



Universiteit
Leiden
The Netherlands

唐宋史研究中的數字化語文學 [The Uses of Digital Philology in Tang-Song History Tr. Lincoln Tsui]

De Weerdt, H.G.D.G.

Citation

De Weerdt, H. G. D. G. (2017). 唐宋史研究中的數字化語文學 [The Uses of Digital Philology in Tang-Song History Tr. Lincoln Tsui]. 唐宋歷史評論 [Tang And Song History Review], 3, 3-19. Retrieved from <https://hdl.handle.net/1887/70164>

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/70164>

Note: To cite this publication please use the final published version (if applicable).

唐宋史研究中的数字化语文学

魏希德 撰 徐力恒 译

在过去的五年左右，研究中国史的学者和研究其他地区的历史学者一样，意识到数字史学并非只是把文本变成量化资料，或只是把这类资料整合起来。^①不少人发现，历史学者利用数字化语文学（digital philology）的方法可以辅助并丰富他们在研究文本和物质文化中使用的分析方法和诠释策略。在这篇短文里，我将概括讨论我接触到的一些用于唐宋文献的数字化研究方法，并讨论我如何将之整合到唐宋政治史和思想史的研究之中。

一 从独立的数据库到数据的关联

大部分历史学者已经接触数字化技术十到二十年了。除了最容易联想到的用计算机打字以外，还有为了搜集资料进行简单的检索。此外，很多学者已经在使用一些常见的办公软件，用来创建表格、统计表、数据库等。这些做法逐渐在辅助或已然取代了过去的读书笔记和卡片。我自己探索数字人文（digital humanities）这个新领域的历程是这样开始的——2007年，我尝试探讨“中国历代人物传记资料库”（CBDB）的资料和数据结构可以为研究南宋（尤其是12世纪）的书信网络带来什么样的资源。我在那一年的“中古中国的群体传记学：对中国历代传记资料库的利用”（Prosopography of Middle Period China: Using The Chinese Biographical Database）工作坊^②提交了一份报告，认定CBDB不适合用来研究书信网络，因为它

① 参见：<http://chinese-empires.eu/blog/digital-interpretations/>。

② 工作坊的内容可参见：<http://www.songyuan.org/Warwick.htm>。

的相关资料还远远不够。我同时指出它应该进一步扩充，收录更多现成的工具书，并容纳更个性化的小规模资料。在这种探索和研读数据库开发工作的相关文献的过程中，我越发意识到把我本人搜集的文献和阅读笔记跟 CBDB 结合，能起到很大的作用。

我在这方面的尝试，源于我对宋人笔记的编纂和刊刻的社会文化史研究。这是我对宋朝时政和历史类文献（例如邸报、会要、地图、奏议等）的接受史进行考察的一部分。我首先编了一个表格，录入目前所见宋代刊刻笔记的所有版本信息。^①我把这个表格变成一份统计表，其中栏目包括作者、笔记原题、编纂时间、刊刻时间、刊刻者、赞助者（如果有的话）、刊刻地点、目前收藏地点（如果该笔记流传至今）和这条信息的出处。有了这些信息，再把它们连到 CBDB 不断增加的人物传记资料，我得以对很多宋人笔记的作者及其生平进行社会背景和地理的分析。我也可以借助哈佛大学地理分析中心的贝明远（Lex Berman）及其同事管理的“中国历史地理信息系统”（CHGIS），运用其提供的地理坐标来画地图，呈现多部宋人笔记的序文和各书目材料中提到的地名。根据宋代刊刻笔记的社会情况，我在一篇论文中指出，从这些数据中，我们可以看到高层官员对笔记的编写和刊刻仍有重要作用，但他们扮演的角色不再像过去那样占据主导地位。^②而这些笔记的主要内容却没变，仍是关于朝廷政治。因此，我所看到的情况和张晖等学者的观察不同——他们认为笔记的作者逐渐变得不再是高官，其中包括了更多地方士人，其中有的人甚至没有官位。^③

在把自己搜集的数据跟其他数据库连接起来之后，我继续扩展这种研

① 网址为：<http://chinese-empires.eu/static/media/uploads/database/index.html>。

② 参看我的“The Production and Circulation of ‘Written Notes’ (biji),” in Michela Bussotti and Jean-Pierre Drège eds., *Imprimer autrement: Le livre non commercial dans la Chine impériale*, Geneva: Librairie Droz, 2015, pp. 19–47. 全文见：https://www.academia.edu/1640813/2015_book_chapter_The_Production_and_Circulation_of_Written_Notes_biji_.In_Imprimer_autrement_Le_livre_non_commercial_dans_la_Chine_imp%C3%A9riale_.Ed._Michela_Bussotti_and_Jean-Pierre_Dr%C3%A8ge_pp._19-47._Ecole_pratique_des_hautes_%C3%A9tudes._Geneva_Librairie_Droz_2015。同时可以参看我的 *Information, Territory, and Networks: The Crisis and Maintenance of Empire in Song China*, Cambridge, MA: Harvard University Asia Center, 2016, 第六章。

