doi:10.16104/j.issn.1673-1883.2019.01.017

《文始》《说文》数据库的建构及意义

许良越
(西南民族大学文学与新闻传播学院,成都 610041)

摘要:章太炎的《文始》是汉语语源研究史上的一部重要著作,在许多方面都具有开创性的贡献。然而,由于《文始》组织结构的层级分类性与所收字头信息的复杂交错性,长期以来难于对这部著作进行全面整理,由此造成对其成就意义的莫衷一是。借助计算机数据库信息处理技术,通过构建"《文始》《说文》全文数据库",不仅能够实现梳理分析的穷尽性与查询统计的便利性,而且还能以全面准确的量化数据还原章氏字源学理论本有的实际价值,并可为基于数据库方法的典籍文献研究提供建模设计参考。

关键词:章太炎;《文始》;《说文》;数据库
中图分类号:H139 **文献标志码**:A **文章编号**:1673-1883(2019)01-0080-05

The Creation and Significance of the Database of *Wenshi*(文始)and *Shuowen*(说文)

XU Liangyue
(School of Literature and Journalism, Southwest Minzu University, Chengdu, Sichuan 610041, China)

Abstract: *Wenshi*(文始), a classic work by Zhang Taiyan, has made pioneering contributions in many respects in the history of Chinese etymology. Nevertheless, due to hierarchical classification of the organization and the complicated relationship of the character information, there has been no comprehensive and exhaustive study on *Wenshi*(文始)for long, resulting in a debate on its achievement. With the aid of database information processing technology to create a joint retrieval system of the complete monographs of *Wenshi*(文始)and *Shuowen*(说文), we can not only make convenient and thorough analyses of these monographs and pass objective evaluation of the achievement of Zhang Taiyan's in Chinese etymology with complete and accurate quantized data, but also provide a modeling reference for the research of classic literature based on database technology.

Keywords: Zhang Taiyan; *Wenshi*(文始); *Shuowen*(说文); database

21世纪的语言学研究已经进入了一个理论创建和技术应用互补的时代,对各类语言文字的信息处理成为衡量学科现代化水平的重要标志。在这样的时代背景下,将以数据库方法为核心的计算机信息处理技术导入传统的语言文字研究中亦是趋势的必然。

一、《文始》在汉语语源研究中的重要价值

《文始》始撰于1908年,初次发表于1910年,书成于1913年,是章太炎在旅日时多次批阅和讲解《说文》的基础上形成的。章氏撰作《文始》的目的是为了"求语根""明语原",以证明汉字的"孳乳浸多之理"。《文始》一书的性质决定了该书的价值,总体说来,主要有以下几个方面:

(一)章太炎语言文字学的代表作

章氏的语言文字之学,是以《说文》字源研究为核心的一个完整的学说体系。这一学说体系的理论基础,首先表现于《国故论衡》中的《语言缘起说》《转注假借说》《成均图》和《古双声说》。这四篇文章分别从同源字族的音义起点、义转孳衍的方式原则、音转变化的表现规律三个方面组成了一个系统的汉语字源学。在《文始》中,章氏又将"语言缘起理论"化作初文与准初文,将"转注假借理论"化作变易与孳乳,将"声转韵转理论"化作《成均图》及"纽目表",并以此为指导进行了构建《说文》同源字族的实践。所以,《文始》一书全面而集中地体现了章氏语言文字之学的根本理论和具体实践,是章氏众多语言文字学著作中当之无愧的代表作。

收稿日期:2018-11-09
基金项目:四川省社会科学规划项目:章太炎字源学的词族验证与价值重估(SC18B150)。
作者简介:许良越(1978—),男,江苏苏州人,讲师,博士,研究方向:汉语语源学研究。

(二)传统语源研究的最高水平

《文始》之前的汉语语源研究,大都是零散随意的系流考源,或是连类而及的排比疏释,尚缺乏完整明确的理论指导和系统全面的系联求源。章氏在继承清人"义存乎声""因声求义""声近义通"的基础上,"总集字学、音学之大成",在《文始》中提出了"形体声类,更相扶胥"的形音义互求观点。根据汉语与汉字的特殊关系,以汉字字族的研究来探讨汉语语源的问题,做到了依凭字形但绝不拘泥于字形,在汉语语源研究史上首次以系统整体的角度出发,将看似无关的单个汉字类聚为音义来源上皆有联系的汉语字族,从而揭示了汉字孳生嬗变的演进过程和语词之间的亲缘关系,代表了传统语源研究的最高水平。

