DOI:10.15941/j.cnki.issn1001-0424.2018.12.008

“中华字库”工程第七包“两汉吴魏晋简牍文字”数据库建设研究

孟 忻

【摘 要】国家新闻出版重大科技工程项目“中华字库”项目工程分为28个包，吉林大学古籍研究所项目组负责第七包“两汉吴魏晋简牍文字”的搜集和整理。并按照网络数字化、出版业的要求，全面、系统地汇编两汉吴魏晋统治时期使用的简牍及墨书材料上的文字，建立古汉字及古代少数民族文字的编码和主要字体字符库。

【关键词】中华字库 两汉吴魏晋时期 简牍文字

Abstract: The major science and technology project “Chinese Characters” on the national press publication is composed of 28 subdivisions. The Research Institute of Chinese Ancient Books, Jilin University, has been in charge of the division 7, which related to the characters written on Han-Jin bamboo slips. To meet the requirement of the digitization and publication, the project aims to build up an overall and systematic compilation of the handwritten characters from Han-Jin and bamboo slips. The project also aims to build up the encoding scheme and script type table of ancient Chinese characters and ancient minority scripts.

Key words: Chinese characters Han—Jin period bamboo slips

0 引言

“中华字库”工程第七包“两汉吴魏晋简牍墨书文字”材料（以下简称“汉简”）的搜集与整理2011年11月进入了实施阶段[1]。在其后近8年的时间里，该项目工程将对这部分的文字尽可能地进行系统性的整理。以实际文本的原始图像为基础，确定标准的规范形体，标注各类属性，有序地分层级排列，建立字际间某种特定的联系，并按照出版印刷、网络数字化的需求，给予确切的编码，建立符合各种需求的汉字、各少数民族文字[2]、非字符号的主用字体字符库。这是文字资源计算型情报数据库的一大创新。

1 “汉简数据库”的信息源

上世纪初至今，两汉吴魏晋简牍大批出土，总数量近40万余支，数量远远超过其他类型的古文字材料，分布在甘肃、湖南、内蒙古、山东、台湾，国外也有收藏：英国、日本、法国、瑞典、美国等。比较重要的有居延汉简、银雀山汉简、睡虎地汉简、武威汉简、凤凰山汉简、马王堆汉简、西安未央宫遗址汉简、八角廊汉简、阜阳汉简、悬泉置汉简、大通上孙家寨汉简、张家山汉简、走马楼汉简、尹湾汉简、虎溪山汉简、水泉子汉简、孔家坡汉简、荆州印台汉简、东牌楼汉简、朝鲜平壤贞柏洞汉简（以上属于两汉）；新疆古尼雅遗址魏晋简、古楼兰遗址魏晋简、郴州苏仙桥遗址吴晋简、走马楼吴简（以上属于吴魏晋）等。其内容主要包括：书籍、簿籍、书檄、律令、案录、检楬、遣策与告地策等[3]。公布汉简文字的图版、释文与研究成果的图书很多，调查、收集资料的工作量和难度很大，获取高清图像的技术性很高，研判汉简文字释读的专业性很强，很多汉简疑难字已经成为汉简研究深入发展的障碍，必须加以系统研究和释读。对汉简中的异体字和俗体字准确加以判定，并建成相对完善的资料库，汉简文字图版与文字的数据化处理必须高水平地应用现代科技。

2 “汉简数据库”的内涵

本项目的目标是：在搜集、整理汉简资料成果的基础上，全面、系统、准确地清理汉简原始资源，在工程总体设计的统一平台下，运用计算机技术，对汉简文字进行全方位的整理，建成完善的汉简文字字形表、属性数据库和资源库，并提供汉简文字编码方案[4]。具体工作内容如下：

本项目采集材料的边界为两汉、吴、魏、晋简牍及墨书材料，不包含马王堆汉墓出土的帛书文字材料。墨书材料指除简牍文字材料之外的，书写在陶器、漆器、墙壁、砖石等载体之上的两汉、吴、魏、晋文字材料。

根据本项目的目标和任务应交付的成果物分别为：建立汉简文字图档资料库、汉简文字原形总表提交合乎要求的图片格式的资料库，可根据工程要求按照年代、材料载体等属性建立不同的子资料库；汉简文字释文文本库、汉简文字字形总表、汉简文字隶定字形表、字库制作所需的字稿可提交相应的文本文档库；汉简文字原形字属性数据库可根据工程总体组设计的采集平台，提交相应的数据库原始资料[5]。

