

数字人文的跨界、融合与对话

——第九届上海国际图书馆论坛数字学术与人文研究分会场“快闪报告”综述

金家琴 夏翠娟

【摘要】“快闪报告”是第九届上海国际图书馆论坛(SILF2018)“数字学术与人文研究”分会场的—个特别环节,12位分别来自哈佛大学、伦敦学院大学、北京大学、中山大学、南京大学、南京理工大学、复旦大学、上海交通大学、华东师范大学和上海图书馆的报告人,围绕“技术变革范式”“数据塑造世界”和“需求决定导向”三个方面的议题,从“方法与工具”“基础设施建设”和“用户体验”等角度,分享各自在数字人文领域的研究和实践成果,为数字人文未来的发展提供了难得的参考和依据。

【关键词】数字人文;上海国际图书馆论坛;综述;人文研究

【作者简介】金家琴,女,上海图书馆(上海科学技术情报研究所)文献提供中心,副研究馆员,研究方向:数字人文、知识服务,作者贡献:文章主要内容整理撰写,E-mail:jqjin@libnet.sh.cn;夏翠娟,女,上海图书馆(上海科学技术情报研究所)系统网络中心数字人文项目主管,高级工程师,研究方向:知识组织、数字人文、开放数据、关联数据,作者贡献:设计文章框架、提纲和写作思路,撰写前言和结语,整体润色与审校(上海 200031)。

【原文出处】《图书馆杂志》(沪),2018.12.29~38

0 引言

“快闪报告”是第九届上海国际图书馆论坛(SILF2018)“数字学术与人文研究”分会场的—个环节,由上海市图书馆学会学术委员会数字人文专业委员会承办,在组织形式上较为特别,由12个涉及数字人文各个方面热点话题的短报告组成,每个报告限时5分钟,共1小时。12位报告人分别来自哈佛大学、伦敦学院大学、北京大学、中山大学、南京大学、南京理工大学、复旦大学、上海交通大学、华东师范大学和上海图书馆,专业涉及计算机、历史、地理和图书情报。正如该环节的主持人夏翠娟所言:“只有数字人文这样的跨学科研究领域,才能把这么多不同专业的学者聚集到一起,让视角跨越专业与学科的限制,开始对话和交流,并逐渐形成一种互相依赖、彼此依靠的共生关系。”

这12位报告人按照报告的内容被分成了三组,分别是技术、资源和用户,这也是数字人文实践的三

个重要支撑点。自数字人文兴起以来,有一部分理性的学者不断进行反思和批判,这也是学术发展必不可少的一部分。因而,这一环节从北京大学王军教授的技术探索开始,以南京大学王涛教授的理性反思结束,形成了一种技术和人文南北对话的局面。

每位报告人在短短的5分钟时间内,可以演示自己的项目并陈述自己的观点,如果时间有宽余,还可以向后面的报告人提一个问题或者对前面的报告人做一个点评。12位报告人合作完成了每人5分钟的“快闪报告”,围绕“技术变革范式”“数据塑造世界”和“需求决定导向”三个方面的议题,分享各自在数字人文领域的研究和实践成果。

1 技术变革范式

在这个部分,几位报告人分别介绍了可视化、GIS、数据关联在数字人文中的应用,以及数字人文智能平台的建设,阐释了技术的发展如何促进人文研究范式的变革和转型。

1.1 王军的《基于符号分析法的宋代政治网络可视化研究》

在《基于符号分析法的宋代政治网络可视化研究》报告中,北京大学信息管理系的王军教授介绍了团队近年来的部分研究成果,如唐代300年仕人的迁徙路线、宋到明几百年的儒家理学传承路线和禅宗法传承可视化平台。以使用中国历代人物资料库(CBDB, China Biographical Database Project)数据集为例,王军教授分享如何通过分析变量“政治对抗”(负关系-)与“政治奥援”(+)关系,以可视化的方法展现宋代政治网络^[1]。比如设定时间维度在公元960年—1279年间,通过数据清洗,共获得了1788位宋代历史人物及其构成的2882个政治关系,并生成了宋代政治网络全貌。

又比如根据宋代政治核心人物如“秦桧”“蔡京”“王安石”“朱熹”和“司马光”等在“无向/有向性”度中心度上较高的排名,进一步构建出宋代的相党政治。

基于上述分析,关于宋代政治网络得到了如下结论:“从平衡系数来看,王安石变法革新以来,北宋政治结局动荡不断陷入新旧(改革派与保守派)的党争林立的政治格局,各党争势力之间的对抗与合作性不断增强,其势愈演愈烈直到南宋初才稳定下来。虽然全宋时代不同时期的斗争强度整体上比较一致,但在政治对抗性的规模上南宋显著高于北宋,自安石变法起始北宋奥援合作局面占据主体,而章惇恢复新法后宋代政治对抗斗争局面成为主体,斗争规模扩大显著增速至1170年基本稳定。”^[1]

