统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-03-26

第二章 数据收集、整理和清洗

2.1 数据目标

2.2 数据收集

2.3 资料整理和数据清洗

2.4 数据的数据库化

2.5 数据质量

2.6 抽样设计

2.7 抽样分布和抽样误差

2.8 问卷设计技术

2.4 数据的数据库化

数据库化的类型

一手资料的数据库化

二手资料的数据库化

若干实例

为什么需要把数据进行数据库化？

> 数据不仅整理好了，也清理好了，是不是就可以分析研究了呢？

- 采用手工计算的情形几乎已经消失了。数据的数量与复杂程度，已经超出了人们运用大脑、纸和笔，直接处理的程度。
- 调查数据的分析与研究，从计算机应用普及以来，就已经主要依靠计算机了。运用计算机是最有效和最快捷的方式。
- 运用计算机就需要满足计算机对数据的要求，那就是**数据库**。
    - 清理整理好的数据，要变成计算机可以读取并进行运算的数据格式，通常这一类的格式都是数据库格式。计算机应用程序不同对数据库的格式要求也不相同。

数据库化的目的就是为了便于分析和使用。基本的要求是通过数据库化，让调查数据格式化、结构化，符合统计分析、计算的要求。

数据库化的类型

数据的数据库化，就是把得到的变量、变量属性或者标签输入计算机，变成结构化的**数据矩阵**。从数据库化的目标来分，主要有如下两类：

- **计算机网络系统的数据库化**，主要是用于存储数据，有各种类型的数据库应用程序。
  - 常见的结构化数据库SQL数据库有有多种，比如开源的免费的My Circle。
- **分析计算用的数据库化**，主要是通过建立数据库，用于统计分析软件的计算。
  - 我们这里所学的就是这一类数据库化。

我们主要学习常用的运用于计算机**单机**统计计算与分析用的数据库化。

大数据的数据库化有不一样的特点和需求。

数据的数据库化示例

SPSS是社会科学统计计算运用比较多的一个大型统计计算软件

SPSS数据库的**数据视图**:

- 每一行代表样本,
- 每一列代表变量
- 中间单元格表示数据取值

SPSS数据库的**变量视图**

- 每一行代表一个变量
- 每一列表示变量的性质和特征

A.调查数据的数据库化（主要步骤）

问卷调查的数据，在完成了问卷的审核、归档、清理以后，在用于分析软件的分析之前，就需要把它转化为数据表示的数据库。通常有三个步骤：

- 第一步，**编码**。
  - 在清理工作中，这项工作应该已经完成了，不过在数据入库之前还需要审核。
- 第二步，**数据录入与转化**。
  - 如果是纸版问卷调查，这个时候就需要录入数据。建议采用专门的**录入软件**进行录入，尽量避免录入中出现的差错，进而降低调查误差。
  - 如果是计算机辅助调查，这个时候就需要转化数据。无论是内容转化还是格式转化，也建议尽量采用可靠的工具，避免出现差错。
- 第三步，对录入完成和转化完成的数据，做基本的**检验和清理**。
  - 最容易出现的差错就是错行、错列造成数据的混乱。

A.调查数据的数据库化（编码）

编码：就是把调查问卷的每一道访题用符号或者数字组合代码换，包括对每一道访题的选项或应答赋值。

- 每一道访题的编码就是数据库表中的变量。
- 应答赋值就是数据库表中的**变量值**，这个只有各种属性，就是数据库表变量视图中的各种标签，又称为**变量标签**。

A.调查数据的数据库化（编码示例）

我们来看例子，这是Self PS中的一道访题，

> 【问题】1.5，请问您希望孩子念书，最高念到哪一个程度？（共7个选项）。
> 【选项】A.小学；B.初中；C.高中；D.专科、职高、技校、大专；E.大学本科;F本科以上;G.不必念书

对这套**访题**我们可以这样编码，访题可以编为B15。（为什么这么编?）

对选项的编码，就以选项的编号做编码。

A.调查数据的数据库化（编码）

问卷调查数据的编码，一般有三种方法：

- 第一：**原始编码**，就是直接运用问卷的编码。
  - 通常这种方法仅仅用在访题数量极少，应答非常简单的情况下。
- 第二：**先编码**，在调查开始之前，编码工作就已经做好了。
  - 通常这种方法会用在基本上都是封闭访题的情况下。
- 第三，**后编码**，就是在问卷调查完成以后再做编码。
  - 只要是有开放访题，一般都会采用这种编码方式。

编码部相当于问卷数据的一个**索引**，把变量、变量值，变量标签关联起来，类似于一本问卷数据字典。

A.调查数据的数据库化（录入）

- 对于简单的问卷调查，可以运用常用的办公软件或统计分析工具来做录入，
  - MS Office Excel
  - Mac Numbers
  - SPSS
  - Stata、statistica、R…
- 对于相对庞大复杂的问卷调查，需要使用专门的数据录入软件。
  - 商业收费的SPSS Data Entry模块
  - 免费的EpiData

专用录入软件的能提高录入效率，并减少录入误差：

- 可把纸版问卷计算机界面化，把纸版问卷完整呈现在计算机屏幕上。
- 可通过对跳转、阈值、变量类型等的控制，尽量减少录入所带来的误差。
- 在录入完成以后，还可以直接把录好的数据导出为数据库表文件。

A.调查数据的数据库化（检验和清洗）

针对的已经数据库化的数据，通常需要运用**统计分析方法**进行检验和清洗：

- 第一，录入错误清理。可以把双录入的数据输出为一个清理数据库，核对录入中出现的冲突数据。
- 第二，编码清理。对不在编码值范围的变量值进行清理。
  - 假设性别属性值的编码原本只有0和1，如果在数据表中出现了其它值，那就一定是哪里有错误了，就需要清理并且改正错误。
- 第三，逻辑清理。主要是针对基本事实逻辑的清理。
  - 比如样本为男性，在是否怀孕的访题下，变量值说明他有怀孕记录，这就是逻辑错误

A.调查数据的数据库化（检验和清洗）

数据库检验和清洗还需要注意如下问题：

- **离群值**：偏离了日常理解的范围，但实际上可能是有效值的一部分。
  - 男性怀孕令人奇怪，女性怀孕就没有什么让人奇怪的了，对不对？
  - 女性16岁-49岁之间怀孕都是正常的。如果数据显示有一位七十岁的老奶奶怀孕了， 有没有可能呢？
- **极大值**和**极小值**， 都是需要再次确认的变量值。
- **无应答**的处理，通过分析已经应答的数值，确定对无应答的处理方式，比如差值。
- 变量的**再编码**，在数据的清理中也可以产生**衍生变量**。
  - 比如受教育程度或者年龄的重新分组
  - 比如依据受教育程度和收入来建构社会经济地位

A.调查数据的数据库化（清单）

正常的完成了数据库化的问卷数据，至少应该包括以下的文件：

1. **调查问卷**（已经有了）
2. 调查问卷的**数据库编码手册**（已经有了）
3. **两个数据库**，一个是完成问卷的数据库，一个是未完成问卷的数据库。
4. **样本数据库**，通常抽样完成以后，一定有一个数据库。这个数据库包括了用于抽样的变量、抽样单位、分层变量、权重变量等，这些应该是分析研究之前已经有的数据。
5. **抽样报告**、**实施报告**，这两份报告用于判断数据质量，制订分析策略。
6. 完成的、未完成问卷数量的**统计表**。通常用表格方式展示出来。
7. **数据清理报告**，对变量的可分析性要进行说明。

B.访谈调查数据的数据库化（主要步骤）

对**访谈调查**的数据，在完成了访谈笔记的整理、格式化、归档、清理之后，在用于分析之前也需要把相关的信息录到数据库中。虽然不一定可以像问卷调查数据那样完全的数据库化，至少**访谈记录**与**整理信息**应该数据库化。

- 第一步，编码。
  - 记录信息的编码（重点工作）
  - 记录内容的编码（如果要进行**文本分析**，则需要此步骤）
- 第二步，录入。
  - 录入访谈记录信息，便于检索，也便于查找。
  - 如果要做内容分析，访谈内容就需要全部地录入。
- 第三步，清理。一般需要逐行核查。**内容数据**是没有办法采用统计分析方法的进行核查。

B.访谈调查数据的数据库化（编码）

访谈数据的编码有两类：

- **访谈记录信息**的编码。基本变量有记录编号、访谈时间、地点、人物、主题、位置图。如果有日志信息，也需要把日志信息加入其中。
- **访谈记录**的编码。如果希望编码的程度可以直接应用到**内容分析软件**的分析，那么就需要学习专门的课程，不同的分析软件对编码的要求是不一样的。

B.访谈调查数据的数据库化（录入）

访谈数据的**录入工具**：

- 要是涉及到数字数据的，就可以使用Excel、SPSS、Stata、Statistica、r等等
- 对文本数据，就可以使用Word，当然也可以使用Numbers和Pages。
- 对访谈内容，还可以采用**内容分析软件**，比如Nvivo、Aquad、ATLAS.ti和Qualrus。

访谈数据录入的**几个要点**：

- 录入策略问题。对于访谈记录信息的录入，尽量采用标准化的格式，目的是便于交换、便于交流。
- 文本格式问题。一般可以先转录为纯文本格式，注意纯文本格式有一个编码问题，最好采用通用的编码，比如Unicode。

B.访谈调查数据的数据库化（清单）

访谈数据的数据库化产出也有一份清单，至少要有以下的**数据文件**：

1. 调查提纲或者访谈提纲，或者访谈设计。
2. 访谈记录的整理、清理的数据库
3. 访谈内容的数据库
4. 访谈记录的数字化，也就是数字化的过程及报告
5. 最后还有清理报告

C.观察数据的数据库化（主要步骤）

观察数据怎么数据库化呢？主要也是三个步骤：

- 第一步，编码。
  - 观察调查数据的编码与其他编码不一样的地方在于观察记录信息比访谈记录信息要丰富得多。当然对观察记录的内容，如果希望用作分析素材，也需要编码。
- 第二步，录入。
  - 在大多数情况下，主要录入观察记录信息，同样，如果要把观察记录的内容作为统计分析的素材，那么也需要把它录到数据库中。
- 第三步，清理。
  - 同样在录入完成之后，要对已经录入的数据进行核查，如果有观察记录的内容，就需要对已经数据库化的内容做仔细的核查，确保内容准确。

