DOI:10.13530/j.cnki.jlis.2022013

数字人文视域下的古籍数字化与古典知识库建设问题*

陈　力

摘　要　“数字人文”是在传统人文研究中引入数字技术及方法来研究人类社会各种文化现象的新型跨学科研究范式。在涉及古代社会与文化研究领域,数字人文研究除了采用人工智能、大数据分析等研究方法以外,还需要一些基础条件,包括如何让计算机利用和理解古代文献和古代文化,古籍数字化和古典知识库建设就是数字人文研究所必须的基础条件。古籍数字化主要涉及两个方面的问题:一个是计算机编码汉字,尤其是异体字和异形字的编码问题;另一个是计算机图像识别并转换为字符的能力问题。古典知识库是对古籍内容进行数据统计、信息和知识挖掘的基础,需要从语言、时间、地理、体系化、联结古今等维度筹划古典知识库建设,以助力数字人文的研究。图1。表1。参考文献16。

关键词　数字人文　古籍整理研究　古籍数字化　古典知识库

分类号　G256

Digitalization of Ancient Books and Construction of Classical Knowledge Repository from the Perspective of Digital Humanities

CHEN Li

ABSTRACT

“Digital humanities” (DH) is a new paradigm of interdisciplinary research which introduces digital techniques and methods in traditional Humanities studies to study cultural phenomena in human society. At present, “automatic punctuating”, “automatic tagging”, “automatic translating” and “automatic collating” of ancient books labeled as DH in traditional cultural research are paid great attention to. Yet necessary basic conditions of the research are deficient, such as the large-scale corpus and related knowledge tools. As a result, essential data of DH projects are almost self-made, which is not only inefficient, but also unexchangeable and incompatible. It also contradicts ideas of big data and the interdisciplinary research that are advocated by DH. Accordingly, this paper makes some preliminary discussion of basic conditions of DH research in traditional cultural domain.

The most basic conditions of DH research are the digitalization of ancient books, which in essence is the digital conversion of physical books, and OCR of texts is the key process of digitalization. The OCR of texts is to recognize and convert images of Chinese characters using computer encodings. It involves two aspects: one is

*本文系国家社会科学基金重点项目“中国文献学史”(编号:19ATQ003)的研究成果之一。(This article is an outcome of the key project “Research on History of Chinese Philology” (No. 19ATQ003) supported by the National Social Science Foundation of China.)

通信作者:陈力,Email:chenli@nlc.cn,ORCID:0000-0002-8309-4510 (Correspondence should be addressed to CHEN Li, Email:chenli@nlc.cn, ORCID:0000-0002-8309-4510)

computer encodings of Chinese characters, and the other is computer image recognition and conversion to characters. In view of current techniques, we believe that apart from OCR of texts, to improve the recognition rate also needs techniques in corpus, vocabulary, knowledge repository, artificial intelligence and final manual recheck. Additionally, as to the standards and unification of characters in ancient books, a possible solution is to establish a comparison table of standardized form of Chinese characters and their variants for computers to do data analysis and character input and output.

The requirements of textual processing of digitalization of ancient books in a usual sense and of DH research are different. The former generally demands reproduction of the original content, while the latter requires processing of content, including data statistics, information and knowledge mining and so on. However, the processing of punctuations, quotations, names of persons, places and books in ancient books is quite complex and diversified, which calls for a large-scale corpus to support. The knowledge repository is an intelligent system based on knowledge. The uppermost function of classical knowledge repository is to connect the ancient and modern times and help computers to understand ancient culture and literature correctly. The objectivity of "knowledge" is emphasized, which provides interpretation of the "knowledge" in accordance with the ancient thinking and culture based on the understanding of ancient culture. We should give priority to knowledge repositories from the perspectives of language, time, geography, systematization as well as attaching the past to the present. Knowledge repositories include not only those of names of persons, places, functionaries, books and renowned items that belong to "explicit knowledge" but also those that have "tacit knowledge" characteristics such as repositories of academic histories, intellectual histories, literature, art and other "indefinable" repositories. 1 fig. 1 tab. 16 refs.

KEY WORDS

Digital humanities. Ancient books collation and studies. Digitalization of ancient books. Classical knowledge repository.

1　问题的提出

"数字人文"是在传统人文研究中引入数字技术及方法来研究人类社会各种文化现象的一种新型的跨学科研究范式。"数字人文"不是简单的人文学科资料的数字转换,也不仅仅是在研究中引入数学建模、套用数学公式,而是利用数字技术和方法,通过对文献的数字化处理并进行数据分析、信息和知识挖掘从而获得新的知识,或者使碎片化的知识系统化、使隐性化的知识显性化。