1.2 陈刚的《构建城市历史地理学研究的时空GIS基础框架》

近年来,随着“数字人文”研究领域及理念的不断深入,促进了中国历史地理学及信息化研究领域的重要进步。南京大学地理与海洋科学学院陈刚副教授的报告《构建城市历史地理学研究的时空GIS基础框架》,回顾了“数字人文”促进历史地理信息化研究的表现:(1)涌现出新型研究成果,产生了一批具有重要学术影响力的历史GIS基础数据库及信息系统;(2)产生出新的学科增长点,壮大了学科研究队伍;(3)革新了研究方法与研究理念。比如应用GIS技术,建设历史地理数据库、编制电子地图集和研制历

史地理信息系统,已成为历史地理学研究不可或缺的技术手段;在大数据时代,借助GIS、文本分析(Text Analysis)、社会网络分析(Social Network Analysis)等技术,历史地理学融入“数字人文”发展潮流,进一步促进跨学科合作。但同时,陈刚副教授也认为,当前历史地理信息化面临着诸多问题与挑战,数据库建设还存在一定不足,提出未来的发展之路是基于时空GIS基础框架理念研制的新一代HGIS(Historical GIS),如台湾“中研院”所研发的SinicaView,哈佛大学地理分析中心建立的数字化合作研究平台WorldMap。

时空GIS基础框架是多源、多媒体历史地理数据的软硬件集成环境(平台),包括数据采集、加工、分析、交换及Web服务所涉及的标准、技术、设施、机制等的总称。由基础历史地理数据库、历史地理数据目录与交换体系、历史地理信息公共服务体系和技术标准及运行保障体系组成,主要特征包括统一时空基准、统一时空数据模型、统一要素分类与编码、统一元数据与交换标准、统一Web信息服务、统一用户环境和统一工具集。

陈刚副教授还以南京大学数字人文与超媒体GIS实验室开发的“六朝建康历史地理信息化”项目为例,总结了在时空GIS基础框架支持下,六朝建康历史地理信息化将进一步整合历史文献、考古资料与野外考察等历史地理学研究方法与内容。统一时空基准统与数据标准、多源、多尺度、多媒体数据、历史地理数据仓库和历史地理信息综合应用平台。

1.3 陈涛的《关联数据应用服务平台(LDSP)研究探索》

上海图书馆博士后工作站的陈涛博士后的报告《关联数据应用服务平台(LDSP)研究探索》,分享了上海图书馆在关联数据服务方面的探索——“关联数据服务平台”(LDSP)。关联数据服务平台由四个服务模块组成:关联数据转换服务LDTS、搜索服务LDQS、发布服务LDPS和知识服务LDKS。除了转换服务外,LDSP的三大服务模块都可以独立提供服务,只要提供SPARQL Endpoint即可接入。

转换平台是独立的,可以离线完成,所以目前并没有嵌入LDSP。转换后的数据放入Triple Store数据库,并和其他在线的SPARQL Endpoint一起接入服

务平台。

检索平台可以检索很多的集成数据集,目前可以检索SinoPedia、CBDB、VIAF数据。LOC和人名规范库也将陆续接入。检索服务提供了关键词检索,主要有多源检索、多样浏览和多点关联等特色。除此之外,检索服务还支持接入节点的资源URI检索,目前已接入Getty、VIAF、LOC、Nobel、方志库、文献库以及上海图书馆的一些开放数据。

发布平台可以提供内容协商服务,主要有RDF/HTML、RDF/XML、JSON-LD、NT格式。很多数字人文系统已经使用了LDSP的发布服务,比如CBDB关联数据平台42万的人物数据、上海图书馆的人名规范库80多万的人物数据、上海图书馆的关联数据书目系140多万的书目数据、方志集成平台中的6万多本方志数据以及上海图书馆开放数据平台中的开放数据等。通过这些链接可以自动获得数据的多种格式,实现资源的内容协商。在需要提供元数据的地方加上这样的链接就可以实现(http://sinopedia.library.sh.cn/{co}/{context}/{resource_path})。

知识平台则完美体现了关联数据的优势,实现不同数据集数据的在线整合与知识发现。知识服务提供了不同数据集之间的知识图谱和知识发现。比如CBDB关联数据平台,就可以在线整合上海图书馆人名规范库、古籍系统和VIAF平台,从而获得更多的知识。使用知识服务非常方便,只要在页面中嵌入链接就可以轻松实现(http://sinopedia.library.sh.cn/static/Resources/graph/oad.html?{res}{resource_uri})。

