数字人文的研究范式与平台建设

Research Paradigm and Platform Construction of Digital Humanities

刘圣婴[1] 王丽华[2] 刘炜[3] 刘倩倩[3]
LIU Shengying WANG Lihua LIU Wei LIU Qianqian

（1.华东师范大学图书馆，上海，200062； 2.上海大学文化遗产与信息管理学院，上海，200444； 3.上海图书馆，上海，200031）

摘要：[目的/意义] 数字人文的兴起正在带来人文研究的范式变革，数字人文平台作为向各学科人文学者提供研究素材的基础设施，同时也是数字化研究方法的承载者，平台建设的推动能够丰富数字人文的方法论体系，促成一种新的数字人文研究范式的确立。**[研究设计/方法]** 考察了数字人文的特点，提出了数字人文研究的范式框架；结合目前国内数字人文年会上展示的项目与论文成果，讨论了中文数字人文资源平台的建设问题，着重研究了功能需求和发展趋势；最后，以上海图书馆正在开发中的“历史人文大数据平台”为案例，阐述了这些思考成果的具体应用。**[结论/发现]** 具体将数字人文平台分为文献层、数据层、接口层、工具层和展现层等层次结构，使其各司其职且相互依存。针对中文数字人文的方法学特点，归纳了平台的不同类型，并对如何具备系统先进性、资源完整性、功能完备性、工具丰富性与用户友好性提出了设计原则。**[创新/价值]** 提出了由技术、过程和行为构成的数字人文研究范式，在数字人文平台中将这三个方面与人文资源相结合，成为数字人文研究基础设施的基本组成；对于数字人文平台建设则提出，除了需要关注技术架构之外，还需要将以领域知识为特征的内容架构单独提取进行设计和实现，并探讨了以语义技术（知识图谱）进行实现的基本做法。

关键词： 数字人文；研究范式；平台建设；方法论共同体；数字学术；基础设施建设

中图分类号： G251

引用本文： 刘圣婴，王丽华，刘炜，等 . 数字人文的研究范式与平台建设 [J]. 图书情报知识，2022，39（1）：6-29.（Liu Shengying, Wang Lihua, Liu Wei, et al. Research Paradigm and Platform Construction of Digital Humanities [J]. Documentation, Information & Knowledge, 2022,39（1）：6-29.）

Abstract: [Purpose/Significance] Digital humanities platform, as an component of infrastructure which provides study materials for humanities scholars in various disciplines, is the carrier of digital research methods. It can contribute to the enrichment of digital humanities methodological system and the establishment of a new research paradigm of digital humanities. **[Design/Methodology]** This paper investigated the characteristics of digital humanities, and proposed a paradigmatic framework for humanities research.With the consideration of the projects and papers presented in domestic digital humanities annual meetings, the construction of the Chinese digital humanities resource platforms was discussed, which emphasized the functional requirements and development trends. Finally, by taking the Digital Humanities Platform of Shanghai Library as an example, applications of the thoughts and ideas were illustrated. **[Findings/Conclusion]** The digital humanities platform is divided into literature layer, data layer, interface layer, tool layer and presentation layer. According to the methodological characteristics of Chinese digital humanities, this paper summarizes different types of platforms. It also puts forward some design principles on how digital humanities platforms can have advanced systems, complete resources, comprehensive functions, abundant tools and user-friendliness. **[Originality/Value]** This paper proposes a research paradigm of digital humanities consisting of technology, process and behavior. In the digital humanities platforms, the combination of the three aspects with humanities resources are the basic components of the digital humanities research infrastructure. In addition to the focus on the technical architecture, it is also necessary toextract the content architecture featured by domain knowledge specially for design and implementation. Besides, the best practices of implementation with semantic technology(knowledge graph) have been discussed.

Keywords: Digital humanities; Research paradigm; Platform construction; Methodological commons; Digital academy ; Infrastructure construction

1 引言

人文学科是所有科学之肇始，是人文精神之依托，被称为知识分子的必备和基础素养。无论是古希腊的七艺（文法、修辞、逻辑、算数、几何、天文、音乐），还是春秋的六艺（诗、书、礼、乐、易、春秋），其所创立的知识教育体系在今天多归属于人文学科范畴，致力于培养区别于万物的所谓“人性”。而当今社会建立起与工

[基金项目] 本文系国家社科基金重大项目 “文化遗产智慧数据资源建设与服务研究”（21&ZD334）的研究成果之一。（This is an outcome of the major project “Resources Development and Service of Cultural Heritage Smart Data”(21&ZD334) supported by National Social Science Foundation of China.）

[通讯作者] 王丽华（ORCID:0000-0002-2399-3848），博士，副教授，研究方向：数字人文、公共图书馆，Email:Wanglh@shu.edu.cn。（Correspondence should be addressed to WANG Lihua, Email:Wanglh@shu.edu.cn, ORCID:0000-0002-2399-3848）

[作者简介] 刘圣婴（ORCID:0000-0002-8707-8569），硕士，助理馆员，研究方向：数字人文，Email:syliu@library.ecnu.edu.cn；刘炜（ORCID:0000-0003-2663-7539），博士，研究馆员，研究方向：智慧图书馆、数字人文，Email:wliu@libnet.sh.cn；刘倩倩（ORCID:0000-0002-8111-5154），硕士，馆员，研究方向：数字人文、数据处理，Email:qqliu@libnet.sh.cn。

业文明相匹配的极其复杂又高深的现代教育，看似造就了大量知识丰富的“专家”，但却带来了知识分子整体上的消失，不仅缺乏对人的价值以及人类未来命运的思考者，连培养基本的责任与担当都成了奢望。在这个机器智能和生命编辑的时代，人文主义遭遇越来越严重的危机，我们比任何时候都更加需要和呼唤世界意义的守护者[1]。

在这样的背景下，数字人文诞生了。

作为信息技术在人文领域的应用，数字人文目前仍处于非常早期的发展阶段。虽然其历史可以追溯到计算机刚开始用来做文字处理的上世纪中叶，迄今已有七十余年，但“数字人文”一词是2004年随着*A Companion to Digital Humanities*一书的出版才得以定名的，当前还不具有公认的定义，甚至连边界在哪里也众说纷纭、莫衷一是。即便如此，鉴于数字化社会的到来已势不可挡，印刷品不再是知识生产与传播的主要媒介。在这个背景下，图灵奖获得者Tony Hey等敏锐地提出“科学研究的第四范式”概念[2]，指出当所有的研究素材和方法都数字化之后，“数据驱动型研究”就水到渠成，人文科学也概莫能外，数字人文必然是人文研究的未来。

数字人文是各门具体人文科学采用数字方法的汇聚和总结，是一种“方法论共同体”（Methodological Commons）。目前这个共同体已开始具备库恩所说的共同的“学科范式”特征，随着专业教育和学科体系的建立，数字人文逐渐从各种方法、技术的大杂烩，开始形成具有一定理论结构和研究规律的独特领域，该领域的研究者正在从对数字人文能不能成为一门“学科”心存疑虑而争论不休，转而开始专注于各类专门问题的探讨和整体共性方法论的总结。当然这与近年来数字人文研究基础设施的不断完善有关，除了大量的数据资源以最新的技术不断赋能研究人员之外，我们还拥有了颇具影响力的协会、学会和专业期刊，定期召开国际或地区性会议，具有稳定的基金支持，尤其是形成了本-硕-博的专业教育体系。目前的薄弱环节是基础设施的建设和提供者与新兴的数字人文研究者之间缺乏沟通对话，导致数据资源相关的平台建设和系统的标准规范尚未建立，正在成形的方法论体系缺乏实践检验，因此未能尽快成熟并得到公认。

以汉学（中国传统学术）研究为代表的中文数字人文研究也处在一个刚刚起步的阶段。早期的数字图书馆或数字典藏成果为当下的数字人文研究提供了重要的数据支持，然而从整体上看仍不系统，缺乏规划，各学科发展也很不平衡，研究成果较为零散、微观，多是对数字技术的简单应用、对过去研究的重复验证，或者是对西方研究的一种单纯模仿，还缺乏有影响力的、独创性的成果。究其原因，图书馆等人类记忆机构在数据基础设施建设方面的滞后是一个重要瓶颈[3]。相比西方国家，我们在数据获取方面的困难要大得多：数据系统之间缺乏联通，付费墙壁垒高耸，造成数据获取的不充分和不完整，或者缺乏必须的数据格式（如中文文献大多以图像方式提供，文本奇缺），影响到项目的成本、成果的水平，以及对数字人文研究方法的归纳总结和教育机构相关人才的培养等，这已成为中文数字人文发展的严重制肘。

本文试图基于中国目前对于数字人文的理论研究，探讨一种开放的数字人文服务平台设计，将数字人文研究范式与提供其支撑的基础设施建设联系起来，使其互相借鉴和促进，不仅满足一般人类记忆机构将数字典藏系统升级为基于数据的服务设施，发挥其全部潜能。重点在通过灵活可迁移的云平台架构设计，以及可互操作、热插拔、容器化的应用App生态建设，使所有机构的平台之间能够实现互联互通，并探讨应用关联数据、知识图谱、实体识别、机器学习等技术，提供人文研究各类文本、图像、社交网络、地理信息和可视化等通用工具的支持，长远支持数字人文项目的全生命周期管理。相信这样的总体性设计能够有助于数字人文方法论体系的丰富探索和尽快成型，从而帮助数字人文研究范式尽早确立。

2　数字人文：一种人文研究的新范式

2.1 数字人文催生人文研究范式转型

人文研究一般是人文学者针对特定问题，综合利用各种材料，透过一定方法，经过研究过程而得出结论并发表交流的完整流程。素材和方法是人文研究的两大要素。传统人文研究的素材可分为文献（文本或图像）、实物和抽象物（概念、角色等）等。传统人文研究的方法通常不是非常严格，一般依靠思辨和写作就能得出结论、完成研究，这也是为什么人们经常诟病“人文学科”缺乏科学性的原因。数字人文带来了方法

学的进步，我们首先可以从方法研究入手，从中找出数字人文研究可重复、可循证的一般规律。

数字人文来自于对人文研究进入数字时代所产生的方法学共同体的归纳，而根据提出科学范式概念的科学哲学大师托马斯•库恩的理论，学科共同体是学科范式的主要特征，因此我们可以认为，研究数字人文方法其实就是在探讨人文科学研究的一种新范式。从分析人文研究的素材和方法入手，我们可以初步掌握数字人文研究范式的基本轮廓。

把人文研究方法分为技术、行为和过程三个方面，有助于考察人文研究的基本方法范式。传统人文研究虽然很少涉及技术，但也绝非没有，例如考古研究中的探方、测量，以及在人文研究中被普遍采用的卡片摘录技术等，如果把社会科学也算上（社会科学与人文科学本身并无明显界限），各类调查、访谈、口述历史、民族志等研究方法都涉及大量的技术，早期数字人文的许多方法其实都来自于用计算机实现手工的工作。研究过程可以认为是研究行为的按一定顺序的组合，相同的技术和行为可以组合成不同的过程，对不同人文学科研究所产生的效果是不同的。以下会有文字专门讨论具体的研究“行为”（见2.3）。

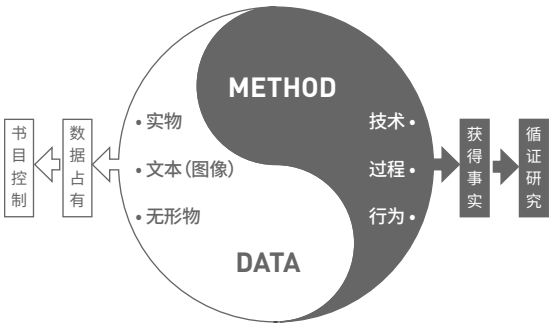

图1 人文研究的基本范式：数据 + 方法
Fig. 1 Fundamental Paradigms of Humanities Research: Data + Method

人文研究方法的技术、行为和过程在数字人文中借助信息技术的进步得到很大的发展，尤其是层出不穷的信息处理技术，可以说这三个部分正在成为数字人文研究新范式的重要内容，成为数字人文领域最重要的主题之一。图1展示了对这种人文研究范式的解构。

数字人文研究的“原料”可以分为数字文本、数码图像或由数字对象构成的“模型”，有学者称之为“数据态”。其中数字模型可以很简单，某个文本数据库可以代表某个人文主题的全部素材，也可以很复杂，复杂到作为某个真实系统的模拟（即所谓数字孪生，Digital Twins）。

数字人文的方法有两类，一是传统方法的计算机实现，例如搜索、分析、比较等，利用计算机只是比传统方法要快很多而已，最著名的数字人文研究案例——罗伯特•布萨神父编制托马斯•阿奎纳全集索引就是这样的例子；二是由计算机技术产生的特殊方法，例如统计、分析、聚类和可视化等，布萨神父最后建立了托马斯•阿奎纳索引服务，就属于对传统人文方法的一种突破。

