项目结题验收单

专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	基于 VBA 的数据清洗程序的开发				
主持人	彭丽	职务/职称	馆员		
所在单位	（加盖单位公章）				
专家意见	课题组按照课题要求，认真开展相关研究和实践，开发的小程序，为我馆学科服务工作中需要的数据提供了有力的技术支持，很好地解决了我校文献的人员贡献度和二级单位贡献度归属问题，为我部门后期的学科服务工作提供了有力的数据保障，大大节省了以往人工统计的时间，提高了工作效率。该程序目前已经投入我校图书馆学科服务的日常使用中。 验收合格，同意结题。 （如需要可增加页数）				
专家签字					
职务/职称	研究馆员	副研究馆员	副研究馆员		

项目编号：2021052

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：基于 VBA 的数据清洗程序的开发

项目关键词：VBA；学科服务； 数据清洗；程序开发

项目单位（盖章）：

通信地址：四川省成都市温江区惠民路 211 号

项目主持人：彭丽

联系电话：028-86290938

电子邮件：23289447@qq.com

提交日期：2022 年 5 月 11 日

基于 VBA 的数据清洗程序的开发

——以四川农业大学为例

关键词：VBA；学科服务； 数据清洗；程序开发

1 研究背景、目的及意义

1.1 研究背景

随着信息处理技术的不断发展，各行各业已建立很多计算机信息系统，积累了大量的数据。为了使数据能够有效地支持日常运作和决策，要求数据可靠无误，能够准确地反映现实世界的状况。数据是信息的基础，在信息表达中起着至关重要的作用。然而人们常常抱怨“数据丰富，信息贫乏”，究其原因，一是缺乏有效的数据分析技术，二是数据质量不高，如数据输入错误、不同来源数据引起的不同表示方法、数据间的不一致等，导致现有的数据中存在这样或那样的脏数据。随着各行各业对数据精准表达信息需求的日益提高，对数据的清洗需求也日渐旺盛，各种数据清洗工具也应运而生，如特殊领域的清洗工具有 IDCENTRIC、PUREINTEGRATE、 QUICKADDRESS、 REUNION 、TRILLIUM 等；用于消除重复记录的清洗工具有 DATACLEANSER、 MERGE /PURGE LIBRARY、MATCHIT、ASTERMERGE 等；支持数据仓库的 ETL 处理，如 COPYMANAGER、DATASTAGE、EXTRACT、WERMART 等。高校图书馆也不例外，图书馆在科研数据管理服务上起着重要的作用。现今，学科服务是现代化图书馆一项不可或缺的创新服务，随着高校科研人员对科学数据管理服务需求越来越大，数据清洗必不可少。高校图书馆数据清洗研究主要集中在文献数据的清洗，已有的数据清洗研究更多的是进行数据清洗策略、方法、流程、框架模型、实施方案 等理论探索，较少涉及具体技术实现。

1.2 研究目的

目前国内高校图书馆的数据管理与服务中，数据清洗环节过度依赖现有系统平台中的数据清洗功能，图书馆或图书馆员参与度较低；而现有的图书馆集成管理系统、发现系统、机构知识库等系统平台中的数据清洗功能不够完善，不能满足深度数据挖掘分析及精准服务推送的应用需求。现有第三方数据清洗平台工具较少，且大多数文献数据分析工具也不具有数据清洗功能。一旦现有系统平台不能满足需求，数据清洗工作多以手工或半自动化方式实现，费时费力。目前我馆学科服务中计算我校研究人员个人贡献度和我校二级单位贡献度所需数据全靠人工逐条筛选，给学科服务工作带来很大困扰。本研究从我馆学科服务工作的实际需求出发，开发一个针对性强、简单实用的小程序，为了利用现有的计算机技能和熟悉的软件环境来提高工作效率，本课题拟在 EXCEL 软件环境中编写 VBA 小程序。

1.3 研究意义

作为一款桌面型数据处理软件，Excel 主要面向日常办公和中小型数据集的处理，但在面对海量数据的清洗任务时却是难以胜任的，即使是小型数据集在使用前也存在需要规范化的问题。通过在 Excel 中进行数据清洗的实践操作，有助于帮助读者理解数据清洗的概念和知识，并掌握一定的操作技巧，为后面进行大数据集的清洗打好基础。