目前,国内"数字人文"研究如火如荼,尤其是在中国传统文化研究领域,这对于繁荣人文社会科学研究、促进"新文科"的发展,有着积极的作用。

不过,在"数字人文热"的背后,一些基础性的也是深层次的问题逐渐显现出来。在中国传统文化研究领域,冠以"数字人文"的古籍"自动标点""自动标注""自动翻译""自动校勘"等倍受重视,但却缺乏这些研究所必须的基础条件,如大型语料库和相关的知识工具。由于缺乏必要的基础条件,许多数字人文研究项目的基础数据几乎完全靠自己准备,不仅效率低下,各项目之间数据不能交换、互用,而且与数字人文研究本身所要求的大数据、跨学科理念相悖。

数字人文研究所需要的条件很多。就知识库而言,有一些是专业性的,还有一些则是通用性、基础性的,通用性、基础性的知识库如何建设,需要学术界充分讨论。本文仅就中国传统

文化领域数字人文研究所涉及的基础条件做一点初步的探讨。

早在20世纪40年代计算机刚刚发明时,人们就开始考虑利用计算机把一种语言翻译成另外一种语言。在中国,著名的藏学家于道泉先生就曾进行过这方面的研究,并于1956年发表了《谈谈翻译机械化》[1],1982年发表了《藏文数码代字》[2]。不过,于先生的研究主要还是直接通过不同语言之间词汇和语法的对应关系来进行处理,至于藏文数码代字,则是将每一个藏文字母赋予一组不同的数字,其原理与今天的Unicode编码相同。

1973年,国外就有学者利用计算机进行甲骨残片缀合的研究,但影响不大。1975年,四川大学考古学家、科幻作家童恩正先生与计算机专家张陞楷、陈景春合作,继续尝试利用计算机对甲骨碎片进行缀合,并发表了《关于使用电子计算机缀合商代卜甲碎片的初步报告》[3]。在童先生等人的研究中,他们设定了六项限制条件:时代、字迹、骨板、碎片、卜辞、边缘,利用条件匹配的原理进行处理。实验选择了263片甲骨碎片作为样本,最后的缀合率在40%上下。这个结果并不太理想,因为"用人工录制标本信息工作量大而且不准确",同时,当时的计算机设备也不够先进。不过无论如何,童恩正先生等人的研究对我们今天来说仍然是有启示意义的。

1980年,美国华裔学者陈炳藻首先利用计算机对《红楼梦》的词汇使用习惯进行了统计分析,最后得出了结论:根据前八十回与后四十回词汇使用的频率比较,两部分基本上是相同的,因此,前八十回与后四十回的作者可能都是曹雪芹[4,5]①。

可以说,童恩正先生和陈炳藻先生的研究,有的内容已经与今天我们所说的"数字人文"研究十分接近了,也是通过对数据的处理,从而发现和获得新的知识,只是当时没有使用"数字人文"这个概念而已。他们当初的研究和研究方法没有能够进一步深入、提升,与他们当时所处的时代和技术条件等因素有关。例如,童恩正先生他们所利用的计算机及其相关设备性能低下,远不能与今日相比,同时,他们所能利用的甲骨只有区区263片,涉及确定时代、字迹、卜辞内容的基础数据量过小,并且都依靠他们自己人工进行预处理,这样,当进行基于内容的数据分析、挖掘时,就显得捉襟见肘,最后得出的结论可信度也不高。同样,陈炳藻先生的研究,仅仅是根据《红楼梦》一百二十回中用字用词的数据统计、对比来进行,用于比较研究的文献也仅限于满族文学家文康的《儿女英雄传》,如果能够把数据分析的范围扩大,也许会更有说服力。可惜的是,陈炳藻先生当时并不具有这个条件,即使是《红楼梦》一百二十回的七十多万字的文本,也是陈先生自己处理的。

童恩正、陈炳藻先生的经历告诉我们,开展数字人文研究,尤其是涉及大数据处理,需要一些基础条件。如果缺乏这些基础条件,研究是难以进行的。

2 "古籍数字化"问题

数字人文研究最基础的条件莫过于古籍数字化。利用数字技术对古籍文本进行处理,我们通常称之为"古籍数字化"。古籍数字化的本质是对实体古籍进行数字转换,内容包括:古籍计算机编目、古籍影像扫描、古籍文本识别、古籍利用辅助工具研发等。古籍数字化是一个系统工程,从古籍的著录、主题与分类标引、文本识别到查询利用、网络传播与共享等各个步骤、各个环节都属于这个系统的一部分,需要高度

① 1987年,陈大康先生发表了《从数理语言学看后四十回的作者——与陈炳藻先生商榷》(《红楼梦学刊》,1987年第1辑,第293-318页),得出了与陈炳藻先生不同的结论。

集成、无缝衔接。遗憾的是,国内不少古籍数字化项目由于业务管理和流程的条块分割,编目、扫描、文本转换、平台发布等工作多是由不同的团队进行,协调不够,有些功能缺失,标准化和开放性不够,很少考虑其他研究项目包括数字人文研究的数据调用和资源共享。

古籍文本的OCR(光学字符识别)一直是古籍数字化的关键环节,主要内容是用计算机所能使用的编码汉字与古籍中的汉字图像进行识别和对应转换,目的在于使计算机能够对文献内容进行处理。它涉及两个方面的问题:一个是计算机编码汉字的问题,另一个是计算机图像识别并转换为字符的能力问题。

