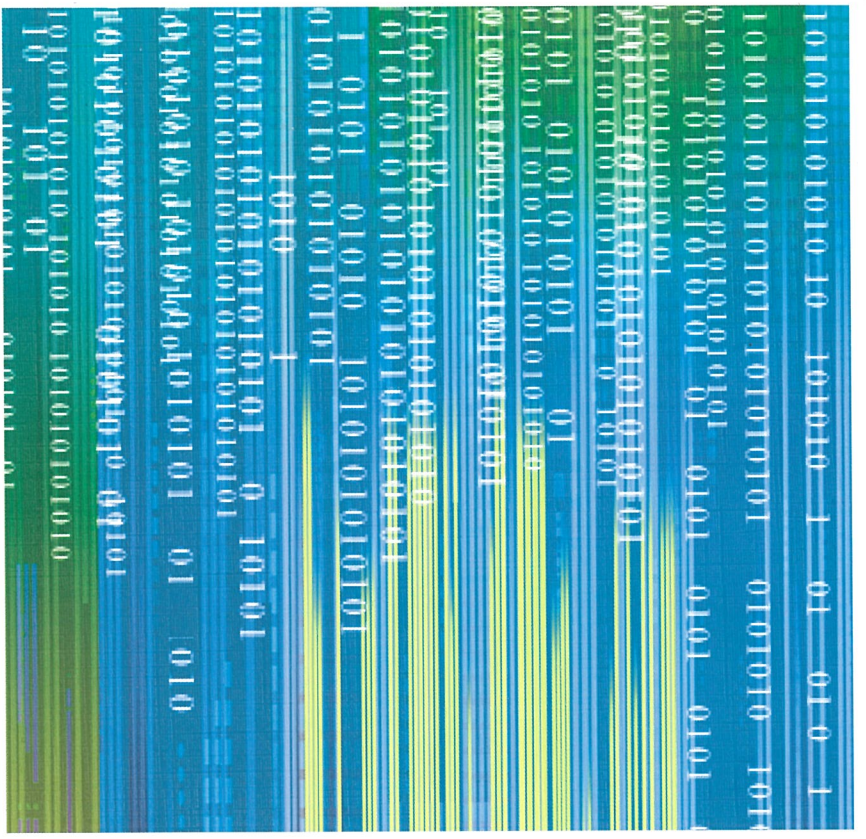
大数据与中国历史研究

第③辑



Big Data and the Study of Chinese History

付海晏 主编

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

目　录

·中国历史研究中的数据库建设·

·专题论文·

·学位论文·

·讲座实录·

·研究动态·

·史料介绍·

从历史记录到结构化人物传记数据：中文材料的半自动处理方式*

徐力恒　王宏甦

一　导言与文献回顾

（一）中国历代人物传记资料库（CBDB）项目

本文研究的将史料文本转化为人物传记数据的方法，是由中国史学研究领域中最著名的一项大型数字人文项目付诸实践的。中国历代人物传记资料库致力于组织中文史料中的海量传记资料，是一个包括将近491000名个人传记信息（截至2021年5月）的关联型数据库（relational database），可用于统计、社会网络、地理空间等各种分析方法。该数据库由哈佛大学、台北中研院、北京大学从2005年开始共同主持，① 其团队成员主要来自哈佛大学和北大，包括人文与计算机领域的专家。该项目利用计算机技术对中文传记资料进行定位与编码，并将这些数据提供给研究者。这些数据既可以在线获取，也可以在微软Access数据库格式中离线使用。该项目也提供用

---

* 原刊为 Lik Hang Tsui and Hongsu Wang, "Semi-Automating the Transformation of Chinese Historical Records Into Structured Biographical Data," in Rebekah Wong, Haipeng Li, and Min Chou, eds., *Digital Humanities and Scholarly Research Trends in the Asia-Pacific*, Cambridge, Hershey, PA: IGI Global, pp. 228 - 246。中文译本校订时作者曾略做增订、更新。

① Harvard University, Download CBDB Standalone Database, 2020, https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database。

于系统互操作的应用程序接口（Application Programming Interface，API）。[①]该项目的合作者还将数据库的条目转化为 macOS 系统自带的辞典软件文件。[②]相较于针对个别人物的研究，CBDB 项目团队成员通过给定格式的文本工作，从而系统地挖掘有关众多历史人物的大量数据。因为 CBDB 记录的信息包括人物的居住地、求学地、任官地、担任的官职、父母、婚姻对象及其人际关系等，人物生活的所有这些方面都能被联结到各个规模非常大的历史人物群组。

