人文大数据及其在数字人文领域中的应用

陈静
南京大学艺术学院，江苏 南京 210031

摘要
人文大数据是指基于数字化或者数字生成的，被认为是人文艺术范畴的大规模数据集。与科学、工程及社会科学数据相比，人文大数据是一种“深层数据”，其来源更加混杂、格式更加多样、维度更加多元、数据层次更加复杂、内涵更加丰富，因此在数据分析过程中存在较大困难。针对人文大数据的特点，结合数字人文研究应用中的关键问题，突出人文大数据作为一个集体概念的复杂情况及可能存在的误区，彰显人文大数据的价值。

关键词
人文大数据；数字人文；深度数据；智慧数据

中图分类号：J0-05 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022086

Humanities big data and its application in the field of digital humanities

CHEN Jing
School of Arts, Nanjing University, Nanjing 210031, China

Abstract
Humanities big data refers to large-scale data based on digitized or digitally generated data that is considered to be in the realm of humanities and arts. Compared with science, engineering and social science data, humanities data is a kind of "deep data" with more mixed sources, more diverse formats, more diverse dimensions, more complex data levels and richer connotations, so there are greater difficulties in the process of data analysis. Focused on humanities big data and its characteristics to identify the key issues in the application of humanities big data research, the complex situation of big data as a collective concept was highlighted, as well as the possible misunderstanding, while highlighting the value of humanities big data.

Key words
humanities big data, digital humanities, deep data, smart data

0 引言

长期以来，人文学者习惯称呼其研究对象为文本、图像，或是音乐、电影，而非数据。他们主要通过书籍、图书馆、档案馆、博物馆，甚至是手工记录和拍摄等途径获取材料，再通过经验性阅读、主观分析和语言解释的方式加以研究。尽管人文研究中也会涉及一定的信息采集和基于数据分析的定量研究，但人文学者习惯处理基于印刷（print-based）或者实物的材料，并将之视为唯一可信且权威的依据，再以经多年训练和研究获得的学识为基础，展开具有强烈经验色彩的个人研究。这种传统研究除了强调人文研究需要长时间知识生产的积累、承袭外，还高度依赖学者作为个体对材料的占有和处理能力，以及材料本身的原真性和有效性。甚至在一定意义上，材料的质量、真假以及丰富性对于一项研究具有决定性的意义。然而，从20世纪中叶以来，信息通信技术（information and communications technology，ICT）及相关基础设施已经深刻地改变了人文学者获取材料、分析内容、书写文本、组织学术交流的方式，重塑了当前的学术生态环境。数字技术及数字化使印刷物、手写书稿、非正式出版档案、绘画、照片、视频、声音文件、建筑、雕塑、壁画、纺织物、器物等多种材质、多种类型的人造物从物质实体变成了虚拟数字，与大量数字生成（digital-born）的内容一起成为人文学者的新研究对象。数字档案库、文本和图像数据库的出现使人文学者可以不用亲自到访千里之外的图书馆、博物馆、档案馆就可以获取所需要的信息；搜索和下载功能使研究者可以在成千上万的资料中快速地浏览和找到有效信息并“据为己有”；文本处理和管理软件改变了研究者组织材料、撰写文章的方式，使研究过程更多地成了“界面操作”；甚至研究者的思考方式也受网络化知识组织方式的影响而变得超文本化，使研究者更多地关注到不同议题和材料之间的关联性[1]。这样的新一轮知识生产方式的变革最集中的体现就是“数字人文”（digital humanities）作为一种跨学科研究领域的出现。数字人文强调将数字科技与人文研究进行结合以推动人文研究转型，“其面对的是未来的知识体系及方法的构建，其回应的是大数据时代基于学者导向（research oriented）的研究需求与基于资源共享的网络基础设施建设（cyberinfrastructure），其建设的是面向数字出生（born-digital）新生代人类的认知方式系统与路径”[2]。尽管作为一个新兴的研究领域，“数字人文”诸多议题尚在讨论之中，但从其发展历程来看，数据的获取和数据本身都对相关研究的开展及研究方法的提出起到了至关重要的作用。特别是大数据和人文数据的关系，以及大数据研究方法在数字人文研究领域中的应用，也是近年来数字人文研究中的焦点问题。本文将聚焦“人文大数据”这一具体对象，将之放置于“数字人文”的研究框架与范围内，对其来源及产生方式进行描述，并通过与自然科学和社会科学数据进行比较，对其特点进行说明，进而对数字人文因人文大数据及其方法的特殊性而呈现出的多样性问题进行探讨。

1 人文大数据

大数据指的是超出了常用软件工具在可容忍的时间内捕获、管理和处理数据能力的数据集。自21世纪中叶以来，数据的收集和处理已经成为计算机、生物医学、信息科学、经济金融等学科的基本

