



UNIVERSITY  
OF TRENTO - Italy



Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

# The human microbiome and person-to-person interactions

---

Document Data:

December 8, 2024

Reference Persons:

Andrea Policano, Roan Spadazzi, Vladyslav Husak

© 2024 University of Trento

Trento, Italy

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



# **Index:**

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Purpose Definition</b>	<b>1</b>
2.1	Informal Purpose . . . . .	1
2.2	Domain of Interest . . . . .	1
2.2.1	Space . . . . .	2
2.2.2	Time . . . . .	2
2.3	Scenarios . . . . .	2
2.4	Personas . . . . .	3
2.5	Competency Questions . . . . .	3
2.6	Concepts Identification . . . . .	5
2.7	ER model definition . . . . .	5
<b>3</b>	<b>Information Gathering</b>	<b>7</b>
3.1	Data value datasets . . . . .	7
3.2	Knowledge datasets . . . . .	8
3.3	Data cleaning, filtering, standardization . . . . .	9
3.4	Finalized datasets . . . . .	10
3.5	Knowledge Layer, Schemas, Ontologies . . . . .	11
3.5.1	Dataset (from Schema.org) . . . . .	11
3.5.2	Country (from Schema.org) . . . . .	12
3.5.3	Person (from Schema.org) . . . . .	12
3.5.4	BioSample (from bioschema.org) . . . . .	12
3.5.5	Taxon (from bioschema.org) . . . . .	12
3.5.6	SGB (Species-level Genome Bin) (from the original data) . . . . .	12
3.5.7	Phenotype (from the original data) . . . . .	13
3.5.8	Transmission (from the original data) . . . . .	13
<b>4</b>	<b>Language Definition</b>	<b>13</b>
4.1	Concept Identification . . . . .	14
4.2	Dataset filtering . . . . .	16
<b>5</b>	<b>Knowledge Definition</b>	<b>17</b>
5.1	ER/EER model formalization . . . . .	18
5.1.1	EType modeling . . . . .	18
5.1.2	Object properties . . . . .	19
5.1.3	Data properties . . . . .	19
5.2	Teleontology . . . . .	20
5.2.1	ETypes . . . . .	20
5.2.2	Object properties . . . . .	21
5.2.3	Data properties . . . . .	21

5.3	Aligned Teleology . . . . .	22
5.3.1	ETypes . . . . .	22
5.3.2	Object properties . . . . .	22
5.3.3	Data properties . . . . .	23
5.4	Finalized Teleology . . . . .	23

## Revision History:

Revision	Date	Author	Description of Changes
0.1	27.10.2024	Roan	Document created
0.2	29.10.2024	All	Purpose Definition, Informal, DoI, Scenarios, Personas, CQ
0.3	02.11.2024	Andrea, Roan	Concepts Identification, ER model definition
0.4	14.11.2024	All	Dataset cleaning, filtering
0.5	16.11.2024	Roan, Vlad	Information gathering
0.6	24.11.2024	All	Language Definition
0.7	06.12.2024	All	Knowledge Definition

# 1 Introduction

In the age of big data, the ability to organize, manage, and interpret biological information has transformed how we understand complex systems like the human microbiome. This project aims to systematically organize and interlink data on the transmission of human microbiome species, social relationships, and environmental factors within a Knowledge Graph (KG). By capturing these intricate relationships, this KG will facilitate detailed queries on how microbes are shared across different populations, family structures, and social interactions, offering insights into population-specific microbiome dynamics and their potential health implications.

The project will utilize the iTelos methodology [1], a structured framework that streamlines the Knowledge Graph Engineering (KGE) process by emphasizing reusability and documentation. This approach enables the reuse of project resources for future applications, minimizing the effort required to develop new KGs for similar purposes. Following iTelos principles, the project will focus on creating reusable, well-documented resources that capture valuable data on microbiome transmission across various social and environmental contexts.

This project aims to serve as a valuable resource for researchers interested in microbiome transmission, potentially supporting further applications in public health, clinical research, and specialized education.

## 2 Purpose Definition

### 2.1 Informal Purpose

The purpose of this project is to create a system that enables users to explore the transmission dynamics of the human microbiome together with various social interactions and environmental elements that can affect it. It is done by also supporting comparative analysis (ex. different bacterial populations present in different samples) and exhibiting key transmission patterns, highlighting the influence of factors such as cohabitation and geographic location on microbial diversity and bacterial phenotypes. The resulting Knowledge Graph will be used to embed both the information about microbiome species shared between individuals, the phenotypic information of the transmitted microbial species, and the information about the social relationships between the considered individuals.

### 2.2 Domain of Interest

The Domain of Interest encompasses the composition of gut bacterial metagenomes in individuals and their transmission rates within different populations worldwide.

### 2.2.1 Space

The study spans a global geography, including diverse regions across Africa (e.g., Ghana, Tanzania, Ethiopia), the Americas (e.g., USA, Argentina, Colombia), Europe (e.g., Germany, Italy, Spain, United Kingdom, Sweden, Finland, Luxembourg), and Asia-Pacific (e.g., China, Fiji). This broad sampling from Western and non-Western countries comprehensively represents varied genetic and environmental backgrounds.

### 2.2.2 Time

The raw metagenomics datasets, which were used as the basis of this project, were constructed from **2014** to **2021**.

## 2.3 Scenarios

We identify the following possible usage scenarios for our project:

1. A microbiologist wants to study the differences in the human microbiome between different geographical ethnic groups to assess population-specific dynamics. To do so, they are looking for a way to retrieve population-specific metagenomic data that specifies, for each ethnicity, the main bacterial strains found.
2. A clinician, given an instance of a bacterial infection, is taking care of a patient presenting severe gastric problems, the patient is known to have a twin living abroad. The clinician wants to see if the reason for the health problems is associated more with the microbial environment or geographical context.
3. A university student attending a microbiology course has to make a presentation about what are the main ways in which bacteria are transmitted between individuals but is unsure of where to access the information they need straightforwardly.
4. A Dental Hygienist professor wants to teach their students the differences in transmission rates between different bacterial strains in the oral cavity. To do so, they need a one-to-one link between bacterial strains and how often they are transmitted.
5. A researcher studies social interactions and wants to measure the level of interactions by microbial transmission between individuals. He needs access to existing knowledge in a structured form about microbial transmission.

Although we identified scenarios that space as much variability as we deemed necessary, our starting focus is directed especially towards the first two scenarios (microbiologist and clinician), with the possibility of expanding in a later phase our Knowledge Graph by letting it tackle more information.

## 2.4 Personas

Based on our usage scenarios, we can distinguish between two groups of Personas:

- **Researchers & Students:**

1. Francesca - 28 - (Scenario 1) - A microbiologist doing PhD at the University of Toronto. Her scientific interest is the research of clinical significance of microbiome population in the human gut. Currently, she is researching the relationship between microbial composition in the gut and phenotypic features, therefore she wants to explore the available data on microbial transmission to identify the crucial part of the microbiome that is strongly related to phenotype and transmitted.
2. Herald - 67 - (Scenario 5) - He is a Professor of Social sciences at the University of Oxford. He is interested in research that captures social interaction with microbiome assays. He wants to revise the existing data on relationships between transition and social interactions.
3. Franco - 20 - (Scenario 3) - He is a university student following the course of Microbial Genomics, taught by Nicola Segata at the University of Trento, in the CIBIO department. After completing the theoretical section of the course, he was assigned to a group work to present a topic of interest; the workgroup settled on the topic of microbial transmission.