从研究过程来看，数字技术和网络交流对过去从收集资料到成果发表简单的线性过程带来了很大冲击，其过程比传统人文研究要复杂得多，可以是来回反复的交互过程，成果发表和交流形式也多利用网络或社交媒体，具有迅速、便捷、容易追踪但转瞬即逝的特点，目前甚至还没有很好的计量与评价方法[4]。

无论是传统方法的计算机实现，还是由于计算机技术发展带来的新方法，如果从目前各类具体数字人文研究项目来考察，或者从不同具体人文学科在走向数字人文过程中的表现来看，其技术、过程和行为三个方面都可以归纳出许多不同的特征。图1虽然呈现了包括传统人文和数字人文在内的人文研究的统一范式，然而它并没有区分这些不同特征。应该说不同人文学科在迈向数字人文过程中的不同特点，不同学科在使用素材或研究方法方面的不同，都会对该学科领域基于数据的研究范式带来影响。例如文学或语言学偏重于利用文本处理技术，历史学则关注实体对象的时空呈现及相互关系，哲学需要将文本抽象为特定语义的概念，等，当然这类不同可以看成是数字人文通用方法细分要素的不同配方组合。这里引入图2，就是要展示数字人文方法受到技术体系和方法体系（指过程和行为）的双重影响，而作用于各门不同人文学科。当然这里讨论的还只是数字人文研究方法的一个一般性思考框架，目前无论是具体的人文学科，还是一般性的数字人文，其方法体系都没有定型，还处在发展变化中，也有待进一步挖掘整理。

2.2 传统人文与数字人文的比较

（1）研究过程方面

传统人文研究对于素材的收集、加工、处理是研究过程的开始，这是人文研究很重要的有机组成部

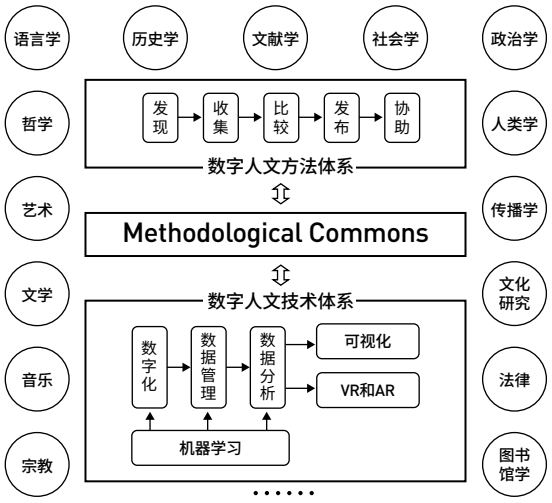

图2 数字人文相关技术体系和方法体系

Fig. 2 Technology System and Method System Related to Digital Humanities

分；而数字人文可以将资料汇集、处理的通用部分独立出来，作为研究基础设施的一部分，由专门的图书馆、档案馆等相关机构去完成，这就区分了基础设施建设工作和数字人文研究工作。目前数字人文领域大量的工作其实是基础设施建设工作，可以看到中文期刊数字人文的论文发表中大量来自图书馆信息档案学科，就是这个道理。但基础设施建设并不能代替数字人文研究，前者的目的是为了促进后者。

（2）素材内容方面

传统人文通常通过管理和操控载体化的文献取得内容，限于手工处理的效率，研究的广度、深度都受到限制；而数字人文研究基于数据，平台通常就能提供细粒度的知识组织，甚至建立了语义联系，使得材料的操控变得较为容易，能够进行更大范围深入研究，跨学科研究也更为容易。

（3）研究方法方面

传统人文研究大都采用定性的思辨方法，通过联想、比较、逻辑推理、思想实验等进行叙事或阐释；而数字人文可以采用建立模型和定量方法，进行文本分析、内容分析、时空分析、社会关系分析、统计聚类、可视化展示等，从某种程度上为人文研究提供了一定的可重复可验证的科学性保证。

（4）技术应用方面

传统人文研究可能会采用田野调查、问卷访谈等；而数字人文可以运用更多计算机技术，如机器学习、神经网络、语义标注、文本分析、量化分析、聚类算法等。

（5）科研协作方面

传统的人文研究大多是学者个人或小规模团队透过多年皓首穷经、苦思冥想，忽然顿悟，取得些许进展；而数字人文更强调大规模协同和社会网络交互，甚至大量采用众包方式，网络平台能否提供相应能力就显得非常重要。

（6）成果交流方面

传统人文基本上以出版图书或发表论文为最高标准；而数字人文可以同时推出网站、数据集、工具、软件、课件、博客文章、可视化作品、多媒体电子书等，专著和论文可以只是副产品。当然数字人文的基础设施可以更丰富和全面，包含计算设施、云平台、资源库、语料库等。

2.3 数字人文研究的行为范式

人文学者的研究行为可以类比于自然科学研究中的实验行为，是数字人文研究范式的重要来源。本文把数字人文方法区分成技术、过程和行为三个方面，研究方法是由研究行为在技术的支持下通过一定的过程组合和迭代而实现，因此人文学者的行为范式非常值得研究，可以认为人文学者在使用数字方法进行学科问题的研究过程中，其共性的行为方式就构成了行为范式。传统人文研究者可能都有独特的行为方式，同一个学派可能会基于相同的方式，而数字人文的价值就在于将其一般化，提取出共性的行为并以一定的技术进行实现，同时进行标准化。因此研究行为成为数字人文研究范式中非常独特的组成部分，本文称之为数字人文研究的行为范式，具体的行为国外称为“学术原语”（scholarly primitives）[5]，可区分为搜索、收集、阅读、协作、比较、发布等类型，每一种行为类型还可进一步分为子行为，例如搜索可以分直接搜索、浏览、探索、存取、链接等；收集可以分为爬取、汇聚、组织等；阅读有浏览、评价、远读、细读、互读等；协作有建立网络、咨询、分享等，如表1所示。

人文研究的具体行为在数字人文平台中都可以以一定的技术加以实现，这些行为与实现技术之间的关系参见表2。每一个子行为都可以开发成目前业界流行的“微服务”，以更加适应灵活先进的云原生计算环境。

传统人文的研究过程通常是从占有材料开始，然后经过发现事实、提出假设、收集资料、分析比较、归纳

表1 基本的研究行为
Table 1 Basic Research Behaviors

搜索（调查）Searching（*investigating*） 直接搜索 Direct searching 链接 Chaining 浏览 Browsing 探索 Probing 存取 Accessing	收集 Collecting 爬取 Crawling 汇聚 Gathering 组织 Organizing
阅读（阐释）Reading（*interpretation*） 浏览 Scanning 评价 Assessing 远读 Distant Reading 细读 Close Reading	协作 Collaborating 建立网络 Networking 咨询 Consulting 分享 Sharing
交叉原语 Cross-cutting Primitives 记录 Note-taking 翻译 Translating	比较 Comparing 传递 Delivering

表2 数字人文研究行为及其技术实现
Table 2 Digital Humanities Research Behavior and CorrespondingTechnical Implementation

技术类别	具体实现技术（不限于）	支持行为
数字化技术	扫描；拍摄；采样；捕捉；图形设计；3D 建模	收集 Collecting
数据管理技术	文本编码；语义描述；本体建模；数据库设计；多媒体搜索；语义搜索；数据看护；名称实体提取 API	搜索 Searching 发现 Discovering
数据分析技术	文本分析（词频、共现、关联、向量、概率）；聚类分类；主题分析；内容挖掘；时序分析；地理空间分析；社会关系分析	阅读 Reading 注释 Annotation 阐释 Interpretation
可视化技术	信息美学；知识地图；主题图；关联呈现；场景模拟；历史仿真	比较 Comparing 说明 Story Telling 发布 Delivering
VR/AR 技术	人机交互技术；脑机界面；认知技术；互动测量；游戏化学习；计算机竞技	协作 Collaborating 说明 Story Telling 发布 Delivering
机器学习技术	自动分类；图像视频音频识别和分析；个性化服务；精准推送；深度学习；超级计算；机器绘画；机器作诗	所有行为

整理，得出结论并进行发表交流。数字人文研究由于素材更多、数据量更大、时空跨度都可能不同以往，因此研究过程可能会变得非常复杂，更多的在提出假设之后需要建立模型，然后将分析比较等研究过程，透过技术手段操控模型中的数据和各类参数来验证、修改或推翻假设，最后得出结论。因而如何利用计算技术实现研究目标也需要有一定的计算思维基础。当然，其前提是数字人文平台能够支持这样的复杂性。

什么是数字人文或什么是好的数字人文，目前还很难划定一个清晰的边界或给出明确的标准。尽管很多人认为，仅仅采用搜索引擎查找资料，或用文字处理软件从事研究而撰写的人文研究成果并不能算是数字人文，但为什么搜索了专门的数据库、用了可视化软件或一些分析工具就可以是数字人文成果呢？Unsworth 认为[6]需要利用数字技术对人文问题进行“表征、建模或模仿”（a practice of representation, a form of modeling or mimicry），才算数字人文（人文计算），然而这个界线也是模糊的，可能未来我们能够划清界线，但那时可能设定界线已经变得没有意义了。但无论如何我们可以认为，从现在开始，人文研究赖以进行的基础已经不是“文献”，而是数据，由此带来基础设施、平台方法乃至评价标准都开始完全不同。我们现在还站在数字人文的门口，新的“范式”正在成型，生逢其时，这是我们的幸运。

2.4 数字人文平台建设现状

数字人文平台是为数字人文研究服务的，也是实现数字人文研究范式的重要的基础设施之一。平台建得好不好最终要通过数字人文研究成果来检验。因此在建立之初首先需要了解数字人文研究人员的需求，了解数字人文研究的一般规律，以及方法、过程和行为，否则也无法设计出好的数字人文平台。当然，数字人文平台“兼容”传统的人文研究是一个前提条件，在很大程度上数字典藏系统应该就能满足需求，然后可以进一步升级开发“真正的”数字人文平台，向人文学者全面提供基于数据的研究基础设施服务。

目前的数字图书馆系统可以看成是一种初级版本的数字人文平台。由于其大都只是将传统的文献扫描成图像，结合元数据库提供有限途径的查询，功能十分有限，基本上只是传统图书馆的一种载体转换，无法满足数字人文研究的进一步需要。虽然有一些平台已开始提供一些工具，例如分词、标点、批注、词云、格式转换、实体提取、人物关系呈现及可视化等，并采用了众包理念，但总体上还较为简单，集成了一些成熟度不一的功能，没有结合人文学者的领域和场景，用户体验不够好。

现有的数字人文平台存在的最大问题还是技术上的，在内容管理上尚未采用知识图谱为代表的语义数据管理技术，还是关系数据库或者全文数据库；在体系结构上虽然已注意借鉴云计算技术，但还没有充分考虑以微服务和容积技术为基础的云原生架构，也没有考虑技术架构和内容架构分离的设计。因此很难满足人物、地点、时代、事件或特定事实主题的资料查询需求，人物或实体之间逻辑或关联关系的延伸查询需求，时空主题范围的统计分析需求以及可视化呈现的需求等。现在的认知计算技术结合了机器学习和人工智能，已经能够提供语词概念或图像实体的提取与分析、特征比较、相似性聚类等，数字人文平台完全可以应用最新技术，实现最新功能。从平台的角度来看，还有较大的提升空间。

3 数字人文研究的中文资源与研究方法

人工智能专家李飞飞曾说："作为科学家，最吸引我的是能够不断去拓宽人类知识的边界，不断问新的问题，并且发明工具来解决这些问题"。数字人文带给人文研究最有价值的地方，也就是它能够极大地拓展我们提问题的能力，从而拓展人文研究的新疆域。它使研究者能够面对海量甚至是"全量"数据进行研究，能够利用各种工具对数据进行分析、比较、挖掘、关联。这些数据是传统人文学者终其一生都不可能看完的，方法手段也是传统手工所无法想象的。因此，数字人文的价值不仅在于它提供了研究的素材，同时也给予了强大的工具和新的方法。以下从中文研究资源和方法两个角度，简述数字人文相关情况。

3.1 中文数字人文基础资源现状

史料乃人文研究之本，而所有人类活动纪录皆可为史料。图书馆等记忆机构自古以来不仅是人类思想纪录的保留地，也同时是人文思想的孵化所。著名的亚历山大图书馆以收藏人类所有知识为己任，但其鸿富的收藏是为了聚集天下英才从事研究写作和知识传授，在其不长的历史时期聚集了数百位先贤哲人，为中世纪乃至一千多年后的文艺复兴留下了非常宝贵的知识财富。海量的资源提供了极其丰富的知识基础，使畅游其中的学者具有完全不同的起点，站在巨人的肩上他们才更有智慧。中文资源亦是如此，渊远流长，历经两千余年流传，培育并滋养了灿烂的中华文明。

自上世纪九十年代以来，中国传统学术相关资源的数字化已获得长足发展，目前通过网络已基本上皆可尽知。然而中文数字典藏的最大特点是以扫描图像为主，总体上转换成文本的数量不及三成，且质量良莠不齐；另一个特点是大多数典藏资源都分散于各家出版机构或数据库厂商，研究机构很少提供典藏资源的开放服务；第三个特点是所有系统提供的功能都很简单，大多只能进行少量字段的查检。虽然也有部分商业化特藏库做得不错，提供全文搜索，并且从文本质量到图文对照都比较人性化，然而总体来说与国外一些数字人文平台的水平无法比肩。当下的技术已经提供了可能性，我们理应做得更好。