研究手段。甚至有学者指出，大数据带来的是一次新的认识论和范式转型，从知识驱动（knowledge-driven）转向数据驱动（data-driven）。而数据驱动的主要特征就是数据密集（data-intensive）、统计探索（statistical exploration）和数据挖掘（data mining）[3]。套用“大数据”的通用定义，即“超出了常用软件工具在可容忍的时间内捕获、管理和处理数据能力的数据集”，人文大数据可以被定义为“基于数字化或者数字生成的，被认为是人文艺术范畴的大规模数据集”。人文领域中的大数据可以分为两类：一类是通过对人文对象数字化（数据采集）的方式获取的各类数据，这类数据以美术馆、图书馆、档案馆和博物馆等文化机构的文化遗产数据为代表，规模庞大且类型多样，在被数字化之前就已经具备了体量大、数据类型多样且价值高等特点，这些数据主要来自手抄或印刷文献、器物、建筑、绘画、模拟方式记录的声音、视频等人造物，代表人类物质与精神文明的历史成就；另一类则是数字技术出现以后不断生成的数字文本、图像、视频、音频以及3D模型等基于各类数字软件的多媒体数据，这类数据以博客、Facebook、Instagram这类网络社交媒体的文本和图像为代表，体现了更宽泛意义上的“数字文化”（digital culture），是数字化时代对人类文化艺术活动的记录。此外，还有一些数据，在传统意义上被认为是非人文社会科学领域的数据，但其被应用到了人文研究之中，因此也开始被研究者认为是人文大数据，如地理及空间信息数据。历史地理信息系统早在20世纪90年代已经出现，其旨在运用地理信息系统（geographic information system，GIS）来研究历史问题。近年来历史地理信息系统得到了进一步发展，从社会史向其他人文科学领域拓展，形成“人文GIS”，与“空间人文”形成了共谋。后者主要的特征之一就是向人文内容进行渗透，更深一步地对人文材料内部进行挖掘（如对文学作品中蕴含的地理及空间特征进行的研究）。而在人文研究领域，也有学者开始将地理系统或者空间作为研究方法，开展“文学地图”或“在地研究”。这些都是人文研究在数字技术时代，尤其是大数据时代出现的新现象。

2 人文大数据的“大”与“小”

人文数据可以很大。若将人文艺术领域跨越千年的各种类型的材料都加以数字化，那所形成的数据集将相当可观。以世界上最大的图书馆——美国国会图书馆为例，截至2021年，该馆馆藏超过17.3亿件，其中2 200万件藏品在“美国记忆”（American memory）项目的资助下被数字化，按照估算大概是9 PB，包括从公元10世纪至今的、来自66个国家的印刷书籍、期刊、照片、录音、报纸、地图、电影、手稿、法律文书、个人叙述、软件、网页、网络档案库和3D对象等多种格式的文件。尽管这个数字化数量已经相当惊人，但尚不能代表人类文明的总量。类似“美国记忆”的数字化项目在过去几十年间一直在进行中，积累了大量的人文大数据，也为相关研究者开展进一步的数据分析提供了基础条件。另一个人文大数据的例子是谷歌的N-gram项目，以让·巴蒂斯特·米歇尔为首的研究团队与谷歌图书合作开展的“基于百万数字图书的文化量化分析”基于谷歌大规模数字化书籍的语料库开展计算分析，并以可视化方式呈现人类文化的发展趋势。研究团队使用自然语言处理中较常用的N-gram模型，以单个词或多个词为单位，对来自全世界的大学图书馆的1 500万本数字化图书中

的，从1800年到2000年的500万本，共计7种语言500亿字的文本进行了统计分析，对英语词汇量变化、英语语法的变迁、集体记忆与健忘、大众声望、审查检测等文化议题进行解读。由于该项目是基于200年间的词频波动进行观察的，因此得出的一些结果是非常具有启发性的。例如英语书籍中最常使用的词汇实际上比权威字典的要多，而且常用词中大约63%的英文词汇在齐夫定律（Zipf's law）的测量下是低频使用词，更有52%的词汇是没有被收录到词典中的。这种通过对大数据集进行定量分析，从而学习人类文化的方式被命名为“文化测量”（cultural analytics）模式，相关成果于2011年在*Science*上发表[4]。此后，不少学者也将此模式用于不同的文化数据集[5-6]。例如卡莱弗·李塔鲁（Kalev Leetaru）对30年间全球的本地新闻进行了调性和地理分析，并成功预测了2011年在阿拉伯半岛发生的重大政治事件及该事件发生的地点[7]。这种规模的数据集使从大规模尺度上对文化事件、趋势、变化进行计算测量成为可能，实现了传统人文学科无法企及的效果。

但对于人文数据而言，大数据的5个V（volume、variety、veracity、value、velocity）中的“volume”（体量大）是一个相对的概念。对于很多人文研究来说，数据集不会很大（如文本数据），几十万字甚至上百万字的文本也不过以KB为单位，相比生物数据之类的大数据而言，算得上小。但是，这些文本包含的内容及其可供研究的问题，并不能用体量来衡量。回顾数字人文的发展历史，很多“小”文本语料扮演了非常重要的角色。

