斯坦福大学的数字人文学进展与学校图书馆的角色

杨继东

数字人文学的发展简况

- 数字人文学（digital humanities，台湾学界译成“数位人文学”）是电子计算机普及以后发展起来的将现代的机器计算技术运用于人文学（humanities）和社会科学（social sciences）的研究方法和手段。目前对它的定义有很多种，但是大致分成比较广义的和狭义的两种。广义的定义把所有利用电脑进行的人文社科研究行为都包括在内（比如利用历史或文学文献的全文数据库进行关键词检索），目前中国大陆的学界基本采用这种定义。而狭义的定义则认为数字人文学的基本特征是利用现代计算技术实现一些单用人脑很难或不可能实现的研究分析。本报告的“数字人文学”采用的是这种比较狭义的定义。
- 数字人文学起源很早，在现代电子计算机被发明后不久就开始了。第一份该领域杂志《计算机与人文学》（Computers and the Humanities）创办于1966年。但是数字人文学的真正进展出现于1990年代，也就是个人电脑逐渐普及以后。进入21世纪以来，它的发展更加迅猛。由于年轻一代的学者很多都在现代信息技术的年代里长大，他们拥有老一代学者不具备的高等数学和计算机编程技能，因此大大推动了这个领域的发展。
- 相对于面向西方历史文化和社会的数字人文学研究来说，对东亚包括中国的数字人文学进展比较落后。其原因当然跟东亚语言文字尤其是汉字的特性有关。由于汉语的单音节性质，汉字的巨大数量，以及书面汉语的字词之间没有空格等因素，用计算机对汉语文献进行分析需要采用与针对西方字母文字文献很不相同的技术。但是近十多年来，在美国和中国大陆以及台湾地区，使用数字人文学方法展开的中国研究也突破了一些技术瓶颈，呈现迅速发展的趋势。

数字人文学的主要方法和技术

- 对社会人文信息的空间和时间分布进行的分析。这种分析一般要利用到地理信息系统（GIS 即 geographical information system）和数字地图。
- 文本挖掘（text mining）也就是对大规模的全文数据库（full-text database）和语料库（corpus）进行分析，其具体表现包括词频统计，用自然语言处理（NLP 即 natural language processing）的各种模式进行的语言学分析，对专有名词比如人名、地名、职位、亲属关系名称的自动辨认和归类，对同形异义或同义异形词汇的统计和分析，对特殊句式的统计和类比，等等。
- 社会网络分析（social network analysis）是利用组合数学的图论（graph theory）和网络理论（network theory）对社会结构进行的分析。这种方法以节点（nodes）代表社会成员（个人或社会单位），以网络图形显示和研究节点之间的各种关联。利用计算机可以使得这种分析变得非常高效和可视化，也可以发现研究者很难通过阅读发现的系统关联。
- 从事数字人文学研究需要运用大量的数学知识和专门技能。一个研究课题的解决往往取决于一个或者多个特殊的算法（algorithm）的制定和应用。研究者还需要对很多前沿的编程和互联网技术有点基本的知识，比如掌握数字地图制作（digital mapping）技能对于利用地理信息系统必不可少，对HTML和XML等标记语言的熟悉是进行文本挖掘的基础，社会网络分析是建立在对相关图论和网络理论模式的理解上的。很多研究者还需要自己设计各种应用程序编程接口（API）以处理特殊的数据集。

大学图书馆对数字人文学发展的意义

- 数字人文学的研究必须基于大量的文本、图像、地理信息等数据。尽管现在有相当多的数据可以从互联网免费获得，但是大量适合学术研究的数据还是商业化的，大学图书馆依然是整个学术生态中不可或缺的一个组成部分。就像购买印刷的书刊一样，图书馆对数据的投资（包括购买和自建）是保障本校数字人文学发展的基础。
- 图书馆具有比较强大的IT队伍。通过近几十年来的数字图书馆建设，大学图书馆在元数据的编纂、文献数字化、服务器的管理、机构数据库的建设、对网络资源的整合、各种学术搜索方法的开发利用、不同语言文字的处理和编码等等方面都积累了丰富的实践经验和理论总结。而这些经验和技能是大多数人文社会科学学者所没有的。尽管有些学者具有相当不错的IT知识和技能，但是他们以个人之力往往难以应付比较大的研究项目，因此需要专业的服务。
- 从可持续性（sustainability）角度说，数字人文学者们在各种基金赞助下做出来的项目在达到一定阶段成果后往往很难延续。跟传统印刷出版的学术成果不一样，数字学术的成果很多是在网上生存的，需要经常性的技术维护和更新。在这方面，大学图书馆可以扮演一个关键作用，提供适用的平台并提供日常的技术支持。

