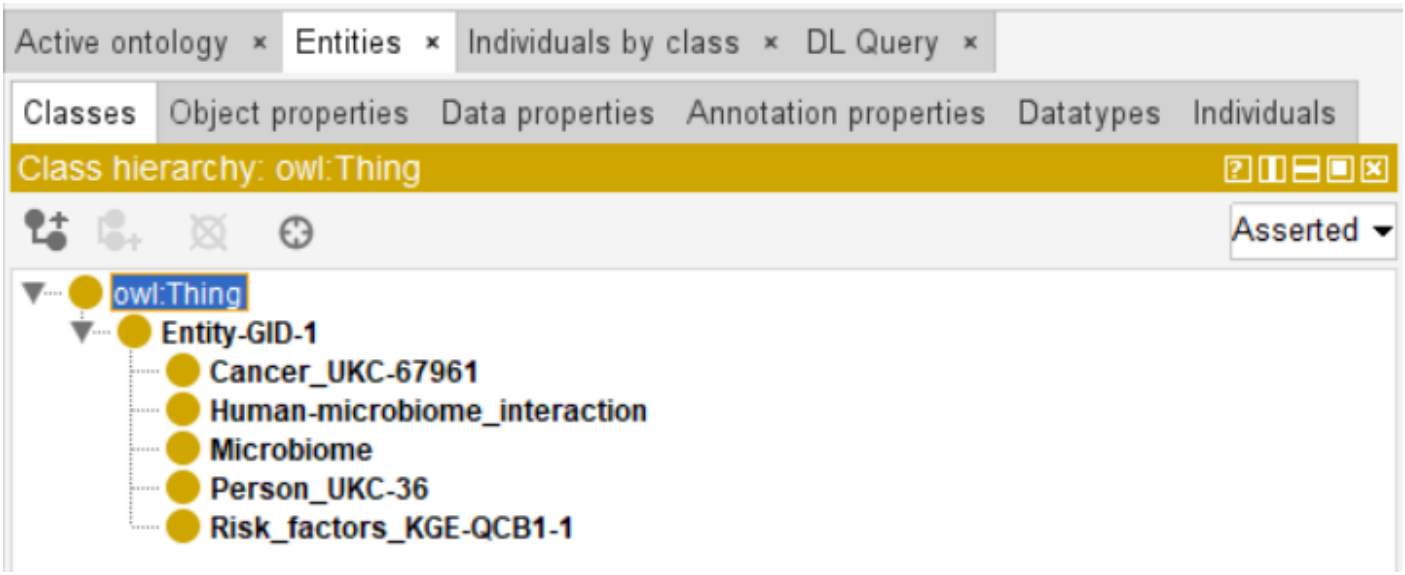


Figure 4: **Class Hierarchy** shows the class hierarchy under `owl:Thing`. Key classes include Cancer, Human-microbiome_interaction, Risk_factors_KGE-QCB1-1, Person_UKC-36, Microbiome, and Entity-GID-1. These classes represent high-level entities relevant to the ontology.

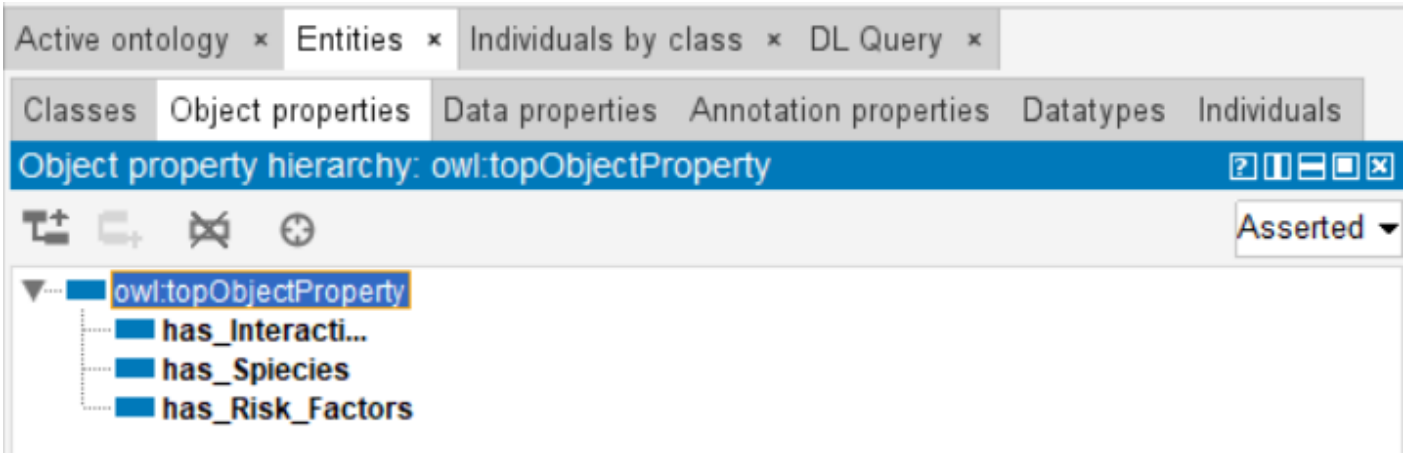


Figure 5: **Object Property Hierarchy** displays object properties associated with the ontology:

- *has_Species*: Represents reltionships to specific microbial species having domain as perosn.
- *has_Risk_Factors*: Links person with certain risk factors.
- *has_Interaction*: Describes correlations between entities (person and microbiome); domain here is human-microbiome interaction and range is person+microbiome.