C.观察数据的数据库化（编码）

观察数据的编码主要包括两个方面：

- 观察记录信息的编码。基本变量包括记录编号、观察的时间、地点、事件、主题，还有观察媒体（望远镜/摄像机/眼睛）。如果有**日志信息**，也可以把日志信息列入其中。
- 观察记录内容的编码。即使观察记录的内容不会作为统计分析的素材，最好还是录入为数据化的文本文件，便于交流。

C.观察数据的数据库化（录入）

观察记录的录入：

- 文本数据、数字数据的录入。采用word或pages录入。
- 图片数据的录入。可以采用类似于Adobe的Lightroom之类的数据库。可以先扫描，再录入记录信息。
- 视频数据的录入，则可以运用类似于Adobe Premier之类的编辑库。
- 音频数据的录入，也可以寻找适用的音频数据库。

C.观察数据的数据库化（清单）

一份完整的数据库化的**观察数据**的数据库， 至少要提供以下的数据文件：

1. 观察提纲或者观察设计；
2. 观察记录的整理、清理数据库；
3. 观察内容数据库；
4. 观察记录数据数字化、数据库化过程的数据；
5. 清理报告。

D.文献数据的数据库化

文献数据一般情况下原本就来源于数据库。因此，运用原来数据库的数据，是文献数据库的特点。文献数据的数据库化包括三个步骤：

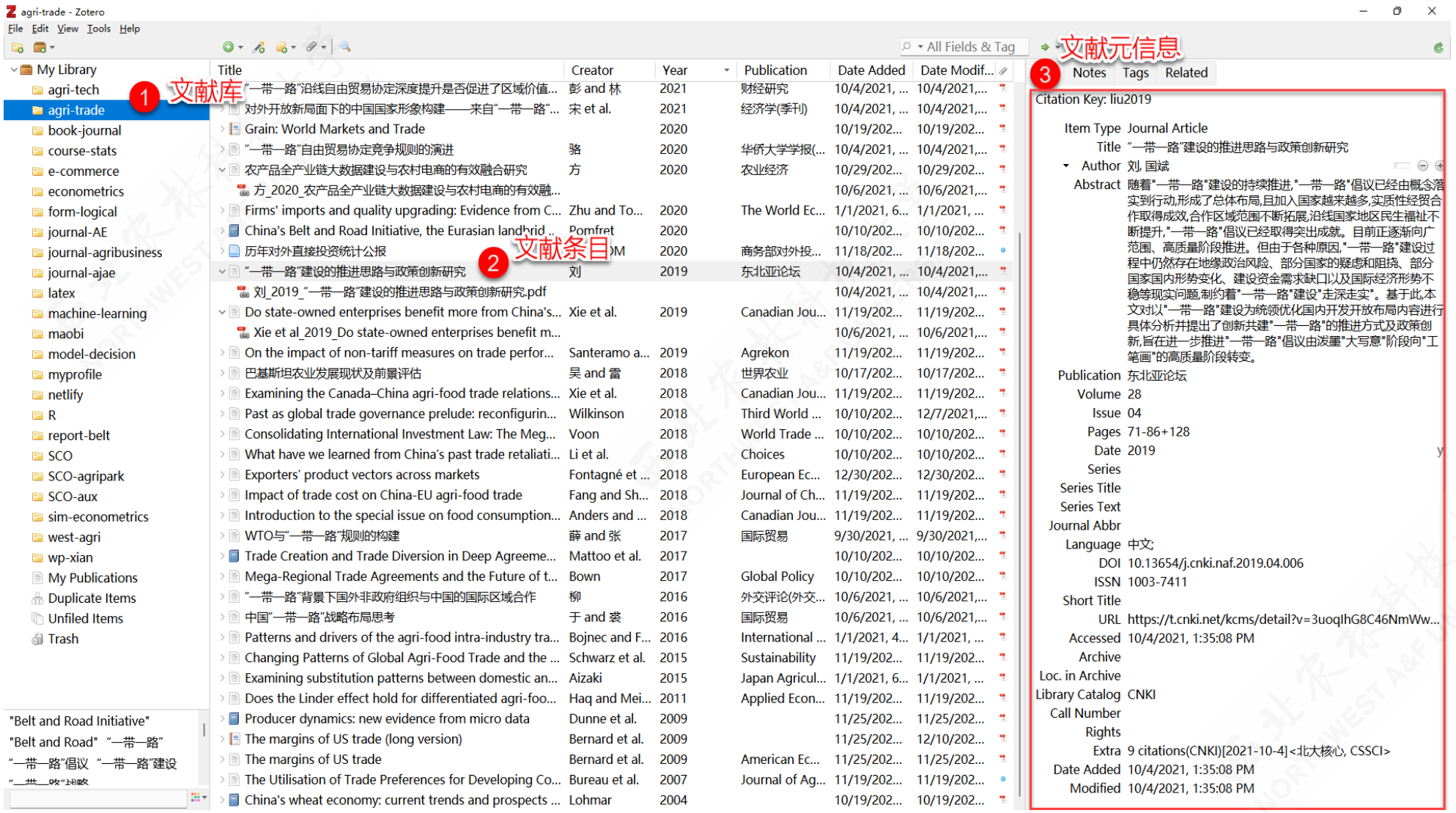
1. **编码**。指的是**文献信息**的编码，而不是**文献内容**的编码，文献信息就是**编目信息**，文献内容就是文献记载的内容。
2. **录入**。就是把原来数据库的文献编目信息和文献内容抄录到研究用的文献数据库中去。
3. **清理**。就是在数据录入完成以后，对录入的数据进行核查、清理，包括完整性检查。

D.文献数据的数据库化

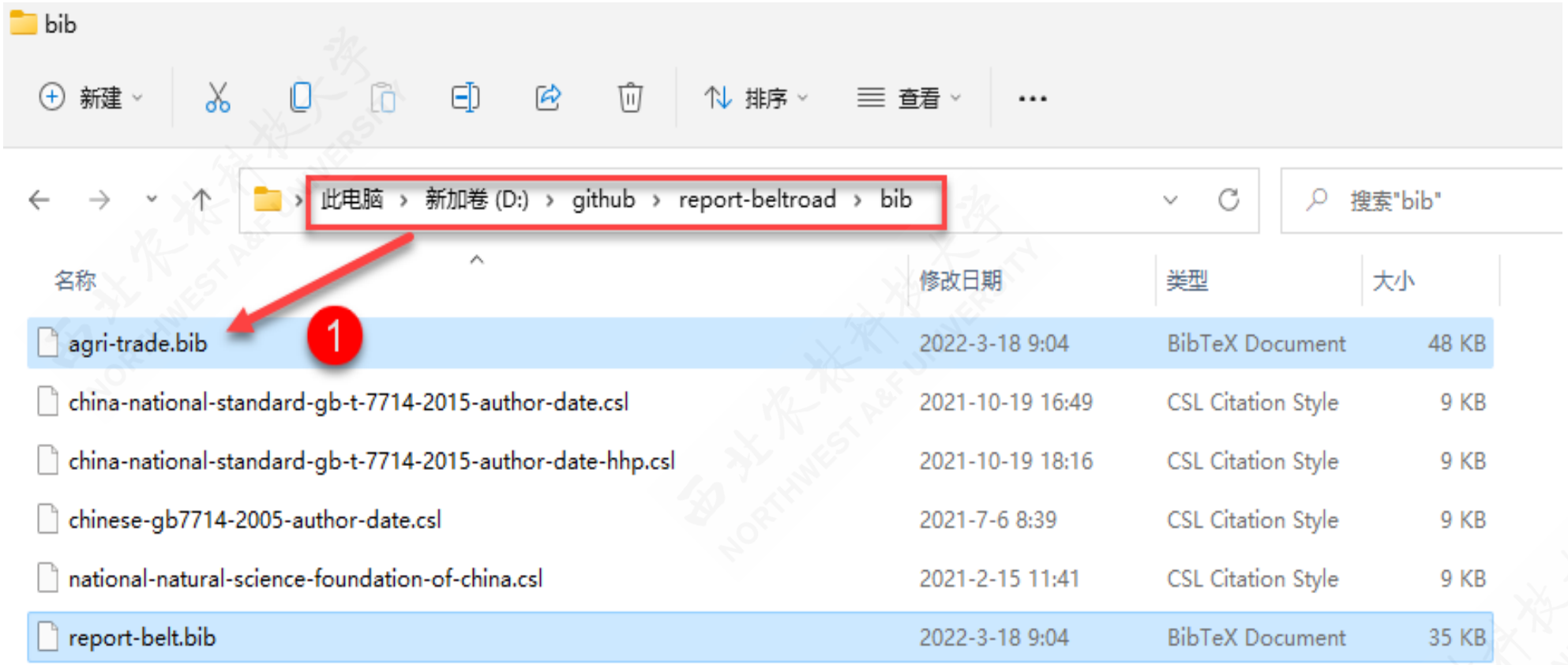
为了确保同学们已经掌握了文献编目信息，我重复一遍文献的编码。

- **文献记录信息**的录入和管理。
  - 基本变量主要有作者、篇名、时间、载体、存放、DOI，或者ISBO，或者ISNN等。
  - 文献记录的编码可以直接运用文献记录的原始编码，一些数据库化的数据，比如jasdo，还支持编码的数据直接导出。
  - 专门的信息录入和文献管理软件：Zotero、Endnote和papers。
- **文献内容信息**的录入和管理。
  - 主要管理的是文献内容、阅读笔记、思路图谱、总结要点等
  - 专门的内容录入和关系管理软件：onenote、Mindmanager、印象笔记等。