2018年3月，哈佛大学包弼德教授在上海哈佛中心组织召开了"中国历史研究的网络基础设施国际研讨会（International Conference on a Cyberinfrastructure for Historical China Studies）"[7]，遍请当今与中文资源及平台界相关人士和机构代表，进行了为期三天的研讨，共有近60场各类会议（sessions and panel discussions），142人次发言，几乎将中文传统学术资源一网打尽。包教授将主要的中文传统学术资源库分为三类（见文末附表1）：平台与工具类、文字/文本图像数据库类以及数据库类（主要是专题或文本库），悉数邀请其代表参会。

包弼德教授的列表展示了中文数字人文资源的建设现状，应该是非常全面了。传统人文学者在从事研究时大部分时间都在遍访资源，常常必须通过打听或者高人指点，有时是偶然机缘，才有可能获得一些线索，是不是合用还要经过人工实际翻看，查找资料与研究者本人的学养、经验都很有关系，没有经验的初学者甚至都无法查到合适的资料，查到了有时也不能判断。对于传统人文研究来说，检索材料的过程经常是作为正式研究过程的一部分，而不是准备。

中文传统学术资源其实是有限的，转换成数据库之后也不会增加。但是转化成数据库之后能够在很大程度上降低人工检索的难度。因此数字人文学者能够在更大范围、更准确地查到所需资料，消除专家与普通研究者存在的信息不对称，让"资料（平台）面前人人平等"。这样的话，查找资料的过程可以从研究过程中

独立，学者能够把更多的时间和精力花在本学科的问题研究上，而不是数据获取上。这是数字人文的最大好处之一。

据笔者不完全估计，目前中国传统学术研究常用的资源大致有：

古籍：根据目前对于古籍的定义，不重复的应不超过20万种，版本数不超过50万种，已基本完成数字化扫描，其中四分之一（约5-6万种）大致完成了文本化，约不超过100亿字。已实现文本化的古籍有很多失去了版本信息（或被加工出版机构根据一种或数种所谓“权威版本”进行加工）。

民国图书：保守估计不重复约有15万种，已基本完成数字化扫描，文本化数量应在300-400亿字，但大多分散在各出版机构。

现代图书：不重复至少500万种，基本都有数字化版本，但并非文本化，其中一多半以CEBX（Common e-Document of Blending XML，基于混合XML的公共电子文档）格式存在，总量约上千亿字。

近代期刊：至少2万种，约800万页，基本完成数字化扫描，但文本化只有50亿字左右。

近代报纸：总量约100万拍，基本完成数字化、文本化（如申报等一些大报）约30亿字左右。

现代期刊：近30年的期刊基本都已经文本化，主要为CNKI等数据库商所掌握。

现代报纸：近30年经汉字照排的报纸基本都有文本，一些大报（如人民日报）也已完成了文本化，但因格式和版权问题，能得到开放应用的很少。

档案馆藏：经过近十多年来国家的大力投入，数字化已基本完成，而且绝大多数在数字化时已经完成了文本化。

博物馆（美术馆）馆藏：真正的数字化（保存级）近年来刚刚开始，许多藏品需要3D建模，随着技术的成熟成本逐渐降低，规模逐渐增大。

如果说包弼德教授的中文传统学术资源列表还不能包罗万象的话，近年来各类收藏机构的中国传统学术资源数字化已经全面展开，数据库已成为中国传统学术研究者检索资料的主要途径。但矛盾的是学者们并没有感到查找资料比以往更方便。这主要有如下问题：

（1）系统较为封闭。就如同古代藏书楼，宝贝秘不示人，是无法得到充分利用的。很多系统甚至不开放元数据，无法让学者查询是否有某些资料。虽然大量的中国传统学术资料都已过了版权保护期，但国内的公藏机构也大都不开放，恐怕被人盗取，还有不少出版机构拿来影印或重新出版，使其又变成“有版权”出版品，依旧在“付费墙”后面，依然没解决开放问题。而中国大陆以外地区的典藏机构近年来逐渐公开了大量资源（见附表2）。

（2）系统之间互不联通。资料分散在各处，必须分别去查，很多甚至没有上网，寻访依旧不易，找到后经常需要手工抄录，然后再进行对比、分析等工作，有时只查元数据并不能满足需求，系统中缺乏研究所需的关键信息，如版本、格式等。

（3）资料准确率低。讹误很多，数字化会放大错误，且缺少修正机制。

（4）使用便捷性差。只是解决了“知道”和“得到”问题，后续所有工作都还是手工的，并不能体验到计算机能够提供的更多好处，例如保存、统计分析等。

以中文数字图书馆（或称为数字典藏）建设为主的数字人文基础设施建设正方兴未艾，目前几乎所有的人文研究都需要从数据获取和整理开始做起，因此大量的数字人文项目其实还是数字典藏项目，这类项目被David Golumbia称为狭义的数字人文，是最容易获得资助的。我们从2020年中国数字人文年会（2020 China Digital Humanities Conference ,CDH2020）的获奖项目（见表3）中可以看到这类项目的一些特点：

（1）数字化逐渐让位于数据化；知识库逐渐增多。

（2）独特的领域应用做得更好，利用技术也很到位，能够提供更多的研究支持。

（3）“低端果实”（low hanging fruit）较多，主要是一些以数字化方式重复已知的结果，或以可视化方式展示历史、人物、事件等主题等。当然其中做得好的，也包含大量的研究成份，以及很多设计和数据处理工作量，也不是没有意义。

（4）以教育、普及和技术培训为目的的项目也有不少。这类项目经常会昙花一现，无法在基础设施中沉淀下来。

从总体上看，当前中国传统学术研究相关材料分布极广，技术各异，标准不一，数据质量良莠不齐，整合有相当难度，利用极为不便。

数字人文研究的素材其实不止于历史资料。当今数字时代大量的数字原生材料，例如美国国会图书馆

表3 CDH2020获奖项目情况
Table 3 Some Information about the Award-winning Projects at CDH2020

项目编号	项目名称	项目类型	主题领域	学科	资源类型	内容覆盖
xm11	家谱知识服务平台	平台建设,工具提供	家谱载体	文献	文本,图像	广
xm28	数位人文分析系统与个人 DH 研究平台	平台建设,工具提供	综合性	历史,综合	文本,结构化数据	广
xm01	CBDB 查询系统第二版	系统建设,工具提供	人物数据	历史	结构化数据	广
xm44	高迁古村数字记忆网站	系统建设	特定主题	社会学	综合	专
xm79	宋元学案知识图谱可视化系统	知识图谱,可视化	特定主题	历史	图谱	专
xm19	IIIF 敦煌壁画数字叙事系统	IIIF 系统建设,叙事工具	特定主题	敦煌学,壁画,艺术,历史	图像,文本	专
xm10	董其昌数字人文展示系统	可视化,数据组织	特定主题	艺术,历史	图像,文本	专
xm67	中国古代皇室家族树	可视化	特定主题,人物数据	历史	文本	专
xm09	唐宋文学编年地图	可视化,数据组织	特定主题	文学,历史	结构化数据	专
xm24	中国多世代人口数据库	系统建设	数据库	人口学,历史	结构化数据	广
xm59	南京地区侵华日军慰安所的 AR 故事地图	AR,GIS	特定主题	历史,战争史	综合	专

收藏的Twitter档案和中国国家图书馆保存的新浪微博，都是很有价值的资源，很多人文社会科学研究都可以在其中找到宝贵的数据资料，但对这些原生数字资源如何收集组织管理，并提供利用，目前似乎并没有找到很好的方法，而且从各国的实践来看当前也不是图书馆档案馆等人类记忆机构当然的职责所在，将来有可能与传统数字人文素材之间的历史联系会中断，产生一段材料的真空期。我们现在应该开始重视这个问题，把数字资源的保存组织也纳入到数字人文平台建设的内容中去统一考量。

3.2 中文数字人文主要研究方法

分析CDH2020的获奖优秀论文（见表4），可以大致了解目前国内数字人文研究通常采用的方法和研究水平。年会一共评出18篇获奖论文，其中一等奖3篇，二等奖5篇，三等奖10篇。18篇获奖论文中有10篇关于基础设施或技术研究，后者涉及建模技术、语义化聚类等，只有8篇可以算做人文主题的探讨，包括阐释学或叙事研究、色彩研究、文化批评等，其中有一篇严格算来也并非数字人文研究，只是它以“数字人文研究”这一现象作为研究的对象，是一篇以非数字人文方法研究数字人文主题的文章。

从表中可以看到，有不少论文是关于资料收集、建库、开发系统、提供功能或方法研究的论文，如编号09130001、06190011等，其中一等奖的三篇论文都是关于数字人文方法、平台和框架研究，并深入到具体人文学科内部，以学科特征为立足点的探讨，比过去泛泛而谈数字化、平台开发或研究方法进了一步，但依旧是数字人文基础设施建设探讨，而不是严格意义上的、以数字方法针对人文问题的研究。这些论文也呈现了一个有意思的现象，即基础设施与技术探讨常常是由跨学科团队完成，而人文主题则多由领域专家独自实现。

这种以基础设施和方法探讨为主的研究现象说明，当前的数字人文研究还处于一个尚未成熟的初始阶段，说明基础设施建设尚未到位，数字人文方法也没有系统成型。不论是人文学者、技术专家，还是资源提供者，都热衷于探讨如何建立更好的研究平台。目前数据获取、加工、组织和平台工具的开发和提供还是主要矛盾，在可以预见的未来，一旦基础设施基本到位，数字人文的研究将真正由人文学者主导，并以人文学科的问题为引领。

当然这也要求基础设施建设与人文学者研究之间逐渐形成一个明确的界线，人文研究的一般方法与具

表4 CDH2020获奖论文主题及研究方法

Table 4 Topics and Research Methods of the Award-winning Papers at CDH2020

论文编号	论文关键词	论文类型 / 采用方法
08150012	“冷门绝学”,音韵学	文本分析,语义提取
09130001	文学,语料库,基础设施	系统开发
06190011	藏学,框架,整合	系统开发
08040001	馆藏利用,方法研究	叙事,方法探索
08120003	哲学,方法研究	文本,主题建模
08150002	历史,还原	叙事,文本提取
08150021	图片,广告	聚类,神经网络
08180003	视频,文化遗产	视频信息提取
09180001	哲学	主题建模,方法研究,人工智能
09080001	时间抽取,报刊资料	实体识别,文本分析
08180002	色彩研究,山海经	实体识别,文本分析
07020001	文化研究,西游记,影视作品	文本分析,比较研究
07150001	诗歌,李白,意象	文本分析,情感分析
07260001	文学,历史,语境	文本分析,GIS 分析,阐释
08140005	理论,隐喻,知识生产	思辨,阐释
08140008	历史,文学	阐释,文本分析
08150005	历史,还原	阐释
08150008	数据处理,人称代词	神经网络

体人文学科的特定方法之间也需要有一定的分野，这样才有利于形成规模和分工协作，而传统人文研究是没有这个界线的，人文学者承担了从资料收集整理到结果交流发布的所有过程，使得研究一直处于零散、琐碎、凭借个体经验和难以合作的原始状态。

不同人文学科的研究对象和问题不同，对应于计算机所存储的媒体类型和处理方式也不同，这或许是造成研究方法是否具有通用性的根本分歧。例如文本是几乎所有人文学科进行研究最常用的材料类型，它也是计算机所能处理的最常见的信息类型，这一点数字人文界毫不陌生，因为罗伯特•布撒神父的工作几乎伴随了计算机文本处理技术进步的全过程，而布撒神父的专业是神学，却是利用计算机实现了属于图书馆学的索引编制技术。文本对于语言学来说就是最直接的素材，理所当然地会利用各类查询（例如追溯肇始源头）、统计（频度研究如词云，或共现研究）、比较（词性、变化）等“行为”来研究语言现象，中文自然就有切词、句读的需求；文学稍有不同，它更多地涉及文体、风格、修辞、情感方面的问题，有时也会引伸出去，探讨作者或虚构人物的关系、时代背景或文学批评；文学有时也会涉及到文献版本的比较、考证、鉴定等，这却又是图书馆学的传统内容；哲学、神学、政治学等虽然也是通过文本进行研究，但更多的却是将文本当作一种抽象概念，思想史、观念史研究中需要应用大量的抽象概念，这些概念常常可以建立起一种复杂的语义或逻辑联系，从而辨别社团、思潮及流派谱系等，这种联系正好是语义技术的强项，应用本体语言完全可以将复杂的语义联系进行代码化，从而就具备了“机读”的能力，可以充分利用计算机的优势进行管理和利用。历史学、考古学等相对来说就更加复杂，它们通常是将文本作为实体对象及其关系的容器，从中可以提取丰富的场景和事件，提供叙事和阐释的根据，或构建社群、谱系。对于这类文本，计算机也可以利用机器学习和知识图谱等技术，构建一系列“数字孪生”模型，从而可以让历史学家像坐上时光机器一般穿越到历史故事中去，甚至可以利用不同的假设来推演可能的结果。