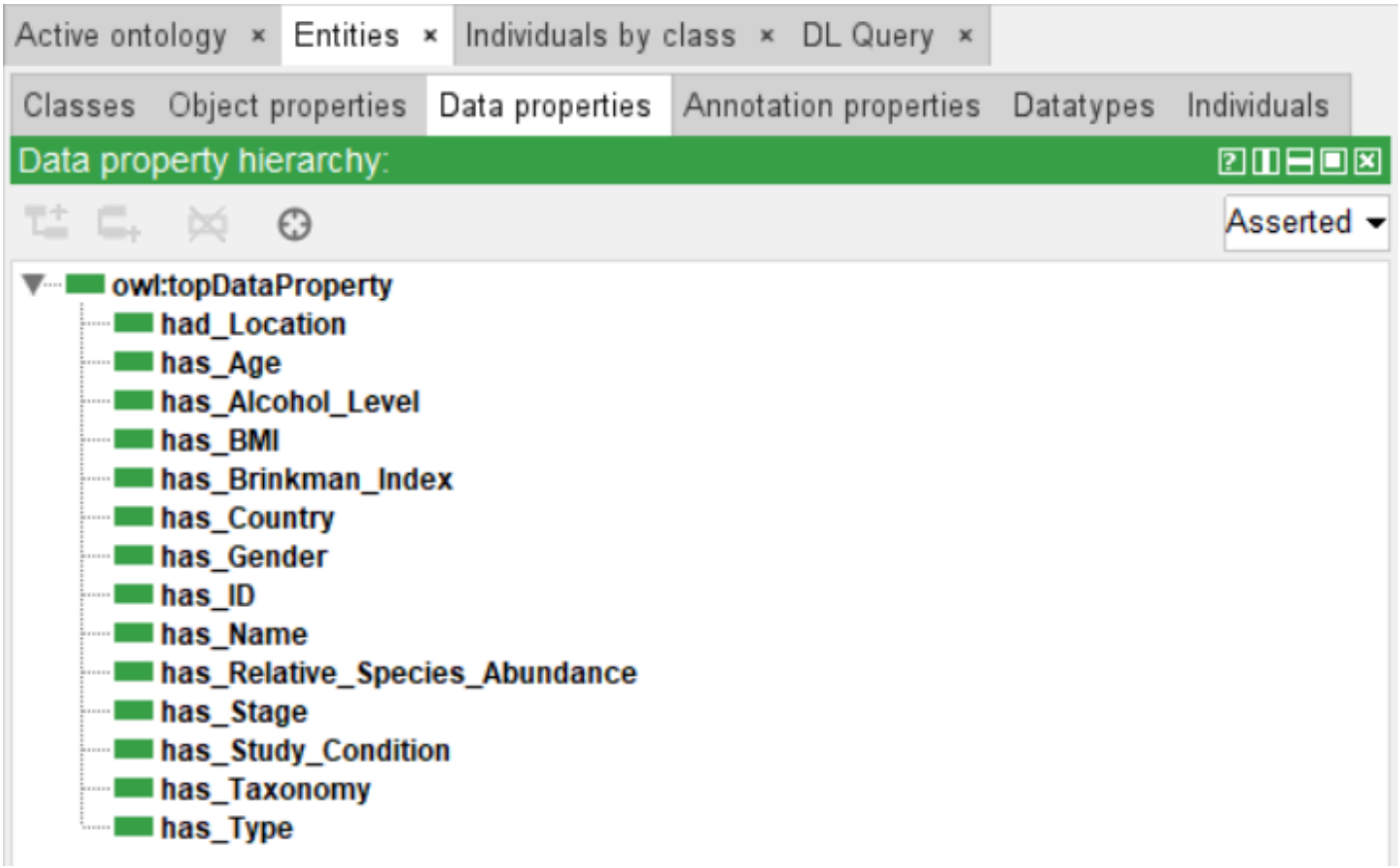


Figure 6: **Data Property Hierarchy**: lists data properties, which capture specific attributes and measurements:

- Properties like *has_Age*, *has_BMI*, *has_Gender*, and *has_Study_Condition* describe individual characteristics.
- Microbiome-specific properties such as *has_Relative_Species_Abundance* and *has_Taxonomy, has_name* detail microbial data.
- Other factors like *has_Brinkman_Index* (for smoking history) and *has_Alcohol_Level* are included as risk factors properties.
- *Has_stage and has_name* properties are properties for Cancer. (has_name here is a shared property for cancer and microbiome).

