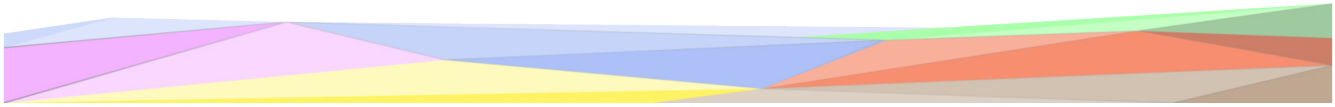
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2025 - Ecological Relations Between Members of the Microbiome

Document Data:	Reference Persons:
January 31, 2025	Virginia Leombruni, Eleonora Giuliani, Marc Shebaby



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	January 31, 2025	Author1	Document created

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 an important role in enhancing the reusability of the resources handled and produced during the process. A clear description of the resources as well as of the process (and single activities) developed, provides a clear 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 1: Purpose Definition** – Outlines the project's objectives and focus, specifying primary goals and intended outcomes. It highlights the motivation behind constructing a knowledge graph to model microbiome interactions and potential disease correlations.
- **Section 2: Information Gathering** – Details the data sources, collection methods, and resources used, emphasizing the importance of rigorous and diverse data acquisition. It includes considerations on the dataset's limitations, such as the presence of only one type of cancer.
- **Section 3: Language Definition** – Presents the linguistic resources used in structuring the knowledge graph, including key concepts, controlled vocabularies, and data formats, ensuring accessibility and integration.
- **Section 4: Knowledge Definition** – Describes the structured knowledge incorporated, including ontologies, databases, and microbiome-related datasets. It discusses how these resources contribute to the conceptual framework of the project.
- **Section 5: Entity Definition** – Defines core entities such as microbiome species, individuals, and diseases. It describes the mapping strategies used to integrate structured knowledge into the knowledge graph, ensuring consistency across datasets.
- **Section 6: Evaluation** – Assesses the effectiveness of the project in meeting its objectives. It discusses the inability to statistically validate species' relative abundances due to arbitrary thresholds and suggests future improvements, such as integrating machine learning techniques.
- **Section 7: Metadata Definition** – Outlines metadata related to language, knowledge, and data resources. It also describes the *People_Metadata* document, which compiles essential information on project contributors and affiliations.
- **Section 8: Open Issues** – Discusses unresolved challenges, including dataset limitations, the need for statistical validation of microbiome abundance classifications, and the potential for integrating time-series data to analyze species-species interactions over time.

1 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 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. 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. Observing such shifts 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:

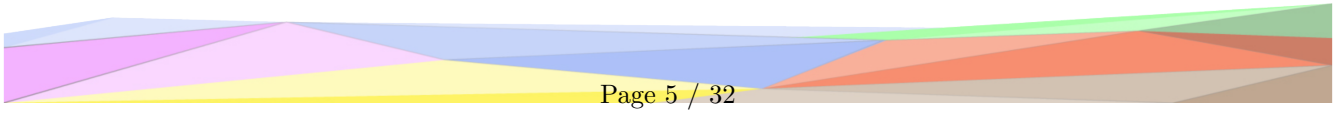
- **Subject id (SRR):** Identifier code unique for each Person.
- **Age:** Specific age of the Person.
- **Age category:** Age group classification.
- **Gender:** Can be male or female.
- **Country:** Ethnicity of the Person.
- **Study:** Refers to the specific research study that has been considered.

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.

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

Table 1: Extraction of Entities Based on CQs and Focus/Popularity Classification

The ER Model

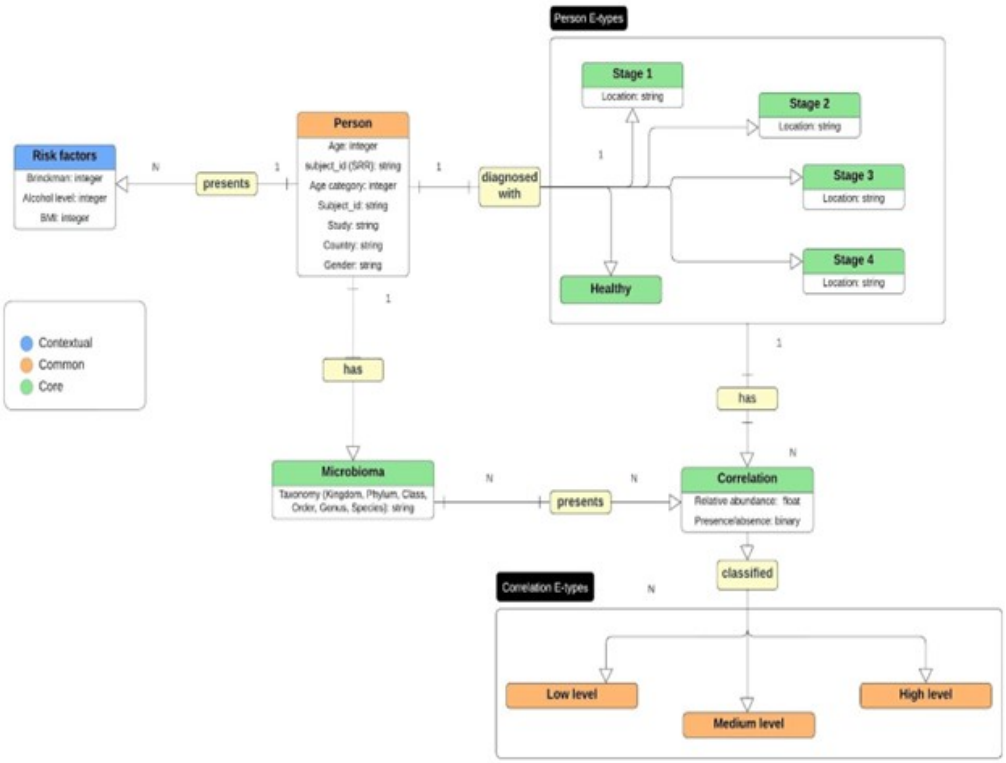


Figure 1: The ER Model

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, 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 above the average value.
- A **low level** occurs if the abundance is below the average value.

To determine these correlation categories, the relative abundance values are discretized based on the mean across all species for a particular group of persons. The grouping is identified based on tackling specific competency questions during the execution of the SPARQL queries.

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.

2 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 2) 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 2: 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.

```r
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")
```

Figure 2: Code

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 3) 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, aicc, brinckman_index, alcohol_numeric
Relative abundance	samples

Table 3: 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 4) 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, subject_id, body_site, diseases, age, age_category, gender, country, BMI, disease_location, ajcc, brinckman_index, alcohol_numeric

Table 4: 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.

3 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.

ConceptID	**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
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; 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

Table 5: Language concepts for e-types

ConceptID	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.

Table 6: Language concepts for e-types relations (object properties)

ConceptID	Word-en	Gloss-en
UKC-44477	Taxonomy	A classification of organisms into groups based on similarities of structure or origin etc.
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 (m^2).
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.

Table 7: Language concepts for e-types attributes (data properties)

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.