示例：Zotero的文献管理和使用I



示例：Zotero的文献管理和使用2



示例：Zotero的文献管理和使用3

agri-trade.bib

D:\github\report-beltroad\bib\agri-trade.bib

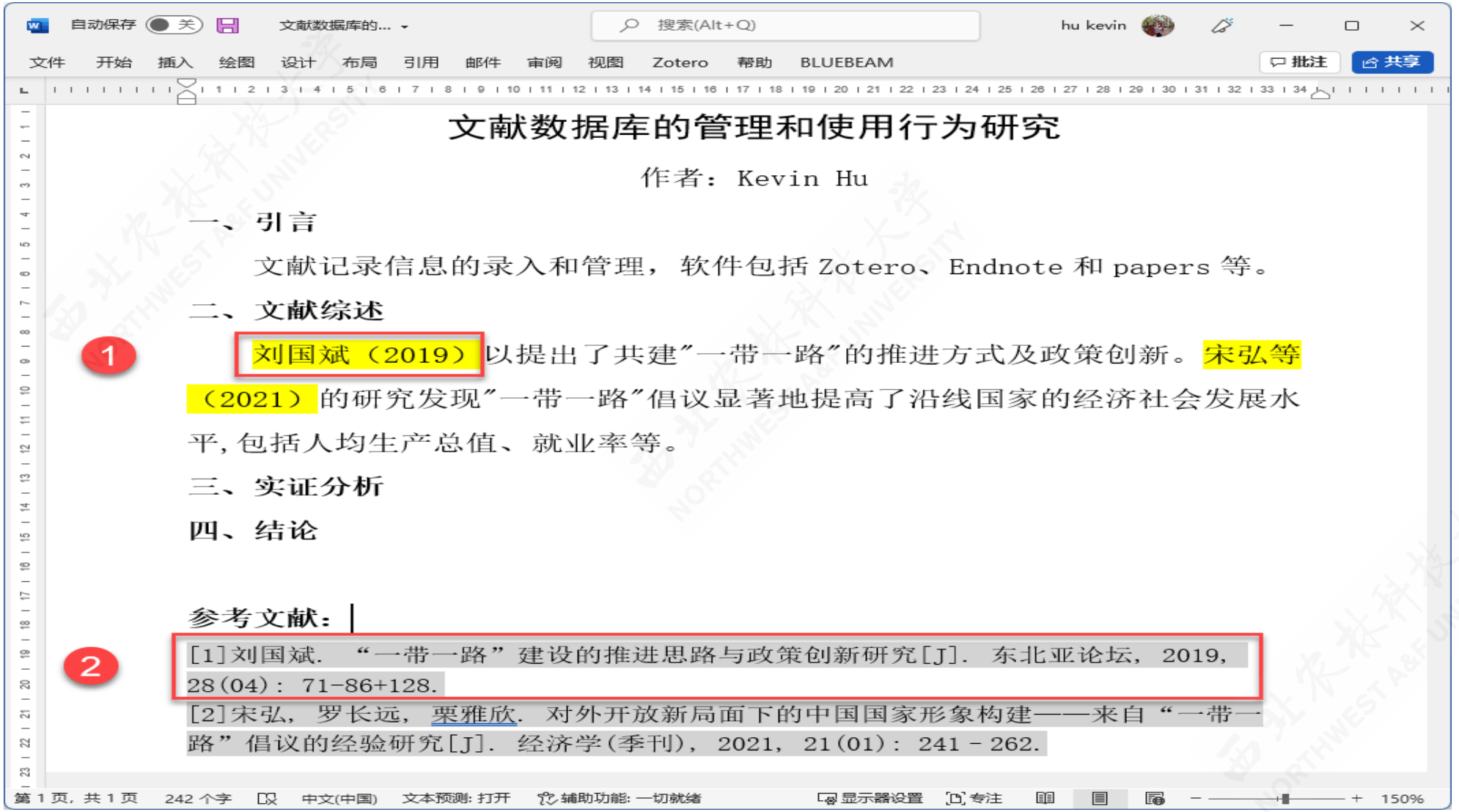
```
\\zotero-local\\AUTH\\柳\\柳_2016_"一带一路"背景下国外非政府组织与中国的国际区域合作.pdf}
}

@article{liu2019,
  title = {{``一带一路''建设的推进思路与政策创新研究}},
  author = {刘, 国斌},
  year = {2019},
  journal = {东北亚论坛},
  volume = {28},
  number = {04},
  pages = {71-86+128},
  issn = {1003-7411},
  doi = {10.13654/j.cnki.naf.2019.04.006},
  abstract = {随着"一带一路"建设的持续推进,"一带一路"倡议已经由概念落实到行动,形成了总体布局,且加入国家越来越多,实质性经贸合作取得
  langid = {chinese},
  keywords = {"Belt and Road Initiative","一带一路"建设,International Regional Cooperation,Policy Innovation,Secondary Core Area Cc
  annotation = {9 citations(CNKI)[2021-10-4]{$<$}北大核心, CSSCI{$>$}$}},
  file = {D\:\\google\\zotero-local\\AUTH\\刘\\刘_2019_"一带一路"建设的推进思路与政策创新研究.pdf}
}

@book{lohmar2004,
  title = {China's Wheat Economy: Current Trends and Prospects for Imports},
  shorttitle = {China's Wheat Economy},
  author = {Lohmar, Bryan Thomas},
  year = {2004},
  publisher = {{USDA, Economic Research Service}},
  file = {D\:\\google\\zotero-local\\AUTH\\Lohmar\\Lohmar_2004_China's wheat economy.pdf}
}

@article{luo2020,
  title = {{``一带一路''自由贸易协定竞争规则的演进}},
  author = {骆, 旭旭},
  year = {2020},
```

示例：Zotero的文献管理和使用4



6.痕迹数据的数据库化（简要）

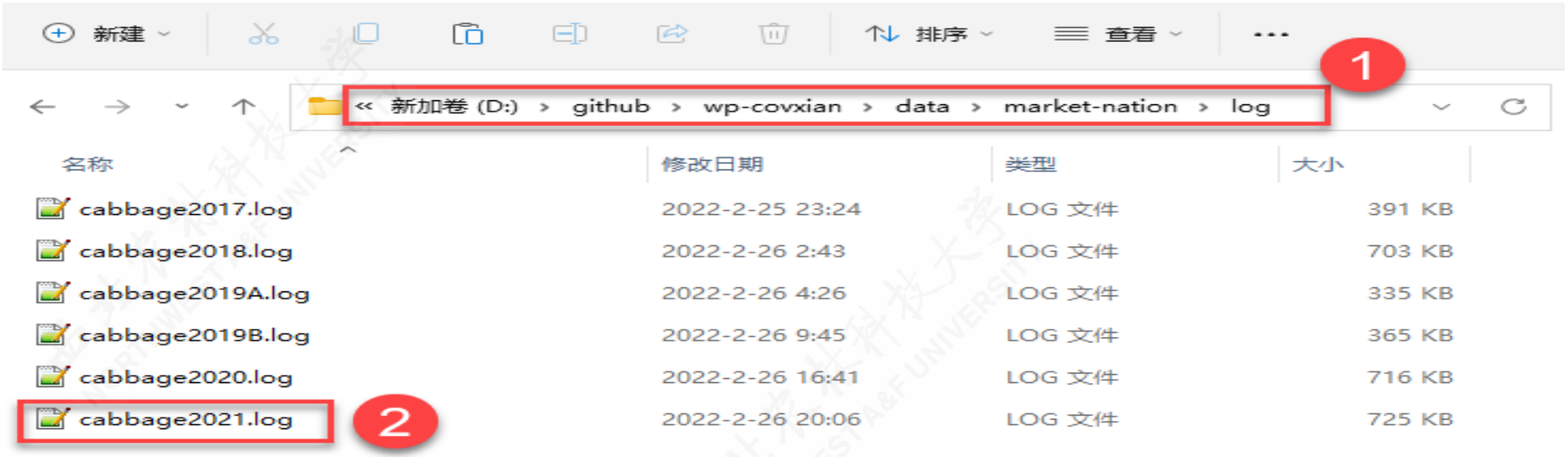
痕迹数据的数据库化，无论是Map-Reduce的产出，还是网页爬取的数据的整理、清理时的产出，都是**基于变量**的数据，还没有把变量数据串起来，变成**基于样本**的数据。

样本在变量上的变异是分析工作的基础，数据库化需要做的工作就是把变量数据串起来，变成类似于样本数据的数据。串起来的方法很多，技术性也很强，基本上依靠**脚本**来完成。

如果从大数据中抽取数据，由于无需数据录入，故数据库化只有两个步骤可做：

1. **编码**。通常原有的数据就已经有编码了，这个手续要做的就是要么确认使用原来的编码，要么呢，因为特殊的原因，需要重新编码，何去何从，完全取决于计算的需要。
2. **清理**。与其他调查数据的清理不同，这里主要是在确认编码以后，确认数据的可计算性，也就是格式化、结构化在转化中没有发生问题，以及是否可以直接运用于分布式并行计算或者单机计算。

示例：数据爬虫下的自动化日志记录I



示例：数据爬虫下的自动化日志记录

cabbage2021.log

```
== D:\github\wp-covxian\data\market-nation\log\cabbage2021.log ==========
Log Path: data/market-nation/log/cabbage2021.log
Program Path: D:/github/wp-covxian/scrape-nation.Rmd
Working Directory: D:/github/wp-covxian
User Name: huhua
R Version: 4.1.2 (2021-11-01)
Machine: BOOK2 x86
Operating System: Windows 10 x64 build 22000
Base Packages: stats graphics grDevices datasets utils methods base
Other Packages: RSQLite_2.2.10 dbplyr_2.1.1 DBI_1.1.2 tidyselect_1.1.2 rvest_1.0.2
                xml2_1.3.3 RSelenium_1.7.7 logr_1.2.8 htmlwidgets_1.5.4 webshot_0.5.2
                xmerit_0.0.11 rvg_0.2.5 officedown_0.2.3 officer_0.4.1 renv_0.15.2
                here_1.0.1 foreign_0.8-81 glue_1.6.1 jpeg_0.1-9 png_0.1-7
                wooldridge_1.4-2 DT_0.20 lubridate_1.8.0 ggrepel_0.9.1 ggthemes_4.2.4
                magrittr_2.0.2 gridExtra_2.3 scales_1.1.1 forcats_0.5.1 purrr_0.3.4
                readr_2.1.2 tibble_3.1.6 ggplot2_3.3.5 tidyverse_1.3.1 stringr_1.4.0
                dplyr_1.0.8 tidyr_1.2.0 openxlsx_4.2.5 knitr_1.37 bookdown_0.24
                rmarkdown_2.11
Log Start Time: 2022-02-26 16:43:35
------------------------------------------------------------------------------

2021年1月至2月。第541页/541页已完成（大白菜）file saved at time:2022-02-26 16:43:59

NOTE: Log Print Time:  2022-02-26 16:43:59
NOTE: Elapsed Time in seconds: 24.5376532077789

2021年1月至2月。第540页/541页已完成（大白菜）file saved at time:2022-02-26 16:44:02

NOTE: Log Print Time:  2022-02-26 16:44:02
NOTE: Elapsed Time in seconds: 2.52399897575378

2021年1月至2月。第539页/541页已完成（大白菜）file saved at time:2022-02-26 16:44:04
```

