UNIVERSITY
OF TRENTO - Italy

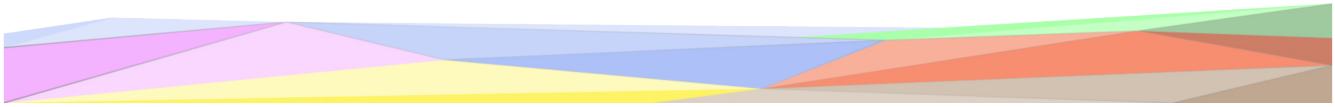
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2024 - The human microbiome and person-to-person interactions

---

Document Data:	Reference Persons:
February 8, 2025	Andrea Policano, Roan Spadazzi, Vladyslav Husak



Index:

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, Vladyslav	Information gathering
0.6	24.11.2024	All	Language Definition
0.7	06.12.2024	All	Knowledge Definition
0.8	12.12.2024	Andrea	Updated Knowledge Definition
0.9	25.01.2025	Andrea, Roan	Entity Definition, Karma
1.0	03.02.2025	Andrea, Roan	Evaluation
1.1	05.02.2025	Roan	SPARQL Queries
1.2	07.02.2025	Vladyslav	Metadata
1.3	08.02.2025	Andrea	Conclusion

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 Microbiologìa. 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 relative abundances of *S. parasanguinis* between Westernized and non-Westernized populations?
    1.3. Which bacterial strains show a significant correlation with negative 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 bacteria 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

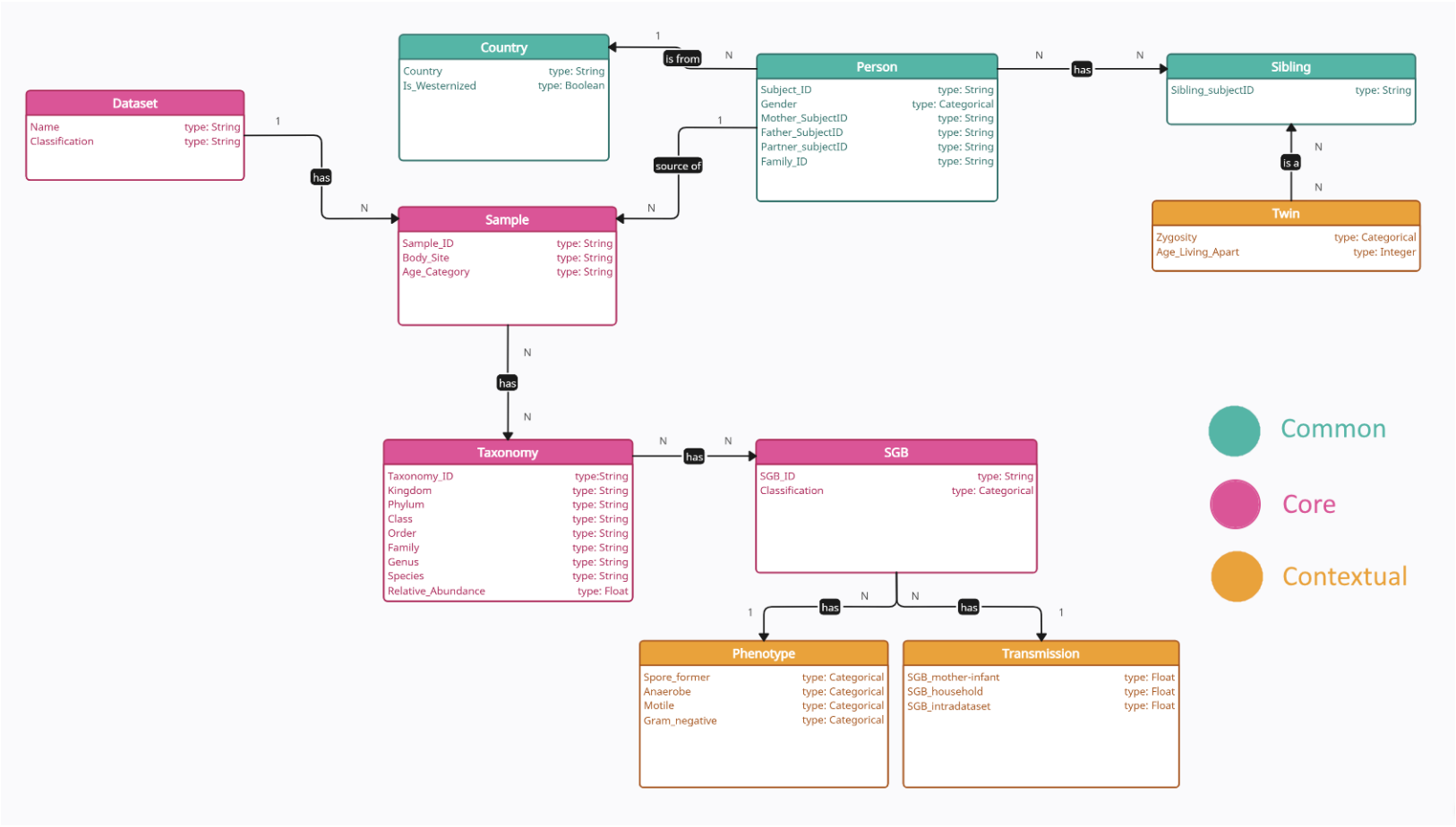


Figure 1: ER Diagram; in green the common ETypes, in red the core ones and in yellow the contextual ones.

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 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 representing 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 for 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
BritoIL_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* (original papers: [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]). 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	YSZC120
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eu...	19.61446	14.37897	0.00131	
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lac...	8.39258	0.00000	0.01835	
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ru...	7.89178	6.47709	0.00100	
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales...	7.37049	2.23896	0.00578	
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales...	5.18316	12.39338	5.65945	
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales...	5.05453	12.89000	4.92746	
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales...	4.93714	3.87848	3.49791	
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ru...	4.36169	1.89050	1.72699	
k__Bacteria|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.
- **intraInterRatio** (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) [20], 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 [21], 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	**Dataset**	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	**Sample**	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	Body Site	extraction point of a sample from a person
UKC-36	**Person**	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	**Country**	the territory occupied by a nation
KGE24-QCB2-2	Westernized	a population that has adopted Western cultural norms, values, and lifestyles.
UKC-52619	**Sibling**	a person’s brother or sister

UKC-53445	**Twin**	either of two offspring born at the same time from the same pregnancy
http://purl.obolibrary.org/obo/GENO_0000133	Zygosity	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	Age Living Apart	the age twins started to live apart
UKC-44477	**Taxonomy**	a classification of organisms into groups based on similarities of structure or origin etc
UKC-42545	Kingdom	the highest taxonomic group into which organisms are grouped; one of five biological categories: Monera or Protoctista or Plantae or Fungi or Animalia
UKC-43156	Phylum	(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
http://purl.obolibrary.org/obo/OHMI_0000468	Relative 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
KGE24-QCB2-4	**SGB**	Species-level Genome Bins - microbial genomes from metagenomic data, representing species-level groupings
UKC-26778	**Phenotype**	what an organism looks like as a consequence of the interaction of its genotype and the environment
https://purl.dsmz.de/schema/SporeFormation	Spore former	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
https://purl.dsmz.de/schema/GramNegativeBacteria	Gram negative	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	**Transmission**	the transfer of microorganisms (bacteria, viruses, fungi, etc.) between individuals, environments, or species, influencing their microbiome composition
KGE24-QCB2-6	mother-infant transmission	the microbial transmission between mother and her infant (i.e. transmissibility)
KGE24-QCB2-7	household transmission	the microbial transmission within people living in the same dwelling (i.e. transmissibility)
KGE24-QCB2-8	intradataset transmission	the microbial transmission within metagenomics samples of same dataset (i.e. transmissibility)
KGE24-QCB2-9	*is from*	a person having origin, association, or ownership related to a source, place, or entity
KGE24-QCB2-10	*source of*	a person being the origin of one or more samples
KGE24-QCB2-11	*has phenotype*	have or possess a phenotype
KGE24-QCB2-12	*has sample*	a dataset having or possessing one or more samples
KGE24-QCB2-13	*has SGB*	have or possess a SGB
KGE24-QCB2-14	*has sibling*	a person having one or more siblings
KGE24-QCB2-15	*has taxonomy*	have or possess a taxonomy
KGE24-QCB2-16	*has transmission*	have or possess a transmission type
KGE24-QCB2-17	*has abundance*	a taxon having a relative abundance in a sample

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. A keen eye must be offered also to highlight the top-level alignment to the schema, which represents a cornerstone for the validation of our ER model, and it is essential for its adaptability and reusability in the long run. From the alignment itself, we can obtain insights regarding the validity of the model and the fact that it conforms to logical, structural and semantic constraints, producing, in the end, an accurate representation of the data that is being modeled. Additionally, during the validation phase, this operation is fundamental to minimize the remaining ambiguities, enhancing the trustworthiness of the model. The schema alignment is also required to facilitate the reuse and the scalability. With the adhering to a well-defined schema, the model becomes modular, making it easier to modify, adapt and adjust depending on the evolving needs and requirements. This also simplifies the incorporation of new features while maintaining the same structure, enhancing the compatibility with other systems that follow the same standard. The schema alignment offers us better communication and documentation of the model’s design, ensuring not only its effectiveness and reliability in this project scenario but also its use in future instance and researches, thanks to its adaptability, making it a valuable asset.

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 [22] tool was used, a free and open-source software for the creation and handling of ontology-based applications. We align [23] 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 this 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 hierarchal 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>*;
- 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 (Fig. 6).

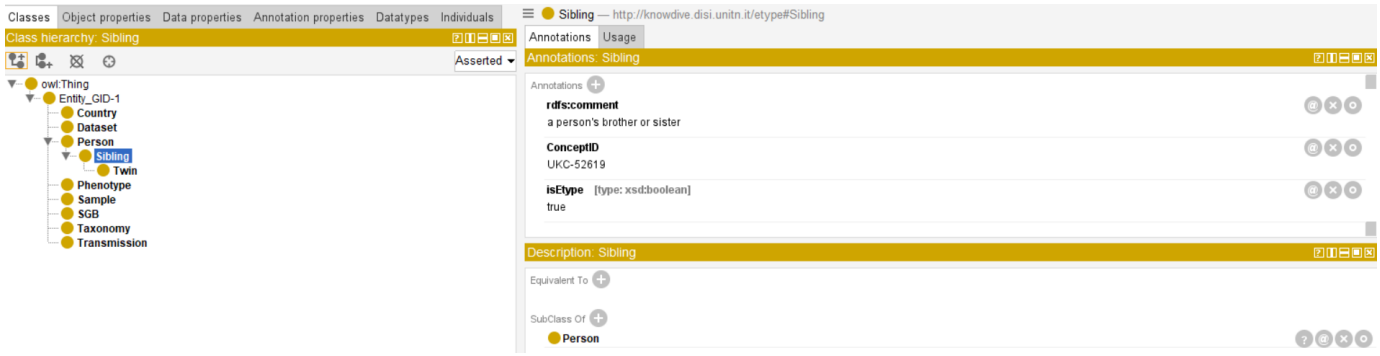


Figure 6: A snapshot of the ETypes modeled in Protégé. Sibling is highlighted; notice the annotations for the definition and the concept ID and the sub-super class relationships.