图像是艺术、考古、人类学、民族学等人文学科不可或缺的资源类型，计算机可以从色彩、图案、纹理等风格特点进行研究，也可以对其进行模式识别，或者对各类实体对象进行识别、比较、分析统计等，帮助得出结论。图像资源尤其对于中国传统学术研究有着无比重要的意义，比西方数字人文研究的意义要大很多。首先，因为中文传统学术典籍目前还不可能都转为文字，OCR的准确度不够，成本巨大，而且操作系统对汉字标准字符集的支持数量也不够用；其次，中文数字人文研究通常只依靠纯文本是不够的，还需要有图像所负载的丰富信息作为辅佐，才具有“循证”价值；最后，最新数字人文平台如IIIF所提供的图像管理能力，能够使图像比纯文本更方便研究。IIIF对图像的管理方式还可以进一步应用于视频、音频等媒体形态，将来还可以有3D模型、交互式数据格式等，这样就完全超越了仅仅由文本组成的平台，成为一个多模态服务平台，数字技术提供的强大工具能够使人文学者超越传统人文研究基本上只是依赖文本和少量图像的局限，对素材的操控能力得到很大的增强。

大数据技术目前在社会关系分析和可视化呈现方面已经非常成熟，另外机器学习及人工智能技术提供了大量数据加工的自动化能力，这使得数字人文更适合进行跨学科、地域、族裔、语言的大规模比较和综合研究，这在传统人文的研究中是不可想象的。

上述讨论的大多是计算机技术赋能人文研究的通用能力，这种通用与专用的分野是不确定的，随着计算机信息处理能力的提高，所能提供的工具肯定会发生变化，最极端的方式说不定是人工智能方法能够取代一切，学者所需做的只是提出并阐释问题而已，余下的工作都交给机器即可，或者至少可以半自动地帮助学者一步步获得结论或解决问题。

4 平台需求与方案设计

当今时代已不再可能举全国之力穷天下收藏，兴建四库全书那样的项目，开发包罗万象的知识平台，现在甚至连某一学科或主题领域的资源都不可能一网打尽。因此我们在构建数字人文平台或开发人文资料数据库时首先应考虑自身的优势和特点，选取一定的文献类型或学科主题，充分考虑服务对象特点和需求，设定有限目标，并做好长期建设的准备。

目前很多中文传统学术资源收藏机构已经开发了一些颇具特色的数字人文平台，如CBDB、DocuSky、MARKUS等，应用了许多先进理念和最新技术。本文希望着眼于未来互联互通，对构建一个整体化的中文数字人文研究的基础设施提出一些设想。包弼德教授曾在2018年提出过类似的想法，他建议构建一个“中国研究的基础设施网络”，希望通过各国中文资料收藏机构的密切合作，开发一个通用平台[3]，使中文资源能够互联互通，进一步促成共建共享。

这是一个非常有远见、有现实意义和可行的建议，但这个平台不必是“一个”平台，而可以是整个中文基础设施共同构成的分布式网络服务，即可以由相关中文资源收藏和研究机构各自建设，但遵循共同制订的技术标准和互操作协议，这样就保证了资源获取和服务的互联互通；同时制订一定的合作机制和业务模式，这样又能够促进互惠互利和可持续发展。

因此，本文探讨的平台即是一个在功能上力求完善、能够满足当下需求的独立的数字人文平台，又同时在体系架构上兼顾了基于最新语义互联网技术的互联互通，是一个尚未实现但完全具有可操作性的设计方案（如图3所示）。以下分系统先进性（4.1）、资源完整性（4.2.1）、功能完备性（4.2.2）、用户友好性（4.2.3）和工具丰富性（4.3）等5个方面进行阐述。

4.1 应用系统先进性

应用系统的先进性可以从两个维度来考察：系统维度和应用维度。系统维度主要指系统架构的先进性，又可以分为技术架构和内容架构。应用维度是指所开发的应用系统是否能提供满足机构需求的完整的解决方案。技术架构提供基于计算机技术的功能实现，内容架构主要提供基于数据管理的知识视图，这两者共同满足数字人文研究需求，解决数字人文研究的痛点问题。应用系统则主要包括馆藏业务管理系统、长期保存/典藏系统、知识库系统、服务应用展示系统四个方面（见图4）。

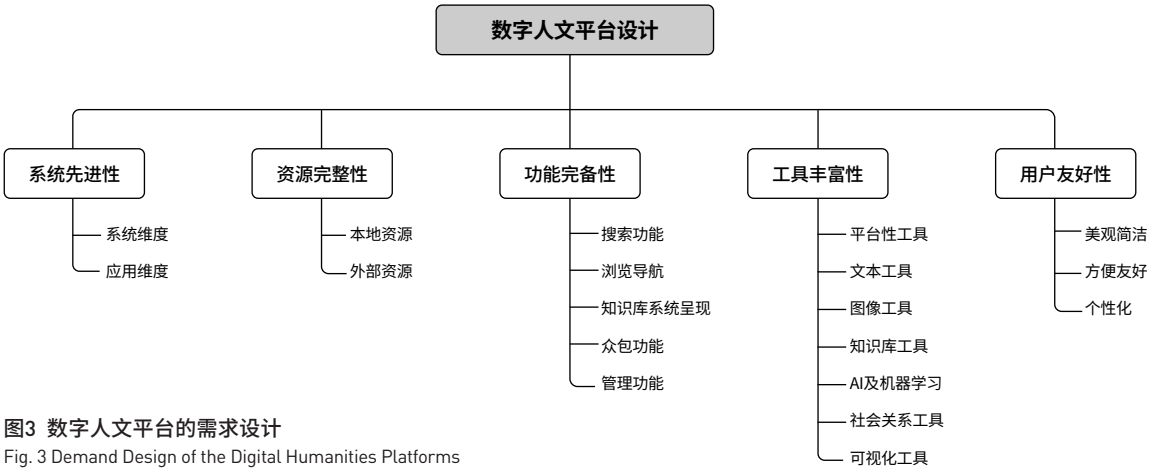

图3 数字人文平台的需求设计
Fig. 3 Demand Design of the Digital Humanities Platforms

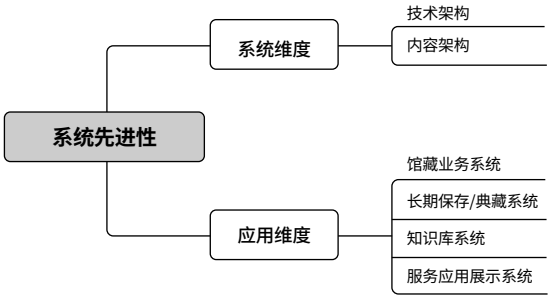

图4　应用系统先进性
Fig. 4 Advanced Systems of the Digital Humanities Platforms

4.1.1 系统维度

（1）技术架构

系统维度首先看技术架构。目前以微服务、容器、容器编排、服务网格、开发运维一体化（DevOps）、无服务器架构等理念为特征的新一代“云原生”技术正在席卷互联网应用。拥有传统IT无法比拟的优势，可以帮助用户高效享受云技术的灵活性，使应用进一步微型化、轻型化，支持更加灵活的松散耦合，更加独立于底层基础设施平台，从而能实现热插拔、平滑、快速开发、迅速扩展、稳定运维、高容错等，大大降低应用成本，提高运行效率。目前云原生已经成为云时代最新的技术标准。

当前还没有数字人文机构采用云原生技术，但图书馆领域正在流行的“下一代图书馆服务平台”（Next Generation Library Service Platform，NGLSP）普遍采用微服务架构，尤其是美国开放图书馆基金会（Open Library Foundation，OLF）支持的开源FOLIO平台（Future of Libraries Is Open，FOLIO）更是支持了云原生技术进行部署实施，其前后台分离的设计和“平台+App”的架构有助于形成一个开放的软件应用生态（见图5），数字人文平台可以作为图书馆服务平台的一个有机组成部分，共用其中某些模块（例如用户管理、资源管理等），也可以单独拆分出去完全独立，通过API进行互操作。

该设计可以进一步支持目前如日中天的技术概念，即“中台”技术（见图6），可形成独立的业务中台、技术中台、数据中台和AI中台。所谓中台，可以理解为将一些能够重复调用的系统资源（数据资源、计算资源、软件及算法模块等资源）独立并共享出来，支持平台中的各类前台或其他应用模块灵活调用，在技术架构上具有无可比拟的先进性。当然该技术毕竟发展还不到十年，其成熟度和标准化程度还不是太高，微服务带来的应用复杂性还难以预料和掌控，这也是新技术必然带来的风险。

参考上述图书馆服务平台的系统架构，一个独立的数字人文平台可以包含文献层、数据层、接口层、业务层（或称服务层，包含各类工具调用）以及展现层等，依次提供技术、资源、平台、服务和界面等相关功能，如图7所示。随着基于文献的数字人文服务逐渐向基于数据的服务转变，文献也可以看成一种特殊的数据类型，纳入数据管理统一的数据格式模块，内外部文献和数据可以通过一定的协议规则进行发现和获取，并通过标准接口进行整合，各类平台内服务和外部服务也可以通过制定行业标准进行规范化整合，从而达成数字人文平台的互操作，于是可以很好地实现包弼德教授关于人文资源互联互通、共建共享的设想。

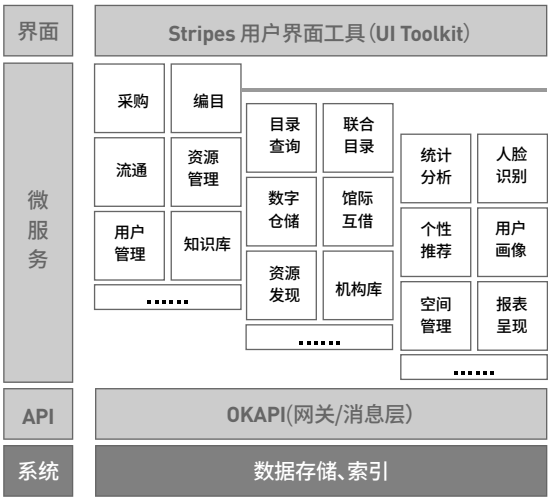

图5　下一代图书馆服务平台 FOLIO 的系统架构
Fig. 5 System Architecture of the Next Generation Library Service Platform FOLIO

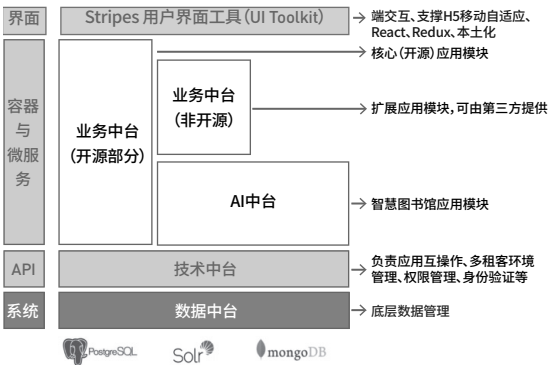

图6　下一代图书馆服务平台 FOLIO 的中台设计
Fig. 6 The Middle Platform Design of the Next Generation Library Service Platform FOLIO

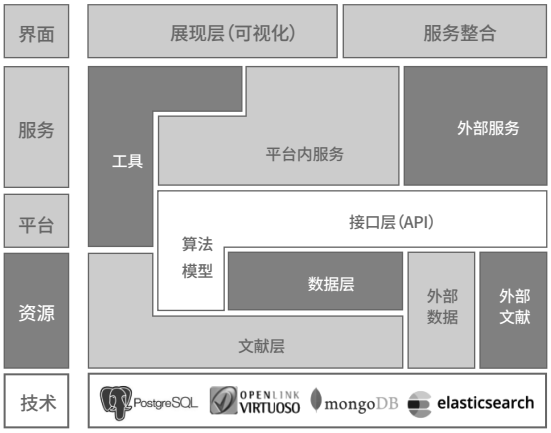

图7 数字人文平台系统架构图示

Fig. 7 System Architecture of the Digital Humanities Platforms

从数字人文的应用场景来看，上述系统架构有一定的独特性，可以很好地支持和解决一些其它技术很难解决的问题：

① 知识单元的标识及其管理问题。所有对人文研究具有独立意义的实体或信息单元，如文献，或人、地、时、事、物、事件、概念，以及各类属性和取值词表等，都需要有独立的标识（即ID），并统一ID编码标准，通常用http URI，其相互之间的关系如有必要可以通过建立本体知识库来管理。当然建立过程可以采用自动抽取加人工辅助校验方式。

