项目规划

**一：总体目标**

设计一个通用数据模型，实现将不同来源的数据库中的电子病历数据通过提取、加载、转换到目标库中，实现数据标准化。

**二：数据来源**

数据包括医院的中文电子病历数据，其中一部分是电子病历中包含的结构化数据，另一部分是通过NLP识别后讲电子病历中非结构化数据转化成的结构化数据。

**三：拟用模型**

拟用健康观测数据科学和信息学组织（OHDSI）开发的观察医疗结果合作项目通用数据模型（OMOP CDM ）作为标准化模型。开源工具多，可实现性强。

拟用Kettle作为ETL工具。纯java编写，且有文档介绍。

**四：总体流程**

1. 源数据：了解源数据电子病历结构，了解数据库中表格字段名称
2. 数据预处理：数据去重、处理缺失值、处理混乱数据

非结构化数据转化为结构化数据

1. 后结构化数据处理：针对NLP识别处理之后的后结构化数据，进行整个转换，便于后期处理。
2. 源数据分析：White Rabbit工具

应用此工具扫描源数据，生成扫描报告，了解原数据库结构。

1. 建立表格和字段的映射关系：Rabbit in a hat工具

跟据扫描报告，建立原数据库于模型数据库的映射关系。

1. 标准术语映射：Usagi工具

对于电子病历中的医学术语，建立与标准术语的映射关系。

1. 标准术语映射表中存在的本地术语：直接建立映射。
2. 标准术语映射表中不存在但是存在同语义的本地术语：语义归类映射
3. 标准术语映射表中不存在并且也么有同语义的本地术语：参考当地医学规范，进行术语讨论。改进标准术语映射表。

注：拿到的数据是中文电子病历数据，进行语义映射可进行翻译（官网有说明）

1. ETL实现---工具Kettle

跟据上述生成的需求文档，对元数据库中的数据进行提取、转换、加载。

1. 标注化后数据的存储。

对提取、转化加载后的数据进行存储。（1.存储到关系型数据库中。2.建立本地存储模型，写java接口用于导出数据）。

1. 数据质量评估

评估数据仓库中和元数据库中的数据差异程度，用来判断通用数据模型的质量性能。

同时对整个流程进行审查，保证数据的准确性和可用性。

**五: 问题与思考**

1. 数据来源

我们数据来源是中文的电子病历数据，OMOP CDM是英文的通用数据模型，虽然有论文将韩语、德语等其他语种电子病历应用到此模型，但中文数据应用比较少。要考虑语种差异。

1. 工具的使用

OMOP 提供的工具主要适用于英文，针对中文数据需进行修改和优化---模拟

1. 后结构化数据的处理和使用

将电子病历中的结构化数据转化为非结构化数据之后，针对其存储格式，考虑如何进行后期使用和整合。

1. 专业术语映射

本地术语为中文术语，模型术语为英文术语，由于中文标准化术语未统一，故标准化术语映射环节需医学相关人员参与制定。

1. 结果

若采用OMOP CDM的工具及流程，能完成数据标准化，能否达到毕设要求。（工具未开发，用的现成的）。