本项目需采集的资源数量预估为：两汉吴魏晋简牍 20 万支，其中大约 12 万支不清晰。在这 12 万支中，大约有 10 万余支需要对图像进行处理，5 000 余支需要目验，3 000 余支需要获取底片，2 000 余支需要重新拍照。以每支简牍平均 15 字计算，最终能够有效提交单字材料在 200 万字左右。墨书材料大约在 800 件左右，以平均每件 30 字计算，需处理的单字材料在 2.4 万字左右。汉简文字字头约为 5 000 个，其中有 30％包含需要单列字头的异体字，这样总的字头数约为 6 500 个（5 000＋5 000＊30％＝6 500）。以每个字头平均收录 10 个单列字头的异体字计算，提交字库工程总的图片格式的汉简文字原形字大约在 65 000 个左右。

需编码的字形制成各种字型表，根据工程的标准来定型，再由“中华字库”厂商制作成精品字库，再按国际标准化组织的要求，研制出不同种类的文字编码方案，提交给国家相应机构，申请纳入 ISO/IEC10646 国际标准[6]。

3 “汉简数据库”框架

本包项目共有“汉简资料调查与搜集”、“汉简图像采集”、“汉简释文”、“汉简文字整理”、“汉简文字筛选与数据加工”、“汉简字稿制作”等任务，这些研发任务互相关联、交叉推进，见图 1。

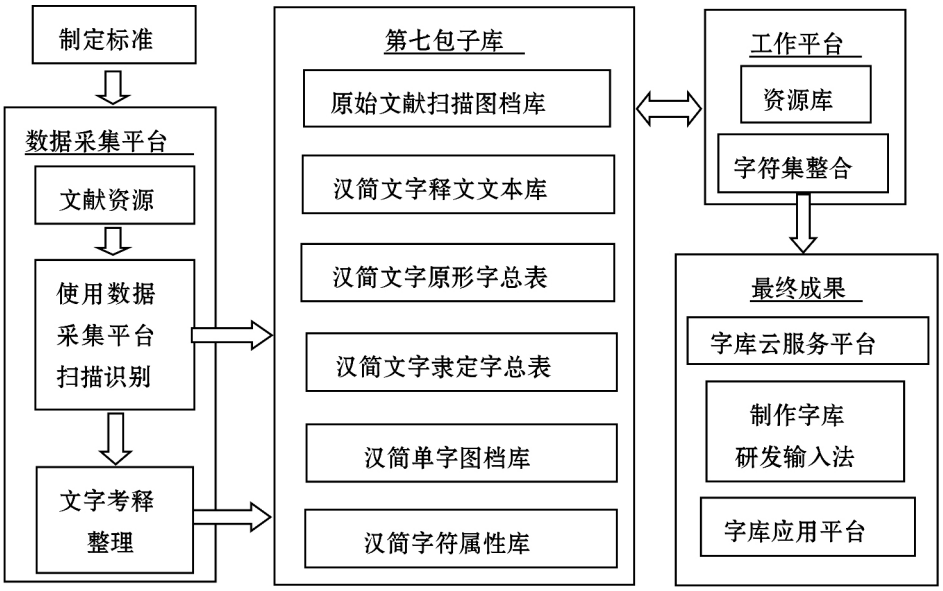

图 1 “汉简数据库”基本框架

3.1 “汉简”图录的收集、整理

本课题首先调查每一批汉简的著录情况，尽力搜集各种图录，包括两汉魏晋简的发掘报告，以及各种两汉魏晋简牍图录，以便于相互比较、择取。

3.2 汉简收藏单位的实地调查与高质量实物照片的拍摄

至今已发现汉代简牍 164 批、帛书 2 批、壁书 1 批，另外还有陶器、壁画上的墨书文字。目前已完全发表简牍材料的仅有 93 批，还有 71 批简牍没有发表。

目前已经公布的两汉魏晋简资料，有很多图版质量不高。具体情况大约有两种：一是原简牍实物的保存状态好、字迹清晰，但因当时拍摄、印刷技术水平不理想，导致图版不够清晰。二是原简牍的保存状态不好，出土后不久即已经碳化或破碎，字迹无法分辨。本项目将通过与各批汉简的收藏机构进行合作，以便取得原始的照片（底片），并尽可能地利用新购买的红外设备拍摄照片或扫描，以取得清晰的图像资料。

