人文新知

“中国历代人物传记资料库”(CBDB)的历史、方法与未来

包弼德　王宏苏　傅君劢　陈　松　柳　舟　朱厚权

摘　要　“中国历代人物传记资料库”(CBDB)项目始于2005年,其目标是通过创建在线与离线的关系型数据库,记录史料中保存下来的历史人物的职业、亲属关系、社会关系等数据。CBDB大量使用计算机技术进行数据收集与管理,以便学者对人物、人群、地域、职官等多方面数据进行交叉分析。利用这一资料库,学者可以进行数据可视化并开展统计分析、网络分析以及空间分析。CBDB是一个面向多方合作的项目,它没有截止日期。它将持续收集各个历史时期的各种数据。通过诸方合作,一个全新的“中国历代人物传记资料库”在线系统即将向公众开放。

关 键 词　关系型数据库;　群体传记学;　数据挖掘;　空间分析与网络分析;　项目可持续发展;　关联数据

分 类 号　K23

作者简介　包弼德(Peter K. Bol)(通讯作者),哈佛大学东亚语言与文明系Charles H. Carswell讲席教授,Email:peter-bol@harvard.edu;　王宏苏,“中国历代人物传记资料库”高级项目经理;　傅君劢(Michael A. Fuller),加利福尼亚大学尔湾分校东亚语言与文学系教授;　陈松,巴克内尔大学东亚研究系副教授;　柳舟,北京大学《儒藏》编纂与研究中心博士研究生;　朱厚权,北京元引科技有限公司创始人,曾任中文在线数字出版集团股份有限公司研发中心总经理兼教育集团副总裁。

0　导言

本文是对“中国历代人物传记资料库”(以下简称“CBDB”)项目的介绍。“导言”部分将介绍CBDB的历史和发展目标。鉴于CBDB中的数据尤其适合群体传记学研究,CBDB的首位项目经理、专长为宋代社会史的陈松将在本文的第二部分介绍这一史学方法的来龙去脉和主要学术成果。CBDB数据库架构的设计者、宋代文学专家傅君劢(Michael A. Fuller)随后将解释为何使用“关系型数据库”为历史人物的生平建模。CBDB团队使用计算机程序从电子文献中挖掘数据,目前采用了包括神经网络模型在内的多种数据挖掘技术,研究先秦哲学的柳舟将对此进行具体说明,包括使用CBDB数据进行研究的三种分析方法:统计分析、网络分析和空间分析。高级项目经理王宏苏将介绍新开发的各种在线系统,以及CBDB与其他在线系统的多种交互方式。在中国,CBDB团队与“中文在线”建立了合作伙伴关系。正如朱厚权所述,作为CBDB商业化分支的“引得”系统将实现CBDB数据、历史文献和可视化图表的有机整合。文章最后以本团队对CBDB未来的思考作结。

CBDB项目源自郝若贝(Robert Hartwell,1932–1996)的工作。郝若贝是专攻宋史的社会史学家。他在遗嘱中将其在研究工作中建立的传记资料库赠予了哈佛燕京学社,这一资料库即为CBDB之前身。CBDB项目

于2005年正式启动,目前由哈佛大学费正清中国研究中心、台湾中研院与北京大学中国古代史研究中心共同负责建设。自启动以来,许多基金会为本项目提供了资助,很多数据库的开发者分享了他们的数据,而众多来自中国的历史学家和文学研究专家也为CBDB的开发贡献了力量。① 今天,CBDB的开发与维护正从依赖基金会资助转向以商业化推广获得新的资金来源。

CBDB的目标是将历史文献中出现过的每个人物的生平信息全面收集汇总,再以适合大批量查询和分析的数据格式将这些信息提供给研究人员。为实现这一目标,CBDB将传记资料存储于一个数据库中,而研究人员可以通过设定诸如时间、地点、人物关系之类的各种变量对数据库执行查询。CBDB收录了传记史料中经常出现的各类人物的生平信息,如姓名、籍贯、生卒年、入仕途径、职官、亲属关系、社会关系、平生著述。为了便于查询,CBDB在对这些传记资料进行编码后将其存储于数据表中,譬如,CBDB对一个人的名讳、表字、行第、室名、别号、谥号等不同名号加以区分并分别编码。关于社会关系的历史记载则更加复杂,为此,CBDB使用了近250种代码来区分不同类型的社会关系。编码和存储工作完成后,用户就可以便捷地对这些资料进行查询了。用户可以执行一个简单的查询,比如在数据库中查看哪些人物曾登进士第;也可以创建一个涉及多张数据表的复杂查询,比如先在数据库中查找在特定地方、特定时段登进士第的所有人物,然后在此基础上进一步查询其中多少人具有两代以内的亲属关系。

无论是数据收集方式还是研究用途,CBDB都与传统的传记词典截然不同。CBDB开发团队欢迎对特定历史人物进行个案研究的学者分享他们的数据,但是团队本身并不会对任何历史人物进行细致入微的考证。CBDB的目的在于大规模地汇总传记资料,因此数据的收集方式是系统性的,主要利用那些按照明确一致的编排体例登载人物生平信息的文献资料(如传记资料索引、地方志、缙绅录等),这些史料的编纂体例为使用计算机程序高效地从历史文献中提取传记数据提供了便利。随着项目的推进,CBDB中每个历史人物的相关数据量自然而然地不断增加,CBDB是一个开放式项目。尽管它对于特定历史人物的个案研究亦有参考价值,但是就初衷而言,CBDB的设计是为了让用户可以一次性研究成千上万个历史人物。当研究对象涵盖数千位历史人物时,数据库中偶尔出现的舛误便不会影响就宏观历史图景所做出的结论。

在CBDB的工作流程中,团队面临着众多挑战。首先,随着数据量的增长,同名同姓的历史人物与日俱增(譬如,目前CBDB收录有54位名为"王佐"的人)。如果知道每个人的生卒年、籍贯和字号,那么人物消歧就不成问题。可是,从一开始就掌握判定某个人物身份所需的全部数据是很少有的情况,在无法判定的情况下只能将同名同姓但身份不明的人暂时作为不同人物录入数据库。随着数据的不断完善,发现和删除重复录入的人物就成了团队从不间断的工作任务。其次,CBDB的大部分数据都是关于唐代至清代的,如果希望把数据覆盖面系统性地延伸到唐代以前,将需要整理更多信息,以扩充职官代码表和行政区划代码表。如果希望把20世纪的历史人物也收入数据库,则还需要创建全新的代码表,例如为所有的大学、高中、报社、出版机构等编制代码。再次,团队尚未对家谱中的数据进行系统性的提取和录入。家谱中保存的传记资料,可靠程度参差不齐,如果要利用家谱来扩充CBDB的数据量,则势必要对相关材料的来源和流传情况详加考订。最后,团队还需要开发一个众包平台,为那些希望为CBDB贡献才智的用户提供一个便捷的窗口。