② 支持多种协议的跨网域搜索发现或获取链接。例如OAI-PMH规范，各类RESTful+JSON的API规范、联邦检索页面分析规范等。

③ 微服务的容器及编排规范。

④ 多种数据类型的管理，包括底层关系数据库、图数据库（包括三元组语义数据）、对象数据、流媒体的管理。

⑤ 复杂但统一的用户及授权管理，包括远程访问管理。

云计算的极致状态是完全去中心化的分布式计算，目前的最新发展是以区块链应用为特征、被称为Web3.0的一套新的网络平台，这使得所有人文资源在底层都可以应用区块链技术进行确权和保护，包括二次文献上链，对象数据采用IPFS、Arweave等去中心化网络存储方式提供永久存储，同时对每一个馆藏单元赋予非同质化通证（Non-Fungible Token，NFT），这就解决了既要保护，又要最大程度开放的矛盾。只要设计出合理的运作模式，就能以某种智能合约方式形成去中心化自治组织（Decentralized Autonomous Organization，DAO），从而实现完全的自我运作，其他对于数字人文平台所有的附加需求都可以围绕这个Web3.0的资源体系进行设计开发。目前这种设计还十分超前，虽然技术都已成熟，但应用尚属首次，有些还是纸上谈兵，尤其在文化遗产领域尚未有任何具体实现。目前整个以Web3.0为基础的元宇宙应用非常缺乏具体的应用场景，人类记忆机构的文化资源正好可以为其提供丰富的想象和精彩的实现。

（2）内容架构

内容架构是数字人文应用系统非常独特的架构，也是语义技术逐渐成熟带来的一种能力，它通常通过领域驱动设计（Domain Driven Design，DDD）而获得。数字人文平台的内容架构反映了平台中的数字化知识内容的语义结构，这个结构可以以知识本体、关联数据、知识图谱等方式进行形式化描述和表达，例如以各类描述词表对人物、地点、时间、事件和各类对象的各类属性和关系进行编码，使计算机可以对表达知识的这些语义数据（可以理解为RDF数据）进行操作，从而可以认为这些数据是机器可“理解”的，以至于可以认为整个知识库中的大量内容都是真实世界的一种映射，甚至可以能够让机器进行一定的“事实推理”。传统的数据库只能对字符串或二进制数据（如图像数据）进行操控，如全文检索也就是一种完全基于字符的匹配。数字人文平台对于信息资源的描述和组织可以认为是一种“数据化”过程，这一过程不一定完全依靠人类来做，很多都可以通过目前越来越成熟的机器学习和人工智能来实现。一旦机器能够读“懂”存储的信息所蕴含的知识内容，数字人文平台就能帮人文学者做很多事情，可以成为能力超强的“研究助理”，它不会遗忘任何一个知识细节，并且具有超快的计算能力。

有这样一些需求涉及内容框架：

① 一致性/相似性计算。

② 工作流定义对研究流程的支持。

③ 各类图像功能（如图像查询、对比、标注等）的支持。

④ 文本与图像关联（可提供加工平台，或研究对比）。

⑤ 提供证据链服务（记录从底层文献到研究结果的整个过程中实体来源及变化，包括引用参考等）。

⑥ 海量数据可视化支持（远读）。

⑦ 事实的可信度计算及排序（需建立可迭代的可信度模型）。

⑧ 众包数据加工平台的数据管理。

⑨ 数据系统迭代进化的支持（数字化、文本化、数据化（实体提取、建立关联等））。

内容架构是以“数据”为基本单位，这里的数据是指能够被计算机处理的（即经过形式化，或至少是代码化的）、具有独立标识（例如URI）的最小语义单元，目前表示为RDF的关联数据是一种最佳实践，其它有不少简化方法（例如采用图数据库技术实现的、不要求数据有全网域唯一标识的“知识图谱”）虽然也能实现一些功能，但并不属于具有一定完备性的知识库系统。基于数据的系统能够进行组合、嵌套、递归从而成为更大的“数据”，也可以有自己的标识，从而可以以各种格式组合成各种知识单元发布于各类媒体中。

人文平台中的知识内容既然以“数据”的方式存在，就应该符合当前在研究数据管理实践中被广泛认可的FAIR原则，即科学数据应具有可查询（Findable）、可获取（Accessible）、可互操作（Interoperable）并且可重利用（Reusable）等性质：

① 可查询指数字人文平台中的数据应该很容易被人或者机器查询到。这有赖于相关的数据集或者数据服务是否以清晰明确的方式进行标识、描述、注册和索引。给数字资源分配一个唯一永久标识符是一项基本要求，同时数字资源应该有充分的元数据注释，数字资源的主要特征应该以标准格式被记录，应该在公开的数据库存储和索引等。

② 可获取指数字人文平台中的数字资源的获取方式应该进行清晰定义，包括如何获得受保护数据的使用授权。在理想情况下应该是一种自动化的方式进行获取数据的验证，判断是否符合授权条件，至少元数据应该是无条件可获取的，即使在原始数据已经不再提供服务的情况下也应该能够获取元数据。

③ 可互操作是指如果同一个实体对象有两个或者更多的数据进行表达，系统应该可以自动进行指代或整合。网络服务可以自动判断它与目标数据之间是否兼容。这要求数据资源或者网络服务的描述具有语义上足够的清晰度。

④ 可重用是指要根据研究领域的标准，对数据的来源信息进行记录和跟踪。这些来源出处信息包括准确的数据描述、取用方式和应用许可等。这样，无论人还是机器都可以判断目标数据资源是否可以重用，可以以怎样的方式进行重用等。

这四个原则与关联数据的五星原则很类似，因此如果采用关联数据技术，则很容易满足FAIR原则。但并不是所有数字人文平台都能够很方便地利用关联数据技术，其中涉及实现的复杂性、效率和成本等问题，以及语义技术本身的成熟度问题，因此目前的数字人文平台大多采用最成熟可用的技术，以关联数据甚至智慧数据为代表的语义技术是一个未来发展方向。

4.1.2 应用维度

数字人文平台大多由人类记忆机构，如图书馆、博物馆、美术馆、档案馆等进行建设和维护。作为数字人文基础设施的主要组成机构，他们的主要业务和服务都是围绕人文资源展开的，一个较为完整的平台通常可以分为四个层次：

（1）馆藏业务管理系统

这主要指对物理藏品或数字藏品的载体，从收集、入藏到转移、剔除或损毁的整个生命周期过程的管理，包括藏品管理系统。它提供了所有馆藏内容最初的来源和版本信息，是循证研究的源头，并通过业务过程的管理保证整个馆藏体系是一个不断发展变化的“活”的有机体。

（2）长期保存/典藏系统

即上述业务管理系统中的藏品管理系统的数字化版本，通常是能够保留最真实和完整信息的保存级数字文件，借助显示或其它设备，能够还原物理藏品的内容或形态，高级形式可以看成是每个馆藏的“数字孪生”，可供研究人员进行各种实验、模拟和深度研究。当然，任何数字化版本都不可能保留原始对象的所有信息，总是会有所损失，所以依赖技术的不断进步，未来可能需要对馆藏进行再次数字化。这类系统目前主要采用关系型数据库加文件系统的方式实现，更为先进的采用了NoSQL数据库的大数据方式，基于云服务架构。而现在应该采用云原生架构加数据中台方式，这样就能够提供底层藏品管理系统与上层知识库系统之间的桥梁，同时提供大量的API供知识库系统和服务应用展示前台调用[8]，这些API可以以标准方式发布于互联网，从而实现数字人文平台的全网域互操作。鉴于将来的数字人文研究都是基于数据的研究，有了这样的典藏系统，就可以解决绝大多数人文学者在研究、教学中的需要。

（3）知识库系统

目前似乎还没有一个恰当的术语来描述这样一种系统，最接近的词汇可能就是“语义知识库系统”，指应用了语义万维网技术对领域知识建立相互关联的知识体系，其知识单元是采用RDF形式（即主-谓-宾结构）描述的语义判断，而整个知识大厦是用知识本体语言OWL或OWL2组织起来，其背后的数学基础是一元谓词逻辑。数字人文平台的内容架构主要是由知识库系统提供的。其简化版就是采用关联数据的系统，更简化的一个版本是目前十分热门的利用“知识图谱”技术所支持的系统。这类系统在人工智能领域属于“符号学派”，与过去的专家系统同属一类，是将人的知识代码化形成规模之后，就具备了某种智能，现在与连结学派和概率学派有融合的趋势，作为人工标注或结构化的数据提供机器学习，从而具有自动获取知识的能力。数字人文平台需要大量的底层“知识库”来支撑各类数据的语义解释和关联关系，例如人名、地名、机构名、朝代、官职、谱系、辞典、词表等，几乎所有的工具书都可以提供知识关联，所有的知识生产都是建立在过去知识的基础上，与这些底层知识库都可以建立起逻辑联系，最强大的是这些知识库都是以某种方式在整个互联网上提供共享，所有基于知识库和标准描述方式的术语词表都可以达成全网域的语义互操作。

（4）服务应用展示系统

这是数字人文平台中绝大多数功能得以实现和展现的前台，也是各类工具与后台数据进行连结的中介，通常以桌面或移动应用，以及浏览器方式提供。所有的搜索、浏览、展示（包括可视化）、众包和用户空间功能都在这里以App方式提供，这样有助于达成大量的第三方应用App的开发和发布，形成一个开放强大的数字人文应用和工具的生态环境，从而很容易实现包弼德教授提出的为第三方数据、第三方工具、第三方图书馆定制免费公开的元数据访问和数据共享的规范和方案[3]。

4.2 平台的资源、功能和界面需求

资源完整、功能完备、界面友好，是任何一个信息系统的基本要求。当然，不同的系统对这三个方面的具体需求是不同的。一个好的数字人文平台至少要在这三个方面达到最低要求，同时要注意三者之间的平衡。

4.2.1 资源完整性

人文研究者在选定了研究问题之后，第一步就是要查询资料。很多机构在建设数据库或提供查询时只从自己已有的或订购的资源入手，这是不够的，还必须考虑到是否有办法提供外部资源的发现，甚至直接获取。要实现这一点，就要应用元数据收割方案，例如OAI-PMH，或开发标准或个性化的API，其中涉及很多考虑因素和资源互操作的具体技术，包括利用知识库系统实现不同系统间的语义互操作，如图8所示。

4.2.2 功能完备性

数字人文平台需要考虑很多与过去数据库检索系统不同的功能，过去的系统主要是以文献为主要内容，根据数据库字段（即高级检索）或全文检索能够定位到具体的文献，再通过链接解析或其他方式获得原文。而数字人文系统由于提供了以“数据”为基础的存储、关联和查询能力，因此多了与“知识库”相关的很多语义功能，而且在搜索、浏览、管理等方面都能够全面支持基于知识的操作（例如SPARQL查询、分面组配等），有时甚至还包含逻辑推理的功能实现（如启发式搜索），如图9所示。