3.3 制定汉简文字的整理规范和标准

汉简文字整理规范和标准主要包括：

(1) 字形选择的原则，是规定从汉简文字中选取进入汉简文字原形字形总表的字形的操作原则。例如是尽量将汉简中的文字全部收入字库还是有选择地收录，对于模糊字形的处理方式是通过相关软件修描还是保持原样等。

(2) 属性标注原则及规范，是对收入汉简文字原形总表的字进行属性标注，明确规定需要标注的属性项（如汉简名、篇名、简号或行数、著录情况、所选照片的出处、出土年、出土地、现藏地或收藏者、汉简文字原形在照片上的坐标位置、字音、部首、对应隶定字形及释文等）以及每个属性项的标注规范。

(3) 订立部首和归部原则，是制定对提取出的汉简文字原形分类时所要遵循的原则。

以上三个标准由总体组确定。

需要各古文字分包协商确定的整理标准有：

(1) 古文字学界一般采用《说文》部首为汉简文字字形订立部首，并明确指定《说文》未收字的归部原则。如根据汉简文字本身特点制定合适的部首表或部首系统，则需要研制立部原则和归部原则，并充分考虑如何与其他类型古文字部首表或部首系统相适应、衔接的问题。

(2) 除与后世隶楷字形相同的字之外，隶定字形有两种情况，一是可释但写法与后代隶楷写法有异的字的隶定，二是不可释之字的隶定。完全不能隶定的字，只收入原字形表。

(3) 排序原则是对经过分部归部之后，对部首的排列顺序及同一部首内的文字排列顺序指定明确的原则，以便于排序和检索。

在遵循总体组以及各古文字分包确定的整理规范与标准的情况下，本课题对“异体”概念的界定，比学界一般的理解要宽泛得多。除了字的大小、笔势外，其他方面有差异的字形都要看作异体。同一个字只要笔画有区别，都算是异体，均予以收录。个别笔势不同的也作为异体看待（如竖笔很长的“年”、“令”等字）。

3.4 建立汉简文字字形数字化图档库

甄别与鉴定已出版资料中最清晰的图片，与收藏单位协商重新拍摄，获取已公开和未发表简牍的清晰图版，想法获取原物已经毁损无法重新拍照的简牍原始拍照图像或原始摹本。

把确定为提取汉简文字原形的照片（图版）按照总体组研制的统一标准全部扫描成电子图档，并用适当的方式进行编目及目录管理，建立汉简文字字形数字化图档库。

3.5 建立汉简文字释文文本库

对已公布的汉简，选择比较权威的释文本作为工作底本，根据已有汉简字形考释、语词释义及相关研究成果，对照图版，同时利用项目组在文字考释方面的优势，对疑难字做出考订，以此对工作底本做出全面的清理和校订。对于未公布的汉简，则做出新的释文本。在此基础上，参照简牍形制、内容等关联信息确定成组汉简的分篇和编联。

在收集已有汉简文字释文文本的基础上，充分吸收学术界最新研究成果，研制释文字形处理规范标准，并按统一的标准建立与汉简文字原形数字化图档库对应的汉简文字释文文本库，并实现两者的关联。

3.6 编制汉简文字原形字总表

首先在提取出来的字图中，选取符合本项目要求的全部字形；然后根据本项目指定的归部原则和排序原则对入选的字进行排序，形成汉简文字原形总表。

3.7 编制汉简文字隶定字总表

将选定的汉简研究资料扫描成图档，从图档中提取单字字图，建立汉简文字隶定字形总表及属性数据库，并与汉简文字字形属性数据库进行关联。对于已有考释的字，选取一个最合理的意见进行收录，不能选出合理释法的字，应在已有基础中选择一个较为合适的在隶定属性中加以说明。

3.8 提取汉简文字单字的图档

利用计算机工具对入选的汉简图像上的文字提取单字字图，提取单字字图的技术标准按总体组的统一规定执行。切割并提取字图时要给每个字图提供坐标位数据，以便准确定位及版式还原。并给每个字图按照统一的设计分配一个唯一的标示码，便于关联检索及后续处理。