- **Healthcare professionals:**

4. Karmen - 32 - (Scenario 2) - an infectiologist at Maribor clinic. He is working on microbial infections and for better treatment prescription he needs to know the possible origin of a pathogen that caused the disease.
5. Juana - 46 - (Scenario 4) - She is a Dental Hygienist working in collaboration with the Sociedad Española de Microbiología. She was asked to give some lessons regarding the link between the microbial environment and dental healthcare. She wanted to highlight the effect of the transmission of different strains in the oral environment.

## 2.5 Competency Questions

Given the scenarios and personas, we created a list of CQs that would align to the previously described heterogeneity, while also avoiding unnecessary intricacy or complexification of the following ER model. Competency Questions span different scenarios and sometimes share some similarities when involved in different fields, but are heterogeneous in same-scenario situations.

### 1. Francesca, Microbiologist, PhD student

- 1.1. I'm studying the gut microbiome. Can I get a list of the most commonly transmitted bacterial species within family households?
- 1.2. What are the transmission rates of *S. parasanguinis* between Westernized and non-Westernized populations?
- 1.3. Which bacterial strains show a significant correlation with Gram staining?

1.4. Does cohabitation affect the presence of *Prevotella intermedia* in the gut microbiome?

## 2. Herald, Social Sciences Professor

2.1. How do the transmission rates differ between siblings and parents of gut bacteria?

2.2. Which transmission types are most associated with Westernized society?

2.3. Which is the most frequent microbial transmission within a family?

2.4. How does the cohabitants number influence the diversity of the gut microbiome?

## 3. Franco, university student of Microbial Genomics

3.1. I need to learn about microbial transmission. What are the primary mechanisms of bacterial transmission between individuals?

3.2. Which bacteria are more likely to be found in saliva compared to stool samples?

3.3. Which parental relationships mostly influence bacterial transmission in families?

3.4. Can I find examples of anaerobic bacterial strains commonly transmitted in Italian families?

## 4. Karmen, Infectiologist

4.1. My patient has a twin living abroad in Germany. Which of the bacterial infections is related to the environmental difference between Slovenia and Germany?

4.2. How does the gut microbiome diversity differ between patients with *S. aureus* in Slovenia?

4.3. If the patient comes to Italy which spore-forming strains should I look preferably?

## 5. Juana, Dental Hygienist

5.1. I want to teach about bacterial transmission. Which bacterial strains exhibit the highest transmission rates in the oral mouth in Westernized and non-Westernized populations?

5.2. Can I get data on how bacterial transmission rates differ between infants and older children?

5.3. How does the transmission mode between individuals affect the abundance of *Treponema Denticola* in the oral cavity?

5.4. Does the presence of other people in the household significantly affect the Dental microbiome?

## 2.6 Concepts Identification

CQ	Common entities	Core entities	Contextual entities
1.1	Family	Sample, SGB, Taxonomy	Transmission
1.2	Country	Dataset, Sample, SGB, Taxonomy	Transmission
1.3		Sample, SGB, Taxonomy	Phenotype
1.4	Country, Person	Dataset, Sample, SGB, Taxonomy	Transmission
2.1	Person, Sibling	SGB, Taxonomy	Transmission
2.2	Country	Dataset	Transmission
2.3	Person		Transmission
2.4	Person	Dataset, SGB	Transmission
3.1			Transmission
3.2		Sample, SGB, Taxonomy	Transmission
3.3	Person, Sibling	Sample, SGB	Transmission
3.4	Country	Dataset, Sampe, SGB, Taxonomy	Phenotype
4.1	Country, Sibling	Dataset, Taxonomy	Twin
4.2	Country	Dataset, Sample, SGB, Taxonomy	
4.3	Country	SGB, Taxonomy	Phenotype
5.1	Country	Dataset, Sample, SGB, Taxonomy	Transmission
5.2	Person	Dataset, Sample, SGB	Transmission
5.3		Dataset, Sample, SGB, Taxonomy	Transmission
5.4	Person	Dataset, Sample, SGB, Taxonomy	Transmission

## 2.7 ER model definition

Based on the defined Competency Questions, we created the following Entity-Relationship diagram. We took into consideration all scenarios and personas, prioritizing the microbiology and clinical areas of interest. In order to do so, we extracted from Valles-Colomer, et. al (2023) [2] the files containing the necessary data for the ER-diagram development, found in the Supplementary Tables of the article. We carefully inspected each file, while also noting down information, variables, and features that could be of interest. During the ER-diagram creation, we kept in mind the possible ways to diversify and group those features/ETypes/properties aligning them with the formalized purpose.

We define the following Entity Types and distinguish them between Common (very general ETypes that could be possibly used in other Domains of Interest and that are not necessarily bound to the purpose), Core (specific ETypes that are very relevant to the purpose and constitute

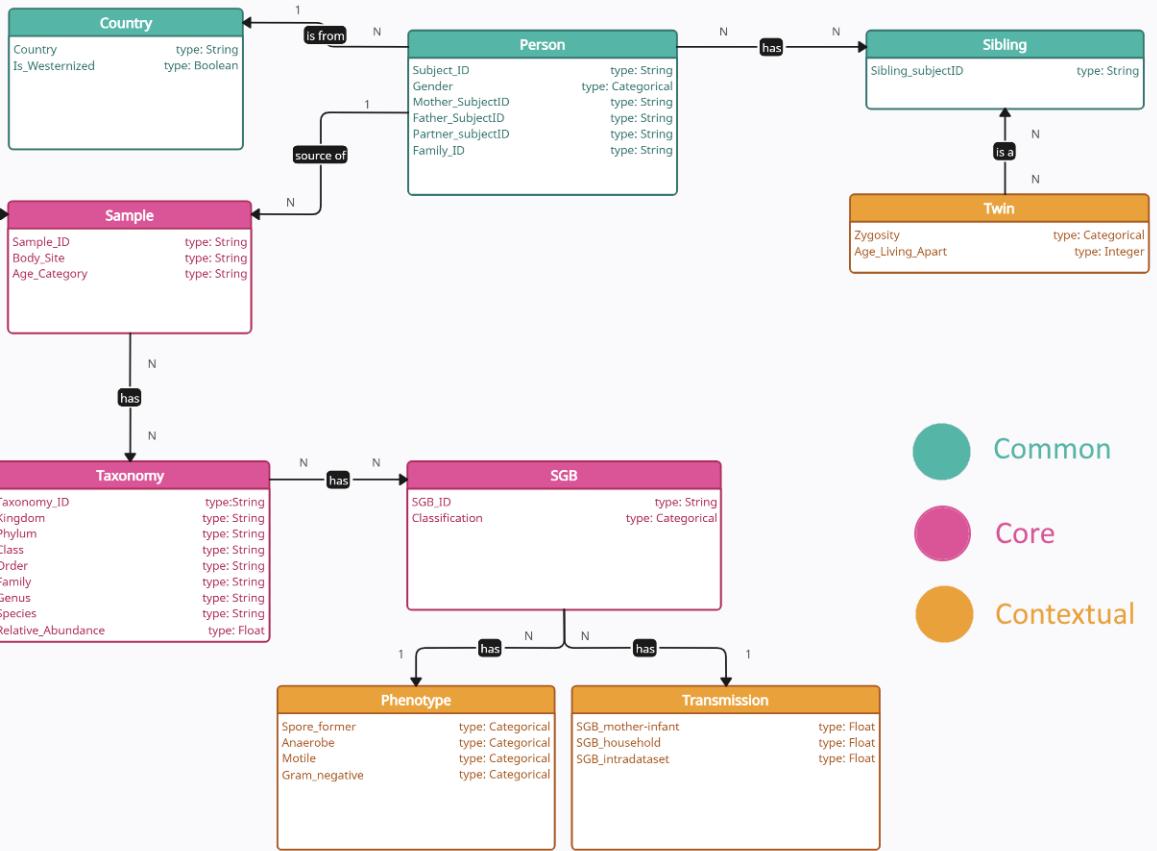


Figure 1: ER Diagram; in green the common ETypes, in red the core ones and in yellow the contextual ones.