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) (Fig. 7).

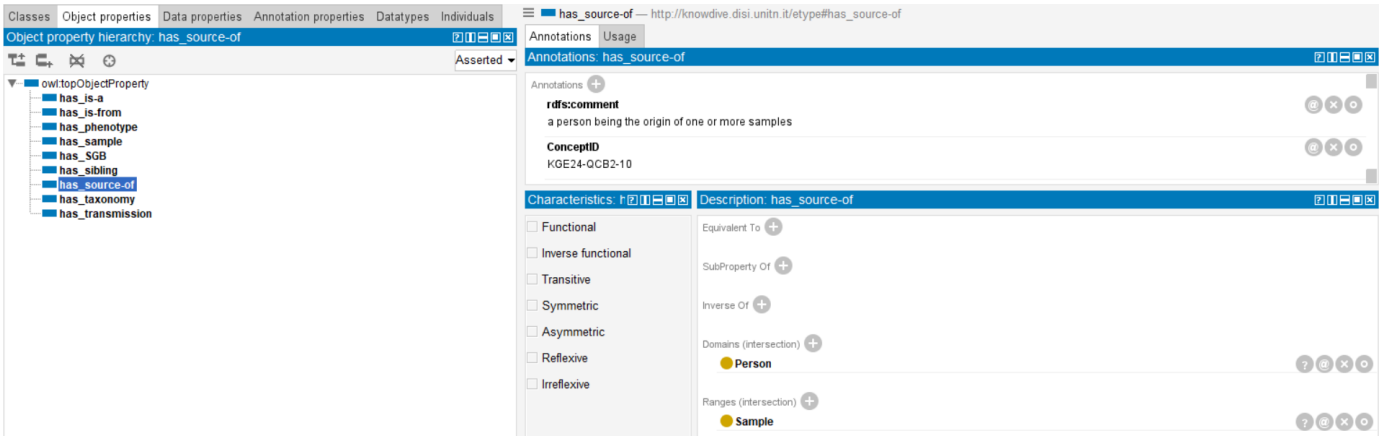


Figure 7: A snapshot of the Object Properties modeled in Protégé. ”Source-of” is highlighted; notice the annotations, domain, and range, in this case, a Person is a source of Sample(s).

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 is belongs to;
- We set the Range(s) of the property as the type this variable can have (for example, Boolean, String, Float...) (Fig. 8).

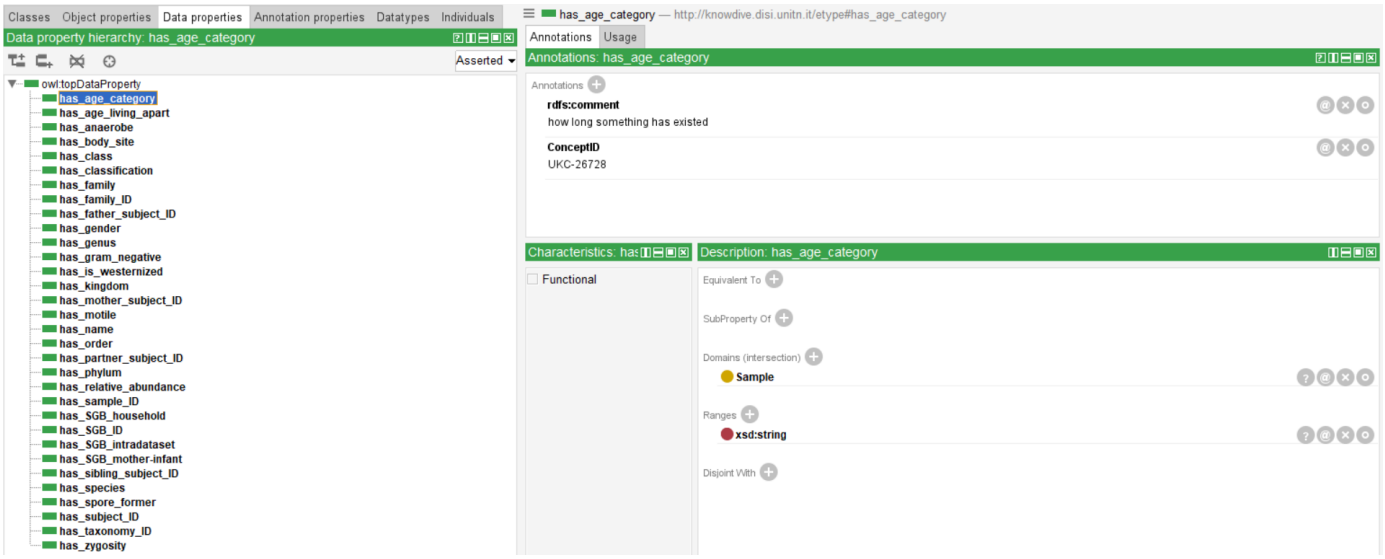


Figure 8: A snapshot of the Data Properties modeled in Protégé. ”Age Category” is highlighted; notice the annotations, the domain (in this case, a Sample has an Age Category), and the range (Age Category is a string).

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 (Fig. 9).

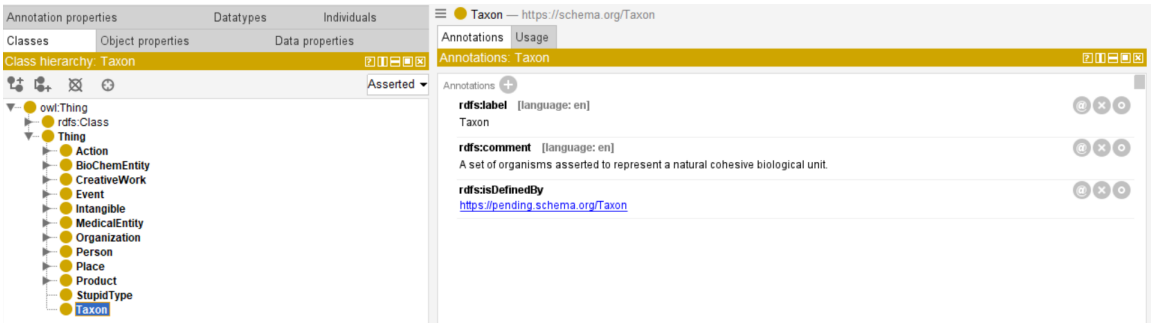


Figure 9: A snapshot of the ETypes present in teleontology (Protégé). ”Taxon” is highlighted; this EType is a part of bioschemas.org, and will be further aligned with EER.

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 (Fig. 10).

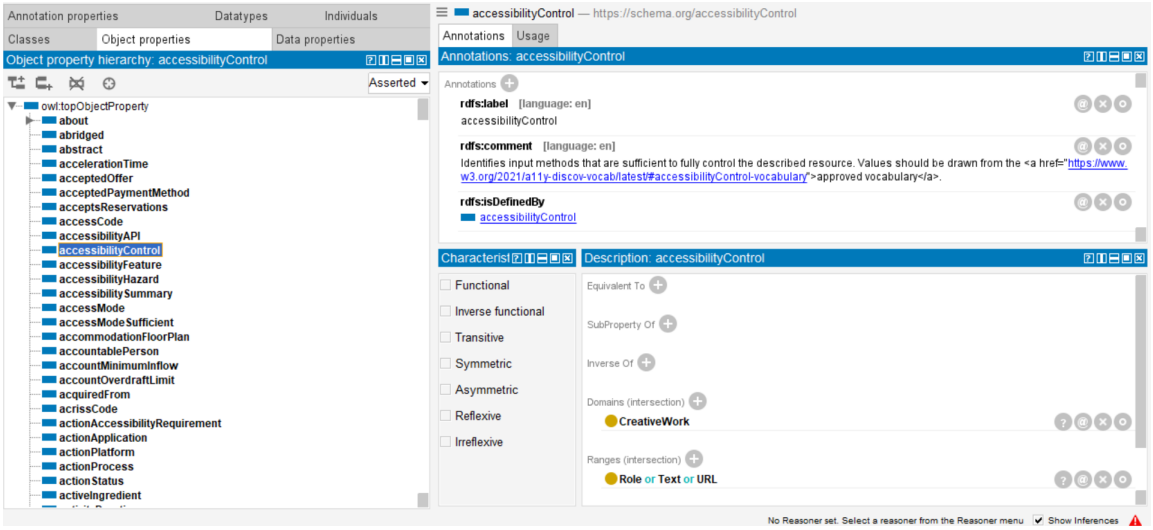


Figure 10: A snapshot of the Object Properties present in teleontology (Protégé). ”accessibilityControl” is highlighted; this property belongs to core schema.org properties.

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 (Fig. 11).

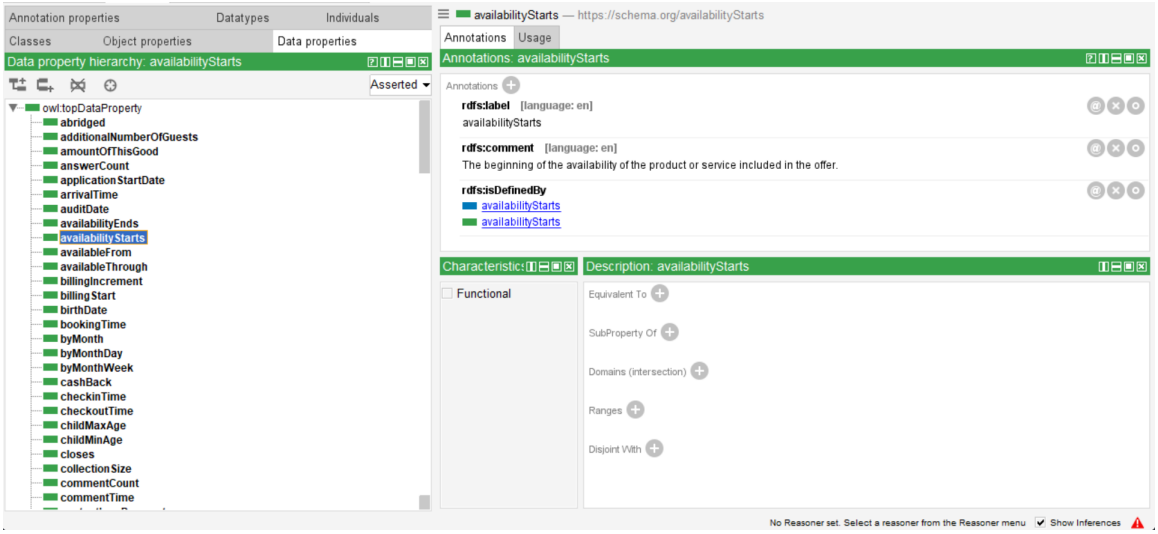


Figure 11: A snapshot of the Data Properties present in teleontology (Protégé). ”availabilityStarts” is highlighted; this property belongs to core schema.org properties.

