NUS Industrial & Systems Engineering - Systems Design Project

ACCELERATING P&G'S PRODUCT INNOVATION WITH A MICROBIOME SEARCH ENGINE

Group 6 Members | Joel Lim, Yu Minghui, Qi Ailin, Liang Yiyun
Supervisors | Dr Lam Tzehau (P&G), A/P Chen Nan (NUS)

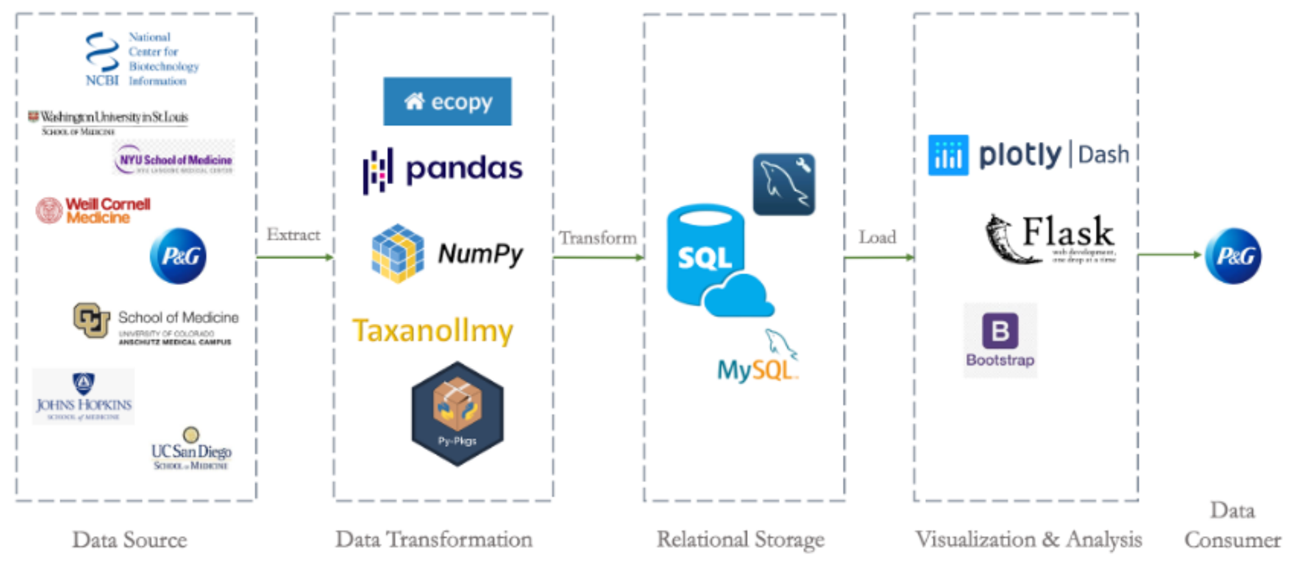
PROBLEM OVERVIEW

Metagenomic data, an important source to discover new genes for P&G's product innovation, has been growing rapidly among global research institutions and P&G research and development. Researchers at P&G do not have easy access to these data, as different datasets come in various file formats, with inconsistent or incomplete data fields, posing difficulties in searching, browsing, and analysing the data, which seems essential to conduct researches, derive innovative ideas, and make product relative decisions.

PROJECT OBJECTIVE

The project aims to provide solution in data standardisation, centralisation, and automated data management to enhance data accessibility and reusability. This is proposed to be achieved by developing tools that can compile diverse and unorganized metagenomic datasets on a centralised platform for P&G's research and development. The tools should consider not only the functional efficiency of cleaning, storing, managing, and querying the data, but also the easy-to-use practical aspects to make it intuitive enough for any end-users who may not have a programming background to utilise the data to derive valuable insights. In brief, our team targets to establish a data pipeline that contains essentially a relational database and a user-interactive data visualisation dashboard.

DATA PIPELINE



The overall data pipeline designed by our team consists of:

1. Data gathering from various sources
2. Data wrangling to accommodate relational database schema
3. Data storage, test and validation within relational database
4. Data visualisation and analysis on front-end dashboard

The following sections detail each part of the data pipeline.