本研究开发的小程序运行于 EXCEL 软件环境中，将数据呈现和后台程序融于一体，使得学科馆员能更直观，更简便的操作数据，无需在不同的软件环境中切换数据文件。此研究在节省学科馆员人工清洗数据时间，减少人工成本，缩短报告时间，提高工作效率等方面具有重要的理论和现实意义；同时，作为一个专门针对图书馆服务中需要解决的科研人员贡献度和二级单位贡献问题而开发的数据清洗程序，希望这种开发途径和程序本身能为其他高校图书馆服务提供一点参考。

2 研究内容及方法（思路、方法、具体内容）

2.1 需求分析

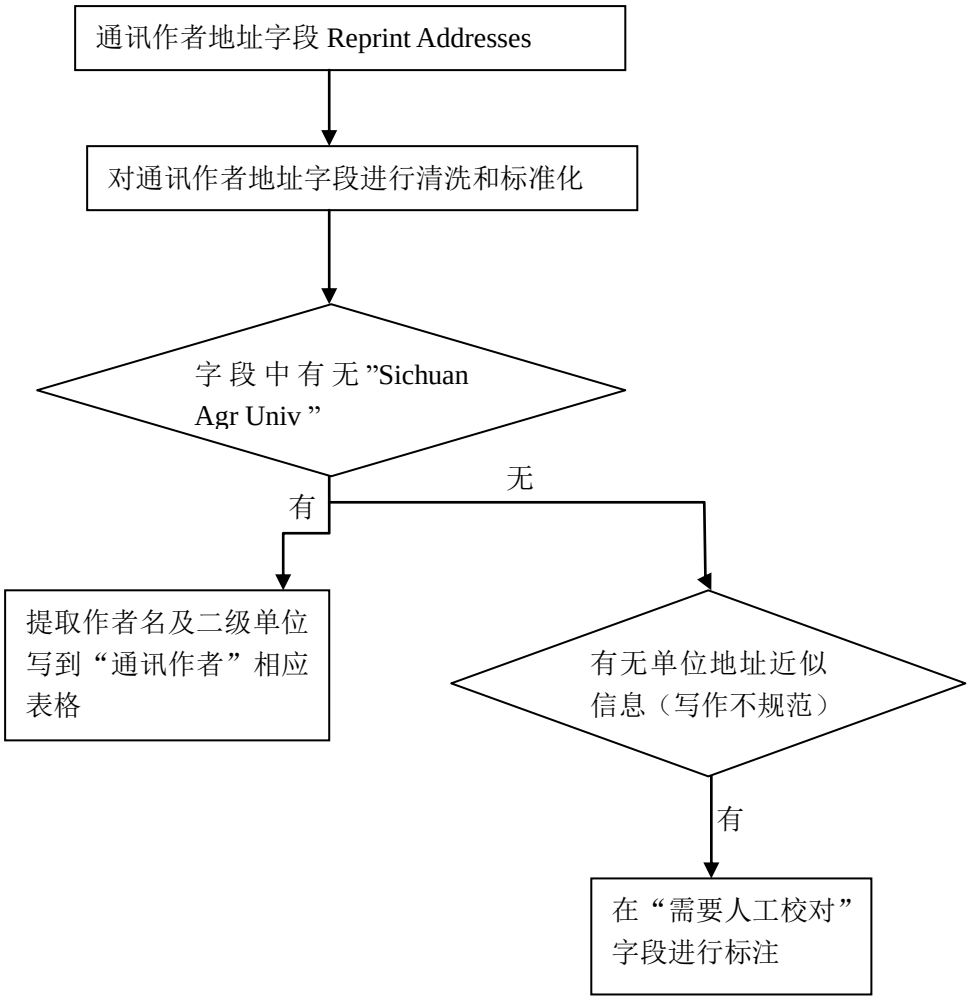
需求分析是程序开发过程中至关重要的一个步骤，做好需求调研是获取正确的程序需求的前提。在馆领导和学科馆员处获取程序需要实现的基本功能，在数据处理专员处调研其在日常人工操作、统计过程中需要处理的数据情况（包括数据的来源，数据的获取、数据的特点等）、遇到的具体问题、如何解决的、还有什么待解决的问题等等。做好需求调研，尤其是数据专员处的调研，可以帮助后面的程序实现过程提高效率，更精准的完成开发任务。

