第三章 结构化数据库的设计与开发

> 本章以 **《现代汉语词典》** 和 **《康熙字典》** 为例介绍结构化语料库的数据建模、数据处理、数据库建设及应用等问题。

一. 关系数据库建模

1.1 简介：

所谓数据建模，就是根据研究对象的结构类型特点，结合自己的研究需要，为其设计出一个合适的数据模型的过程。在关系数据库中，数据模型可以简单的理解为由字段（列）和记录（行）组成的数据表结构。一个关系数据库可能包含多张表，表与表之间通过 **键** 建立联系。

1.2 关系数据库示例

为节约存储空间、方便后期维护，在进行关系数据库建模时，常将重复出现的信息置于单独的表中，随后在数据库中通过外键进行链接。

例：学生选课表关系数据库结构

单表式结构

学号	姓名	年级	专业	所选课程	课程编号	任课老师
M0031	张三	大二	中文	现代汉语	c003	王老师
M0031	张三	大二	中文	音韵学	c012	张老师

可将词表重新组织为以下三张表，这种处理方案有两大好处：

1. 节省存储空间：表一中学生信息只出现一次，表二课程信息也只出现一次，表三通过学号和课程编码可关联学生信息和课程信息。
2. 方便维护：假如如课程信息发生变化，如现代汉语任课老师改变，在多表结构中，只需要在课程信息表中修改任课老师（一次）即可，而如果采用单表式结构，需要修改所有含有此课程的记录。

表一：学生信息表

学号	姓名	年级	专业
M0031	张三	大二	中文
M0045	李四	大二	中文

表二：课程信息表

课程名称	课程编码	任课老师
现代汉语	c003	王老师
音韵学	c012	张老师

表三：学生选课表

学号	选课编码
M0031	c003
M0045	c012

1.3《现代汉语词典》数据建模

1.3.1 《现代汉语词典》的结构

- 分字条、词条两部分。
- 字条可粗分为3块：`字头`、`拼音`、`释义`（释义可根据需求继续细分）。
- 词条可分为`词头`、`释义`两部分。
- 词条隶属于某个字。

1.3.2 用户需求分析

- 通过字头检索汉字拼音、释义。
- 通过拼音检索符合条件的汉字。
- 通过词头检索词条释义。
- 词头和字头的隶属关系在检索系统中并不重要。

1.3.3 数据结构：

- 字典表

zitou_id	字	拼音	释义
0001	阿	ā	〈方〉前缀。（1）……

- 词典表

citou_id	词	释义
0001	阿昌族	我国少数民族之一……

1.4 《康熙字典》数据建模

1.4.1 《康熙字典》的结构

- 卷首：序、凡例、检字、辨似、等韵
- 正文：分`正集`、`补遗`、`备考`三部分（篇）
    - 每部分下按子、丑、寅、 卯十二地支分上中下列集。
    - 每集下按部首排列字条，正集215部。
    - 每个部首下所辖字根据部首外笔画数升序排列。
    - 每个字条分字头、释义两部分。

1.4.2 用户需求分析

- 根据字头查找释义。
- 由字头查找其出处，以便回溯纸本校对。
- 其他需求：引书统计、

1.4.3 数据建模

1.4.3.1 单表式结构

zi_id	篇	集	部	部首外笔画数	字头	释义
Z00001	正集	子集上	一	0	一	……

1.4.3.2 《康熙字典》的多表式结构

单表结构中的`篇`、`部首`等信息是重复信息，可将其提取出来单独建表，由于《康熙字典》数据结构相对简单，并不一定要采用多表式结构。多表式结构在数据库中便于管理（1.2中所列优点2），但是在Excel中做数据分析时，单表式结构因数据内容直接在表中展示，方便人工观察，因此更加合适。

- **篇表：**

Pian_ID	篇名
P01	正集
P02	补遗
P03	备考

- **部首表：**

bu_ID	部首
B001	一

- **集表**

JI_ID	集
J001	子集上

- **《康熙字典》字表：**

zi_id	篇	集	部首	部首外笔画数	字	注释
Z00001	P01	J001	B001	0	一	……

二、数据预处理

2.1 数据规整化

- 使文本格式统一，去除无关信息。
- 使用制表符、换行符将数据预处理为文本式表格形式。

2.1.1 相关查找替换模式

- 出处切分：

  集+部： 查找 `(^9【[!】]{2,3}】)(【[!】]{2,3}】^9)` 替换为 `^9正集\1^9\2`

篇+集+部： 查找 `(^9【[!】]{2,3}】)(【[!】]{2,3}】)(【[!】]{2,3}】^9)` 替换为 `\1^9\2^9\3`

2.2 文本导入excel

方法一：粘贴法

直接将使用制表符和换行符分割的文本数据粘贴到excel即可。

方法二：Excel外部数据导入

`数据` —— `获取外部数据` 根据数据来源可使用不同的数据导入方法。

三、Excel数据处理与数据表建设

3.1 数据处理

Word和Excel在数据处理中有不同的优劣势，可根据材料的具体情况灵活选用处理工具。

3.1.1 数据表问题发现及修复

- 在表中仍有发现篇、集切分错误的问题，需根据错误特点进行修复。
- 字头列中发现有部分四字节字被错误转换为 `？？` 的情况，因有两个字头字段，可使用正常的一列恢复错误的列。

3.2 数据表建设

- 创建 `篇目` 表，并导入篇目数据。
- 创建 `字` 表，导入数据。
- 创建 `部首` 表和 `集` 表，由 `字` 表提取部、集信息，并导入数据。
- 为 `篇`、`部`、`集` 表分别给定id，作为主键。
- 用 `篇`、`部`、`集` 表中id替换 `字` 表中的对应内容。

四、Access数据库建设

4.1 数据导入

数据——外部数据——Excel

4.2 建立表间关系

拖放操作建立各表主键、外键之间的联系。

五、Access数据库应用

Access数据库建设完成后，可以有多种不同的应用方法：

5.1 查询

- 通过查询向导或查询设计建立查询
- **进阶：** 学习`SQL`——结构化查询语言(Structured Query Language)

5.1 窗体

窗体是Access软件内部可以实现的一种类似软件式的简单程序，可以方便展示查询结果，实现一些简单的检索功能。

5.2 高级应用

- Office内部通过VBA开发简单的检索程序。
- 第三方程序设计语言（VB、C、VB.NET、java等等）均可调用、查询、更新Access数据库。