1 群体传记学

通过系统地收集和编排大量历史人物的传记资料(截至2020年5月,CBDB总共收录了约47万人的传记资料),CBDB为群体传记学研究(Prosopography)提供了一个数据宝库。群体传记学将历史学家的关注点从制度史、事件史和思想观念史引向历史人群。它鼓励历史学家通过对历史人群的研究来回答政治活动、官僚体制的运作实态、社会结构以及宗教和文化嬗变等方面的问题。

英国历史学家劳伦斯·斯通(Lawrence Stone)对群体传记学下过一个经典定义:群体传记学就是要"通过

① 本项目的详细信息参见:https://projects.iq.harvard.edu/chinesecbdb/關於我們。

对一群历史人物的生平进行综合研究,进而发掘出他们共同的背景特征”。[1]由此可见,群体传记学与历史传记写作在研究对象和方法上有着显著的不同。历史传记力求对特定人物的独特的生平经历面面俱到地加以描述,并经常对传主的内在特征和特殊品质(如动机、性格等)予以剖析;而群体传记学所关注的则是“或大或小的一个历史人群在其生平经历中具有一般性、普遍性和共性的那些特征”。[2-4]群体传记学的研究方法是归纳法。群体传记学家首先需要界定其研究范围,亦即所要研究的目标人群;其次要就这一目标人群的一些可观测特征(如生卒年份、家庭背景、教育背景、经济状况、职业、宗教信仰等)提出一系列整齐划一的问题;最后,需要就这些问题有系统地进行数据收集,并根据这些数据对目标人群的相关特征加以概括。

现代语义上的“群体传记学”一词最早出现于 1897 年出版的《罗马帝国人物志》(*Prosopographia Imperii Romani*)一书的书名中。20 世纪 60 年代,群体传记学已经吸引了从事欧洲中世纪史、现代史和当代史研究的诸多学者。与此同时,群体传记学的研究范围也大大扩展,关注的对象从最初的权力精英扩大到妇女、中下阶层乃至社会边缘人群。群体传记学的发展是史学领域中两种学术潮流互相激荡的结果:历史学家们一方面试图在正式的制度安排和意识形态之外寻找更具说服力的历史解释,另一方面也越来越多地借用社会学和人类学发展出来的各种概念和方法对历史现象加以分析。[1]

尽管“群体传记学”在汉语中没有对应的概念,但是 20 世纪以来许多中国史学者所采用的研究方法显然与欧洲史学界的群体传记学传统遥相呼应。不论是私家著述还是官修史书,传记写作在中国都有着悠久的历史传统。[5]95-114今日所见的各类史料中保存了大约 60 万篇人物传记[6]148-157,充分利用这些史料进行群体传记学研究对中国史学的发展具有重要意义。20 世纪以来,海内外许多历史学家利用这些史料探讨了诸如明清时期精英阶层的构成与社会流动等一系列重大历史问题。这里特别值得一提的是郝若贝于 1982 年发表的一篇具有开创性意义的长文[7],其中一节对宋代宰执和财政官员的社会身份和出身地域进行了详尽的分析。这篇论文引发了一场关于宋代政治精英的热烈讨论,时至今日,这场讨论仍在为唐宋社会史研究注入活力。[8]郝若贝的这项研究在方法论上的贡献尤需注意。首先,他大量使用墓志史料,有系统地从中收集传记数据,并通过创建以人物为核心的数据库对这些资料加以编排和整理[9]89-121,而这个数据库正是 CBDB 的前身。其次,郝若贝在分析传记数据时,不再满足于此前群体传记学家常用的简单的统计归纳法,而是将空间分析引入其中,在施坚雅(G. William Skinner)的地理宏区理论框架中探讨宋代官僚的社会身份与出身地域。除此之外,郝若贝还研究了宋代官僚家庭的婚姻关系,尽管他未能对其婚姻网络的结构特征加以分析。郝若贝的上述工作昭示着以利用计算机技术进行数据收集、整理和加工为特点,高度关注空间分析和网络分析的史学研究新篇章正在群体传记学领域开启。

在欧洲史研究领域,群体传记学也呈现出同样的发展态势。21 世纪初,拜占庭史学家迪翁·斯迈思(Dion Smythe)将群体传记学研究中的新趋势与传统的群体传记学方法进行比较,勾勒出了新式群体传记学的若干突出特点。在他看来,“旧式”群体传记学探讨的是一个历史群体的外在特征(如教育背景或所任官职),而“新式”群体传记学在此之外还关注某个群体中个体之间的相互关系,以及由这些关系构成的交错重叠的社会网络,因此,“新式”群体传记学大大扩展了群体传记学的研究范围,使其研究课题不再局限于寻找特定历史人群“共同的背景特征”。[10]177-178无论是“旧式”的还是“新式”的,群体传记学最大的魅力在于它所提供的分析框架和分析手段,将微观层面上对一个个历史行为人的分析与宏观层面上对政治、社会、文化结构及其嬗变过程的分析有机地结合了起来。使用群体传记学方法的历史学家的工作始于收集和分析历史上一个个人物的传记数据,但其最终目标则是通过对这些数据的归纳和概括,更深入地理解由这些人物所组成的特定历史人群的普遍特征及其人际网络的结构性特点。

在过去的十几年中,从网络视角出发的群体传记学研究在中国史领域大量涌现,加深了人们对中国历史上的政治、社会、文化、宗教精英的理解。谭凯(Nicolas Tackett)的研究表明,在 9 世纪中国占据主导地位的政治精英具有两个鲜明特征:其一是几乎无一例外地累世居住于长安或洛阳及其附近地区;其二是排他性的婚姻网络,这个婚姻网络以唐代的两京为中心,将至少五分之三的政治精英家族紧密地联结在一起。[11]107-145陈松对两宋郡守的抽样分析进一步揭示,直至 11 世纪中叶,中国政治精英的婚姻网络仍然表现出类似特征,如