二手数据的数据库化（实例分享）

研究议题：旱区农业科技资源配置情况研究。具体**研究内容**如下：

- 2.1 科技装备
  - 2.1.1 农业机械动力
  - 2.1.2 农用拖拉机
  - 2.1.3 农用灌溉机械
  - 2.1.4 农用收获机械
  - 2.1.5 农业化学要素
- 2.2 科技投入
  - 2.2.1 公共财政投入
  - 2.2.2 RD研发投入
- 2.3 科技计划
  - 2.3.1 重大基础类科技计划
  - 2.3.2 国家自然科学基金
  - 2.3.3 农业综合开发投入
- 2.4 科技条件
  - 2.4.1 国家工程技术研究中心
  - 2.4.2 国家重点实验室
- 2.5 科技服务
  - 2.5.1 国家农业科技园区
  - 2.5.2 技术示范转移机构
  - 2.5.3 高技术产业和科技企业

资料和数据

研究对象：旱区16个省份——北京、天津、河北、山西、内蒙古、辽宁、吉林、黑龙江、山东、河南、西藏、陕西、甘肃、青海、宁夏、新疆

文本资料：政府公开资料、公共信息、图书、文献...

数据资料：统计年鉴、网页数据、商业数据库信息...

资料整理

文件夹管理：

1 **文献资料**文件（material）：收集到的各种相关资料（.xlsx、.word、.pdf、.html、.png等）

2 **粗制**的数据文件（raw data）：摘录、数值化（.xlsx）

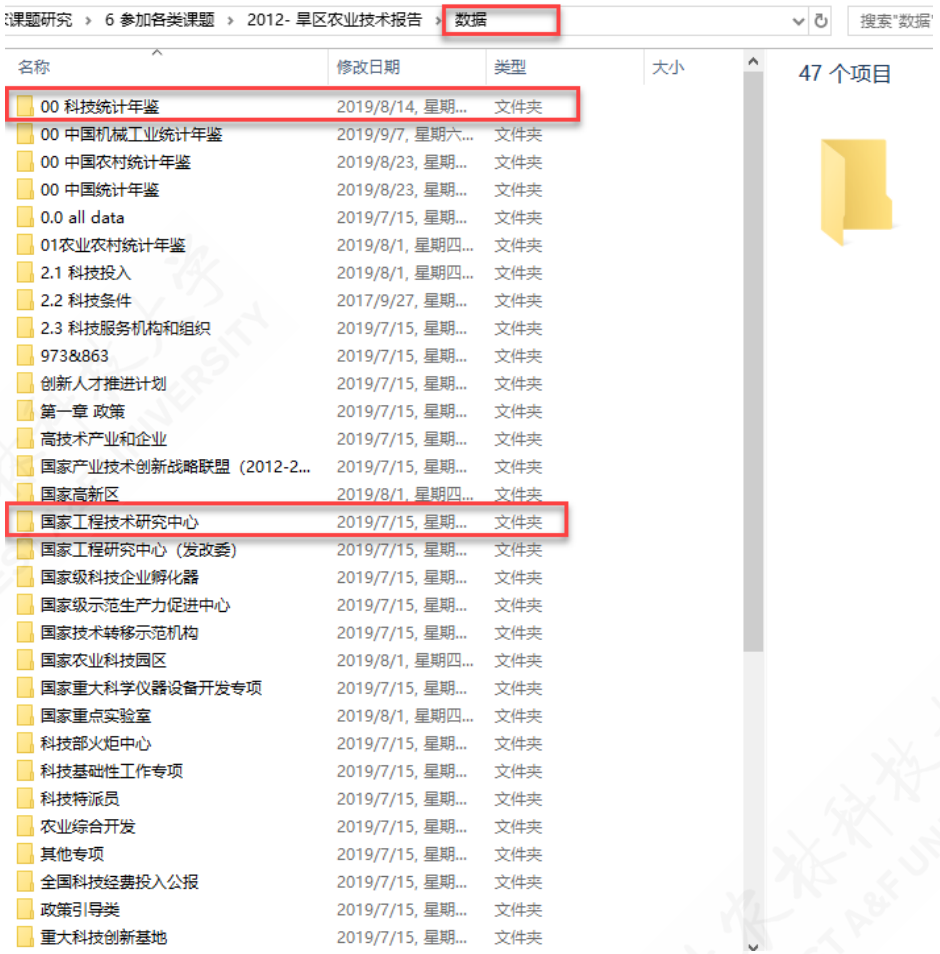
3 **提取**的数据文件（extract data）：整合、合并（.xlsx）

4 **加工**的数据文件（process data）：更新、维护（.xlsx）

5 **分析**的数据文件（analysis data）：调用、子集化（.xlsx）

0.文献资料1

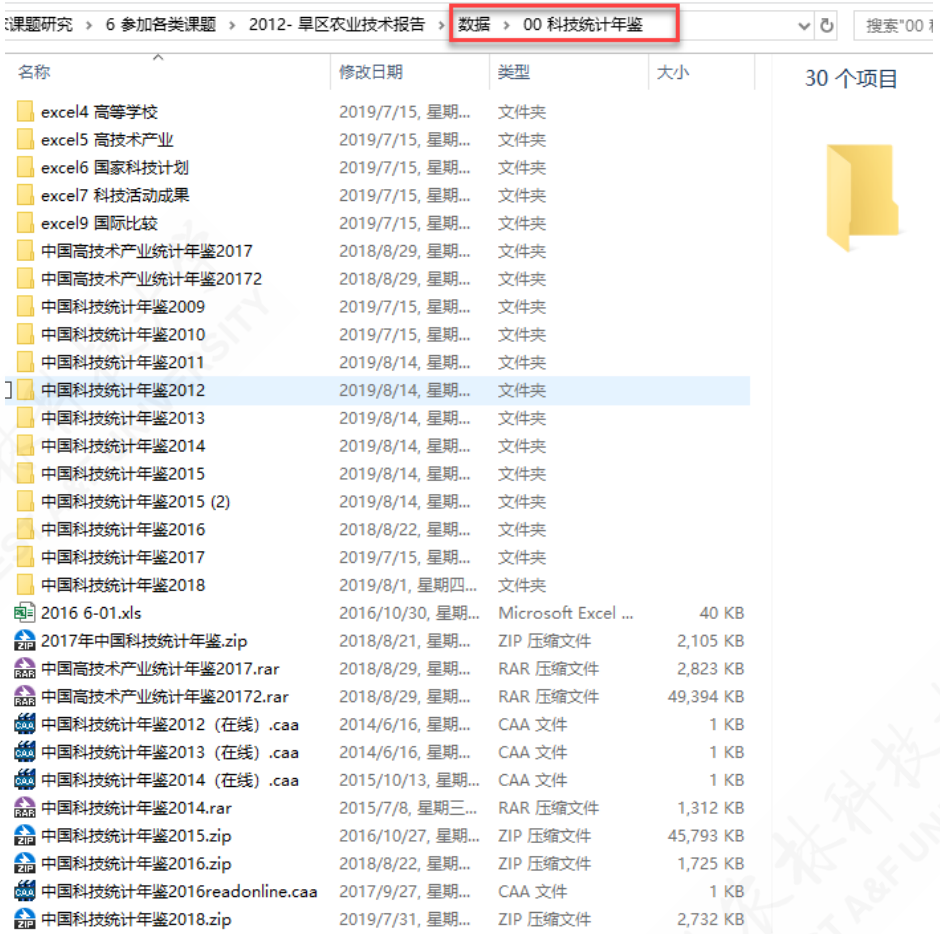
- 囊括了研究涉及的全部材料
- 分门别类在各个文件夹下
- 形成**目录树**
- 文件以原始状态存放
- 格式各种各样