(三)开创了基于《说文》的汉语字源学

传统的《说文》研究,大多集中于体例特点、六书理论、形义系统、声义系统,而都未能从发生学的角度对《说文》进行梳理系联。章氏撰作《文始》,先从《说文》中独体与准独体的字形入手,归纳出同源起点的初文准初文;再依据《说文》的谐声体系和读若读如,总结出同源音变的韵表纽表;又根据《说文》中的训释说解和文献例证,概括出同源义转的孳乳变易。在此基础上,对《说文》中保存的先秦字料做了全面的系联求源,将其重新整理编排,使隐含其中的汉语字族首次得以显现,用新的方法开辟了《说文》研究的新领域,为《说文》研究提供了一种全新的视角,开创了基于《说文》形音义体系的汉语字源学。

(四)我国第一部汉语同源字典

《文始》撰成于我国语言文字研究由旧的"小学"向新的"语言文字之学"过渡的时期。它的出现第一次明确建立了汉语同源字族构建的理论原则、系联的方法模式、具体的音义标准,为汉语同源字典的编撰奠定了基础。《文始》在编排体例上完全打破了传统字典辞书或从字形出发以部首统领汉字,或从字义出发以义类统领汉字,或从字音出发以韵目统领汉字,而代之以形音义三者兼顾的做法:以音排序、据形立目、按义归类。章氏首创的这种全新的同源字族编排框架,合于汉语同源字词音变成词、词分造字、音近义通的特点,解决了同源字词的类聚编排问题,成为后世各类汉语同源字典通行的编撰体例,是我国真正意义上的第一部汉语同源字典。

在汉语语源研究史上,《文始》是一部承上启下、继往开来的重要著作。它既是对传统语源观念的全面总结,又是章黄学派字源理论的最重要成果,更对其后汉语语源研究产生了深远影响。

二、《文始》研究数据库化的重要意义

数据库是按一定结构组织、并可以长期储存在计算内的、具有某些内在含义的、在逻辑上保持一致的、可共享的大量数据集合[1]。数据库技术是目前使用计算机进行数据处理的主要方式,在以大批量数据的存储、组织和使用为基本特征的领域里,数据库有着广泛的应用。在《文始》研究中引入数据库处理技术,其必要性主要体现在以下几个方面。

(一)全面梳理的需要

章氏字源学理论的系统性与《文始》整体结构的紧密性,决定了不可能从主观随意的举例性讨论中看出《文始》的原有价值,而必须在全面系统的梳理分析以后才能予以揭示与说明。然而,《文始》对《说文》同源字族的组织编排,并非是平面化单线性的前后罗列,而是立体化多线性的层级分类;加之其收字规模较大,每个字又各有音义字源等若干方面的信息,相互之间呈现出复杂的关系网络。这些原因使得要全面梳理《文始》,仅靠人工是难以完成的,必须借助计算机数据库技术才能真正实现穷尽性的全面梳理。

(二)定量分析的需要

长期以来对《文始》的研究,由于缺乏全面系统的梳理,使得无论是赞成的一派还是批评的一派,对其评价都常常是概念性的、印象性的和举例性的。遗憾的是,学者们往往把这种基于零散材料而得出的局部性表现,夸大为整部《文始》的全局性特点,由此造成一直以来对《文始》成就价值的不断争议。要想全面、客观的评价《文始》,给出让人信服的定性结论,首先必须对《文始》进行严格意义上的定量分析;而要进行定量分析,不重不漏地获得精确的量化数据,就必须以数据库方法为基本分析工具。

(三)对比查询的需要

《文始》既以《说文》为汉语同源字族的系联对象,《说文》中保存的汉字的各类信息,自然成为《文始》字族系联的基础和依据。因此,为了充分展现《文始》中章氏归纳字族的原因和理由,在全面梳理《文始》的同时,还必须经常查询比对《说文》中的相关信息。如果采取人工频繁检索的方式,不仅速度慢,易出错,而且还难于做全面的对比统计。反之,使用数据库技术中的联表查询方法,将《文始》与