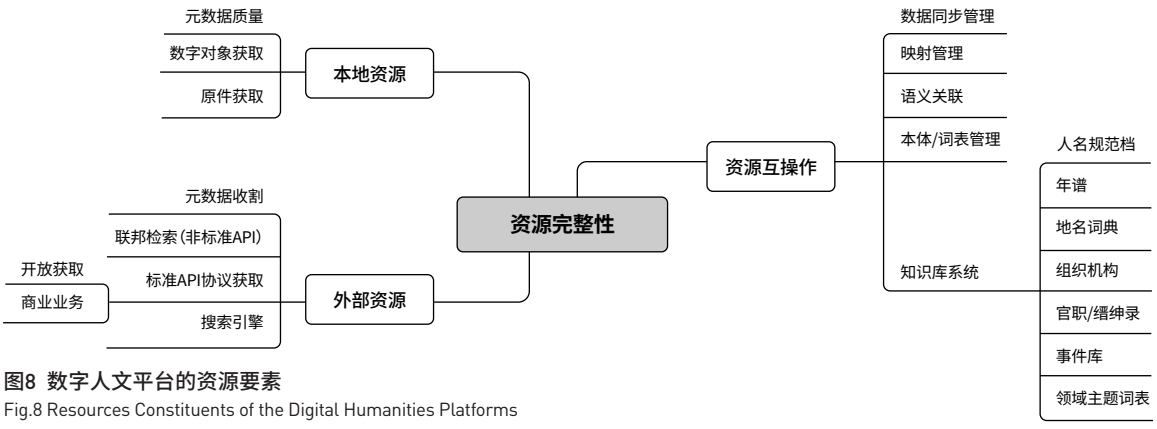

图8 数字人文平台的资源要素

Fig.8 Resources Constituents of the Digital Humanities Platforms

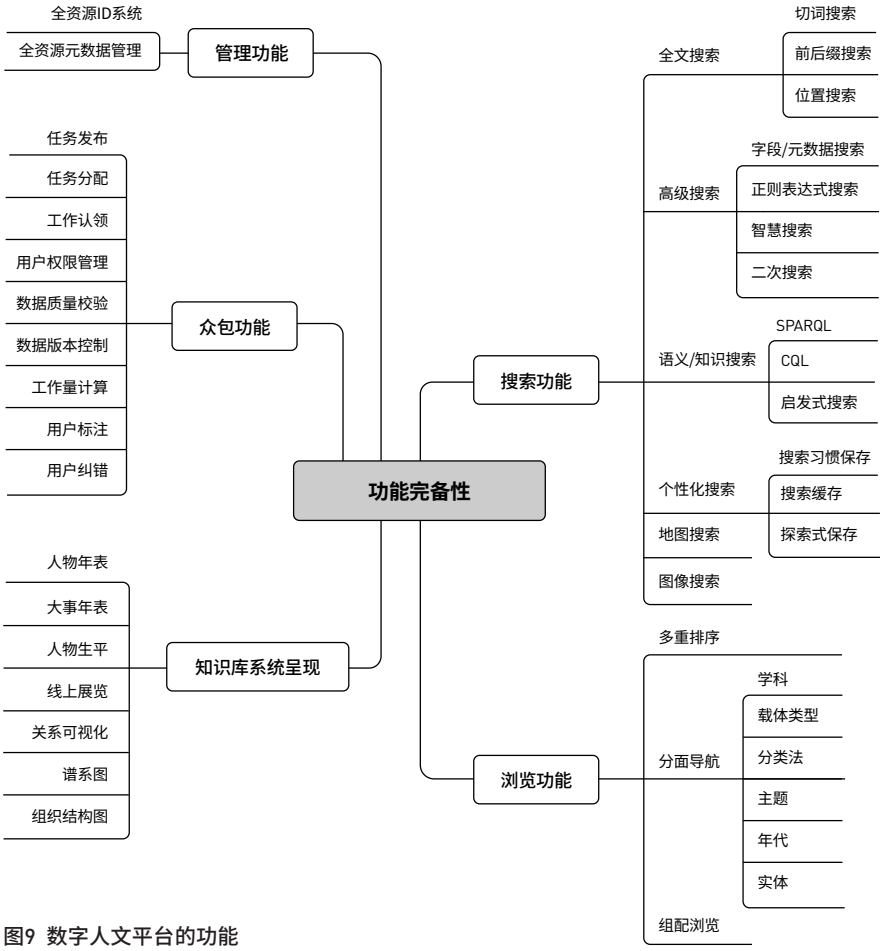

图9 数字人文平台的功能
Fig. 9 Comprehensive Functions of the Digital Humanities Platforms

数字人文平台还有一个特质是要利用众包让用户参与到系统的建设中来，这是当前几乎所有数字人文应用都采取的方式，因为仅仅通过图书馆或相关机构工作人员的工作是不可能实现海量高质量数据加工的。

4.2.3 用户友好性

当前的信息系统对用户友好性的要求越来越高，这也是对系统界面提出的要求，除了一般的方便友好、美观简洁之外，能否提供良好的个性化服务成为系统能否留住用户的重要特性，而且个性化服务大量采用了人工智能技术（见图10）。当然，由于个性化的前提是需要有用户注册登录等用户管理功能，且对用户的行为也会进行一定的收集，这涉及到用户隐私问题，平台在设计开发时必须考虑到隐私保护与个性化之间的平衡，很多研究工具的提供应该能同时支持本地脱机版和上传网络版两种不同的运行方式，当然两者在功能细节上可以有所不同。

4.3 平台的工具开发

利用大量的数字人文工具进行研究是数字人文区别于传统人文最重要的特点之一。工具是方法的重要组成，成熟的方法往往通过工具的开发而得以固化，并且负载了大量前人的经验总结。传统人文研究能够独立的工具不多，且资料的收集、阅读和加工处理往往是一体化、个人化的，工具很难独立于资料，有的甚至很难独立于研究团队。这也是为什么有许多人文社会科学学派往往是得益于独特的方法。

工具要求越丰富越好，但这里讨论的只是人文研究可能用到的具有一定通用性的工具，以及这些工具的常见功能，数字人文学者可以通过这些工具的组合，结合资源和研究过程，发展出自己独特的方法。这些工具可以有一定的独立性，但依附于平台能够更好地发挥作用，因此平台将致力于深入研究人文学者的需求，推出大量的标准规范，从而让大量第三方都能够开发自己的独特工具，甚至工具与资源或知识库的结

合体，从而有助于形成一个应用生态，以及工具App市场。

这里将工具划分为平台性工具（包括数据工具、IIIF、GIS、文献计量工具、阅读工具、社会关系工具）、文本工具、图像工具、知识图谱工具、机器学习工具和可视化工具等六大类（如图11所示）。上述分类的合理性需要进一步探讨，其中涉及的内容也远不是对各类工具的穷尽例举，仅仅作为一个讨论的基础，供具体进行工具开发和平台建设时参考。

（1）平台性工具

这里的平台是指网络上可以实现一定的功能、有特定输入输出的环境，平台性工具就是依附于平台的软件工具，或自身就是一个独立的工具，它通常需要结合一定的数据，与一些组件配合，并经过一定的流程才能达到目的。例如IIIF（国际图像互操作框架）就是一个功能强大的综合性图片平台，由多个服务器灵活组合而成，它本身就可以成为数字人文的服务平台，这里之所以作为一种工具，因为它提供了大量的关于图像的操作功能，如搜索、缩放、旋转、标注、比较等，可以应用于人文研究，非常强大。类似的还有数据处理平台工具、GIS平台工具、文献计量平台工具、社会网络分析工具以及阅读平台工具等。

（2）文本工具

文本是数字人文利用最多的资源类型，文本工具也是数字人文工具中种类最多、使用最频繁的工具，也是目前开发最成熟的工具类型。上图列出的是常用工具，一些综合性的文本工具，如“远读”“细读”则列在平台性工具类目下。

（3）图像工具

通常所有的图像扫描、处理软件都可以作为数字人文的图像工具，这里仅列出数字人文项目非常常用的工具类型，如图像特征提取工具、图像分类/聚类工具和基于图像的搜索工具等，图像平台IIIF已作为平台类工具列出。

（4）知识图谱工具

知识图谱是数字典藏向数字人文进化的关键技术之一，这里将关联数据、语义万维网技术都归入知识图谱。这类工具包括了实体提取、URI赋值、词表模式、本体构建等语义化工具，本体/词表管理、语义映射、RDF语义数据存储等语义管理工具以及SPARQL、启发式搜索、分面呈现等语义搜索、展示和利用工具等。

（5）机器学习工具

当前，数字人文的大量应用都用到了人工智能领域的机器学习技术。从OCR到实体提取，从神经网络到深度学习，无一不能应用于数字人文研究的各个过程。机器学习最大的特点是离不开数据，尤其是海量的数据，因此数字人文平台中的数据是其产生作用的前提条件，而由数据训练出来的机器学习模型又可以应用于更广泛的数据中，这是它的运作方式，也是它的价值所在。

（6）可视化工具

可视化是数字人文进行数据操控、展示和结果呈现必不可少的工具，也是数字人文区别于传统人文的重要特质。可视化虽然有很多工具，但现在基于互联网的工具已成为主流，正在成熟起来。它后台连接的数据可以是平台上已有的数据，或者挖掘出来的数据，

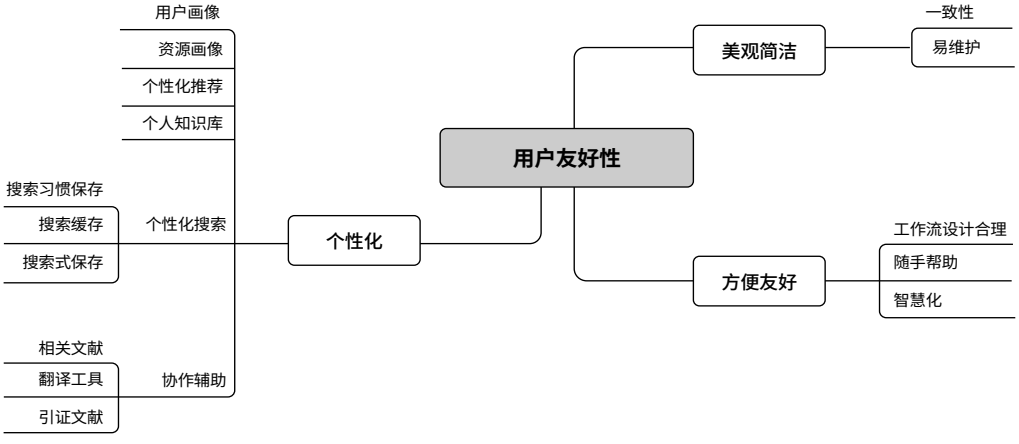

图10 数字人文平台的用户体验

Fig. 10 User Experience of the Digital Humanities Platforms

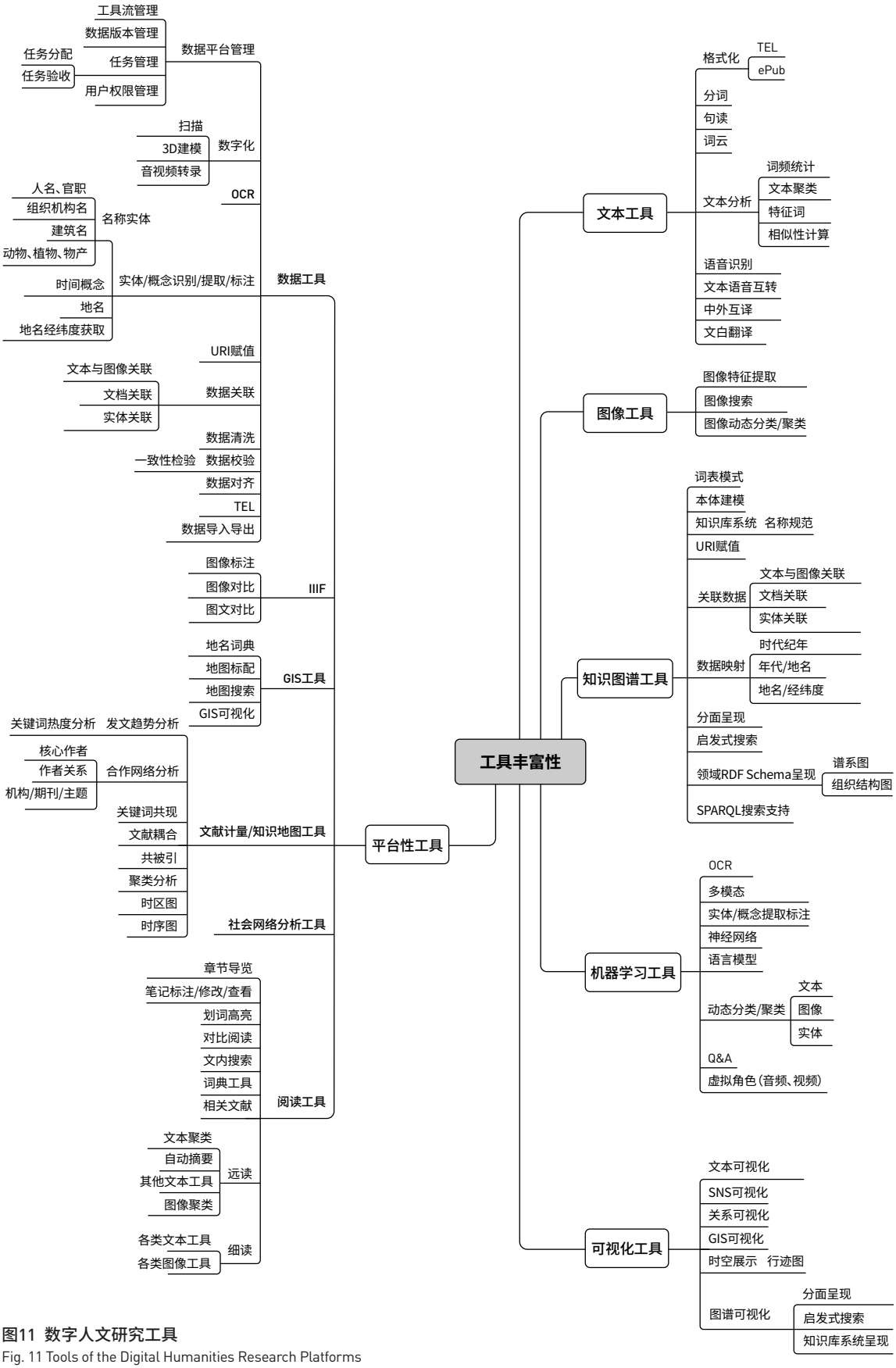

图11 数字人文研究工具

Fig. 11 Tools of the Digital Humanities Research Platforms

或者是用户上载的数据，是否支持多种应用方式取决于平台架构设计的灵活性。

5 案例：历史人文大数据平台

上海图书馆正在建设的历史人文大数据平台，就是应用上述理念和技术，依托自身资源，向全社会提供一个先进、开放、全面的数字人文服务平台。打造这个平台主要有三个目的：一是升级原有的数字图书馆系统；二是提供基于“知识”的数字人文服务；三是试验一些互联互通共建共享的新协议与新模式。其实就是作为对前述数字人文发展趋势进行应对的一种尝试。