4 Knowledge Definition

The kTelos Phase

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.
  - **OHMI:** A biomedical ontology that represents the entities and relations in the domain of host-microbiome interactions (OWL file).
  - **DOID:** 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 3).

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 4).

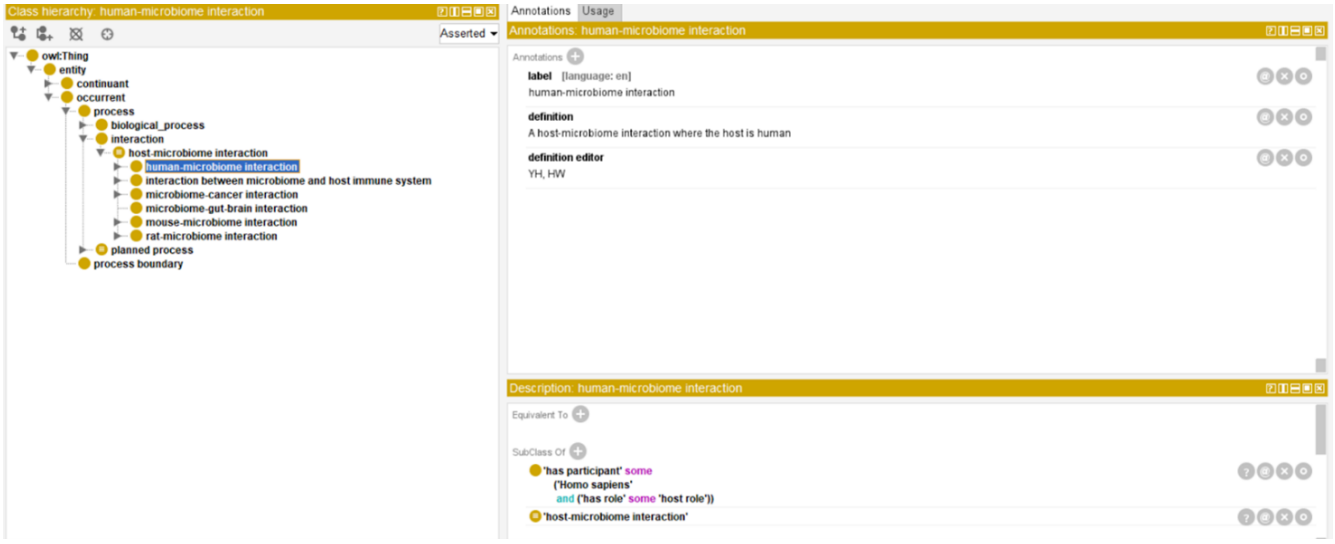


Figure 3: Located class of interest "human-microbiome interaction" in OHMI.owl (reference schema).

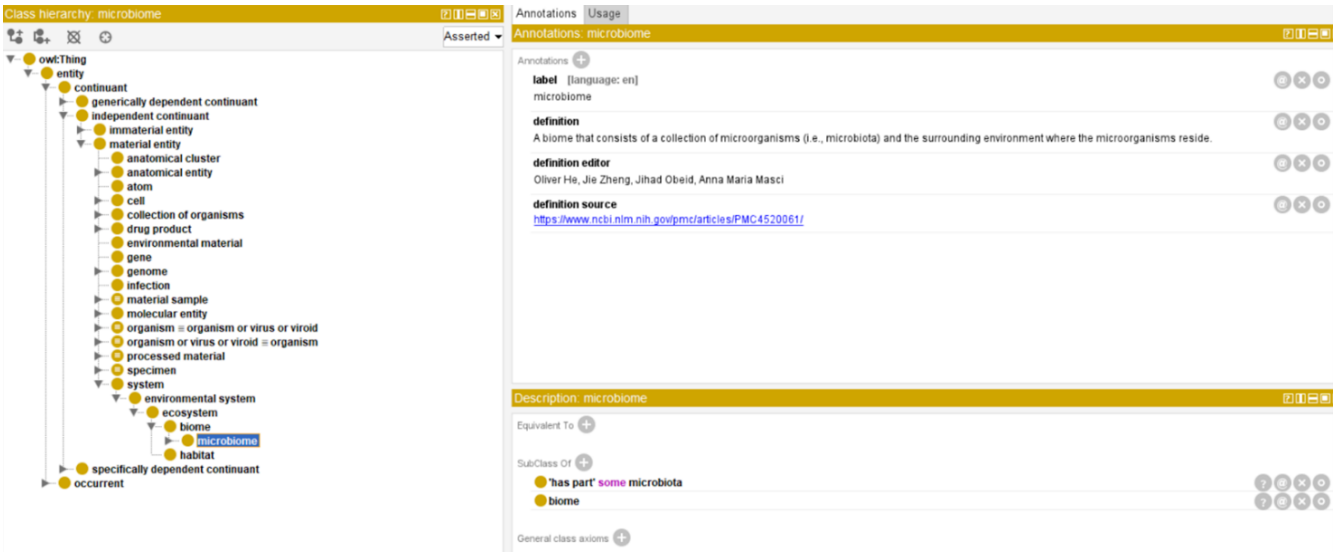


Figure 4: Microbiome Class found in the reference schema under the general class "continuant" in OHMI.owl

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 5).

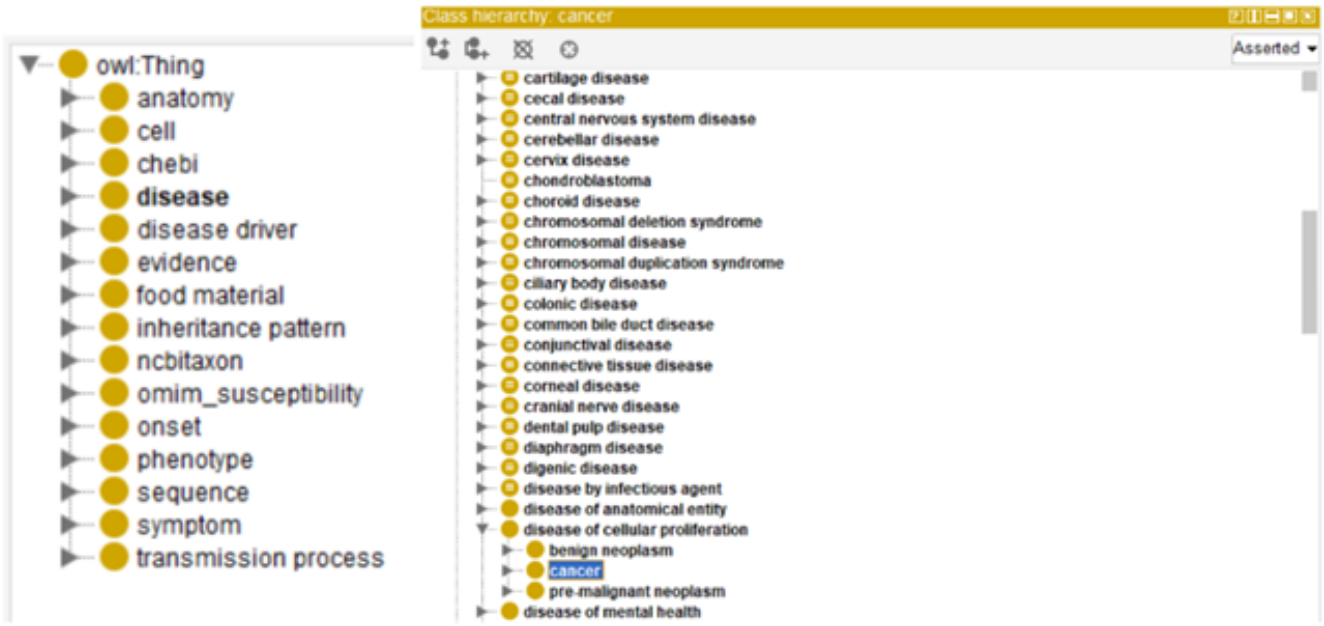


Figure 5: "Disease" as a general class in DOID.own (image on the left). "Cancer" in the DOID.owl as a subclass of "disease of cellular proliferation" (image on the right).