CBDB 本质上是一个关联型数据库。换言之，相对于将所有数据加载到其中的单个数据表，它是由许多连接在一起的不同数据表组成的，从而对中国历史人物生活史的不同方面进行分类和编码。与非结构化的纯文本数据库不同，历史人物关系数据库既可以帮助用户发现与定位个别历史人物信息，也有助于在整体水平上系统分析历史人物的特点。通过大范围的数据搜集，CBDB 提供了许多考察过去个人与群体生活之方法。其生成和利用数据的方法正在改变群体传记学（prosopography）的历史研究法。[③]群体传记学探讨的是人物群体的共同特征。它通过把人物生活中共同方面的现象搜集整合成数据，揭示了历史上的社会群体问题，让学者能够更好地理解个人和群体之间的关系。有效利用群体传记学研究方法，尤其是在使用计算机辅助分析的情况下，支持学者对历史文本及其所记录人物关系的探索。[④]在填充诸如 CBDB 一类人物传记数据库的数据时，它不仅为学者提供了更大规模、更全面的数据，还提供了更智能的数据。[⑤]

---

① 徐力恒：《唐代人物大数据：中国历代人物传记资料库（CBDB）和数位史学》，谭国根、梁慕灵、黄自鸿主编《数码时代的中国人文学科研究》，秀威资讯科技股份有限公司，2018，第121—139页。

② Harvard University, *Download CBDB Standalone Database*, 2020, https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database.

③ Stone L., "Prosopography," in F. Gilbert and S. R. Graubard, eds., *Historical Studies Today* (New York: Norton, 1972), pp. 107 - 140. 也可参见徐力恒、王涛《数位人文：跨界与争鸣》，蒋竹山编《当代历史学新趋势：理论、方法与实践》，台北，联经出版事业公司，2019，第539—565页。

④ K. Verboven, M. Carlier, and J. Dumolyn, "A Short Manual to the Art of Prosopography," in K. S. B. Keats-Roha, ed., *Prosopography Approaches and Applications: A Handbook* (Oxford, UK: Prosopographica et Genealogica, 2007), pp. 35 - 70.

⑤ M. L. Zeng, "Smart Data for Digital Humanities," *Journal of Data and Information Science* 1 (2017): 1 - 12; 曾蕾：《图档博领域的智慧数据及其在数字人文研究中的角色》，《中国图书馆学报》2018年第4期。

不仅是研究中国古代史，研究澳大利亚史[①]、盎格鲁－撒克逊时期的英格兰[②]、拜占庭帝国[③]、埃及中王国[④]、伊斯兰史[⑤]、耆那教史[⑥]、日本史[⑦]、台湾史[⑧]等领域，建设和分析数据集和数据库方面都存在类似的人物数据库项目。这些关于其他地区历史的项目，尤其是涉及非拉丁语字母语言文献者，通常在处理和分析实体纸本史料时面临相似的问题和可能性。

（二）建立用于唐代中国群体传记学数据的挑战

本部分讨论的半自动化实践方式是 CBDB 唐代群体传记学子项目的成果。该项目在 2015—2017 年由唐研究基金会（Tang Research Foundation）赞助，美国唐研究学会（T'ang Studies Society）也为这个项目提供了资助。唐代是中国古代最伟大的王朝之一。最近一些利用历史数据来考察社会趋势的研究成果，说明了将唐代人物传记转化为系统化、结构化的数据如何极大地推动了对这段历史的研究。[⑨]针对 CBDB 唐代群体传记学子项目，该项目团队通过调查唐代人物的任官记录、墓志、史书记载和文学材料，致力于对有关唐代人物的职业、旅行和社交网络等信息展开穷尽式的大规模

---

① *BDA Online—Biographical Database of Australia*, 2018, https://www.bda-online.org.au/.

② *Prosopography of Anglo-Saxon England*, 2010, http://pase.ac.uk/index.html.

③ J. Martindale, *Prosopography of the Byzantine Empire*, 2014, http://www.pbe.kcl.ac.uk/.

④ *Persons and Names of the Middle Kingdom*, 2018, https://pnm.uni-mainz.de/name/2748.

⑤ M. Romanov, "Algorithmic Analysis of Medieval Arabic Biographical Collections," *Speculum* S1 (2017): S226－S246.

⑥ P. Flügel, "Jaina-Prosopography I. Sociology of Jaina Names," in *Select Papers Presented in the 'Jaina Studies' Section at the 16th World Sanskrit Conference Bangkok, Thailand & the 14th World Sanskrit Conference Kyoto, Japan* (New Delhi: D. K. Publishers & Distributors, 2017), pp. 187－267.

⑦ *JBDB—Japanese Biographical Database*, 2018, https://www.network-studies.org/#!/; L. Born, "Leveraging the Japanese Biographical Database as a Digital Resource for Education and Research," in *Proceedings of the 8th Conference of Japanese Association for Digital Humanities* (Tokyo: Japanese Association for Digital Humanities, 2018), pp. 24－26.