“数字人文”在西方一般被认为有两个源头：人文计算（humanities computing）与文本批评（textual critics）。而文本批评以电子编辑（electronic editing）为代表[8]。人文计算的开创往往会追溯到意大利神父罗伯托·布萨（Roberto Busa）在1949年开启的、与国际商业机器公司（International Business Machines Corporation，IBM）合作的The Index Thomisticus项目。这个项目主要是利用IBM当时基于穿孔卡和磁带存储的计算机对中世纪神学家托马斯·阿奎纳（Thomas Aquinas）写作的及与其相关的179部、1 000多万字古典文本进行处理，半自动地生成中世纪拉丁文字词的索引[9]。该项目在20世纪70年代出版了56卷7万多页的印刷物，其中包括10卷索引（index）、31卷托马斯·阿奎纳作品索引大全（concordances）、8卷相关作者的索引大全以及7卷原初文本的重印本。该项目在1989年以CD-ROM形式出版后，在2005年发布了在线版本，在2006年启动了对全部语料库的语义分析。整个项目持续多年，耗费巨大，除了成吨的卡片以外，还有长度达到1 500 km的磁带、1万小时的计算机工作时长和100万小时的人工工作时长[10]。无论是从文本还是从技术上而言，这个项目都是具有开创性意义的， 其塑造了一种新型的人文学者与科学家（工程师）合作模式的典范，也奠定了计算机处理人文文本的一些共性，如文本分析以语料分析为基础、半自动化或者自动化程序处理、索引作为语料的基础数据、多学科的跨学科性等。但倘若纯粹地从数据量上来看，这个“不仅是第一个，也是有史以来最大的数字人文项目之一，尽管按照今天的标准，其结果可能被认为是‘小’”[9]——其光盘内的数据不过1.4 GB。但可以确定的是，由此开启的是人文研究，乃至知识生产历史中的一个新时代。托马斯·阿奎纳项目的开启和实施，不仅标志着人文计算作为一个新兴领域的出现，更标志着人文研究中使用计算机运算的技术已经形成一套理论化的思考，

也开启了一系列基于文本索引的语料库和程序的计算语言学项目，其中包括伦敦大学学院（University College London）和擎天计算实验室（The Atlas Computer Laboratory）开发的COCOA二代、牛津语汇索引程序OCP和希腊语库TLG等。这些文本处理程序主要致力于语料库的建设与对文本创建、维护和存储方面的程序进行联合开发与推广。这种取向在1950—1960年影响了不少文学研究者利用计算机处理机器可读文本的内容，对大体量的作品做出分析，如关于联邦党人信件的作者研究堪称经典。

由另一个源头即文本批评所延伸出来的数字人文脉络则更关注从文献学的角度利用信息技术对文本进行深度编辑与标注。最重要的成果是文本编码倡议（Text Encoding Initiative，TEI）的《电子文本编码和交换指南》（guidelines for electronic text encoding and interchange）。TEI是一个集体开发和维护数字形式的文本表示标准的联盟，其主要成果是一套规定了机器可读文本的编码方法的准则。该准则主要被应用于人文学科、社会科学和语言学领域。对于数字人文领域而言，TEI提供了一种机器读取人文文本的规范标准，因其灵活性、综合性和可扩展性等特点，在很多图博档机构中得到了应用。此外，文本批评非常重视对文本的深度挖掘，因此尤其强调通过标注的方式对非结构化数据进行结构化，或生成元数据，在元数据的基础上进行数字存档和知识再生产。例如罗塞蒂档案（the Rossetti Archire）或威廉姆·布莱克（William Blake）档案这样的项目就很好地践行了这样的路径。特别是对于文本物质性的重视，使这些档案在数字化的过程中尽可能地考虑到了印刷文本的专有属性，并通过数字标注的方式加以呈现[11]。在此类项目中，对象本身的数量并不多（如威廉姆·布莱克档案中收录的作品数量不过100多幅），但每一幅的元数据不仅包括了作品信息数据，还包括对图像内容的标注和文本内容的转录。这种对小数据集展开的深度标引和研究，也形成了数字人文中的重要内容。特别是随着20世纪90年代中后期数字技术的更新迭代、数字化内容的不断增加，计算语言学逐渐从人文计算中独立出去，这种研究趋势得到了更广泛的应用，影响遍及各个人文学科，也显示着“数字人文”新阶段不再延续早期的发展路径。大约在2000年以后，“数字转向”（digital turn）时代到来，个人计算机变得十分普遍，成为大多学者可以方便使用的设备，如OMEKA、Voyant这样的专门面向人文学者的数据档案化、文本分析可视化的工具也被开发了出来。