1.4 熊泽泉的《华东师范大学数字人文智能平台建设》

对于以史料为研究对象的人文学者来说,如何从大量史料中挖掘信息是目前面临的一大问题。科研工作者需要对史料中的细节进行深层次挖掘,对资源和技术都有较高要求。拥有古籍资源的机构,如图书馆,由于缺乏技术支持,基本只是实现了古籍的数字化保存功能,为古籍文献的知识挖掘带来了一定的障碍。对于另外一些掌握OCR技术的公司来说,会利用机器学习来优化识别模型,但是缺乏训练样本,前期只能通过算法和人工结合获取数据,通过不断迭代,周期较长。另外,很多研究者无力购买向

汉王这样的大公司的OCR技术产品,以至于大量的资源散落保存在各个自建数据库里,无法真正发挥价值。

为了更好地发挥史料的价值,降低数字人文的研究成本,让研究者更加专注于史料内容的深入挖掘,打通数字人文研究领域的各个环节,让图书馆等机构都能参与进来,华东师范大学、上海图书馆和汉王数字科技正在共同研发数字人文智能平台DHAIP(参见图1)。据华东师范大学图书馆熊泽泉馆员的报告《华东师范大学数字人文智能平台建设》中介绍,DHAIP是针对史料的自动识别、元数据著录和全文索引的智能平台,目的是要满足资源方、技术方和研究者三方的使用需求,同时让普通大众也能从此受益。

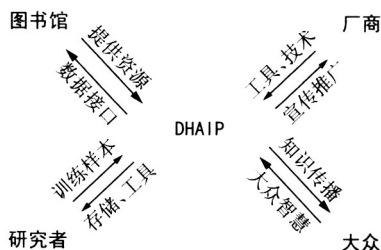


图1 DHAIP平台

DHAIP有4个功能模块:文字识别、众包模块、知识挖掘和API接口开放。

文字识别模块是DHAIP的核心模块,主要是基于汉王公司深度学习的OCR识别模型,完成图像预处理,版面分析,字符识别(文本切行分、单字符切分、单字符识别)和后处理(词库、自动分词、语言模型)。

众包模块是利用上海图书馆数字人文平台现有的众包模块,有任务发布平台、任务抄录平台和管理平台。

知识挖掘模块又分为社会网络分析、GIS空间分析、科学计量分析和文本分析。

API接口开放模块提供资源揭示、数据关联、数据共享于复用、和使用统计,达到三方资源共用。

2 数据塑造世界

2.1 王蕾的《地方文献的范畴及其数字人文视域》

中山大学图书馆王蕾博士的报告《地方文献的范畴及其数字人文视域》,从人文学者的角度,分享了数字人文之认识、地方文献范畴及演变、面向地方

历史文化研究与传承之地方文献数字人文路径。

数字人文的实践有三种模式:(1)以资源为导向,即基于特定文献资源的数字人文实践;(2)以研究为导向,即基于特定领域研究的数字人文实践;(3)各类分析工具开发,如文本挖掘、可视化、智能交互、学习等技术工具。其中,无论以资源为导向,还是以研究为导向的数字人文实践,均包含四大要素,即资源范畴的界定与收集;资源数据化、本体化、模型化建设;人文研究范式与方法的应用与革新;人文学者、资源整理者、技术开发者的合作模式。王蕾博士认为数字人文实践过程中有四个关键环节,包括:领域研究的认知与分析、资料收集、数据整理以及数据分析与呈现。传统的人文方法和范式在这些关键环节和建设过程中均具有深刻的影响力(参见图2)。

传统的地方文献范畴是指地方史料、地方人士著述、地方出版物,没有触及民间物质文化遗产和非物质文化遗产的资料。王蕾博士认为这种文献资源的缺失会影响数字人文开发的广度和深度。随着地方历史文化研究的日益深入,传统地方文献范畴应进一步扩展至地方特色资源、地方物质文化遗产及非物质文化遗产资料等资源。王蕾博士介绍了目前团队在民间历史文献数字化及数字人文建设中的关键思路。重点阐释了:(1)民间历史文献、田野考古资料的文本分析、基本概念及关系研究与本体建设,以及分类体系研究;(2)非物质文化遗产资料,以地方非遗信息资料、传承人抢救性记录资料,含口述资料及实践、教学与综合性记录视频资料、实物数字化资料等为对象的基本概念及关系研究;(3)研究资料的基