⑧ S. H. Sie, H. R. Ke, and S. B. Chang, "Development of a Text Retrieval and Mining System for Taiwanese Historical People," in *Pacific Neighborhood Consortium Annual Conference and Joint Meetings* (*PNC*), 2017, pp. 56－62. 张素玢：《建置"台湾历史人物传记资料库"（TBDB）的尝试与初步成果》，《人文与社会科学简讯》第 3 期，2018 年。

⑨ N. Tackett, *The Destruction of the Medieval Chinese Aristocracy* (Cambridge, MA: Harvard University Asia Center, 2014). 中译本为〔美〕谭凯《中古中国门阀大族的消亡》，胡耀飞、谢宇荣译，社会科学文献出版社，2017。

数据搜集工作。其来源包括原始材料和现代学者有关唐代人物的研究成果，下文将对此进行介绍。

目前用于研究中国历史的数据库，通常是利用光学字符识别（optical character recognition，OCR）来处理数字文本的全文数据库，但这类文本数据库项目一般不会将文本转化为可用于大数据分析历史的结构化数据。[①]由于OCR技术目前的识别准确率难以令人满意，这类电子化项目有的甚至不会或尚未将图像处理成全文字文本，过去国家清史纂修工程对清代档案的电子化便是如此。[②] OCR识别手写史料的效用相当有限，阻碍了档案文献的电子化。在OCR技术提高到在没有过多人工监督下计算机就可以准确有效提取手写中文文本之前，学者还是集中对那些印刷本的中文文本进行电子化，才能确保效率，毕竟人工整理原始中文文本的工作量非常庞大，尤其是那些模糊不清的历史文书。

古汉语与现代汉语在组织和分析电子化材料方面也存在一些技术挑战。汉字的数量与差异性也增加了从文献中数字化挖掘与提取中文文本的难度。而且许多相关的资源属于特定的学科和领域，在国际数字人文学术研究中受到的关注较少。诚然，一些研究已经提出了解决部分语言困难和提高汉字电子化的算法，例如Kim等人聚焦于韩国历史文献中的汉字分词。[③] 然而，对电子化特定的中文史料提出实际解决方案的研究仍然非常少，需要我们进行更多思考。

虽然我们意识到学界亟须研究手写本文稿电子化的最佳方案，但是在现有技术下，这种电子化所需要的人力远超CBDB项目所能承担的水平。因此，CBDB项目团队在处理唐代子项目时，选择聚焦于相对容易电子化的印刷文本。在CBDB项目团队开始这些唐代群体传记学的工作之前，已然存在

---

① H. De Weerdt, "Isn't the Siku Quanshu Enough? Reflections on the Impact of New Digital Tools for Classical Chinese," 2014, http://chinese-empires.eu/blog/isnt-the-siku-quanshu-enough-reflections-on-the-impact-of-new-digital-tools-for-classical-chinese/；申斌、杨培娜：《数字技术与史学观念——中国历史数据库与史学理念方法关系探析》，《史学理论研究》2017年第2期。

② L. Mao and Z. Ma, "'Writing History in the Digital Age': The New Qing History Project and the Digitization of Qing Archives," *History Compass*, Vol. 10, No. 5, 2012, pp. 367-374，参见胡恒《数据库建设与清史研究》，《清史研究》2016年第4期。

③ M. S. Kim, K. T. Cho, H. K. Kwag, and J. H. Kim, "Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents," in S. Marinai and A. R. Dengel, eds., *Lecture Notes in Computer Science*, Vol. 3163, 2004.

大量印刷而成的重要史料，但它们需要被电子化、模型化以及被处理为电子数据。这类可操作的历史文献集将是本文讨论的重点。

二 电子化和提取人物传记数据

（一）选择合适的材料进行电子化

CBDB 项目挖掘了多种文本，而这些文本每种都经历了一个多阶段过程。通过将印刷的历史工具书转化为数据，这种有效的流程能更方便学者利用中国史学界的研究成果。例如，对于政府在京城、州一级以及其他地方的官员任命情况，我们利用目前最详尽的研究成果，并利用计算机数据提取信息，将其转化为可检索且结构化的文本。通过详细分析 CBDB 唐代中国子项目的工作流程和电子化操作，以下将介绍和分析电子化与数据采集的原创方法。

在设计这一子项目时，我们选取了以下图书以供唐人传记资料系统电子化之用：

①傅璇琮等编撰《唐五代人物传记资料综合索引》，中华书局，1982。

②郁贤皓、胡可先：《唐九卿考》，中国社会科学出版社，2003。

③郁贤皓：《唐刺史考全编》，安徽大学出版社，2000。

④吴廷燮：《唐方镇年表》，中华书局，1980。

⑤（清）徐松撰，孟二冬补正《登科记考补正》，北京燕山出版社，2003。

⑥吴汝煜等编著《唐五代人交往诗索引》，上海古籍出版社，1993。

这六部工具书因为系统完备、内容可靠、格式简明、便于获取而被选中。第一，它们的编纂目标都在于系统地概括某类唐代历史人物群体及其社会交往，并且都有文献出处可供学者追溯。其中包括大量的中央和地方官，也包括地方人物、诗人以及获得科举功名者。选择这些系统涵盖重要人物群体的工具书，能确保我们全面撒网，并在材料允许的情况下，尽量穷尽对 CBDB 唐代人物传记数据的系统搜集。[①] 第二，这些著作也是唐史研