本程序主要针对我馆学科服务工作中，计算我校研究人员贡献度和二级单位贡献度的统计需求来进行程序设计和开发。经调研，目前我校对研究人员和二级单位贡献度的归属规则如下：论文权属遵循唯一性原则，即一篇论文只归属一个作者，作者归属定位顺序为：通讯作者>第一作者>其他作者，通讯作者的优先顺序是：末位通讯>其他通讯作者；二级单位的定位顺序依据作者归属而定，即通讯作者所属单位＞第一作者单位＞参与作者单位，共同通讯的以最后一个通讯作者的机构定位。因此，针对每篇文献，本程序需要提取：我校所有通讯作者及其所属二级单位；为方便后续其他需求，需要提取我校所有第一作者及其所属二级单位；在通讯作者和第一作者都不为我校研究人员的情况下，需要提取排名最靠前的我校研究人员及其排位和所属二级单位。

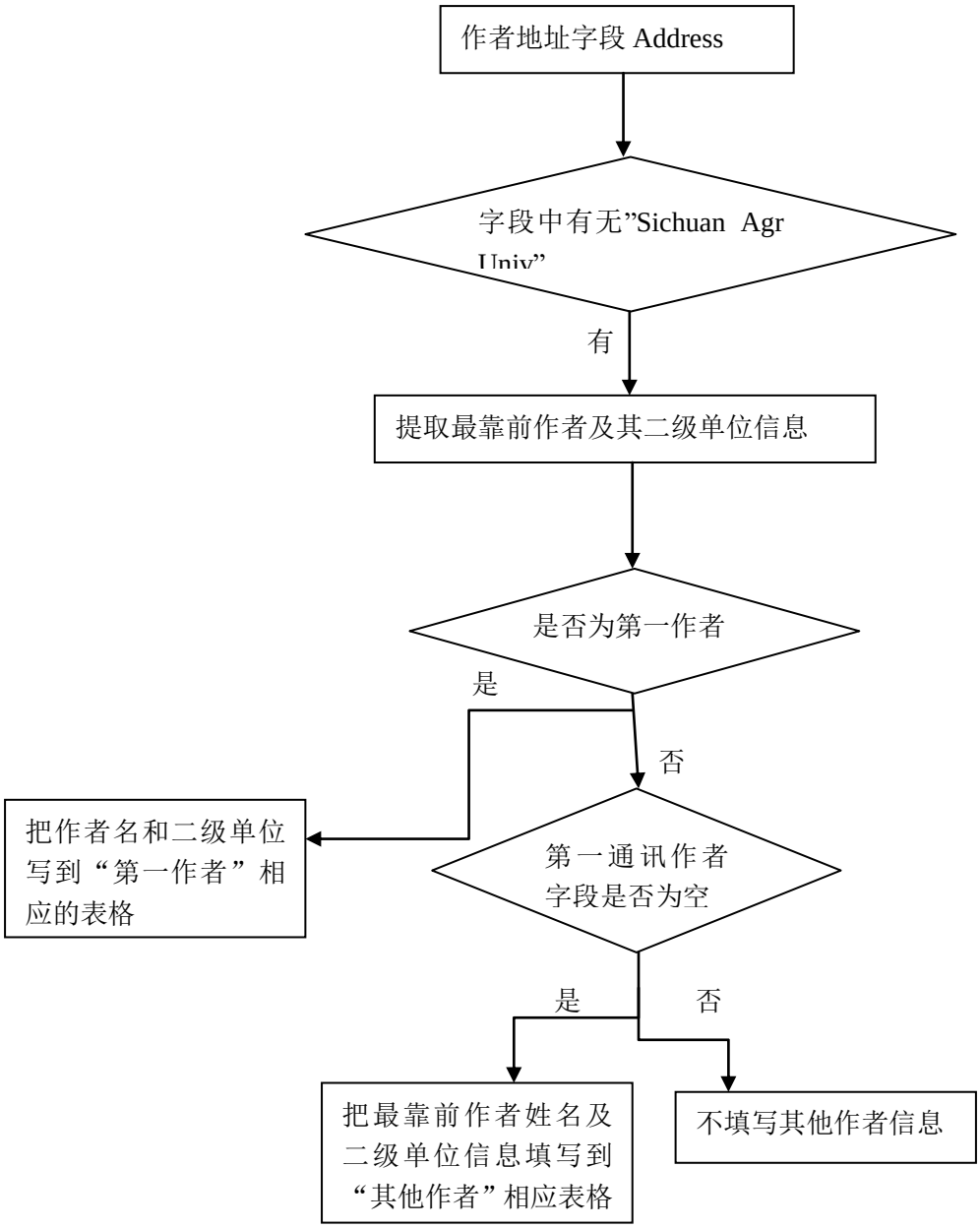
2.2 程序设计

设计是根据需求调研的结果，对程序的技术实现由粗到细进行设计。根据需求分析阶段的调研结果，本程序分成两个模块，一个是通讯作者及其所属二级单位信息的提取，一个是非通讯作者（第一作者和其他作者）及其所属二级单位信息的提取，得到如下数据流程图。

（1）通讯作者及其所属二级单位信息提取