从西方形成的人文计算到数字人文这个脉络来看，实际上我国在20世纪下半叶就开展了大量基于语料库的计算语言学研究，如从1979年到1983年，就有4个大型的现代汉语语料库项目在我国落地，即武汉大学的汉语现代文学作品语料库（1979年，527万字）、北京航空航天大学的现代汉语语料库（1983年，2 000万字）、北京师范大学的中学语文教材语料库（1983年，106.8万字）和北京语言学院（1996年更名为北京语言大学）的现代汉语词频统计语料库（1983年，182万字）[12]。这些数据库和之后的国家级语料库、大规模真实文本语料库等专业数据库主要针对语言学方面的研究。面向更多领域学者的中文学术数据库多为图博档甚至是商业公司开发的基于典籍的文本图像或者全文数据库，如由香港迪志文化出版有限公司推出的文渊阁四库全书的电子版、由北京大学等高校与北京爱如生数字化技术研究中心合作建立的“中国基本古籍库”[13]。与此同时，还有一些人文学者从研究需求出发开发的数字项目，如北京大

学中文系开发的全唐诗分析系统与全宋诗分析系统、先在香港中文大学后迁至台湾政治大学的“中国近现代思想史研究专业数据库(1830—1930)”等。在这些项目中,数据规模虽大,类型各有不同,但数据库限定性比较强,往往只能进行检索,无法下载或者进行更深入的研究。关于此类问题,在近年来关于文献数字化的相关讨论中已经非常多了。尤其是研究者们已经关注到了以往数字化工作中的一些问题,如传统的古籍数字化大多是对原始纸质文献的图片展示,仅可检索编目数据,对内容仅以浏览为主,缺少全文提供,用户也无法按照自身的研究需求对数据进行深度挖掘和再利用等[14]。相较而言,“中国历代人物传记数据库”(China biographical database,CBDB)和德龙(Donald Sturgeon)开发的“中国哲学书电子化计划”(Chinese text,Ctext)则兼顾了大数据与人文研究的属性。虽然CBDB的单机下载版总共不过几十MB(SQlite格式),但其中收录了超过52万位历史人物的传记资料,每个人物条目都包含了人名、时间、地址、职官、入仕途径、著作、社会区分、亲属关系、社会关系、财产、事件等数据,可供学者们开展统计分析、地理空间分析与社会网络分析等[15]。值得一提的是,CBDB不仅涉及了中文文献的数字化、数据化(datafication)、数据清洗、数据分析、数据库搭建、软件开发以及数据可视化等一系列的数据全流程工作,而且非常仔细、详尽地记录和说明了整个数据库的发展历史、技术开发和数据处理过程,对其他人文大数据项目的建设极具参考价值。

3 人文大数据的多样性与语境化

葛剑雄教授曾经在讲座中提到,“运用现代科学技术,我有两个衡量标准,那就是,首先它最后的精确度有没有其他方法加以验证,其次它的结果有没有意义,能不能改变一个重大的学术论断。我发现大数据在历史研究中还是没有太大必要,因为我们掌握的数据不够,而且很多是二手甚至三手数据,盲目运用的结果就是可信度越来越低,误差也会越来越大,到最后还是需要人来做出判断和取舍,这是没有必要的”[16]。这里他谈到关于大数据应用于历史研究的必要性,首先谈到的是数据的量不足,其次是数据的可信度低。关于数据的量,这点前文已经讨论过,对于人文数据而言,量并不是最重要的,过度强调大,其实是对大数据的一种化约式(reductive)的误读。实际上,大数据的多样性(variety)和真实性(veracity)往往发挥着更加重要的作用。

首先,人文大数据的来源决定了这些数据从一开始就会是多种多样的。例如美国国会图书馆在线上发布时,不仅考虑到原真性,发布了文件数字化后的图像文件,还考虑到了人文研究者的分析需要,提供了数字文件的元数据,以及包括了XML格式的标记数据和TXT格式的全文数据,这体现了人文数据的多样性和特殊性。异质的数据往往同时被应用于同一个人文研究项目中,而学者就是要利用这些异质数据集之间的联系和重叠进行各种推断。对于人文大数据而言,多样性还意味着这些数据集结构的多样性。很多时候,这些不同的数据集无法被整合成一个统一的数据集,然后用一种方法来分析。甚至,同一种算法针对不同的数据集也可能需要训练不同的模型。但人文数据的异质性是人文大数据最明显的优势,也是人文学科数据最大的挑战。有了这些来源不同、格式不同的数据,研究者才能更加灵活地组合,以便从

中获取最大的研究效果。这也是人文大数据与社会科学大数据、科学大数据的区别之一。

其次，大数据的真实性和准确性需要一定的人工干预。虽然更大量、多样的数据才可以弥补以往小样本、抽样数据的片面与偏差，但正如葛剑雄教授所言，对数据的盲目应用往往是导致数据误差的重要原因。布萨神父在论及他为何在阿奎那项目开始后试图引入计算机时回忆说，“我相信计算机的速度和准确度将对这项研究中涉及的数据汇编工作有很大帮助”，但他也关注到了任何关于语言学数据的解释都是归纳式的，更多的是基于已有的经验证据及支持可靠结论的文献的完整度，因此布萨非常关注源数据的质量[9]。中国学者在处理大规模真实文本语料时也发现了类似的问题，如宋柔在统计语料库中的词语接续对时发现，随着语料库规模的增大，新增加的接续对中的垃圾逐渐会占大部分甚至绝大部分。垃圾主要分布在统计到的低频度接续对中，主要来源是分词中专名识别错误[12]。实际上数据一旦达到一定的规模，其中难免存在错误、冗余数据，对于传统的统计学或者数据科学来说，合理范围内的偏差是可以接受的，但对于人文研究而言，会因为文本在光学字符识别（optical character recognition，OCR）过程中出现的乱码而被批评。在这个问题上，如何在尽可能扩大数据规模的同时，兼顾数据的多样性，并确保其真实性，就成为人文大数据处理中的关键。大部分的数字人文项目会特别关注数据准确性的问题。