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 Protégé, a tool used for ontology development. The data properties were maintained, but the new Concept Labels were specific to cancer and microbiome composition.

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

- **Entities (classes)**
- **Object properties**
- **Data properties**

Images of the Protégé visualizations are provided below:

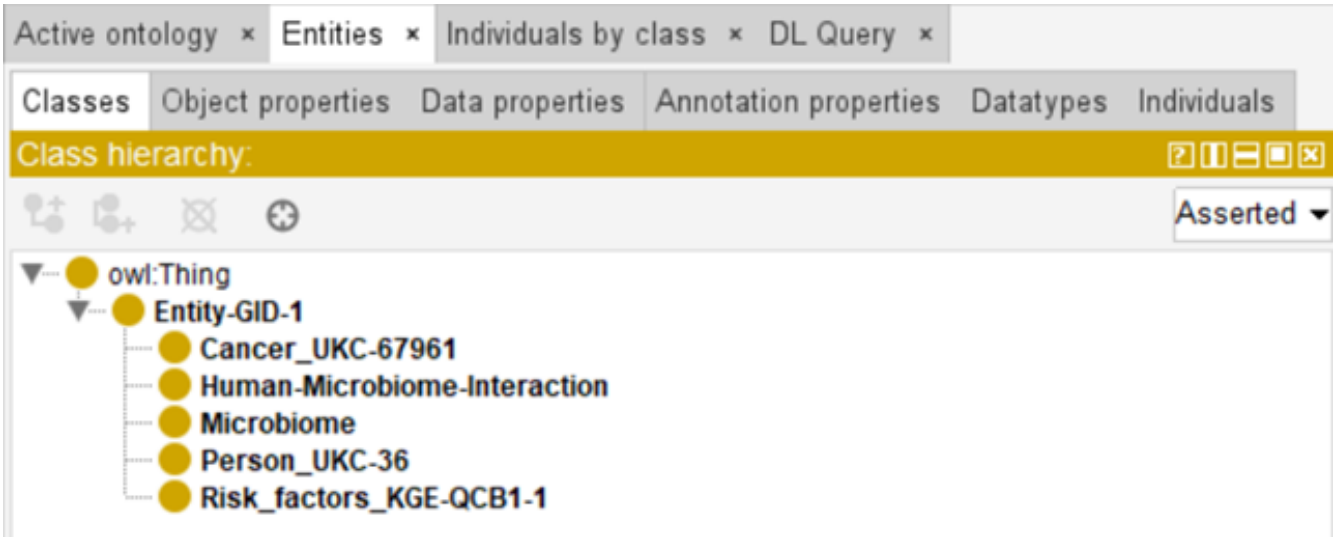


Figure 6: Class Hierarchy shows the class hierarchy under `owl:Thing`. Key classes include `Cancer`, `Human-Microbiome_Interaction`, `Risk_Factors_KGE-QCB1-1`, `Person_UKC-36`, `Microbiome`, and `Entity-GID-1`. These classes represent high-level entities relevant to the ontology.

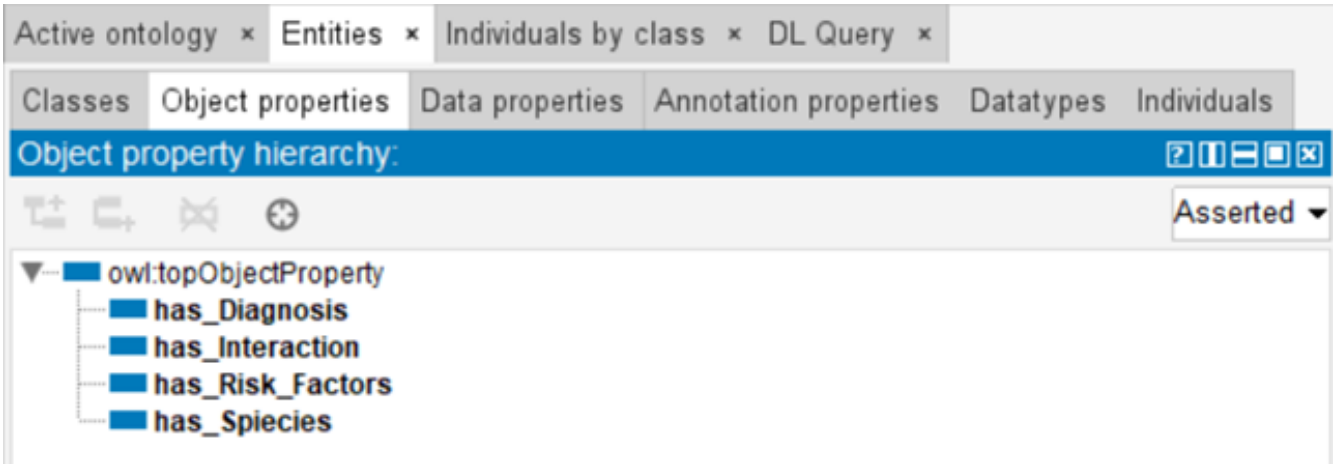


Figure 7: Object Property Hierarchy displays object properties associated with the ontology:

- **has_Species**: Represents relationships to specific microbial species having domain as a person.
- **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.
- **has_Diagnosis**: Refers to the diagnosis of a specific person with a particular type of cancer..

