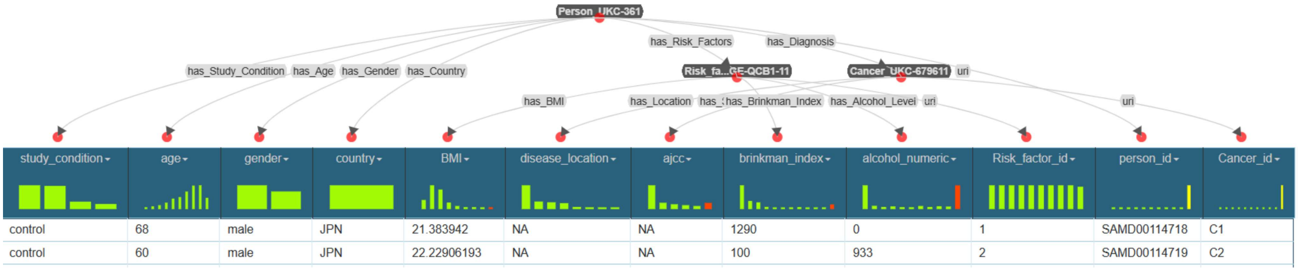
5 Entity Definition

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

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

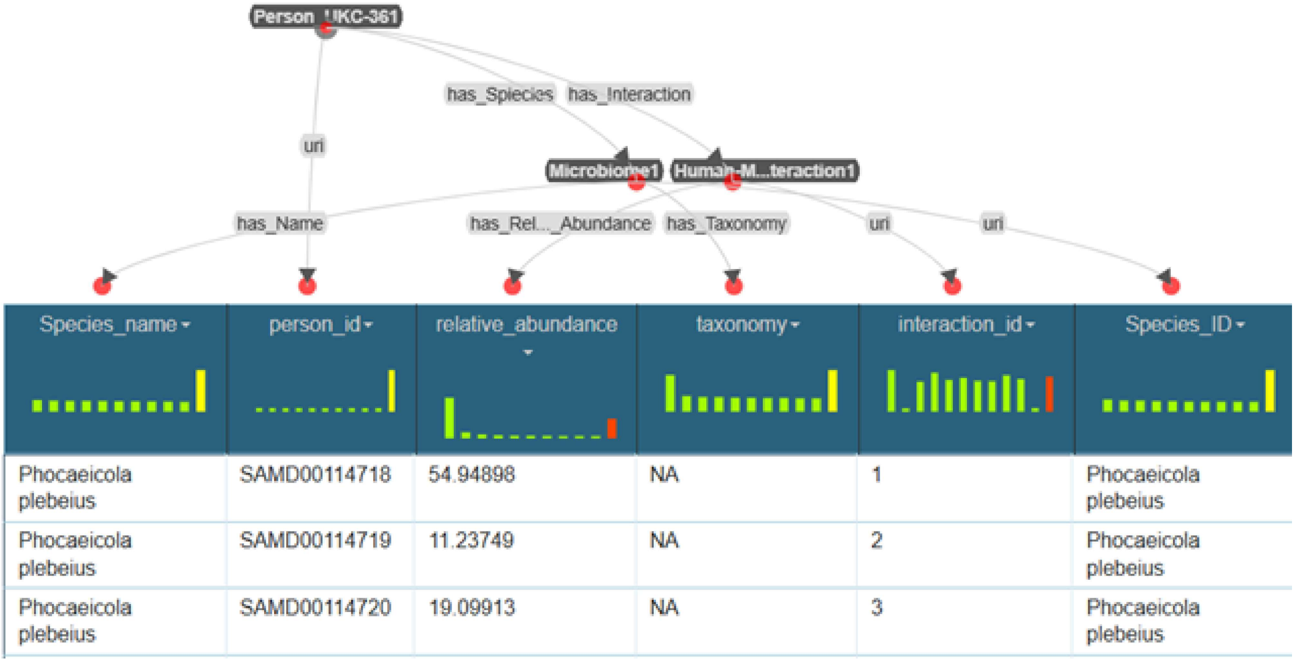
Entity Definition sub-activities:

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



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

Figure 10: Karma visualization



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

Figure 11: Karma visualization