再次，人文数据需要语境。这种语境一方面体现在人文数据不仅仅是被提取和计算的对象，也要被放回原初语境，如放回文本的上下文中进行观察和解读；另一方面则是因为人文研究谈及的社会或历史“语境”是非常大的范围。在概念史研究学界曾经有过一场争论。金观涛、刘青峰两位老师在1997年启动了一个名为“特定现代中文政治概念形式的量化研究”的项目，意图对新文化运动期间最具代表性的12个中文期刊中的文章进行量化统计和分析。随后，两位老师意识到现代重要政治观念的研究开展是可以通过对更大范围内的文本进行检索和分析进行的，由此建立了“中国近现代思想史研究专业数据库（1830—1930）”，并将基于该数据库的相关研究以《观念史研究：中国现代重要政治术语的形成》为名出版，其中包括了对近代思想史中多个（组）现代重要观念进行的基于关键词的研究。此后有学者提出，基于数据库对历史进行研究受到数据库收录资料的限制，其中很多资料没有被收录，会影响到研究的真实性。很多语境化的信息，如信息及观念的传播方式、物质构成、商业运作、读者获得途径、读者的阅读接受情况等，无法用精准的时间或数字来表现，企图用数据多少或出现频率来揭示，不但存在极大的难度，更存在致命的缺陷。两位老师随后在回应中明确回复，其所做的研究也都是在数据库所收录的文献范围内开展的，因此如若认为更大规模资料的收录会影响目前的研究结果，则需要进行实际的研究加以验证。而且，以关键词为中心的观念史研究是典型的人文学科，只不过引进了数据库方法：“数据库在人文研究中只有辅助作用，它为研究者提供了极大的便利，也提出了更高的要求。它只是在对关键词的使用情况和类型分析这一素材收集和整理环节上提供了工具，而研究者在此基础上，要以人文学科的基本范式和自己的研究素养来分析这些资料”[17-18]。这场论辩中批评者的主要怀疑点在于一定数量的数据（哪怕是一亿两千字的数据量）及基于该数据

集的一种统计分析能否体现历史的真实？其实回到大数据本身，或许就能有更好的理解。不存在任何数据集是“全数据”，事实上，可能永远都没有办法做到全数据。那么基于大数据的研究与所有以往的研究一样，都是在一定的范围内基于一定的对象进行的研究，因此局限性是不可避免的。那么这里实际上要回答的是，基于部分数据，而且是相当大的数据集的研究是否有效？这个答案也是毋庸置疑的，实际上，哪怕是基于某一种单一来源的数据集，当体量大到一定程度时，从数据的角度而言，其与基于多个数据来源的小数据集的研究都一样具有意义。衡量的标准不在于数据本身，而在于研究的结论本身。而验证结论的方法是定量还是定性也是没有唯一性的。但提出批评是需要一定的条件的，尤其是对定量分析的批评，最好是要建立在对同样数据集的验证实验的基础上，而这一点往往更多地体现在自然科学研究中，而非人文研究。同时，数据、文本的语境与历史、社会的语境并非同一层面。正如批评者所言，并非所有的历史、社会语境都可以文本化、数据化，因此，也并非所有的人文研究都需要依赖数据分析。在这个意义上，有学者在讨论“什么不是数字人文”“什么是数字人文”以及“什么是好的数字人文”中都提到了，数字人文或者说基于人文大数据的人文研究，重要的并不是工具或者方法论本身，而是究竟用这样的数据和工具解决什么样的人文问题。人文性在数字人文研究中是第一位的。可以说，这样的讨论体现了人文学界对于大数据及大数据研究方法的一种内省和警觉。正如葛剑雄教授提出的，要考量“它的结果有没有意义，能不能改变一个重大的学术论断”，人文研究的问题还是要回到人文的领域里进行检验。