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 (Fig. 12).

- 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);.

5.3.2 Object properties

There are 3 Object Properties aligned (Fig. 13).

- 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);

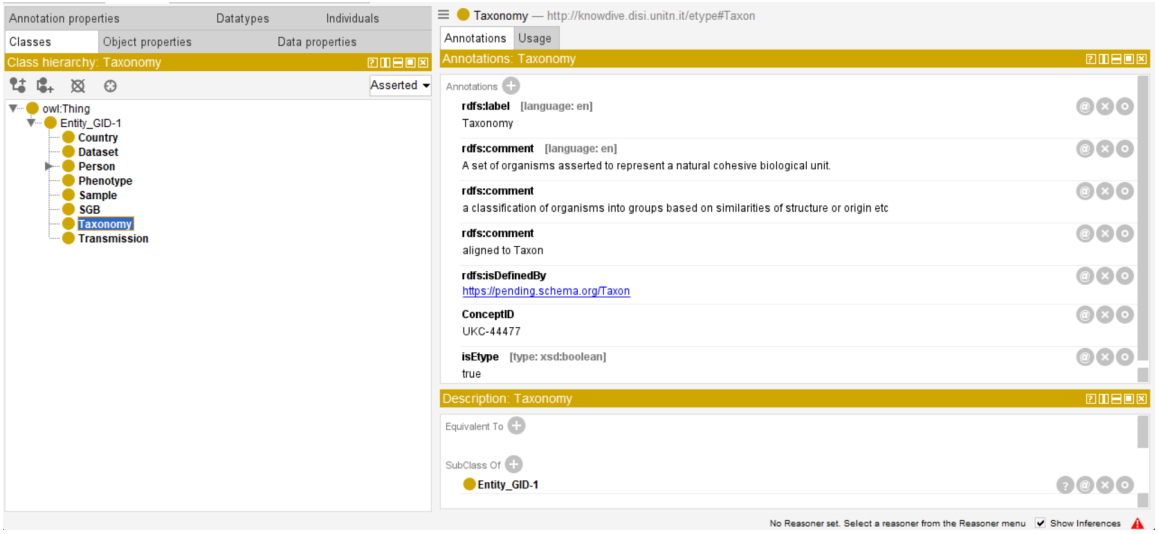


Figure 12: A snapshot of the ETypes present in the teleology (Protégé). ”Taxonomy” is highlighted; this property was aligned with Taxon from schema.org.

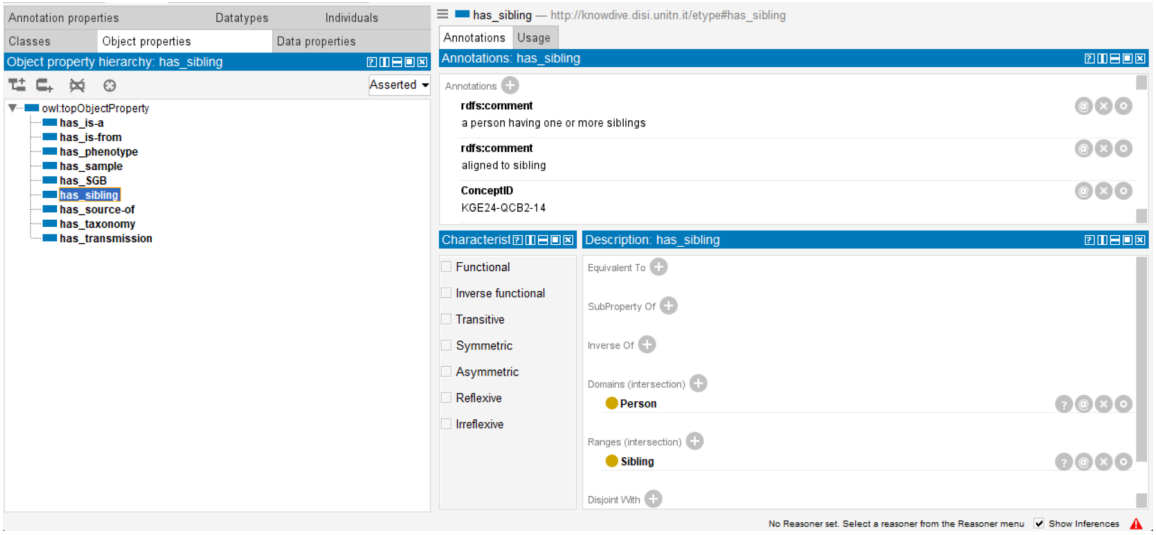


Figure 13: A snapshot of the Object Properties present in the teleology (Protégé). ”has_sibling” is highlighted; this property was aligned with the sibling from schema.org.

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 (Fig. 14) 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 [24] ("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.

6 Entity Definition

We are now approaching one of the final phases of the project: here our objective would be to identify, define and describe the entities we have already encountered, while also giving the reader a precise and robust clarification of the attributes, characteristics and relationships of the entities that are inherent to the problem domain. The final objective is merging the knowledge and data layers into a single illustrative structure: the Knowledge Graph (KG).
From the previous sections we have seen how to handle the heterogeneity derived from the sources and formats according to our design. Here, instead, our input will be the cleaned and aligned data resources and the teleology obtained from the previous sections, the building blocks necessary to ultimate the final KG.
Even so, before we start the actual procedure, we have to take into account a characteristic of the teleology itself: the fact that it does not account for all the data-values heterogeneity it is associated to. Therefore, to fix this and be sure we are producing a final output that is suitable, we have to undergo a 3-steps procedure, dedicated to the entity's matching, identification and mapping. In this phase of the project a new tool will be used: Karma [25], capable of correctly mapping the cleaned and aligned entities derived from the first 2 steps of the procedure.

6.1 Entity Matching

Considering that the real-world entities can be represented in different datasets through different properties, we have to deal with the need to find the right collection of properties among those that can be present across different datasets and the need to set those properties to the correct value across the different representations. The main advantage here is that, taking into

account the previous phases of our work (and the middle-out approach employed), most of the misalignment between the different ETypes and Entities is already solved. Entities represent correctly the same concepts, there are no conflicts between attribute values nor inconsistencies and everything follows the main outline previously described also in the EER diagram.

6.2 Entity Identification

The entity identification step is focused on identifying the entities that cover a central role with respect to our problem domain, while checking whether or not there is an already present identifier or identifying set that can facilitate our work.
After selecting the Entities that we deem relevant to our system's goals, we end up having the Entity Identification table **2**. Notice that some entities previously described do not have an identifier, due to them not being necessary (as entities) for our purpose; specifically:

- Twin: has the same ID of Sibling, as the person is the same, but just acquires the status of "Twin", clear in the "zygosity" property.
- Phenotype: its properties become directly linked to the SGB they belong to (and therefore to its ID). Even though it would be useful to have identifiers for the different phenotypes, the relevant presence of NAs in the phenotype table leads to treating them as properties rather than an identifying set.
- Transmission: is directly linked to the SGB it belongs to: even if there are different types of transmission, their combinations could be a unique identifying set. On the other hand, the SGB ID already covers this "uniqueness" and heterogeneity.

Entity	**Identifier**
Dataset	Name
Sample	sampleID
Person	subjectID
Sibling	sibling_subjectID
Country	country
Taxon	Taxonomy
SGB	SGB

Table 2: Table representing the identifiers of the relevant Entities. Notice that contextual Entities (Twin, Phenotype, Transmission) are not present due to them inheriting or being connected to other IDs. For example, Twin inherits the Sibling identifier, Phenotype and Transmission the SGB one.

6.3 Entity mapping

In this last section we aim at merging the teleology with the relative information values present in the datasets by mapping the entities to their respective data representations and roles inside the system. What we want to do here is mainly map the interactions between entities through the previously described relations, while also inter-linking them to the corresponding columns and properties of the datasets. Moreover, another relevant objective is to represent everything

modeled in Karma into a model graph that can be useful both for better understanding the final output and be implemented in further studies and research.
Karma is an integration tool that allows the user to quickly and easily integrate data coming from different sources, in order to model the information according to the teleology of choice (that was created in Section **5**). In addition to that, it also helps by visualizing entities relationships across different datasets as well and tracks changes or modification done to the graph to further improve the reusability of our work.
The outputs of Karma (Figure **15**, **16**, **18**) are 6 main RDF (Resource Description Framework) Turtle files that we fused together coupled with an RDF for each of the 22 datasets described in Section **3.4**.