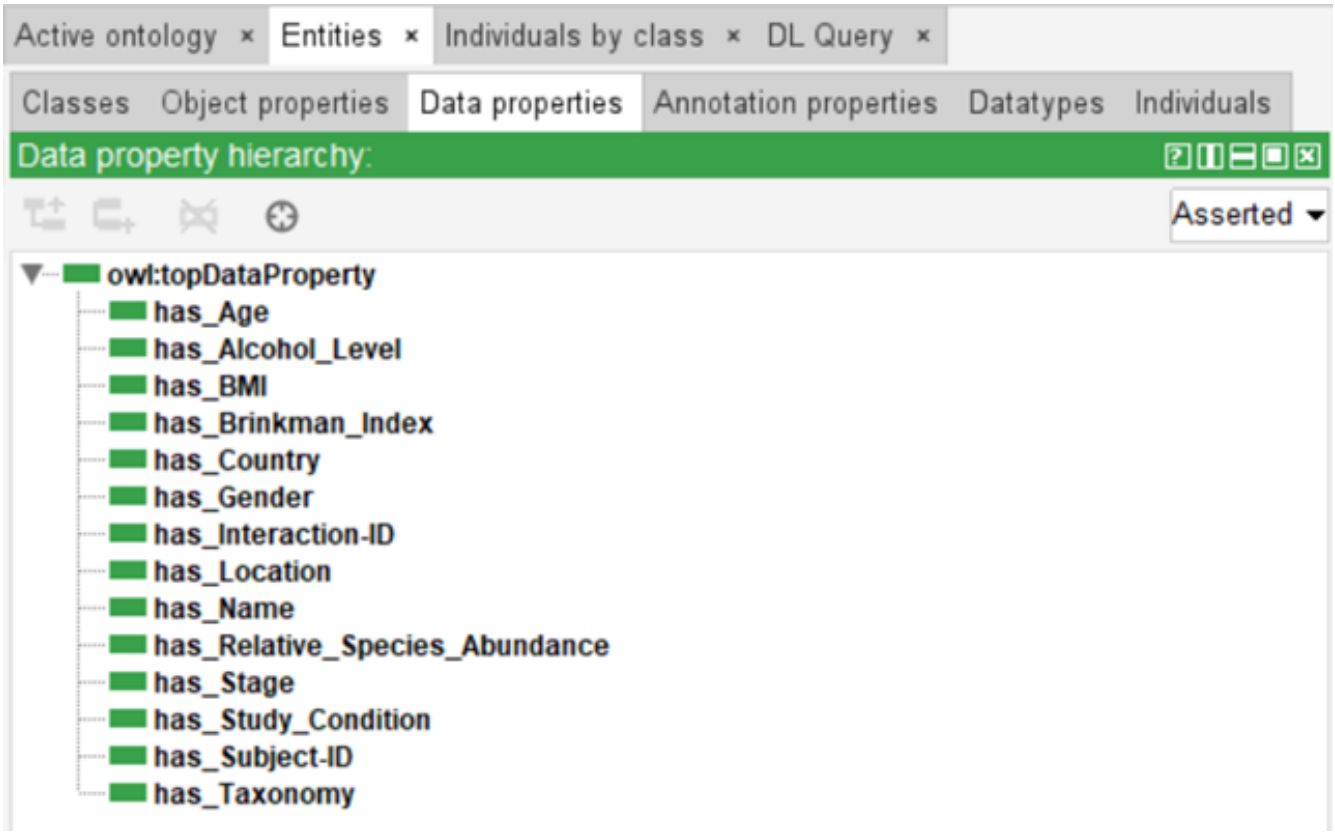


Figure 8: Data Property Hierarchy: lists data properties, which capture specific attributes and measurements:

- Properties like `has_Age`, `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), `has_BMI` and `has_Alcohol_Level` are included as risk factor 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 9). After alignment, we removed the general terms that are not relevant to our purpose scope and data such as “Continuant”, “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 9). As a result, our spreadsheet will also need to be updated.

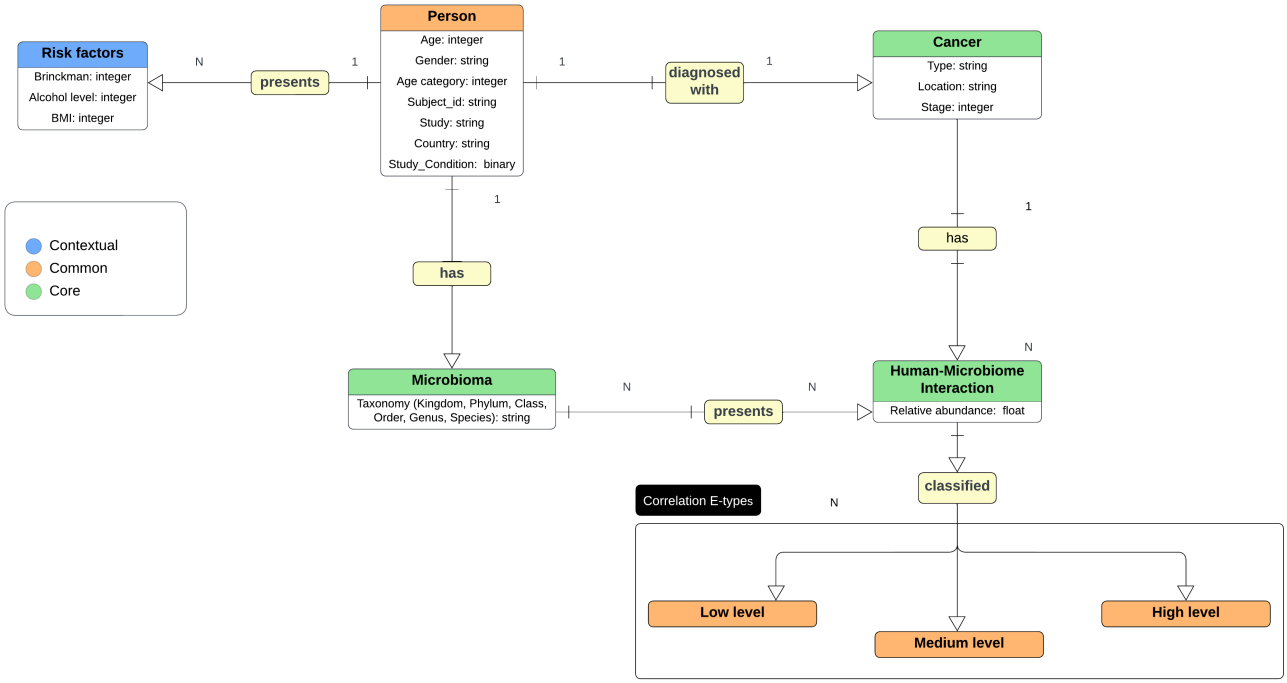


Figure 9: The ER model

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.

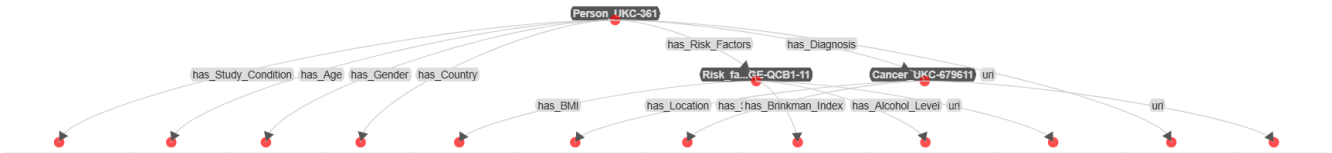
5 Entity Definition

Our goal in developing the final structured knowledge graph is to integrate our established teleology with our CSV files. This step is crucial for identifying entities, data properties, and object properties while ensuring that data from different sources are managed effectively to avoid heterogeneity issues.