至少五分之二的郡守因血缘和婚姻处于同一张庞大的亲属关系网下;不过,和9世纪相比,11世纪中叶中国政治精英的婚姻网络明显更具开放性,在这些互为亲戚的郡守中,尽管五分之二的人来自于开封、洛阳及其周边地区,但是换个角度看,这也意味着有半数以上的人其出身为首都圈以外的地方。[12]101-152 13世纪初的情况则迥然不同,这种以首都为中心的庞大的全国性婚姻网络已然解体,取而代之的是较小规模的区域性血亲和姻亲网络。不过,全国性婚姻网络的解体未必意味着精英阶层内部凝聚力的削弱。魏希德(Hilde De Weerdt)[13]325-394和包弼德[14]相继在他们的研究中指出,宋明两代的地方精英通过学术交往和文字酬唱,在广阔的地理范围内建立起了厚实的社会网络,并以此巩固了"士"的身份认同和内在凝聚力。以上这些研究只是冰山一角,近年来,越来越多的学者从网络的视角另辟蹊径,开展了大量群体传记学研究。这个日益壮大的研究群体中不仅有历史学家,还有众多的社会科学家和计算机科学家。他们的研究多姿多彩,涵盖了宋代党禁的性质、南宋馆阁官僚的仕履特征、中国佛教史上的关键人物、清代至20世纪中国官僚体制内的权力分配和政治恩庇等众多主题。[15-18]

上述这些研究很多在不同程度上利用了CBDB提供的数据。对于从事群体传记学和网络分析的学者而言,CBDB最大的吸引力在于其将分散的历史记录加以整合的能力。群体传记学和历史传记在选题和方法上有着显著的不同。一般说来,要想撰写一部出色的历史传记,历史学家只能选取那些在史料中记载翔实的杰出人物作为他们的研究对象。历史上的大多数人物,包括精英阶层的大多数人物均声名不彰、默默无闻,有关他们生平经历的历史记载零星且分散,群体传记学的优点正在于它能够将这些散见于大量历史文献的传记资料汇总起来,进而揭示特定历史人群的共同特征及其内部成员之间的相互关联。群体传记学的这一研究方法使CBDB这样的关系型数据库别具吸引力。以北宋仁宗年间的两位宰执官——欧阳修(1007-1072)和庞籍(988-1063)为例,此二人的关系其实颇为密切:欧阳修的长媳和庞籍的次媳都是吴待问的孙女,为堂姊妹。不过这层亲属关系鲜有史料提及,历史学家只能通过将数篇墓志铭中的只言片语拼凑起来才能发现这层关系(图1)。相比之下,这些资料被录入数据库之后,CBDB只需几秒钟就能发现这层关系以及欧阳修的数百个血亲姻亲。正因为像CBDB这样的关系型数据库具有将散见于史料的大量传记信息拼接在一起的强大能力,英国历史学家凯瑟琳·济慈-罗汉(Katharine Keats-Rohan)说,群体传记学和计算机实为"天作之合"[19]。①

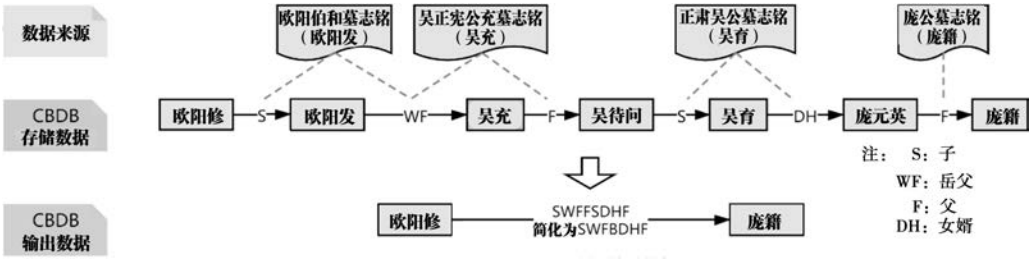

图1 CBDB发掘出的欧阳修、庞籍间的亲属关系

2 关系型数据库的价值

关系型数据库有一个显著特点,即对于想要研究的那批数据,它可以根据数据中隐含的结构特征建立数据模型,对数据进行结构化处理。由此得到的结构化数据包含了实体(entities)与实体关系(relations)两个部分。在对中国古代史进行建模时,CBDB的做法是先确定一系列跨越不同历史时段并具有关键性意义的实体(如人物、地址、社会关系结构、亲属关系结构、官僚机构、社会组织等),然后追踪这些实体之间的相互关系。

通过从实体和实体关系两个方面对传记资料进行结构化处理,CBDB使数据编排和数据分析变得既高效

① 20世纪80年代以来,在尝试了编排群体传记数据的不同方法后,很多学者认为关系型数据库乃是最合适的数据模型,对于大规模的群体传记学项目来说尤为如此。

又灵活。在此可以将 CBDB 采用的关系型数据结构和电子表格做一个简单的对比。电子表格由行、列两个维度组成,不同的列用来存储不同类别的数据,而每一行则是一条具体的数据记录。因此,电子表格中的每个单元格都提供了一条数据记录中关于某个特定数据类别的特定数据。这些单元格中有时候会被塞入大量数据,这时要想将这些数据与表格中的其他信息关联起来就变得非常棘手,表 1 便是如此。

表 1　　电子表格的数据编排方式(以司马光为例)

姓名	生卒年	任官	社会关系
司马光	1019 – 1086	(1) 1059 度支勾院; (2) 1085 门下侍郎; (3) 1086 左仆射兼门下侍郎; ……	(1)[元祐党];(2) 安惇[被 Y 陷害];(3)晁补之[祭文由 Y 所作];(4) 陈荐[为 Y 作祭文];(5) 陈敏[受 Y 之弟子礼];(6) 程颐[推荐];(7)丁度[为 Y 作祭文];(8) 范纯礼[是 Y 的恩主];……

要想让这些数据使用起来更加便捷,就必须严格遵循一个单元格中只存入一个数据的原则。也就是说,要将内容复杂的单元格分离出来,把它变成一张单独的数据表,用于记录实体之间的相互关系。在上例中,则需要另建一张数据表来记录人(即司马光)与官僚机构之间的关系,还需要再建一张数据表来记录人与人之间的社会关系;为了提升数据录入的效率并尽可能地避免讹误,可以用代码替换掉那些一再重复的人名(司马光、安惇、晁补之等)、官名(度支勾院等)以及对社会关系所做的文字描述(“祭文由 Y 所作”等),见图 2。