DATA PROCESSING & DATABASE

This section involves the first three parts of the data pipeline.

1. Data Source

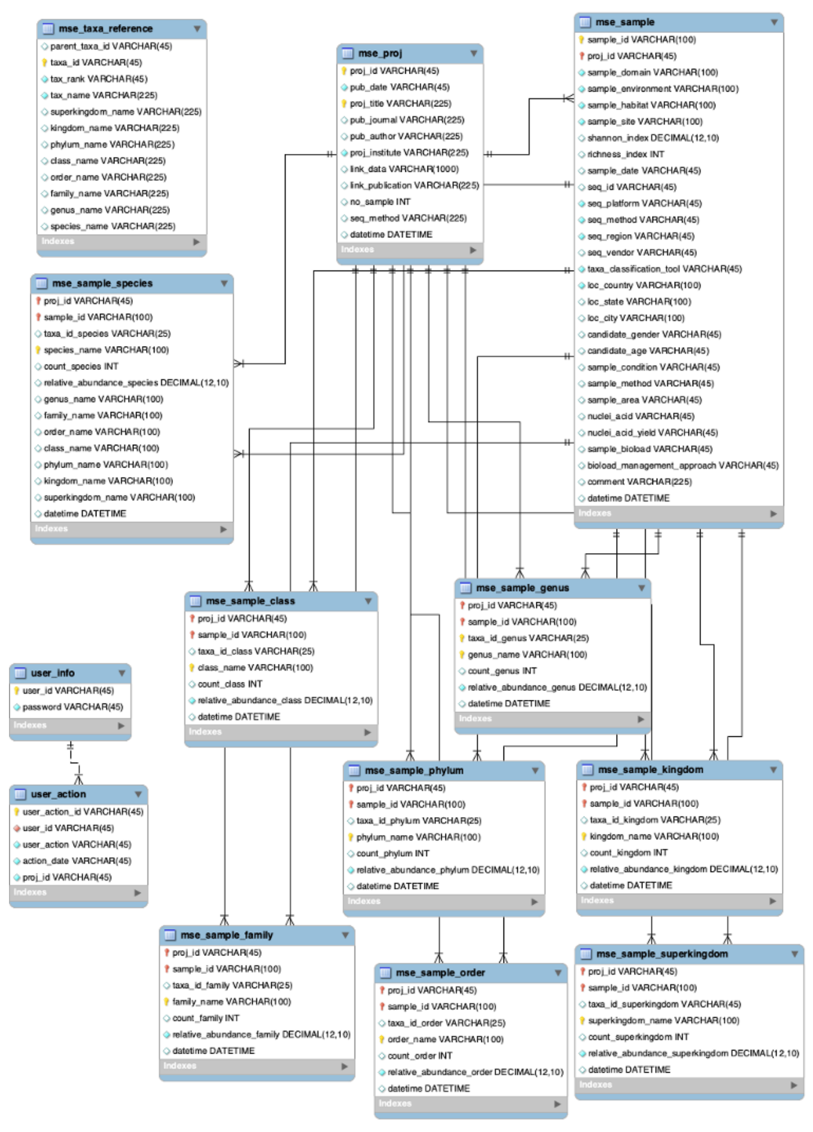
Data was provided by P&G, consisting of publicly available research and private research done by the company.

2. Data Transformation

Data cleaning and processing was done to ensure the data fits the database structure. The team also created a Python package, MSEDataInput, to facilitate the uploading of new data to the database.

3. Relational Storage

All data is stored and managed in a MySQL relational database, according to the schema designed by our team.