4 深层数据与智慧数据

那么，理想的人文大数据是什么样的？不妨从与社会科学的比较开始分析。通常社会学、经济学、政治学、传播研究和营销研究被认为更适合使用定量方法（即用于分析数据的统计、数学或计算技术），而人文学科，如文学研究、艺术史、电影研究和历史，则倾向于使用诠释学、参与观察、厚重描述、符号学和细读等方法。对于社会科学和人文研究而言，数字技术与大数据所带来的学科影响则以计算社会学（computational social science）和数字人文为代表。尽管两者在研究对象和研究方法上有相同与交叉，如皆以数字技术及数字文化为对象、都会涉及数据处理方法的应用，但两者也存在区别，如数据获取和处理的方式、研究问题的提出等。而从数据的层面来说，列夫·马诺维奇（Lev Manovich）将前一类可以适用于定量分析的，与大群人或团体有关的数据称为“表层数据”，将后一类与更为小众的群体有关的数据称为“深层数据”[19]。他指出尽管基于大规模数据的社会计算（social computing）研究往往能提供关于人类在数字文化时代的行为和表现得更广泛的数字图景（digital landscape），但计算机在理解文本、图像、视频和其他媒介意义与语境方面具有的局限性，使这些研究都只能是基于简化维度的分析，甚至会受到错误数据的影响。而他所设想的理想状态则是将人所具有而计算机所不具有的理解和解释能力与计算机运用算法处理大规模数据的能力结合起来。这一点其实在有关“智慧数据”的讨论中也有所体现。

曾蕾、王晓光、范炜与克里斯托弗·绍什（Christof Schöch）分别曾撰文讨论

过智慧数据。曾蕾等指出智慧数据是“实现大数据特征中最后一个‘V’——价值（value）的方法，即通过对任何规模的可信的、情境化的、相关切题的、可认知的、可预测的和可消费的数据的使用来获得重大的见解和洞察力，揭示规律，给出结论和对策”。借此他们提出，“智慧数据通常带有自描述机制，背后有领域本体作支撑，使这些数据符合特定的逻辑结构和形式规范，而且可以支持推理，由此形成智慧的基础，产生可预测和可消费的数据”。同时，还因为“智慧数据较强的可解释性，支持逻辑推理从而使之可以用于多种用途和支持多种互操作，并且具有很强的可追溯能力，能够满足人文研究范式的需要。”他们通过图博档中关于关联数据、图像深度标引和非物质文化遗产数据的元数据等议题来说明智慧数据具有的特性。绍什关于智慧数据的定义则更加简洁，即“我建议首先将大数据看作相对非结构化的、混乱的和隐含的、体积相对较大的、形式多样的。相反地，我建议将智能数据看作半结构化或结构化的、干净的和明确的，以及体积相对较小、异质性有限的。”两种定义从不同方面指向了智慧数据的价值和属性，可以帮助人们理解为什么在人文研究中学者会强调智慧数据。这恰恰是因为人文研究对数据的要求更高、更加苛刻，而人文数据，尤其是第一类通过数字化生成的人文数据，其数据的结构化程度、清洁度和可量化效果都是由数据生成过程，甚至是投入人力的多少来决定的。

5 计算很重要，但不是全部

随着大规模数据集的出现和数据分析方法的更新，计算的问题也越来越多地受到了学者的关注。在文学界，以佛朗哥·莫雷蒂（Franco Moretti）为代表的学者，包括马修·乔克斯（Matthew Jockers）、马修·威尔肯斯（Matthew Wilkens）和安德鲁·派珀（Andrew Piper）等在内，支持运用主题建模、网络分析等从海量数字化文学资料库中挑选出的语言与形式的宏观模式。尤其是莫雷蒂基于对大量小说文本信息（如标题）的统计分析形成的“远读”（distant reading）理论及研究方法对数字人文乃至整个人文学界影响深远。但从实际效果而言，莫雷蒂的“远读”方法也并没有真正从根本上解决布萨1949年提出的问题：如何用计算机使学者们快速而准确地深入研究诸如真实性、文本批评、风格、年代和翻译等一系列问题。在美国现代文学协会出版物（Publication of the Modern Language Association，PMLA）2017年组织的一次关于“远读”的讨论中，莫雷蒂对此作出了回应。他部分地赞同了苏真（Richard Jean So）教授对其的批评——“（莫雷蒂）所做的不过是对其语料的一个统计描述”，同时还指出安德鲁·派珀所提出的实现一种“模型的模型”（model of a model）是未来必然的发展路径。他指出，苏真等人及芝加哥大学文学实验室正在进行的“模式”的研究将完全改变理论所具有的可能性，将会改变历史与文学研究的关系，尤其是改变文学研究的时间性框架，历史将成为文学研究的前提[20]。而“模型的模型”或者说“模式”正是计算文学努力通过量化计算实现的方法论尝试。赵薇指出，从莫雷蒂的概念模型到后来的文学实验室的计算批评，“实证研究”与文学阐释、文化批评被有机地融合在一起。量化文学研究的本质是根据研究的需要，选取合适的测量尺度和有效的测量手段，只有这样才能真正发现问题[21]。