由于 CBDB 中的数据是按照关系型数据库的架构编排的,从事群体传记学研究的学者可以根据自己的研究需要,从任意一个角度切入并进而探索不同实体(人物、地址、职官等)之间的联动关系。又如人们可以利用 CBDB 探讨福建出身的官员是否比浙江出身的官员更偏爱与本地家族联姻、其婚姻模式是否依任官品级而有不同、是否随时间而发生变化等一系列问题。这些问题都涉及亲属关系、仕宦情况、地域差别、历史变迁等多个方面。换言之,它们探讨的乃是人物、职官、地址、亲属关系四个实体之间错综复杂的关系(下页图 3)。

通过把多张数据表串联起来,CBDB 可以方便地查找出一群人物的出身地域、血亲姻亲、所任职官及彼此间的社会交往等多方面的传记数据,并将这些数据导出到地理信息系统和社会网络分析软件中用于做进一步的分析。也就是说,正因为 CBDB 是一个关系型数据库,所以其数据可根据研究需要按照不同格式重新编排,既可输出为电子表格,也可以输出为 XML 或图形数据库文件。

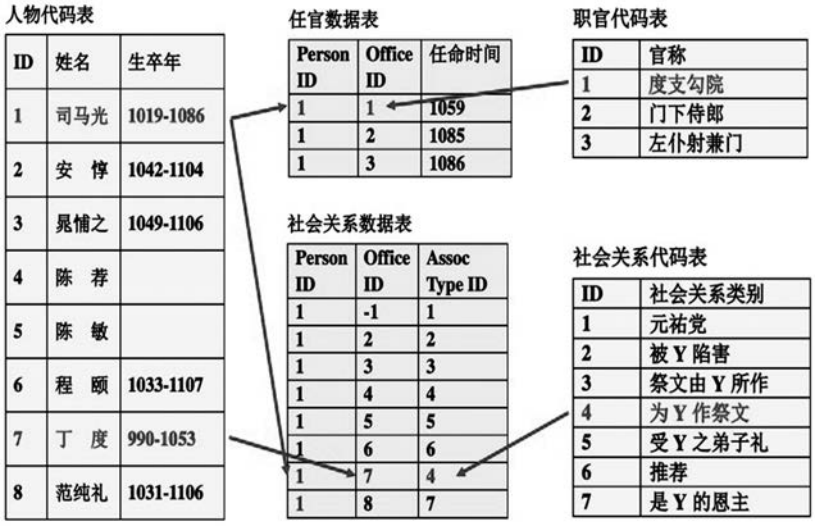

图 2　CBDB 数据编排方式示意图(以司马光为例)

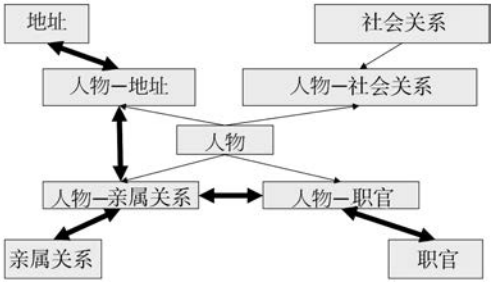

图3 CBDB中的实体关系模型示意图

3 从历史文献中提取数据

CBDB的数据来源主要是包含大量的、集中的人物生平信息的历史文献,例如传记资料索引、正史列传、墓志铭、地方志等综合性的传记资料,文学作品中涉及人物生平和关系往来的内容,例如祭文、序、记、书信等,以及官方文书中的年表、会要等人物资料。①

从数据收集的角度来说,这些文献的主要特征在于对人物生平信息的记录非常集中,这些信息通常在一个句子或者一个段落中序列性地出现。如光绪年间的《孝丰县志·职官志》中对于该地区某县令的记录——"郭治江西泰和举人嘉靖二年任陞嵩明知州见名宦",其中的姓名、籍贯、入仕方式、任职时间、之后的官职等信息正是CBDB希望收录的,而这些信息几乎连续地出现在一条记录中。这样一类信息的提取,是一个典型的序列标注的任务,在自然语言处理领域已经积累了大量有效的方法和模型,因此可以借助这些深度学习的方法实现数据收集的自动化。

进行模型训练的第一步是制作训练数据集,即人工标注一部分所需要的信息,提供神经网络的"学习样本"。在人工标注的过程中,尽量优先使用正则表达式来批量标注具有类似表达结构的信息,再通过纯人工的方法标注其他信息。

有了训练集以后,需要将训练数据向量化,再输入模型中。在自然语言处理领域,语言文字的向量化已经相对成熟,能够很好地使向量化的文字依旧表现其语义上的关联。相关预训练模型十分丰富,并不断取得新的进展,序列标注等各类经典任务的准确度得以提高。例如,在最近的地方志标注任务中,CBDB团队直接使用了BERT(Bidirectional Encoder Representations from Transformers)[21],将这一预训练模型的词向量结果作为输入特征。

在模型选择方面,目前,一个先进且成熟的用于序列标注的模型是Bi-LSTM-CRF(Bidirectional-Long Short Term Memory-Conditional Random Field)[22]。该模型在LSTM模型的基础上不仅增加了双向的学习,能够更好地记忆上下文的信息,同时在LSTM输出的预测基础之上增加了CRF(Conditional Ranclom Fields)层,更好地利用了句子整体层面的信息。在地方志标注任务中,CBDB团队选择将BERT的词向量输入到Bi-LSTM-CRF模型中,通过大量人工标注的训练数据让模型学习到某种标注的规律并最小化损失函数。训练好的模型在测试数据集上的F值能够达到80.06%,表现良好,由此能够继续自动地批量处理大量未标注的地方志文本,大大提高数据收集的效率。

CBDB团队一方面需要针对不同的历史文献,有针对性地训练不同的标注模型,使之更集中地学习该文献的语词和表述,提高标注的准确度;另一方面,则试图探究不同文本之间的迁移学习的方法及更少需要人工标注的半监督、无监督学习的方法,进一步提高数据收集的效率。

① CBDB的数据来源具体参见:https://projects.iq.harvard.edu/chinesecbdb/资料来源。

4 数据分析

文本挖掘技术的进步极大地提高了从历史文献中提取传记资料的效率,使 CBDB 的数据量迅速增长,为群体传记学研究开拓出广阔的前景。与此同时,随着各种简单易用的软件不断涌现,对大型数据集进行统计分析、网络分析和空间分析的技术门槛也不断降低,越来越多的学者正努力将 CBDB 为历史研究开拓出的前景化为现实。

