INLS 623
ETL Lab
Deliverable 1
07/14/2022
Somang Song

PubMed is a web interface integrating a search engine powered by the NIH. It is a biomedical literature database which central database contains the full text of publications in the database. Individuals can use PubMed to do direct searches by entering search terms on web pages.

National Library of Medicine provides document type definition (DTD) that defines the structure of data in the MEDLINE XML files. A document database is a nonrelational database designed to store and query data as JSON-like documents. Therefore, PubMed needs database schema and database loading tools to work with SQL (Structured Query Language) server. This is why it imported as 0 columns when I tried to import JSON files to SQL for the assignment.

It is obvious that converting PubMed tat into SQL is challenging. This is because SQL is a relational database management system, a type of database that stores and provides access to data points that are related to one another. It is one of the systems that are easy to quantify and rank order results therefore, an ideal tool for both quick and complex queries.

The reason why it's not readily solvable is documented databases store all information for a given object such as PubMed identifier (PMID) in a single instance in the database, and every stored object can be different from every other. On the other hand, relational databases generally store data in separate tables that are defined by the programmer. This is the reason why UNION the two tables pubmed_raw and pubmed_raw2 was impossible due to the different number of columns.

Converting XML DTD to a relational database is not readily solvable because of its own characteristics restrict the ability to inquire. Therefore, it should provide the schema that XML describes the structures and attributes.

References

Dynomant, E. (2017, October 18). *MEDOC: a Python wrapper to load MEDLINE into a local MySQL database*. arXiv.Org. https://arxiv.org/abs/1710.06590

Oliver, D. E., Bhalotia, G., Schwartz, A. S., Altman, R. B., & Hearst, M. A. (2004). Tools for loading MEDLINE into a local relational database. *BMC Bioinformatics*, *5*(1), 146. https://doi.org/10.1186/1471-2105-5-146

Ye, F. (2009). Converting XML DTD to Database. *2009 International Workshop on Intelligent Systems and Applications*. https://doi.org/10.1109/iwisa.2009.5072794