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).

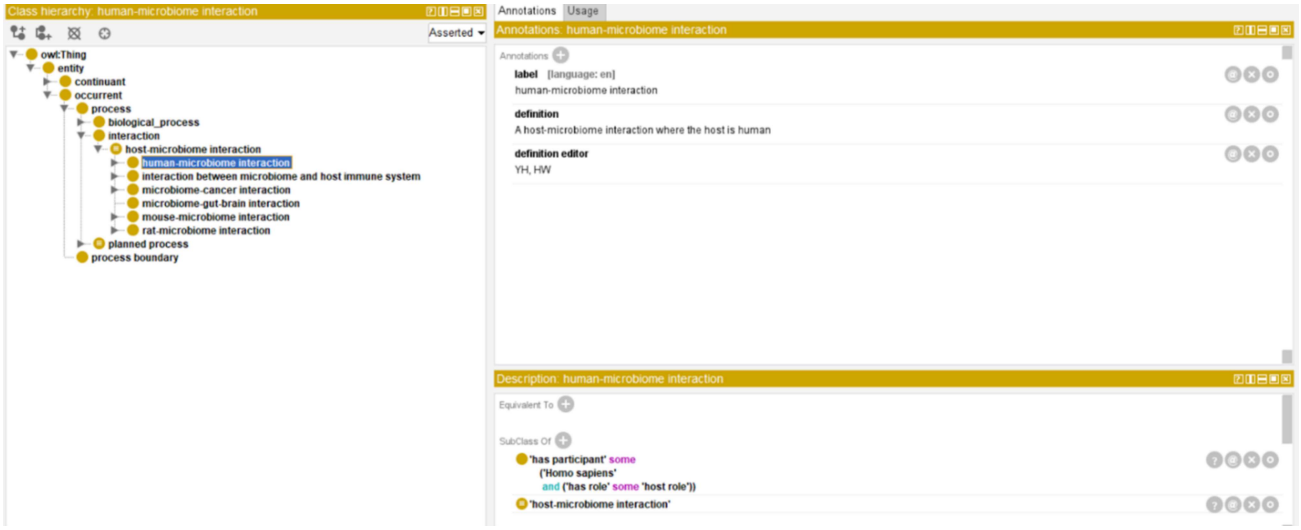

Figure 3: Located class of interest "human-microbiome interaction" in OHMI.owl (reference schema).

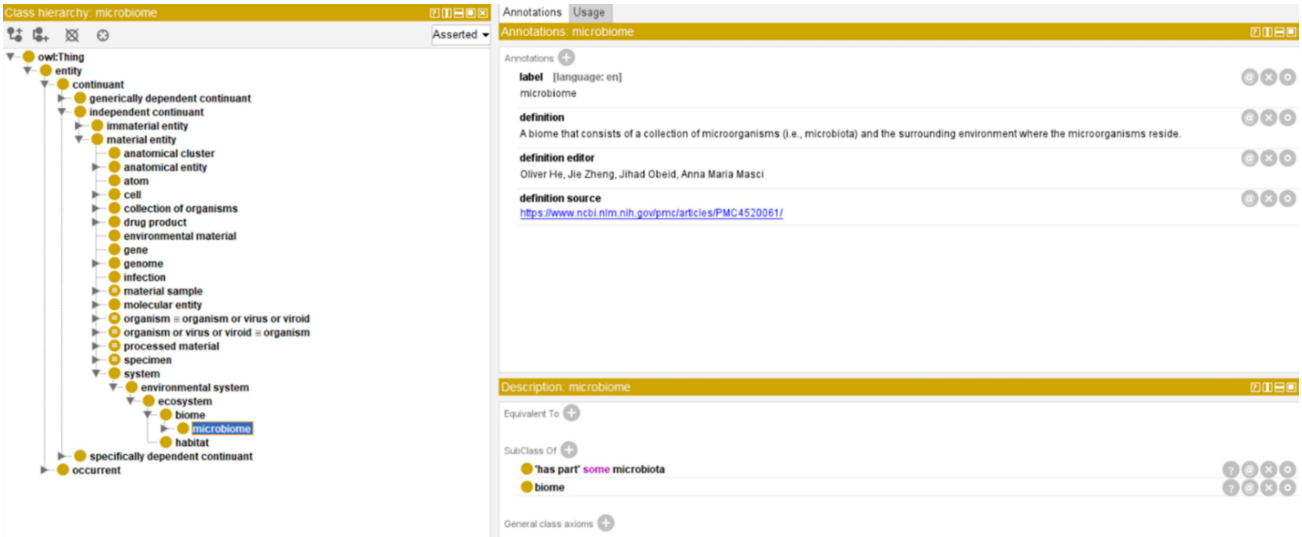

Figure 4: Microbiome Class found in the reference schema under the general class "continuant" in OHMI.owl

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).