Initially, we decided to modify the CSV file containing the relative abundances of species for each individual. Since the original file used person IDs as row names, it was not suitable for data integration. To properly model the interaction between species and individuals along with their respective relative abundances, we restructured the CSV file so that person IDs appear as a column. Given that a single person can be associated with multiple species, person IDs may appear multiple times across different rows. This can be achieved by the 3 following steps:

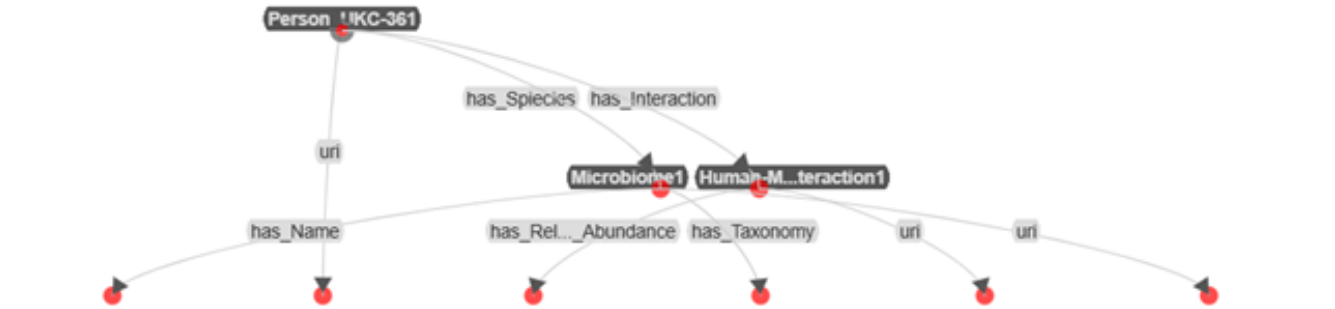
Entity Definition sub-activities:

- Entity matching The person entity exists in both files with different object properties. To ensure consistency, the person_id is used as a unique identifier to match the same individuals across datasets.
- Entity identification Within the first dataset, several entities exist, such as Risk Factor, Cancer, and Person. To ensure proper identification, two additional columns were inserted as unique identifiers for the Risk Factor and Cancer entities. Additionally, the Microbiome entity in the second dataset was uniquely identified using the species name. In Karma, these unique identifiers are referred to as URIs (Uniform Resource Identifiers).
- Data mapping As seen in figure 10 and in figure 11, with the help of Karma the teleology is directly linked to the data including all Entities with their URIs, and properties. The turle-RDF files(KG) produced by karma were merged as one to have one final graph for exploitation.



study_condition	age	gender	country	BMI	disease_location	ajcc	brinkman_index	alcohol_numeric	Risk_factor_id	person_id	Cancer_id
control	68	male	JPN	21.383942	NA	NA	1290	0	1	SAMD00114718	C1
control	60	male	JPN	22.22906193	NA	NA	100	933	2	SAMD00114719	C2

Figure 10: Karma visualization



Species_name	person_id	relative_abundance	taxonomy	interaction_id	Species_ID
Phocaeicola plebeius	SAMD00114718	54.94898	NA	1	Phocaeicola plebeius
Phocaeicola plebeius	SAMD00114719	11.23749	NA	2	Phocaeicola plebeius
Phocaeicola plebeius	SAMD00114720	19.09913	NA	3	Phocaeicola plebeius

Figure 11: Karma visualization

6 Evaluation

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

- The final Knowledge Graph information statistics (like, number of etypes 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 KG's suitability to satisfy the project purpose.

After completing all the phases of the construction of the knowledge graph and obtaining the final knowledge graph, an essential and important step is the evaluation. iTelos provides various methods to assess the execution. Specifically, both the reusability (secondary objective) and the ability of the knowledge graph to satisfy the competency queries (primary objective) are assessed.

One possible metric to evaluate the knowledge layer that can be applied is the Coverage, defined as the extent to which a portion of knowledge is represented in the knowledge graph.

Teleontology vs CQs

At the EType level:

$$\mathrm{Cov}_E(\mathrm{CQ}_E) = \frac{|\mathrm{CQ}_E \cap \mathrm{T}_E|}{\mathrm{CQ}_E} = \frac{5}{5} = 1$$

Where CQ_E represents the number of etypes extracted from the CQs and T_E is the number of etypes of the Teleontology.

At the Property level:

$$\mathrm{Cov}_p(\mathrm{CQ}_p) = \frac{|\mathrm{CQ}_p \cap \mathrm{T}_p|}{\mathrm{CQ}_p} = \frac{18}{19} = 0.94$$

specifically, we have 14 data properties and 4 object properties

Where CQ_P represents the number of properties extracted from the CQs and T_P is the number of properties of the Teleontology.

Teleontology vs Reference Ontologies (ROs)

The OHMI ontology consists of 1025 entities and 133 properties, whereas the DOID ontology consists of 18,839 entities and 45 properties. The shared entities between them are 165. Then the evaluation is of the following. At the EType level:

$$\text{Cov}_E(\text{RO}_E) = \frac{|\text{RO}_E \cap \text{T}_E|}{\text{RO}_E} = \frac{3}{19699} = 0.0001$$

Where RO_E represents the number of etype extracted from the ROs and T_E is the number of properties of the Teleontology.

At the Property level:

$$\text{Cov}_p(\text{RO}_p) = \frac{|\text{RO}_p \cap \text{T}_p|}{\text{RO}_p} = \frac{18}{178} = 0.1$$

Where RO_p represents the number of properties extracted from the ROs and T_p is the number of properties of the Teleontology.

Connectivity

On the other hand, evaluating the Data layer involves measuring the connectivity of the knowledge graph (KG). This analysis is performed across two key dimensions, both quantified using the connectivity metric:

- **Entity connectivity:** This measures the degree of interconnection between entities within the graph, emphasizing the richness and density of relationships between them.
- **Property connectivity:** This evaluates how thoroughly entities are linked to their corresponding properties, reflecting the completeness and detail of the information represented in the graph.

	Person	Cancer	Risk Factors	Microbioma	Human-Microbiome Interactions
Person	4	258	253	403480	53611
Cancer		2			
Risk Factors			3		
Microbioma				2	53611
Human-Microbiome Interactions					1

Table 8: The KG's evaluation - Data layer

SPARQL

To usefully exploit our knowledge graph, SPARQL was used to query and answer the competency questions. Useful information about relative abundances of species in different scenarios were retrieved. To answer CQ-6 and CQ-4.

```
SELECT (MAX(?mean) AS ?total_Mean)
WHERE {
    {
        SELECT ?person (AVG(xsd:float(?relative_value)) AS ?mean)
        WHERE {
            # Get species of interest
            ?species rdf:type etype:Microbiome .
            ?interaction rdf:type etype:Human-Microbiome-Interaction .
            ?interaction etype:has_Relative_Species_Abundance ?relative_value .
            ?species etype:has_Interaction ?interaction .
            ?person rdf:type etype:Person_UKC-36.
            ?person etype:has_Spiecies ?species .
            ?person etype:has_Interaction ?interaction .
            ?person etype:has_Study_Condition ?status.
             FILTER(str(?status)='CRC').
        }
        GROUP BY ?person
```