文献资料文件夹

0. 文献资料1-1

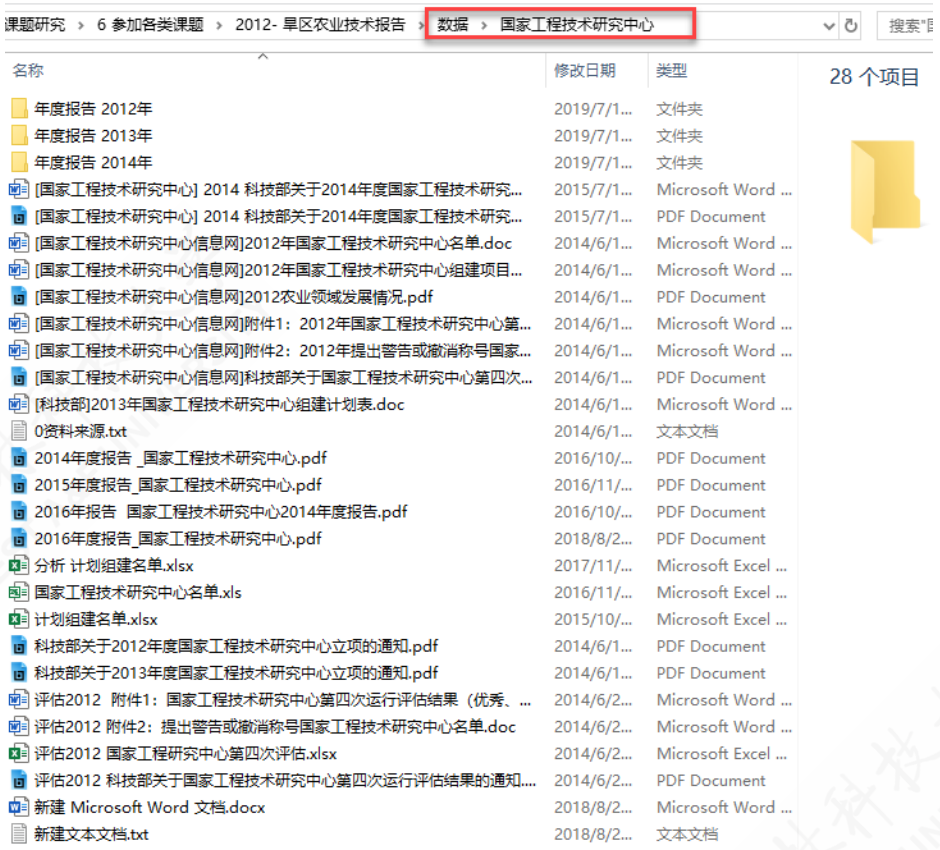
- 历年的《中国科技统计年鉴》
- 数据来源：人大经济论坛；中国知网-统计年鉴数据库
- 部分年鉴**数值化**（.xls）
- 部分年鉴仅是**数字化**（.caj）
- 每本年鉴都有**目录**
- 年鉴中仅部分内容跟研究相关



子文件夹

0.文献资料1-2

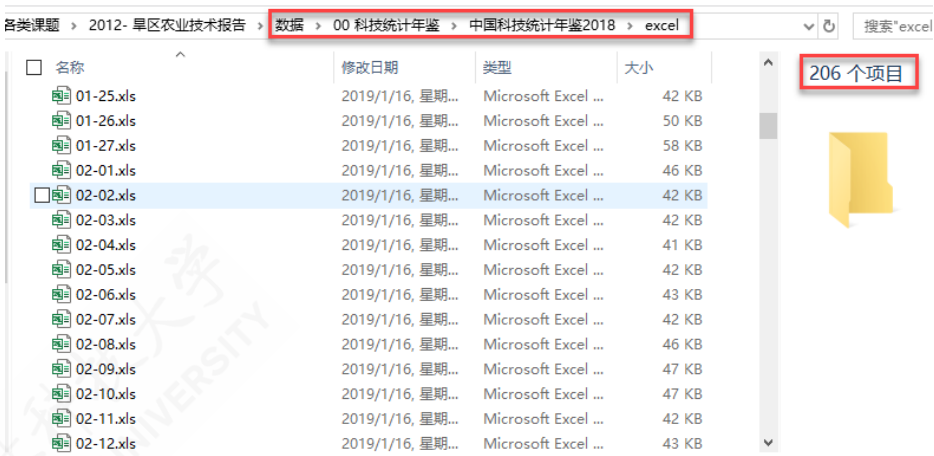
- 历年《国家工程技术研究中心》资料
- 数据来源：科技部网站
- 部分资料以**年度报告**呈现（.pdf）
- 部分资料以**公开网页**呈现（.html、.doc）
- 资料发布时间不确定
- 资料非标准化，需手工收集整理



文献《国家工程技术研究中心》

0.文献资料1-1-1

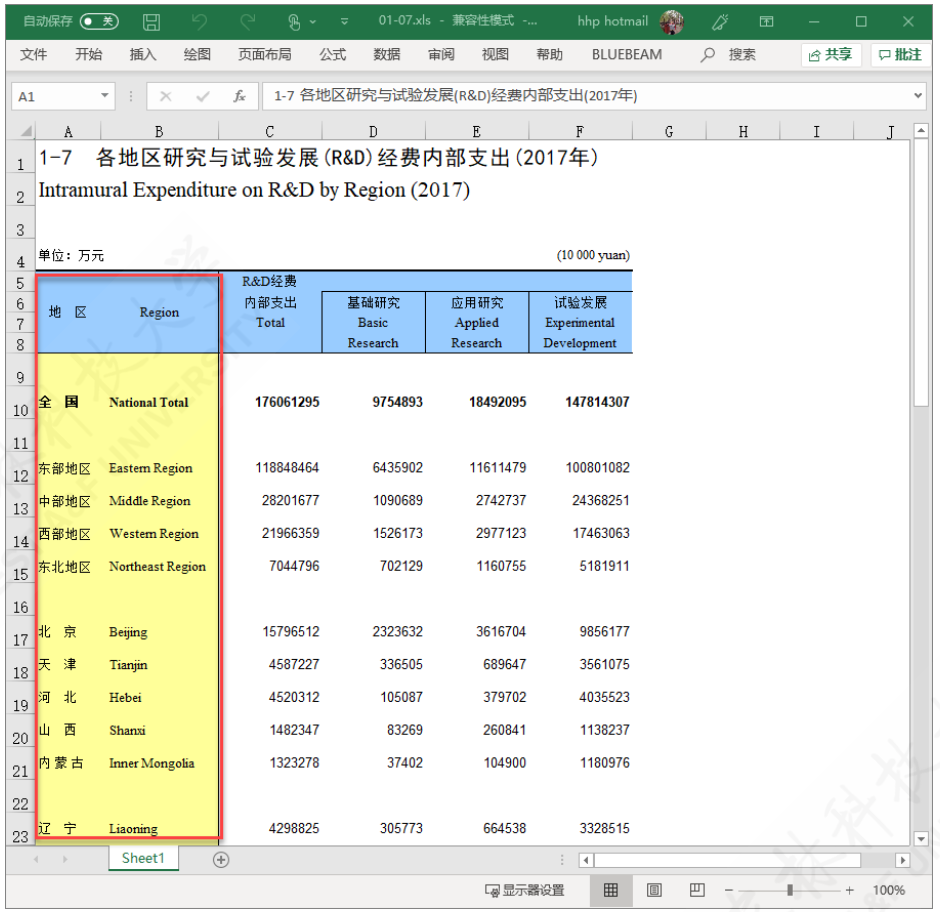
- 《中国科技统计年鉴2018》
- 数据来源：人大经济论坛
- 该年鉴已**数值化**（.xls）
- 年鉴统计资料依次以.xls格式呈现
- 具体文件含义可以查看**目录**
- 年鉴中仅部分.xls跟研究相关，需要提取出来



文献《中国科技统计年鉴》

0.文献资料1-1-1-1

- 《中国科技统计年鉴2018》
- 数据来源：人大经济论坛
- **“表1-7 2017年中国RD支出类型”**
- 原始表格有**各种“烦人状况”**!
  - 看**行**：空行？字符有空格？意外字符？
  - 看**列**：列变量？中英文？跨多行？
  - 看**单元格**：数值（number）还是文字（character）？



1-7 各地区研究与试验发展(R&D)经费内部支出(2017年)

Intramural Expenditure on R&D by Region (2017)

单位：万元 (10 000 yuan)

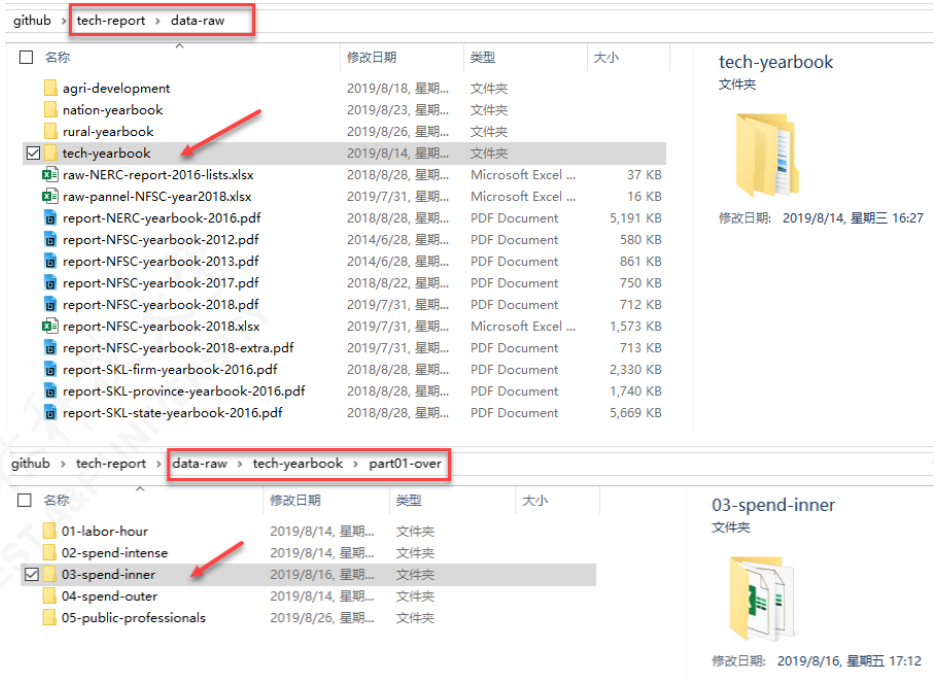
地区 Region	R&D经费内部支出 Total	基础研究 Basic Research	应用研究 Applied Research	试验发展 Experimental Development
全国 National Total	**176061295**	**9754893**	**18492095**	**147814307**
东部地区 Eastern Region	118848464	6435902	11611479	100801082
中部地区 Middle Region	28201677	1090689	2742737	24368251
西部地区 Western Region	21966359	1526173	2977123	17463063
东北地区 Northeast Region	7044796	702129	1160755	5181911
北京 Beijing	15796512	2323632	3616704	9856177
天津 Tianjin	4587227	336505	689647	3561075
河北 Hebei	4520312	105087	379702	4035523
山西 Shanxi	1482347	83269	260841	1138237
内蒙古 Inner Mongolia	1323278	37402	104900	1180976
辽宁 Liaoning	4298825	305773	664538	3328515