4.1 统计分析

有些群体传记学研究的任务相对简单,研究者关注的只是某一特征在特定历史人群中的分布情况,因此使用描述性的统计学指标对相关数据进行综合概括就足够了。例如,人口史学家若想知道男女两性在不同历史时期的死亡年龄是否存在明显差异,只需使用 Excel 或类似的电子表格对死亡年龄这一变量在不同历史人群中的集中趋势(即平均值、中位数和众数)和离散程度(如标准差)进行计算即可。而在一些更加复杂的研究中,研究人员可能需要使用多变量分析(如相关分析和回归分析)甚至概率模型,方能了解目标人群中两个或多个特征之间的相互关系。譬如,一个人在家中的排行对他在科举中的表现是否有影响?一个人的殿试名次与他一生中所任的最高官职之间是否有关联?要回答这些问题,研究人员必须将 CBDB 数据导入专门的统计软件(如 SPSS 和 R)进行分析。利用这些软件,研究人员还可以绘制各种统计图表(如直方图、交叉表和散点图),以更直观的方式对数据的分布特征和相关程度予以呈现。

当然,图表本身并非研究结论。研究人员必须将数据置于相关的学术语境中,并审慎地对其学术意涵加以解读。一个严谨的学者必须充分了解 CBDB 的数据来源,因为原始文献中存在着的各种偏差都会不可避免地被带入 CBDB 中。不同文献中的偏差在 CBDB 中既可能相互抵消,也可能被进一步放大。

4.2 网络分析

统计分析旨在对特定历史人群的外在特征(或称属性)进行概括归纳,网络分析则提供了另一种思考历史的方法。网络分析关注的是由不同人物之间的相互联系构建起来的具有特定结构特征并能够为每个个体的行动赋予力量或施加约束的社会关系网络。

CBDB 收集了大量的亲属关系和社会关系数据,并为查询和输出这些数据提供了多种选项。用户可以查询任何两个人物(或任何两个群体)之间直接或间接的社会联系;也可以指定某个特定的历史人物(或者从别的查询中导入已生成的人物列表),然后由此出发进行循环搜索——先搜索指定人物的社会关系人,再进一步搜索这些关系人的社会关系人,以此类推;还可以通过设置各种参数来限定搜索的范围,比如只查找具有特定社会交往关系(如师生、书信往来)的人物或者只查找那些与特定地方有关的人物。一般建议用户不要把那些与指定人物隔了三四层中介的社会关系人纳入搜索范围,毕竟一个人的朋友的朋友的朋友对他来说恐怕与陌生人无甚区别。

在亲属关系查询中,用中介人的数量来限定搜索范围则没有太大的意义。比如,在搜索亲属关系时,CBDB 可能先找到指定人物的女儿,然后再找到这个女儿的哥哥。表面上看起来指定人物与他女儿的哥哥之间隔了一层中介(他的女儿),但是他女儿的哥哥即是他的儿子,他们之间的关系并不比他与女儿的关系更疏远。为了解决这个问题,CBDB 用英文字符对不同类型的亲属关系进行编码(例如,以“D”表示女儿,以“B”表示兄弟,以“S”表示儿子)。通过对这些英文字符的连缀和简化[如规定“DB”(女儿的兄弟)可以简化为“S”(儿子)],CBDB 得以对循环搜索中找到的任何两个人物之间的亲属关系进行准确的表达(参见图 1)。在此基础上,CBDB 还设计了一个四值度量法,从辈分、兄弟姊妹、婚姻多个维度表达亲属关系的亲疏远近。

在查询完成以后,CBDB 还为检索数据的输出提供了多种格式。从 CBDB 输出的亲属关系和社会关系数据可以直接导入 Gephi、Netdraw、Pajek、UCINET 等各种软件做进一步分析。Gephi 是一个特别受人文学者欢迎的网络数据可视化平台,它是开源软件并具有可扩展性(即用户可以通过自制的插件扩展其功能)。它绘制的社会网络图十分美观,同时也提供了一些基本的分析功能,比如用户可以使用不同算法(如 k - 核分析、模块度分析等)查找网络中内部联系密切的子群或使用不同指标(如中介中心度、亲近中心度和特征向量中心度等)评估每个行为人在网络结构中的重要程度。如果用户希望对 CBDB 输出的亲属关系和社会关系数据做更进一步分析,则应当考虑 UCINET 或 NetworkX。尽管 UCINET 不提供可视化选项,但是它的分析功能是迄今为止最为完善的。对 Python 或 R 语言有一定了解的用户,也可以考虑使用 NetworkX 和 iGraph 等开源程序包。这类程序包为分析复杂网络提供了很多经典算法,尤其适用于对超大型网络数据集的分析和制图。除此之外,从 CBDB 输出的亲属关系和社会关系数据中还有每个人物的地理坐标,因此用户也可以使用 ArcGIS 或 QGIS 等空间分析软件来探索亲属关系网和社会关系网的空间分布特征(图 4)。

4.3　空间分析

在社会网络分析中,对于研究者来说有意义的是实体之间的相互关系;而在空间分析中,有意义的则是事物在地球表面的空间分布。用户可以使用 ArcGIS 或 QGIS 等地理信息系统软件将 CBDB 数据投射到地图上。在几乎全部查询操作中,CBDB 都提供了与地理信息系统兼容的输出格式。用户可以选择以这种格式输出检索结果,然后将它直接导入地理信息系统软件中,再使用不同的符号样式加以呈现(如图 5 即是采用分级符号绘制的)。使用地理信息系统绘制的地图不同于传统的纸质地图,在地理信息系统中用户可以添加新的图层并根据自己的需要显示或隐藏已有图层。哈佛大学的“WorldMap”平台上就有很多使用地理信息系统绘制的有关中国的在线地图。①