3.9 建立汉简文字字形数据库

依据本项目研究确定的属性项及属性标注规范进行属性标注，建立汉简文字字形数据库。

3.10 全方位的属性数据库制作

完备搜集已经出现过全部汉简牍文字原形字和隶定字的基础上，编制与之配套的数据和全方位属性库。

这些课题所要进行的工作大致可分为三个层面：(1) 尽可能完整、清晰地取得并处理简牍资料的图像记录；(2) 尽可能全面、精确地复原和解读简牍文本；(3) 尽可能规范、系统地制作简牍文字数据库、原形字表和隶定字表，并形成字稿。图像记录是工作的基础，数据库、字表制作和完成字稿是最终目标，文本复原和研究既是重要的阶段性成果，又是实现最终目标的必要准备。不过，这三者并非绝对分作前后衔接的三个阶段。资料的调查、采集和拍摄工作主要集中于前期，但在此后也会根据课题发展的需要，穿插进行；文本考定与数据库、字表制作之间更将保持有效的协同、互动，从而在这两方面都达到最佳效果。释文的整理、字表的研制既是基础性的研究工作，也是对前期工作的总结，是课题的阶段性成果之一。而后期的单字整理是配合整个工程综合数字化开发的桥梁，通过这些工作可以将前期工作所收集到的图片信息完成规范化的分类、命名，从而架起与字表、数据库之间的关联。

4 “汉简数据库”的建设

4.1 汉简牍资料调查、采集与处理

本课题的工作主要是完成 93 批汉简资料的调查、采集与图像处理工作，即确定全部汉简牍的发掘、发现地、遗存情况及其收藏、保护单位和现存状况；完成全部汉简牍资料的图版、照片的采集工作；尽可能多进行汉简牍的红外拍摄和扫描；对采集到的图版、照片进行规范的数字化处理。

本课题拟采用的技术路线核心是以红外拍照与红外扫描为主，对已有照片或底片、出版物中的彩色或黑白图版进行数字化处理，辅以常规数码拍照及扫描等各类先进数字化技术手段在简牍文字搜集、整理中的综合使用；使用相应计算机软件、程序，完成简牍文字资料基础处理（优化、切割、命名、分类）和系统整合（包含归档、排序、关联）。

项目组将在“中华字库”工程提供的工作平台和数据采集平台上展开工作，同时注意更新整个项目中使用相应计算机程序、软件的版本，以期获得更新的技术支撑。

在具体操作过程中，课题组还将特别注意原始资料和研究文献的充分占有。原始资料是指简牍出土状况、形态特征与书写其上的文字符号。前两项内容，尽量查阅发掘记录、考古报告，并对收藏的简牍实物进行观测。文字符号信息，在尽可能收集先前照片、底片的同时，花大力气获取红外图像。研究文献除国内发表之外，还将大力收集、消化在国外发表的文献，把国际学者的研究成果吸收进来。

本课题的主要工作是获取简牍图像，这是整个项目的工作基础。通过与简牍收藏机构的密切合作，最大限度地获取最高质量的简牍图像，将为简牍文字辨识与文本复原尽量准确、可靠提供保障，也为原形字表及其属性数据与字稿制作提供优良素材。

本课题工作将先期展开，以能为其他课题提供资料。其他课题在实施期间，发生简牍图像方面的疑问或需求，本课题将尽可能补充获取简牍图像。在项目实施期间新发表的汉简牍资料，本课题也尽量获取高质量图像，以修订、扩充和完善数据，提高本项目的结项质量。

项目可望在下列方面取得创新性成果：完成迄今为止最全面的汉简牍文字资料的搜集和整理，获得一套最完整、最明晰的数字图像记录。

4.2 汉简牍释文整理与研究

本课题主要完成 93 批汉简牍资料的释文编写体例确定、海内外研究成果的收集、撰定汉简释文以及释文文本数据库的编订等工作。具体来说，是对已公布的汉简，选择比较权威的释文文本作为工作底本，根据已有汉简字形考释、语词释义及相关研究成果和各种简牍图版，发挥项目组在文字考释方面的优势，对疑难字做出考订，从而对工作底本在释文、断读、编联等方面做出全面修订，形成代表学界最新进展的更加准确、可靠的释文文本。在此基础上，编制尽可能全面、准确的释文文本资料库。