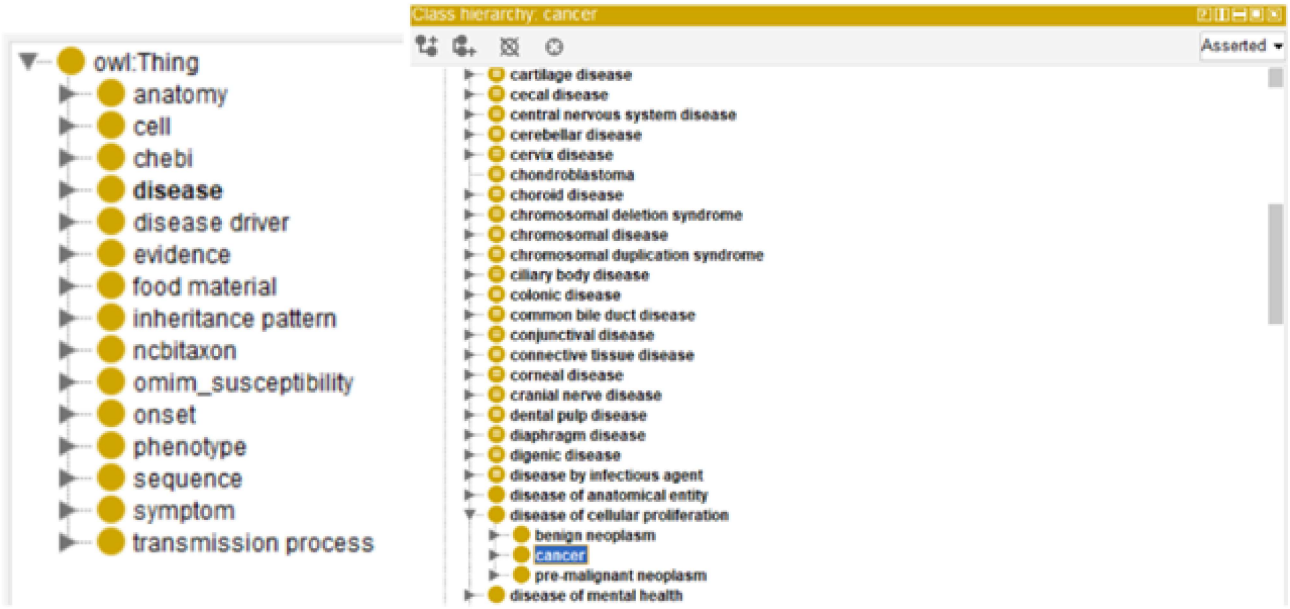

Figure 5: "Disease" as a general class in DOID.own (image on the left). "Cancer" in the DOID.owl as a subclass of "disease of cellular proliferation" (image on the right).

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:

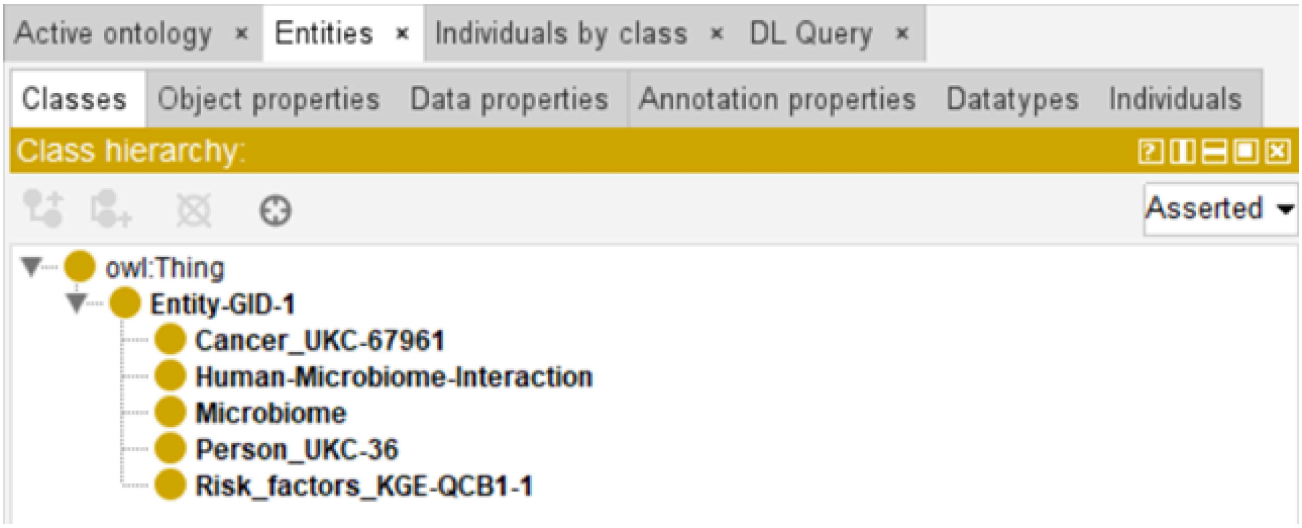

Figure 6: Class Hierarchy shows the class hierarchy under `owl:Thing`. Key classes include `Cancer`, `Human-Microbiome_Interaction`, `Risk_Factors_KGE-QCB1-1`, `Person_UKC-36`, `Microbiome`, and `Entity-GID-1`. These classes represent high-level entities relevant to the ontology.

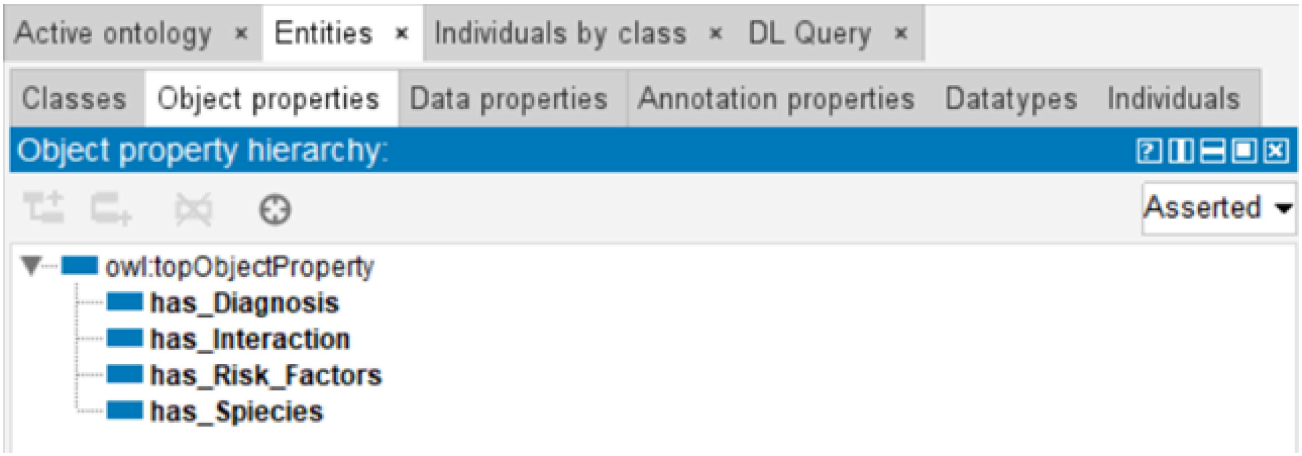

Figure 7: Object Property Hierarchy displays object properties associated with the ontology:

- **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..