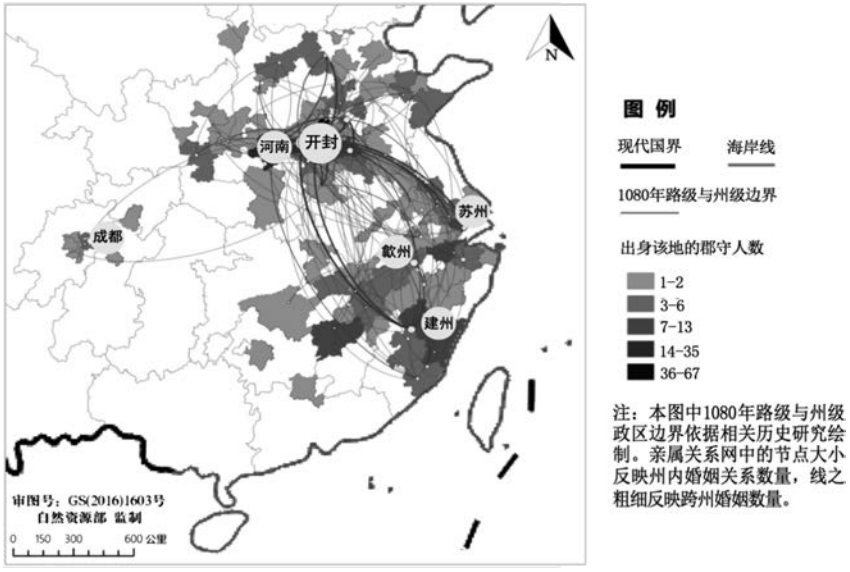

图 4　基于 CBDB 数据绘制的 1040 年代在任郡守籍贯与亲属网络空间分布图

① 如“ChinaMap”中的当代中国地图(https://worldmap.harvard.edu/chinamap/)及“ChinaXmap”6.0 版中两宋时期的中国地图(https://worldmap.harvard.edu/maps/7086)。

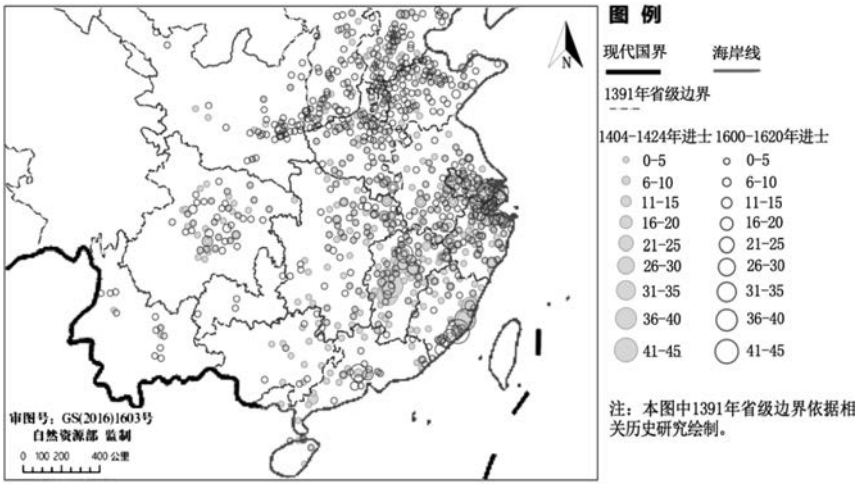

图 5　基于 CBDB 数据绘制的永乐年间与万历后期进士籍贯分布图

CBDB 中的空间数据既有关于历史人物的(如籍贯、任官地),也有关于社会机构的(如书院、寺庙的地理位置)。图 5 对比了永乐年间(1404—1424 年)和万历后期(1600—1620 年)进士的籍贯分布情况。绘制这样一张地图,除了需要从 CBDB 获取关于入仕途径的数据外,还需要从中国历史地理信息系统(CHGIS)官网和 CHGIS Dataverse 下载其他历史地理数据用于绘制底图。① 这些历史地理数据中特别值得一提的有模拟地形特征的数字地形图(DEM)②、施坚雅领衔开发的有关 19 世纪中国的各种空间数据③,以及复旦大学历史地理研究中心为哈佛—复旦中国历史地理信息系统项目开发的各种空间数据。最新发布的“中国历史地理信息系统”第 6 版④,收入了清嘉庆二十五年(1820 年)和宣统三年(1911 年)两个年度省、府、县三级政区治所和边界的时间切片数据,及反映公元前 221 年至公元 1911 年府县两级政区变化情况的时间序列数据。

5　全新的在线系统

5.1　新版在线录入系统

CBDB 新版在线录入系统⑤是重要的数据导入工具,由时为北京邮电大学博士研究生的傅群超于 2017 年开发并开源,当前由台湾中研院卢建安和王祥安维护和持续开发。在线录入系统最常用的场景是导入难以由程序批量处理的数据。通过对录入逻辑和录入值有效性检查的设计,在线录入系统可将用户在录入界面上输入的内容自动进行标准化,并导入 60 余张表中。

在线录入系统还是众包录入的重要工具。众包录入方法有两种:一种是登录后直接输入(仍在开发);另一种是通过 API 提交众包数据。这两种众包录入的内容都不会直接修改数据库。录入系统的专家用户可直接列出所有有待处理的众包记录(下页图 6)并对其审核,审核通过的数据将自动被写入在线录入系统的数据库。

① CHGIS 官网:http://chgis.fas.harvard.edu/;CHGIS Dataverse 网址:https://dataverse.harvard.edu/dataverse/chgis。

② https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E1FHML

③ https://dataverse.harvard.edu/dataverse/hrs

④ http://chgis.fas.harvard.edu/data/chgis/v6/

⑤ CBDB 在线录入系统访问网址:http://47.114.119.106:8000/basicinformation;开源地址:https://github.com/cbdb-project/cbdb-online-main-server。