③ 此处提到的是张晖《宋代笔记研究》，华中师范大学出版社，1993。——译者注

究方法，开始处理笔记的电子全文。把数据连接到这些全文，可以让我在研读笔记的过程中看到更丰富的信息。我也因而发现，要分析王明清《挥麈录》的信息网络，最好是利用笔记的文本建一个数据库。王明清经常在这部书里提到他跟谁有过对话、他的消息来自何人、他读到的书作者是谁，还有他拜访了哪些藏书家，等等。我对全书的电子版做了标注，对书中的每条消息来自何人，关于什么话题，或关于什么文献，都一一做记录。当我这样做的时候，我还在标注里为笔记提到的每个人加上 CBDB 的人名编号，然后把电子文本连到 CBDB。这样就可以从 CBDB 导出上百条在《挥麈录》提到的人物的年代、籍贯、仕途等信息。尽管我有时候要为其增补一些额外信息，或纠正其中的一些资料（类似做法往往是必要的），但 CBDB 的好处在于，它收录了昌彼得、王德毅等编的《宋人传记资料索引》^① 和李之亮编的《宋代京朝官通考》^② 等参考资料，而这些资料可以帮助我们重现王明清的信息网络，并分析其社会、地理面向。^③

这项研究是我的新书《信息、疆域、网络：宋帝国的危机和维系》（*Information, Territory, and Networks: The Crisis and Maintenance of Empire in Song China*）第二部分包含的内容。我在书中提出，中国政治体生成并散布信息的方式在宋代发生了结构性的转变。本来，各种体裁的文献，如疆域图、会要、邸报、军事地理书等都由中央官员编纂，并由同一批人使用、参考。不过，到了 12 世纪，阅读和评论这些文献的人是来自宋帝国在各地居住的各种文化精英。我在制度变化、法律史、文化生产等方面分析朝堂和士人之间在政治沟通上的变化。我系统地分析了笔记和其他文献中记载士人阅读各种材料的做法，并在此基础上提出士人生成文本数量的大规模上升显示两种特点，对其后的中国历史在大规模帝国的形成和维系有重要作用。首先，士人在他们对中国整体的讨论中普遍表达出一种帝国观，一种对理想中中华共同体的愿望。其次，这些文本在规模颇大的圈子中流传和被人讨论，涉及的人数众多，背景多元，分布广。正是利用了数

① 昌彼得、王德毅等编《宋人传记资料索引》，鼎文书局，1974～1976。

② 李之亮编《宋代京朝官通考》，巴蜀书社，2003。

③ CBDB 的资料来源可以参看：<http://projects.iq.harvard.edu/cbdb/cbdb-sources>。这些参考资料当然不会全无缺漏，但它们都是相对重要的。

字化手段，使我得以看到官方文书的接受和士人网络呈现何种具体情况。

然后我进一步拓展这种尝试，和徐力恒、陈韵如、李允中、尤紫锡等研究生选读了更多笔记。我们的原始数据可以在我为这部书架设的网站上下载。^① 在何浩洋博士的协助下，我也把这些文本和数据建成一个在线平台，为笔记中的信息网络提供互动的可视化功能（见图1）。读者可以据此验证我书中的论点，也可以对数据和文本进行其他分析，比如可以：

- 画出热点图（heat maps）和群集图（cluster maps）呈现信息来源人的籍贯和为官地点等
- 比较不同笔记提到的人的时代分布
- 查阅信息来源人是否有为官经历，以分析他们的背景
- 查阅不同信息来源人和其他人物在笔记中出现的次数
- 查阅这些人在哪些笔记段落中出现，并查阅原文



图1 为特定宋人笔记建立的互动分析平台

当前界面为“地图视图”，显示信息来源人的籍贯。最下面的工具栏除“地图视图”外，包括“全文视图”、“表格视图”、“柱形图”和“饼图”等选项。

在探讨笔记中提到的社会关系时，我对笔记这一类文献和它在宋代的变化有了新的认识。这种语境化的阅读方式，连带其他种类的数字化阅读

^① 网址为：<http://chinese-empires.eu/reference/information-territory-and-networks/>。

方式（例如我书中第8章对不同版本进行比对的语料库语言学分析 [corpus linguistic analysis]），改变了我原来细读这些文献时得到的认识。这些方法使我能够重新检讨王明清等人在笔记中如何表达他对当时宋金关系的独特见解，并将之放到当时士人普遍看法的脉络中。学者进行这种脉络化的工作，总是很有挑战性的。

二 大规模的文本挖掘

我最初探索数字化研究方法时，以大规模和中等规模的史料挖掘的做法为主。我当时集中研究了特定文本（例如一位作者编纂的一系列笔记）或少数几个作者的特定著作。不过，数字化研究方法其实可以让我们做更多元化的工作。对我来说，数字化阅读能为我们实现一些独特功能——它允许学者的眼光在不同规模之间切换，可以时大时小。其实，学者已经习惯对艺术品和文物采取这种方法，他们有时放眼在其中的细部，有时同时研究一大批物件。^① 而我们对文字材料也可以做同样的事情。数据审编（digital curation，又译数据管理）的做法让我们得以扩大或缩小研究的规模，不过我们目前还没找到最妥善地处理数据审编的方式。