① 当然，“穷尽”是相对而言的，指的是尚无其他现存工具书能在同一类别中涵盖范围更广的历史人物，所以是现有的最佳选择。

究者比较重视的工具书。这并不意味着上述作品毫无疏漏，但至少表明它们是系统搜集相关人物传记信息方面的最佳成果。[①]这自然有助于保证源于这些文献的人物传记数据的质量。第三，这些著作基本上是用中文排版印刷的现代工具书，因此较易电子化，也便于进行文本挖掘——它们的记录基本上是著录历史人物及相关信息的表单。在每部工具书中，信息都表现为规整的格式。从内容上说，它们可以比较容易地转化为数据，而不需要过多繁重的操作。

为详细说明上一特点，不妨以吴廷燮《唐方镇年表》中记载方镇节度使的一页为例（见图1）。在这部书的各卷之下（本例中，该页截取自凤翔节度使部分），节度使之名皆按时序逐年排列，其下提供了充分的史实依据。

在此书的许多条目中，编者还注明了史料来源或提供了史实考证。不可否认，这类格式便利了批量电子化工作。只要项目团队成员辨明该书呈现人物传记记载格式的基本特征，就可以设计出一种机制，从而选取书中数百条目。与人工搜集并录入数据条目相比，这种方法显然更加高效。最后但也同样重要的是，这些著作要比较容易获取。因为这些工具书的文本在子项目进行之时尚未被电子化，项目团队首先获取的是这些书的印刷本，然后进行进一步处理，具体方案将在下文介绍。

为了在CBDB中扩充唐代中国的地理数据，团队成员还与历史地理专家合作，为唐代群体传记学子项目处理了其他图书，包括郭声波教授的《中国行政区划通史·唐代卷》。[②]团队采用了单独的另一工作流程来电子化并处理这些图书，徐力恒在2018年的一篇文章中对此有详细介绍。[③]在同样的子项目中，CBDB项目团队还处理了一批唐代的原始材料，尤其是记载唐代人物传记的墓志。这是通过一系列文本标记技术实现的，两位作者已在2016年的一篇论文中提供了相关解释说明，本文不再赘言。[④]

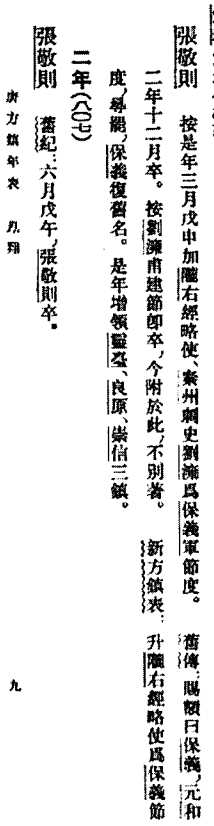
---

① 徐力恒：《唐代人物大数据：中国历代人物传记资料库（CBDB）和数位史学》，谭国根、梁慕灵、黄自鸿主编《数码时代的中国人文学科研究》，第121—139页。

② 郭声波：《中国行政区划通史·唐代卷》，复旦大学出版社，2012。

③ 徐力恒：《唐代人物大数据：中国历代人物传记资料库（CBDB）和数位史学》，谭国根、梁慕灵、黄自鸿主编《数码时代的中国人文学科研究》，第121—139页。

④ H. Wang and L. H. Tsui, "Creative Uses of MARKUS in the China Biographical Database Project," 2016, https://dh.chinese-empires.eu/forum/topic/5/creative-uses-of-markus-in-the-china-biographical-database-project.



张敬則
十八年（八〇二）
張敬則
十九年（八〇三）
張敬則
二十年（八〇四）
張敬則
永貞元年（八〇五）
張敬則
元和元年（八〇六）
張敬則　按是年三月戊申加隴右經略使、秦州刺史劉澭爲保義軍節度。舊傳：賜額曰保義，元和二年十二月卒。按劉澭南趙節即卒，今附於此，不別著。新方鎮表：升隴右經略使爲保義節度，舉關，保義復舊名。是年增領鹽臺、良原、崇信三鎮。
二年（八〇七）
張敬則　舊紀：六月戊午，張敬則卒。

唐方鎮年表　凡翔　九

图 1　《唐方镇年表》中记载 9 世纪初期凤翔节度使的一页

说明：内容包括贞元十四年（798）至元和二年（807）凤翔节度使张敬泽的细节史料。

资料来源：吴廷燮《唐方镇年表》卷 1，第 9 页。

（二）优化 OCR，以便提取和处理历史数据

当项目团队取得前述六部工具书时，首先会把它们的书页拆解和剪切。进而项目团队会使用 ImageTrac 6300 扫描仪扫描这些拆散的书页。机器会一次性扫描多份书页，一分钟可扫描约 286 页。由此扫描仪可以生成高分辨率的双面书页数字图像。扫描仪以三种格式保存图像，分别是 JPEG 2000、TIF 以及 PDF 文件。如果放入扫描仪的书页彼此叠合在一起，扫描仪会在开始扫描之前自动检测，并把重叠书页分开。这项功能减少了扫描过程中人工的参与。