Table | Raw response | Pivot Table | Google Chart

Filter query results | Compact view ☐ | Hide row numbers ☐

	total_Mean
1	"3.4462788"^^xsd:float

Figure 12: Query 1: Retrieving the maximum average of a person's relative abundance with all the species

```
select DISTINCT ?person ?alcohol ?cig_level ?species ?occurrence ?mean_of_species
where{
{
select ?species (COUNT(?species) AS ?occurrence) (AVG(xsd:float(?relative_value))AS ?mean_of_species)
where {
   #get species of interest
 ?species rdf:type etype:Microbiome .
 ?species etype:has_Name ?name .
    #get event of interest (relative abundance)
  ?interaction rdf:type etype:Human-Microbiome-Interaction .
  ?interaction etype:has_Relative_Species_Abundance ?relative_value .
  FILTER(xsd:float(?relative_value) > 3.44).
  ?species etype:has_Interaction ?interaction .
    #get person of interest
  ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Spiecies ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.

  ?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
  ?person etype:has_Age ?Age .
  ?risk etype:has_Alcohol_Level ?alcohol .
  ?risk etype:has_Brinkman_Index ?cig_level .
  FILTER(xsd:float(?cig_level) > 600).
  FILTER(xsd:float(?alcohol) > 300).
  ?person etype:has_Risk_Factors ?risk .
  ?cancer rdf:type etype:Cancer_UKC-67961.
  ?cancer etype:has_Stage ?stage.
  FILTER(str(?stage)!='NA').
  ?person etype:has_Diagnosis ?cancer.
}
    GROUP BY   ?species #?cig_level ?alcohol
}
 ?person rdf:type etype:Person_UKC-36.
    ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Spiecies ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.

  ?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
  ?person etype:has_Age ?Age .
  ?risk etype:has_Alcohol_Level ?alcohol .
  ?risk etype:has_Brinkman_Index ?cig_level .
  FILTER(xsd:float(?cig_level) > 600).
  FILTER(xsd:float(?alcohol) > 300).
  ?person etype:has_Risk_Factors ?risk .

  ?cancer rdf:type etype:Cancer_UKC-67961.
  ?cancer etype:has_Stage ?stage.
  FILTER(str(?stage)='iv').
  ?person etype:has_Diagnosis ?cancer.
}
ORDER BY DESC (?occurrence)
```

Figure 13: Query 2: Answers CQ-6, CQ-4 and CQ-1 by retrieving the species with high relative abundance (greater than the mean of relative abundances for cancer) for cancer patients having stage ”iv” cancer, high alcohol level and high brinkman index level. The species are ordered by their connectivity number with the cancer patients. Additionally the mean for the species is returned.

	person	alcohol	cig_level	species	occurrence_of_species	mean_of_species	stage
1	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Bacteroides%20uniformis	"12"^^xsd:integer	"11.230103"^^xsd:float	"iv"
2	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Bacteroides%20uniformis	"12"^^xsd:integer	"11.230103"^^xsd:float	"iv"
3	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Bacteroides%20uniformis	"12"^^xsd:integer	"11.230103"^^xsd:float	"iv"
4	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Bacteroides%20uniformis	"12"^^xsd:integer	"11.230103"^^xsd:float	"iv"
5	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Eubacterium%20rectale	"11"^^xsd:integer	"6.4065185"^^xsd:float	"iv"
6	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Eubacterium%20rectale	"11"^^xsd:integer	"6.4065185"^^xsd:float	"iv"
7	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Eubacterium%20rectale	"11"^^xsd:integer	"6.4065185"^^xsd:float	"iv"
8	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"^^xsd:integer	"8.460612"^^xsd:float	"iv"

Figure 14: Results of query 2; much of bacteriodes are found to be linked with cancer patients that have high alcohol and cigarettes levels.

9	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"^^xsd:integer	"8.460612"^^xsd:float	"iv"
10	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"^^xsd:integer	"8.460612"^^xsd:float	"iv"
11	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"^^xsd:integer	"8.460612"^^xsd:float	"iv"
12	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Prevotella%20copri	"10"^^xsd:integer	"37.641094"^^xsd:float	"iv"
13	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Prevotella%20copri	"10"^^xsd:integer	"37.641094"^^xsd:float	"iv"
14	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Bacteroides%20stercoris	"9"^^xsd:integer	"9.614393"^^xsd:float	"iv"
15	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Bacteroides%20stercoris	"9"^^xsd:integer	"9.614393"^^xsd:float	"iv"

Figure 15: Extension of results for query 2

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>


select DISTINCT ?person ?species ?relative_value ?stage#(AVG(xsd:float(?relative_value)) AS ?mean)
where {
    {select ?name
        where{
         #get species of interest
 ?species rdf:type etype:Microbiome .
 ?species etype:has_Name ?name .
    #get event of interest (relative abundance)
  ?interaction rdf:type etype:Human-Microbiome-Interaction .
  ?interaction etype:has_Relative_Species_Abundance ?relative_value .
  FILTER(xsd:float(?relative_value) > 1 && xsd:float(?relative_value) <5).
  ?species etype:has_Interaction ?interaction .
    #get person of interest
  ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Spiecies ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.
    FILTER(str(?status)!='CRC').
    }}
 ?species rdf:type etype:Microbiome .
 ?species etype:has_Name ?name .
    #FILTER(?name !='Prevotella copri' ).
    #get event of interest (relative abundance)
  ?interaction rdf:type etype:Human-Microbiome-Interaction .
  ?interaction etype:has_Relative_Species_Abundance ?relative_value .
  FILTER(xsd:float(?relative_value) > 10).
  ?species etype:has_Interaction ?interaction .
    #get person of interest
  ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Spiecies ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.
    FILTER(str(?status)='CRC').
 # ?cancer rdf:type etype:Cancer_UKC-67961.
  #?cancer etype:has_Stage ?stage.
  #?person etype:has_Diagnosis ?cancer
}
#GROUP BY (?person)
```

Figure 16: Query 3

Figure 16 answers general CQs and CQ-9 by identifying species associated with cancer. It highlights species that show a high relative abundance in individuals with cancer and a low abundance in healthy individuals.

	person	species	relative_value
1	http://localhost:8080/source/SAMD00164889	http://localhost:8080/source/Bacteroides%20uniformis	"10.04378"
2	http://localhost:8080/source/SAMD00114811	http://localhost:8080/source/Prevotella%20sp%20CAG5226	"10.12263"
3	http://localhost:8080/source/SAMD00115010	http://localhost:8080/source/Prevotella%20sp%20CAG520	"10.12896"
4	http://localhost:8080/source/SAMD00114775	http://localhost:8080/source/Faecalibacterium%20prausnitzii	"10.13096"
5	http://localhost:8080/source/SAMD00164867	http://localhost:8080/source/Bacteroides%20uniformis	"10.13915"

Figure 17: Results of Query 3: Highlights that species with low relative abundance in Healthy individuals are highly expressed in individuals diagnosed with cancer.