文献-中国科技统计年鉴2018

-表1-7 RD支出类型

A. 粗制数据（raw data）!

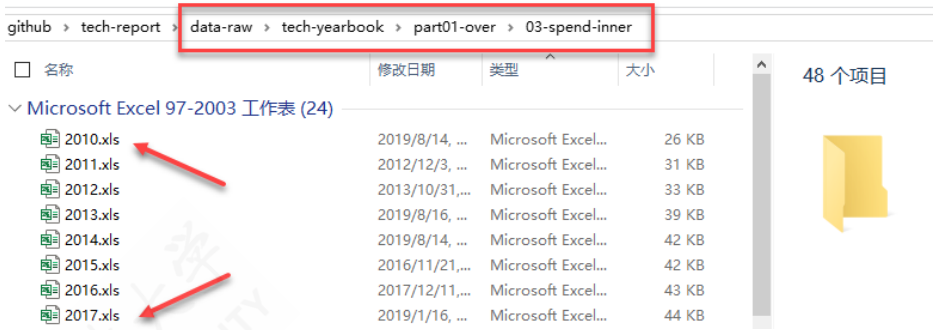
- 《中国科技统计年鉴2010-2018》
- 数据来源：人大经济论坛
- 各年年鉴**整合**
- 不按**年份**，而按**内容**来管理文件夹
- 文件夹命名坚持用**英文**！



重新整理后的科技统计年鉴文件夹

A. 粗制数据（raw data）2

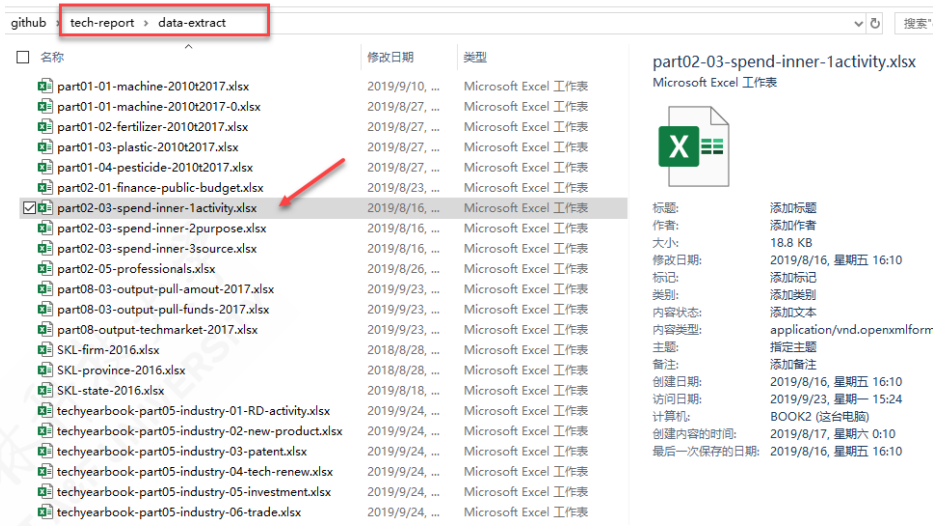
- 《中国科技统计年鉴2010-2018》
- 数据来源：人大经济论坛
- **表 中国RD支出类型”**（.xls）
- 取你所需！
  - 每年的表格来自每年的**年鉴**
  - 每年的表格**单独命名**
  - 文件命名要有**规则**
  - 确保每个文件的行列数据**保持一致**！



历年的RD支出类型（2010-2017）

B. 精制数据（extract data）I

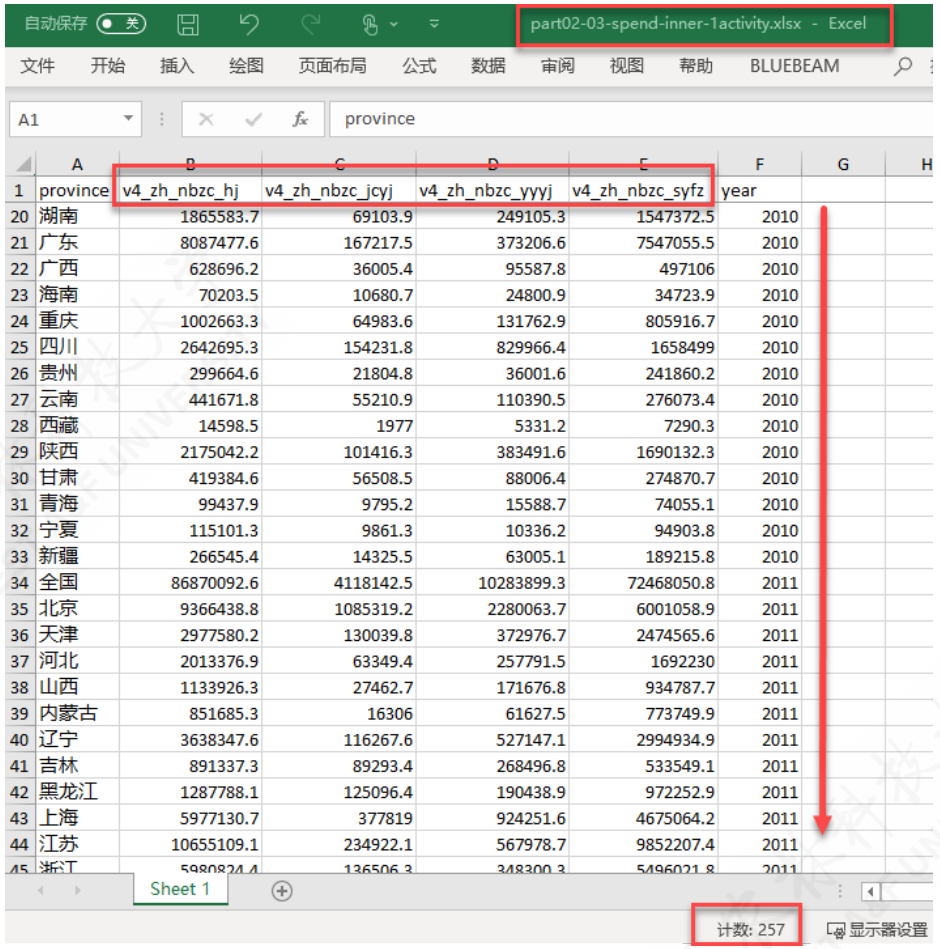
- 《中国科技统计年鉴2010-2018》
- **表 RD支出类型（2010-2017）**
- 数据来源：人大经济论坛
- 依次读取整合每一年的**表 中国RD支出类型.xls”**
  - 统一**变量命名**
  - 分别写入**年份**信息
  - **行合并**年度文件数据
  - 确保数据是**正确读取**的！



提取整合后的RD支出类型（2010-2017）

B. 精制数据（extract data）2

- 《中国科技统计年鉴2010-2018》
- **表 RD支出类型（2010-2017）**
- 数据来源：人大经济论坛
- 基本保持原来的数据形态：
  - 看**行**(257行)：无空行、地区字符正确标准
  - 看**列**：列变量统一命名
  - 看**单元格**：全部是数值（number）

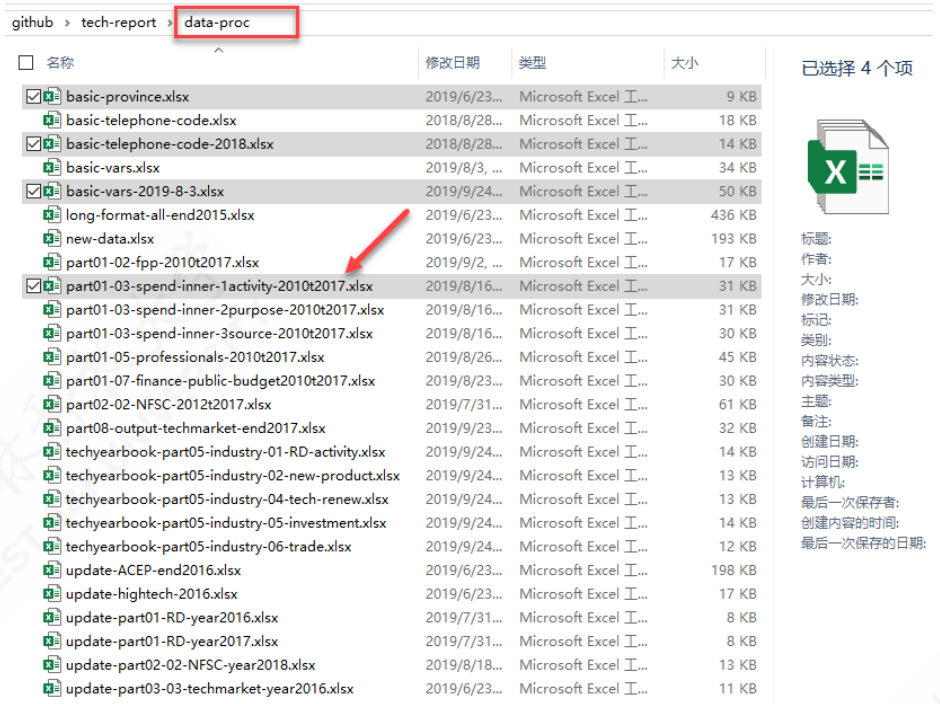


	A	B	C	D	E	F
1	province	v4_zh_nbzc_hj	v4_zh_nbzc_jcyj	v4_zh_nbzc_yyyj	v4_zh_nbzc_syfz	year
20	湖南	1865583.7	69103.9	249105.3	1547372.5	2010
21	广东	8087477.6	167217.5	373206.6	7547055.5	2010
22	广西	628696.2	36005.4	95587.8	497106	2010
23	海南	70203.5	10680.7	24800.9	34723.9	2010
24	重庆	1002663.3	64983.6	131762.9	805916.7	2010
25	四川	2642695.3	154231.8	829966.4	1658499	2010
26	贵州	299664.6	21804.8	36001.6	241860.2	2010
27	云南	441671.8	55210.9	110390.5	276073.4	2010
28	西藏	14598.5	1977	5331.2	7290.3	2010
29	陕西	2175042.2	101416.3	383491.6	1690132.3	2010
30	甘肃	419384.6	56508.5	88006.4	274870.7	2010
31	青海	99437.9	9795.2	15588.7	74055.1	2010
32	宁夏	115101.3	9861.3	10336.2	94903.8	2010
33	新疆	266545.4	14325.5	63005.1	189215.8	2010
34	全国	86870092.6	4118142.5	10283899.3	72468050.8	2011
35	北京	9366438.8	1085319.2	2280063.7	6001058.9	2011
36	天津	2977580.2	130039.8	372976.7	2474565.6	2011
37	河北	2013376.9	63349.4	257791.5	1692230	2011
38	山西	1133926.3	27462.7	171676.8	934787.7	2011
39	内蒙古	851685.3	16306	61627.5	773749.9	2011
40	辽宁	3638347.6	116267.6	527147.1	2994934.9	2011
41	吉林	891337.3	89293.4	268496.8	533549.1	2011
42	黑龙江	1287788.1	125096.4	190438.9	972252.9	2011
43	上海	5977130.7	377819	924251.6	4675064.2	2011
44	江苏	10655109.1	234922.1	567978.7	9852207.4	2011