实现这三个目的有两条现实可行的路径：其一，从现有的数字图书馆系统出发，也就是从目前上海图书馆馆藏特色资源出发，升级技术架构和内容架构：技术架构全面微服务化、容器化和平台化，支持外部资源与服务通过各种标准或非标准方式（推荐RESTful API）接入；内容架构进行“数据化”改造，支持“基于知识的服务”。其二，从数字人文研究者的角度出发，规划所有人文资源的整合方案，从提供资源到提供平台环境（包括工具），努力实现主要数字人文应用场景的“一站式”服务。

5.1 平台的建设规划

上海图书馆走上数字化道路已经有四分之一个世纪。从1996年位于上海淮海中路的“新馆”开馆，就开始古籍数字化项目，并且参与了中国最早的由国家图书馆牵头的“试验性数字图书馆计划”，成立专门部门，每年耗费巨资进行特色资源的数字化工作，从无间断。

仅仅数字化是不够的，提供知识服务是图书馆的根本宗旨。早期重视数字化，但对于数字典藏系统的建设并没有充分重视，因此数字资源的整合服务一直没有充分开展。到2016年，上海图书馆尝试以最具特色的馆藏家谱资源为案例，开始了以服务为导向的系统开发尝试，取得了不错的效果，迄今家谱系统一直是数字典藏中利用效果最好的资源之一。

为了建设具有知识关联的数字人文服务系统，底层知识库平台建设是必不可少的，这也是数字人文基础设施最困难的内容。近几年我们还陆续构建了人名规范、地名规范、地理名称规范、机构规范等规范知识库，可以支持目前列入计划的特色资源库的底层知识关联，并开始开发一些工具，提供众包、标注、分析、可视化等功能。

正是由于有了底层知识库的支持，上海图书馆的特色资源库才有可能做一个全面规划，将来各类数字人文系统可以在一个统一的平台上，我们称之为历史人文大数据平台。虽然这一平台尚未建成，但已经经过了初步尝试，证明了技术和工程上的可行性和可能性，且数据也有一定规模。目前，我们除家谱库外，正在开发的还有古籍库（包括精品善本库）、碑帖库、地方志库、手稿尺牍库、名人档案库（如盛宣怀档案、张佩纶档案等）、民国资源库（包括书刊报）等，这些文献如按照数字人文研究的要求，可以建立无数个基于各类学科或主题的知识库，可以汇总在一个平台上提供满足各类需求的统一服务，通过一定的开放链接协议，可以将全网域的各类资源连为一体，组成一个虚拟中文数字人文平台。

5.2 平台的应用场景

对于一个资源众多、用户复杂、目标多重的服务平台来说，“主页”概念是不适用的。历史人文大数据平台虽然设计了一个主入口，但它的作用只相当于“游客中心”甚至是“疏散中心”，主要起到宣传、导航、资源发现和用户培训的作用。任何一个简单的搜索，都可以返回所有资源库中（甚至外部联邦检索或搜索引擎）的命中内容，这样能够让随便逛逛的读者也有所收获，同时用户对自己感兴趣的主题可以通过哪些资源库获得有一个非常直观的认识，使带有目的的读者能够迅速找到属于自己的入口。

平台对所有的专题库（包括文献库、知识库和工具库三类）都有一个入口，其中大多数文献库都以元数据库加扫描图片方式提供，个别有全文，知识库和工具库都支持响应式H5接口，可嵌入各类App。

我们把平台用户分为四类：普通用户、专业用户、系统用户和机器用户，普通用户是无需用户认证即可来“随便逛逛”的用户，平台会有很多线上展览、人文讲座、推广活动、技能培训等内容发布。专业用户是平台服务的主体，通常是经过注册的研究人员或大学师生，也可能是相关机构中的个人用户（登录为单位用户或以IP控制方式提供权限管理），这类用户除非使用主页中的搜索框进行资源发现（搜索框在各相关页

表5 历史人文大数据平台提供的服务

Table 5 Services Provided by the Digital Humanities Platform of Shanghai Library

场景类别 \ 用户类别		普通用户（非注册）	专业用户（注册）	系统用户
Epic 叙事	Story 故事	Feature 功能		
搜索	简单搜索 / 全文搜索	精确或模糊匹配可选		
		支持提问自动分词		
		支持前后缀搜索		
		智慧搜索（框式搜索）：I feel lucky （用各种技术猜测用户需求的检索）		
	高级搜索	分字段限定	支持专业检索，即逻辑表达式在一个搜索框内实现高级搜索	
			支持字段按数据类型（如年代）限定，支持范围限定	
			位置限定搜索	支持正则表达式搜索
				统计式搜索
	知识搜索		基于实体名称 / 概念的搜索，如人、地、机构、事件、物体等的名称	
			启发式搜索	
				SPARQL/CQL 搜索（部分）
	图像搜索		基于输入图像的匹配（只支持部分图像库）	
	二次搜索	在任意结果集中再进行限定搜索（支持逻辑式限定）		
浏览	排序	检索结果支持多字段排序（分主次），支持拼音 / 笔画 / 时间等排序方式，支持正序 / 逆序		
	分面组配浏览导航	按结果集分面情况选择		
	知识导航		按知识本体呈现结果	
	地图浏览	检索结果具有地理或时代属性的，可以在地图上进行时空呈现		
	对比		不同检索结果集比较呈现，可多窗口（参数可选）对象比较	
	结果可视化	结果集基本属性的可视化	尽可能实现多种可视化	可视化参数可调
下载	元数据下载至个人空间		下载至个人空间，再支持本地下载	可订制 Schema 数据格式，与高清对象数据一起打包下载
	对象数据下载		需经授权才能下载	
	参考文献格式下载		可选择一定的参考文献格式，批量下载	
阅读	结构导览	目录章节导览、跳转、超链接		
	文内搜索	搜索词高亮并可导航（下一个）		
	词典工具	可外挂多语种 / 专业词典		
	文本朗读	自动机器语音朗读 / 生僻字词朗读（随词典工具）		
	书签 / 高亮	非专业注册用户存于本地，专业用户上载个人空间，并提供社会化（分享）选项		
	批注 / 修改		以 W3C 标注标准方式（RDF）实现，可分享	
	对比		选择不同对象多窗口打开比较和分别批注	
	文献推荐		阅读及写作过程中不断推荐相关文献，推荐模型相关参数 / 阈值可选	
高级阅读 / 阅读工具	细读		文本标注：根据词表（内建 / 挂接词表或用户导入词表）标注	
			文本标注：句读、分词、词性标注等	
			实体识别并标注	
			文白翻译	
		文本互译（外挂翻译工具）		
			文本格式化，格式标注 / 转换（如 TEI、ePub 等，参见数据加工工具）	
	遥（远）读	特征词 / 词云生成		
			自动摘要（模型参数可调）	
			文本 / 图像相似性计算、聚类分析	
			风格分析	
			情感分析	
	共读		支持用户书签 / 标注 / 批注等信息的共享和挖掘	

续表5

场景类别 \ 用户类别		普通用户（非注册）	专业用户（注册）	系统用户
Epic 叙事	Story 故事	Feature 功能		
众包平台	任务管理			任务发布、分配
	用户管理			注册及权限管理、资金监管
	工作认领			工作量计算、统计反馈
	质量管理			质量校验、纠错、版本控制
可视化	文本可视化	各种词云	各种高阶词云、关系图、热力图	参数可调
	关系可视化		各类 d3/eCharts 图示	参数可调
	GIS 时空	行迹图	时空叙事 / 现地研究	
	图谱可视化		各类知识本体可视化，包括人物生平 / 大事年表 / 互动展览 / 组织机构图示等	

面也都会出现），一般无须从主入口进入，只要浏览器保留了登录Cookie，域名会直接将其定位到他自己的个性化页面，该页面已经将其经常使用或可能用到的专业资源入口与各类服务功能集成在一起了，每个用户的专业入口都是个性化的，与“我的空间”捆绑，用户如果不满意，也可以在“我的空间”中修改参数设定。系统用户是那种参与数据加工或项目研发的用户，这是“平台性”的具体体现，作为平台，不是一个私有的封闭系统，而应该有一定的开放性，属于整个社区，允许大家参与共建、分享成果，因此必然有一类用户通过贡献内容、参与数据加工或功能开发而具有更多的权限。最后一类“机器用户”就是指通过API或其他接口直接消费数据的计算机程序，这样能将平台与互联网上其他应用连为一体，使“一站式”服务成为可能。

平台提供的所有服务可以分为“场景”“故事”和“功能”三个层次，分别对应三类用户，提供不同的功能组合，详细如表5所示。“场景”可粗略地对应于前述数字人文研究的“行为”，例如搜索、浏览、下载、阅读等，故事是组成场景的若干种应用，而“功能”是平台提供的最小单位的模块，通常对应于目前云原生应用架构中的“微服务”。这里的服务基本都还是用户直接可用的部分内容，后台其实还有大量的微服务，由于与平台用户并无直接关系，这里就不详述了。

6 结语

数字人文平台建设的愿景是让人文研究不再困难。从雅典学园到文艺复兴，从鲁国杏坛到康梁变法，两千年来人文学者的创造性思考从来都是依靠个体的博览群书与博闻强记，依靠师徒私授或学派论战，思想的诞生、学说的完善，以及对社会实践的影响主要依靠的是个人的能力，人文知识的产生、发展和传播的整个过程是偶然、不清晰和不确定的，每位学者都要从最原始的篇章学起，遍历所有典籍并考察整个源流，穷极一生只能成为专家而无法成就大家，而数字人文正在第一次给人文研究带来革命。针对人文研究的完整过程，数字人文已能够分而治之：首先，使资料查询和获取不再困难，然后使知识存储、传播和利用不再困难；其次，让分析、比较，形成观点不再困难；最后，使结果展示、交流和争鸣不再困难。人文学者不再是单打独斗而是集团作战，无须管中窥豹而是直接综揽全局尽情把握，人文研究的规律与方法将得到更好的揭示，人文成果的发表形式将不限于书刊，人文学说的比较与评价将更方便地在实践中得到检验和反馈，为人文研究提供的服务能力将更快地得到迭代和提高。照此发展下去，那么问题来了：如果数字人文充分采用了人工智能技术，推向极致，可能机器也能自动进行人文研究。此时的人文，还是人文吗？其实数字人文的终极意义还是在于以科技强化人文，而不是将人文变成被动机械的对象，进行去价值化和无意义化。最终的意义呈现，其主体是人类自身。当所有的人文都是数字人文时，“数字”与“人文”才能够真正合为一体，那时“数字”的工具性特征便不再重要，人文研究此时便能回归本源，真正彰显人类的价值和生命的意义。这也是我们要用尽所有先进技术，推进数字人文平台的开发与建设的根本原因所在。

作者贡献说明

刘圣婴，王丽华：提出研究思路，论文撰写与修改；

刘炜：论文拟题，修改与定稿；

刘倩倩：收集资料，撰写论文。

参考文献

[1] 尤西林 . 阐释并守护世界意义的人——人文知识分子的起源及其使命 [M]. 上海：华东师范大学出版社，2017.(You Xilin. The Person Who Interprets and Protects the Meaning of the World: The Origin and Mission of the Humanistic Intellectual[M]. Shanghai: East China Normal University Press, 2017.)

[2] Hey T,Tansley S, Tolle K. 第四范式数据密集型科学发现 [M]. 潘教峰，张晓林等译 . 北京：科学出版社，2012.(Tony Hey,Stewart Tansley，Kristin Tolle. The Fourth Paradigm: Data-intensive Scientific Discovery [M]. Translated by Pan Jiaofeng, Zhang Xiaolin, et al. Beijing: Science Press, 2012.)

[3] 包弼德，夏翠娟，王宏甦 . 数字人文与中国研究的网络基础设施建设 [J]. 图书馆杂志，2018,37(11)：18-25.(Peter K. Bol, Xia Cuijuan, Wang Hongsu. The Digital Humanities and a Cyberinfrastructure for China Studies[J].Library Journal,2018,37(11)：18-25.)

[4] 王晓光 . “新技术”和“文科”不能简单相加 [N]. 光明日报，2020-12-29(14) .(Wang Xiaoguang."New Technology" and "Humanities" Cannot Be Simply Added Together [N]. Guangming Daily, 2020-12-29(14) .)

[5] Unsworth J. Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? [J/OL]. Humanities Computer: Formal Methods, Experimental Practice(2000-05-13) [2022-02-12].https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

[6] Unsworth J. What Is Humanities Computing, and What Is Not? [J/OL].Jahrbuch für Computerphilologie(2006-12-07) [2022-02-12].https: //www.ideals.illinois.edu/handle/2142/157.

[7] 王宏德，庄惠茹，魏家惠 . 出席“中国历史研究的网络基础设施国际研讨会”：看数位人文研究之重要发展 [J]. 国家图书馆馆讯，2018(2)：36-39.(Wang Hongde, Zhuang Huiru, Wei Jiahui. Attending the "International Conference on the Cyberinfrastructure for Historical China Studies": Retrospect on the Important Development of Digital Humanities Research[J]. National Central Library News Bulletin, 2018(2)：36-39.)