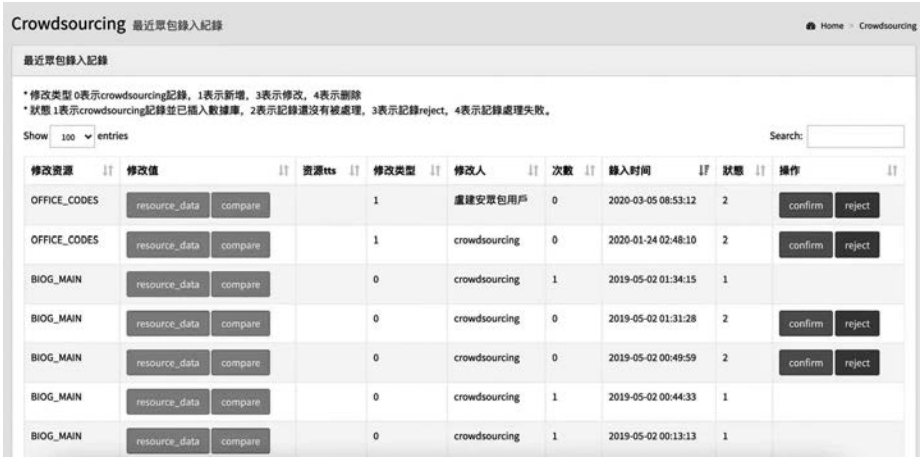

图 6　CBDB 在线录入系统中通过 API 进行众包输入的记录

5.2　新版在线查询系统

CBDB 提供了两种新版在线查询网页以方便用户使用。一是商业版①,由中文在线公司“引得”系统开发团队于 2018 年开发并维护。此系统为商业闭源,但开放注册使用,除批量下载查询结果需所在机构购买外,其他在线查询、可视化功能均免费使用。二是开源版②,由 CBDB 开源社区成员之一——北京大学信息管理系小组开发并维护。开源版本中相对激进地添加实验性功能,并欢迎任何用户通过 GitHub 的 issue 和 pullrequest 功能参与到开源社区中,提供功能性建议并向开源社区贡献代码。

5.3　API

“API”的中文全称是应用程序接口,它常用于项目间合作的场景中。CBDB 团队公开了 API 使用文档③,任何项目都可以在得到许可后,将 CBDB 的 API 引入自己的项目。当前与 CBDB 通过 API 合作的项目有“人名权威资料库”④、“明清妇女著作数字计划”⑤、“码库思古籍半自动标记平台”(MARKUS)⑥、“DocuSky 数位人文学术研究平台”⑦、“中国哲学书电子化计划”(CText)⑧。CBDB 不仅提供 API 供其他项目调取数据,调用 API 的开发者还可以通过可写入数据的 API 直接添加或修正 CBDB 数据源中的数据。

5.4　关联数据

CBDB 与上海图书馆合作开发了 CBDB 关联数据平台⑨。此平台中,CBDB 中的人物已与 DBpedia、VIAF、上海图书馆人名规范库进行了互联,并向用户提供基于人物的知识图谱检索及 SPARQL 查询。使用 SPARQL 查询时,用户不需要安装任何软件,就可以直接在网页平台上进行复杂条件的自定义查询(图 7)。

---

① http://www.inindex.cn/

② https://github.com/cbdb-project/cbdb-online-query-app

③ CBDB 可读写 API:https://github.com/cbdb-project/cbdb-online-main-server/blob/develop/API.md。以人名或 ID 查询人物所有资料的“重型” API:https://projects.iq.harvard.edu/cbdb/cbdb-api。

④ http://archive.ihp.sinica.edu.tw/ttsweb/html_name/

⑤ http://digital.library.mcgill.ca/mingqing/

⑥ https://dh.chinese-empires.eu/markus/beta/

⑦ https://docusky.org.tw/DocuSky/ds-01.home.html

⑧ https://ctext.org/

⑨ http://cbdb.library.sh.cn/spardled

5.5 开源社区

前一版在线系统于2008年由一家商业公司与CBDB团队合作建立,无论是技术、算法还是界面设计,在当时都是先进的。随着时间推移,数据结构、查询需求、技术都在发展,传统商业公司很难在十余年中以学术项目支付得起的价格来长期更新系统,因此CBDB团队希望新版在线系统由整个项目的开源社区共同维护。

CBDB的开源录入、查询系统基于GitHub。GitHub提交分支机制非常适合多人或多机构合作,并且GitHub有完善的版本管理机制,非常方便追溯历史版本。此外CBDB采用了持续集成(Continuous integration),此机制保证GitHub上的代码和生产环境的代码一致,没有任何不透明的开发行为。最后,GitHub拥有良好的贡献记录机制,所有向社区贡献代码者都是开源系统的参与者、所有者。团队希望CBDB是所有贡献者共同的项目。

图7 CBDB、上海图书馆关联数据平台上运用SPARQL上进行查询示例

注:此处查询的是登记为唐代剑南道及其所有下属地址的人物最常与哪些地方的人物交往。

6 未来计划

CBDB并不设定项目的结束期限,这个被期许永不止息的项目的发展目标是不断增加历史数据。当前,项目将产出大量数据,分别来自对地方志和清代朱卷的数据挖掘,及对《缙绅录》数据集的合并。随着更多墓志资料的电子化,有望从中获取大量亲属和职官数据,如已完成的唐墓志数据采集。CBDB团队也将从明清诗文集中提取更多社会关系,以增加社会网络关系数据的覆盖范围,正如已完成的唐、宋、元代数据提取。

在持续收集新数据的过程中,CBDB团队会根据实际情况扩展数据模型来捕捉新的数据类别,这些新类别对中国历代人物数据的建模非常重要。例如,之所以有“事件”实体,只是起初团队认为此实体可能重要,但事实证明,到目前为止并没能有效建立“事件”数据,当然,未来也可能会系统地收集到“事件”数据。相似地,限

于当前数据源所包含的数据类别,CBDB 并没有设计中国历代人物商业活动的数据模型。未来收集明清新数据时,商业活动可能会变得重要,届时会根据实际情况增加新模型来捕捉这些数据。

CBDB 是合作性项目,非常欢迎学者个人、大学、商业公司参与合作。近几年,团队在开发众包录入平台方面获得了重大进展,通过众包系统,研究者和志愿者可以很方便地向 CBDB 提交数据。众包录入可以非常有效地录入那些难以用计算机处理的数据,譬如社会网络关系。另外,CBDB 也计划和开放的文本库,比如"中国哲学书电子化计划"进行数据互联。

CBDB 致力于分享不同格式的数据,以满足各种学术研究和工程的需求。前文提到 CBDB 团队已向公众提供了可读写的 API,可读式 API 提供了 CBDB 中所有重要数据点(datapoints)的查询功能,但可写式 API 当前只可向基本信息表和职官信息诸表写入待审核的数据。在接下来的两年内,所有重要数据点的可写功能将被开放。

另一种数据分享格式是关联数据。同样,前文已提到在上海图书馆的帮助下,CBDB 通过开放的在线平台向公众提供关联数据。通过此平台,用户可进行人名和人物传记地址的复杂条件查询。未来,上海图书馆的同仁将进一步为用户提供标准化的入仕种类、任官官职、任官地址和社会关系等查询入口。

除此之外,CBDB 一直在项目网页中向用户分享制作数据的中间产物。① 最常见的分享格式是 Excel 大表,原始数据在这张大表中已被表格化和标准化。今年即将发布的数据是《中国丛书综录》第二册的书名和作者数据。