its skeleton) and Contextual (even more specific ETypes, extremely related to the purpose that add value to it and generally would not be found in other applications used this way):

#### • Common ETypes:

- Country: a very general way to represent the "space" location of someone or something: in our case the origin of our samples/datasets. This will be useful combined with other ETypes to assess differences (for example, bacterial transmission rates) based on the origin of our dataset.
- Person: a general EType to represent someone, together with its gender and parental relationships. This is needed as the same individual can produce multiple Samples; furthermore, it is useful in order to link individuals to one another, specifying possible transmission rate differences when comparing parents and children.
- Sibling: a more specific, but still general, EType to represent brother-sister relationships. This has been specified into another EType rather than an "Individual" property to assign to it the Contextual "Twin" EType.

#### • Core ETypes:



- Dataset: a collection of Samples, under a name and a classification regarding its collection.
- Sample: our hub EType where much of the knowledge flows through: it is a collection of genetic sequences and other metadata. It is needed to tackle the most relevant competency questions and to obtain much of the data that follows.
- SGB: Species Level Genome Bins are groupings of genome data corresponding to the closest (known or unknown) microbial species (or taxonomic classification). They are derived from grouping DNA sequences obtained from metagenomic samples into bins that represent closely related genomes. Another important hub EType that will lead to other entities closer to the purpose.
- Taxonomy: a way to represent the bacterial taxonomic classification of known or unknown SGBs. It is needed to output the scientific name of the SGBs found in a sample.

- **Contextual ETypes:**

- Transmission: contains specific values associated with the transmission types and rates in different contexts for each SGB. Very relevant to our purpose in distinguishing the most/least transmitted bacterial strains.
- Phenotype: contains categorical variables that give insights into the phenotype of each SGB that are useful to answer competency questions strictly related to the phenotype itself.
- Twin: a more specific type of sibling; to our purpose it represents a way to distinguish differences in transmissibility between twins of different zygosity or years of living apart.

Regarding the relationships between ETypes, we identify straight-forward ones such as "Twin is a Sibling" or "Individual is part of Family" that tackle the family relationships and others like "SGB has Transmission/Phenotype/Taxonomy" to subdivide our purpose based on the Competency Questions we want to answer. For example, if we need to compare different phenotypes for a group of SGBs we rely on the related ETypes "SGB" and "Phenotype" (CQ 1.3, 3.4, 4.4). We can apply a similar reasoning to other "has" relationships.

## 3 Information Gathering

This section describes the second primary input for the project: the data source list. This phase will document and describe every resource (language, schema and data values) needed and used in the next steps in order to produce standardized datasets that satisfy the formalized purpose.

### 3.1 Data value datasets

All the data necessary to our purpose is derived from two sources:

- The person-to-person transmission landscape of the gut and oral microbiomes[2]: a paper regarding the human microbiome and its transmission within and across populations. It contains 35 Supplementary Tables comprising data regarding datasets of the study, samples, taxonomies, SGBs, transmission rates, phenotypes, correlation statistics and other data values.
  - **Sample Metadata and Overview (Tables S1–S2):** Contains metadata for 9,700 samples, including demographic data, body sites, and age categories. These tables also link samples to individuals and families, enabling relational analyses.
  - **Microbiome Profiling and Taxonomic Data (Tables S4–S9):** Provides microbial taxonomy data such as phylum, genus, and species classifications. It includes microbial group identifiers (SGBs), diversity indices, and profiles for specific environments like stool and saliva.
  - **Statistical Analyses and Diversity Patterns (Tables S3, S10–S19):** These tables focus on statistical tests (e.g., PERMANOVA) and diversity metrics to examine variability within and between groups. They also analyze microbial sharing rates and group-level comparisons.
  - **Transmissibility Data (Tables S9, S16, S20, S30–S32):** Includes detailed metrics on microbial transmissibility, such as mother-infant, household, and intradataset sharing. **Table S16** and **S20** evaluate transmissibility through the **Intra\_inter\_ratio** metric, showing how microbial sharing differs within and between groups. **Table S30** provides age-specific transmission rates, particularly for mother-offspring pairs.
  - **Functional Traits and Phenotypes (Tables S34–S35):** Describes microbial traits such as spore formation, motility, and anaerobic status. These features offer insights into microbial adaptability and ecological roles.
  - **Supplementary and Specialized Data (Tables S21–S29, S33):** Includes comparisons of microbial sharing across populations, environments, and relationships, such as twin studies and inter-household sharing rates.
- curatedMetagenomicData[3]: an R package containing datasets that include the relative abundance of a microbial organism (expressed with a taxonomic assignment) for each sample.

### 3.2 Knowledge datasets

- Schema.org: describing hierarchies between Entity types together with their properties.
- Schema.org LOV: schema of the schema.org vocabulary.
- Bioschemas.org: uses Schema.org markup specifically designed for Life Sciences. It includes structures named Profiles over Types together with properties that are either designed by Bioschemas or deriving from Schema.org.

### 3.3 Data cleaning, filtering, standardization

Out of these 35 tables, we decided to keep 11 (Figure 2, cleaned and modified each one in the following way:

- Table S1: containing information about the datasets and their classification; we kept only the datasets that had a corresponding one in the *curatedMetagenomicData* R package. Some studies of the paper (like the Italian and Chinese datasets) were brand new and were not imported yet to the R package due to them being unpublished. We removed the datasets having *NA* values in "PMID", then kept only the columns "Dataset", and "Dataset\_classification" (Figure 3). Some datasets (like Ghana, Tanzania) have different names in the R package but are present: this is fixed by exporting the R datasets with the same names as the ones present in the Supplementary Table 1.

Table	Description
Table S1	Summary of the 9715 samples included in the study by dataset.
Table S2	Metadata of the 9715 samples included in the study.
Table S4	List of profiled SGBs in stool samples, taxonomic classification, and strain identity thresholds.
Table S5	List of profiled SGBs in saliva samples, taxonomic classification, and strain identity thresholds.
Table S9	Gut SGB mother-infant, household, and intrapopulation transmissibility.
Table S16	Gut SGB mother-infant transmissibility (Chi2 tests, two-sided).
Table S20	Gut SGB household transmissibility and twin transmissibility (Chi2 tests, two-sided).
Table S30	Oral SGB mother-infant, household, and intrapopulation transmissibility.
Table S31	Oral SGB mother-infant transmissibility (Chi2 tests, two-sided).
Table S32	Oral SGB household transmissibility (Chi2 tests, two-sided).
Table S34	SGB predicted phenotypical traits.

Figure 2: Legend table containing information regarding the Supplementary Tables we kept.

