2 Information Gathering

After completing the first phase of the project, which involved defining the objectives and building the ER model, we now move on to the information-gathering phase. In this second phase, we identify and accurately describe the sources and structure of the available data while performing a thorough cleaning and filtering process. The goal is to produce standardized datasets ready for the subsequent stages of analysis.

Sources Identification

For this project, the data source used is the R package `CuratedMetagenomicData` (CMD). This package provides curated and standardized human microbiome data, useful for innovative analyses. It also includes a range of tools that provide information on gene families, marker abundance, marker presence, pathway abundance, pathway coverage, and relative abundance for samples collected from various body sites. The taxonomic abundances of bacteria, fungi, and archaea for each sample were calculated using MetaPhlAn3, while the metabolic functional potential was determined with HUMAnN3. The manually curated sample metadata and standardized metagenomic data are made available as `(Tree)SummarizedExperiment` objects. Additionally, specific metadata for each sample is collected based on the study's objectives. In total, the package includes over 26 studies comprising 5716 samples and 34 diseases. (Table 2) shows the different types of studies available in the package, along with their respective properties.

Dataset Name	Body Site	Disease	# Total Samples	# Case Samples	Average Reads per Sample (std) (M)	Size (Tb)	# Reads (G)	Reference
AsnicarF_2017	Stool, milk	None	26	-	21.4 (19.8)	0.2	0.5	7
BritoIL_2016	Stool, oral	Other condition	312	-	67.4 (51.8)	5.6	21.0	8
Castro-NallarE_2015	Oral	Schizophrenia	32	16	61.0 (25.2)	0.5	2.0	9
ChngKR_2016	Skin	Atopic dermatitis	78	38	15.8 (7.5)	0.3	1.2	10
FengQ_2015	Stool	Colorectal cancer	154	93	53.8 (8.5)	2.3	8.3	11
Heitz-BuschartA_2016	Stool	Type 1 diabetes	53	27	44.5 (0.9)	0.5	2.4	12
HMP_2012	Several	None	749	-	51.5 (44.8)	9.4	38.6	4
KarlssonFH_2013	Stool	Type 2 diabetes	145	53	31.0 (17.6)	1.4	4.5	13
LeChatelierE_2013	Stool	Obesity	292	169	69.0 (23.2)	4.0	20.1	14
LiuW_2016	Stool	Other condition	110	-	58.3 (26.8)	1.8	6.4	15
LomanNJ_2013	Stool	Shiga-toxigenic *E. coli*	43	43	9.2 (12.1)	0.1	0.4	16
NielsenHB_2014	Stool	Inflammatory bowel diseases	396	148	53.9 (20.2)	3.5	21.4	17
Obregon-TitoAJ_2015	Stool	Other condition	58	-	47.1 (20.9)	0.6	2.7	18
OhJ_2014	Skin	None	291	-	24.7 (38.1)	2.2	7.2	19

Table 2: Datasets available in the package

Datasets Collection

For the purposes of this study and to simplify the analysis, we selected the dataset `2021-10-14.YachidaS_2019`, which focuses on the disease 'Carcinoma Cancer' and contains data on 712 species and 616 samples. We extracted the relative abundance data, which indicates the quantity of each species found in each sample while excluding unrelated data on genetic markers. Additionally, we retrieved the sample metadata and converted the datasets into CSV files to ensure the data's integrity and quality. The conversion process is illustrated in the code below, which generates two CSV files. Moreover, we also created a function in R to map the species names with their taxonomy and added the latter as a new field in the CSV file.

```r
library(curatedMetagenomicData)
#data_check_all=curatedMetagenomicData(".*.relative_abundance",dryrun = FALSE, rownames = "short")
data <- curatedMetagenomicData("YachidaS_2019.relative_abundance",dryrun = FALSE, rownames = "short")

relative_abundance <- as.data.frame(data[[1]]@assays@data@listData[["relative_abundance"]])


person_data <- as.data.frame(data[[1]]@colData@listData)


relative_abundance$taxonomy <- matched_species

#write relative abundance to csv
write.csv(relative_abundance, "relative_abundance.csv")
#write person data to csv
write.csv(person_data, "person_data.csv")
```

Figure 2: Code

The first file contains all the relevant metadata associated with each sample, providing detailed information about their characteristics. This metadata includes key variables such as the study name, subject ID, body site, study condition, age, gender, BMI, and disease location, among others. These attributes offer important contextual insights that allow us to better understand the conditions under which the samples were collected and the individual features of each sample. The second file, on the other hand, outlines the relative abundance of each species across the samples, offering a clear representation of the species distribution within the dataset. This file provides data on the specific quantities of various microbial species found in each sample, helping to reveal patterns and variations in microbial composition across different conditions or patient groups.

Together, these files form the essential data foundation that will support and guide the subsequent stages of our analysis. By combining detailed metadata with species abundance data, we ensure that the analysis is both comprehensive and contextually grounded, allowing for meaningful interpretations of the microbial data concerning the study's objectives. (Table 3) summarizes the two key files generated in the information-gathering phase of the project.

CSV file	Columns
Metadata	study_name, subject_Id, body_site, study_condition, diseases, age, age_category, gender, country, non_westernized, sequence_platform, PMID, number_bases, minimum_read_length, median_read_length, curator, BMI, disease_location, aicc, brinckman_index, alcohol_numeric
Relative abundance	samples

Table 3: Metadata and Relative Abundance Files

Data cleaning and standardization

Once the data were collected and displayed, the next step involved filtering and retaining only the information that was relevant to the specific purpose of this study. This process ensured that the dataset was focused and manageable, eliminating any unnecessary details that could introduce noise into the analysis. As part of the cleaning procedure, certain columns from the metadata file were removed, as they were not essential for the current analysis. Additionally, instead of discarding the NA (missing) values for the metadata attributes, we decided to preserve them in case the data was reused and updated in subsequent studies. By retaining these values, we maintain the flexibility to incorporate additional data or refine the metadata as new information becomes available. (Table 4) summarizes the metadata columns retained after cleaning, highlighting key attributes relevant to this study while preserving NA values for potential future data integration.

CSV file	**Columns**
Metadata	study_name, subject_id, body_site, diseases, age, age_category, gender, country, BMI, disease_location, ajcc, brinckman_index, alcohol_numeric

Table 4: Metadata Columns Post-Cleaning and Standardization

Schema Generation

In the Schema Generation phase, we define a conceptual and logical structure to organize collected data for effective analysis. The objective is to create a schema that represents the relationships between various variables, ensuring both data and metadata align with the project's analytical goals. The objects and their properties for the schema are already reported previously in the ER description, so we expect the data layer of the knowledge graph to be composed of nodes representing the instantiation of the objects (person and species) and the edges that model the relationship between them. Additionally, our schema includes an event type, Correlation, represented as a node that links a Person with Species. This Correlation event node holds properties that classify whether a species has a high association with a disease, providing a basis for inferring inter-species relationships through methods like Pearson correlation. Therefore, representing our knowledge graph in this way facilitates interoperability and enables data reuse for future studies.

Formal resource generation

During the Formal Resource Generation phase of the project, we focused on creating and organizing the formal resources necessary for the subsequent stages of the analysis. These resources include the finalized datasets, data processing and analysis scripts, and detailed documentation that ensures consistency and reproducibility throughout the entire analysis process.