础数据建设研究。

王蕾博士提出未来面向地方历史文化研究与传承的地方文献数字人文建设实践可探索和推进建立集跨类型资源整合系统、区域社会历史文化呈现系统,以及研究与分析资源共建系统于一体的数字人文系统或平台。这个过程主要有两个难点:一是地方社会历史人物数据库。探索提取并积累地方历史文献资料中的人物资料,并进行人物信息与关系的元数据揭示,建立地方民间历史人物数据库,实现人物信息和社会关系的分析与呈现。二是地方虚拟GIS探索。通过地方历史文献考证与田野调查,采集地名GIS坐标信息,建构虚拟的地理关联数据;提取地方历史文献内容中的时空信息予以数据编码,为地方历史文献资源的时间、地点、图形检索的可视化技术发展提供理论探索。

2.2 赵思渊的《如何激活民间文书:元数据加工、再组织及分享》

2012年,上海交通大学开始地方文献的整理工作,建设中国地方历史文献数据库。目前大约有35万件来自南方地区的各省份民间文献的收藏,34万件全部扫描电子化,12万~13万件有metadata。

数据库的metadata是针对所有收藏的资料做的非特异性设计。但是对于历史学来说,有非常多的专题研究,如果每个专题研究都另外建一个数据库,显然是不可行的。元数据的价值如何再挖掘?元数据建设如何在史料基础信息与专题研究之间平衡?上海交通大学人文学院历史系赵思渊副教授的报告《如何激活民间文书:元数据加工、再组织及分享》,

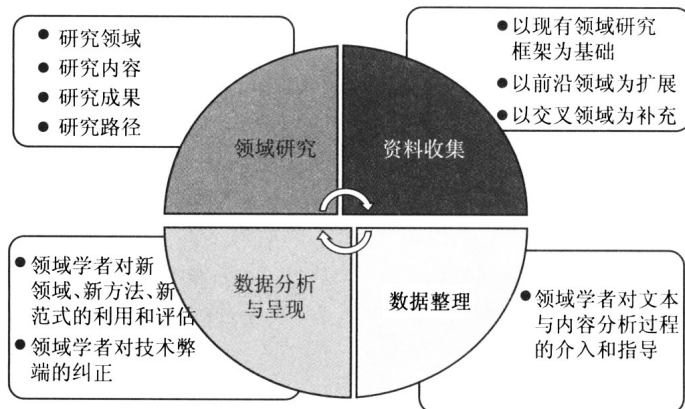


图2 传统的人文方法和范式在数字人文建设中的深刻影响力

以自主开发的上海交通大学馆藏民间文书编目数据库为样例,探讨如何花最小的成本,利用已有的metadata去做民间文书的专题研究工具。

民间文书编目数据库是以契约类的文书为研究对象的开发工具,提供基本的文献编目检索、提供契约类文献的交易关系分析、提供文书人物信息的社会网络分析。单件或者批次文献现在只能提取亲属关系。亲属关系描述:(h1)家族:事主栏位内的人名描述内包含亲属称谓用字。人名描述即每个人名后括号内的内容。亲属称谓列表:父;母;叔;伯;侄;姪;兄;弟;姻;甥;(h2)宗亲:(b2.1)人名第1字(姓)相同的情况下第2字或第3字相同。(b2.2)事主栏位内的人名描述内包含亲属称谓用字:房;族;亲。(b3)补充规则:一个人名同时符合(b1)与(b2.1)时,判定为(b1)家族。

2.3 李旻的《以人物和组织为中心——“历史数据”数据库的设计与实践》

李旻的报告《以人物和组织为中心——“历史数据”数据库的设计与实践》分享了自己如何通过数据库保存平时学习的知识点。李旻认为关系数据库因为利于使用、定义清晰、结构固定,仍是历史数据标准化的重要手段。在自建数据库时,首先是对历史记录的重新解构,梳理知识点。以人物和组织(人群)为中心的信息结构,将人物和人群分作两大类。数据以之前的理论为基础,特别针对政治史研究。2002年开始建设,截至2018年10月,收录历史上的

“人”11万条,“人群”3.1万多条,各类“关系”近50万条,并记录其属性,已初具规模。

2.4 杨敏的《上海图书馆近代报刊文献数字资源基础建设及服务》

上海图书馆杨敏副研究馆员的报告《上海图书馆近代报刊文献数字资源基础建设及服务》,从数字人文基础数据建设方面,介绍了上海图书馆中国近代报刊文献数字资源建设所做的努力。