Dataset	Dataset_classification
PasolliE_2019_Madagascar	Households
PehrssonE_2016_PER	Households
PehrssonE_2016_SLV	Households
CosteaPI_2017_KAZ	Longitudinal_set
LouisS_2016	Longitudinal_set
MehtaRS_2018	Longitudinal_set
NielsenHB_2014_ESP	Longitudinal_set
AsnicarF_2017	Mother_offspring
BackhedF_2015	Mother_offspring
ChuDM_2017	Mother_offspring
FerrettiP_2018	Mother_offspring
ShaoY_2019	Mother_offspring
TettAJ_2019_Ethiopia	Mother_offspring
WampachL_2018	Mother_offspring
YassourM_2018	Mother_offspring
Brittoli_2016	Mother_offspring and Households
Ghana	Mother_offspring and Households
Tanzania	Mother_offspring and Households
CosteaPI_2017_DEU	Mother_offspring and Households
AsnicarF_2021	Twins
XieH_2016	Twins
HMP_2019_ibdmdb	NA

Figure 3: Table S1 containing the datasets we kept together with their classification.

- Table S2: containing information about the samples and individuals of the study. We filtered out all samples (rows) that belonged to datasets we would not be using and kept the columns you can see in Figure 4.

sampleID	subjectID	body_site	age_category	gender	familyID	mother_subjectID	father_subjectID	partner_subjectID	sibling_subjectID	zygosity	age_twins	country	non_Westernized	Dataset
MV_FEI1_t1Q14	MV_FEI1	stool	<=1y	female	AsnicarF_2017_MV_1	MV_FEM1	NA	NA	NA	NA	NA	ITA	no	AsnicarF_2017

Figure 4: A row of table S2 containing the information we decided to keep for each sample. It includes sampleID and subjectID, information about the person, its parental relations, provenance and related dataset.

- Tables S4 and S5: containing information regarding SGBs and taxonomic classification of each one (respectively of stool and saliva samples). We reworked this tables in order to keep the entire taxonomy as a needed ID to link it with the tables of the R package, but also by creating new columns to represent each taxonomic level independently to tackle specific queries (ex. Kingdom, Phylum, Class etc.).

- Tables S9 and S30: containing useful information about the transmissibility of each SGB in different contexts (including mother-to-offspring, in households and intra-datasets). These tables were not changed.
- Table S34: containing information about the phenotype of each SGB (Spore\_former, Anaerobe, Motile, Gram\_negative). This table was not changed.
- Tables S16, S20, S31, S32: extra tables containing Chi-Square statistics about relatedness between SGBs and the Person's age, transmission mode vs environment, SGBs and datasets (households vs twins).

The reworked tables were all exported as `.tsv` files for standardization. Furthermore, we extracted data about the relative abundance of each bacterial species (meaning the percentage representing how much the species is present within a sample with respect to all others) for each sample using the R library `curatedMetagenomicData`. The following procedure can be repeated for each of the datasets listed in Figure 3 to get the relative abundance of a specific bacterial species, represented by a taxonomic classification (rows), for all samples (columns) of a dataset.

```
# load the R library
library(curatedMetagenomicData)
# load a dataset; curatedMetagenomicData("dataset_name.relative_abundance")
current_dataset <- curatedMetagenomicData("2021-03-31.XieH_2016.relative_abundance", dryrun = FALSE)
# load the relative abundance table
rl <- current_dataset$`2021-03-31.XieH_2016.relative_abundance`@assays@data@listData$relative_abundance
# export .tsv of the relative abundance table
write.table(rl, file = "XieH_2016.tsv", sep = "\t")
```

	YSZC12003_35365	YSZC12003_35366	YSZC12003_35387	YSZC12003_35388
k_Bacterial p_Firmicutes c_Clostridia o_Clostridiales f_Eu...	19.61446	14.37897	0.00131	
k_Bacterial p_Firmicutes c_Clostridia o_Clostridiales f_Lac...	8.39258	0.00000	0.01835	
k_Bacterial p_Firmicutes c_Clostridia o_Clostridiales f_Ru...	7.89178	6.47709	0.00100	
k_Bacterial p_Bacteroidetes c_Bacteroidia o_Bacteroidales...	7.37049	2.23896	0.00578	
k_Bacterial p_Bacteroidetes c_Bacteroidia o_Bacteroidales...	5.18316	12.39338	5.65945	
k_Bacterial p_Bacteroidetes c_Bacteroidia o_Bacteroidales...	5.05453	12.89000	4.92746	
k_Bacterial p_Bacteroidetes c_Bacteroidia o_Bacteroidales...	4.93714	3.87848	3.49791	
k_Bacterial p_Firmicutes c_Clostridia o_Clostridiales f_Ru...	4.36169	1.89050	1.72699	
k_Bacterial p_Firmicutes c_Clostridia o_Clostridiales f_Lac...	3.02199	0.19740	0.05757	

Figure 5: A snapshot of the datasets deriving from the R library `curatedMetagenomicData`. The rows represent taxonomical classifications, the columns the sample IDs that are dataset-specific.

### 3.4 Finalized datasets

After cleaning, exporting and standardizing all of our data, we end up with the following datasets:

File	Source	Content	Related ETypes
datasets.tsv	Table S1	Dataset and classification	Dataset
samples.tsv	Table S2	Sample Metadata	Sample, Person, Country, Sibling, Twin
taxonomy_SGB_stool.tsv	Table S4	Stool SGBs	Taxonomy, SGB
taxonomy_SGB_saliva.tsv	Table S5	Saliva SGBs	Taxonomy, SGB
transmission_rates_1.tsv	Table S9	Transmission rates	Transmission
transmission_rates_2.tsv	Table S30	Age-specific transmission rates	Transmission
phenotypes.tsv	Table S34	Phenotype of each SGB	Phenotype
SGB_age.tsv	Table S16	Relatedness SGB-age	SGB
SGB_transmission.tsv	Table S20	Relatedness SGB-transmission	SGB
transmission_env.tsv	Table S33	Relatedness transmission-environment	Transmission

Moreover, there is a `.tsv` file for each dataset listed in `datasets.tsv`. For example, "`XieH_2016.tsv`" will contain data about the dataset named XieH\_2016, specifically the relative abundance of each species for each sample of the dataset. In total, we have 22 datasets that are also present in the `curatedMetagenomicData` R library. These datasets were added in a second moment with respect to the ones present in the previous table as we noticed we were missing a very important link present in our ER model: the one between the Sample and the Taxonomy. Specifically, we needed a straight-forward way to have, for each sample, all the species that were found inside it: this is given by the relative abundance, a numeric value that measures the percentage of a bacterial species inside a sample. Of course, this gives us the possibility to also say whether a species is present in a sample or not just by looking at the percentages (0% meaning that that specific species is not present in the sample).

### 3.5 Knowledge Layer, Schemas, Ontologies

The cleaned dataset contains various entities and properties essential for analyzing microbial transmission and diversity. For most entity types, no existing schema from Schema.org or Bioschemas.org is available, so we use the schema derived directly from the dataset. However, for some types, we align with existing standards where applicable.

#### 3.5.1 Dataset (from Schema.org)

- **name** (Text, from Schema.org): Name of the dataset.
- **classification** (Text, from the original data): Classification of the dataset.

### 3.5.2 Country (from Schema.org)

- **name** (Text, from Schema.org): Country associated with samples or individuals.
- **isWesternized** (Boolean, from the original data): Whether the country is considered Westernized.

### 3.5.3 Person (from Schema.org)

- **identifier** (Text, from Schema.org): Identifier for an individual.
- **gender** (Text, from Schema.org): Gender of the individual.
- **mother\_subjectID, father\_subjectID** (Text, from the original data): Identifiers for parents.
- **sibling\_subjectID** (Text, from the original data): Identifier for siblings.
- **familyID** (Text, from the original data): Identifier linking the individual to a family.