（2）非通讯作者及其所属二级单位信息提取



2.3 程序开发、运行和维护

利用 Excel 软件环境，编写 VBA 脚本程序，将设计阶段的逻辑用程序代码实现。并用下载的 WOS 数据检测程序是否实现需要的功能，根据测试情况，对程序进行反复修改、调试，直到满足需求为止。

3 结论与建议

程序包含两大模块，一个是我校通讯作者及其所属二级单位信息的提取，一个是我校非通讯作者（第一作者和其他作者）及其所属二级单位信息的提取。按照论文权属唯一性原则，为我校文献的人员贡献度和二级单位贡献度计算很好的解决了归属问题，为我部门后期的学科服务工作提供了有力的技术支持和数据保障，大大节省了以往人工统计的时间，提高了工作效率。目前程序已经投入我校图书馆学科服务的日常使用中。但程序仍然有待改进的地方，一是可移植性不高，二是可操作性不强。在运行程序前需要一些准备工作，必须要按照设定的格式去准备好基础数据以及工作表名，在程序运行前需要修改程序中当前数据表的最大行的行号，否则程序无法运行或运行不正确。但这些操作都很简单，有详细说明文件，操作者只需在运行前按照步骤准备好基础数据即可。以期这种程序设计思路和解决问题的方法能为同行们提供一点参考。