报纸:近代报纸的元数据是从正文、广告和图像三个方面作著录。正文方面,对报纸的类别、栏目、新闻来源和新闻发生地作著录。以《字林西报》谋篇文章为例,它的类别是“新闻”,栏目是“News from the outports”,新闻来源为路透社,新闻发布地为广州。通过这些元数据标引,可以对题名、新闻发生地等信息进行文献查找。广告方面的著录主要包括广告标题、广告语、广告发布者、广告产品、广告类别和广告栏目。基于报纸的著录,上海图书馆形成了中国近代中英文报纸的全文数据库,包括《新闻报》《时报》等大报以及一批小报资源。

期刊:完成2万多种,1000多万篇数字化著录。基于此进行深度挖掘,数据库精选近300种文学类期刊,共计66万页,进行全文本识别。读者不仅可以从标题、作者、刊名等字段进行检索,而且可对该库进行篇内全文本查询,并注入了字典功能。

图片:上海图书馆对近代文献资源内海量的图片单独进行著录,建设了“图述百年——中国近代文

示例:“人”唐太宗李世民

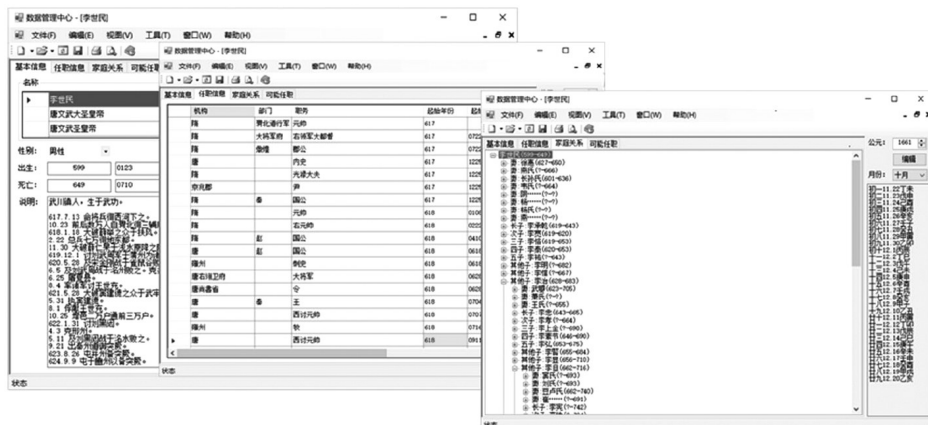


图3 “历史数据”数据库示例

数据库发布地址: <http://BiographicDB.fudan.edu.cn/BookStore>。

献图库”。海量的近代文献中,除了对文章中独立的图片的核心元数据进行著录,并实现了图片与文章、图片与期刊、图片与图片期刊的关联。目前已经推出了60万副图片。

3 需求决定导向

3.1 赵宇翔的《数字人文研究中的众包模式探索——任务匹配、用户参与及游戏化设计》

近几年崛起的数字人文领域对公众科学项目的实施有了更多迫切的现实需求,文化遗产机构(GLAMs)逐渐认识到将其馆藏数字化传播和利用的必要性。南京理工大学经济管理学院赵宇翔教授在《数字人文研究中的众包模式探索——任务匹配、用户参与及游戏化设计》报告中,以“基于科研众包模式的公众科学项目”为例,探索了如何建设一个众包模式的平台。

“任务匹配”阶段主要任务是从价值密度角度划分了三种不同类型数字人文类公众科学项目:质变式涌现型公众科学项目、量变式涌现型公众科学项目和非涌现型公众科学项目,对这三种类型的特征做分析。

赵宇翔教授分析了数字人文类公众科学项目众包的各个组成要素:发包方是以科学家、教育培训人员、技术评估人员等组成的科研团队为主;接包方是非职业的科学爱好者和普通志愿者;平台则是第三方综合型平台,自设专项型平台,并做了匹配对比分析。

“用户参与”是指冷启动阶段用户的广泛参与和持续启动阶段持续参与。赵宇翔教授总结了“用户参与”的三种驱动因素。一是平台驱动因素:感知有用性、感知易用性、平台社交性、平台专业化、游戏化元素、平台可视化。二是任务驱动因素:任务易操作性、任务自主性、任务趣味性、任务情境性、任务反馈性。三是志愿者驱动因素:好奇心和兴趣、乐趣、科学贡献、归属感、虚拟奖励、自我效能、感知学习、声誉。并结合上海图书馆盛宣怀档案抄录项目,从机构组织、平台设计和任务设计的角度对两个阶段提出了相应的激励对策。