### 3.5.4 BioSample (from bioschema.org)

- **identifier** (Text, from Bioschemas.org): Unique identifier (sampleID) for a sample.
- **samplingAge** (Text, from bioschema.org): The age of the object when the Sample was created.
- **bodySite** (Text, from the original data): Anatomical site where the sample was collected.

### 3.5.5 Taxon (from bioschema.org)

- **identifier** (Text, from Bioschemas.org): Unique identifier for the taxonomic entity.
- **scientificName** (Text, from Bioschemas.org): The currently valid scientific name of the taxon.
- **alternateScientificName** (Text, from Bioschemas.org): Synonym or alternate scientific name for the taxon, if available.
- **parentTaxon** (Text, from Bioschemas.org): Closest parent taxon of the taxon in question.
- **childTaxon** (Text, from Bioschemas.org): Closest child taxa of the taxon in question.
- **taxonRank** (Text, from Bioschemas.org): The taxonomic rank of this taxon.
- **relativeAbundance** (Float, from the original data): Relative abundance of the taxon in the sample.

### 3.5.6 SGB (Species-level Genome Bin) (from the original data)

- **SGB\_ID** (Text): Identifier for the species group bin.
- **classification** (Text): Type of SGB classification (e.g., kSGB, uSGB).

### 3.5.7 Phenotype (from the original data)

- **sporeFormer** (Boolean): Indicates if the microbial species forms spores.
- **anaerobe** (Boolean): Indicates if the microbial species is anaerobic.
- **motile** (Boolean): Indicates if the microbial species is motile.
- **gramNegative** (Boolean): Indicates if the microbial species is Gram-negative.

### 3.5.8 Transmission (from the original data)

- **motherInfantTransmissibility** (Number): Rate of transmission from mother to infant.
- **householdTransmissibility** (Number): Rate of transmission within a household.
- **intralInterRatio** (Number): Ratio comparing intra-group and inter-group transmissibility.

This schema reflects the dataset's structure while incorporating standards from Schema.org and Bioschemas.org where applicable, providing a robust framework for representing microbiome-related entities and relationships.

## 4 Language Definition

The language definition phase has, as main objective, to formally define the concepts and information that will be present and used in the final Knowledge Graph to satisfy the project purposes. In order to do so, we have to consider and identify all the concept structures such as ETypes, relations and the object properties to be formally defined. This operation was carried out with the help of the Universal Knowledge Core (UKC) [4], aligning each identified concept with the integrated search engine to examine if the concept was already present or not. The results of this section can be seen in Table 1. If a concept was not found in UKC, it was either looked into using a different ontology or, if still not found, it was identified and defined by us. Due to the context of the project, some of the elements taken into account are very purpose-specific (for example, the EType Transmission and its properties) and therefore had to be fully defined and specified. Since our study presents many words and concepts that are specific for the biological field, we had to search for other ontologies to retrieve the needed information with which we completed our Language Resource table. We identified 3 ontologies, all coming from NCBO BioPortal [5], depending on the concepts that still needed identification and definition:

- **D3O**: the DSMZ Digital Diversity Ontology, used to retrieve details related to microbial-specific phenotypes.
- **GENO**: the Genotype Ontology, used to represent the genetic variations described in genotypes, linking them to phenotypes; in our case the twin's zygosity.
- **OHMI**: the Ontology of Host-Microbe Interactions, a biomedical ontology used to represent relationships between host-microbe interactions; in our case, the bacteria's relative abundance.

## 4.1 Concept Identification

In the end we were able to produce the Language Resource Table in which each identified concept was listed and represented in a table having 3 columns:

- Column 1: ConceptID; it has three different "formats" based on whether the word was found in UKC (ID: UKC-number), found in other ontologies (ID: link) or defined it ourselves (ID: KGE24-QCB2-number).
- Column 2: containing the labels or words associated to the concept
- Column 3: containing the glossary/definition for each word

Furthermore, we decided to color-code our table for better visualization in the following way:

- In black the concepts aligned and found in UKC;
- In blue those aligned and not found in UKC but in different ontologies;
- In red those concepts not found in any other resource and that we defined ourselves;
- In bold we can see the E-Types, in *italics* the relationships and as plain simple text the Properties.

ConceptID	Word-en	Gloss-en
https://purl.dsmz.de/schema/Dataset	<b>Dataset</b>	A structured collection of data, often presented in tabular or other formats, that is used for analysis, research, or reference purposes. Datasets can consist of numerical, textual, or categorical information
UKC-2	Name	a language unit by which a person or thing is known
UKC-42540	Classification, Group	a group of people or things arranged by class or category
UKC-46463	<b>Sample</b>	all or part of a natural object that is collected and preserved as an example of its class
UKC-26728	Age_Category, Age	how long something has existed
KGE24-QCB2-1	<b>Body Site</b>	extraction point of a sample from a person
UKC-36	<b>Person</b>	a human being
UKC-27174	Gender	the properties that distinguish organisms on the basis of their reproductive roles
UKC-51131	Mother	a woman who has given birth to a child (also used as a term of address to your mother)
UKC-49598	Father	a male parent (also used as a term of address to your father)
UKC-52872	Partner	a person's partner in marriage
UKC-43042	Family	a social unit living together
UKC-45187	<b>Country</b>	the territory occupied by a nation
KGE24-QCB2-2	<b>Westernized</b>	a population that has adopted Western cultural norms, values, and lifestyles.
UKC-52619	<b>Sibling</b>	a person's brother or sister

UKC-53445	<b>Twin</b>	either of two offspring born at the same time from the same pregnancy
<a href="http://purl.obolibrary.org/obo/GENO_0000133">http://purl.obolibrary.org/obo/GENO_0000133</a>	<b>Zygosity</b>	An allelic state that describes the degree of similarity between features in a 'single locus complement', within the genome of a cell or organism (i.e., whether the alleles or haplotypes that reside at the same location on paired chromosomes are the same or different)
KGE24-QCB2-3	<b>Age Living Apart</b>	the age twins started to live apart
UKC-44477	<b>Taxonomy</b>	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 Protocista or Plantae or Fungi or Animalia
UKC-43156	Phylum	(biology) the major taxonomic group of animals and plants; contains classes
UKC-43160	Class	(biology) a taxonomic group containing one or more orders
UKC-43163	Order	(biology) taxonomic group containing one or more families
UKC-43166	Family	(biology) a taxonomic group containing one or more genera
UKC-43171	Genus	(biology) taxonomic group containing one or more species
UKC-43176	Species	(biology) taxonomic group whose members can interbreed
<a href="http://purl.obolibrary.org/obo/OHMI_0000468">http://purl.obolibrary.org/obo/OHMI_0000468</a>	<b>Relative Abundance</b>	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
KGE24-QCB2-4	<b>SGB</b>	Species-level Genome Bins - microbial genomes from metagenomic data, representing species-level groupings
UKC-26778	<b>Phenotype</b>	what an organism looks like as a consequence of the interaction of its genotype and the environment
<a href="https://purl.dsmz.de/schema/SporeFormation">https://purl.dsmz.de/schema/SporeFormation</a>	<b>Spore former</b>	The ability of certain microorganisms to form spores, a resistant structure that allows survival in adverse conditions. Spore formation is a significant trait for microbial classification and survival strategies
UKC-6534	Anaerobe	an organism (especially a bacterium) that does not require air or free oxygen to live
UKC-82548	Motile	(of spores or microorganisms) capable of movement
<a href="https://purl.dsmz.de/schema/GramNegativeBacteria">https://purl.dsmz.de/schema/GramNegativeBacteria</a>	<b>Gram negative</b>	Bacteria that do not retain the crystal violet stain used in Gram staining. Gram-negative bacteria have an outer membrane and are often resistant to certain antibiotics