关于第一个问题,从理论上说,要进行古籍文本的识别,则古籍中出现的所有汉字在计算机中都应有相应的编码,并且在以后的发布、利用环节能够被有效使用。这个问题看起来简单,但实际上却非常复杂,原因就在于汉字本身的复杂性。

二三十年前,计算机所使用的汉字编码字符集收录的字数偏少是困扰古籍数字化的主要因素。1980年颁布的国家标准GB2312收录的汉字仅6 763个,1993年颁布的GB13000.1-93(GBK)也只收录了21 003个汉字,2000年3月颁布的GB18030收录了27 484个汉字。到2005年颁布的GB18030-2005,已收录汉字70 244个,到2021年9月,Unicode14.0版正式发布,其中收录的汉字已超过了9.3万个,汉字字符太少的问题已基本解决,但一个新的问题又出现了,这就是古籍用字的规范与统一。

据民间学术网站《汉字宝典》的不完全统计,在古籍中出现过的汉字包括异体字已超过15万个,但如果算上古籍各种写本、印本存在着大量写法有细微差别而导致计算机OCR识读时可能发生偏差的"异形字",数量将会成倍地增加。无限制地增加字符集中汉字的数量并不能解决古籍的用字问题,反而会带来新的问题。

古籍的用字,本来是不多的,下面的统计可以给我们一个大致的印象(见表1)。

表1　古籍用字统计①

书　名	成书年代	总字数	不重复用字
易经	先秦	21 055	1 363
尚书	先秦	28 073	2 025
诗经	先秦	37 438	2 989
论语	先秦	15 935	1 355
孟子	先秦	35 417	1 913
仪礼	先秦	56 809	1 529
左传	先秦	195 792	3 290
荀子	先秦	75 815	2 726
风俗通义	东汉	58 843	3 350
论衡	东汉	212 050	3 630
列子	晋	30 900	2 329

① 以下各书据《先秦两汉古籍逐字索引丛刊》(香港中文大学中国文化研究所编,台湾商务印书馆1992—2002年出版)统计。不同的版本,统计数容有差异。

但是,当我们在进行古籍 OCR 时,所需要处理的字形达数十万个,其原因主要是古籍在抄写、刻印的过程中,除去抄错、刻错的以外,抄写者、刊刻者的书写习惯差异造成了一个汉字对应多个字形的情况。例如,在敦煌写本中,大量使用"俗字",如"多"有 14 种字形,"恶"有 27 种字形[6]。这主要是由于敦煌写本的书写者大多是文化水平不太高的经生,多一笔少一笔、长一划短一划十分常见,有的书写潦草,胡乱连笔、任意变形。古代书坊刻书也大量使用俗字,与国子监等官方刻书机构通常使用"正字"不同,这主要是由于书坊刻书的主持者文化水平不高、出于节省成本等商业因素,而读者对这些问题又不是十分敏感。明代刻书,常常使用"古字",并且很多时候是自己生造出来的古字,以附庸风雅、故弄玄虚。正是由于这些原因,《史记》本来使用的汉字不到 5 200 个,但是,如果算上不同的版本,累计起来,《史记》各种版本使用过的汉字字形,远远超过此数。还有一个问题就是避讳。除了皇帝之讳外,最难处理的是民间个人私讳,因为使用私讳的通常只有本家族之人,私讳缺笔、变体字大量存在,也造成了汉字系统越来越复杂,字形越来越多。

关于第二个问题,一方面,古籍字形过多,有些差别极其微小,并且很容易与其他字相混淆;另一方面,计算机的文字识别能力有限。二者之间的矛盾需要另辟蹊径,寻求解决的办法。

汉字的 OCR 技术经历了几十年的发展,对于现代排版印刷体图书来说,已经进入了实用阶段。但是,对于古籍来说,则远远没有达到可用的程度。受古籍复杂多变的汉字字形、复杂的版式等因素制约,目前,对于比较简单、规范、整齐的古籍,四川大学与阿里巴巴达摩院共同研发的"汉典重光"OCR 平台,其正确识别率也只能达到 97.5%,换言之,其差错率是万分之二百五十,是国家图书出版文字差错率标准的二百五十倍[7,8],而这是目前业内最高的识别率。

古籍 OCR 技术的难度在于古籍版面复杂、文字字形多样、大小字注及行间注并存,至于随文批校等就更不必细说,这些都是在进行 OCR 识别时必须要面对的问题。

完全依靠 OCR 来解决古籍的文字识别,从目前的技术来看,是不可能达到最低的国家标准的。因此,笔者认为,要提高计算机对于古籍的文字识别率,还需要结合语料库、词表及知识库、人工智能等多方面的技术,最后再辅以人工复校,这其中语料库、词表知识库可以发挥很大的作用,并且可以作为工具库为数字人文研究的其他方面所互用。举例说明,在古籍中,除了"一字多形"以外,"一形多字"的情况也十分普遍。所谓"一形多字",即一个字形对应多个不同文字,如"二",可以对应数字 2,也可以对应"上下"的"上",也可以作为重复字符号。又如"于",其字形有:

亏　亏　亏

"亏"字形除对应"于"外,还可对应"考""巧";"亏"字形除对应"于"外,还对应亏欠的"亏"。有一些手写字,只有细微的区别,例如:"憧"与"幢"、"枸"与"拘"、"麈"与"麈"。还有一些字,由于字形过于接近,再加上古代刻字工人书写雕刻习惯、雕版磨泐、纸墨低劣、刷印模糊和鼠啮虫穿等情况,使得计算机在识别时会发生误读,如"芊""芊""芋""芋"。所有这些,要让计算机能够正确地识别,必须要根据上下文关系来判断,特别是借助古籍语料库、古代人名库、地名库、职官名库、名物库、各种语言类词典(如成语词典、俚语词典、方言词典等),让计算机可以根据上下文的关系来学习、补齐字句。目前,海内外都有不少语料库①,只是太过分散,并且开放性不够。中华书局自 2003 年起即着手建设"中华古籍语料库",海外学者在建设相关工具库方面也做出了很大成绩,如中国历代人物传

① 参见:https://www.zhihu.com/question/20447189,查询于 2021 年 11 月 15 日。

记数据库(CBDB)①。

还有一个问题,就是古籍用字的规范与统一。当我们阅读古籍时,在多数情况下,对古籍中的异体、异形字并不那么敏感。计算机则不同,每一个字符都有一个独立的编码,当我们进行文字检索时,计算机后台是根据其编码来进行匹配的,每一个字形都有一个独立的 Unicode 编码,如:

"峰(5CF0)"或作"峯(5CEF)"
"群(7FA4)"或作"羣(7FA3)"
"略(7565)"或作"畧(7567)"
"曾(66FE)"或作"曾(66FD)"
"黑(9ED1)"或作"黒(9ED2)"
"高(9AD8)"或作"髙(9AD9)"
"舍(820D)"或作"舎(820E)"
"頤(9824)"或作"頥(9825)"

像"峰"与"峯"、"群"与"羣"、"略"与"畧",笔画甚至笔顺都是完全一样的,只是偏旁部首位置不同,但它们都有各自的 Unicode 编码,在计算机中,它们完全是按不同的字符来处理的。还有一些字,虽然是同一种书,但有不同的版本,每个版本的用字都可能不同,如古籍中最常见的"历"字,常见的有以下几种写法:

曆(66C6)　歷(6B77)　歷(6B74)
厤(53A4)　歷(20AB1)

"曆"本义为"历法""年历",在古籍中使用频率最高的是年号,如"聖曆""大曆""寶曆""鳳曆""慶曆""天曆""萬曆""永曆"。但是,由于清乾隆皇帝名"弘曆",清代刊印的古籍中往往将"曆"改为"歷"或"厤""歷""歷"。它们之间只有细微的差别,但 Unicode 编码完全不同,如果在一个古籍数据库中,同一种书可能有多种版本,或者引用同一句话,有的可能作"曆",有的可能作"歷",有的可能作"歷",有的可能作"厤",有的可能作"歷"。当进行全文检索时,用什么字作为检索词,就成了一个问题。

一个可行的解决办法,就是建立一个汉字正字与异体字、异形字的标准对照表,以供计算机进行数据分析和汉字的输入、输出使用。2013 年 6 月 5 日,国务院公布了《通用规范汉字表》,该表共收录汉字 8 105 个,全部为简体字。《通用规范汉字表》还有附表《规范字与繁体字、异体字对照表》,但只收录了与 2 546 个规范字相对应的 2 574 个繁体字,对于古籍数字化来说,这是远远不够的。2021 年 10 月,国家标准《古籍印刷通用字规范字形表》(GB/Z 40637-2021)发布,该标准规定了古籍印刷通用字收字和宋体字形规范原则,给出了 14 250 个古籍印刷通用字的字形、字音以及在国际编码字符集 ISO/IEC 10646 中的码位,适用于传世古籍的印刷出版,以及现代书刊的繁体版印刷。这个标准对于古籍出版来说是十分必要的,但对于古籍数字化来说,可能还是不够的,需要进一步扩充。十年前,国家曾启动过大型数字化基础工程"中华字库工程"②,主要是解决汉字和少数民族文字字形(包括古文字)的标准化及其输入、输出问题。搜集、整理的汉字包括:甲骨文、金文、简牍帛书及其他古文字、石刻,行书、草书、版刻楷体字、宋元及明清印本文献用字,现代出版物用字及符号,少数民族古文字及现行文字,等等,也提到了要制作一个中间字库,但立项时未确定承担单位。一方面,"中华字库工程"现在尚未结项,如何建立一个正字字库(中间字库)并与其他不同字体字形文字关联起来,似乎还暂未有结果。因此,作为数字人文以及古籍数字化的基础条件,需要尽快建立一个适度且基本够用、标准的汉字对照表。

