Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

The human microbiome and person-to-person interactions

Document Data:	Reference Persons:
November 2, 2024	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

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 the creation of 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 exhibits 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 the different scenarios and sometimes share some similarities when involved in different fields, but are heterogeneous in same-scenario situations.

1. **Francesca, Microbiologist, PhD student**

   1.1. I'm studying the gut microbiome. Can I get a list of the most commonly transmitted bacterial species within family households?

   1.2. What are the transmission rates of *S. parasanguinis* between Westernized and non-Westernized populations?

1.3. Which bacterial strains show a significant correlation with Gram staining?

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

2. **Herald, Social Sciences Professor**

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

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

2.3. Which is the most frequent microbial transmission in different social groups?

2.4. How does social network size 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 transmitted via oral contact compared to vaginal transmission?

3.3. Which environmental factors mostly influence bacterial transmission in families?

3.4. Can I find examples of pathogenic 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. Are there any bacterial strains linked to Maribor's geographic locations that could explain fewer vomiting symptoms?

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

4.4. If the patient comes to Italy for which pathogenic 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	Geography	Dataset, Sample, SGB, Taxonomy	Transmission
1.3		Sample, SGB, Taxonomy	Phenotype
1.4	Geography	Dataset, Sample, SGB, Taxonomy	Transmission
2.1	Individual, Family, Sibling	SGB, Taxonomy	Transmission
2.2	Geography	Dataset	Transmission
2.3			Transmission
2.4			Transmission
3.1			Transmission
3.2		Sample, SGB, Taxonomy	Transmission
3.3			Transmission
3.4	Geography	Dataset, Sampe, SGB, Taxonomy	Phenotype
4.1	Geography, Sibling	Dataset	Twin
4.2	Geography	Dataset, Sample, SGB, Taxonomy	
4.3	Geography	Dataset, Sample, SGB, Taxonomy	
4.4	Geography	SGB, Taxonomy	Phenotype
5.1	Geography	Dataset, Sample, SGB, Taxonomy	Transmission
5.2		Dataset, Sample, SGB, Individual	Transmission
5.3		Dataset, Sample, SGB, Taxonomy	Transmission
5.4		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.

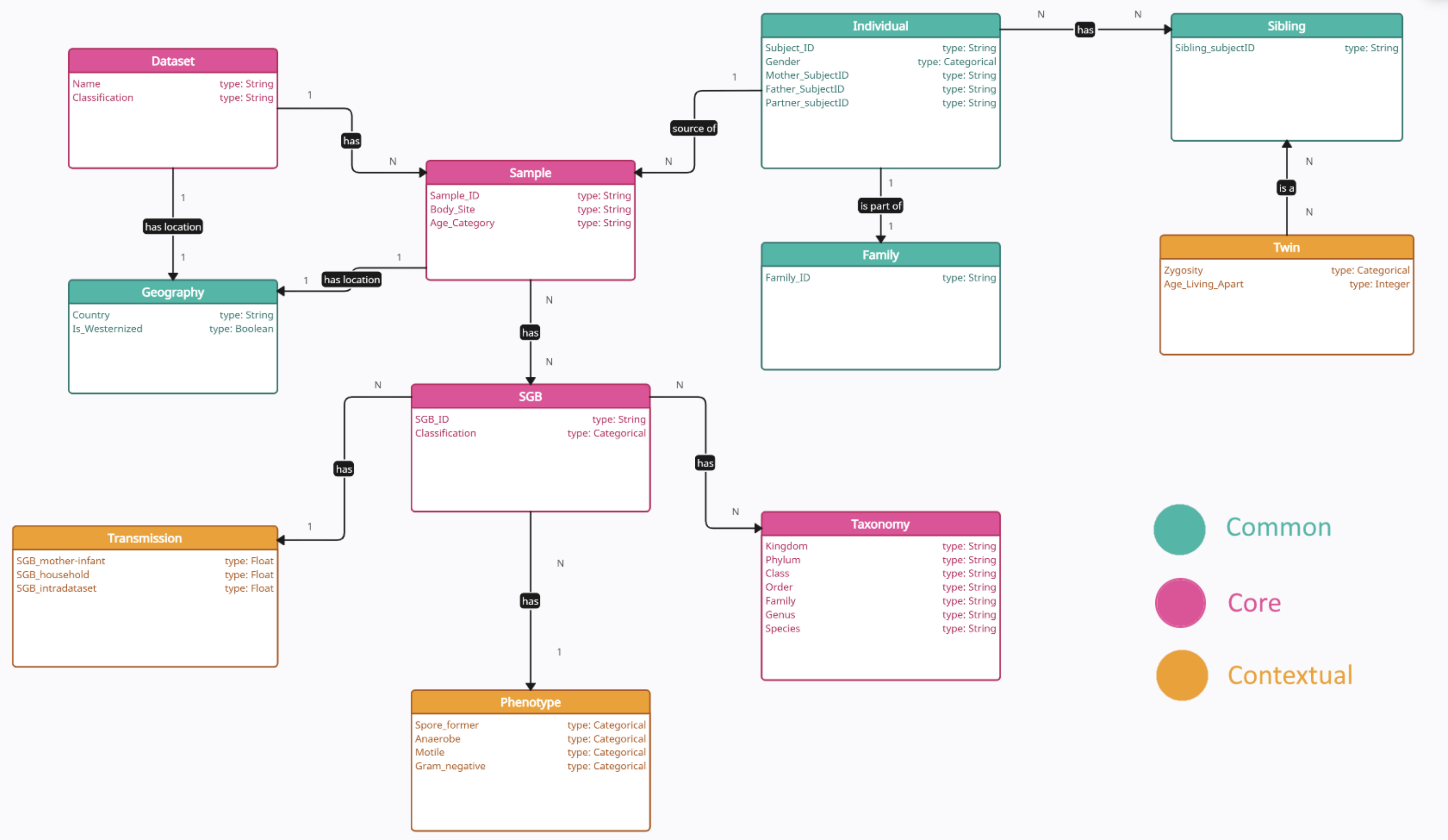

Figure 1: ER Diagram; in green the common ETypes, in red the core ones and in yellow the contextual ones.

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 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**:
  - Geography: 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.
  - Individual: a general EType to represent someone, together with its gender and parental relationships. This is needed as the same individual can produce multiple Samples; furthermore, it is useful in order to link individuals to one another, specifying possible transmission rate differences when comparing parents and children.
  - Family: a general EType to represent a group of individuals. It is an alternative way to make comparisons between families instead of individuals.

- 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 to the transmission types and rates in different contexts for each SGB. Very relevant to our purpose in distinguishing 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.

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.