《说文》在数据表层面关联起来，不仅能使信息查询简单易行，而且还能随时得出详细准确的统计数据。

（四）知识库建设的需要

知识库是指为方便和有效地使用与管理大量的知识，而把人类已经具有的知识以一定的形式表示存储到计算机中所构成的系统，又称知识库系统。它是任何基于知识的智能系统的基础[2]。知识库系统的建设或是基于人工智能（专家系统），或是基于数据库技术。利用数据库技术将《文始》和《说文》中的信息以二维关系表的形式表征与存储后，就能以“《文始》《说文》全文数据库”为语料数据平台，根据一定的规则表达式推导出相应的结果记录集，形成“《文始》《说文》知识库”，从而不仅使检索查询《文始》更加便捷，也能使其中的内在理论体系得以明晰表现。

在《文始》研究中引入数据库技术，是科学而系统的整理《文始》所必备的基本条件。这种处理方式既能满足高效梳理《文始》的实际需要，又符合文献典籍数字化的发展趋势，并能在传统学科研究方式的革新方面做出新的有益探索。

三、《文始》《说文》全文数据库的设计

构建“《文始》《说文》全文数据库”，首先必须选定录入数据的文献版本和实现数据库的数据库管理系统。为了达到最优化的系统设计，在文献本版的选择上，《文始》采用的是上海古籍出版社出版的《章太炎全集七》中殷孟伦先生的标点本[3]，并参校浙江图书馆的木刻校勘本[4]；《说文》采用的是中华书局出版的大徐本[5]，并参考臧克和、王平、刘志基开发的“《说文解字》全文检索系统”[6]；在数据库管理系统的选择上，采用的是目前占据主流地位的关系型数据库。

按照关系型数据库的设计理论及建模方式[7]，“《文始》《说文》全文数据库”的构建流程可分为以下几个步骤：

（一）概念结构设计

概念结构设计，就是用ER图（实体关系图）为工具，对《文始》与《说文》的组织结构关系进行概念结构分析，抽取出其中的实体、属性及联系，以构建出二书的概念结构模型，完成从现实世界到信息世界的第一级抽象。

《文始》的组织结构，实质上是一种层次结构。这种层次结构具体体现为：全书分作若干卷，每卷下编排有若干条，每条下派生出若干级，每级下归并为若干类，每类下类聚起若干字。根据《文始》在组织结构上的特点，用ER图抽象出其概念结构模型如下：

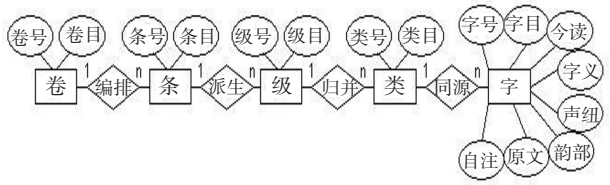

图1 《文始》组织结构ER图

《文始》各实体及其属性说明如下：

《文始》“卷”实体，有2个属性：“卷号”是“卷”的数字序号，“卷目”是以古韵二十三部的对转关系为各卷的名称。

《文始》“条”实体，有2个属性：“条号”是“条”的数字序号，“条目”是以每条的第一个字头作为各条的名称。

《文始》“级”实体，有2个属性：“级号”是“级”的数字序号，“级目”是据各条中字头与初文或准初文的远近关系确定的派生层级。

《文始》“类”实体，有2个属性：“类号”是“类”的数字序号，“类目”是归并各条中具有相同变易或孳乳来源的字头后得出的字源类别。

《文始》“字”实体，有8个属性：“字号”是“字”的数字序号，“字目”是各条所收的具体字头，“今读”是对字头的现代汉语注音，“字义”是字头的本义说明，“声纽”是字头所属的上古音声纽，“韵部”是字头所属的上古音韵部，“原文”是《文始》对字头的说解，“自注”是字头为《文始》自注中的字头。

《说文》的组织结构，也是一种层次结构，具体体现为：全书分作若干卷，每卷下编排有若干部，每部下类属若干字，每字下分列出若干形。根据《说文》在组织结构上的特点，用ER图抽象出其概念结构模型如下：

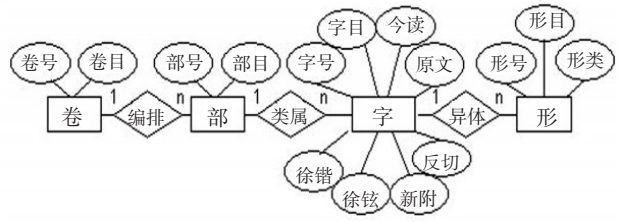

图2 《说文》组织结构ER图

《说文》各实体及其属性说明如下：

《说文》“卷”实体，有2个属性：“卷号”是“卷”的数字序号，“卷目”是各卷的次序名称。

《说文》“部”实体，有2个属性：“部号”是“部”的数字序号，“部目”是各部的部首名称。

《说文》“字”实体,有8个属性:“字号”是“字”的数字序号,“字目”是各部所收的具体字头,“今读”是对字头的现代汉语注音,“原文”是《说文》对字头的说解,“反切”是徐铉所加的《唐韵》反切,“新附”是字头为徐铉后增的新附字,“徐铉”是徐铉对字头的注释,“徐锴”是徐锴对字头的注释。