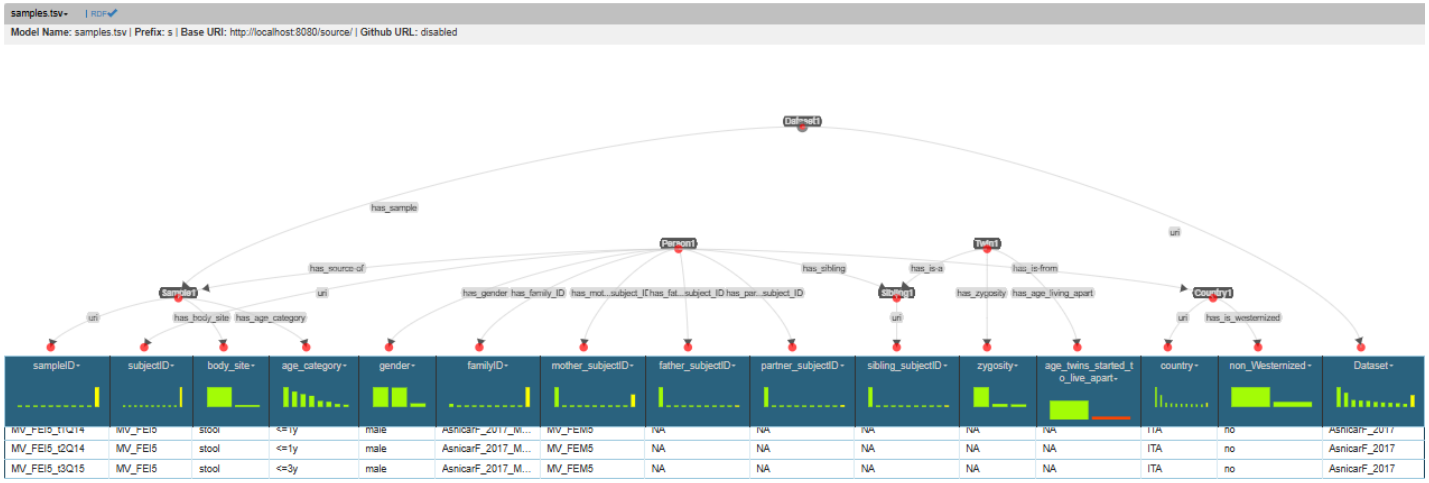


Figure 15: A snapshot of the Karma tool GUI and how we modeled the *samples.tsv* table by including the following entities: Sample, Person, Country, Sibling, Twin, Dataset.

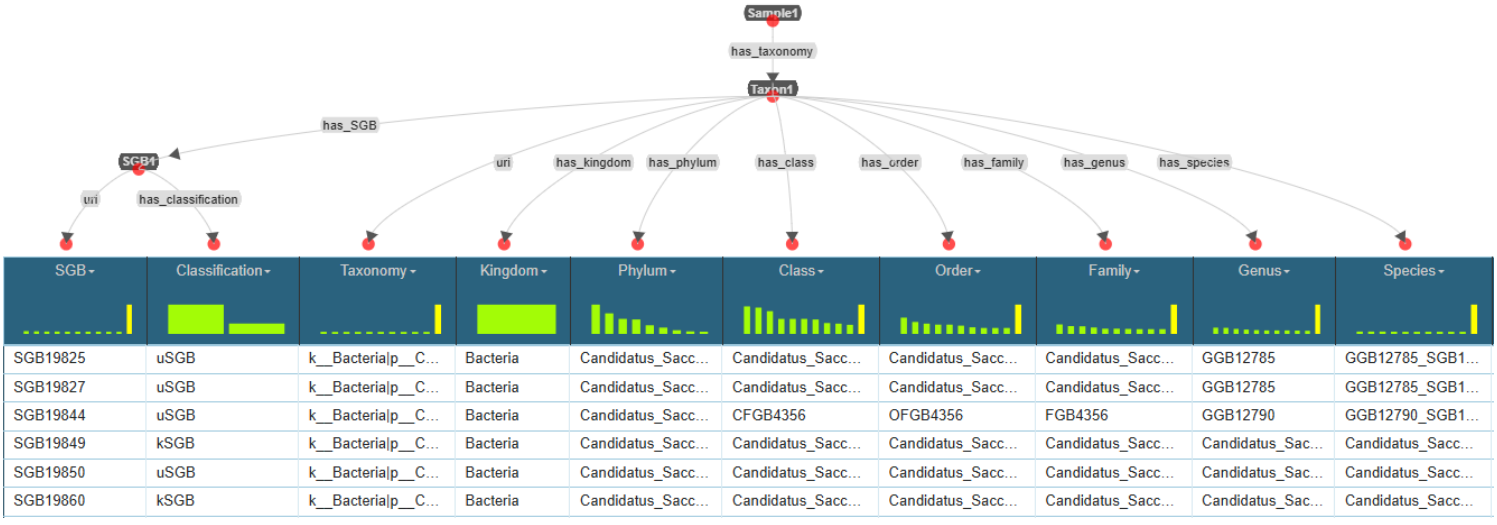


Figure 16: A snapshot of the Karma tool GUI and how we modeled the *taxonomy_SGB_saliva.tsv* table by including the following entities: Taxon, SGB, Sample.

6.4 Dataset fixing

When associating entities in Karma and modeling connections between them, we noticed that the datasets derived from R that are described in section **3.4** need a re-formatting in order to make them alignable with the tool itself. Specifically, the original tables (Figure **5**) had in the first column the Taxon IDs and in the first row the Sample IDs, while in the middle the corresponding relative abundances. This causes problems when modeled in Karma as we cannot properly setup the correct links between the two involved entities and the property values inside, therefore we apply the following modifications in the structures of these datasets:

- Transpose the tables, having columns and rows inverted (rotation):

```
df = pd.read_csv(input_file, sep='\t', index_col=0)  # Read the TSV file
df_transposed = df.transpose()  # Transpose the dataframe
df_transposed.to_csv(output_file, sep='\t')  # Save the transposed dataframe
```

- Transform the tables, having only three columns of combinations of Sample and Taxon IDs with the relative abundance (reshaping). This corresponds to transforming a wide-format *.tsv* into a long-format one (Figure **17**):

```
df = pd.read_csv(input_file, sep='\t')  # Read the TSV file
# Melt the dataframe to long format
df_melted = df.melt(id_vars=["sampleID"], var_name="Taxonomy", value_name="relative_abundance")
df_melted.to_csv(output_file, sep='\t', index=False)  # Save the reshaped dataframe
```

The Python scripts used to perform these transformations are available on GitHub ([24]).

sampleID	Taxonomy	relative_abundance
MV_FEI1_t1Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	59.3501
MV_FEI2_t1Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	0.0
MV_FEI3_t1Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	85.42333
MV_FEI4_t1Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	46.70438
MV_FEI4_t2Q15	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	0.5477
MV_FEI5_t1Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	0.0
MV_FEI5_t2Q14	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	0.02243
MV_FEI5_t3Q15	k__Bacteria-p__Proteobacteria-c__Gammaproteobacteria-o__Enterobacterales-f__Enterobacteriaceae-g__Escherichia-s__Escherichia_coli	0.0

Figure 17: A snapshot of the new datasets deriving from the two previously described transformations. The three columns respectively represent the sampleID, the Taxon ID and the relative abundance of that taxon in that sample.

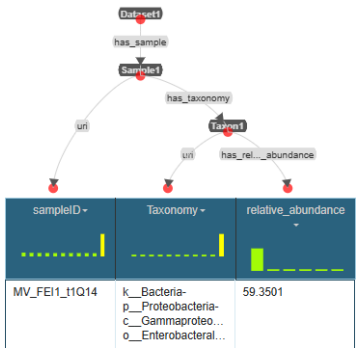


Figure 18: A snapshot of the Karma tool GUI and how we modeled the updated dataset TSVs by including the following entities: Taxon, Sample, Dataset.

6.5 Final considerations

Due to the relevant size of our datasets, especially regarding the 22 tables containing links between all samples and all bacterial taxons found inside them, we decided to limit the amount of data used just for computational efficiency purposes. In principle, all of the 22 tables are modeled in the same way, and what distinguishes them is the amount of data they contain. Our reasoning is that, if the final queries can work with a single dataset, they would work with multiple ones as well. The structure of the final KG would remain almost unchanged; since we have datasets containing sets of samples, including some of them only limits the amount of outputted data, but keeps the final KG working as intended. We will show that the queries work during the evaluation step even on a subset of the datasets as a ”proof” that integrating more data can be easily done by adding all Taxons found inside a Sample together with the relative abundance.

7 Evaluation

As we are approaching the final part of the project, we are still missing some relevant stepping stones that are needed to have a full scheme of our work. In the previous phase we managed to produce the final KG and explain how it is possible to expand it by adding more datasets having the same shape as described in Section **6.4**. Now we are tackling the evaluation step of this project; this can be achieved by considering two main objectives:

- Purpose Satisfaction: is the KG able to satisfy/answer the competency questions proposed at the start of the project?
- How much reusable is the final KG produced in different scenarios?

We will now go ahead in describing how our final knowledge graph looks like and the metrics that are used to evaluate this work.

7.1 Reworking the ER and KG

When testing some SPARQL queries we noticed that the property ”*relative_abundance*”, that was supposed to link a float value to each taxonomy inside a sample, behaved in a different way than what we expected. Specifically, the main problem lies in the way this property is modeled and related to the Sample and Taxon ETypes: since there was not a direct link between Sample and ”*relative_abundance*”, running queries involving this property outputted wrong combinations of the triplets Sample-Taxon-”*relative_abundance*”. To fix this, a small modification to the original EER diagram was made, and consequently all the previous phases aligned accordingly.
We decided to add a new Contextual EType ”**Abundance**” linked to both Sample and Taxon via a ”***has_abundance***” object property. Moreover, this new EType has a single data property: ”*relative_abundance*”, that was moved from ”Taxon” to this new EType (Figure **19**). Since this is more of a ”logical” modification, rather than a conceptual one, the Language definition does not change, as the term ”Abundance” in this context shares the same definition of

”*relative_abundance*”. We added a new definition for the new ”has_abundance” object property.

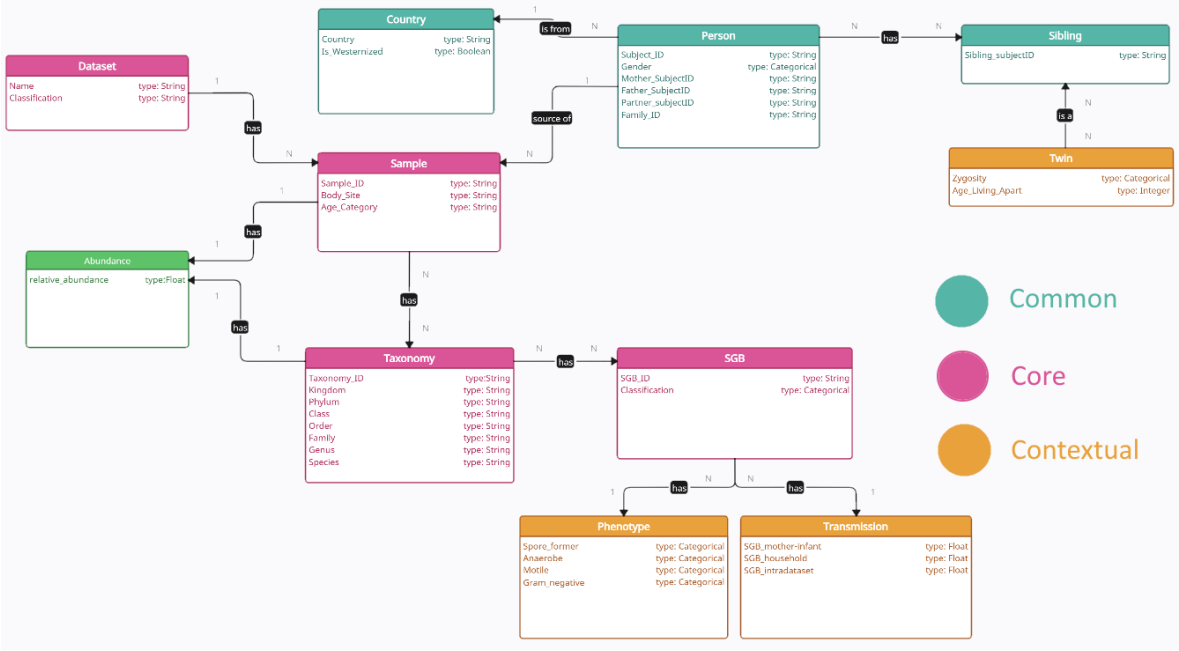


Figure 19: Reworked ER Diagram; notice the new EType ”Abundance” in green. This new EType can be classified as a Contextual one.