汉简字形隶定的难点和症结在于存在数量众多的俗体字和异体字。而本项目组成员多为古文字学者，长于疑难字形的考释，可以最大限度保证汉简释读的准确性。同时，图像采集等相关课题组能够提供质量更高的汉简图版，可以看到更清晰的汉简单字图像，为准确释文提供了重要保障。

本课题的主要工作是整个分包的重要阶段性成果，又是实现最终目标的必要准备。只有对简牍文字原形作出正确的释读，对简牍文本作出正确复原，原形字表及其属性数据的制定和字稿确定，才具有正确的学术基础。文本复原、探讨与数据库、字表制作之间将保持有效的协同、互动，从而在这两方面都达到最佳效果。文本的复原也与字表的制作密切相关，在文本复原的基础之上，才能确定原形字头，进而编辑字形表。同时，在字形表的制

作过程中的新发现、新成果，也可以对文本复原的成果加以验证和修订。

全面、准确的汉简牍释文文本不仅是本项目中的重要一环，同时也可以为相关领域提供支持：新的汉简释文可以为历史学家提供一个精确的释文文本；为语言学家提供一个出土汉简文献的语料库。

4.3 汉简牍原形字表编制

本课题拟以古文字方面学术研究与先进的科学技术相结合，对两汉魏晋简牍资料进行科学整理，利用课题组成员专业知识、字表编纂经验，科学设计并制作汉简文字字形表。

本课题将在“汉简资料调查与搜集”、“汉简释文整理与研究”这两个课题组的支持下，克服汉简照片（图版）或欠清晰、释文意见或有歧异的问题，在更清晰照片和更准确的释文文本基础上开展工作。对于字头的排列，课题也将在已有各种方法的基础上，推敲出更加科学、合理的方案。

课题可望在下列方面取得创新性成果：对汉简文字原形字形进行全面系统的搜集和整理，建立起完备的汉简文字原形字形表和数据库。

4.4 汉简隶定字表编制

本课题主要完成汉简资料中所出现的隶定字形的确定、收取范围、排序原则、字表编制等工作。这是整个分包中基础性研究工作的一个重要方面，主要工作在这个项目的中期展开，在图片采集、处理以及释文写定的基础上，收集已经出现的全部隶定字形资料，并依照一定的原则排序，将其纳入原形字头之下，最后制作成隶定字总表，这是项目的最终成果之一。

本课题将编制完整的两汉简牍隶定字表，为古文字学家提供一个完备的字形表，为字库编码技术人员提供一份汉简文字规范字表。

4.5 汉简单字整理

本课题主要完成简牍资料全部单字图片的切割、命名、分类与储存，全方位的属性数据库制作等。该课题是整个分包的后期综合性工作，在完备搜集已经出现过全部两汉魏晋简牍文字原形字和隶定字的基础上，完成与之配套的数据资料库，同时与相关分包沟通、合作，研制出适合先秦、秦汉文字特点的编码方案和码表，实现整个分包的最终目标。

文字整理的各包成果最终都要进入到字符集整合里。标准研发的成果既为字符集整合服务，很大一部分也要通过字符集整合来实现和验证。系统总集成是对采集平台、工作平台、应用平台、输入法等技术开发包的集成，包括字库制作，既要提供给各文字包来用，也要最后应用于字符集整合。

5 “汉简数据库”的系统使用功能

“中华字库”包括95款汉字经典字库，一款汉字字符库。“汉简”工程采用轮廓的方式来描述字体，造字采用高曲线度方式，造出的字编制成国际编码标准提案，多数收录在汉字字符库中。字库适于各种精密度的输出设备。安装在Windows系统中，支持各种中文应用软件。字典字库的编码为GB－2312，符合国家标准，内建造字系统，利用系统强大的图形功能，可以随时造出您所需要的任何文字[7]。它不是一个简单的电脑字库，实际是一个凝聚着汉语文字学家、历史文献学家、计算机网络工作者及其它专业人员研究成果的文字资源数据库。

6 结语

“汉简”作为特定历史时期的产物，不但具有科学价值，而且具有其历史价值和艺术价值。“中华字库”工程项目囊括了以往的字库成果和迄今为止的汉语研究成果，吸引了全世界汉语信息使用者的目光。

