Steps to follow to recreate the Microbiome Knowledge Graph:

1. **Data extraction**
2. Launch https://portal.hmpdacc.org/.
3. For 'Get Started By Exploring' button, select 'Data'.
4. From the filters on the left side, select the following:

Projects: Integrative Human Microbiome Project

Body Site: feces

Studies: IBDMDB

1. From the pie chart for 'File Counts by File Type', select 'abundance\_matrix' in the pie chart.
2. Navigate to the 'Files' tab and add all files to cart.
3. Move to the cart and download both 'File Manifest' as well as 'Sample Metadata'.
4. For the two files, 'sample\_id' is a common field. Using this, merge the two files to create one single Excel file.
5. **Data preparation**
6. Load conda microbiome\_env.yaml to make sure that all needed python libraries are available.
7. Create '*<cohort\_name>*.xlsx' files based on the 3 cohorts: UC, CD, nonIBD.

**Note**: Groups are only created to process large amount of files in smaller steps. This step can be omitted if below step can be run for all the files to be downloaded, if processing capacity is adequate.

1. Run python script "prepare\_biom\_file\_from\_url.py". It has one parameter: excel file that has to have column "urls" containing urls for biom files to download

>python prepare\_biom\_file\_from\_url.py *<filename>*.xlsx

1. Biom files are downloaded into the folder "data".
2. Biom files are convreted into tsv format and slightly modified to make them ready for neo4j loading.
3. Processed tsv files are in the folder "import\_data".
4. **Data pre-processing**
5. Run python scripts ‘Pathway\_preprocessing.py’, ‘Taxonomy\_preprocessing.py’ and ‘Enzyme\_preprocessing.py’. Follow the instructions in the .py files to run them for all 3 cohorts one by one.
6. Processed, merged excel files are downloaded in .csv per cohort per category (pathway/enzyme/taxon). Pathway files are called ‘<cohort\_name>\_processed.csv’, Taxonomy files are called ‘<cohort\_name>\_tax\_processed.csv’ and Enzyme files are called ‘<cohort\_name>\_enz\_processed.csv’.
7. Convert them to .txt files to make them suitable for loading onto Neo4j.

**Note**: This could be bypassed by doing either of the following: export statement in .py files could be changed to .txt instead of .csv, or Neo4j loading query could be changed to take in .csv files instead of .txt files)

1. **Data loading**
2. In Neo4j Desktop, create a new project. Inside the project, create 3 Local DBMSs (one for each cohort).
3. Place the necessary pre-processed files in the following path:

C:\Users\<name-of-user>\.Neo4jDesktop\relate-data\dbmss\<long-unique-name-for-each-local-dbms-created>

1. From ‘Cypher queries’, load sample-level information from 'Master\_List.xlsx' to create the following:

nodes: Subject, Site, Visit number, Diagnosis, Sex, Race, Age &

relationships: Subject -> Site, Visit num, Diagnosis, Sex, Race, Age

1. From ‘Cypher queries’, load cohort-level information from each of the processed Excel files to create the following:

nodes: Pathway,Enzyme, Genus

relationships: Subject -> Visit num -> Pathway -> Genus, Subject -> Visit num -> Enzyme -> Genus & Subject -> Visit num -> Genus

**Note**: Please increase dbms.memory.heap.max\_size upto 8G for each Neo4j database.