对历史学者来讲，只要想到涉及一个人群、一个广阔的地域，或各个世纪的长时段，规模的重要性很明显。我会从目前进行中的研究计划中举出两个例子，解释数字化研究方法如何让我们进行大规模的研究。第一个例子和政治史有关，^② 第二个则和城市史和科技史相关。

宋代政治史和思想史的研究者长期以来热烈讨论11~13世纪的大规模党争。对这个关键问题的大部分研究著作要么集中分析编年史料对这些事件的刻画，要么集中在一些人物之上（有时候是重要人物，也有时候是比较边缘的人物，例如黄宽重先生近来重点研究的孙应时）。利用这些角度研究党争问题，不容易理解集体人群是怎么聚合起来变成政治联盟的。每

① 例如 Florian Kräutli 开发的工具，网址为：<http://www.kraeutli.com/index.php/2016/04/08/timeline-tools/>。

② 参看我在莱顿大学的就职演讲，全文见：<http://media.leidenuniv.nl/legacy/oratie-de-weerdt-eng.pdf>。中文翻译见魏希德《重塑中国政治史》，《汉学研究通讯》2015年第2期，第1~9页。

个士人在生涯中的不同阶段都参与到人脉的建立和维系。不论是准备科举考试和应考的时候需要这样做，寻求官职和迁转，甚至是投身其他事业，都需要这样做。既然这种活动对于士人的仕途来讲是必需的，而且是种普遍现象，那么历史学者自然不只需要了解中央朝廷中的党争，也需要知道地方上的情况如何。我认为，要研究党争如何渗透到地方上去，必须将现存的相关史料一网打尽。因此，在一些博士后和博士生的协助下，我开始分析不同党籍人士的相关著作中同时提到过哪些人名，看看这能如何帮助我们研究这个问题。通过分析 12 世纪三个关键时刻，我们希望探讨党争在宋代的社会和文化演变中发生了什么变迁这个问题。

在一次实验中，我们整理了《元祐党籍碑》（12 世纪前十年）、秦桧（1090 ~ 1155）迫害的人员名单（12 世纪 40 年代）和庆元党禁（12 世纪 90 年代）等三个名单中的所有人物，然后搜罗每个时代在世士人的所有传世文章。^① 然后，我们分析三个名单中的人在哪些文章中有同时出现的。由于坊间已经有大量材料被电子化，所以这种大规模的分析是可以做到的。我们以 CBDB 的“指数年”（index years）为标准，^② 囊括了 1104 年、1142 年和 1196 年这三年前后各三十年的所有宋人文集中的篇章。1104 年前后三十年的相关史料包含 59969 篇作品，包括 2231 位作者的 23701759 字；1142 年前后三十年的相关史料包含 47040 篇作品，包括 1139 位作者的 18780575 字；1196 年前后三十年的相关史料包含 52593 篇作品，包括 2598 位作者的 23446605 字。由于电脑对文本进行自动检测的技术仍有很大局限，这些数据需要经过人工仔细处理，拿掉不相关的资料。

我们据此可以得出三组文集史料都出现的人物究竟有哪些。我们得出这些数据之后，可以用来做什么？其中列出什么作品论及不同党派的人员，并注明次数是多少。我们能够以此发现哪些作品是我们过去研究中没有足够重视的，应该得到更多注意。通过这些党争成员同时出现的关系，

① Hilde De Weerd, Chu Ming - Kin and Ho Hou - leong, "Chinese Empires in Comparative Perspective: A Digital Approach," *Verge: Studies in Global Asias*, Vol. 2, No. 2 (Fall 2016), pp. 58 - 69. 全文见: <http://www.jstor.org/stable/10.5749/vergstudglobasia.2.2.0058>。

② “指数年”是 CBDB 推算历史人物年代的指标，标准参见 http://projects.iq.harvard.edu/files/cbdb/files/131009_rules_for_index_years.xls?m=1438184906。——译者注

我们还可以进行社会网络分析，考察党争成员的网络重合度为何，有哪些小群体，并且其中有哪些人物占据重要地位，能起到把不同人群连起来的作用。我们可以由此回到一手史料和二手材料，进一步研究这些人物和我们过去忽略的一些小群体。我们也可以进一步考察有哪些因素影响这种同时出现的人，及其中小群体的成员构成，例如籍贯、家庭关系、仕途等。另外，我们希望以后能记录史料如何呈现不同时期的党争成员之间的社会关系，并有什么样的变化。这项研究仍在进行之中，但初步结果显示几个群体在党派结构上相当不同——元祐党人名单是小群体的松散结合，庆元党禁则是紧密地和一个核心人群连在一起的，而秦桧迫害的人员之间的关系则不明显。这显示社会网络分析可以用来展示社会关系有什么影响，也可以用来反映社会关系的缺席有哪些影响。我们也打算借助中国历史的大型群体传记数据库和文本数据库，推进目前分析历史上社会网络的做法，尤其是发展取样（sampling）的方法，用以比较社会网络中共同出现的人（co-occurrence）和同时代人群的随机抽样人群。这项工作显示，道学群体是一个联系很紧密的群体，因为当我们利用同样方法抽样分析庆元党人中的一百个人在 52593 篇作品中共同出现的程度的时候，发现其要远远低于道学群体。