《说文》“形”实体,有3个属性:“形号”是“形”的数字序号,“形目”是与字头对应的各类字形,“形类”是《说文》对各字形的归类。

虽然《文始》与《说文》有着密切的关系,但因为二者是两本独立的著作,所以在以上概念结构的设计中,没有对二书作合并处理,而是仍使其各自保持独立性,分别建立概念结构模型。

(二)逻辑结构设计

逻辑结构设计,就是根据关系数据理论的转换规则,将《文始》与《说文》ER图中的实体、属性及实体之间的联系转换为相应的关系模式,以构建出二书的逻辑结构模型,完成从信息世界到数据世界的第二级抽象。

《文始》ER图中有“卷”“条”“级”“类”“字”5个实体,相应地可转换为5组关系;各个关系中的属性为相应实体中的属性;各实体之间皆为1∶n联系,因此可将各实体1端的主键纳入n端实体中作为外键,建立关系之间的联系。经转换后,《文始》的逻辑结构可用以下关系模式表示(加下划线的属性为主键,下同):

文始卷(卷号,卷目)

文始条(条号,条目,卷号)

文始级(级号,级目,条号)

文始类(类号,类目,字号,级号)

文始字(字号,字目,今读,字义,声纽,韵部,原文,自注,类号,说文字.字号)

《说文》ER图中有“卷”“部”“字”“形”4个实体,相应地可转换为4组关系;各个关系中的属性为相应实体中的属性;各实体之间皆为1∶n联系,将各实体1端的主键纳入n端实体中作为外键,建立关系之间的联系。经转换后,《说文》的逻辑结构可用以下关系模式表示:

说文卷(卷号,卷目)

说文部(部号,部目,字号,卷号)

说文字(字号,字目,今读,原文,反切,新附,徐锴,徐铉,部号)

说文形(形号,形目,形类,字号)

由于在概念结构的设计中对《文始》与《说文》采取的是分别建模的方式,二者实际上并无关联。但在实际操作时,《文始》与《说文》必须先相互关联才能实现联合检索。对此问题采取的解决办法是:将“说文字”关系模式中的主键“字号”加入“文始字”关系模式中作为外键,从而实现《说文》与《文始》的1∶n联系。这种1∶n联系,是由于《文始》所收的《说文》字头有重出现象造成的。此外,为了对应《文始》中收录有,但《说文》中却未收的字,尚需在《说文》的各关系模式中,增加一个序号为“0”的记录,其所有属性取空值,表示该字为《说文》所未收。按照这样的处理方式,即可实现对《文始》《说文》的联合查询。

(三)规范化分析

规范化分析,就是根据关系规范化理论,对《文始》与《说文》逻辑结构中的关系模式进行分析,确定各关系模式中属性之间的函数依赖关系和达到的范式等级,以检测系统设计的优劣程度。

运用规范化理论,可以看出,在以上关系模式中,除“文始类”和“说文部”2个关系模式外,其余的7个关系模式,主键都是本模式的唯一决定因素,所以这7个关系模式都属于BC范式,在函数依赖的范畴内,规范化程度已经达到了最高。

至于“文始类”和“说文部”2个关系模式,考虑到在实际检索操作时,可能需要提取作为“类目”或“部目”字头的各项信息,所以在逻辑结构设计时将“文始字.字号”属性和“说文字.字号”属性分别加入这两个关系模式中,以便能直接检索“类目”和“部目”字头的相关信息。这样做虽然降低了这2个关系模式的范式等级,但却提供了查询时的便捷性,属于必要的冗余属性。

(四)表结构创建

表结构创建,就是根据关系数据库管理系统的要求,将《文始》与《说文》逻辑结构中的关系模式转换为相应的数据表形式,并对数据表中的属性名称、数据类型、长度大小、取值范围等问题作出规定与说明,以建立存储数据的基表结构,是对整个设计流程的全面总结和最终表示。