图像扫描后会被处理以清除图像噪声。这些图像噪声涉及书页中影响光学字符识别率或者扰乱人物传记信息数据化的任何标记。这些内容包括页眉[①]的章节标题和页码，也包括用于中文出版物的标点符号，例如专名号和书名号等。通过专门设计的算法，方框中的页眉从书页中被剪切移除（见图2）。我们利用一项广度优先搜索（breadth-first search）算法，把图像中的连续线识别为非字符组件。这些在电子化过程中不需要的标点符号标记也会从图像中清除出去。由于不同工具书页的页眉形式有所差异，每部书都需要量身定制方法，但一部书只要设计一种就足够，不需要逐页处理。这意味着项目团队为每部书都设计并采用了独特的算法。简而言之，我们发展了多种算法来优化OCR的精确度，尤其是用于带有现代标点符号和格式的书。在此过程中，项目团队会记录页码以便在数据库中保留史料来源。史料来源会在最终输出的CBDB数据库条目中显示。

在完成这些图像处理步骤后，下一步是将图像置于OCR的处理过程。项目团队将已被处理和清理的高分辨率图像输入OCR应用软件ABBYY Fine Reader。由此，工具书中的繁体字文本从电脑图像中被识别并提取出来，每页文本都会被保存下来，以方便之后的检索和校对工作。

项目团队利用了一个专门设计的校对系统，对这些可检索的文本进行人工核验。尽管上述OCR软件所用的算法是项目团队试验过的最好的一种，但由于团队所处理的文字包括许多易引起误认的繁体字，这超出了目前我们所知的任何OCR软件的能力。有鉴于此，以人眼检查来校对这些可检索文本是至关重要的。上述校对系统可以让CBDB项目团队的编辑小组成员将被标记的文本和图像文件并置对勘，并在必要时校改文本（见图3）。

由于编辑小组成员在多地同时操作，因此项目团队在线管理校对系统，以便编辑者可以随时随地通过互联网访问。文本数据储存于服务器中并在线保存，以便同时工作。因为只有文本中的数据点才会进入数据库，所以编辑小组成员不必校对全文，只需要聚焦校对那些数据点。在大多数情况下，编辑小组成员所做的注释和史料全文都未校对，因为它们不会进入CBDB的数据中。再以《唐方镇年表》为例，只有其中的姓名、官名以及任命时间会经编辑者校对。只关注这些将纳入数据库条目的信息，大大减轻了

① 竖版中文书籍的页眉一般在页面的最左或最右侧，因此算法可根据这些特点而制定。

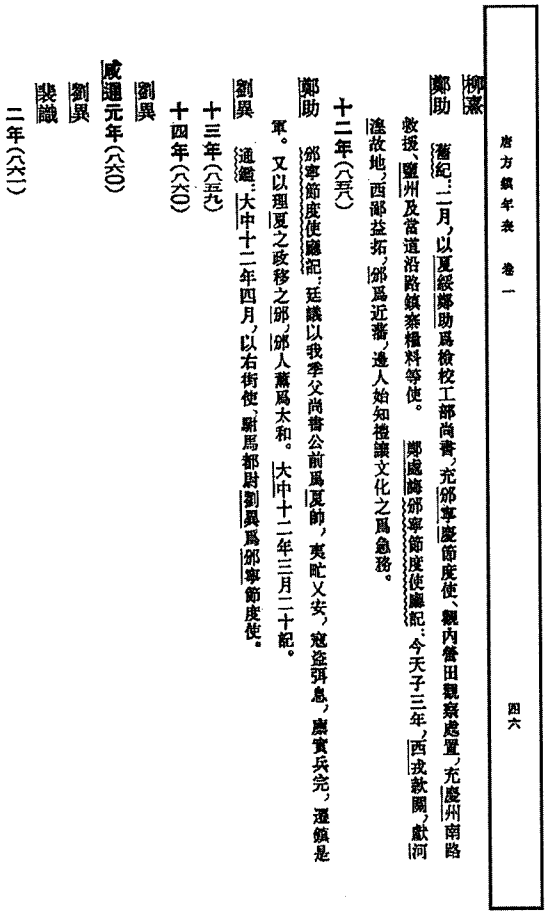


图 2　图像扫描后去页眉

说明：在图像处理过程中，右侧方框中的页眉从书页中被识别出来，并剪切移除。页眉的内容包括书题、卷数以及页码。

资料来源：吴廷燮《唐方镇年表》卷 1，第 46 页。

负责校对的编辑小组成员的负担，并且提高了校对工作的准确率。同时，该校对系统记录了编辑者频繁更正的汉字，并在积累足够的记录后在文本中以红色高亮显示，有助提醒校对者特别注意某些 OCR 效果很差的字符。该机制有利于降低校对工作的错误率。为避免误校，当前 OCR 程序中的字符由机器自动校对，但这展现了一种可能性，即根据人工校对的训练数据（training data）可以开发出一种评估每个汉字 OCR 准确率的自动化程序。这将是一种有望被纳入上述校对系统的功能。