这些提取数据的方式对研究很多历史问题都有帮助。我和我的一位博士生熊慧岚目前正在研究修筑城墙的长时段历史。我们利用一系列“正则表达式”（regular expressions）^①，描述地方志和文集中城记、门记等史料一般提到修筑城墙的文字表述，藉此提取数据。据此我们建了一套数据集，涵盖从宋代到清代修筑城墙的活动记载。这可以让我们在空间和时间上呈现城墙的修筑活动、维护、坚固程度、工程材料、使用人力、涉及、规模、地点、作用等方面。这种研究方法对其他历史学者也有用处，例如城市史、军事史专家在研究城市规划和军事科技时就可以利用这样的数据集，考察修筑城墙和火药技术发展的关系。^②

① 正则表达式可以帮助我们全文数据中提取具有特殊写法的信息。比如中文史料中常见的日期多以年号开始，后接年、月、日，学者编写正则表达式时就可以按此格式，把一批文本中的日期信息全部抽取出来。——译者注

② 例如 Tonio Andrade, *The Gunpowder Age: China, Military Innovation, and the Rise of the West, 900–1900*, Princeton: Princeton University Press, 2016。

这样的大型数字化研究计划可以用于大规模的考察，也可以让用户集中考察具体文本中的史料、人物、地点和修筑事件。当然，为了达到这种目标，历史学者需要面对一些新的挑战。他们要投入大量时间在数据审编之上。所有数字化计划都是如此，因为自动化方法不是完美的。另外，历史学者进行这种大规模分析的时候，可能要调整自己的期待值。我们进行数字化的研究时，需要调整研究的习惯。例如，当进行大规模分析时，可能要忍受试验和失败，也可能得容忍一定程度的不准确数据和凌乱。而且，这种工作也不可能是单枪匹马可以完成的。要得到新的认识，就必须得到不同专长的学者合作，一起发展适合人文研究的数字化研究方法和平台。^①

三 关于成为开发者

也是因为体会到这一点，引发我主持开发“码库思”（MARKUS），一个用于古代汉语文本分析和阅读的平台。^②我过去在申请一个研究政治沟通的项目时，决定聘用一位计算机专业的博士后，以我们对宋代私人著述的标记为基础，主力开发可视化和分析信息网络的工具。^③当这个项目行将结束之时，我可以断定花这么多时间和资源在数字化研究之上是让我非常满意的决定。历史学者通常会把建立网站、数据库和电子平台的工作假手于他人，承包给学校内部或学校以外的开发商。这些开发商通常只能投入有限的时间，而且会利用现成模板来解决他们客户需要处理的问题。当我和一位全职的博士后、对人文研究感兴趣的计算机专家合作时，我终于可以开始处理我研究《挥麈录》时遇到的问题。这些问题属于中国数字化研究中的一些关键的结构性问题。

第一，正如我在别处指出的那样，^④20世纪90年代初以来就有一些大型商业公司建立数据库，为历史学者提供中文古籍的全文数据库。不过，

① 关于合作方面的问题，参看我在这篇博客文章的讨论：<https://cpianalysis.org/2016/06/09/collaborative-innovation-and-the-chinese-digital-humanities/>。

② 网址为：<http://dh.chinese-empires.eu/beta/>。

③ 项目网站的网址为：<http://chinese-empires.eu/>。

④ 我曾在以下演讲更详尽地讨论这一点。参见：<http://leccap.engin.umich.edu/leccap/viewer/r/az07QY>。

它们没有在数据发现、可视化和文本分析等方面花太多精力，但这些全部是让我们能更好地利用电子媒介做研究的工具。这些数据库提供的检索功能仍然非常有限，检索结果的处理也很原始，提供的参考工具也不多，对于检索结果的文本输出也有诸多限制，让人难以接受。当下的研究者往往需要从多个文本库中搜集和处理材料，但这些数据库的设计仍然用著作种类、作者、书名、地点（例如对地方志而言）等限定学者的检索条件。^①不同数据库提供各有出入的文献种类，使得不同数据库仍在不断出现，但内容往往存在重复。它们之间没有相互操作性（interoperability），使得研究者难以从多个数据库取得它们需要的资料。

第二，那些为了给研究者在处理文本和整理笔记上提供方便的软件也存在一些限制。当我为了研究政治沟通而试用各种用来标注笔记和书信的方法时，我一开始用过一个为社会科学研究者设计、商业的量化分析软件，然后也改用人手标记过一些文件。前者包含一些方便的标记和可视化功能，但由于它是商业软件，所以它和其他数据库（对我们来说就是 CBDB 和 CHGIS）缺乏相互操作性。另外，由于它没有完善的输出功能，所以其他软件不能读取和使用它的文件。特定的文件格式也使得分享和可持续性变成问题。这些软件在处理东亚语言上也有问题，虽然这些问题近来逐渐得到解决。利用一些标准格式（例如简称为 TEI 的文本标码标准）手动地标记文件可以解决这些问题，但又会带来其他不足。例如，很多标记步骤是重复进行的，其实可以自动化处理；而且，由于编辑的界面很原始，这种手动标记方法对于研究者来说在视觉上不直观。