```
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>


        SELECT ?person ?species ?relative_value ?cig ?bmi ?status
        WHERE {
            # Get species of interest
            ?species rdf:type etype:Microbiome .
            ?species etype:has_Name 'Helicobacter pylori'.
            ?interaction rdf:type etype:Human-Microbiome-Interaction .
            ?interaction etype:has_Relative_Species_Abundance ?relative_value .
            ?species etype:has_Interaction ?interaction .
            ?person rdf:type etype:Person_UKC-36.
            ?person etype:has_Spiecies ?species .
            ?person etype:has_Interaction ?interaction .
            ?person etype:has_Study_Condition ?status.
             #FILTER(str(?status)='CRC').
            ?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
            ?person etype:has_Age ?Age .
            ?risk etype:has_BMI ?bmi .
            ?risk etype:has_Brinkman_Index ?cig .
            ?person etype:has_Risk_Factors ?risk .
        }
```

Figure 18: Query 4

Figure 18 answers CQ-8 by retrieving the individuals whose microbiome includes *Helicobacter pylori*.

	person	species	relative_value	cig	bmi	status
1	http://localhost:8080/source/SAMD00114899	http://localhost:8080/source/Helicobacter%20pylori	"0.00157"	"0"	"22.18934911"	"control"
2	http://localhost:8080/source/SAMD00164772	http://localhost:8080/source/Helicobacter%20pylori	"0.00337"	"570"	"25.40281608"	"adenoma"
3	http://localhost:8080/source/SAMD00164834	http://localhost:8080/source/Helicobacter%20pylori	"0.01394"	"360"	"22.14532872"	"CRC"
4	http://localhost:8080/source/SAMD00164893	http://localhost:8080/source/Helicobacter%20pylori	"0.00399"	"0"	"18.7961895"	"adenoma"

Figure 19: Shows that out of the four persons that contain the Helicobacter pylori one person with cancer has the most relative abundance among the others.

```
SELECT ?person ?status ?cig ?bmi
       (SUM(?ecoli_abundance) AS ?Escherichia_coli_abundance)
       (SUM(?gnavus_abundance) AS ?Ruminococcus_gnavus_abundance)
       (ABS(SUM(?ecoli_abundance) - SUM(?gnavus_abundance)) AS ?Abs_Diff_Ecoli_Gnavus)
       (ABS(SUM(?ecoli_abundance) - ?mean_gnavus_crc) AS ?Abs_Diff_CRC)
       (ABS(SUM(?ecoli_abundance) - ?mean_gnavus_noncrc) AS ?Abs_Diff_NonCRC)
WHERE {
    # Get species of interest
    ?species rdf:type etype:Microbiome .
    ?species etype:has_Name ?name .
    FILTER(?name IN ('Ruminococcus gnavus', 'Escherichia coli'))
    ?interaction rdf:type etype:Human-Microbiome-Interaction .
    ?interaction etype:has_Relative_Species_Abundance ?relative_value .
    ?species etype:has_Interaction ?interaction .
    # Ensure valid number format
    BIND(xsd:float(?relative_value) AS ?rel_value)
    # Assign values conditionally
    BIND(IF(?name = 'Escherichia coli', ?rel_value, 0) AS ?ecoli_abundance)
    BIND(IF(?name = 'Ruminococcus gnavus', ?rel_value, 0) AS ?gnavus_abundance)

    # Person details
    ?person rdf:type etype:Person_UKC-36.
    ?person etype:has_Spiecies ?species .
    ?person etype:has_Interaction ?interaction .
    ?person etype:has_Study_Condition ?status.
    ?person etype:has_Gender 'female'.