提取整合后的RD支出类型（2010-2017）

C. 加工数据（process data）I

- 《中国科技统计年鉴2010-2018》
- **表 RD支出类型（2010-2017）**
- 数据来源：人大经济论坛
- 需要继续对数据形态加工变形
- 目标是**标准化**的数据集！！？

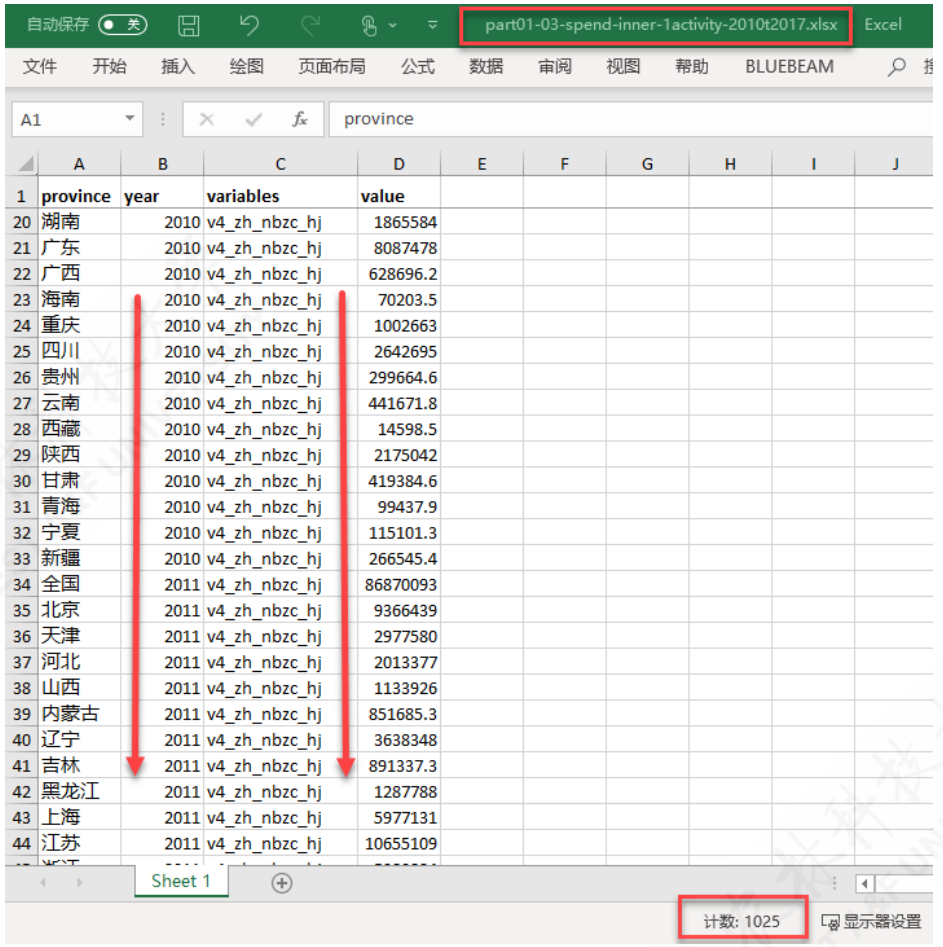


加工变形后的RD支出类型

（2010-2017）

C. 加工数据（process data）2

- 《中国科技统计年鉴2010-2018》
- **表 RD支出类型（2010-2017）**
- 数据来源：人大经济论坛
- 这是一份**标准化**的数据集！！！
  - 看**行**(1025行)：按年度(year)、按省份(province)
  - 看**列**：4个**变量**被折叠对方为1列(variables)！
  - 看**单元格**：全部**数值**被折叠对方为1列(value)！



part01-03-spend-inner-1activity-2010t2017.xlsx

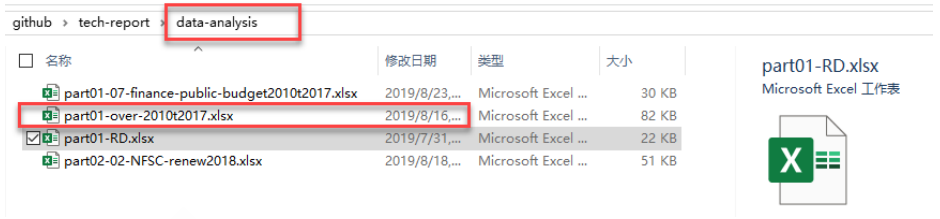
	province	year	variables	value
20	湖南	2010	v4_zh_nbzc_hj	1865584
21	广东	2010	v4_zh_nbzc_hj	8087478
22	广西	2010	v4_zh_nbzc_hj	628696.2
23	海南	2010	v4_zh_nbzc_hj	70203.5
24	重庆	2010	v4_zh_nbzc_hj	1002663
25	四川	2010	v4_zh_nbzc_hj	2642695
26	贵州	2010	v4_zh_nbzc_hj	299664.6
27	云南	2010	v4_zh_nbzc_hj	441671.8
28	西藏	2010	v4_zh_nbzc_hj	14598.5
29	陕西	2010	v4_zh_nbzc_hj	2175042
30	甘肃	2010	v4_zh_nbzc_hj	419384.6
31	青海	2010	v4_zh_nbzc_hj	99437.9
32	宁夏	2010	v4_zh_nbzc_hj	115101.3
33	新疆	2010	v4_zh_nbzc_hj	266545.4
34	全国	2011	v4_zh_nbzc_hj	86870093
35	北京	2011	v4_zh_nbzc_hj	9366439
36	天津	2011	v4_zh_nbzc_hj	2977580
37	河北	2011	v4_zh_nbzc_hj	2013377
38	山西	2011	v4_zh_nbzc_hj	1133926
39	内蒙古	2011	v4_zh_nbzc_hj	851685.3
40	辽宁	2011	v4_zh_nbzc_hj	3638348
41	吉林	2011	v4_zh_nbzc_hj	891337.3
42	黑龙江	2011	v4_zh_nbzc_hj	1287788
43	上海	2011	v4_zh_nbzc_hj	5977131
44	江苏	2011	v4_zh_nbzc_hj	10655109

Sheet 1　计数: 1025

加工变形后的RD支出类型（2010-2017）

D.分析数据（analysis data）I

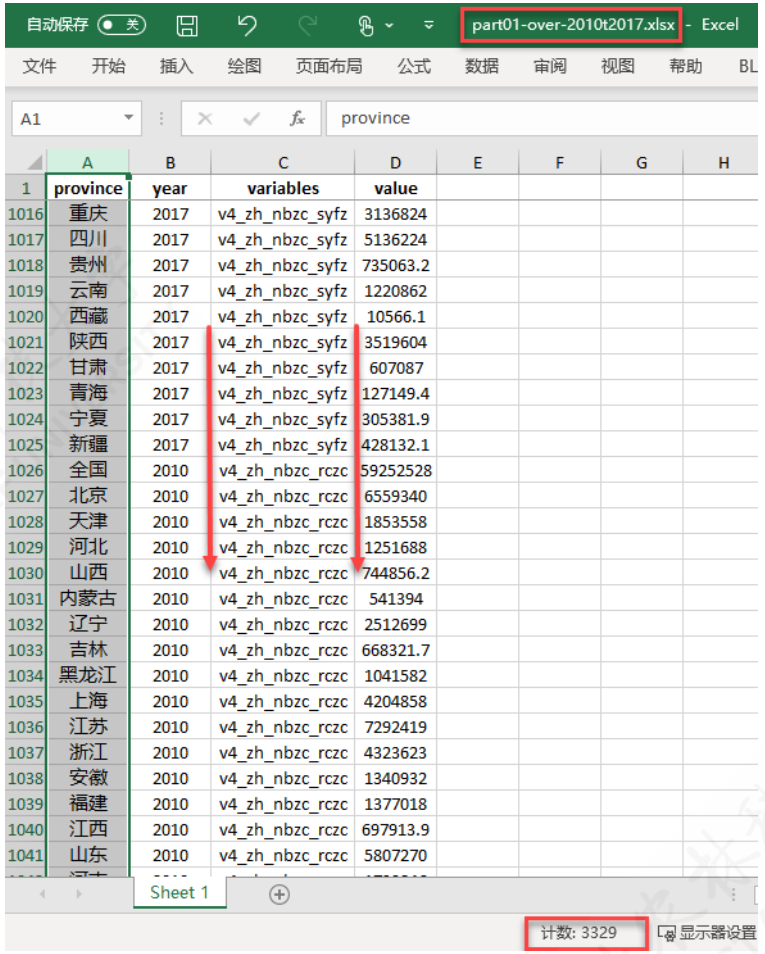
- 《中国科技统计年鉴2010-2018》
- **完整的RD数据集**(part01-over-2010t2017.xlsx)
- 数据来源：人大经济论坛
- 每一个数据子集被加工完成后，需要继续进行整合
- 目标是一个**标准化**的完整数据集！！？