古籍数字化还有一个基础条件,就是对古籍进行标准化计算机编目。所谓"标准化",是指所有的古籍收藏机构、个人,在进行古籍计算机编目时,都应当使用标准的语言编码、编目格

① 参见:https://projects.iq.harvard.edu/chinesecbdb,查询于 2021 年 11 月 27 日。

② 参见:"中华字库工程",https://baike.baidu.com/item/%E4%B8%AD%E5%8D%8E%E5%AD%97%E5%BA%93%E5%B7%A5%E7%A8%8B/9193968,查询于 2021 年 10 月 28 日。

式、著录规范,包括统一分类、主题词或关键词等。古籍计算机编目之于数字人文研究,其意义不仅在于揭示馆藏,方便人们发现和获取,它还给数字人文研究建立了一个文献的标准体系,防止因版本不同而出现结果的误差。同时,古籍编目数据库本身也可以作为古籍数字化、数字人文研究所需要的词表或知识库。目前,由国家古籍保护中心组织开展的全国性古籍普查已近尾声,各藏书机构的古籍普查登记目录正陆续出版,期待一个统一、规范的全国古籍联合目录早日问世。此外,《中国古籍总目》(网络版)出版工程(一期)也已由国家正式立项。

至于古籍的分类、主题词或关键词标引,涉及的问题更多,当另文讨论。

3 "古典知识库"问题

一般意义上的古籍数字化与数字人文研究对古籍文本处理的要求是不同的。在进行古籍数字化时,通常只需要忠实还原文献的内容即可;而进行数字人文研究,问题就要复杂得多了,需要对古籍的内容进行处理,包括数据统计、信息和知识挖掘等。例如眼下热门的古籍自动标点,涉及引文、人名、地名、书名的处理等问题,要让计算机系统判断哪些是引文,引文的起点在哪里、终点在哪里,后台就需要一个庞大的语料库来支撑。这看起来不算太困难,但实际情况常常并不是那么简单。因为古籍中的引文,有许多并不忠实原文(况且还有版本的问题,原文本来可能就不同),有的只是撮取大意而已,面对这类问题,可能需要后台有相关的知识库支撑,如成语典故知识库,古代的各种类书在这方面可以发挥重要的作用。至于人名、地名、书名,如果只是简单的自动标点,给出专名号即可,后台有一个简单的人名字号库、地名库、书名库,即可以帮助计算机分析判断哪些是人名,哪些是地名,哪些是书名,从而划出专名号。但是,如果我们需要利用古籍进行更进一步的数字人文研究时,面临的问题就更复杂了,因为古代人名字号相同的很多,需要判断究竟是哪个人的名字。早在梁朝,梁元帝就曾编纂过《古今同姓名录》,目的就是区分古籍中的同名人物。书名、地名相同的也很多,仅《隋书·经籍志》即著录了七家《晋书》、七家《晋纪》,加上《晋中兴书》《东晋新书》和裴松之《晋纪》、郭季产《续晋纪》,后世统称"十八家晋书"。地名是一个非常复杂的问题,同一地名,不同时期、不同时代可能在不同地方,全世界都是如此,这主要是由于族群迁徙、宗教、历史等原因造成的,如美国地名很多与英国及欧洲的相同。在中国古代,这种情况也十分普遍,从商人先祖契到商王盘庚之前,四处迁徙,其都城有"前八后五"之说,每迁一处,都称该地为"商",直到商王盘庚迁都于"殷"(今河南安阳小屯),才"更不徙都",同时也称"商"。如果不加区分,把甲骨文中所有叫"商"的地名都定位在今天的河南安阳小屯,显然是错误的。1914 年北洋政府内务总长朱启钤提交的《拟改各省重复县名撮取理由分别说明》指出:唐代有重名县十九个,宋代有三十个,明代有四十二个,到清末时"二县同名者七十四,三县同名者十有二,四县同名者四,五县同名者三,六县同名者一"[9]。至于像"五里店""十里铺""马鞍山"之类的地名,几乎各地都有。这类问题,都需要有相关的知识库作为支撑。

知识库(Knowledge Base),顾名思义,是基于知识的智能系统。社会知识是一个内涵十分庞大、复杂的系统,有的比较简单、清晰,例如人名、地名等事实类的知识,著名的"中国历代人物传记数据库"(CBDB)虽然结构复杂,但其内容是清晰、明确的;而有的则非常复杂,涉及思想、情感类的知识,其表达并让计算机能够调用是非常困难的,例如,梅、兰、竹、菊作为植物,它们的内涵是十分清晰、明确的,在建设相关的知识库时,很容易处理。但涉及它们在古代文化中的象征意义时,情况就变得复杂了。以梅为例,《永乐大典》中关于梅的专题占了六卷,涉及"梅"的词条共 90 个,包括不同品种、不同颜色、