KGE24-QCB2-5	<b>Transmission</b>	the transfer of microorganisms (bacteria, viruses, fungi, etc.) between individuals, environments, or species, influencing their microbiome composition
KGE24-QCB2-6	<b>mother-infant transmission</b>	the microbial transmission between mother and her infant (i.e. transmissibility)
KGE24-QCB2-7	<b>household transmission</b>	the microbial transmission within people living in the same dwelling (i.e. transmissibility)
KGE24-QCB2-8	<b>intradataset transmission</b>	the microbial transmission within metagenomics samples of same dataset (i.e. transmissibility)
KGE24-QCB2-9	<i>is from</i>	a person having origin, association, or ownership related to a source, place, or entity
KGE24-QCB2-10	<i>source of</i>	a person being the origin of one or more samples
KGE24-QCB2-11	<i>has phenotype</i>	have or possess a phenotype
KGE24-QCB2-12	<i>has sample</i>	a dataset having or possessing one or more samples
KGE24-QCB2-13	<i>has SGB</i>	have or possess a SGB
KGE24-QCB2-14	<i>has sibling</i>	a person having one or more siblings
KGE24-QCB2-15	<i>has taxonomy</i>	have or possess a taxonomy
KGE24-QCB2-16	<i>has transmission</i>	have or possess a transmission type

Table 1: This table represents the purpose-specific concepts, formalized and defined. The terms in **bold** represent the ETypes, the ones in *italics* the relations and all the others are the properties of the corresponding ETypes. Furthermore, the words in black are the ones also present in UKC (and have the UKC-ID), the ones in **blue** are present in other ontologies (with the ID being the link to that concept) and the ones in **red** are defined by us (and have ID: KGE-QCB2-number).

## 4.2 Dataset filtering

Dataset filtering aims at aligning all the concepts previously identified with the data layer resources (3.1). This also means filtering out all resources or elements that are not formally defined. In our case, this step is trivial as we aimed to identify (and put in our language resource table) all concepts that are found in our data layer. To be more precise, we go through each table presented in 3.4:

- *dataset.tsv*: fully defined with "Dataset", "Name", and "Classification".
- *samples.tsv*: fully defined ("Sample", "Age", "Body Site", "Person", "Gender", "Mother", "Father", "Partner", "Family", "Country", "Westernized", "Sibling", "Twin", "Zygosity", "Age Living Apart").
- *taxonomy\_SGB\_stool.tsv* and *taxonomy\_SGB\_saliva.tsv*: fully defined ("Taxonomy", "Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species", "SGB", "Classification").
- *transmission\_rates\_1.tsv* and *transmission\_rates\_2.tsv*: fully defined ("Transmission", "Age", "mother-infant transmission", "household transmission", "intradataset transmission").

- *phenotypes.tsv*: fully defined ("Phenotype", "Spore former", "Anaerobe", "Motile", "Gram negative").
- *SGB\_age.tsv*, *SGB\_transmission.tsv*: fully defined ("Age", "SGB", "Classification").
- *transmission\_env.tsv*: not defined as this dataset was dropped; it did not provide relevant information to properly answer the Competency Questions.
- *(dataset\_name).tsv*: each is fully defined ("Taxonomy", "Name", "Sample", "Relative abundance").

## 5 Knowledge Definition

Once all the sub-tasks related to the Language Definition phase (4) were completed, the next step in the project was to define the knowledge behind our KG and the data associated with it. Our aim is formalizing our ER/EER model, creating the Teleontology related to the KG, and align both to produce a final object created by the overlapping of the two: the Teleology.

In this phase, we implement the kTelos methodology, which is fundamental for the reuse of the knowledge resources and their implementation. The final aim of this part is to unify all our representations of the information we are dealing with as much as possible and enhance the interoperability of the final KG. Being already in possession of the ER/EER model, the main problem encountered in this step of the study is the absence of a Teleontology. This type of ontology is able to integrate the concept of teleology while focusing mainly on the end goals, organizing the structure of the knowledge in different ways:

- Hierarchical structured way via IS-A/PART-OF relationships
- Connections established by object properties
- EType-related data properties

Our approach to creating a Teleontology from scratch was to apply what we learned in the previous sections of the project: to build it we needed the concepts found in schema.org, grouping the entities and properties found to later be aligned against the formalized ER/EER model. To carry out the previously described steps, the Protégé [6] tool was used, a free and open-source software for the creation and handling of ontology-based applications. We align [7] the formalized ER/EER model, stored as an OWL file, to the created Teleontology, producing an OWL output file containing the final Teleology by:

- Identifying the leaf ETypes,
- Dropping the ETypes for which we do not have data and the more general ones,
- Adding to the leaves all the corresponding purpose-specific information and properties.

## 5.1 ER/EER model formalization

Following the EER model specified previously (fig. 1) we formalize it using Protégé by modeling ETypes (classes), objects, and data properties to create a final OWL file. With respect to the original EER, the following things have been changed in the Protégé implementation to align the knowledge layer with the language one:

- The data property "Country" in the EER was renamed to "Name"
- All data properties are standardized: they start with a lowercase character and are separated by "\_" if multiple words are present, especially noticeable for some ID properties (for example, Sibling\_SubjectID became sibling\_subject\_ID).

Once these edits were taken into account, we formalized our ER/EER model on Protégé, starting with ETypes, then object properties (relations), and finally data properties.

### 5.1.1 EType modeling

The first thing done was creating a "parent" EType common for all the ones we want to model named "Entity\_GID-1", then, for each EType found in the informal ER/EER model we:

- Create a new subclass of "Entity\_GID-1", sibling of the other ETypes (with the possibility of having more hierarchical relations; for example, Twin is a Sibling, therefore a subclass of the latter) and name it with the corresponding EType name also found in the Language Resource Sheet;
- Rename its IRI (Internationalized Resource Identifier) as [http://knowdive.disi.unitn.it/etype#<EType\\_name>](http://knowdive.disi.unitn.it/etype#<EType_name>);
- Annotate it with the definition and concept identifier found in the Language Resource Sheet;
- Creating a new boolean annotation "isEType" and setting it to True.

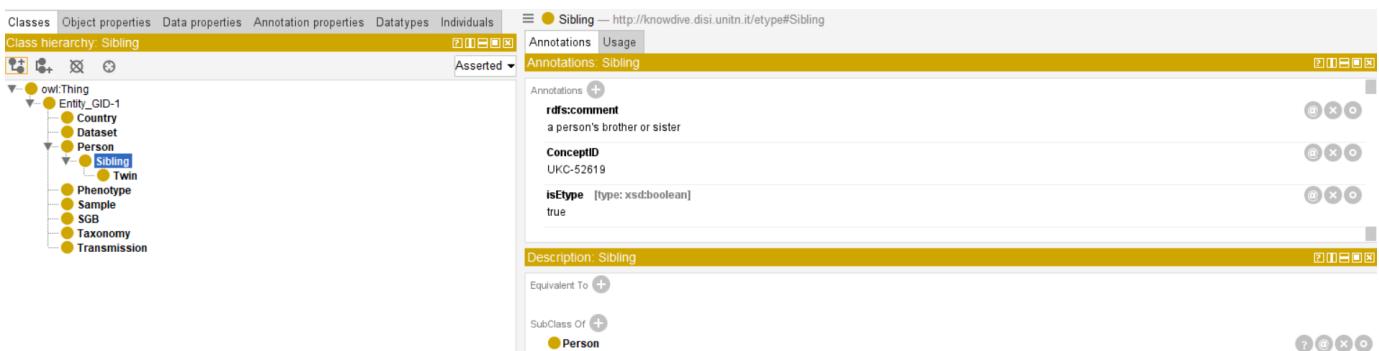


Figure 6: A snapshot of the ETypes modeled in Protégé. Sibling is highlighted; notice the annotations for the definition and the concept ID and the sub-super class relationships.