聚合各个子数据集为一个完整RD数据集（2010-2017）

D. 分析数据（analysis data）2

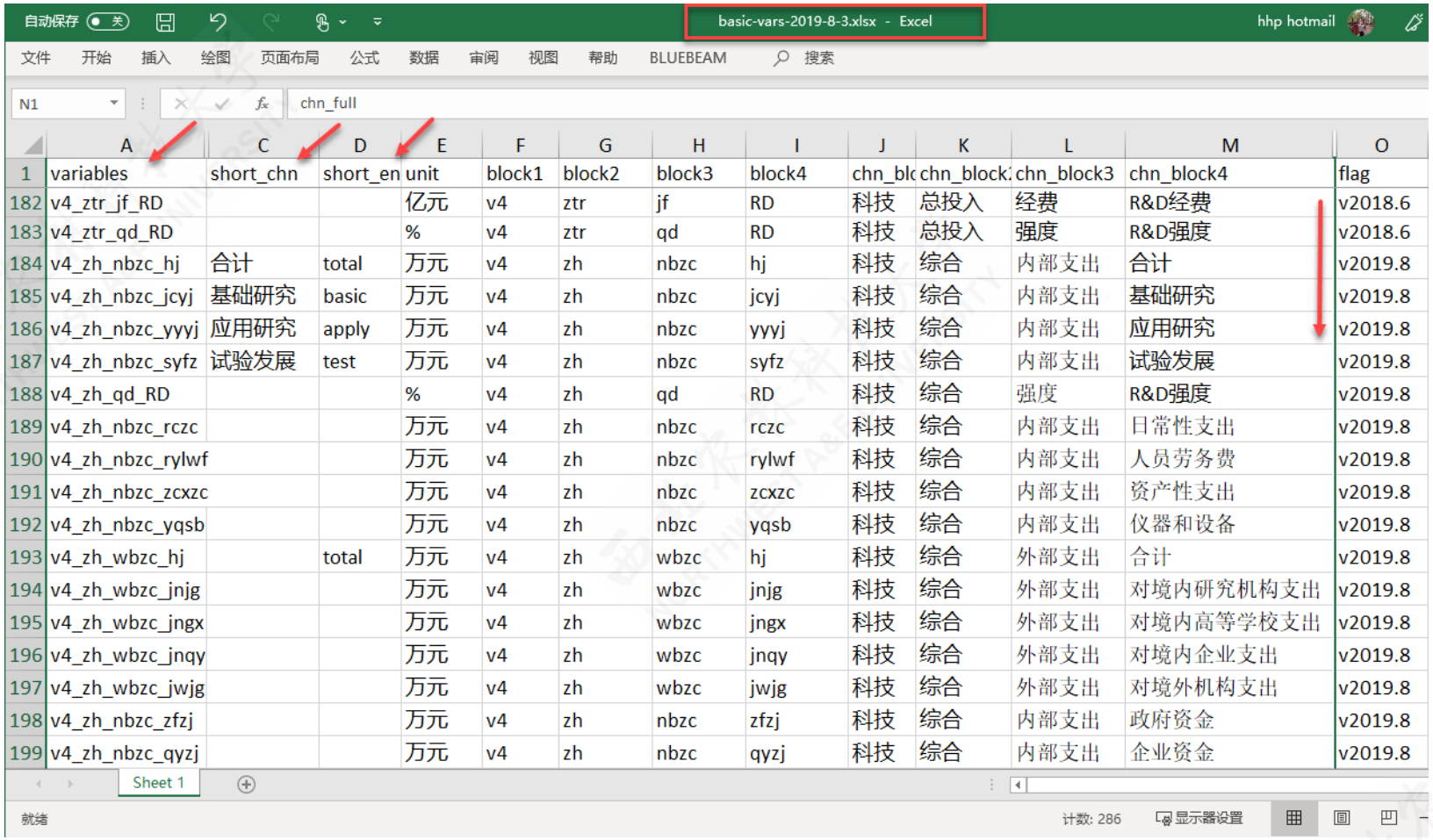
- 《中国科技统计年鉴2010-2018》
- **完整的RD数据集**(part01-over-2010t2017.xlsx)
- 数据来源：人大经济论坛
- 这是一份**完整的**、**标准化**的数据集！！！
  - 看**行**(3329行)：按年度(year)、按省份(province)
  - 看**列**：全部**变量**被折叠对方为1列(variables)！
  - 看**单元格**：全部**数值**被折叠对方为1列(value)！



	A	B	C	D
1	province	year	variables	value
1016	重庆	2017	v4_zh_nbzc_syfz	3136824
1017	四川	2017	v4_zh_nbzc_syfz	5136224
1018	贵州	2017	v4_zh_nbzc_syfz	735063.2
1019	云南	2017	v4_zh_nbzc_syfz	1220862
1020	西藏	2017	v4_zh_nbzc_syfz	10566.1
1021	陕西	2017	v4_zh_nbzc_syfz	3519604
1022	甘肃	2017	v4_zh_nbzc_syfz	607087
1023	青海	2017	v4_zh_nbzc_syfz	127149.4
1024	宁夏	2017	v4_zh_nbzc_syfz	305381.9
1025	新疆	2017	v4_zh_nbzc_syfz	428132.1
1026	全国	2010	v4_zh_nbzc_rczc	59252528
1027	北京	2010	v4_zh_nbzc_rczc	6559340
1028	天津	2010	v4_zh_nbzc_rczc	1853558
1029	河北	2010	v4_zh_nbzc_rczc	1251688
1030	山西	2010	v4_zh_nbzc_rczc	744856.2
1031	内蒙古	2010	v4_zh_nbzc_rczc	541394
1032	辽宁	2010	v4_zh_nbzc_rczc	2512699
1033	吉林	2010	v4_zh_nbzc_rczc	668321.7
1034	黑龙江	2010	v4_zh_nbzc_rczc	1041582
1035	上海	2010	v4_zh_nbzc_rczc	4204858
1036	江苏	2010	v4_zh_nbzc_rczc	7292419
1037	浙江	2010	v4_zh_nbzc_rczc	4323623
1038	安徽	2010	v4_zh_nbzc_rczc	1340932
1039	福建	2010	v4_zh_nbzc_rczc	1377018
1040	江西	2010	v4_zh_nbzc_rczc	697913.9
1041	山东	2010	v4_zh_nbzc_rczc	5807270

聚合各个子数据集为一个完整RD数据集（2010-2017）

数据和变量关联与管理



basic-vars-2019-8-3.xlsx - Excel

N1 | chn_full

	A	C	D	E	F	G	H	I	J	K	L	M	O
1	variables	short_chn	short_en	unit	block1	block2	block3	block4	chn_blo	chn_block3	chn_block3	chn_block4	flag
182	v4_ztr_jf_RD			亿元	v4	ztr	jf	RD	科技	总投入	经费	R&D经费	v2018.6
183	v4_ztr_qd_RD			%	v4	ztr	qd	RD	科技	总投入	强度	R&D强度	v2018.6
184	v4_zh_nbzc_hj	合计	total	万元	v4	zh	nbzc	hj	科技	综合	内部支出	合计	v2019.8
185	v4_zh_nbzc_jcyj	基础研究	basic	万元	v4	zh	nbzc	jcyj	科技	综合	内部支出	基础研究	v2019.8
186	v4_zh_nbzc_yyyj	应用研究	apply	万元	v4	zh	nbzc	yyyj	科技	综合	内部支出	应用研究	v2019.8
187	v4_zh_nbzc_syfz	试验发展	test	万元	v4	zh	nbzc	syfz	科技	综合	内部支出	试验发展	v2019.8
188	v4_zh_qd_RD			%	v4	zh	qd	RD	科技	综合	强度	R&D强度	v2019.8
189	v4_zh_nbzc_rczc			万元	v4	zh	nbzc	rczc	科技	综合	内部支出	日常性支出	v2019.8
190	v4_zh_nbzc_rylwf			万元	v4	zh	nbzc	rylwf	科技	综合	内部支出	人员劳务费	v2019.8
191	v4_zh_nbzc_zcxzc			万元	v4	zh	nbzc	zcxzc	科技	综合	内部支出	资产性支出	v2019.8
192	v4_zh_nbzc_yqsb			万元	v4	zh	nbzc	yqsb	科技	综合	内部支出	仪器和设备	v2019.8
193	v4_zh_wbzc_hj		total	万元	v4	zh	wbzc	hj	科技	综合	外部支出	合计	v2019.8
194	v4_zh_wbzc_jnjg			万元	v4	zh	wbzc	jnjg	科技	综合	外部支出	对境内研究机构支出	v2019.8
195	v4_zh_wbzc_jngx			万元	v4	zh	wbzc	jngx	科技	综合	外部支出	对境内高等学校支出	v2019.8
196	v4_zh_wbzc_jnqy			万元	v4	zh	wbzc	jnqy	科技	综合	外部支出	对境内企业支出	v2019.8
197	v4_zh_wbzc_jwjg			万元	v4	zh	wbzc	jwjg	科技	综合	外部支出	对境外机构支出	v2019.8
198	v4_zh_nbzc_zfzj			万元	v4	zh	nbzc	zfzj	科技	综合	内部支出	政府资金	v2019.8
199	v4_zh_nbzc_qyzj			万元	v4	zh	nbzc	qyzj	科技	综合	内部支出	企业资金	v2019.8

Sheet 1 | 就绪 | 计数: 286

变量命名是一门学问！

数据和变量关联与管理

- 原始文件没有变量？
- 变量形式与其含义？
  - **唯一识别变量名**(variable)：v4_zh_nbzc_hj、v4_zh_nbzc_jcyj、v4_zh_nbzc_yyyj、v4_zh_nbzc_syfz
  - **中文变量名**(short_chn)：合计、基础研究、应用研究、试验发展
  - **英文变量名**(short_eng)：total、basic、apply、test
- 变量命名如何动态调整？
  - 备注变量系统的**版本号**(flag)：v2018.6、v2019.8、v2019.9

本节结束