不同形态的梅,也有以梅为名的水果及制品(按现代植物分类可能是错的,但这正是古代"知识"与现代"知识"差异之所在),这些属于古代所谓"名物"的范畴,可以看作是事实类的知识库;另外还有古人咏梅、写梅、画梅的艺术作品,可以把它们当作专业的语料库。这些,在建立知识库时,都比较容易处理。但是,梅在中国传统文化中的象征意义十分重要而复杂,像林逋《山园小梅》诗"疏影横斜"之梅与陆游《卜算子·咏梅》词"无意苦争春,一任群芳妒"之梅,其意蕴显然是不同的,这类文学作品常常是"只可意会,不可言传",而涉及哲学、思想范畴的概念,不同学者之间的差异、细微之处更不易区分。说到这一点,有一个观念需要明确:在古籍数字化、数字人文研究中,计算机终究只是辅助研究工具,并不能完全替代人的大脑。

知识库如何建设?需要考虑哪些因素?清华大学刘石、孙茂松教授曾经提出过设想:

> 借助于中国历史上一切古典学的研究成果,在保障古籍文献内容完整性及内部逻辑性的基础上突破文献原有结构,周密地设定主题词表,专业地提取各种实体,如年代、地域、人物、社团、著述、事件等等,构建实体的相关属性及不同实体间的关系,通过这些实体及相互关系对文献进行深层组织和知识管理,这就是我们所构想的"中国古典知识库"[10]。

这个构想无疑是十分宏大美好的。不过,从超过20万种古籍中提取各种实体并构建实体的相关属性及不同实体间的关系,涉及的问题、方面很多,考虑到实施的可能性,工作需要拆分,即在一个统一的框架之下,通过构建一个个具体的专业知识库,最后形成综合性的古典知识库集群。

中国传统文化中的"名物制度",包括事物如动物、植物、器物之名和历代各种制度、风俗习惯、思想观念,后人在整理古代文献时,由于时代的差异,需要借助一些工具作为参考。清代著名学者汪中说:

> 古之名物制度,不与今同也,古之语不与今同也,故古之事不可尽知也[11]。

另一位差不多同时的学者章学诚也说:

> 校雠之先,宜尽取四库之藏,中外之籍,择其中之人名地号,官阶书目,凡一切有名可治、有数可稽者,略仿《佩文韵府》之例,悉编为韵,乃于本韵之下,注明原书出处及先后篇第,自一见再见以至数千百,皆详注之,藏之馆中,以为群书之总类[12]。

同样,今天在进行古籍整理、古籍数字化以及数字人文研究时,也需要借助一些工具,只不过,传统时代这些工作需要人力完成,而今天我们可以通过计算机来完成,因此,古典知识库最主要的功能是沟通古今,让计算机正确地理解古代文化和古代文献。

知识库可以是某一专题的相关知识,过去常常又被称之为"专家库",它强调的是"知识"的客观性,它的功能是向人们提供客观、专业的知识。需要注意的是,"客观""专业"的知识未必是"科学"的,譬如巫术,巫术反科学的性质毋庸置疑,但巫术是中国古代政治与社会生活中的重要组成部分,不了解它是什么,不清楚巫术的概念、术语,就不能理解它在古籍中的意义,对人是如此,对计算机也是如此。除了巫术以外,古代的许多"知识""常识""习惯",在今天看来显然是不"科学"的,但它们却是中国古代文化的重要组成部分。因此,古典知识库所强调的不是"古典知识"的科学性,而是"古典知识"的客观性,即从理解古代文化特点的角度给予这些"知识"以符合古代思想、文化的解释。

人名知识库、地名知识库、职官知识库、书名知识库、名物知识库等是属于"显性知识"的知识库,还有一些是具有"隐性知识"特点,如涉及学术史、思想史、文学艺术以及其他"不可言状"的知识库,这些需要从以下维度重点考虑。

(1)语言的维度。中国历史悠久、地域辽阔,不同民族、不同时代、不同地区的语言十分复杂。东汉王充《论衡·自纪篇》说:"经传之文,圣贤之语,古今言殊,四方谈异也。"因此征集各地方言在古代是国家治理的一个重要内

容。“周、秦常以岁八月遣輶轩之使,求异代方言,还奏籍之,藏于秘室。”[13]西汉扬雄所撰《方言》,便是中国第一部专释方言之书。同一个字、同一个词,在各地有不同的含义,“党、晓、哲,知也。楚谓之党,或曰晓,齐宋之间谓之哲。”[14]东晋葛洪《抱朴子·钧世》说:“古书之多隐,未必昔人故欲难晓,或世异语变,或方言不同。”[15]方言的使用在小说和地方文献中最为突出,例如清末小说《海上花列传》就大量使用吴方言,而民国小说《死水微澜》就大量使用四川方言。不同方言所带来的最常见问题就是同物异名,如红薯在不同地区,或称甘储、甘薯、朱薯、金薯、番茹、红山药、玉枕薯、山芋、地瓜、甜薯、红苕、白薯、阿鹅、萌番薯等。一般的名物异称,可以参照传统图书馆学“名称规范控制”来解决。但是,语言问题非常复杂,古人论及如何理解六经,谓“古文读应《尔雅》,故解古今语而可知也”[16]。今天在建设知识库时,则需要语言学家的深度参与。