然而，并非所有的学者都能接受对人文数据进行量化分析。一篇于2017年10月

15日发表在美国《高等教育纪事报》网站上名为《数字人文搞砸了》(The digital-humanities bust)的文章引发了广泛争论[22]。作者提摩太·布伦南是明尼苏达大学双城分校的文化研究、比较文学及英语系教授。在布伦南教授看来,英国剑桥分析公司Ada算法事件体现的是对“数据”和“算法”的盲目乐观主义在现实社会中的受挫。布伦南指出,算法不仅是一系列失败事件背后的推手,也是隐藏在数字人文研究及其20年蓬勃发展的逻辑,数字人文也在这种“非常公开和尴尬”的结果中面临危机与反思。他在历数了这些年来数字人文学者得到的诸多好处(如美国国家人文基金、梅隆基金会提供的大量资金资助,一流期刊文章的背书以及得到晋升终身教职岗位等)之后,提出质疑:数字人文到底有什么成就?布伦南教授认为,数字人文研究对算法的依赖使数字人文学者在面对文本时只看到了通过算法所呈现出的文本的特点(如词频),却无法触及文本中有价值的内容;也同时因为对算法的依赖,数字人文学者无法摆脱计算的局限性,而以此局限性为探寻研究问题的限定。尤其针对书籍内容的量化分析、文学批评中的“远读”策略和“文学模式识别”等,布伦南认为数字人文学者只是看到了表层的数字和数据,但却不能像使用大脑那样使用计算机进行深入的思考:“由于其自身机制,数字‘阅读’从根本上将大脑自然产生的智慧灵感,建立价值形式的建立,以及本能冲动都彻底排除在外。”论其原因,一是因为将“更多信息混淆为更多知识”,数字人文学者无法在其所施用的方法之外进行反思,认识到该方法在认识论上的意义和方法论上的价值;二是“对科学的迷恋,新自由主义的撤资”,占有少量资源或者长期处于学界边缘的年轻学者通过新科技在已经划定格局的学术场域内争取更多的文化资本,获取地位提升。因此,“与其说数字人文是一场革命,不如说数字人文是为了反对主流形式,从而强行将人文从其存在原因中剥离出来的那个楔子”。

文中提到的关于数字人文中的某些局限性也确实是数字人文学界普遍存在的问题,如部分研究还停留在词频的程度上,而且有些数据本身也是经过预先加工的,因而有“作弊”嫌疑,同时很多数据处理的过程也是在人工监督下完成的,因此结果也不那么令人惊喜等。但布伦南一文中的问题也是非常明显的,如“数字人文”在文中被简化为了关于数字的“量化”,而抹杀掉了数字人文中学科、研究问题和领域的多样性;再如苏真和霍伊特·朗(Hoyt Long)关于日本俳句的“文学模式识别”(literary pattern recognition)研究并非只是在检验一个已知结果的正确性,而是通过一种新的计算方式挑战及改变以往对于俳句的认知及研究思考。对于这种误读或者攻击,包括被批评对象特德·安德伍德(Ted Underwood)和霍伊特·朗在内的3位学者在2017年11月1日的《“数字”与“人文”不对立》(“digital” is not the opposite of “humanities”)[23]中做出了回应:首先,量化研究在经济、社会学乃至人文研究中应用已久,数字人文因此“获罪”实在是作者有意为之;其次,仅就量化或者说数字而言,数字人文中所说的“数字”也比作者所说的简单计算词频要广泛得多,例如之前提到的“文学模式识别”,“就已经被用来探讨虚构的本质、文类的周期,以及塑造角色的性别假设等”。这些问题是文学史的核心问题,并且因数字人文得以从一个新的尺度进行讨论。最后作者还指出,数字人文不仅仅意指新的研究手段,也影响到博物馆、新闻、图书馆等机构面向公众传播的新形式。类似的讨论还出现在了历史研究、艺术史研究等领域。

以大数据和计算的方式进行人文研究受到了普遍的争议。但正如埃里克·威斯科特（Eric Weiskott）在对此的回应中提到的，数字技术正在重新创造历史，这个过程和16世纪印刷技术在欧洲出现时发生的情况类似，也同样引起了质疑。而作为一种不可逆转的过程的结果，数字技术改变的不仅仅是知识传递，更是一种新的知识形式的体制建构，并非仅仅是认识论的改变[24]。确实如此，对于数字人文而言，计算并非仅有的手段，但人文大数据却是已经存在且必须要面对的现象。如何更好地利用数字技术与方法对人文大数据开展多角度的研究是比争论是否可以使用数字技术或方法更为实际和迫切的问题。

6 结束语

以上关于人文大数据的讨论，多将人文大数据看作为达到某种研究目的所使用的材料，但事实上大数据本身及大数据分析过程中产生的一系列伦理问题，如ImgaeNet这样的大规模图像数据集中具有的性别、种族偏见问题以及这些问题引发的相关算法缺陷问题、数据收集及清理背后的数据劳动问题等，引发了人文学者的普遍关注。人文大数据带来的问题不仅仅是研究范式的转变，其更成为研究问题本身。但很遗憾的是，目前从事数据科学的研究者们却较少与人文学者就人文大数据及大数据在人文研究中的价值展开直接而深入的讨论，期待此次专题能开启如此契机。

参考文献:

[1] 苏芃. 古代经典的“超文本”阅读法[N]. 光明日报, 2022-10-03.
SU P. “Hypertext” reading method of ancient classical[N]. Guang Ming Daily, 2022-10-03.