我很庆幸能聘用数字人文专家何浩洋做博士后。我和他首先做的是总结我在标记宋人笔记中采用的方法。在 CBDB 的基础上，我们发展出人名、地名、官名和时间信息的自动标记功能。得到这些帮助时，学者可以比使用一般数据库更完善地发现和分析文本，效率又比手动标记高得多。

我们在开发这个工具的过程中不断加入新的功能，并把它设计成更贴

① 相关讨论参见：<http://chinese-empires.eu/blog/isnt-the-siku-quanshu-enough-reflections-on-the-impact-of-new-digital-tools-for-classical-chinese/>。

近历史学者和人文学者的工作流程。我们加入手动标记的功能，使用户不只可以修订、加入标记，还可以更自由地定义标记的类别。我们在这平台上加入词典的参考工具，因为它不只是用来标记文本，也是给研究者用来阅读史料的（见图2）。我们加入了用于特定语言和领域的词典时，用户可以同时看到很多参考文献中的词条解释。系统里也有做笔记的功能，用户能够在其中记录对文本的翻译、笔记和写上待办事项（例如注明以后需要查证的地方）。此外，我们也加入关键词标记的功能，让研究者可以上传自己的关键词名单，甚至利用正则表达式、关键词索引（KWIC）或关键词生成器（一个利用词频分析关键词出现模式的工具）来分析文本。我们也提供了筛选功能，让用户可以利用标记来选取特定段落。

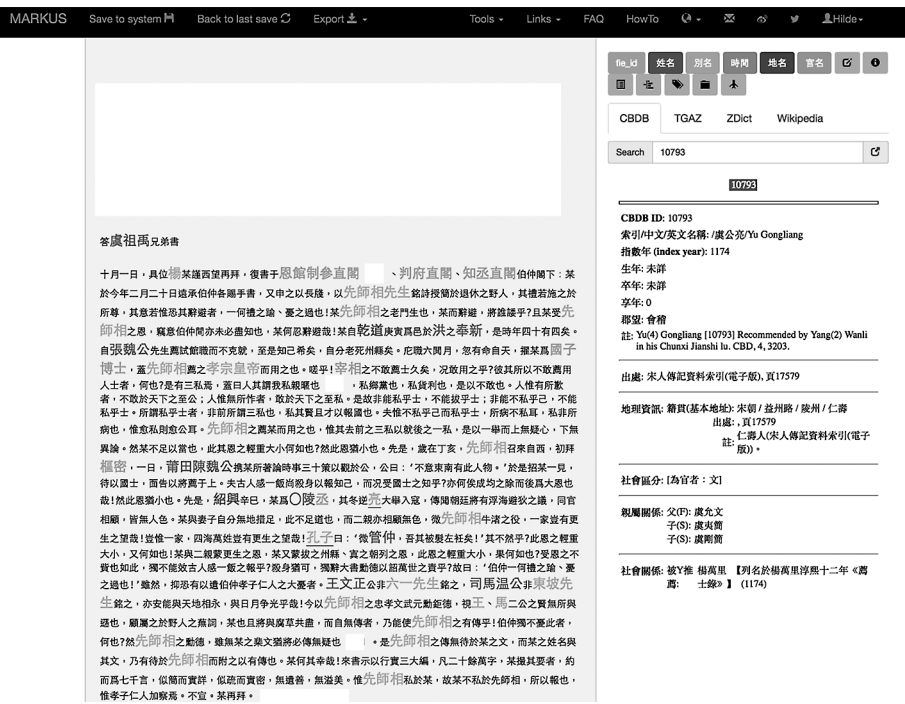


图2 码库思平台界面

图中右栏显示标记选项和参考资料。

此外，我们简化了提取文本的操作，整合码库思和其他数据库，并利用现成的软件来简化为数据进行可视化的过程。在两年时间之内，我们把

码库思变成一个关联的系统。在这个系统里，大量文本的标记和可视化可通过自动化操作。当我们把码库思储存的文件连到 Palladio 和 PLATIN 等可视化工具，^① 用户可以通过我们命名为 VISUS 的界面导入标记时提取的 CBDB 人物资料，并以地图、社会网络图、表格、时间轴、饼图和词云（word cloud）等形式检阅那些资料。用户也可以输出所有数据，利用专门的空间分析、网络分析、统计分析等软件，进行更复杂的分析。为了让导入文本变得更容易，码库思已经连到一些人们常用的开源文本库，例如德龙（Donald Sturgeon）建立的“中国哲学电子书计划”（Chinese Text Project）^②。我们未来还会将码库思跟维习安（Christian Wittern）的“汉籍リポジトリ”（Kanripo）^③ 连接。

我们在未来会继续让码库思灵活地在标记、阅读、探索、分析和做笔记等功能之间切换，更好地贴近研究者的工作流程。我在参与开发时学到很多——在设计这种研究的基础设施然后付诸实行的过程中，我更好地注意到数字媒介和方法的长处和短处。当人们具有批判性和建设性地参与其开发，为它的改善而努力时，可以使之更好地为人文学科服务。这种经验也让我体会到学界和其他界别之间合作的重要性。大部分新加入和未来会加入的功能都来自人文研究者和学生的建议。这些例子包括对社会关系的标记，允许认定不同信息之间存在各种关系；也包括使用数字化手段对文本进行比较，还有利用机器学习提升自动标记的准确率（precision）和召回率（recall）。通过跟国际上的计算机专家（如台湾大学的项洁和杜协昌）合作，我们也开发出同时标记多个文件的功能。