(2)时间的维度。几千年来,语言、名物制度都在不断地变化,同一字、同一词、同一名称,在不同时代有不同的含义。如“棉”,唐宋以前指乔木类的木棉,而唐宋以后则主要指草本的棉花;唐代以前的“糖”主要指饴糖,而唐代以后,则多指砂糖。“睡”,汉代以前专指打盹,《说文解字》:“坐昧也。”而先秦之“睡”作“寐”“寝”,《论语·公冶长》:“宰予昼寝”。“交通”,古义“勾结”“联系”“接触”,后来演变为道路相连,陶渊明《桃花源记》“阡陌交通,鸡犬相闻”,也与后世之交通运输不同。职官、地名古今变化也很大。如“尚书”一职,始设于秦,直到汉代,只是掌管文书的小吏,但到了明清时代,便成了六部首长。因此,在构建知识库时,需要考虑时间、时代的因素。

(3)地理的维度。在古代社会与文化研究中,地理是一个极其重要的因素,前面提到地名的变化,其实问题不止于此,即使是同一地名,由于时间的不同,一地的治所与四至也可能有很大变化。在处理方言、同名等问题时,地理也是一个最重要的维度。因此,以地理信息系统为基础或者通过关联地理信息系统来构建知识库是一个很好的解决方案,中国历代人物传记数据库(CBDB)通过关联地理信息系统就是一个很好的范例。

(4)体系化的维度。知识库是一个系统、一套体系,古典知识库实质上是一个古代文化的知识体系,并且这个知识体系一直处于发展变化之中。构建中国的古典知识库,既需要反映中国古代文化的特点,也要结合中国古代知识体系的结构、分类及其不同知识之间的相互关系来考虑。中国最早对知识进行有意识的分类可以追溯到《尚书》《诗经》各篇的编排上。战国时代成书的《禹贡》《山海经》可以看作是先秦时代关于地理的知识体系,在这个体系中,既有古人确实已经掌握的地理信息和地理知识,也有古人想像中的“世界”模型。大约西汉时期成书的《尔雅》,本是关于儒家经典的“名物训诂”之书,但用今天的眼光来看,实际上就是一个相当完整的上古知识体系(见图1)。

除《尔雅》以外,中国古代的史志、政书、类书等都可以看作是不同时代的专科或者综合性的知识体系。仅以古籍分类为例,西汉末年刘向、刘歆等人对皇家藏书进行系统整理后编制的《七略》,则是一个基于文献的知识体系,在这个知识体系中,知识被分成了六艺、诸子、诗赋、兵书、术数、方技六大类,其下又再细分。从《七略》到《四库全书总目》分类方法的发展变化,几乎就是从西汉到清代中国知识体系分类变化的一个缩影。同样,类书作为古代的知识工具,其内容与特点的变化,更是反映了社会知识结构的变化。关于这一点,我们将另为文讨论,不赘。

(5)联结古今的维度。中华文化是一个有着几千年历史、至今仍然充满活力的文化。古代文化发展到今天,有一个传承与弘扬的问题。因此,数字人文研究的一个重要使命就是将古代文化与当代文化连接起来,让研究者和普通公众更好地理解传统文化,真正让古籍中所蕴含

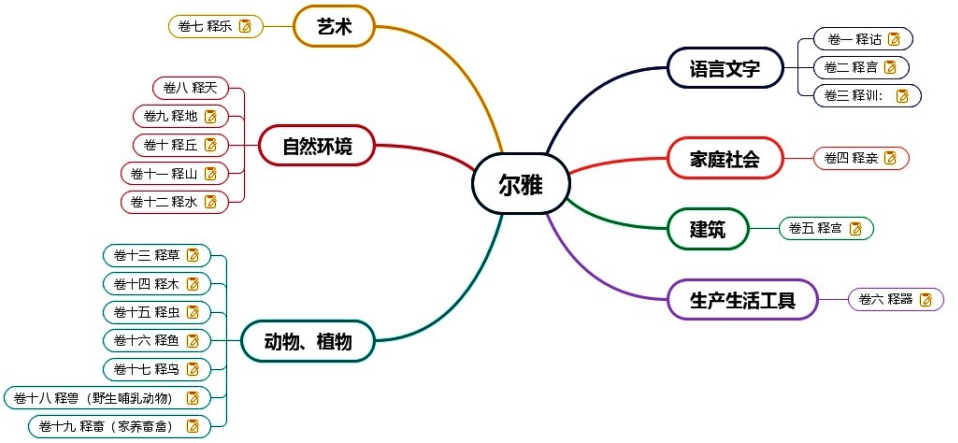

图 1　《尔雅》知识体系示意

的知识活起来。当然,也可以将外国的优秀文化连接起来,以达到中西合璧的目的。同时,几千年来,文化和学术研究一直都在不断地发展进步,如何在知识库中反映文化与学术发展的脉络,充分反映当代人文社会科学的研究成果,这也是构建古典知识库必须要充分考虑的问题。

参考文献

[1] 于道泉. 谈谈翻译机械化[J]. 中央民族学院院刊,1956(18). (Yu D Q. On translation mechanization[J]. Journal of Central Institute for Nationalities,1956(18).)