On the other hand, this modification led to changes in both the teleology modeled in Protegé and on the Entity Definition done with Karma. Precisely, on Protegé we added this new EType as a sibling of the others, defined it and aligned it the same way ”*relative_abundance*” was, added a new object property ”has_abundance” to link this new EType both with Sample and with Taxon and finally moved the domain of the data property ”*relative_abundance*” from Taxon to Abundance. The new finalized teleology now has 12 ETypes, 10 object properties and 32 data properties. Accordingly, on Karma every selected dataset involving relative abundances was re-modeled accordingly (Figure **20**), generating new RDF files.

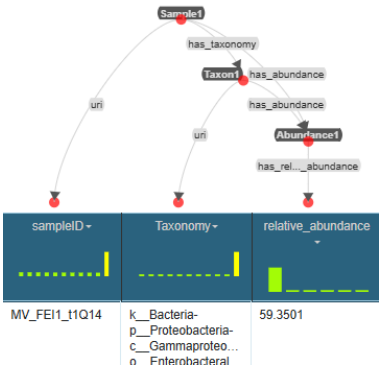


Figure 20: A snapshot of the reworked datasets on Karma. Notice the new EType ”Abundance” and how it is linked to both Sample and Taxon.

7.2 KG information, statistics, visualization

To cover as many Competency Questions as possible, we accurately made a sub-selection of the datasets to specifically include the information needed. Specifically, out of the 22 datasets related only to the relative abundance of the different bacterial species in a sample, we included 6 (*AsnicarF_2017*[4] for the italian samples, *BackedF_2015* and *BritoIL_2016*[6] for the westernized vs non-westernized samples, *CosteaPI_2017_DEU*[8] and *PehrrsonE_2016_SLV*[15] for the German and Slovenian samples, *XieH_2016*[18] for twins).
In total, the KG has 11 ETypes (Table **3**) like the ones planned in the new EER diagram (**19**), 10 object properties linking them and 32 data properties. It would be possible to reduce the amount of ETypes to 9 by injecting the properties of Phenotype and Transmission directly into SGB, but this would lead to a possible loss of contextual information, as adding phenotype or transmission-type identifiers could be possible. Figure **21** shows the class hierarchy visualization made with GraphDB [26] while Figure **23** a visual representation of how the KG is inter-connected and gives the possibility to link Datasets, Samples, Persons with their Country and Siblings, Taxons and SGBs (that include the Phenotypes and Transmission rates).

EType	**Instances**
Dataset	22
Sample	5945
People	3089
Country	16
Sibling	59
Twin	58
Taxon	1686
SGB	874
Phenotype	787
Transmission	589
Abundance	1.727.946
Total	1.741.071

Table 3: Table representing the number of instances for each EType. Total number of statements: 5.341.683

Figure 21: Snapshot of GraphDB class hierarchy viewer; there are 12 classes due to Entity_GID-1 being one.

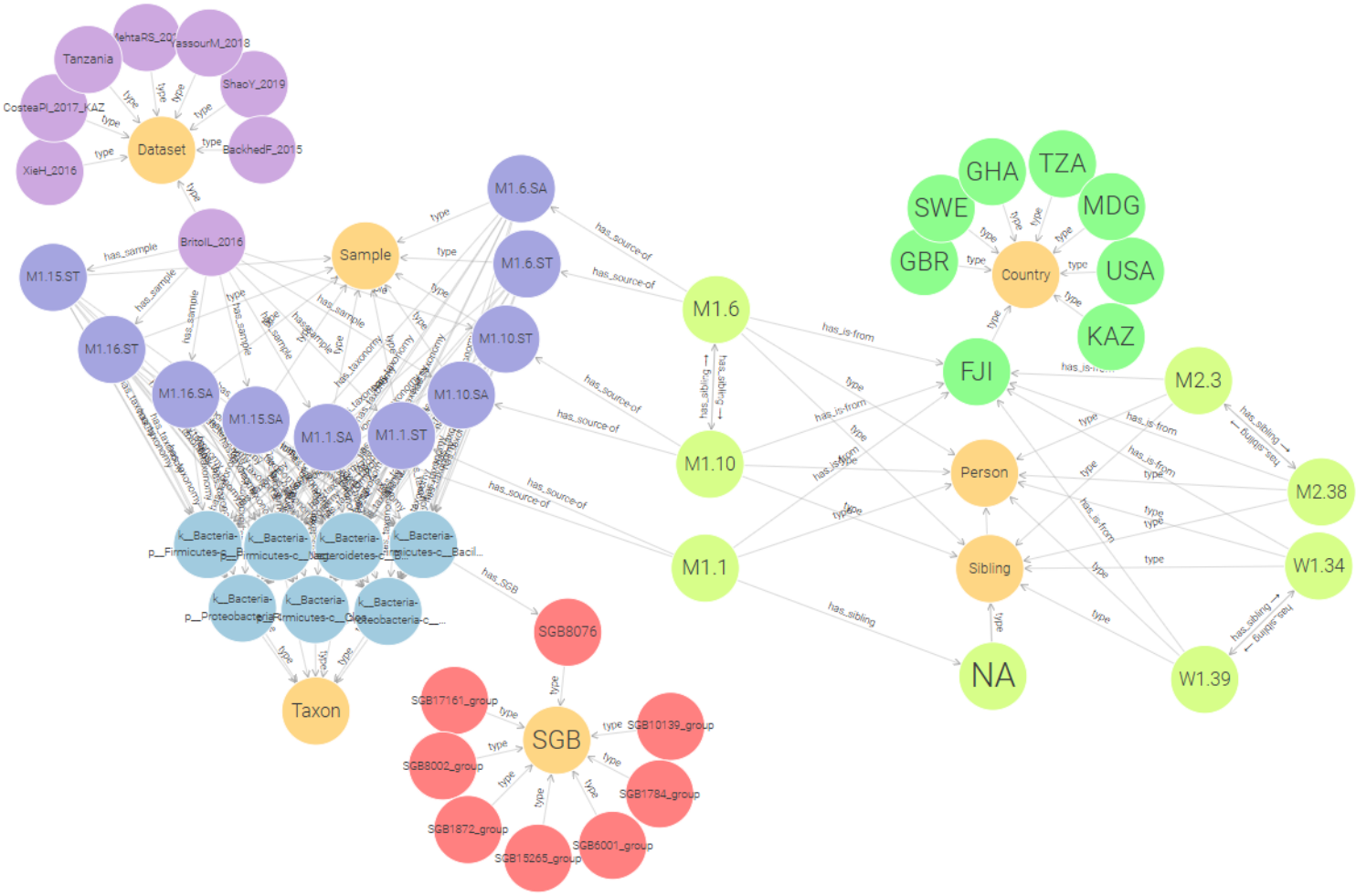


Figure 22: Graph made with GraphDB showing the connections between entities of the KG. Notice the *BritoIL_2016* Dataset (purple) containing Samples (blue) taken from People (light green) with sibling relationships, from Fiji (Country, green). Moreover, Samples contain Taxons (light blue) that correspond to an SGB or a group of them (red).

Starting off with the Purpose Satisfaction, we are going evaluate the KG both at a Knowledge and Data Layer level, with the aim of comparing our KG Teleontology against the CQs proposed at the start, and its level of connectivity.

7.3 Knowledge Layer Evaluation

In the context of the knowledge layer, we consider two objectives: primary, in which we compare the Teleontology with the CQs mentioned in the first parts of the projects with the aim of discerning how much the KG is able to cover the entities and properties that can be extracted from the CQs; secondary, the similar comparison done between the Teleontology and the reference ontologies.

Variable	Meaning
CQ_E	Number of ETypes exctracted from the CQs
T_E	Number of ETypes in the Teleontology
CQ_p	Number of properties extracted from the CQs
T_p	Number of properties in the Teleontology
RO_E	Number of ETypes exctracted from the Reference Ontologies
RO_p	Number of properties exctracted from the Reference Ontologies

7.3.1 EType coverage of the Teleontology

$$\text{Cov}_E(CQ_E) = \frac{|CQ_E \cap T_E|}{CQ_E} = \frac{11 \cap 12}{11} = 1$$

7.3.2 Property coverage of the Teleontology

$$\text{Cov}_p(CQ_p) = \frac{|CQ_p \cap T_p|}{CQ_p} = \frac{34 \cap 32}{34} = 0,941$$

7.3.3 EType coverage with respect to the Reference Ontologies

$$\text{Cov}_E(RO_E) = \frac{|RO_E \cap T_E|}{RO_E} = \frac{914 \cap 12}{914} = 0.013$$

7.3.4 Property coverage with respect to the Reference Ontologies

$$\text{Cov}_p(RO_p) = \frac{|RO_p \cap T_p|}{RO_p} = \frac{183 \cap 34}{183} = 0.186$$

7.4 Data Layer Evaluation

Now that we have explained the mechanism through which we evaluate the purpose satisfaction at the knowledge layer level, it is time to focus on how to do the same on the data one. Here we want to grasp how dense, or in this case connected, the KG is, both at the end of its construction and during the iTelos phases. The connectivity of a KG can be measured in the following ways:

- Entity connectivity: it evaluates how much the entities are connected to one another.
- Property connectivity: it evaluates how much the entities are linked to their properties.