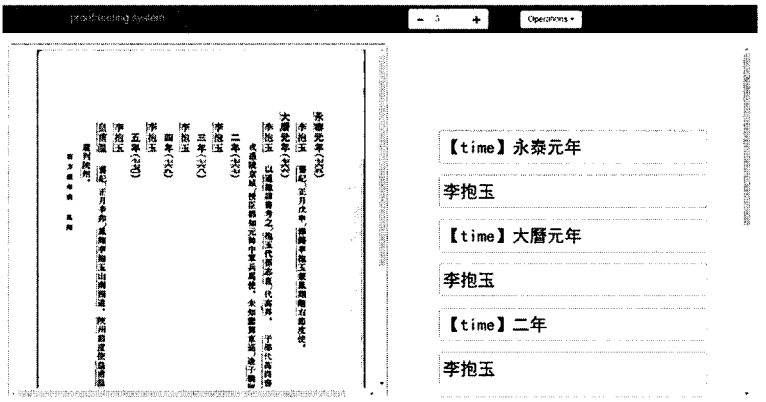


图3 CBDB在线校对系统

说明：顶部的工具栏提供页面切换功能和其他操作功能。左侧显示的是吴廷燮《唐方镇年表》中的一页，右侧显示的则是OCR生成的数据点（data points）。不同类型的数据点被分行显示，而年代信息是在“time”之后显示。

文本被校对完成后，项目团队就会使用分词（word breaks）和停用词（stop words）对其进行汇总和分段，例如“年”这个字。之后，就把文本制成表格，把其中的数据点整齐地分开排列，并根据不同类别记录下来。到这个阶段，这些历史工具书的数据化就完成了。

三 人物传记数据的处理和消歧

尽管工具书的数据化已经完成，但数据在纳入CBDB之前，仍需进行更多处理。本阶段需要进行数据清理，以便使不同材料来源的人物传记条目能使用相同标准，可资比较。举例而言，项目团队会将中国古代年号批量转化为公元纪年，比如武德四年一般会转换为621年。一些采用简体字的工具书也会被仔细地转为繁体字，以便使所有人物传记数据都有统一的编码。[①]

处理这些数据的重要环节之一是人物姓名的编码和消歧。尽管CBDB的自我定位是一个用于群体传记学研究的数据库，但它并不是仅仅搜集史上

① 因为数据库处理的是历史材料，CBDB目前的数据底本均为繁体字。

所有名称出处的那种群体传记学成果，而是一种对已消歧人物的传记信息进行模式化处理和组织的数据库。由于这里讨论的唐代群体传记学项目所选用的工具书内容有重叠之处，这意味着这些著作可能分别记载了相同人物的不同方面，所以在这些记载成为 CBDB 人物传记数据之前，需要先进行消歧工作，项目团队也需要识别和连接同一历史人物的资料。当然，出于中国的命名习惯，同名的情况是相当普遍的，因而这项任务有特殊的挑战。[①]中国历史人物拥有相同姓名的情况并不少见——并非所有同名的人物传记条目都指向同一位历史人物。在项目团队将这些人物传记条目纳入数据库之前，需要首先理清这些条目是否表示同一历史人物。换言之，必须对相同人名条目进行消歧，才能进行编码。[②]

为了使消歧工作高效而准确，同时借助了计算机算法与项目团队中人文学者的历史思辨。这里所指的算法，有助于比较条目中的数据点，以确定在同一姓名的多条传记条目的官名、任命时间、别名等内容是否存在重合。这些成对的条目被分为五个等级，以表示它们属于同一历史人物的可能性，每一等级由带星号的数字标识，从可能性最低的 1 * 到可能性最高的 5 *。将所有相同人名的传记条目分为这五个等级后，可能性最难以确定的成对条目，例如等级 3 *，则会由人工检查。在中国文史方面训练有素的编辑小组成员会逐条检查这些条目，如有必要还会参考更多史料，来确定它们是否为同一历史人物。通过把人工精力集中在这些难度最大的条目上，我们确保人文学者的学术判断首先被用于机器算法最难解决的问题。之后编辑小组成员将进一步解决等级 2 * 和 4 * 的条目。他们通常并不处理等级 1 * 和 5 *，因为基于上述算法，我们已经可以非常确定这些条目消歧的结果。这种半自动化消歧方法大大减少了项目团队处理上百对人物传记条目的工作量。同时，该方法并未牺牲数据条目的准确性。

接下来，项目团队会将正则表达式（Regular Expressions）算法应用于可检索的表格文本，该算法用 Python 编写和修订。这样，就可以基于 CBDB

---

① P. Adamek, *A Good Son is Sad if He Hears the Name of His Father: The Tabooing of Names in China as a Way of Implementing Social Values* (UK: Routledge, 2017).