[2] 徐力恒, 陈静. 我们为什么需要数字人文[N]. 社会科学报, 2017-08-24.
XU L H, CHEN J. Why do we need digital humanities[N]. Social Sciences Weekly, 2017-08-24.

[3] KITCHIN R. Big data, new epistemologies and paradigm shifts[J]. Big Data & Society, 2014, 1(1): 1-12.

[4] MICHEL J B, SHEN Y K, AIDEN A P, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011, 331(6014): 176-182.

[5] PETERSEN A M, TENENBAUM J, HAVLIN S, et al. Statistical laws governing fluctuations in word use from word birth to word death: 10.2139/ssrn.1890569[P]. 2011-01-01.

[6] ROTH S. Fashionable functions: a Google N-gram view of trends in functional differentiation[J].International Journal of Technology and Human Interaction, 2014, 10(2): 34-58.

[7] LEETARUK K. Culturomics 2.0: forecasting large-scale human behavior using global news media tone in time and space[J]. First Monday, 2011, 16(9).

[8] 陈静. 数字人文知识生产转型过程中的困境与突围[J]. 文化研究, 2018(2): 171-185.
CHEN J. The crisis and solution of digital humanities in the transformation of knowledge production[J]. Cultural Studies, 2018(2): 171-185.

[9] ROCKWELL G, PASSAROTTI M. The index thomisticus as a big data project[J]. Umanistica Digitale, 2019.

[10] TASMAN P. Executive, 74[N]. The New York Times, 1988-03-07.

[11] MCGANN J. Radiant textuality: literature after the world wide web[M]. New York: Palgrave Macmillan, 2001.

[12] 冯志伟. 中国语料库研究的历史与现状[J]. 中国语言与计算机学报, 2002, 12(0): 43-62.
FENG Z W. The history and present situation of Chinese corpus research[J]. Journal of Chinese Language and

Computing, 2002, 12(0): 43–62.
[13] 陈静. 当下中国“数字人文”研究状况及意义[J]. 山东社会科学, 2018(7): 59–63.
CHEN J. The research status and significance of “digital humanities” in contemporary China[J]. Shandong Social Sciences, 2018(7): 59–63.
[14] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报, 2016, 42(2): 66–80.
OUYANG J. Visual analysis and exploration of ancient texts for digital humanities research[J]. Journal of Library Science in China, 2016, 42(2): 66–80.
[15] 徐力恒. 中国历史人物大数据[J]. 中国计算机学会通讯, 2018, 14(4):19–24.
XU L H. Big data on Chinese historical figures[J]. Communications of the CCF, 2018, 14(4): 19–24.
[16] 葛剑雄. 大数据在历史研究中没有太大必要[Z]. 2019.
GE J X. Big data is not much needed in historical research[Z]. 2019.
[17] 金观涛, 刘青峰. 简答张仲民先生对拙作的评论[N]. 东方早报·上海书评, 2010–05–30.
JIN G T, LIU Q F. Briefly answer Mr. Zhang Zhongmin’s comments on my book[N]. Oriental Morning Post·Shanghai Book Review, 2010–05–30.
[18] 李里峰. 概念史研究在中国: 回顾与展望[J]. 福建论坛(人文社会科学版), 2012(5): 92–100.
LI L F. The study of conceptual history in China: retrospect and prospect[J]. Fujian Tribune, 2012(5): 92–100.
[19] MANOVICH L. Trending: the promises and the challenges of big social data[M]. [S.l.:s.n.], 2012.
[20] MORETTI F. Franco moretti: a response[J]. PMLA/Publications of the Modern Language Association of America, 2017, 132(3): 686–689.
[21] 赵薇. 从概念模型到计算批评: Franco Moretti之后的世界文学研究[J]. 西南民族大学学报(人文社科版), 2020, 41(8): 181–189.
ZHAO W. From conceptual models to computational criticism: studies in world literature after Franco Moretti[J]. Journal of Southwest Minzu University (Humanities and Social Science), 2020, 41(8): 181–189.
[22] BRENNAN T. The digital-humanities bust: after a decade of investment and hype, what has the field accomplished? [Z]. 2017.
[23] BOND SARAH E, LONG H, UNDERWOOD T. “Digital” is not the opposite of “humanities” –the chronicle of higher education[Z]. 2017.
[24] WEISKOTT E. There is no such thing as “the digital humanitie”[Z]. 2017.

作者简介

陈静(1981–), 女, 博士, 南京大学艺术学院副教授, 主要研究方向为数字人文、数字艺术与数字遗产。

收稿日期: 2022–10–16
通信作者: 陈静, cjchen@nju.edu.cn
基金项目: 国家社会科学基金资助项目(No.21BA026)
Foundation Item: The National Social Science Foundation of China (No.21BA026)