	Dataset	Sample	Person	Country	Sibling	Twin	Taxon	Abundance	SGB	Transmission	Phenotype	Sum *
Dataset	**22**	5911										5911
Sample		**5945**					863973	863973				1727946
Person		5911	**3089**	3050	3050							12050
Country				**16**								0
Sibling					**59**							0
Twin					58	**58**						58
Taxon							**1686**	1727946	955			1728901
Abundance								**1727946**				0
SGB									**874**	589	1574	2163
Transmission										**589**		0
Phenotype											**787**	0
Sum *	0	11822	0	3050	3108	0	863973	2591919	955	589	1574	
												TOT
Entity Connectivity	5911	434942	4016,667	3050	1554	58	864291,3	1295959,5	1039,3333	589	1574	2612985
												TOT
Property Connectivity	11	1981,667	617,8	16	59	29	210,75	1727946	437	196,3333333	196,75	1731701

Figure 23: Table representing Entity (in orange) and Property (in blue) connectivity of each EType.

7.5 SPARQL Queries

Lastly, in order to test our final knowledge graph and verify whether it fully satisfies the initial purpose, SPARQL queries were created. Specifically, to each Competency Question 2.5, a corresponding query was executed via the GraphDB interface after uploading our previously generated RDF files. From a biological point of view all questions make sense, but not all of them can be directly answered with the result of a query. In some cases the result of the query is a strong starting point for more research to follow, therefore some (highlighted in magenta) have been either re-adapted or re-interpreted from the original CQs.

7.5.1 Query 1.1 - return the most commonly transmitted bacterial species within family households

Execution time: 0.1s
Returned results: 26

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?species ?transmissibility
WHERE {
    ?tax rdf:type etype:Taxon;       # gets the Taxons
         etype:has_species ?species; # gets the Species
         etype:has_SGB ?SGB.         # gets SGBs
    ?SGB etype:has_transmission ?tr. # gets transmissions
    ?tr etype:has_SGB_household ?transmissibility # gets household transmission values
    FILTER(?transmissibility != 'NA')             # excludes unknown transmission values
    FILTER(!REGEX(STR(?species), "^GGB")) # excludes unknown species
    FILTER(xsd:float(REPLACE(?transmissibility, ",", ".")) > 0.5) # define "common" as > 0.5
}
ORDER BY DESC(xsd:float(REPLACE(?transmissibility, ",", "."))) # order by most to least common
```

	species	transmissibility
1	"Streptococcus_parasanguinis"	"0,792"
2	"Bacteroides_fragilis"	"0,77777778"
3	"Bifidobacterium_angulatum"	"0,75"
4	"Bifidobacterium_bifidum"	"0,73170732"
5	"Streptococcus_salivarius"	"0,71002132"
6	"Streptococcus_thermophilus"	"0,71002132"
7	"Streptococcus_vestibularis"	"0,71002132"
8	"Rothia_mucilaginosa"	"0,7"
9	"Rothia_sp_HMSC061E04"	"0,7"
10	"Bifidobacterium_adolescentis"	"0,67195767"
11	"Bacteroides_faecis"	"0,66666667"
12	"Bacteroides_cellulosilyticus"	"0,625"
13	"Bacteroides_timonensis"	"0,625"
14	"Megamonas_funiformis"	"0,625"

7.5.2 Query 1.2 - return the mean of the relative abundances of *S. parasanguinis* between Westernized and non-Westernized populations

Execution time: 0.1s
Returned results: 2

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT (IF(?is_westernized = "yes", "no", "yes") AS ?westernized) # flips for clarity
(AVG(xsd:float(?relative_abundance)) AS ?mean_abundance) # compute mean abundances
WHERE {
    ?person rdf:type etype:Person;
            etype:has_is-from ?country;
            etype:has_source-of ?sample.
    ?country etype:has_is_westernized ?is_westernized.
    ?sample etype:has_taxonomy ?taxa.
    ?taxa etype:has_species ?species.
    FILTER(?species = 'Streptococcus_parasanguinis') # filter by species
    ?taxa etype:has_abundance ?abundance.
    ?sample etype:has_abundance ?abundance.
    ?abundance etype:has_relative_abundance ?relative_abundance
}
GROUP BY ?is_westernized # group by westernized status
```

	westernized	mean_abundance
1	"yes"	"0.15536788"^^xsd:float
2	"no"	"0.6270229"^^xsd:float

7.5.3 Query 1.3 - return the bacterial strains correlated with negative gram staining

Execution time: 0.1s
Returned results: 503

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?strain ?gram_staining
WHERE {
    ?SGB rdf:type etype:SGB;
       etype:has_SGB_ID ?strain;
       etype:has_phenotype ?pheno.
    ?pheno etype:has_gram_negative ?gram_staining.
    FILTER(?gram_staining = '1') # selects only negative gram staining
}
```

	strain	gram_staining
1	"SGB13972"	"1"
2	"SGB13976"	"1"
3	"SGB13999"	"1"
4	"SGB14005"	"1"
5	"SGB14007"	"1"
6	"SGB14017"	"1"
7	"SGB14136"	"1"
8	"SGB14193"	"1"
9	"SGB14215"	"1"

7.5.4 Query 1.4 - return the mean relative abundance of *Prevotella intermedia* in cohabitating samples

Execution time: 0.1s
Returned results: 1

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?class (AVG(xsd:float(?relative_abundance)) AS ?mean_abundance)
WHERE {
    ?dataset rdf:type etype:Dataset;
         etype:has_classification ?class;
         etype:has_sample ?s.
    ?s etype:has_taxonomy ?tax.
    ?tax etype:has_species ?species.
    FILTER(?species = "Prevotella_intermedia") # filter by species
    ?s etype:has_abundance ?abundance.
    ?tax etype:has_abundance ?abundance.
    ?abundance etype:has_relative_abundance ?relative_abundance.
}
GROUP BY ?class # group by dataset classification
```

	class	mean_abundance
1	"Mother_offspring and Households"	"0.7376618"^^xsd:float

7.5.5 Query 2.1 - return the transmission rate differences between mother-infant and households

Execution time: 0.1s
Returned results: 61

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?sgb ?mother_infant ?household # also outputs the difference and % difference
((xsd:decimal(REPLACE(?mother_infant, ",", ".")) - xsd:decimal(REPLACE(?household, ",", "."))) AS ?
    difference)
(CONCAT(STR(ROUND(((xsd:decimal(REPLACE(?mother_infant, ",", ".")) - xsd:decimal(REPLACE(?household, ","
    , "."))) / xsd:decimal(REPLACE(?household, ",", ".")) ) * 100)), " %") AS ?percentage_difference)
WHERE {
    ?sgb rdf:type etype:SGB;
       etype:has_transmission ?transmission.
    ?transmission etype:has_SGB_mother-infant ?mother_infant;
            etype:has_SGB_household ?household.
    FILTER (!CONTAINS(LCASE(STR(?sgb)), "geneid"))
    FILTER(?mother_infant != 'NA')
    FILTER(?household != 'NA')
}
```

	sgb	mother_infant	household	difference	percentage_difference
1	loc:SGB10068	"0,44375"	"0,14847591"	"0.29527409"^^xsd:decimal	"199 %"
2	loc:SGB14631_group	"0,806451613"	"0,20325203"	"0.603199583"^^xsd:decimal	"297 %"
3	loc:SGB14809	"0,64"	"0,18181818"	"0.45818182"^^xsd:decimal	"252 %"
4	loc:SGB14824_group	"0,8"	"0,15809524"	"0.64190476"^^xsd:decimal	"406 %"
5	loc:SGB15078	"0,090909091"	"0,27777778"	"-0.186868689"^^xsd:	"-67 %"

7.5.6 Query 2.2 - return the most abundant bacteria in Westernized samples

Execution time: 6m 50s
Returned results: 10

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?species
WHERE {
    ?person rdf:type etype:Person;
            etype:has_is-from ?country;
            etype:has_source-of ?sample.
    ?country etype:has_is_westernized ?is_westernized.
    FILTER(?is_westernized = "no")
    ?sample etype:has_taxonomy ?taxa.
    FILTER (!CONTAINS(LCASE(STR(?taxa)), "geneid"))
    ?taxa etype:has_species ?species.
    FILTER (!CONTAINS(LCASE(STR(?species)), "geneid"))
    ?taxa etype:has_abundance ?abundance.
    ?sample etype:has_abundance ?abundance.
```

```
    ?abundance etype:has_relative_abundance ?relative_abundance
    FILTER(xsd:decimal(?relative_abundance) > 0)
}
ORDER BY DESC(xsd:float(REPLACE(?relative_abundance, ",", ".")))
LIMIT 10
```

	species
1	"Haemophilus_parainfluenzae"
2	"Erysipelatoclostridium_ramosum"
3	"Bifidobacterium_dentium"
4	"Bacteroides_faecis"
5	"Escherichia_coli"
6	"Eubacterium_rectale"
7	"Streptococcus_infantis"
8	"Eubacterium_eligens"
9	"Bacteroides_vulgatus"
10	"Bacteroides_cellulosilyticus"

7.5.7 Query 2.3 - return the most frequent microbial transmission within a family

Execution time: 0.1s
Returned results: 1

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species ?transmission
WHERE {
    ?tax rdf:type etype:Taxon;
    etype:has_SGB ?sgb;
    etype:has_species ?species.
    ?sgb etype:has_transmission ?t.
    ?t etype:has_SGB_household ?transmission.
    FILTER(?transmission != "NA")
}
ORDER BY DESC(xsd:float(REPLACE(?transmission, ",", ".")))
LIMIT 1
```

	species	transmission
1	"Streptococcus_parasanguinis"	"0,792"

7.5.8 Query 2.4 - return the transmission rate differences between different social units

This query can be answered exactly like 2.1. The differences in transmission rates are a starting point to assess the effect of the number of cohabitants (ex. comparing mother-to-son transmission and household transmission) on the presence/absence of certain bacterial species.

7.5.9 Query 3.1 - return the main microbial transmission types

Execution time: 3m 45s
Returned results: 4

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?property
WHERE {
    ?transmission rdf:type etype:Transmission.
    ?transmission ?property ?value.
}
```

	property
1	rdf:type
2	etype:has_SGB_mother-infant
3	etype:has_SGB_intradataset
4	etype:has_SGB_household

7.5.10 Query 3.2 - return the bacteria found more frequently (higher relative abundance) in saliva samples vs stool samples

Execution time: 4.9s
Returned results: 147

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species (AVG(xsd:decimal(?saliva_abundance)) AS ?avg_saliva)
                (AVG(xsd:decimal(?stool_abundance)) AS ?avg_stool)
WHERE {
  {
    SELECT ?species ?body_site (SUM(xsd:decimal(?relative_abundance)) AS ?total_abundance)
    WHERE {
      ?sample rdf:type etype:Sample;
            etype:has_body_site ?body_site;
            etype:has_taxonomy ?taxonomy;
            etype:has_abundance ?abundance.