“汉简数据库”建设，本文只是进行框架性的宏观研究，它涉及文字学、文献学、信息处理技术、计算机技术、图像识别技术、数据库技术、网络通信技术、标准化等文理科多个领域，有很多课题亟待相关学者深入研究。

注释

[1]吉林大学社会科学处．“中华字库”工程第七包研发项目启动[EB/OL]．[2011－11－22]．https：//www.sinoss.net/2011/1122/37824.html

[2]吴晶．保护少数民族语言文字走上科学道路[N]．中国民族报，2012－06－01：1.

[3]国家新闻出版重大科技工程项目介绍—“中华字库”工程第七包“两汉、吴、魏、晋简牍文字的搜集与整理”[J]．吉林大学社会科学学报，2012（1）：封二．

（下转第84页）

区域异质性图书馆联盟应对联盟内各成员馆现有数据资源进行整合，依据成员馆特色馆藏资源建立个性化服务平台，满足不同专业特征、研究兴趣和研究课题的用户的个性化的资源需求，以个性化的服务方式、数据信息资源和定制化的初级研究成果提供个性化的信息服务。

同时，区域异质性图书馆联盟也不应该忘记自己肩负的普及文化的使命，不断开展读书活动、展览、讲座和技能培训等大众化服务，以满足人民群众更高层次文化生活的需求。

3 结语

区域异质性图书馆联盟是一定区域内经济、文化、教育发展的需要，是图书馆社会化服务方式的转变，是一项造福一方的公益事业。联盟在不断挖掘服务方式的同时，要积极深化合作共建、资源共享的目标，实现“互联网＋”环境下区域异质性联盟间合作，促进联盟共同发展。

注释

[1]山东省图书馆[EB/OL].[2018－03－25].http://first.sdlib.superlib.net/admin.

[2]山东省文献信息服务中心（CALIS）[EB/OL].[2018－03－25].http://uas.sd.calis.edu.cn:8090/.

[3]孟祥凤．图书馆联盟数字资源共享平台架构研究[J].图书馆界，2015（2）：6.

参考文献

1 张甫，吴新年，张红丽．国内区域图书馆联盟建设与发展研究[J].情报杂志，2011（8）：138－143.

2 杨丽．广东省图书馆联盟建设与服务调查分析[J].情报探索，2016（2）：46－49.

3 吴玉灵．跨地区、跨系统图书馆联盟建设策略研究——以江西省为例[J].山东图书馆学刊，2016（5）：47.

4 欧阳剑．数字图书馆信息资源的可发现性研究[J].图书馆论坛，2013（1）：32－37.

5 朱前东，高波．德国的图书馆信息资源共享模式[J].大学图书馆学报，2008（5）：43－48.

6 朱俊波．我国中小城市区域图书馆联盟建设研究[J].图书馆建设，2011（1）：85－88.

7 李力文，贾睿，李广生．对天津市高校图书馆文献资源共建共享的一些建议[J].图书馆工作与研究，2008（12）：47－49.

8 赵晖．鉴于“上海协作网模式”的我国跨系统图书馆联盟发展模式的优化问题研究[J].图书情报工作，2010（23）：69－73.

9 李锦兰．国内区域图书馆联盟研究——基于31个省级图书馆联盟网站的调研[J].图书馆工作与研究，2017（3）：56－61.

10 王泽琪，王代礼．我国异质性图书馆联盟建设现状与发展对策[J].图书馆学刊，2014（10）：29－31，35.

柳 霞 山东社会科学院省情与社会发展研究院，研究馆员。

（上接第50页）

[4]于晓海．吉林大学获首个超千万元资助哲社项目[N].中国社会科学报，2011－03－22：5.

[5]新闻出版重大科技工程项目“中华字库”工程申报指南[EB/OL].[2016－11－02].http://http://www.docin.com/p－598877033.html.

[6]张翼飞．古籍数字化中的字符集问题与解决方案[J].出版发行研究，2016（3）：79.

[7]西西软件园整理．中华字库大全介绍[EB/OL].[2011－06－22].http://www.cr173.com/html/12127_1.html.

孟 忻 吉林大学古籍研究所资料室副研究馆员。研究方向：信息组织与检索。