综合前述,合并表示《文始》《说文》表结构如表1、表2所示。

四、《文始》《说文》全文数据库的效用

根据上述“《文始》《说文》全文数据库”的设计方案,以Microsoft Access 2010为系统运行平台,现已完成对《文始》与《说文》全文数据的录入与校对工作。经实际操作表明,利用“《文始》《说文》全文数据库”,能对《文始》全书各卷各条目的收字组成

表1 《文始》表结构

属性名称	数据类型	长度大小	取值范围与说明
卷号	数值	2	卷目序号(前一位)+同卷韵部序号(后一位)
卷目	字符	14	卷名+同卷韵部名
条号	数值	3	条目顺序编号
条目	字符	2	每条中的第一个字头
级号	数值	4	条目序号(前三位)+同条层级序号(后一位)
级目	数值	1	每条层级从1级起始,变易派生在同级,孳乳派生产生新的层级
类号	数值	5	条目序号(前三位)+同条字源类别序号(后两位)
类目	字符	30	各条中字头的字源类别归属(初文/准初文/变易/孳乳)
字号	数值	5	条目序号(前三位)+同条字头序号(后两位)
字目	字符	2	各条中的具体字头(含章氏自注)
今读	字符	6	字头的现代读音
字义	字符	50	字头的本义说明(据《汉语大字典》)
声纽	字符	10	字头所属的章氏古声二十一纽(同时标注王力古声三十三纽)
韵部	字符	10	字头所属的章氏古韵二十三部(同时标注王力古韵三十部)
原文	备注	/	各条中字头下的相应说解(含章氏自注)
自注	逻辑	1	字头是否为章氏自注

表2 《说文》表结构

属性名称	数据类型	长度大小	取值范围与说明
卷号	数值	2	卷目顺序编号
卷目	字符	6	卷目名称
部号	数值	3	部首顺序编号
部目	字符	4	部首名称
字号	数值	4	字头的顺序编号
字目	字符	2	各部中的具体字头(含新附字)
今读	字符	6	字头的现代读音
原文	字符	250	各部中字头下的相应说解
反切	字符	6	字头的《唐韵》反切注音
新附	逻辑	1	字头是否为新附字
徐铉	字符	250	徐铉对《说文》字头的注释
徐锴	字符	250	徐锴对《说文》字头的注释
形号	数值	5	字形顺序编号
形目	字符	2	字头的相应字形(小篆/重文)
形类	字符	150	许慎对字头的字形类别归属(篆文/古文/籀文/或体)

情况、初文准初文情况、韵转声转情况、变易孳乳情况、书证引文情况等方面问题进行严格意义上的穷尽性统计分析,并能随时与《说文》进行关联查询,在此基础上可以形成详细到字头各项信息的数据报表。这些都是传统的训诂疏证式分析方法所无法比拟的。

通过数据库方法处理《文始》《说文》后,不仅使得检索《文始》《说文》更加方便快捷,为汉语同源词研究提供了易于获得的丰富语料资源,而且还能以全面的精准数据纠正先前对《文始》的诸多不当结论,客观还原出章氏字源学理论本身的意义价值。对这方面问题的详细论述,请参看拙文《数据库环境下的〈文始〉研究》[8]。

事实证明,在传统语言文字研究中引入数据库技术,不仅能在文献语料的统计处理上更为方便准确,而且还能通过建模设计过程与量化分析方式,充分展现出研究对象内在的本质特征和外在的表现特点,真正实现了研究手段的科学化和表达形式的精确化。可以预言,基于数据库技术的典籍文献研究,必将开启我国传统语言文字研究的新局面!

参考文献:

[1] 姚普选.数据库原理及应用[M].北京:清华大学出版社,2006:1.
[2] 宋继华,王宁.基于超文本环境的《说文解字》知识库的建立[J].语言文字应用,1999(3).
[3] 章太炎.章太炎全集(七)[M].上海:上海人民出版社,1999.
[4] 章太炎.章氏丛书[M].杭州:浙江图书馆校刊,1919.
[5] 许慎.说文解字[M].北京:中华书局,1963.
[6] 臧克和,王平,刘志基.《说文解字》全文检索[M].广州:南方日报出版社,2004.
[7] 刘志妩,张焕君,马秀丽.基于VB和SQL的数据库编程技术[M].北京:清华大学出版社,2008:15-33.
[8] 许良越.数据库环境下的《文始》研究[A].励耘语言学刊[C].北京:学苑出版社,2016.