“游戏化设计”主要是“在非游戏情境下使用游戏设计元素”,将游戏化元素与众包平台的功能模块

对应起来,实现克服冷启动,增加用户黏性和竞争力的目标。

3.2 王宏魁的《跨学科合作中的人文学者——以CBDB 2018 实践为例》

“数字人文”为人文学者提供了新的研究范式和视角。“数字人文”中人文学者的角色又该如何定位呢?哈佛大学计量社会科学中心王宏魁研究员的报告《跨学科合作中的人文学者——以CBDB 2018 实践为例》,以“中国历代人物资料库(CBDB, China Biographical Database Project)”项目为例,探讨了在人文数据库、数据集的建设和使用过程中,跨学科合作的人文学者主要承担4种角色和任务:一是提问者,检查者。人文学者主要负责提出问题:自动标点;提供语料:对文本进行朝代标注;评判结果:对测试集的结果进行检查和反馈以确定如何平衡Precision和Recall。二是主导者。此项目中,人文学者利用数字技术的工具和软件,撰写正则表达式、读python代码、算法设计讨论和根据史料提出新算法。数字人文如何解决。三是挑战者。人文学者既是旁观者又是批评者。四是转换者。人文学者需要将人文研究需求转化为信息技术专家可理解的计算机语言。基于人文学者的角色和任务,王宏魁研究员提出了对人文学者进行数字人文训练,主要内容包括方法论、技术和工具的使用和如何合作。为人文学者引入新的技术、方法和合作者,培养出既精通多种语言又专业的数字人文学者。

3.3 付雅明的《借鉴认知绘图方法探究图书馆用户体验》

付雅明博士的报告《借鉴认知绘图方法探究图书馆用户体验》,分享了借鉴认知绘图方法探究图书馆用户体验的实践。认知绘图源于地理和心理学研究,也被认为是民族志研究中的典型技术。它已被用于获取用户使用或思考特定资源或场所的方式的可视化表示。在社会科学领域,它以更一般的方式用于描绘人们如何理解世界。“绘图”:形成外部环境认知的过程和这种认知的表征。

认知绘图方法已经成功申请到了研究项目。如在伊利诺伊州大学图书馆的民族志研究:ERIAL项目(<http://www.erialproject.org/project-details/methodology>)。

3.4 王涛的《数字人文也应少谈些反思,多研究些问题》

最后,南京大学历史学院王涛教授,从人文学者的角度,提出《数字人文也应少谈些反思,多研究些问题》。虽然“数字人文”概念自2009年逐渐广为中国学界接受,许多人文学者特别是历史学家,至今对数字人文还是抱有怀疑的态度。王涛教授认为近段时间历史专业的杂志在发数字人文和大数据类的文章,大多是用抽象的陈念在讨论反思的问题,真正意义上用数字人文的方法和理念来解决问题的研究太少。很多从历史学角度提出反思的学者,并没有参与过数字人文的项目。旁观者的体验和实践者的体验是截然不同的,他鼓励更多的人文学者参与到数字人文的实践中来。

4 综述

12位报告人、12个短报告、5分钟报告时间,短小精悍的“快闪报告”成为此次SILF大会难忘的一小时。各位报告人围绕“方法与工具”“基础设施建设”“用户体验”和“反思”4个方面的主题,为数字人文未来的理论研究和实践提供了难得的参考和依据。

4.1 方法和工具

人文领域研究中,数字人文技术展现的可行性与高效性是近年来广泛讨论的热点主题。本次报告对数字人文技术在人文历史研究领域、历史地理信息研究领域的相关应用进行了探讨。

在《基于符号分析法的宋代政治网络可视化研究》报告中,北京大学信息管理系的王军教授用数字人文的相关方法对有关宋代政治历史问题进行了全新的解读。报告介绍了以宋代政治为例,从数字人文视角出发,借助符号分析方法对哈佛大学“中国历代人物资料库”(CBDB)进行实证探索与可视化分析。结合已有的史学问题和相关观点,王军教授分享通过分析变量“政治对抗”(负关系-)与“政治奥援”(正关系+),用可视化的方法展现了宋代政治整体网络分布特征、核心人物的地位与结构拓扑以及不同时期宋代政治网络的时序政治关系演化模式三个层次,为研究宋代党争政治格局提供了一种新的思考方式^[1]。

在大数据时代,借助GIS、文本分析(Text Analysis)、

社会网络分析(Social Network Analysis)等技术,历史地理学融入“数字人文”的跨学科合作,虽然促进了历史地理信息化的研究,但是仍面临着诸多问题与挑战。南京大学地理与海洋科学学院陈刚副教授认为,未来的发展之路是基于时空GIS基础框架理念研制的新一代HGIS(Historical GIS),如台湾“中研院”所研发的SinicaView,哈佛大学地理分析中心建立的数字化合作研究平台WorldMap。南京大学数字人文与超媒体GIS实验室开发的“六朝建康历史地理信息化”平台,就是在时空GIS基础框架支持下,建立的统一时空基准与数据标准、多源、多尺度、多媒体数据、历史地理数据仓库和历史地理信息综合应用平台。