对历史研究来说，这种合作其实不是新事物。20 世纪初唐宋史研究的前辈们为了他们本身和学界的研究需要，编过不少索引、引得、词典和其他工具书，为中国历史研究的发展做出了很大的贡献。开发数字化工具，其实是这种合作的延伸和提升。如果我们希望我们学术单位的资源能完善地利用资源，就应该更积极地参与这种开发。当然，这种参与同时会带来

① Palladio 的网址为 <http://hdlab.stanford.edu/palladio/>。PLATIN 的网址为 <https://github.com/skruse/PLATIN>。

② 网址为 <http://ctext.org/>。

③ 网址为 <https://www.kanripo.org/>。

一些新挑战。在学术单位里找到办法让数字化服务能够运营下去，只是数字史学学者面对的众多挑战之一。学者也面对迫切的需要，为学生设计课程，确保他们得到充足的训练，了解人文相关的数字化研究方法、数字媒介的理论和批判，还有反思传统和数字化语文学的工具。

四 小规模文本标记

从我个人的经验来看，投入时间总结一个研究项目采取的数字化研究方法是值得的，尤其当其他学者采用这些方法在研究取得新进展时。^① 对于研究者本身，这也是值得做的事情。在我做完宋人笔记的研究之后，我开始了两个小型的研究计划——一个是研究杨万里的书信，另一个是对《贞观政要》的研究。前者是一个比较历史项目的一部分，该项目的目标是通过分别细读文人、教士的书信，比较他们政治参与的模式和影响。后者是一个集体项目，目标是把《贞观政要》译成英文，作为“剑桥政治思想史原著系列”（Cambridge Texts in the History of Political Thought）的新书出版。^② 我们开发码库思对这些项目的工作是有帮助的——码库思在经过改良之后，更适用于我们处理的问题。这启发我们继续进行开发，让历史学者能够做小规模的分析。

为了更好地理解杨万里的将近 500 通书信，我决定结合传统做阅读笔记的方法和对文本进行多角度的数字化分析。我把所有书信题目和本文里提到的人名都做了标记，还另外加上了书信的时间和地点，以及收信人的官位（从杨万里的文集的现代整理本和年谱而来）。除了这些元数据（metadata）以外，我还加入了书信的主要话题。我在码库思可以把这些我在阅读书信时做的笺注转成对所有文本的概览，并利用地图、时间轴、社会网络图、表格、标记和词云等，浏览时间、地点、收信人、主题、书信里提到的任何地点、收信人的官位等。我可以把这些跟 CB-DB 导出的数据结合起来，既可以宏观地检阅所有数据的情况，也可以集中地看某时段、主题或具体的书信内容（我们基于 250 通书信的样本分

① 一些文章提供了例子，参见 <http://dh.chinese-empires.eu/forum/category/8/research-blogs>。

② 参看 <http://admin.cambridge.org/academic/subjects/philosophy/philosophy-texts/series/cambridge-texts-history-political-thought>。

析出来的结果见图 3 ~ 图 6)。这样的工作有部分可以在 Excel 软件中完成，但是能把元数据、全文和各种可视化的选项结合，是灵活度高得多的做法。我不断因为这种新方法而重新检讨我过去在没有这些工具时细读史料获得的印象。从我这里附上的截图已经可以看到我们得出的一些整体观察（例如宰执作为收信人的现象非常突出，从吉州发到京城的书信也非常多，等等）。不过，由于这是一项仍在进行的研究，我这里希望说明的只是这种数字化研究方式在方法上和语文学上的优势，而非结论本身。