4 项目成果

《程序运行说明》文件，见附件

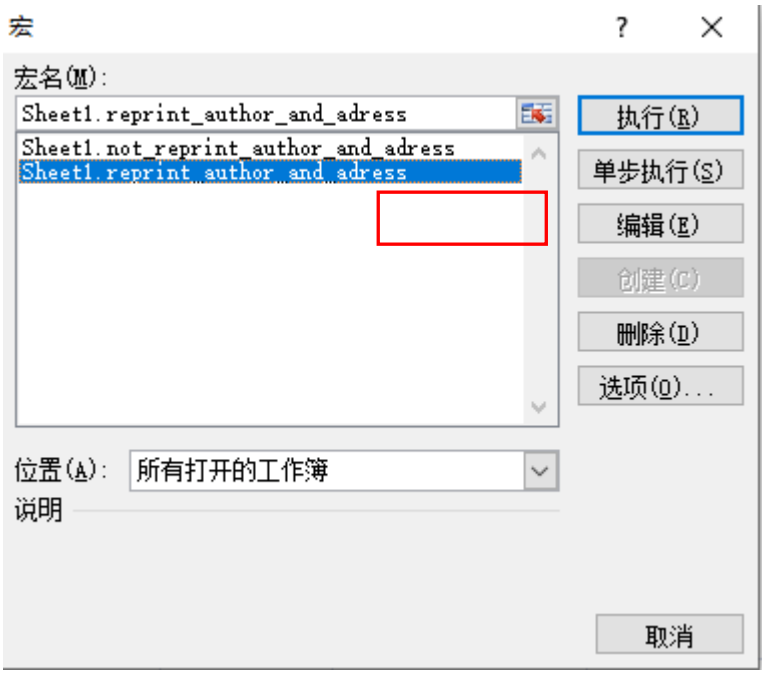
《样本文件及源代码.xls》文件，见附件

拟发表论文《图书馆服务中的数据清洗策略研究》

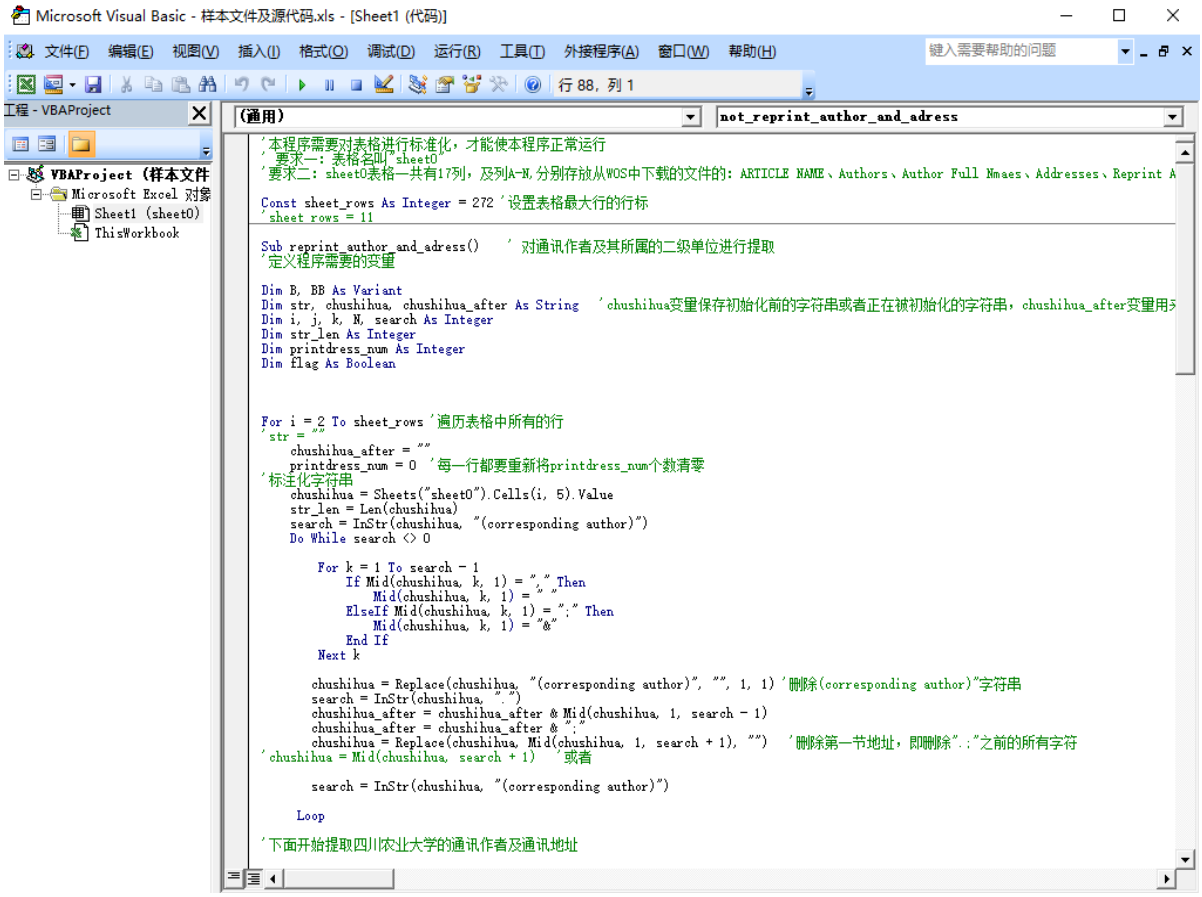
在《样本文件及源代码.xls》文件中，编辑和查看源代码的方式如下：

（1） 用 EXCEL 打开《样本文件及源代码.xls》，切换到“开发工具”选项卡

（2） 点击“宏”按钮，在弹出的对话框中点击“编辑”


