浅析简帛文献的数据库建设

李　静

（湖北 武汉 430072）

简帛文献，大致是指在纸张广泛作为书写载体之前，书写于竹简、木牍、绢帛上的古代文献。数据库技术，是管理信息系统中的一个关键技术，它所研究的问题是如何科学地组织和存储数据，如何高效地获取和处理数据[1]。对海量信息进行收集、筛选、分类、录入，建立快速有效的检索系统，是信息科技时代诸多学科需要解决的问题。简帛学是一门面对过往的学问，一方面埋藏于地下的古代文献不断被发现，另一方面现代信息技术的发展则给这些古代文献提供了新的研究方法和手段，如何在古代文献与现代科技之间搭建桥梁，建立快速有效的信息纽带，实现数据的分类与检索，是进行学术研究重要的技术手段。

简帛文献数据库，从功能上可分为字形数据库和文本数据库，都是面向对象的关系数据库，前者主要实现字形的快速检索、比对，后者主要实现释文文本及参考文献的检索，目前已开发或正在建设中的简帛文献数据库，多数涵盖以上两种功能，又在侧重点上各有不同，以下简单介绍几种常用的数据库：

“汉达文库”之“竹简帛书资料库”：该数据库由香港中文大学中国古籍研究中心制作，收录资料包括《居延汉简释文合校》、《武威汉代医简》、《武威汉简》、《马王堆汉墓帛书》、《散见简牍合辑》、《睡虎地秦墓竹简》、《银雀山汉墓竹简》、《定县汉简》、《阜阳汉简》、《张家山汉简》、《郭店楚简》和《上海博物馆藏战国楚竹书》，以原报告释文为底本，根据研究成果对释文进行修订，据传世文献对释文进行增补，所有校改内容皆加以标识，形成新的释文文本，可以实现全文浏览、图像浏览、简帛文献与传世文献对比、字词句检索等功能[2]。

战国楚简帛文字典型形体检索系统：该数据库由华东师范大学中国文字研究与应用中心开发，收录资料包括《包山楚简》、《长沙子弹库战国楚帛书研究》、《郭店楚墓竹简》、《九店楚简》、《曾侯乙墓》、《战国楚竹书汇编》、《上海博物馆藏战国楚竹书》第一至五册，以及《葛陵新蔡楚墓》，在上述材料范围内以现有考释研究为基础，通过细致的字形认定，提供GBK编码范围内战国楚简帛文字的原始典型形体。字形的选择以形体清晰、结构具有代表性为标准，目前只提供GBK编码范围内通用字头的检索[3]。

简帛金石资料库：该数据库由台湾中研院历史语言研究所文物图像研究室制作，收录资料包括郭店楚简、包山楚简、曾侯乙墓简、睡虎地秦简、马王堆帛书、银雀山汉简、居延汉简、敦煌汉简、两汉镜铭集录、汉代石刻集成、汉印文字征及补遗等六十余种资料，可以进行释文和参考文献检索，检索结果以文本形式呈现[4]。

中国古代简帛字形、辞例数据库：该数据库由武汉大学简帛研究中心开发，包含四个子库，分别是楚简、秦简、汉简和其他，目前包括已发布的全部楚简资料，部分秦简牍、汉简牍资料和古代字书数据，可以实现对单字、偏旁以及辞例的检索，在“简帛网”上共享使用[5]。

作者：李静，武汉大学历史学院、简帛研究中心。

此外还有香港中文大学图书馆制作的“郭店楚简资料库”、“走马楼三国吴简·嘉禾吏民田家莂资料库”,台湾中研院史语所制作的“汉代简牍数位典藏”以及复旦大学出土文献与古文字研究中心制作的“上博简字词全编资料库”和“清华简字形辞例数据库”等。以上列举的各种数据库,给研究者提供了很大的便利,但在使用过程中我们也发现这些数据库的设计或多或少存在一些缺陷,因而影响了检索结果的准确性。这大致可以概括为以下两点:

1.字符编码与输入法

现有的计算机汉字编码是十分有限的,目前已有的GBK和Unicode编码无法囊括所有的汉字,尤其是手写的古文字。我们在写定释文的过程中,对于这类字的通行做法是造字或以图片代替。因这类字不能转化成计算机识别的编码,所以在数据库检索中,这类字无法被准确检索。编码解决的是计算机如何识别和显示古文字的问题,输入法则是解决数据库的使用者能否输入古文字进行检索,这两个因素制约了简帛文献数据库的全面性和准确性。

张再兴先生提出“彻底数字化”原则,即首先对古文字进行造字处理,建立各种类型的古文字字体,解决字形在电脑中的显示问题,其次开发相应的古文字字形输入编码,使古文字字形转变成计算机能够识别的字符,最后研发古文字输入法,解决古文字的输入[6]。这是一种彻底的解决方案。