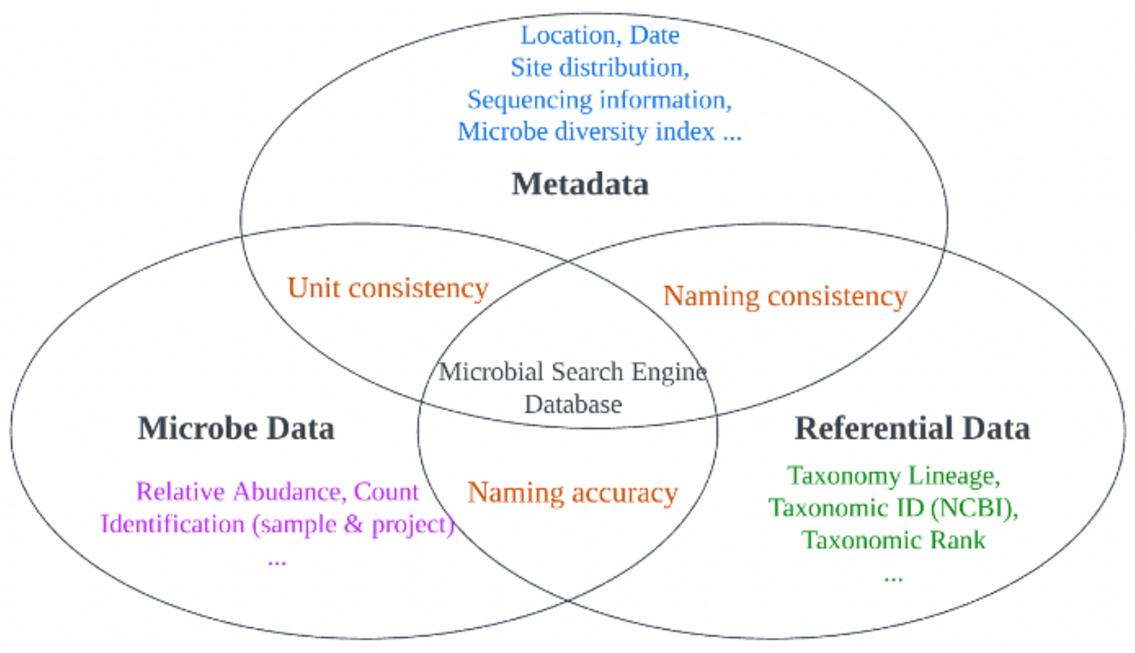


Database Schema

The figure on the left shows the structure or schema of the database. The processed data is split into tables that are related to each other, allowing data to be retrieved quickly and effectively.

Data Distribution

The figure on the right shows the various types of data present in the database and how they are related and processed for uniformity.

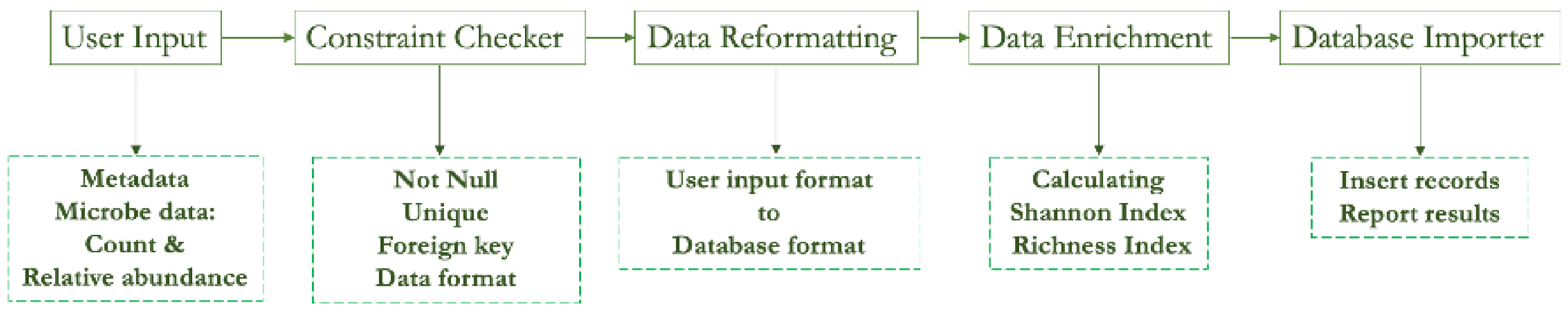


Data types and Relationships in MSE Database

MSEDataInput Package

A package the team created to allow new data to be added to the database automatically.

The figure on the right shows the process by which new data gets processed and added to the database.