关联数据在图书馆领域具有广泛的应用前景,通过采用关联数据技术,图书馆有机会在未来语义网建设中发挥主导性作用。上海图书馆的陈涛博士后在报告中介绍了上海图书馆构建的关联数据应用服务平台(LDSP)及其四个服务模块:关联数据转换服务LDTS、搜索服务LDQS、发布服务LDPS和知识服务LDKS。该平台能够很好地提供基于文献知识内容的揭示、导航和检索,通过开放数据重用和与外部数据的互联,丰富了数据的关联性,实现不同数据集数据的在线整合与知识发现,为基于互联网的数据服务提供了一种基础设施^[2]。

利用数字技术从史料中挖掘深层次的细节信息,对资源和技术都有较高的要求。拥有古籍资源的机构缺乏挖掘古籍文献知识的技术,拥有OCR技术的公司受限于缺乏训练样本,以史料为研究对象的人文学者又无力购买OCR技术产品,以至于大量的资源散落保存在各个自建数据库里,无法真正发挥价值。华东师范大学图书馆熊泽泉馆员介绍了华东师范大学、上海图书馆和汉王数字科技正在共同研发的数字人文智能平台DHAIP。这个平台是针对史料的自动识别、元数据著录和全文标引的智能平台,力图打通数字人文研究领域的各个环节,达到三方资源共用,降低数字人文的成本,满足资源方、技术方和研究者三方的使用需求,促进史学研究。

4.2 数字人文数据基础设施建设

民间历史文献资源的收藏、利用与数据库开发已成为当今学术型图书馆的工作重点。数字人文的

兴起为这一工作提供了研究思路和实现方法。中山大学图书馆和上海交通大学图书馆作为民间文献资源收藏的代表,利用已有的优势资源和研究传统,在展开地方文献数字人文路径的实践方面,已取得了初步的成果与进展。

中山大学图书馆王蕾博士的报告《地方文献的范畴及其数字人文视域》,从人文学者的角度,对数字人文的认识、地方文献范畴及演变、面向地方历史文化研究与传承的地方文献数字人文路径进行了梳理和回顾。中山大学图书馆团队在民间历史文献数字化及数字人文的关键思路是分类体系与基础数据建设。分类体系研究是指民间历史文献、田野考古资料的文本分析、基本概念及关系研究与本体建设和非物质文化遗产资料的基本概念及关系。基础数据建设主要是指研究资料的基础数据建设。未来他们将会建立跨类型资源整合系统,同时也是区域社会历史文化呈现系统,并在此基础上实现研究与分析资源共建系统^[3]。

上海交通大学开始地方文献的整理工作始于2012年建设中国地方历史文献数据库。目前大约有35万件来自南方地区的各省份民间文献的收藏,34万件全部扫描电子化,12万~13万件有元数据。对于历史学来说,面对如此多的专题研究,如何再挖掘元数据的价值,元数据建设如何在史料基础信息与专题研究之间平衡?上海交通大学人文学院历史系赵思渊副教授以自主开发的上海交通大学馆藏民间文书编目数据库为样例,探讨如何花最小的成本,利用已有的元数据去做民间文书的专题研究工具。民间文书编目数据库是以契约类的文书为研究对象的开发工具,提供基本的文献编目检索、提供契约类文献的交易关系分析、提供文书人物信息的社会网络分析^[4]。

高手在民间,复旦大学计算机科学技术学院李旻讲师是历史学爱好者。2002年他开始设计和建设“以人物和组织为中心——‘历史数据’数据库”,保存平时学习的知识点。在自建数据库时,首先是对历史记录的重构,梳理知识点。以人物和组织(人群)为中心的信息结构,将人物和人群分作两大类。截至2018年10月,收录历史上的“人”11万条,

“人群”3.1万多条,各类“关系”近50万条,并记录其属性,已初具规模。

上海图书馆馆藏非常丰富的近代报刊文献资源,并一直致力于近代报刊的数字化进程,使更多的历史文献得以保存和展现。上海图书馆杨敏副研究馆员以报纸、期刊和图片的数字化过程作为案例,介绍了上海图书馆数字人文基础数据建设所取得的成果。近代报纸主要从正文、广告和图像三个方面作元数据著录。期刊方面已完成2万多种,1000多万篇数字化著录,精选近300种文学类期刊进行全文识别,现可提供篇内全文本查询。上海图书馆对近代文献资源内海量的图片单独进行著录,建设了“图述百年——中国近代文献图库”,目前已经推出了60万副图片^[5]。

