数据库管理规则

胡华平

2020-07-13

Table of Contents

# 1.变量体系

**基本要求**：

* 变量系统具有可维护性。
  1. 首先是变量可以持续增加调整。因为研究目的需要而要求增加新数据，从而引入新变量。
  2. 其次是变量系统本身的版本演进有要有记录信息。例如v2018.6、v2019.8。主要是考虑数据和变量系统是分别隔离存放的，否则会引起调用错误。
* 可以实现快速字典化查询。例如：英文变量和中文变量的对应查询。
* 坚持中文和英文“双变量系统”。代码操作中主要使用英文，报告输出主要使用中文。

# 2.数据体系

区分两大类数据模块：

“粒度数据”主要指数据获取时，最小的数据单元。一般而言，不同渠道的来源数据，都是按某一个指标或某一类指标查询获得（或平台或网站），那么每次查询就会获得一个单元的“粒度数据”。“粒度数据”主要处在 **原始数据**(raw data)和**标准数据**（standard data）两个阶段。

“研究数据集”，是指研究者基于特定研究目标，为了获得有力论据而采用的“数据集”。因此“研究数据集”往往是由“粒度数据”组合而成的。“研究数据集”主要处在**加工数据**（process data）、**分析数据**（analysis data）两个阶段。

简单说，“粒度数据”模块的处理工作主要是数据爬虫、抓取、清洗等基础操作技术；而“研究数据集”模块处理工作主要是关心选择哪些变量以形成论据支撑。

区分数据阶段：

1. **原始数据**(raw data)阶段，是数据获取的最初形态，文件类型各式各样，数据呈现方式也各不相同。大多不是规范化表格，不利于数据合并和整理。例如，从搜数网下载的.html表格数据文件；从FAO数据库下载的.csv数据文件；从商务司下载的.pdf数据文件。
2. **标准数据**（standard data）阶段，是根据规则制定的数据形态。主要有两个要求：一是数据呈现方式要求为“标准表”（每1行代表样本、每1列代表变量）；二是数据存放格式要求为.csv（体积小、格式通用、便于读取）。
3. **加工数据**（process data）阶段，是对“粒度数据”中的标准数据（standard data）文件，进行初步整合（主要是行合并，也即**年度数据**的合并）。
4. **分析数据**（analysis data）阶段，是对“加工数据”文件，进一步整合（主要是列合并，也即多个**变量指标**合并）。

**基本要求**：

* 说明文档的编写必须简要和清晰。数据收集整理过程中，要仔细记录统计口径、资料来源、操作人、操作时间、处理过程等内容。
* 变量体系和数据体系要一致。可以根据变量体系对数据体系进行区块化管理。
* 不同阶段的数据文件都要单独存放，目的是便于检查和纠错。例如，年度农业产值汇总数据的**原始数据文件**存放在文件夹data/v1/01-nycz-00-hz/html-source01-01/下，其对应的**标准数据文件**则存放在文件夹data/v1/01-nycz-00-hz/csv/下。
* 文件名、文件夹、和文件路径设定要统一规范。例如，年度**原始数据文件**应命名为2018.html（来源于搜数网查询），其对应的**原始数据文件**命名为2018.csv。

# 3.数据维护

数据维护，主要涉及到数据更新、数据存放、数据查询和数据下载等。数据更新维护和权限分配管理是关键，基本上按照前述相关说明即可完成，但是需要设定人员角色。

* **数据录入人**：根据任务分工，进行粒度数据（包括原始数据和标准数据）处理的人员。
* **数据检查人**：下发数据录入任务的人员，应该要对录入人的数据进行基本核查。
* **数据使用人**：因为研究需要，而提出数据使用要求的相关人员。
* **数据授权人**：对数据知识产权或法律风险进行评估和管理的人员。

# 4.团队协作

大部分现代化团队协作平台往往具有权限管理的功能，利于github平台。

影响团队使用协作平台的主要障碍包括：

* 领导力问题。小组制管理中，小组长的沟通协调能力。项目制管理中，项目负责人的进度控制能力。团队管理中，负责人的方向把控能力。
* 人员技能问题。例如相关团队成员不能熟练掌握分布式协作技能，如git命令。又或者相关人员对统计分析语言（如R或Python）掌握程度高下不一。
* 网络和硬件问题。例如网速过慢（以及无法VPN）、硬件设备不支持或支持不够（例如多外接显示屏）。

协作策略：

* 按“项目制”进行任务分工和推进
* 推广使用现代化协作平台和技术