### 5.1.2 Object properties

Regarding the object properties (the relations between ETypes), we follow a similar reasoning:

- We create a new sibling object property with the corresponding name found in the language definition formatted as `has_<name>`;
- We define is a definition and the Concept ID;
- Rename its IRI to `http://knowdive.disi.unitn.it/etype#has_<property-name>`;
- We set the Domain(s) and Range(s) of each object property (meaning the EType it starts from and the one it ends in).

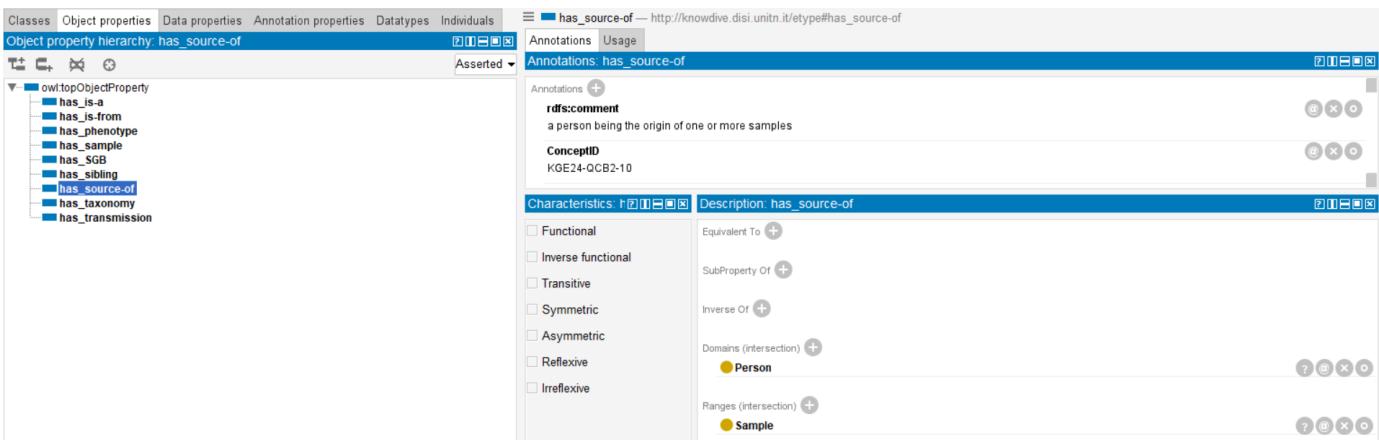


Figure 7: A snapshot of the Object Properties modeled in Protégé. "Source-of" is highlighted; notice the annotations, domain, and range, in this case, a Person is a source of Sample(s).

### 5.1.3 Data properties

For each data property found in the informal ER/EER:

- We create a new sibling data property with the corresponding name found in the language definition formatted as `has_<name>`;
- We define is a definition and the Concept ID;
- Rename its IRI to `http://knowdive.disi.unitn.it/etype#has_<property-name>`;
- We set the Domain(s) of the property as the EType it belongs to;
- We set the Range(s) of the property as the type this variable can have (for example, Boolean, String, Float...).

The screenshot shows the Protégé interface with the following panels:

- Left Panel (Data property hierarchy):** Shows the 'owl:topDataProperty' node expanded, listing numerous properties under 'has\_age\_category' such as 'has\_age\_category', 'has\_age\_living\_apart', 'has\_anerobe', etc.
- Middle Panel (Annotations):** For the 'has\_age\_category' property, it shows annotations like 'rdfs:comment' (how long something has existed) and 'ConceptID' (UKC-26728).
- Right Panel (Description):** For the 'has\_age\_category' property, it shows characteristics like 'Functional', 'Equivalent To', 'SubProperty Of', 'Domains (Intersection)' (Sample), 'Ranges' (xsd:string), and 'Disjoint With'.

Figure 8: A snapshot of the Data Properties modeled in Protégé. "Age Category" is highlighted; notice the annotations, the domain (in this case, a Sample has an Age Category), and the range (Age Category is a string).

## 5.2 Teleontology

Due to the specificity of our project domain, no dedicated teleontology exists. Therefore, we decided to build it starting from the schema.org ontology, which also contains bioschemas.org ontology inside. Because the chosen teleontology is quite redundant we will not go through all its ETypes and properties but just showcase its general structure.

### 5.2.1 ETypes

Schema.org ETypes have a very branched structure with a large number of subclasses. However, the ETypes we are interested in in the alignment process usually are high-level ones.

The screenshot shows the Protégé interface with the following panels:

- Left Panel (Class hierarchy):** Shows the 'owl:Thing' node expanded, listing various ETypes like 'rdfs:Class', 'Thing', 'Action', 'BioChemEntity', etc., with 'Taxon' highlighted.
- Middle Panel (Annotations):** For the 'Taxon' EType, it shows annotations like 'rdfs:label' ([language: en] Taxon), 'rdfs:comment' (A set of organisms asserted to represent a natural cohesive biological unit.), and 'rdfs:isDefinedBy' (<https://pending schema.org/Taxon>).

Figure 9: A snapshot of the ETypes present in teleontology (Protégé). "Taxon" is highlighted; this EType is a part of bioschemas.org, and will be further aligned with EER.

### 5.2.2 Object properties

Schema.org has a large number of Object Properties linking all the ETypes that are part of its skeleton. Most of them are irrelevant in our case and will be dropped during the alignment.

The screenshot shows the Protégé ontology editor. On the left, the 'Object property hierarchy' for 'accessibilityControl' is displayed, with many properties listed under 'owl:topObjectProperty'. In the center, the detailed view for 'accessibilityControl' is shown, including its rdfs:label ('accessibilityControl'), rdfs:comment ('Identifies input methods that are sufficient to fully control the described resource. Values should be drawn from the <a href="https://www.w3.org/2021/rdfy-discov-vocab/latest#accessibilityControl-vocabulary">approved vocabulary</a>.'), and rdfs:isDefinedBy ('accessibilityControl'). On the right, the 'Characteristics' tab is selected, showing options like Functional, Inverse functional, Transitive, Symmetric, Asymmetric, Reflexive, and Irreflexive. Below these are sections for Domains (intersection) and Ranges (intersection), both currently set to 'CreativeWork' and 'Role or Text or URL' respectively. A note at the bottom indicates 'No Reasoner set. Select a reasoner from the Reasoner menu'.

Figure 10: A snapshot of the Object Properties present in teleontology (Protégé). "accessibilityControl" is highlighted; this property belongs to core schema.org properties.

### 5.2.3 Data properties

Schema.org has a significant number of Data Properties. None of them will directly be aligned to our formalized EER.

The screenshot shows the Protégé ontology editor. On the left, the 'Data property hierarchy' for 'availabilityStarts' is displayed, with many properties listed under 'owl:topDataProperty'. In the center, the detailed view for 'availabilityStarts' is shown, including its rdfs:label ('availabilityStarts'), rdfs:comment ('The beginning of the availability of the product or service included in the offer.'), and rdfs:isDefinedBy ('availabilityStarts'). On the right, the 'Characteristics' tab is selected, showing options like Functional, Equivalent To, SubProperty Of, Domains (intersection), Ranges, and Disjoint With. A note at the bottom indicates 'No Reasoner set. Select a reasoner from the Reasoner menu'.