      ?taxonomy etype:has_abundance ?abundance .
      ?taxonomy etype:has_species ?species.
      ?abundance etype:has_relative_abundance ?relative_abundance .
    }
    GROUP BY ?species ?body_site
  }

  # Assign values based on body-site
  OPTIONAL { FILTER(?body_site = "saliva")  BIND(?total_abundance AS ?saliva_abundance) }
  OPTIONAL { FILTER(?body_site = "stool")   BIND(?total_abundance AS ?stool_abundance) }
}
```

```
GROUP BY ?species
HAVING (?avg_saliva > ?avg_stool)  # Only keep taxonomies with greater abundance in saliva
ORDER BY DESC(?avg_saliva)
```

	species	avg_saliva	avg_stool
1	"Neisseria_sicca"	"503.83254"^^xsd:decimal	"0.01166"^^xsd:decimal
2	"Gemella_haemolysans"	"474.79925"^^xsd:decimal	"3.61315"^^xsd:decimal
3	"Streptococcus_oralis"	"453.30628"^^xsd:decimal	"7.03931"^^xsd:decimal
4	"Rothia_mucilaginosa"	"440.37167"^^xsd:decimal	"25.30692"^^xsd:decimal
5	"Streptococcus_parasanguinis"	"373.66370"^^xsd:decimal	"162.80659"^^xsd:decimal
6	"Porphyromonas_gingivalis"	"371.37438"^^xsd:decimal	"0.06050"^^xsd:decimal

7.5.11 Query 3.3 - return the parental relationships mostly influencing bacterial transmission in families

This query cannot be answered by our KG; this question makes sense from a biological perspective, but in our case we do not have a direct correlation between specific parental relationships and the transmissibility. Transmission values are different when considering either mother-to-child transmission or general household transmission, but we do not have the data to also consider other specific cases needed for this query (like fathers, siblings etc.). What we can do is at least see the relevance of mother-to-child transmission with respect to the others.
Execution time: 0.1s
Returned results: 51

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?SGB ?mts_value ?household_value ?intradataset_value
WHERE {
  ?SGB rdf:type etype:SGB;
       etype:has_transmission ?entity.
  ?entity etype:has_SGB_mother-infant ?mts_value;
          etype:has_SGB_household ?household_value;
          etype:has_SGB_intradataset ?intradataset_value .
  FILTER(?mts_value != 'NA')
  FILTER(?household_value != 'NA')
  FILTER(?intradataset_value != 'NA')
  FILTER (!CONTAINS(LCASE(STR(?SGB)), "geneid"))

  FILTER(xsd:decimal(REPLACE(?mts_value, ",", ".")) > xsd:decimal(REPLACE(?household_value, ",", ".")))
  FILTER(xsd:decimal(REPLACE(?mts_value, ",", ".")) > xsd:decimal(REPLACE(?intradataset_value, ",", ".")
      ))
}
ORDER BY DESC(?mts_value)
```

	SGB	mts_value	household_value	intradataset_value
1	loc:SGB1903	"1"	"0,33333333"	"0,2970297"
2	loc:SGB9283	"0,956521739"	"0,29032258"	"0,05335068"
3	loc:SGB1855	"0,946428571"	"0,375"	"0,05369128"
4	loc:SGB1844_group	"0,933333333"	"0,625"	"0,00938967"
5	loc:SGB17256	"0,929577465"	"0,73170732"	"0,30834753"

7.5.12 Query 3.4 - return the commonly transmitted anaerobic bacterial strains in Italian families

Execution time: 0.1s
Returned results: 11

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?SGB ?anaerobe ?transmissibility_household
WHERE {
  ?person rdf:type etype:Person;
        etype:has_is-from ?country.
    FILTER(?country = loc:ITA)
    ?person etype:has_source-of ?sample.
    ?sample etype:has_taxonomy ?tax.
    ?tax etype:has_SGB ?SGB.
    ?SGB etype:has_phenotype ?pheno.
    ?pheno etype:has_anaerobe ?anaerobe
    FILTER(?anaerobe = "1")
    FILTER (!CONTAINS(LCASE(STR(?SGB)), "geneid"))
    ?SGB etype:has_transmission ?trans.
    ?trans etype:has_SGB_household ?transmissibility_household
    FILTER(?transmissibility_household != 'NA')
    FILTER(xsd:float(REPLACE(?transmissibility_household, ",", ".")) > 0.5)
}
ORDER BY DESC(xsd:float(REPLACE(?transmissibility_household, ",", ".")))
```

	SGB	anaerobe	transmissibility_household
1	loc:SGB1853	"1"	"0,77777778"
2	loc:SGB17256	"1"	"0,73170732"
3	loc:SGB17244	"1"	"0,67195767"
4	loc:SGB1860	"1"	"0,66666667"
5	loc:SGB6962	"1"	"0,625"

7.5.13 Query 4.1 - return the bacterial species with more relative abundance in German samples with respect to Slovenian ones

Execution time: 8.6s
Returned results: 130

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
```

```
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species ((ROUND(100 * AVG(xsd:decimal(?abundance_DEU))) / 100) AS ?avg_abundance_DEU)
       ((ROUND(100 * AVG(xsd:decimal(?abundance_SLV))) / 100) AS ?avg_abundance_SLV)
WHERE {
  ?person rdf:type etype:Person;
        etype:has_is-from ?country;
  FILTER(?country IN (loc:DEU, loc:SLV))
  ?person etype:has_source-of ?sample .
  ?sample etype:has_taxonomy ?taxon;
        etype:has_abundance ?abundance.
  ?taxon etype:has_abundance ?abundance;
         etype:has_species ?species .
  ?abundance etype:has_relative_abundance ?relabundance

  BIND(IF(?country = loc:DEU, ?relabundance, 0) AS ?abundance_DEU)
  BIND(IF(?country = loc:SLV, ?relabundance, 0) AS ?abundance_SLV)
}
GROUP BY ?species
HAVING (AVG(xsd:decimal(?abundance_DEU)) > AVG(xsd:decimal(?abundance_SLV)))
ORDER BY DESC(?avg_abundance_DEU)
```

	species	avg_abundance_DEU	avg_abundance_SLV
1	"Bacteroides_vulgatus"	"4.26"^^xsd:decimal	"0.11"^^xsd:decimal
2	"Bacteroides_uniformis"	"2.68"^^xsd:decimal	"0.17"^^xsd:decimal
3	"Bacteroides_dorei"	"1.88"^^xsd:decimal	"0.06"^^xsd:decimal
4	"Eubacterium_eligens"	"1.68"^^xsd:decimal	"0.17"^^xsd:decimal
5	"Alistipes_putredinis"	"1.43"^^xsd:decimal	"0.05"^^xsd:decimal

7.5.14 Query 4.2 - return the differences in the gut microbiome diversity in Slovenian samples with *S. aureus*

Since our datasets do not have a direct measurement of microbiome diversity (such as the Shannon's index) we cannot answer to this query. Although, we can get all Slovenian samples together with the relative abundance of *S. aureus*. (this query does not output results due to *S. aureus* not being measured in Slovenian samples.)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?person ?country ?relative_abundance ?species
WHERE {
  ?person rdf:type etype:Person;
        etype:has_is-from ?country.
  FILTER(?country = loc:SLV)
  ?person etype:has_source-of ?sample.
  ?sample etype:has_taxonomy ?tax.
  ?tax etype:has_species ?species.
  FILTER(?species = "Staphylococcus_aureus")
  ?tax etype:has_abundance ?ab.
  ?sample etype:has_abundance ?ab.
  ?ab etype:has_relative_abundance ?relative_abundance
}
```

7.5.15 Query 4.3 - return the spore forming bacterial strains common in Italy

Execution time: 1.2s
Returned results: 10

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species ?spore_former
(ROUND(100 * AVG(xsd:decimal(?relative_abundance))) / 100 AS ?mean_abundance)
WHERE {
  ?person rdf:type etype:Person;
        etype:has_is-from ?country.
  FILTER(?country = loc:ITA)
  ?person etype:has_source-of ?sample.
  ?sample etype:has_taxonomy ?tax.
  ?tax etype:has_species ?species.
  ?tax etype:has_abundance ?ab.
  ?sample etype:has_abundance ?ab.
  ?ab etype:has_relative_abundance ?relative_abundance.
  FILTER(xsd:decimal(?relative_abundance) != 0)
  ?tax etype:has_SGB ?SGB.
  ?SGB etype:has_phenotype ?phenotype.
  ?phenotype etype:has_spore_former ?spore_former.
  FILTER(?spore_former = "1")
}
GROUP BY ?species ?spore_former
ORDER BY DESC(xsd:decimal(?mean_abundance))
```

	species	spore_former	mean_abundance
1	"Dorea_longicatena"	"1"	"0.77"^^xsd:decimal
2	"Ruthenibacterium_lactatiformans"	"1"	"0.24"^^xsd:decimal
3	"Flavonifractor_plautii"	"1"	"0.21"^^xsd:decimal
4	"Erysipelatoclostridium_ramosum"	"1"	"0.21"^^xsd:decimal
5	"Ruminococcus_callidus"	"1"	"0.13"^^xsd:decimal
6	"Coprobacillus_cateniformis"	"1"	"0.06"^^xsd:decimal
7	"Clostridium_perfringens"	"1"	"0.05"^^xsd:decimal
8	"Clostridium_innocuum"	"1"	"0.04"^^xsd:decimal
9	"Intestinibacter_bartlettii"	"1"	"0.02"^^xsd:decimal
10	"Turicibacter_sanguinis"	"1"	"0.02"^^xsd:decimal

7.5.16 Query 5.1 - return the bacterial strains with highest transmission rates in the oral mouth between westernized and non-westernized populations