宏
宏名(M):
Sheet1.reprint_author_and_adress
Sheet1.not_reprint_author_and_adress
Sheet1.reprint_author_and_adress
执行(R)
单步执行(S)
编辑(E)
创建(C)
删除(D)
选项(O)...
位置(A):
所有打开的工作簿
说明
取消

（3） 进入到源代码编辑器


Microsoft Visual Basic - 样本文件及源代码.xls - [Sheet1 (代码)]
文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 调试(D) 运行(R) 工具(T) 外接程序(A) 窗口(W) 帮助(H)
键入需要帮助的问题
行 88，列 1
工程 - VBAProject
VBAProject (样本文件
Microsoft Excel 对象
Sheet1 (sheet0)
ThisWorkbook
(通用)
not_reprint_author_and_adress
'本程序需要对表格进行标准化，才能使本程序正常运行
'要求一：表格名叫"sheet0"
'要求二：sheet0表格一共有17列，及列A-N,分别存放从WOS中下载的文件的：ARTICLE NAME、Authors、Author Full Nmaes、Addresses、Reprint A
Const sheet_rows As Integer = 272 '设置表格最大行的行标
'sheet_rows = 11
Sub reprint_author_and_adress() ' 对通讯作者及其所属的二级单位进行提取
'定义程序需要的变量
Dim B, BB As Variant
Dim str, chushihua, chushihua_after As String 'chushihua变量保存初始化前的字符串或者正在被初始化的字符串，chushihua_after变量用
Dim i, j, k, N, search As Integer
Dim str_len As Integer
Dim printdress_num As Integer
Dim flag As Boolean
For i = 2 To sheet_rows '遍历表格中所有的行
'str = ""
chushihua_after = ""
printdress_num = 0 '每一行都要重新将printdress_num个数清零
'标注化字符串
chushihua = Sheets("sheet0").Cells(i, 5).Value
str_len = Len(chushihua)
search = InStr(chushihua, "(corresponding author)")
Do While search <> 0
For k = 1 To search - 1
If Mid(chushihua, k, 1) = "," Then
Mid(chushihua, k, 1) = " "
ElseIf Mid(chushihua, k, 1) = ";" Then
Mid(chushihua, k, 1) = "&"
End If
Next k
chushihua = Replace(chushihua, "(corresponding author)", "", 1, 1) '删除(corresponding author)"字符串
search = InStr(chushihua, ";")
chushihua_after = chushihua_after & Mid(chushihua, 1, search - 1)
chushihua_after = chushihua_after & ";"
chushihua = Replace(chushihua, Mid(chushihua, 1, search + 1), "") '删除第一节地址，即删除".;"之前的所有字符
'chushihua = Mid(chushihua, search + 1) '或者
search = InStr(chushihua, "(corresponding author)")
Loop
'下面开始提取四川农业大学的通讯作者及通讯地址

5 参考文献

[1]张玲玲.高校图书馆科研数据管理服务体系构建[D].哈尔滨:黑龙江大学,2021.

[2]樊慧丽,邵波.高校图书馆数据清洗问题与策略研究[J].高校图书馆工作,2017, 37 (182):35-40.

[3]左志林.我国高校图书馆数据馆员研究[J].图书馆建设,2020(1):138-144.

[4]王曰芬,章成志,张蓓蓓,等.数据清洗研究综述[J]现代图书情 报技术,2007(12):50-56.

[5]廖书妍.数据清洗研究综述[J]电脑知识与技术,2020,16(20):44-47.

[6] 赵月琴,范通让.科技创新大数据清洗框架研究[J].河北省科 学院学报,2018,35(2):35-42.

[7]叶鸥,张璟,李军怀.中文数据清洗研究综述[J].计算机工程与 应用,2012,48(14):121-129.

[8] 张晋辉,刘清.基于推理机的 SCI 地址字段数据清洗方法设 计[J].情报科学,2010(5) :741-746.

[9] 马晓亭.基于大数据决策分析需求的图书馆大数据清洗系统设计 [J] .现代情报,2016(9):107-111.

[10]杨文建,邓李君.国外高校图书馆科研数据管理研究进展及其启示[J].国家图书馆学刊,2017,26(05):88-97.

[11]王继娜.国外高校图书馆科学数据管理服务的调研与思考[J].情报理论与实践,2019,42(08):159-167.