4.3 数字人文中的用户关怀

近几年崛起的数字人文领域对公众科学项目的实施有了更多迫切的现实需求,文化遗产机构(CLAMs)逐渐认识到将其馆藏数字化传播和利用的必要性。南京理工大学经济管理学院赵宇翔教授基于科研众包模式的公众科学项目,概述了建设数字人文类公众科学项目的众包平台的运作流程,“任务匹配”、众包各个组成要素和“用户参与”。“用户参与”有三种驱动因素:一是平台驱动因素:感知有用性、感知易用性、平台社交性、平台专业化、游戏化元素、平台可视化;二是任务驱动因素:任务易操作性、任务自主性、任务趣味性、任务情境性、任务反馈性;三是志愿者驱动因素:好奇心和兴趣、乐趣、科学贡献、归属感、虚拟奖励、自我效能、感知学习、声誉。众包平台的可持续发展需要设计相应的激励机制来调动用户的参与意愿,结合上海图书馆盛宣怀档案抄录项目,从机构组织、平台设计和任务设计的角度提出了优化众包平台的激励对策,同时“在非游戏情境下使用游戏设计元素”来提升众包平台的可用性和用户体验,为众包情境下的游戏化模式探索提供了参考依据^[6]。

“数字人文”为人文学者提供了新的研究范式和视角。“数字人文”中人文学者的角色又该如何定位呢?哈佛大学计量社会科学中心王宏甦研究员以“中国历代人物资料库”(CBDB, China Biographical Database Project)项目为例,探讨了在人文数据库、数据集的建

设和使用过程中,跨学科合作的人文学者主要承担4种角色和任务:一是提问者,检查者。主要负责提出问题;自动标点;提供语料:对文本进行朝代标注;评判结果:对测试集的结果进行检查和反馈以确定如何平衡 Precision 和 Recall。二是主导者。利用数字技术的工具和软件,撰写正则表达式、读 python 代码、算法设计讨论和根据史料提出新算法。三是挑战者。既是旁观者又是批评者。四是转换者。将人文研究需求转化为信息技术专家可理解的计算机语言。基于人文学者的角色和任务,王宏魁研究员提出了对人文学者进行数字人文训练,主要内容包括方法论、技术和工具的使用和如何合作。为人文学者引入新的技术、方法和合作者,培养出既精通多种语言又专业的数字人文学者。

用户体验对数字人文服务的发展至关重要,近年来许多数字人文研究者专注于在提升用户体验方面进行理论研究和实践。英国伦敦大学学院信息学院(UCLDIS)付雅明博士借鉴认知绘图方法开展了用户体验的实践探索。认知绘图源于地理和心理学研

究,也被认为是民族志研究中的典型技术。它已被用于获取用户使用或思考特定资源或场所的方式的可视化表示。在社会科学领域,它以更一般的方式用于描绘人们如何理解世界。

参考文献:

- [1]严承希,王军.数字人文视角:基于符号分析法的宋代政治网络可视化研究[J].中国图书馆学报,2018(5):1-16.
- [2]夏翠娟,刘炜,陈涛,等.家谱关联数据服务平台的开发实践[J].中国图书馆学报,2016(3):27-38.
- [3]王蕾,薛玉,肖鹏,等.民间历史文献数字人文图书馆构建——以徽州文书数字人文图书馆实践反思为例[J].图书馆论坛,2018,38(3):30-36.
- [4]赵思渊.民间文书整理与研究中的数字人文方法与文献学本位[J].中国史研究动态,2017(5):1.
- [5]杨敏.近代中国报纸数字资源的建设和利用研究[J].图书馆工作与研究,2014(6):60-64.
- [6]赵宇翔.科研众包视角下公众科学项目刍议:概念解析、模式探索及学科机遇[J].中国图书馆学报,2017(5):42-56.

Cross-Disciplinary Dialogue for Digital Humanities:

A Summary of the "Flash Report" of the Digital Academic and Humanities Research Session of the 9th Shanghai International Library Forum

Jin Jiaqin Xia Cuijuan

Abstract: "Flash Report" is a special part of the Digital Academic and Humanities Research Session of the 9th Shanghai International Library Forum(SILF2018). From the perspectives of "methods and tools", "infrastructure construction" and "user experience", 12 speakers from Harvard University, London College, Peking University, Sun YAT-SEN University, Nanjing University, Nanjing University of Science and Technology, Fudan University, Shanghai Jiaotong University, East China Normal University and Shanghai Library, shared their researches and practical achievements in the field of digital humanities regarding the topics of "Technology Transformation Paradigm", "Data shaping the world" and "demand decision-oriented", which provided valuable reference and foundation for future development.

Key words: Digital humanities; Shanghai International Library Forum; Overview; Humanities research