除共享数据外,CBDB 项目也确保项目中使用的工具是开源的,并对公众开放。② 接下来会将开源的工具用于自动识别人物姓氏,它不仅能识别汉族人姓氏,也可以识别契丹人、满族人、蒙古族人姓氏。

作为 CBDB 商业化分支,"引得"数字人文平台未来将会积极配合项目,优化人物传记库的产品服务,加快完成"中国地名沿革库""中国历代职官库"等结构化数据库的研发进度,与通用文献资源打通,构建古籍命名实体发现系统;并在此基础上,不断优化古籍识别、自动句读、文本标注、文本相似度比对、科研展示图表等古籍整理发布工具系统,为数字人文领域的教学、训练、科研提供精细化服务。

CBDB 团队收集使用 CBDB 数据的论文及学术项目,并在项目网站上发布或建立索引。③ 如果有学者在出版物中使用了 CBDB,并希望通过该项目网页分享自己的研究,可以联络 CBDB 团队。同时 CBDB 团队也鼓励分享在论文或专著中使用到的数据,并建议使用稳定并能长期存续的平台来分享数据,如 GitHub、Dataverse 或 Omeka,或者把数据直接发送给 CBDB 团队发布。

参考文献

[1]Stone L. Prosopography [J]. Daedalus, 1971, 100(1):46 - 79.

[2]Verboven K, Carlier M, Dumolyn J. A Short Manual to The Art of Prosopography [M]//Keats - rohan K S B. Prosopography approaches and Applications: a Handbook. Oxford: University of Oxford, 2007:35 - 69.

[3]Keats - Rohan K S B. Introduction: Chameleon or Chimera? Understanding Prosopography [M]// Keats - rohan K S B. Prosopography Approaches and Applications: a Handbook. Oxford: University of Oxford, 2007:1 - 34.

[4]Oldfield S - J. Narrative Methods in Sport History Research: Biography, Collective Biography, and Prosopography [J]. The International Journal of the History of Sport, 2015, 32(15):1855 - 1882.

[5]Twitchett D C. Chinese Biographical Writing [M]//Beasley W G, Pulleyblank E G. Historians of China and Japan. London: Oxford University Press, 1961:95 - 114.

[6]Wilkinson E. Chinese History: A New Manual [M]. Cambridge, MA: Harvard University Asia Center, 2015:148 - 157.

[7]Hartwell R M. Demographic, Political, and Social Transformation of China, 750 - 1550 [J]. Harvard Journal of Asiatic Studies,

① 参见:https://projects.iq.harvard.edu/chinesecbdb/资料集。

② CBDB 开源工具分享网址:https://github.com/cbdb - project。

③ 使用 CBDB 的论文及学术项目参见:https://projects.iq.harvard.edu/chinesecbdb/下載投影片與論文。

1982, 42(2):365 – 442.

[8]Chen S. The state, the Gentry, and Local Institutions: the Song Dynasty from a Longue Durée Perspective [J]. Journal of Chinese History, 2017, 1(1):141 – 182.

[9]Hartwell R M. A Computer – Based Comprehensive Analysis of Medieval Chinese Social and Economic History [M]//Mair V H, Liu Y. Characters and computers. Amsterdam: IOS Press, 1991:89 – 121.

[10]Smythe D. Prosopography [M]//Jeffreys E, Haldon J, Cormack R. The Oxford Handbook of Byzantine Studies. New York: Oxford University Press, 2008:176 – 181.

[11]Tackett N. The Destruction of The Medieval Chinese Aristocracy [M]. Cambridge, MA: Harvard University Asia Center, 2014: 107 – 145.

[12]Chen S. Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s [M]//Ebtry P B, Smith P J. State power in China, 900 – 1325. Seattle: University of Washington Press, 2016:101 – 152.

[13]De Weerdt H. Information, Territory, and Networks: the Crisis and Maintenance of Empire in Song China [M]. Cambridge, MA: Harvard University Asia Center, 2015:325 – 394.

[14]Bol P K. Changing Literati Networks: Kinship and Collegiality, 1100 – 1400 [J]. Journal of Historical Network Research, forthcoming.

[15]Shang W, Huang W. Investigating the Relationships Between Scholars and Politicians in Ancient China: Taking the Yuanyou Era as an Example [J]. Journal of the Japanese Association for Digital Humanities, 2018, 3(1):33 – 48.

[16]De Weerdt H, Ho B, Wagner A, Qiao J, Chu M. Is There a Faction in This List? [J]Journal of Chinese History, 2020, 4(2):1 – 43.

[17]Bingenheim M. Who Was ‘Central’ in the History of Chinese Buddhism? A Social Network Approach [J]. International Journal of Buddhist Thought & Culture, 2018, 28(2):45 – 67.

[18]Keller F B. Moving Beyond Factions: Using Social Network Analysis to Uncover Patronage: Networks among Chinese Elites [J]. Journal of East Asian Studies, 2016, 16(1):17 – 41.

[19]Keats – Rohan K S B. Prosopography and Computing: A Marriage Made in Heaven? [J]. History and Computing, 2000, 12(1):1 – 11.

[20]Mathisen R W. Medieval Prosopography and Computers:Theoretical and Methodological Considerations [J]. Medieval Prosopography, 1988, 9(2):73 – 128.

[21]Devlin J, Chang M – W, Lee K, Toutanova K. BERT: Pre – training of Deep Bidirectional Transformers for Language Understanding [C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics, 2019:4171 – 4186.

[22]Huang Z, Xu W, Yu K. Bidirectional LSTM – CRF Models for Sequence Tagging [J]. arXiv preprint arXiv/1508.01991,2015.

The History, Methods, and Future of the China Biographical Database (CBDB) Project

Peter K. Bol Wang Hongsu Michael A. Fuller Chen Song Liu Zhou Zhu Houquan

Abstract The China Biographical Database (CBDB) project began in 2005 to create an online and stand – alone relational database with data on the careers, kinship, and social associations of men and women appearing in the Chinese historical record. CBDB gathers data using advanced computational methods and organizes the data to enable analysis between individuals, groups, places, offices, etc. CBDB makes it possible visualize the data in statistical, network, and spatial analysis. CBDB is an open – ended, collaborative project that will continue to gather data on all historical periods. Through its partners CBDB is becoming available to the public through new online systems.

Key words Relational Database; Prosopography; Data Mining; Spatial and Network Analysis; Project Sustainability; Linked Data