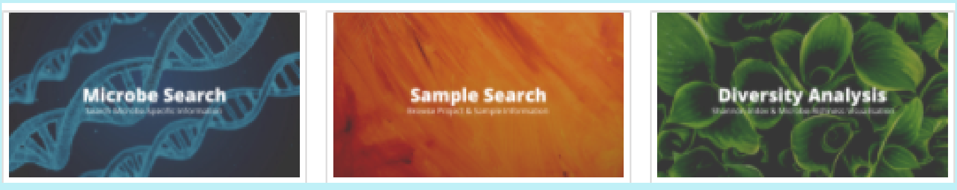


MSEDataInput Package Functionalities

DASHBOARD OVERVIEW

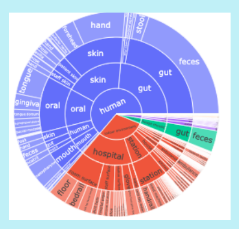
HOME PAGE

The landing page contains a database summary and buttons that direct users to the Search Germs page, the Search Samples and the Diversity Analysis Page.



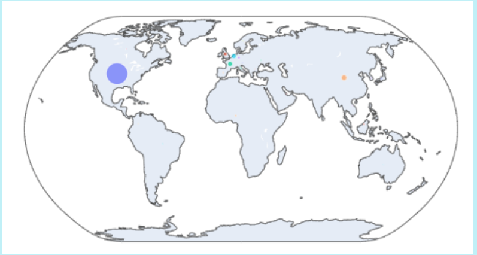
Database Summary

The data summary page includes the following analysis:



Site Distribution

Relative Abundance of Top 10 Microbes



World Map of Samples of all Projects

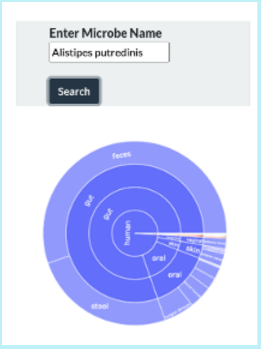
The simple analysis and visualisation tool here gives the user an intuitive overview of the data in the database. It helps to identify what can or cannot be found in this database and upload additional resources to continuously ameliorate the database.

MICROBE SEARCH

When given the name of the microbe and its taxonomy level, users are able to find the corresponding microbe's site distribution, relative abundance, and other information on the samples that contain the microbe.

Sunburst & Relative Abundance

The relative abundance chart and the sunburst chart together are able to help researchers in P&G to spot new white space of their product where unmet needs of customers may be discovered.



Site Distribution

Relative abundance of the searched microbe

Sample Distribution World Map & Project Summary Table

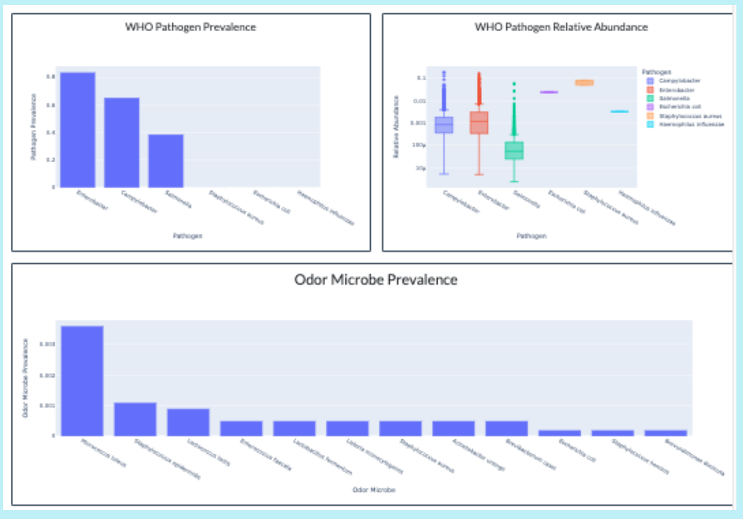
A significant amount of hours can be saved by the researchers from digging through research samples containing the specific microbe or samples from a specific country. They are able to give more pertinent information to the researchers.

SAMPLE SEARCH

As opposed to the microbe search, the sample search page is used when there is not a known microbe and the user is trying to identify what microbe can be found with given search criteria.