[2] 于道泉. 藏文数码代字[J]. 民族语文,1982(3):1-7. (Yu D Q. Tibetan digital characters[J]. Minority Languages of China,1982(3):1-7.)

[3] 童恩正,张陞楷,陈景春. 关于使用电子计算机缀合商代卜甲碎片的初步报告[J]. 四川大学学报(自然科学版),1975(2):57-65. (Tong E Z,Zhang S K,Chen J C. Preliminary report on the use of electronic computers for the conjugation of shell fragments[J]. Journal of Sichuan University(Natural Science Edition),1975(2):57-65.)

[4] 陈炳藻. 从词汇上的统计论《红楼梦》的作者问题[C]//周策纵. 首届国际红楼梦研讨会论文集. 香港:香港中文大学出版社,1983. (Chen B Z. The author's problem of *Dream of the Red Chamber* from the perspective of lexical statistics[C]//Zhou C Z. Chinese papers from the First International Conference on the *Dream of the Red Chamber*. Hong Kong: Chinese University of Hong Kong Press,1983.)

[5] 海炯. 首届国际红楼梦研讨会简况[J]. 社会科学,1980(5):156-157. (Hai J. Brief introduction of the first international conference on the *Dream of Red Mansions*[J]. Journal of Social Sciences,1980(5):156-157.)

[6] 黄征. 敦煌俗字典[M]. 上海:上海世纪出版集团,上海教育出版社,2005:96-97,100-102. (Huang Z. Dunhuang Suzidian[M]. Shanghai: Shanghai Century Publishing Group,Shanghai Education Press,2005:96-97,100-102.)

[7] 图书质量管理规定[EB/OL]. [2021-11-15]. http://www.gov.cn/gongbao/content/2005/content_69258.htm. (Provisions on the quality control of books[EB/OL]. [2021-11-15]. http://www.gov.cn/gongbao/content/2005/content_69258.htm.)

[8] 图书编校质量差错认定细则(2019修订版)[EB/OL]. [2021-11-15]. https://www.sohu.com/a/287913602_757863. (Rules for identifying errors in editing and correcting quality of books (2019 revised edi-

tion)[EB/OL]. [2021-11-15]. https://www.sohu.com/a/287913602_757863.)

[9] 中国第二历史档案馆. 政府公报(1914年2月5日第628号)[M]. 上海:上海书店,1988:152. (The Second Historical Archives of China. Government bulletin (February 5,1914,No. 628)[M]. Shanghai: Shanghai Bookstore,1988:152.)

[10] 刘石,孙茂松. 关于建设"中国古典知识库"的思考[J]. 人民政协报,2020-08-24(8). (Liu S,Sun M S. Thoughts on the construction of "Chinese classical knowledge base"[J]. CPPCC Daily,2020-08-24(8).)

[11] 汪中. 述学·释三九(中)[M]. 北京:中国书店,1925. (Wang Z. Shuxue·Shisanjiu(Ⅱ)[M]. Beijing:Cathay Bookshop,1925.)

[12] 章学诚. 校雠通义内篇·校雠条理[M]. 刻本. 嘉业堂,1919. (Zhang X C. Jiaochou Tongyi Neipian·Jiaochou Tiaoli[M]. Block-printed edition. Jiayetang,1919.)

[13] 应劭. 风俗通义校注·序[M]. 王利器,校注. 北京:中华书局,1981:11. (Ying S. Fengsu Tongyi Jiaozhu·Xu[M]. Wang L Q,annotate. Beijing: Zhonghua Book Company,1981:11.)

[14] 扬雄. 輶轩使者绝代语释别国方言:卷一[M]. 刻本. 浔阳郡斋,1200(宋庆元六年). (Yang X. Youxuan Shizhe Juedaiyu Shi Bieguo Fangyan:volume 1[M]. Block-printed edition. Xunyangjunzhai,1200 (the 6th year of Qing Yuan in Song Dynasty).)

[15] 杨明照. 抱朴子外篇校笺(下)[M]. 北京:中华书局,1991:67. (Yang M Z. Baopuzi Waipian Jiaojian (Ⅱ)[M]. Beijing:Zhonghua Book Company,1991:67.)

[16] 班固. 汉书·艺文志[M]. 北京:中华书局,1964:1707. (Ban G. Hanshu·Yiwenzhi[M]. Beijing: Zhonghua Book Company,1964:1707.)

陈 力 四川大学历史文化学院教授,国家图书馆研究馆员。四川 成都 610064。

(收稿日期:2021-12-02)