Execution time: 1.2s
Returned results: 10

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
SELECT DISTINCT ?SGB ?transmission_mother_infant
?transmission_household ?transmission_intradataset
WHERE {
  ?sample rdf:type etype:Sample;
          etype:has_body_site ?body.
  FILTER(?body = "saliva")
  ?sample etype:has_taxonomy ?tax.
  ?tax etype:has_SGB ?SGB.
  ?SGB etype:has_transmission ?transmission.
  ?transmission etype:has_SGB_mother-infant ?transmission_mother_infant;
    etype:has_SGB_household ?transmission_household;
    etype:has_SGB_intradataset ?transmission_intradataset.
    FILTER(?transmission_mother_infant != 'NA')
    FILTER(?transmission_household != 'NA')
    FILTER(?transmission_intradataset != 'NA')
}
ORDER BY DESC(xsd:float(REPLACE(?transmission_intradataset, ",", ".")))
```

	SGB	transmission_mother_infant	transmission_household	transmission_intradataset
1	loc:SGB8076	"0,214285714"	"0,792"	"0,77303895"
2	loc:SGB8002_group	"0,062992126"	"0,71002132"	"0,64370465"
3	loc:SGB9712_group	"0,117647059"	"0,4966443"	"0,41432769"
4	loc:SGB6939	"0,1875"	"0,46666667"	"0,38110236"

7.5.17 Query 5.2 - return the different bacterial transmission rates between infants and older children

While it is possible to select samples by their age, we do not have data regarding different transmission rates in different age categories. What we can do is returning the most abundant bacterial species in infants vs older children.
Execution time: 43s
Returned results: 230

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species
((ROUND(100 * AVG(xsd:decimal(?abundance1))) / 100) AS ?meanAbundance1day)
((ROUND(100 * AVG(xsd:decimal(?abundance2))) / 100) AS ?meanAbundance1week)
((ROUND(100 * AVG(xsd:decimal(?abundance3))) / 100) AS ?meanAbundance1year)
((ROUND(100 * AVG(xsd:decimal(?abundance4))) / 100)AS ?meanAbundance3years)
WHERE {
  ?sample rdf:type etype:Sample;
          etype:has_age_category ?age_category .
  FILTER(?age_category IN ("<=1d", "<=1w", "<=1y", "<=3y"))
  ?sample etype:has_taxonomy ?taxon .
  ?taxon etype:has_species ?species .
  ?sample etype:has_abundance ?abundance .
  ?taxon etype:has_abundance ?abundance .
  ?abundance etype:has_relative_abundance ?relative_abundance .
  FILTER(?relative_abundance != 'NA')

  BIND(IF(?age_category = "<=1d", ?relative_abundance, 0) AS ?abundance1)
  BIND(IF(?age_category = "<=1w", ?relative_abundance, 0) AS ?abundance2)
```

```
  BIND(IF(?age_category = "<=1y", ?relative_abundance, 0) AS ?abundance3)
  BIND(IF(?age_category = "<=3y", ?relative_abundance, 0) AS ?abundance4)
}
GROUP BY ?species
ORDER BY DESC(?meanAbundance1day)
```

	species	meanAbundance1day	meanAbundance1we...	meanAbundance1year	meanAbundance3ye...
1	"Escherichia_coli"	"3.08"^^xsd:decimal	"3.02"^^xsd:decimal	"3.44"^^xsd:decimal	"0.26"^^xsd:decimal
2	"Bacteroides_fragilis"	"1.06"^^xsd:decimal	"1.49"^^xsd:decimal	"3.58"^^xsd:decimal	"0.92"^^xsd:decimal
3	"Prevotella_copri"	"0.18"^^xsd:decimal	"0"^^xsd:decimal	"0.77"^^xsd:decimal	"0.76"^^xsd:decimal
4	"Bacteroides_vulgatus"	"0.15"^^xsd:decimal	"1.9"^^xsd:decimal	"2.21"^^xsd:decimal	"1.24"^^xsd:decimal

7.5.18 Query 5.3 - Return the transmission rates of *Treponema Denticola*

This query does not return results due to the absence of data regarding the transmission rates of *Treponema Denticola*.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX loc: <http://localhost:8080/source/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?species ?SGB
?transmission_mother_infant ?transmission_household ?transmission_intradataset
WHERE {
  ?tax rdf:type etype:Taxon;
       etype:has_species ?species.
  FILTER(?species = "Treponema_denticola")
  ?tax etype:has_SGB ?SGB.
  ?SGB etype:has_transmission ?transmission.
  ?transmission etype:has_SGB_mother-infant ?transmission_mother_infant;
            etype:has_SGB_household ?transmission_household;
            etype:has_SGB_intradataset ?transmission_intradataset.
}
```

7.5.19 Query 5.4 - return the dental microbial diversity with varying number of people in the household

This CQ is very general, and a single query would not return enough results. Moreover, our data does not possess a measure of microbial diversity. On the other hand, with a specific bacterial species, the solution is very similar to Query 5.3.

8 Metadata Definition

The project's metadata are available in the GitHub repository [24] in the format of Microsoft Excel Worksheet (.xlsx). It contains definitions of all resources employed and produced throughout the project lifecycle. Metadata is fully compatible for distribution via DataScientia. Below is the structured description of metadata for the types of resources produced:

8.1 People Metadata Description

This metadata encompasses information about individuals involved in the project, including:

- **Identifiers:** Names, email addresses, and affiliations.
- **Attributes:** Nationality, gender, and personal webpages (if available).

8.2 Project Metadata Description

This metadata describes the project details, such as:

- **Core Information:** Project title, description, start and end dates, keywords, and project type.
- **Funding and Deliverables:** Funding agency (if applicable), project input data, and outputs like reports and diagrams.
- **Coordination:** Name of the project coordinator and related observations.

8.3 Dataset Metadata Description

This metadata outlines datasets generated or utilized within the project, including:

- **Identifiers:** Dataset license, version, and name.
- **Attributes:** Keywords, creators, publishers, and owners.
- **Details:** Dataset language, size, file format, publication date, and domain.
- **Description:** Summary of dataset content.

This metadata structure ensures seamless sharing and distribution of the project's outputs across relevant data catalogs.

9 Open Issues and Conclusions

Following the realization of all the steps that we presented throughout the whole project, we are now able to give our conclusions and feedback regarding the whole process. As a starting point, the initial schedule that we assigned ourselves was stretched due to complications arising in some occasions, prevalently in the final sections of the project, such as the Entity definition and Evaluation phases. Here issues came to light when working with Karma and SPARQL, due to conflicts in the establishment of relationships and factors (such as the relative abundance) between the elements that we worked on. In order to come up with a working solution we had to dedicate more time and modify our reasoning accordingly. Moreover, our effort was dedicated to understanding the data we were working on and what to keep or remove to satisfy the objectives that we described during the Purpose definition step 2. We were also able to answer or re-interpret the CQs by applying queries on the final KG structure and retrieving what

the model outputted. A better approach would have been directly making more precise CQs, but, since we did not have complete familiarity with the data, we did not know what questions we would have been able to answer.
In the end, we are satisfied with the results we obtained, ranging from the production of the final KG structure and purpose to the queries answered using that model. We deem our work adequate and effective in terms of performance, consistency and re-usability. Even so, there are still open issues that remain, for example:

- Having big datasets and files: we avoided using the complete starting datasets (due to processing time and file sizes) but relied only on a fraction of them. Implementing all datasets would mean having a greater KG from which we would be able to extract more information, covering a bigger domain of knowledge than we already are.
- The Twin's Role: throughout the project we used datasets that were specific for twins; it would be interesting to see, at a biological level, what would happen while considering also datasets only with brothers/sisters not classified as twins. In our case, one of the ETypes, either "Sibling" or "Twin", loses its purpose due to almost all siblings (58/59) being twins.
- Further tuning of the KG: it would be optimal to streamline even more the KG that we produced by implementing the data that we obtain from the transmissibility table into the SGBs properties, without the need of creating a new EType for it. This would not lead to a loss of information, as transmissibility has been measured as a mean quantity for each SGB.

Considering the reusability of this project, we are confident to say that further improvements in the near future can be done, with the scope of augmenting to a greater extent the KG that we constructed, promoting a greater coverage of the this particular field of knowledge which we had the chance to work with. In terms of biological relevance, it is useful to have the possibility of getting relevant information to answer a question with just a query: this can lead to a strong starting ground to develop more articulate questions and follow up researches.

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] Francesco Asnicar, Serena Manara, Matteo Zolfo, Dao T. Truong, Marcel Scholz, Federico Armanini, Paolo Ferretti, Viola Gorfer, Andrea Pedrotti, Alessandro Tett, and Nicola Segata. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems*, 2(1):e00164–16, 2017.

[5] Francesco Asnicar, Sarah E. Berry, Ana M. Valdes, Long H. Nguyen, and Giulia Piccinno et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nature Medicine*, 27(2):321–332, 2021.

[6] Ilana L. Brito, Sami Yilmaz, Kerwyn Huang, and Liang Xu et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, 2016.

[7] Derrick M. Chu and Samuel C. Antony Maria Seferovic Kjersti S. Aagaard Monica Ma, Gregory L. Prince. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nature Medicine*, 23(3):314–326, 2017.

[8] Paul I. Costea et al. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35(11):1069–1076, 2017.

[9] Pamela Ferretti et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe*, 24(1):133–145.e5, 2018.

[10] A. Almeida et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39:105–114, 2021.

[11] Sandrine Louis et al. Characterization of the gut microbial community of obese patients following a weight-loss intervention using whole metagenome shotgun sequencing. *PLOS ONE*, 11(2):e0149564, 2016.

[12] Raaj S. Mehta et al. Stability of the human faecal microbiome in a cohort of adult men. *Nature Microbiology*, 3(3):347–355, 2018.

[13] H. Bjørn Nielsen et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, 2014.

[14] Edoardo Pasolli et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, 2019.

[15] Erica C. Pehrsson et al. Interconnected microbiomes and resistomes in low-income human habitats. *Nature*, 533(7602):212–216, 2016.

[16] Adrian Tett et al. The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe*, 26(5):666–679.e7, 2019.

[17] Linda Wampach et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nature Communications*, 9(1):5091, 2018.

[18] Hailiang Xie et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Systems*, 3(6):572–584.e3, 2016.

[19] Moran Yassour et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe*, 24(1):146–154.e4, 2018.

[20] Giunchiglia F., Bella G., and Nair N.C. et al. Representing interlingual meaning in lexical databases. *Artificial Intelligence Review*, 56:11053–11069, 2023.

[21] 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.

[22] Musen MA and Protégé team. The protégé project: A look back and a look forward. *AI matters*, 1(4):4-12, 2015.

[23] Fausto Giunchiglia and Daqian Shi. Property-based entity type graph matching. *arXiv*, 2021.

[24] Github repository of the project: https://github.com/boreico/kge_qcb_project/tree/main, 2024.

[25] Craig A. Knoblock and Pedro Szekely. Exploiting semantics for big data integration. *AI Magazine*, 36(1):25–38, 2015.

[26] Güting and Ralf Hartmut. Graphdb: Modeling and querying graphs in databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 297–308, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.