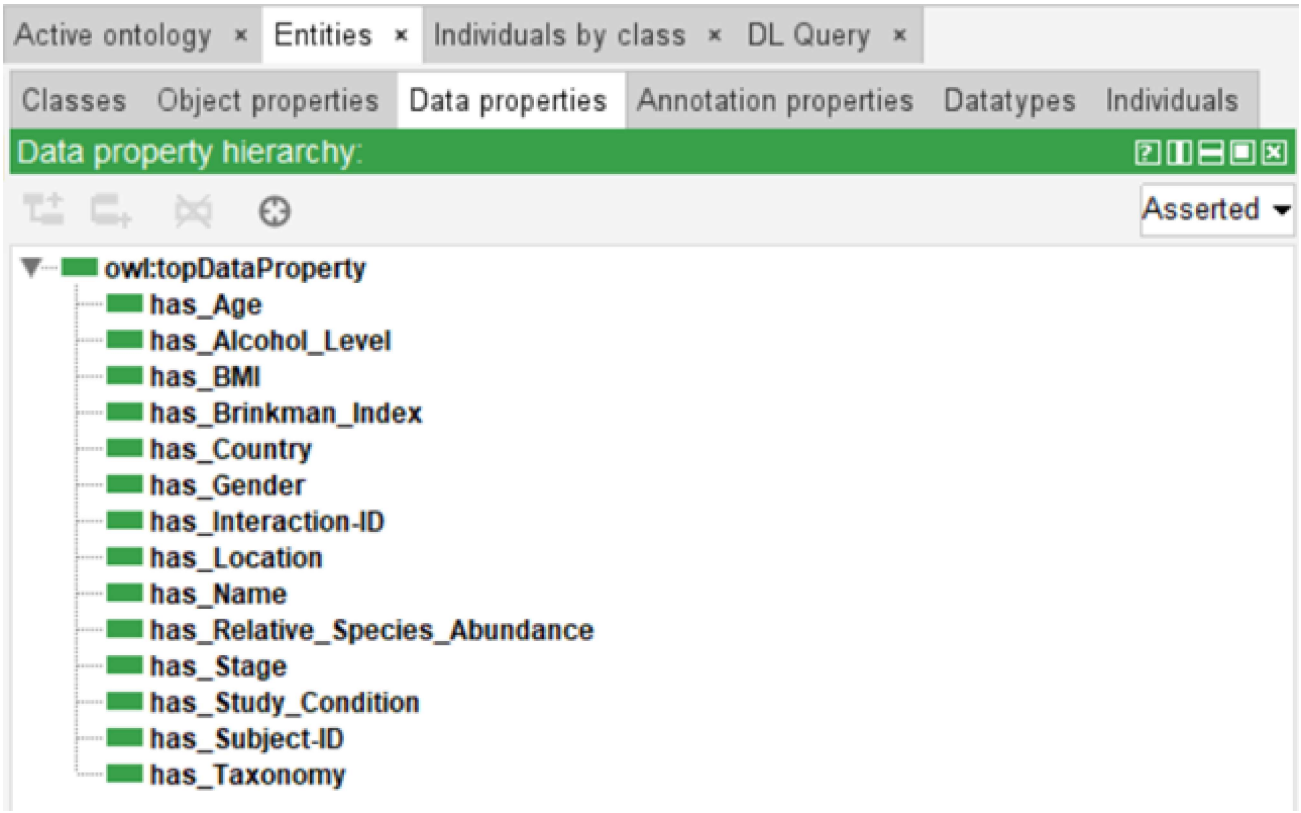

Figure 8: Data Property Hierarchy: lists data properties, which capture specific attributes and measurements:

- 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.

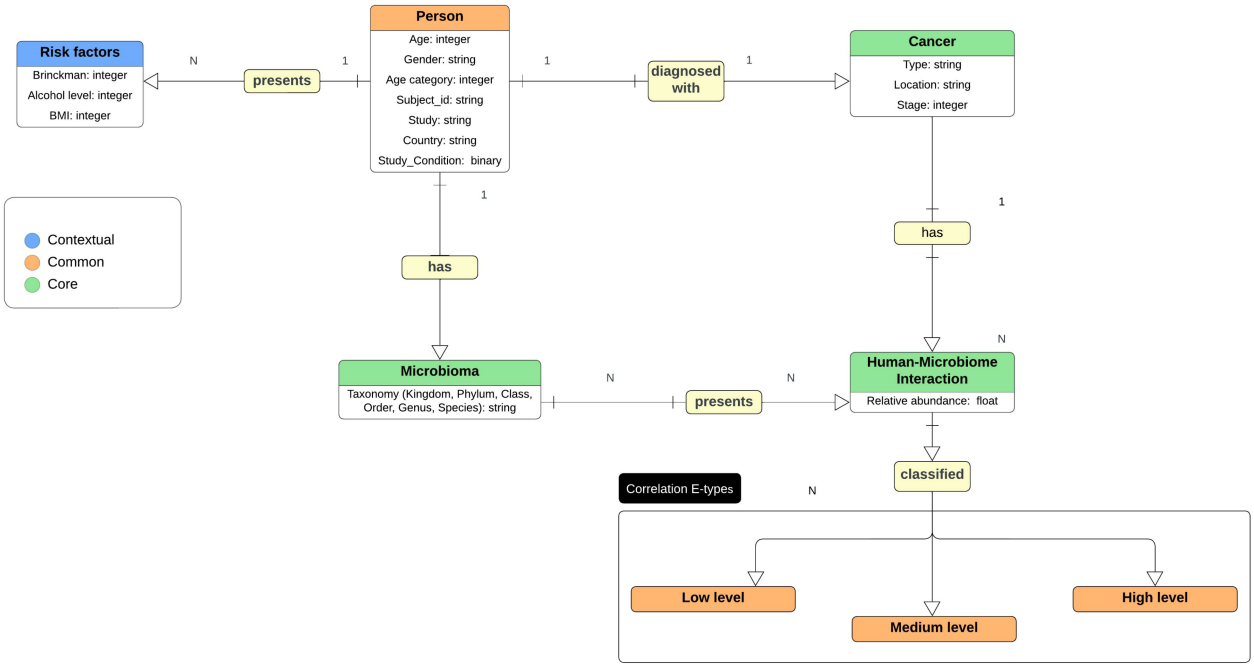

Figure 9: The ER model

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.