② 其他群体传记学项目可能面临着类似挑战，参见 P. Flügel, "Jaina-prosopography I. Sociology of Jaina Names," in *Select Papers Presented in the 'Jaina Studies' Section at the 16th World Sanskrit Conference Bangkok, Thailand & the 14th World Sanskrit Conference Kyoto, Japan*, pp. 187 – 267。

现存有关官名等方面的数据，为文本中的条目进行编码。[①]只需运用常见的应用程序，例如在 Excel 表格之中进行操作，这些表格数据便可以使用公式（formulas）和宏（macros）等功能实现编码。完成这些工作后，它们最终将会被纳入 CBDB 数据库。以上就是唐代群体传记学子项目的数字化工作流程。

四 讨论

（一）半自动化方式的效率

通过使用可被验证的计算机技术，我们使单纯运用人力无法达到的速度让抓取中国人物传记信息成为可能。例如，运用上述工作流程，项目团队只花费了接近 88 个工作时数，就把《唐方镇年表》中的 1567 页人物传记信息转化为结构化数据。相较于人工输入传记条目（即项目团队在 2008 年之前采用的工作方案），本文勾勒的工作流程和处理过程极大地节省了人力。人力主要被投入需要人文专业知识的任务上（例如从含糊的记录中考辨出可信的历史信息），而不是用于手动录入并整理数据。通过这种方法，中国人物传记的数据化工作也远比人工录入和校对大量汉字更为有效。采用这一工作流程后，CBDB 共收入了 52751 名唐代人物数据。所有这些新添加的人物传记数据都利用本文介绍的半自动化方法实现了电子化。简而言之，这种工作流程准确而高效地把现代出版的工具书文本信息纳入了关联型数据库。

非常重要的是，这个工作流程并不以牺牲数据的质量来换取效率。由于 CBDB 是一个旨在促进学术研究的学术项目，其历史数据的准确性与可靠度至关重要。实际上，人文研究者经常质疑数据库处理和记录原始资料的可靠性，他们通常更倾向于亲自复查并考订史料，而不是单纯引用二手文献，或数据库中的记载。因此，为了维持数据库的可信度，提高学者使用的信心，CBDB 项目团队努力确保人物传记数据的质量。在这个唐代群体传记学子项目的工作流程中，作为数据来源的工具书选取了那些被学界广泛认可的作品。在工具书的数字化工作流程中，既运用计算机算法，但只要

① E. Yamangil, S. P. Chen, and P. Bol, *A RegEx Machine*, 2012, https://projects.iq.harvard.edu/files/cbdb/files/a_regex_machine_yamangil_chen_bol.pdf.

有必要，也会使用人工作业——其中包括对易引起误解的汉字进行细致的人工校对，也包括以计算机算法进行人工难以迅速处理的大量人物传记条目消歧工作。类似 CBDB 项目团队处理的中文史料记载往往并不十分规整，也包含相互矛盾之处。在机器学习机制进一步提升其处理这类材料的性能之前，人文专家的持续投入仍然至关重要，甚至可以说是必要的。因此，CBDB 项目高度重视具有学科领域专业知识的学者的参与。

（二）史料数据化的价值

本文的研究方法极大地促进了在 CBDB 中输入新的结构化数据，以便进行群体传记学分析。这种新开发的工作流程将改变我们选择哪些中文文本进行数字化的考量，同样也会影响我们对这些资料进行数字化的最佳工作方案的选择。这也使当代数字化研究成果可以建立在前辈学者多个世纪以来整理人物传记资料的努力之上，包括历史人物的列表、人物传记辞典、人物传记资料索引等。在被转化为可以轻易于数据库和搜索引擎中查询的数据，以及在整体水平上进行可视化和分析之前，这类印刷本的人物传记资料已俨然是个高质量的历史信息“库”了，然而学者一般很难对其进行综合研究。把这些书的记载转化为数据之后，这些著作的学术价值便会得到提升，从而再次焕发活力。本文所阐述的数字化方法，不仅专就唐代而论，也涉及中国历史上留下文字史料的其他历史时期。当更多的这类著作被数字化时，学者可获取、运用的历史数据将会更加庞大且更全面，也将更加智能。CBDB 项目团队已经开始处理唐代中国的其他材料，同时在处理中国历史更晚近的时期，包括宋代到清代的材料。[①]由于这些工作，学者编纂的人物传记参考书不再被印刷形式限制，学者可以方便地获取它们的信息，并利用计算机技术对它们进行各种分析。在电子化过程中记录大量文件，例如原始参考书的页码，学者有需要也可以查找原书出处，进行核对。

（三）未来可供研究的新问题

唐代史料的电子化不只让回答传统历史问题更加便利，也为数字人文

① 徐力恒：《唐代人物大数据：中国历代人物传记资料库（CBDB）和数位史学》，谭国根、梁慕灵、黄自鸿主编《数码时代的中国人文学科研究》，第 121—139 页。

研究者生发出新问题和研究上的新可能。例如，上述中文人名条目的消歧工作流程，对其他大规模电子化和数据化项目也是有价值的，尤其是旨在处理与中国历史人物相关资料的那些项目。实际上，这种消歧技术已经有效地应用到 CBDB 的其他子集，例如系统地从地方志中挖掘官员传记数据。[①]如果史料是以印刷形式存在的，巨量史料中的大规模人名消歧问题在以往是没有可能用人力简便地解决的，尤其是一些常用的中文姓名。我们的数据处理工作流程将为这类问题的解决提供新的思路。