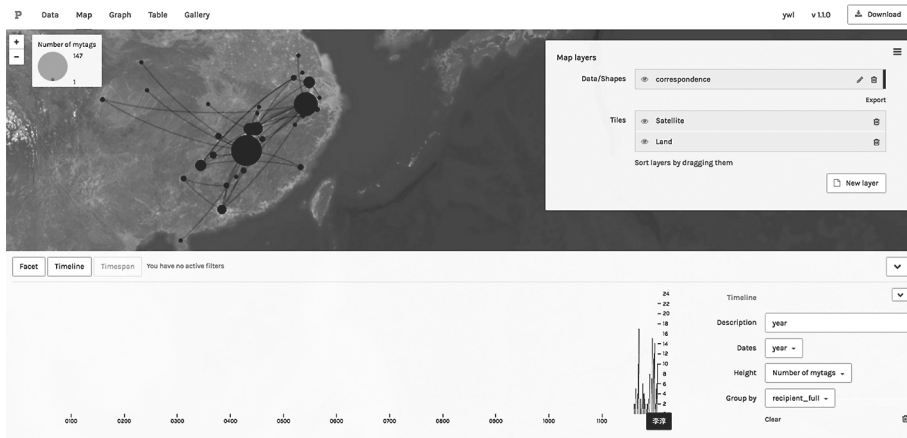


图 3 在码库思标记杨万里书信（其中 250 通作为样本）之后，利用 VISUS 导出资料，在 Palladio 分析其地域和时代分布

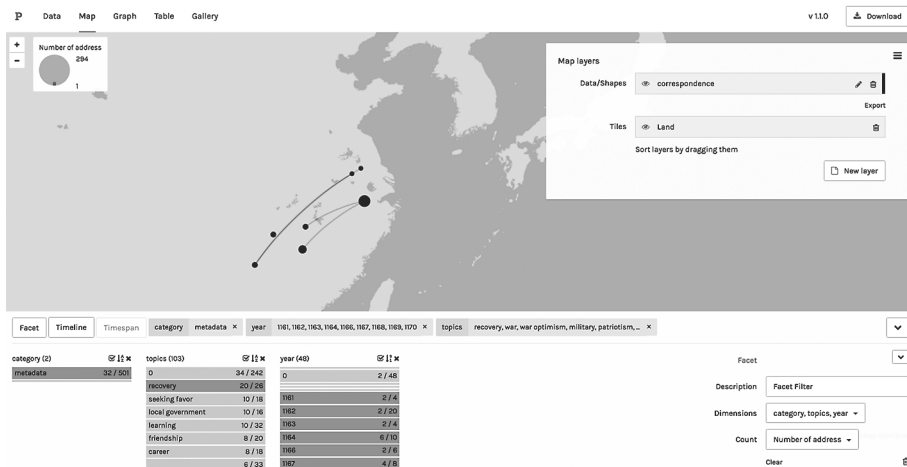


图 4 从 VISUS 导出标记过的杨万里书信（其中 250 通作为样本）之后，Palladio 地图视图中的分面浏览（faceted browsing）界面

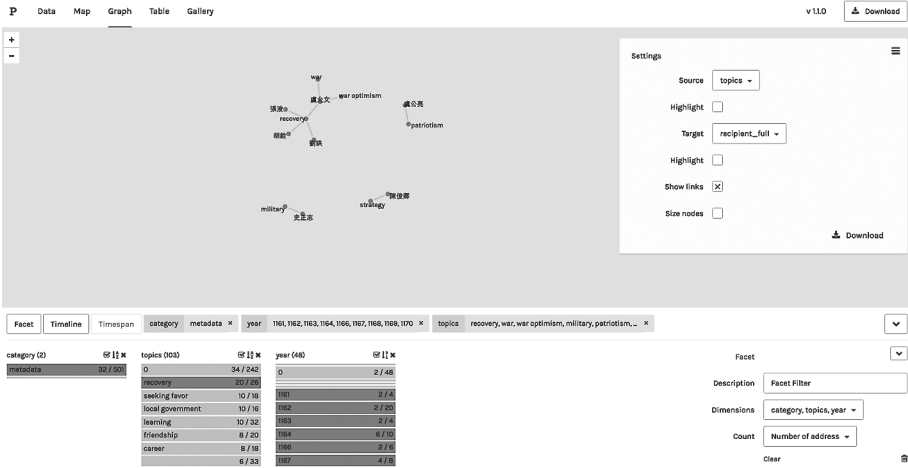


图 5 从 VISUS 导出标记过的杨万里书信（其中 250 通作为样本）之后，
Palladio 社会网络视图中的分面浏览界面

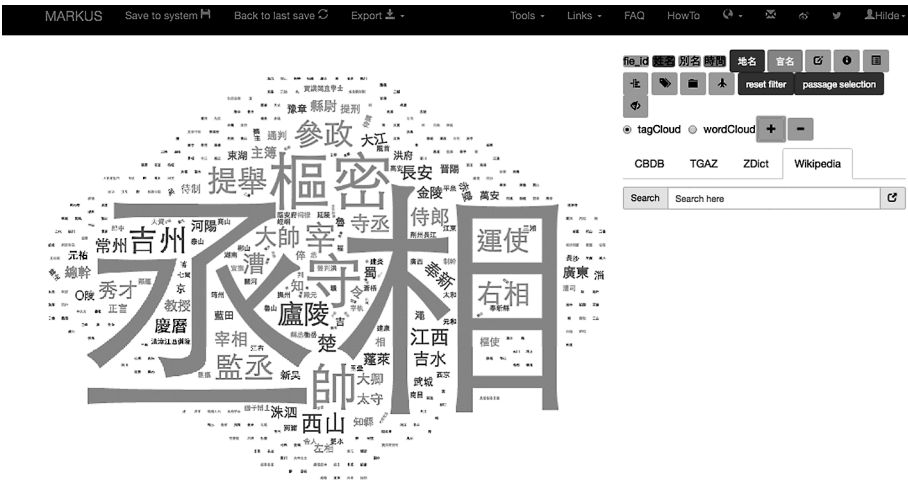


图6 从VISUS导出标记过的杨万里书信(其中250通作为样本)之后,生成出官名和地名的词云

在我们翻译《贞观政要》的过程中，我们进行标记的目的不同，主要是为了编辑文本。由于翻译是集体进行的，我们对整部书进行标记，以进行标准化处理。标记的内容包括官名、地名、人名、书名的各个列表。我们也编制了一个灵活度很高的电子版，让读者可以用更多不同方式读这部书——这么一来，书中段落既可以用时间顺序来排列，又可以按它们在不