    # Risk factors
    ?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
    ?person etype:has_Risk_Factors ?risk .
    ?risk etype:has_BMI ?bmi .
    ?risk etype:has_Brinkman_Index ?cig .
    # Compute mean Ruminococcus gnavus abundance per group
    {
        SELECT ?status (AVG(xsd:float(?relative_value)) AS ?mean_gnavus)
        WHERE {
            ?species rdf:type etype:Microbiome .
            ?species etype:has_Name 'Ruminococcus gnavus' .
            ?interaction rdf:type etype:Human-Microbiome-Interaction .
            ?interaction etype:has_Relative_Species_Abundance ?relative_value .
            ?species etype:has_Interaction ?interaction .
            ?person rdf:type etype:Person_UKC-36.
            ?person etype:has_Spiecies ?species .
            ?person etype:has_Interaction ?interaction .
            ?person etype:has_Study_Condition ?status.
        }
        GROUP BY ?status
    }
    # Bind mean values separately for CRC and non-CRC
    BIND(IF(?status = 'CRC', ?mean_gnavus, 0) AS ?mean_gnavus_crc)
    BIND(IF(?status != 'CRC', ?mean_gnavus, 0) AS ?mean_gnavus_noncrc)
}
GROUP BY ?person ?status ?cig ?bmi ?mean_gnavus_crc ?mean_gnavus_noncrc
```

Figure 20: Query 5

	person	status	cig	gender	Escherichia_coli_...	Ruminococcus_g...	Abs_Diff_Ecoli_Gn...	Abs_Diff_CRC	Abs_Diff_NonCRC
1	http://localhost:8080/source/SAMD00114718	"control"	"1290"	"male"	"1.31309"^^xsd:float	"2.26875"^^xsd:float	"0.9556599855422974"^^xsd:float	"1.3130899667739868"^^xsd:float	"1.0476807355880737"^^xsd:float
2	http://localhost:8080/source/SAMD00114719	"control"	"100"	"male"	"9.1E-4"^^xsd:float	"0.07518"^^xsd:float	"0.07427000254392624"^^xsd:float	"9.10000002477318E-4"^^xsd:float	"2.35986065864563"^^xsd:float
3	http://localhost:8080/source/SAMD00114720	"control"	"1800"	"male"	"0.00525"^^xsd:float	"4.14455"^^xsd:float	"4.1392998695373535"^^xsd:float	"0.005249999929219484"^^xsd:float	"2.355520725250244"^^xsd:float
4	http://localhost:8080/source/SAMD00114721	"control"	"300"	"male"	"16.3262"^^xsd:float	"0.09351"^^xsd:float	"16.232690811157227"^^xsd:float	"16.32620048522949"^^xsd:float	"13.965429306030273"^^xsd:float
5	http://localhost:8080/source/SAMD00114730	"control"	"0"	"female"	"0"^^xsd:integer	"0.19638"^^xsd:float	"0.19638000428676605"^^xsd:float	"0"^^xsd:integer	"2.360770702362060"^^xsd:float
6	http://localhost:8080/source/SAMD00114734	"control"	"900"	"male"	"0.02027"^^xsd:float	"0"^^xsd:integer	"0.020269999280571938"^^xsd:float	"0.020269999280571938"^^xsd:float	"2.340500593185425"^^xsd:float
7	http://localhost:8080/source/SAMD00114736	"control"	"0"	"female"	"0.6486"^^xsd:float	"9.47762"^^xsd:float	"8.829020500183105"^^xsd:float	"0.6485999822616577"^^xsd:float	"1.7121707201004028"^^xsd:float

Figure 21: Results of Query 5

Figure 20 illustrates the process of answering Query 5, which addresses CQ-10 through the following steps:

(A) **Mean Relative Abundance Calculation**: Compute the mean relative abundance of one of the most abundant species, *Ruminococcus gnavus*, specifically for the 'CRC' group in both the 'Healthy' and 'CRC' populations.

(B) **Individual Distance Measurement**: for each person calculate the distance between her *E. coli* relative abundance and their own relative abundance of *Ruminococcus gnavus* computed in step (A).

(C) **Group Distance Comparison**: Measure the distance between each person's *E. coli* relative abundance and the mean relative abundances computed in step (A) for the 'Healthy' and 'CRC' groups.

This stepwise approach provides a structured method for addressing the query and comparing microbial profiles across the two groups.

7 Metadata Definition

In this section, the report collects the definitions of all the metadata defined for the different resources produced throughout the entire process. The metadata defined at this stage describes both the final result of the project and the intermediate results from each phase (language, schema, and standardized values of data sources). Defining metadata is crucial to enabling the distribution (sharing) of the produced resources through data catalogs. Therefore, it is important to describe where these metadata will be published to distribute the resources they describe (e.g., DataScientia catalogs).

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

- **Project Metadata Description**
- **Language Resource Metadata Description**
- **Knowledge Resource Metadata Description**
- **Data Resource Metadata Description**
- **People Metadata Description**

Project Metadata Description

The project metadata includes essential details about the overall project, such as the title, URLs, and descriptions. This metadata will be published and made available through repositories on GitHub.

prjURL	prjKeywords	prjType	prjDescription	prjStartD	prjEndD	prjFundingAger	prjInput	prjOutput	prjCoordinator	prjObservations
https://github.com/Virginia-hub-del/Knowledge-Graph-Engineering	microbiome,risk factors,cancer,interactions	Knowledge Resource Generation	This Knowledge Graph is designed to highlight correlations between the composition of the microbiota in a collection of individuals, specific risk factors, and colorectal cancer diagnoses. It has been developed using the iTelos methodology.	set-24	feb-25	Datascientia foundation	The project utilizes a dataset sourced from the R package CuratedMetagenomicData (CMD). Knowledge resources were derived from BioPortal.	A Knowledge Graph was developed, complete with a full ontology, teleology, and teleontology. Additionally, a GitHub repository, a presentation and a report, including all the details of the full process, were created.	Simone Bocca	

Figure 22: Project metadata

Language Metadata Description

Language resource metadata refers to the language, detailing concepts, creators, descriptions, versions, domains, sizes and file formats.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguage	DatSize	DatName	DatPublication	DatDescription	DatVersion	DatDomain	DatFileFormat
MIT License	https://gitl	microbioma, cancer, individual, risk factors	Virginia Leombruni	Virginia Leombruni	Virginia Leombruni	english	concepts	Language.tsv	29/01/2025	Description of all languages used in the Human-Microbial Interaction Knowledge Graph	1.0	Human-Microbial Interaction	tsv

Figure 23: Language metadata

Knowledge Metadata Description

Knowledge resources metadata contain detailed descriptions of the knowledge resources used in the project, including their description, creator, language, size ect.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguage	DatName	DatPublication	DatDescription	DatVersion	DatDomain
http://creativecommons.org/licenses/by/4.0/	https://www.ebi.ac.uk/ols4/ontologies/ohmi	Biological process, interaction, microbioma-cancer interaction	European Bioinformatics Institute (EMBL-EBI)	Alexander Alekseyenko	European Bioinformatics Institute (EMBL-EBI)	english	Ontology for Host-Microbiome Interactions (OHMI)	09-ott-23	Ontology that represents the entities and relations in the domain of host-microbiome interactions	17/09/2019	Host-microbiome interactions
https://creativecommons.org/publicdomain/zero/1.0/	https://www.ebi.ac.uk/ols4/ontologies/doid	Disease, cancer	Institute for Genome Sciences, University of Maryland School of Medicine	Northwestern University	Institute for Genome Sciences, University of Maryland School of Medicine	english	Disease Ontology (DOID)	2003	Ontology that models the hierarchy of various diseases	18/12/2024	Diseases

Figure 24: Knowledge metadata

Data Metadata Description

Data resources metadata describes the datasets used or produced during the project. This section includes information about data formats, size, and availability.

DatLicense	DatURL	DatKeyword	DatPublisher	DatCreator	DatOwner	DatLanguag	DatSize	DatName	DatPublicatio	DatDescription	DatVersion	DatDomain	DatFileFormat
curatedMetagenomicDa	https://github.com/Virginia-hub-c	person_id, age, gender, BMI, ajcc, brinkman_index, alcohol_numeric	2021-10-14 YachidaS_201	YachidaS	YachidaS	english	55,5 KB	Person_Metadata.csv	29/07/2025	It contains all the relevant metadata associated with each sample, providing detailed information about their characteristics	14/06/2021	Person metadata	csv
curatedMetagenomicDa	https://github.com/Virginia-hub-c	relative_abundance, sample_id	2021-10-14 YachidaS_201	YachidaS	YachidaS	english	1,24 MB	Relative_Abundance.csv	29/07/2025	It provides data on the specific quantities of various microbial species found in each sample	14/06/2021	Relative abundance	csv
###	https://github.com/Virginia-hub-c	species_name, person_id, relative_abundance, taxonomy, interaction_id	Virginia Leombruni	Virginia Leombruni	Virginia Leombruni	english	68,3 MB	Transformed_Data_with_Taxonomy.	29/07/2025	It contains data on the relative abundance of various microbial species in different individuals, each identified by a unique person_id	1.0	Relative abundance	csv

Figure 25: Dataset Metadata

People Metadata Description

We have created an additional file, named *People_Metadata*, which compiles essential information about the project's contributors. This document provides a clear reference to the participation and affiliation of each team member.

comIdentifier	firstName	lastName	email	nationality	gender	affiliation	personalWebpage
eleonora-giuliani	Eleonora	Giuliani	eleonora.giuliani@studenti.unitn.it	Italian	F	Datascientia, Knowdive group, Universita degli Studi di Trento	https://github.com/Ele91463
virginia-leombruni	Virginia	Leombruni	virginia.leombruni@studenti.unitn.it	Italian	F	Datascientia, Knowdive group, Universita degli Studi di Trento	https://github.com/Virginia-hub-del
marc-shebaby	Marc	Shebaby	marc.shebaby@studenti.unitn.it	Lebanese	M	Datascientia, Knowdive group, Universita degli Studi di Trento	https://github.com/Marc-shebaby

Figure 26: People metadata

8 Open Issues

In conclusion, we successfully constructed a precise knowledge graph that effectively represents the interactions between individuals and their respective microbiome species, as well as interactions among species themselves. However, while our teleology includes an entity for cancer, our dataset contained only one type of cancer. This is not necessarily a limitation, as other studies can integrate additional data with multiple cancer types using the same knowledge graph.

A notable weakness of this study is the inability to statistically and significantly evaluate the results of the queries. In particular, the threshold used to classify species' relative abundances as high or low was chosen arbitrarily based on the average, making it unreliable for hypothesis-driven studies. Therefore, future research should incorporate machine learning techniques for a more comprehensive analysis of this knowledge graph, allowing for the extraction of meaningful insights from the data.

Additionally, other studies may be interested in grouping specific classes or families of microbiome species and comparing them by aggregating their relative abundances across different individuals. Another potential extension, that was present in this study due to the lack of data is the integration of time-series data for the same individuals, which could provide deeper insights into species-species interactions over time.