以上所述的中国人物传记工具书的数据化工作流程也并非全无缺陷。遗憾的是，由于现有 OCR 技术的限制，对于手写的中文文稿而言，采用本文所述的工作流程是不符合成本效益的。将该工作流程应用到现代排版的中文图书，尤其是具备相对统一格式的人物传记工具书，才能产生最佳效果。另一局限是，在这种半自动化方式的每一阶段，人工参与依然至关重要。换言之，该工作流程仍需要不少人力来修正每部工具书数据化的具体实践，包括配套的算法。由于保障 CBDB 人文数据的质量至为关键，半自动电子化在可见的将来仍会是主要方法。由于本文讨论的子项目利用上述半自动化方法生成了不少历史数据，也从中积累了训练数据（training data），可以促进 OCR 的自我编程以及史料的数据化。值得探讨的是，将来在其他研究中更大规模的训练数据如何促进由机器智能支持的自动电子化。另外，子项目产生的数据对于唐代史料的处理和电子化分析也会大有帮助。[②]

---

① Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang, and Peter K. Bol, "Mining Local Gazetteers of Literary Chinese with CRF and Pattern based Methods for Biographical Information in Chinese History," in *Proceedings of the Third Workshop on Big Humanities Data*, 2015 *IEEE International Conference on Big Data*, Santa Clara, CA: IEEE, 2015, pp. 1629 - 1638; L. H. Tsui, "How many People have Your Name? Computational Approaches to Name Disambiguation for Chinese Historical Figures," Paper presented at "Digital Humanities Asia Summit: Harnessing Digital Technologies to Advance the Study of Asia and the Non-Western World", Stanford University, Palo Alto, CA., 2018. 参见彭维谦、程卉、陈诗沛《从全文到表格：地方志职官志中职官资料之半自动撷取》，《数位典藏与数位人文》第1期，2008年，第79—125页。

② Chao-Lin Liu and Yi Chang, "Classical Chinese Sentence Segmentation for Tomb Biographies of Tang Dynasty," in *Proceedings of the* 2018 *International Conference on Digital Humanities* (Mexico City: Association of Digital Humanities Organizations, 2018), pp. 231 - 235; 徐力恒：《中国历史人物大数据》，《中国计算机学会通讯》2018年第4期。

结 语

本文论述的技术是历史学者、文学学者、计算机学者以及编程员之间努力合作的结果。作为美国、中国大陆和台湾地区研究机构的共同创举，它也是亚太地区独特的学术项目。尽管数字人文作为一种富有前景的学术范式已经深深影响了国际学界对历史文献的研究与处理，考察非西方和非拉丁字母语言文献的特殊电子化实践的研究仍然不多，相关探究也不够充分。本文综合探讨了相关技术和语言的挑战，以及 CBDB 项目团队如何利用合适的方法来应对这些挑战。当然有些问题仍需要等待进一步的技术革新方能得到彻底解决。通过介绍 CBDB 工作及其在中国历史研究方面的电子化探索，本文希望为中文文本史料的电子化做出一点贡献，并且希望借此展现数字人文研究积极地面向更丰富的文化多样性，为处理非西方国家的文献做出贡献。[①]通过分析数字人文研究在中国历史数据方面的电子化、处理及呈现等工作流程，我们衷心希望更多研究者参与中文史料宝藏的数据化以及利用工作。

［徐力恒，香港城市大学中文及历史学系助理教授；王宏甦，哈佛大学中国历代人物数据库（CBDB）高级项目经理；徐阳（北京大学历史学系博士生）译；徐力恒校］

① S. Mahony, "Cultural Diversity and the Digital Humanities," *Fudan Journal of the Humanities and Social Sciences* 11.3 (2018): 371 - 388.

图书在版编目(CIP)数据

大数据与中国历史研究. 第3辑 / 付海晏主编. --
北京 : 社会科学文献出版社, 2021.9
ISBN 978-7-5201-8564-6

Ⅰ.①大… Ⅱ.①付… Ⅲ.①数据管理-应用-中国
历史-研究 Ⅳ.①K207

中国版本图书馆CIP数据核字(2021)第121107号

大数据与中国历史研究(第3辑)

主　　编 / 付海晏

出 版 人 / 王利民
责任编辑 / 李丽丽　陈肖寒
文稿编辑 / 徐　花　侯婧怡

出　　版 / 社会科学文献出版社 (010)59367256
地址：北京市北三环中路甲29号院华龙大厦　邮编：100029
网址：www.ssap.com.cn
发　　行 / 市场营销中心 (010) 59367081　59367083
印　　装 / 北京玺诚印务有限公司

规　　格 / 开 本：787mm × 1092mm　1/16
印 张：11.5　字 数：178千字
版　　次 / 2021年9月第1版　2021年9月第1次印刷
书　　号 / ISBN 978-7-5201-8564-6
定　　价 / 85.00元

本书如有印装质量问题，请与读者服务中心 (010-59367028) 联系