As a consumer, we make use of the domain language that we generated to define all the words in our ER model, to align our schema with a reference ontology (figure 7). After alignment, we removed the general terms that are not relevant to our purpose scope and data such as “Continuant” and “Occurant” and “disease of cellular proliferation”. Additionally, the usage of reference ontologies highlighted a gap in our former ER model which is the presence of the cancer class.

Based on this, we decided to update our ER model by creating a new entity, Cancer, which includes properties such as cancer stages, location, and type (figure 7). As a result, our spreadsheet will also need to be updated.

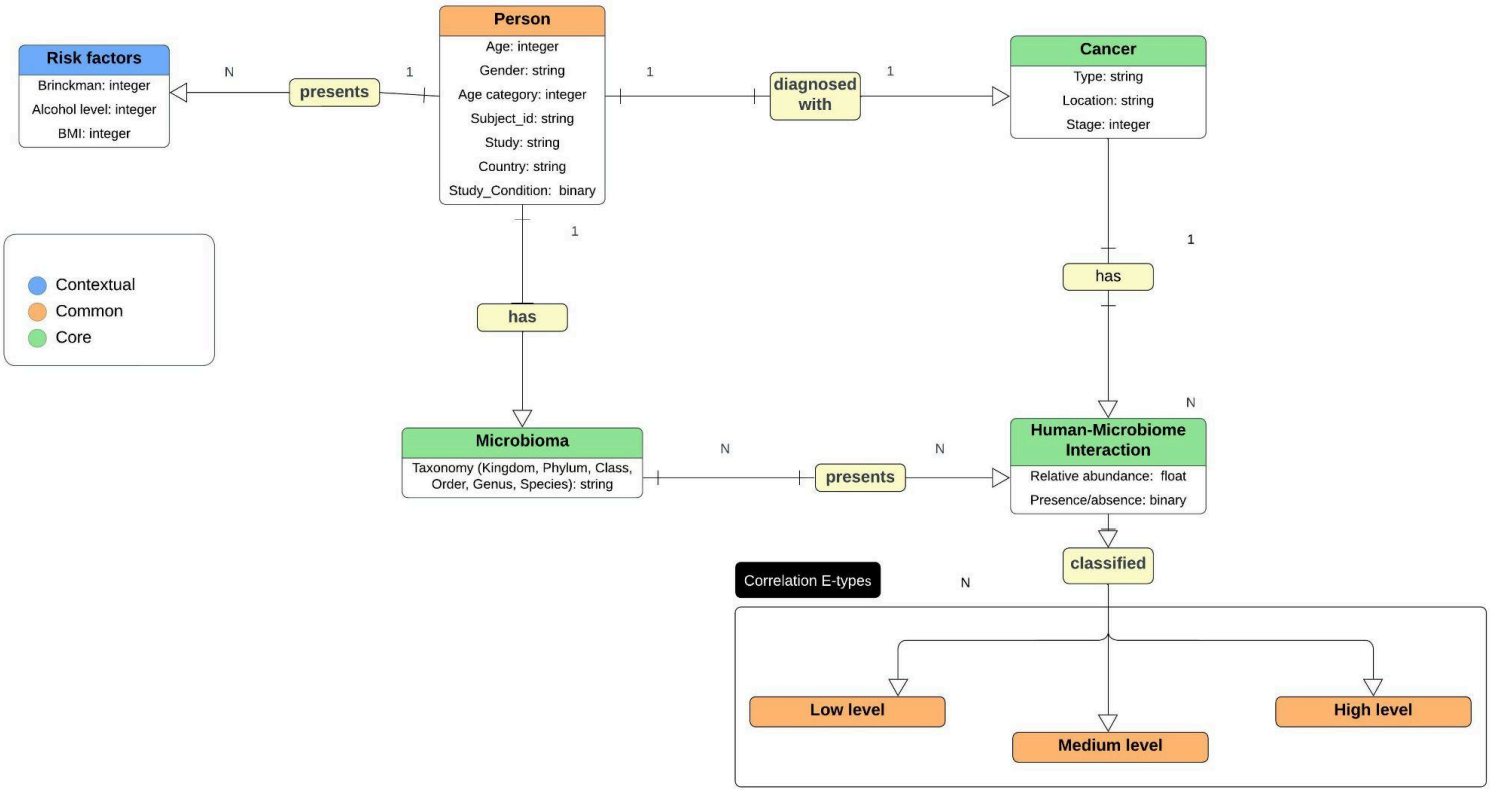


Figure 7: ER model

Thus, with our updated ER model we were able to formalize our constructed schema and align it properly to our chosen reference ontologies so that our generated teleontology can be reused later with extended data.

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.

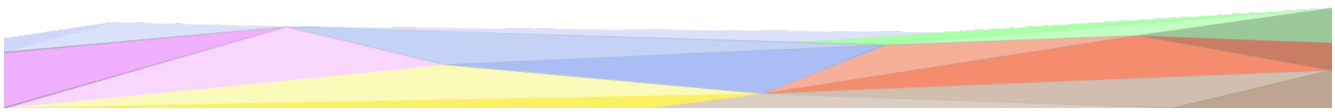
7 Evaluation

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?