[8] 鲁丹，李欣，陈金传 . 基于 API 技术的数字人文基础设施的构建 [J]. 图书馆学研究，2019(13)：42-46.(Lu Dan, Li Xin, Chen Jinzhuan. Construction of Digital Humanities Infrastructure Based on API Technology[J]. Research on Library Science, 2019(13)：42-46.)

附录
Appendix

附表1 中文数字人文代表性数据库和资源网站

Appendix 1 Representative Digital Humanities Databases and Sites for China Studies

名称	创建机构或个人	类型	特点
上海图书馆古籍联合目录及循证平台	上海图书馆	文本、图片	收录有1,400余家机构的古籍馆藏目录，并带有便捷的数字分析和可视化工具
台湾汉学研究中心资源库	台湾汉学研究中心	文本	搜集有散佚海外的中国历代典籍文献影照本
Chinese Text Project	德龙 Donald Stugeon	文本	结合 OCR，专门设计的众包系统和 API，同时提供转录系统，可用于转录和使用前现代中文文字资料
10000 rooms(Tina Lu)广厦千万间项目	耶鲁大学	文本、图片	用于前现代文本研究的在线协作平台
通用型古籍数字人文研究平台	台湾政治大学图书馆与台湾汉学研究中心	文本、图片	以台湾汉学研究中心特藏明人文集为文本，政大社资中心开发数字分析工具
中华书局籍合网	古联 (北京)数字传媒科技有限公司	文本、图片	国内首款古籍整理与数字化综合服务平台。上线资源3,396种，累计约15亿字。
Docusky	台湾大学数位人文研究中心等规划	文本	人文学者进行个人化材料整理与分析的网络平台。文本与工具分离、工具基于网页
MARKUS	Ho, Hou leong Brent, Hilde De Weerdt，欧洲研究理事会资助	文本、图片	强大的中文或韩文文本标记工具，能便捷导出到其他数字平台进行深入研究
法鼓文理学院数位人文项目	法鼓文理学院图书资讯馆	文本、图片	佛典文献之数字整合研究平台，主要功能为文献内容阅读、深度资料搜寻、数字量化分析
“中研院”数位人文平台	台湾“中央研究院”数位文化中心	文本、图片	一个完整的人文科学研究环境，让研究者可以搜集资料、比对文本、统计分析、可视化呈现
国家图书馆出版社数据库群	国家图书馆出版社	文本、图片	影印古籍、民国时期文献等各种稀见历史文献的数字出版

续附表1

名称	创建机构或个人	类型	特点
CNKI（中国知网）	同方知网（北京）技术有限公司	文本	最大的连续动态更新的中国学术文献数据库
CBDB	哈佛大学、台湾“中研院”、北京大学	文本	系统性地收入中国历史上所有重要的传记资料
Kitamoto 实验室	Kitamoto Asanobu	图片	专注于图像数据，研究图像处理、检索、分析的各种方法
搜韵网	陈逸云	文本	研发出深受诗词爱好者喜爱的工具如韵典、诗词校验及诗词检索
台湾大学图书馆数位人文项目	台湾大学图书馆	文本、图片	全文影像资料库，包括淡新档案、台湾古碑拓本、伊能嘉矩手稿、田代安定手稿、歌仔册、狄宝赛文库等
CADAL	浙江大学和中国科学院研究生院等14个单位	文本	整合国内并有重点地引进国际学术机构的各类信息资源与服务。集中发布260万余册（件）数字资源
中文在线	中文在线数字出版集团	文本、图片	拥有数字内容资源超过400万种，签约版权机构600余家，签约知名作家、畅销书作者2,000余位
“中研院”汉籍电子文献	台湾“中央研究院”历史语言研究所	文本、图片	内容包括经、史、子、集四部，754,200,000字，几乎涵括了所有重要的典籍
香港中文大学图书馆数据库	香港中文大学图书馆	文本、图片	香港中文大学图书馆特藏包括善本、手稿和档案文献等，已推行“香港中文大学数码典藏”计划
陈澄波画作与文书	台湾“中央研究院”数位文化中心	文本、图片	台湾画家陈澄波主题网站，利用数字工具，营造立体的时空脉络，视觉化呈现他的作品及生平
EastView	创始人 Kent D. Lee 和 Dima Frangulov	文本、图片	全球信息提供商，提供包括亚洲的报纸期刊、电子书、图片库等资源
北京大学开放研究数据平台	北京大学	文本	完整的数据提交、管理和发布功能；灵活的访问控制、请求与审核机制；规范的版权保护、实名学术社区
台湾学术经典	联合百科电子出版有限公司	文本	汇集全台核心学刊，亦收录独家史料、档案，及具研究价值、可雅俗共赏的杂志经典和优质期刊
瀚唐典藏	北京时代瀚唐科技有限公司	文本、图片	采用超大字元集加工的古籍资料库。涵盖甲骨文、金文、简帛文、印章、石刻等。提供拓片、释文等内容
上海图书馆近代报纸 / 期刊数据库	上海图书馆	文本、图片	目前近代期刊收录种类最多的全文数据库
中国地方历史文献数据库	上海交通大学出版社	文本	内容来自上海交通大学地方文献中心专家的田野调查或市场收购，总量约为35万件，150万页
中华经典古籍库籍合网	中华书局下属古联数字传媒科技有限公司	文本、图片	古籍整理与数字化综合服务平台。《中华经典古籍库》资源已达到近1,900余种，总计约10亿字
书同文	北京书同文数字化技术有限公司	文本、图片	提供数字化中文典籍，数据库基于云服务，并开发了一系列数字人文软件和工具
华艺数位	台湾华艺数位股份有限公司	文本	以艺术资料库为主轴，跨足学术领域，整合台湾、中国大陆两岸学术资源之检索平台
国学大师	郦勇	文本、图片	集典故、古籍、诗词、字词、成语、书法、人物地名等历史资料的大百科
Utah Genealogical Society（美国犹他家谱学会）	犹他州家谱协会	家谱	当今全球最大、最完整的华人族谱数据库
浙江大学学术地图发布平台	浙江大学社会科学研究院与哈佛大学共建	GIS 数据	为用户提供发布、编辑、搜索、查看、定位及分享等功能，总量为300余幅地图、600图层、40余万条
中国历史官员量化数据库——清代	香港科技大学李中清、康文林教授团队	文本	超过400万条记录，以《缙绅录》为参照，从1760年到1912年的345,071名官员的详细名单与官职
唐宋文学编年地图	中南民族大学数字人文资源研究中心	GIS 数据	以历史地图为平台，呈现诗人一生的活动轨迹
莱顿大学数字人文中心数据库	莱顿大学数字人文中心	文本、图片	多语言语料库、手语语料库及分析工具
台湾“中研院”人社中心 GIS 专题中心数据库	台湾“中研院”地理资讯科学研究专题中心	GIS 数据	一系列 GIS 基础平台、资料库，特色的健康主题 GIS
台湾历史人文传记资料库	刘昭麟、Michael A. Fuller（傅君劢）等	文本	台湾版 CBDB，拥有便捷的可视化社会关系分析功能

续附表1

名称	创建机构或个人	类型	特点
Dictionary Databases	A. Charles Muller	文本	数字化的佛教字典、儒道字典，词汇量达104,000左右
明清妇女著作	麦吉尔大学图书馆 Digital Initiatives 团队	文本	独有的明清妇女作家著作角度，集合多所大学图书馆的馆藏
南京大学数字人文项目	南京大学高研院数字人文创研中心	文本、图片	包含多个专题数据库
LoGaRT 地方志研究工具	马克斯·普朗克科学史研究所（MPIWG）	文本、图片	带领研究者从阅读角度跳跃到鸟瞰视角，从大量已数字化的旧方志，进行宏观提问，探究地方性知识
复旦大学中国历史地理信息系统	葛剑雄、包弼德	GIS 数据	中国历史时期连续变化的基础地理信息库
台湾“中研院”史语所数位文化中心	数位典藏内容与技术专题中心	文本	成立于1928年，包含多个平台，总数逾128万、160个成果网站并提供便利的跨资料库检索
上海图书馆家谱数据库	上海图书馆	文本、图片	收藏中国家谱（原件）数量世界第一，支持地图检索，还能“在线修谱”和“上传家谱”
关西大学亚洲研究开放研究中心数据库	关西大学亚洲研究开放研究中心	文本、图片	以数字化长期保存关西大学图书馆所藏东亚文化研究资料为基础，推进数字人文项目

附表2 中国大陆以外地区主要数字人文数据库名录

Appendix 2 List of Major Digital Humanities Databases Outside China's Mainland

数据库名称	网址
东京大学东洋文化研究所汉籍善本全文影像资料库	http://t.cn/aoxNuq
东京大学东洋文化研究所双红堂文库全文影像资料库	http://t.cn/zTyT450
日本国会图书馆古典籍资料	http://t.cn/zOm2To6
日本学习院大学东洋文化研究所所藏古典籍	http://t.cn/8k3IHye
京都大学人文科学研究所在线版四库提要	http://t.cn/aNmUFV
京都大学人文科学研究所东方学数字（デジタル）图书馆	http://t.cn/aoHBzA
早稻田大学古典籍总合	http://t.cn/hGhxOv
京都大学电子图书馆贵重资料画像数据库	http://t.cn/zjyfHfq
中国哲学书电子化计划	http://t.cn/zRO0E0e
台湾“中研院”汉籍电子文献瀚典全文检索系统	http://t.cn/ao1wf9
台湾网上书上网—数位典藏与学习电子书库	http://t.cn/a9lJX9
台北故宫博物院故宫博物院和东吴大学数位古今图书集成	http://t.cn/a3PqU3
首都图书馆古籍插图库	http://t.cn/SIqnOw
《工部局董事会会议录》28册	http://t.cn/8kUQCOh
哥伦比亚大学东亚图书馆 Chinese Paper Gods	http://t.cn/aKkyaI
康奈尔大学华生中国收藏（Wason Pamphlet Collection）	http://t.cn/8k136fl
哥伦比亚大学东亚图书馆古巴华工调查录（6卷）	http://t.cn/8kJiZVX
台湾大学图书馆公开取用电子书（110074条）	http://t.cn/hbJgKn
辅仁大学历史学系等南京教区契约文书数位典藏计划	http://t.cn/8kyFO2y
赖永祥长老史料库	http://t.cn/8FKtJAH
剑桥大学图书馆馆藏中国丛书综合目录检索	http://t.cn/8kZMTQk

续附表2

数据库名称	网址
剑桥大学图书馆电子目录（北平图书馆善本书胶片）	http://t.cn/8F87bo3
剑桥大学图书馆电子目录（景印离藻堂四库全书荟要）	http://t.cn/8F87bou
金陵图书馆《民国丛书》目录索引（第一辑至第五辑）	http://t.cn/zOp8Nex
中国大陆各省地方志书目查询系统	http://t.cn/hcZpm1
东洋学文献类目检索［第7.3α版］	http://t.cn/zTX4c28
漢字データベースプロジェクト（汉字数据库工程）	http://t.cn/SiZBD7
日本全国汉籍检索系统	http://t.cn/aCMEfI
台湾大学台湾历史数位图书馆之研究工具集	http://t.cn/z85Tcsv
清代人口史研究资料库	http://t.cn/zWv04zn
清代档案人名权威资料查询	http://t.cn/zR6uYkd
台湾“中研院”国际电脑汉字及异体字知识库	http://t.cn/hgsrrH
台湾“中研院”语言学研究所等搜文解字	http://t.cn/8kx2foj
台湾“中研院”语言学研究所汉字知识本体	http://t.cn/8FA0ncX
缺字系统	http://t.cn/zThmsVi
“中国文化大学”中华百科全书1983年版	http://t.cn/akc9jE
“中国文化大学”中文大辞典（全10册）	http://t.cn/zO3QT0L
网络资料版华人基督教史人物辞典	http://t.cn/aO5ZEd
金陵图书馆建有十余个具有地方特色的数据库	http://t.cn/8ktw4VO
国际中国学杂志《通报》文章目录汉译	http://t.cn/aYRwYx
Modern Chinese Scientific Terminologies（近现代汉语学术用语研究）	http://t.cn/zRYoShP
海德堡大学汉学文典	http://t.cn/zRjMO6v
海德堡大学早期《申报》索引	http://t.cn/zRlkFch
中文和合本圣经查询系统	http://t.cn/h0XGb
中华福音神学院图书馆中文基督教期刊论文索引	http://t.cn/8Fy1Yix
澳门期刊论文索引	http://t.cn/8FLGer7
台湾 NCL 中国文化研究论文目录	http://t.cn/8kJECgC
台湾民间传说主题资料库	http://t.cn/8F24jwG
台湾人文及社会科学引文索引资料库	http://t.cn/8F240WC
台湾文史哲论文集篇目索引系统	http://t.cn/8kixsvj
台北学研究主题资料库	http://t.cn/8F240Wo
台湾大学法律学院台湾日治时期统计资料库	http://t.cn/zROOF7V
中华佛学研究所中国佛教寺庙志数位典藏	http://t.cn/zOmolCZ
中国佛教数字目录	http://ddz.ee/iIIRt

（收稿日期：2021-04-15）