同版本中的次序来排列。

在数字化研究中，很多进展都是研究开始之前无法预料的，往往令人惊喜不断。这些可能会吸引学者更投入使用数字化语文学的研究方法。在标记王明清《挥麈录》的信息来源人之后，为了看看该书提到什么人，而且都是哪些人提到的，我把其中出现的所有人名都标记了一遍。通过利用不同方法来衡量这些人物的中间性（centrality），我既可以据此印证我本来的估计（例如蔡京的中间性应该很高），又可以确认文本中不太明显的一些现象（例如曾布和宋高宗），甚至可以探索重要人物是跟何种主张联系在一起。关于他们的主张，我们不只可以通过对术语做考察，还可以通过分析群集（clustering），看某些人对其他人物的依附。这些发现可以帮助我们开发新的算法，预测文本包含什么偏见，并帮助历史学者判断他们的研究里需要包括哪些史料，并将之放到更广的思想和政治语境中。

结 语

我希望以上内容能够清晰地说明一点：数字化的研究方法可以用于多种历史课题的研究，从特定文本的细读到大批文本的探索和分析，到利用成千上万篇文字来绘图，其中包含各种丰富的可能性。数字化的分析未必需要成为量化史学，我们现在借助数字化分析可以写出多种多样的历史著作。我们可以利用数字化分析寻找历史文本和物件的内在结构和信息，比如我过去第一个历史地理信息（GIS）方面的研究成果就曾把《历代地理指掌图》凡例中的读法在电子地图上呈现。^① 我们也可以试试实现社会科学化的史学，例如利用社会网络分析、取样、概率方法等研究工具，以测试过去的结论，提出新的问题，突出过去研究中的空白，并提出新的解释，等等。而且，我们在这过程中可以继续使用并丰富研究文史的传统方法，这些方法从近代早期就已经是史学研究的基石。

① 参见：<https://ora.ox.ac.uk/objects/uuid:c0c281cd-4c2d-46ea-94b0-f7758cf0c3e6>。

参考资料

笔者的博客文章：

Collaborative Innovation and the Chinese (Digital) Humanities. University of Nottingham China Policy Institute Blog, June 9, 2016, <https://blogs.nottingham.ac.uk/chinapolicyinstitute/2016/06/09/collaborative-innovation-and-the-chinese-digital-humanities/>.

Isn't the Sikuquanshu enough? Reflections on the impact of new digital tools for classical Chinese. Communication and Empire: Chinese Empires in Comparative Perspective, Feb. 20, 2014, <http://chinese-empires.eu/blog/isnt-the-siku-quanshu-enough-reflections-on-the-impact-of-new-digital-tools-for-classical-chinese/>.

Digital Interpretations. Communication and Empire: Chinese Empires in Comparative Perspective, Feb. 5, 2014, <http://chinese-empires.eu/blog/digital-interpretations/>.

访谈：

魏希德：《我学术生涯的关键人物和关键词》，澎湃新闻，2017年2月9日，http://www.thepaper.cn/newsDetail_forward_1611403。

魏希德：《如何将数位人文工具 Markus 用于历史研究》，澎湃新闻，2017年2月10日，http://www.thepaper.cn/newsDetail_forward_1611410?from=singlemessage&isappinstalled=1。

在线工具：

笔者 *Information, Territory, and Networks: The Crisis and Maintenance of Empire in Song China* 一书的网站和相关的可视化工具，2015，<http://chinese-empires.eu/reference/information-territory-and-networks/>。

播客节目：

“Hilde De Weerd on MARKUS.” 2016年7月31日，DH East Asia Podcast, <http://www.dheastasia.org/2016/07/31/podcast-3-hilde-de-weerd-on-markus/>。

软件：

码库思 (MARKUS): A markup, reading, and visualization platform for classical Chinese texts (与何浩洋共同开发), 2014, <http://dh.chinese-empires.eu/beta/>。

码库思中文教学视频: http://dh.chinese-empires.eu/beta/video_zhcn.html。

笔者的演讲视频：

Digital Perspectives on Middle-Period Political History, 2016年4月5日，密歇根大学中国研究中心: <https://youtu.be/2oxHTEFEa38>。

(其他版本, 哥德堡: http://media.hum.gu.se/filedb/index.php?cdir=TmpVNU1qZz0%3D&c_hash=62c1644730fcf46086164dc08fdcf5e8 和 http://media.hum.gu.se/filedb/?cdir=TmpVNU1qYz0%3D&c_hash=41c9fba5f4490c7c0501ac047752a02b。斯坦福: <https://vimeo.com/168242706>。)

Humanities Tools for Library Resources, 2016 年 4 月 4 日, 密歇根大学图书馆: <http://leccap.engin.umich.edu/leccap/viewer/r/az07QY>。

《文本标记与历史研究》, 2015 年 4 月 29 日, 台湾中研院历史语言研究所: https://www.youtube.com/watch?v=NltG3EjC9_A。

《宋代新资讯结构的形成》, 2015 年 4 月 27 日, 台湾大学中文系: https://www.youtube.com/watch?v=1Xd_mJ9eJHk。