由裘锡圭先生担任首席专家的“中华字库”工程,旨在建成全部汉字及少数民族文字的编码和主要字体字符集,形成汉字编码体系,重点研发汉字输入、输出、传播以及兼容等关键技术。古文字作为汉字发展阶段,利用先进技术对现有资料进行全面、准确的整理,为古文字的输入、输出制定统一可行的技术标准,既为简帛文献整理提供技术方案,也是简帛文献数据库建设实现“彻底数字化”的前提条件。

另一种解决方案是通过输入单字中出现的偏旁进行检索,这里又涉及到偏旁的确立和汉字的拆分原则。刘志基先生倡导通过创建通用古文字偏旁数字平台,以营造有助于偏旁分析法科学运用的数字化环境。偏旁数字平台的创建除了必须穷尽资料和全面吸取既有古文字考释研究成果外,还需要完成“偏旁”的宽容认定,根据构形差异、易导致构形差异的相关属性、造字意图语境中的用法、组合关系特点等完成对偏旁的系统标注[7]。

不论是“彻底数字化”原则,还是偏旁数字平台的建立,都是浩大的工程,需要在统一运筹下实施,并成为行业内认定的标准。目前简帛文献的数据库建设仍然以个体为单位进行,对于古文字的隶定、偏旁的确立和分析仍存在学术分歧,面对计算机语言的精确性、唯一性要求,复合偏旁检索可以避开字体与输入法在技术方面的专业要求,实现现有字库中没有的字或者不识字的检索。我们建议在已有古文字考释研究成果的基础上,以检索的准确性为目标,确定数据库使用的偏旁以及拆分原则,在数据录入过程中始终坚持这一标准,并在数据库制作完成后公布这些信息以便用户查询。

2.释文体例与数据库设计

释文是连接学术研究与信息科技的桥梁,释文水平的高低直接影响数据库的检索效率,简帛文献数据库的建设不是简单的编程正在于此。准确、标准的释文是我们建设数据库的基础,因此基于数据库的释文,首先应当明确释文报告的底本、释文隶定遵循的原则、通假字、异体字的标注等,这些信息应当如简牍报告的凡例一样公布在数据库的相关页面。计算机语言的特点是唯一性及准确性,所以我们建议以原始报告释文为底本,对有明确错误的地方加以修订,释文隶定采用从严原则,对于现有字符集里没有的古文字,随文注出通假字或异体字,按照偏旁使用及拆分标准进行偏旁分析。对于那些已识但现有编码字符集没有的字,既可通过查询通假字或

者异体字进行检索，也可以进行复合偏旁检索。对于不识字，可以通过复合偏旁方式检索。

释文编写是基础工作，由于我们在做这项工作时已经包含了通假字、异体字信息，所以我们可以在建设标准字库（即全文字形数据库）的同时，建设通假字库和异体字库，实现对于这类古文字的快速分类与检索。

通过上述两点，我们可以看出简帛文献数据库制作的首要条件是标准的制定，这是由计算机的特性决定。学术研究是多元化的，我们需要在学术研究与标准之间取得平衡，建立简帛文献数据库的适用规则。此外，新资料不断被发现，新研究成果不断推出，我们需要开发相应的后期维护系统，及时对数据库进行更新与修订，以保证数据库资料的完整性和准确性。

简帛文献数据库的建设始于21世纪初，经过十几年的努力取得了一些规模和进展，为传统学术提供了新的研究方法和手段。虽然在制作标准、资源整合等方面仍存在一些问题，但却是传统学术利用现代科技迈向数字化的开端。随着科技的进步，数据库建设过程中面临的难题将逐步解决，新技术的产生也将对数据库开发者提出新的功能需求，简帛文献数据库建设将日臻完善。

注释：

[1]陈林生、高学东：《数据库技术的现状及其发展趋势》，《管理信息系统》1999年第3期。

[2]网址：http://www.chant.org/。

[3]网址：http://www.wenzi.cn/pages/czjs.asp/。

[4]网址：http://saturn.ihp.sinica.edu.tw/~wenwu/search.htm/。

[5]网址：http://www.bsm-whu.org/zxcl/。

[6]张再兴：《互联网时代出土文献数据库建设的思考与实践》，《中国文哲研究通讯》第21卷第2期，台湾中研院中国文哲研究所2011年6月。

[7]刘志基：《古文字偏旁数字平台与数字化环境下的古文字偏旁分析法》，《中国文字研究》2015年第2期。

（责任编辑、校对：蔡丹）