Figure 11: A snapshot of the Data Properties present in teleontology (Protégé). "availabilityStarts" is highlighted; this property belongs to core schema.org properties.

## 5.3 Aligned Teleology

We aligned the Teleontology and our EER by merging corresponding ETypes and linking the corresponding properties with annotations. We then deleted all non-leaf ETypes and kept only the ones for which we have data. The final teleology is described below.

### 5.3.1 ETypes

We merged the ETypes of our EER with ones that correspond to them in teleontology.

- We merged Taxonomy (schema.org) with Taxon (our EER)
- We merged Dataset (schema.org) with Dataset (our EER)
- We merged Country (schema.org) with Country (our EER);
- We merged Person (schema.org) with Person (our EER);.

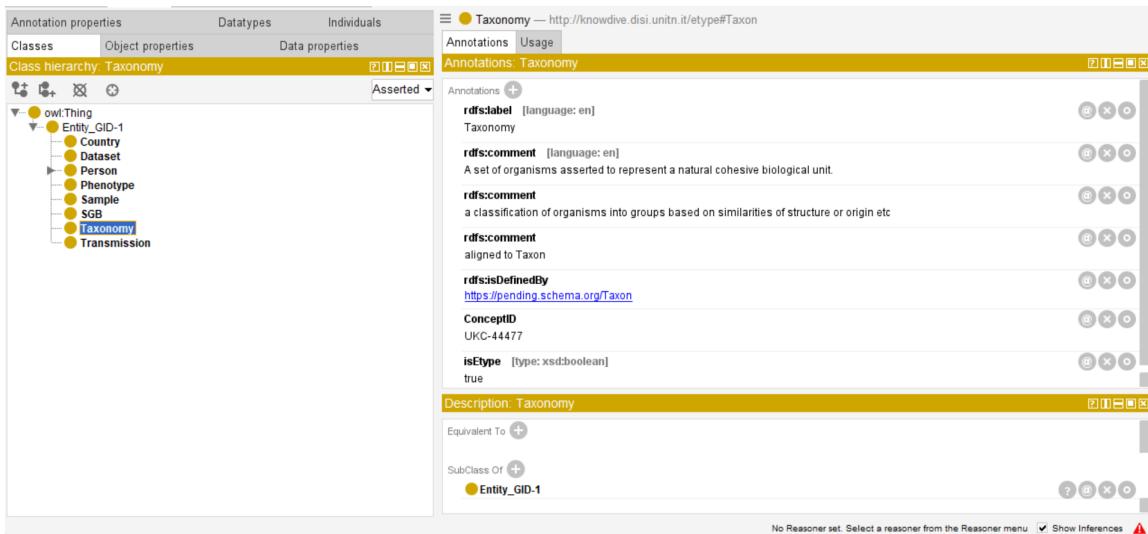


Figure 12: A snapshot of the ETypes present in the teleology (Protégé). "Taxonomy" is highlighted; this property was aligned with Taxon from schema.org.

### 5.3.2 Object properties

There are 3 Object Properties aligned.

- We aligned sibling (schema.org) with has\_sibling (our EER)
- We aligned fromLocation (schema.org) with has\_is-from (our EER)
- We aligned equal (schema.org) with has\_is-a (our EER);

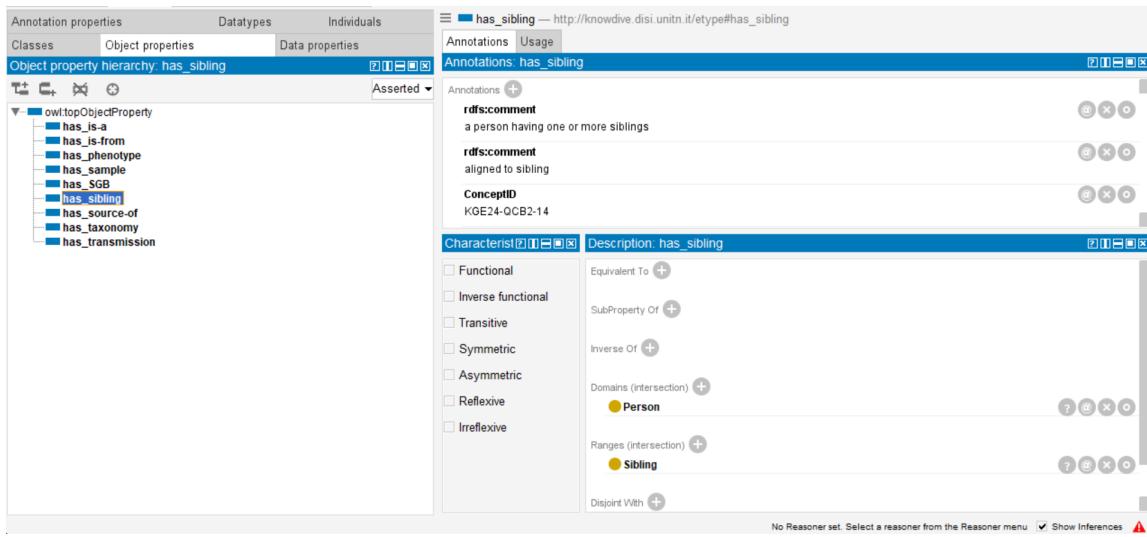


Figure 13: A snapshot of the Object Properties present in the teleology (Protégé). "has\_sibling" is highlighted; this property was aligned with the sibling from schema.org.

### 5.3.3 Data properties

There were no corresponding Data Properties in the schema.org to our EER. So the Data Properties of our EER remain unchanged.

## 5.4 Finalized Teleology

Our final teleology has 11 ETypes, 9 object properties and 32 data properties. All have their ConceptID and definition annotated, together with their alignment status, domain and range. The OWL file was exported in .rdf extention and uploaded on GitHub [8] ("teleology\_KGE24\_QCB2.rdf").

Metrics	
Axiom	451
Logical axiom count	92
Declaration axioms count	222
Class count	11
Object property count	9
Data property count	32
Individual count	165
Annotation Property count	8

Figure 14: Metrics of the finalized Teleology.

## References

- [1] F. Giunchiglia, S. Bocca, M. Fumagalli, M. Bagchi, and A. Zamboni. iTelos - Purpose Driven knowledge graph generation. 2021.
- [2] Valles-Colomer et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature*, 614:125–135, 2023.
- [3] Pasolli, E., Schiffer, L., Manghi, and P. et al. Accessible, curated metagenomic data through experimenthub. *Nat Methods*, (14):1023–1024, 2017.
- [4] Giunchiglia F., Bella G., and Nair N.C. et al. Representing interlingual meaning in lexical databases. *Artificial Intelligence Review*, 56:11053–11069, 2023.
- [5] Whetzel PL, Noy NFand Shah NH, Alexander PR, Nyulas C, Tudorache T, and Musen MA. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, (W541-5), 2011.
- [6] Musen MA and Protégé team. The protégé project: A look back and a look forward. *AI matters*, 1(4):4-12, 2015.
- [7] Fausto Giunchiglia and Daqian Shi. Property-based entity type graph matching. *arXiv*, 2021.
- [8] Github repository of the project: [https://github.com/boreico/kge\\_qcb\\_project/tree/main](https://github.com/boreico/kge_qcb_project/tree/main), 2024.