斯坦福图书馆的CIDR及其服务

- 斯坦福大学在全球的信息技术产业中具有重要地位。硅谷的很多公司的创办和发展都离不开本校提供的人才。斯坦福图书馆的CIDR（Center for Interdisciplinary Digital Research 跨学科数字研究中心）也是美国高校图书馆中最早成立的专门支持数字人文学的部门。
- CIDR目前拥有7名全职人员，他们都拥有很强的计算机背景，擅长编程和网页制作，大多也拥有文科博士学位（在历史、人类、语言、社会、教育学等领域）。除了该部门的核心全职人员，CIDR也负责培训分散在全校各个文科院系的“学术技术专家”（Academic Technology Specialists），并通过他们向全校的文科师生提供数字人文学的帮助。
- “社会科学数据与软件”（Social Science Data and Software）是从属于CIDR的一个小组。该小组有固定的办公室，每天开放八个小时，欢迎各个院系的文科师生前来咨询与数字人文研究有关的任何问题，向他们提供建议和服务。
- “人文学术文本支持服务”（Humanities Text Support Service）是CIDR的另外一项重要服务。该部门接收斯坦福师生在各种数字人文项目中产生的大量文本数据，并提供保存、管理、升级、获取等各项服务。
- CIDR还定期向全校师生以及图书馆员提供各种讲座，介绍各种适用于数字人文学研究的技术和研究实例，比如 Python 语言和文本挖掘等。

CIDR支持的数字人文项目

- 除了上述各种日常服务以外，CIDR还直接为斯坦福文科教授们从事的一些数字人文学项目提供编程和网页设计、维护等技术支持。由于需要这种支持的教授们较多，CIDR成立了一个由教授和技术专家们组成的委员会，每年接收项目申请，并以从中选出若干个给与支持。到目前为止，已经做成了几个在全世界的数字人文学界很出名的项目，比如：
  - “古罗马世界地理空间模型”（ORBIS）。这是以数字地图为基础建立起来的一个古罗马时代的日常交通模型。它可以让研究者迅速找寻到古罗马境内的地名，并计算在不同地方之间以当时的运输手段旅行的路径和时间。
  - “血亲不列颠”（Kindred Britain）。以图像和时间轴线显示几个世纪以来英国人的血缘、社会、空间分布关系。这里包括3万多个人物，其中有很多名人包括莎士比亚和达尔文，也有不少普通人物。对研究英国社会和历史的演变有很大的帮助。
  - “现代中国的墓葬改革”（Grave Reform in Modern China）。此项目由历史系的墨磊宁（Thomas Mullaney）教授主导，以空间和时间的轴线来表现当代中国的墓葬改革过程，以及其中牵涉到的各种社会和文化现象。
  - “城中自然”（City Nature）。通过地理信息系统和文本挖掘表现美国大城市中的公共场地包括公园、市民休憩场所、自然保护区域的演变及其对社会和文化发展的意义。
- 除了图书馆，斯坦福大学出版社从去年开始也发表数字人文学项目，目前已经有一个上线，是通过一些数字化的老照片和地理系统讲述大峡谷的自然和文化历史。

数字人文学发展面临的问题和挑战

- 数字人文学尽管近年来发展迅速，但是也面临不少问题和挑战，首先是学术上的争论。有一些学者认为这种学术方式过于理科化，将众多的人文和艺术概念演变成纯数字的操作，削弱了人文的意义以及对文本的解读。也有人（尤其是研究古代历史文化的学者）觉得数字文本资源的准确性和可靠性有很大问题，在彻底解决之前无法形成被广为接受的研究成果。也有些人认为数字人文学的发展需要很大投入，挤占了很多传统学术研究需要的资源，造成很大的不公平。——连许多从事数字人文学研究的学者们也承认，所有这些批评在一定程度上是合理的。
- 过去几十年里，数字技术的发展和变化速度非常快，很多技术的通用性不高，时效性也很短。这就造成了数字人文学处于不断的快速演变中，难以形成比较固定的学术标准和规范，这就对数字学术成果的评议和审核造成很大困难。
- 数字人文学需要非常复杂的专业技术支撑，这对大学图书馆的工作人员是巨大挑战。即使对于像斯坦福这样位于硅谷拥有大量IT人员的图书馆来说，我们也深切感到人员配置的不足，无法满足师生们的所有要求。
- 大型数字人文学项目的可持续性依然是个问题。图书馆尽管在这方面可以提供一些帮助，但是网站的维护、数字学术成果的出版尚未找到一个各方都接受的可持续的模式，还需长期的探索。

数字人文学发展的前瞻

- 尽管有上述问题和挑战，过去20多年来数字人文学的发展还是非常迅猛的，随着具备现代信息技术知识的年轻一代人文社科学者的人数日益增长，对这个领域的兴趣在可预见的将来肯定还会高速增长下去。
- 数字人文学显然不会取代传统的人文学和社会科学研究方法，但是在某些以大量数据为基础的领域，比如社会学和人口学等，可能会成为主流方法。在另外一些以解读文本为基础的领域，比如历史和文学，仍然会是一种辅助手段。
- 近年来在数字人文学界逐渐形成一个共识：很有必要建立该领域的“基础设施”（infrastructure）。这种基础设施一般是指一个具备各种技术手段的通用网上平台（platform），它可以处理各种语言的文字数据以及图像、地图等资料，并能让并学者们通过一些简单明了的API实现常用的计算和分析。欧盟资助的DARIAH（Digital Research Infrastructure for the Arts and Humanities）就是这样一个“基础设施”。也有一些学者、图书馆和公司在努力建设适合专门领域的数字人文学基础设施，比如在中国研究尤其汉学方面，就有一些这样的尝试。
- 数字人文学使大学图书馆找到了新的服务方向和发展机会，但同时也对图书馆形成了新的挑战。图书馆员有必要密切跟踪相关领域的研究动向，增强自己的数字信息技术业务能力，以应对这些挑战。