Pathogen & Odor Organism Identified

Based on the user search, instead of displaying all resulting microbes which may not be harmful, only the WHO pathogen and odor organism's prevalence and relative abundance are displayed, as P&G products mainly tackle pathogens and odor organisms.



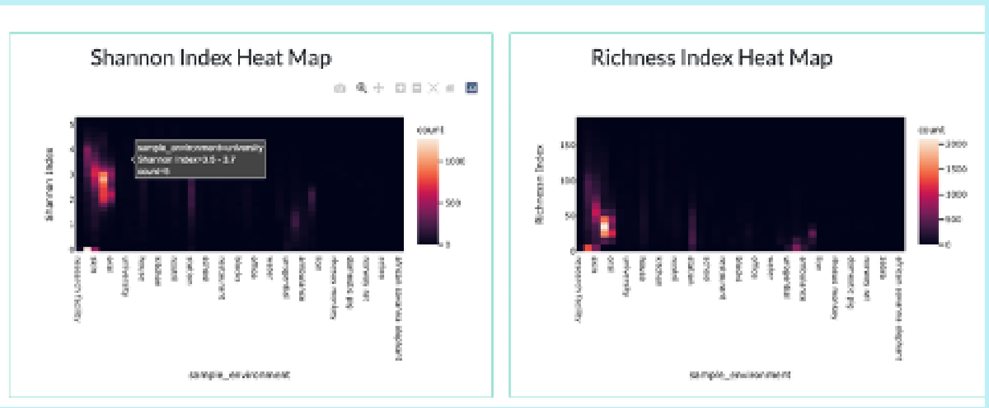
Pathogen & Odor organism Identified

Microbes Ranked by Mean Relative Abundance

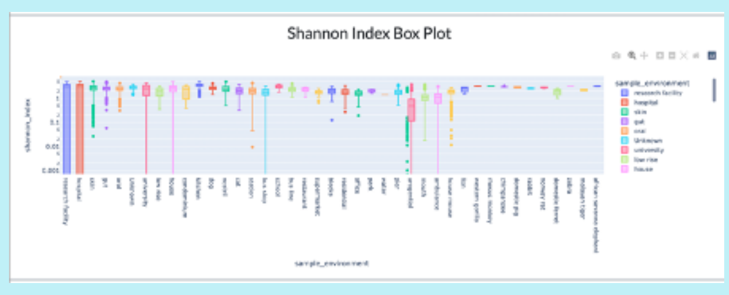
The tables are used to cross-reference with the pathogen and odor organisms prevalence table.

DIVERSITY ANALYSIS

A diversity index is a quantitative measure that reflects how many different types (such as species) there are in a dataset or a community



Shannon index and richness index heat map



Shannon index box plot

Shanon Index

Shannon diversity index tells you how diverse the species in a given community are. It rises with the number of species and the evenness of their relative abundance.

Richness Index

Richness Index simply quantifies how many different types of microbes the dataset of interest contains.

KEY OUTCOMES & ACHIEVEMENTS

Project Outcomes & Benefits

The team created a database that serves as a centralised location for all of P&G's metagenomic research data to be stored and easily retrieved, and a dashboard that allows for intuitive data visualisation and data analysis to be completed.

The solution contributes to P&G in expanding their market reach as they may use these tools to further their research capabilities for continuous product innovation and development. It will also enable researchers to tap into unseen areas and white spaces in product research by providing convenient and powerful access to data.

Project Achievements

The team has successfully met P&G's expectations of addressing the data issues they encountered by providing a valuable and timely solution to their problems.

With this new tool, researchers at P&G will be able to conduct research faster, more conveniently and more effectively. This tool not only centralises the data for any user to access, but the dashboard also provides useful data visualisations and summaries, allowing users to rapidly understand the data they are working with.

Skills Developed

Over the course of this project, the team has further developed their skills in the following areas:

- Programming & Software Tools - Python, SQL; MySQL, Plotly-Dash Framework
- Project Management: Planning, Scheduling, Stakeholder Management
- Product Mangement and Development, Agile Methodologies

Our combined experience as ISE students has enabled us to use and grow our programming, project management and communication skills to develop a useful solution for P&G by leveraging on each member's strengths and weaknesses.