■ 教育数字化·高等教育大数据专题　　DOI:10.15998/j.cnki.issn1673-8012.2022.04.003

高等教育大数据建设路径
——美国的经验及其对中国的启示

常桐善

（1.华中科技大学　教育科学研究院，武汉　430074；2.加州大学　校长办公室，奥克兰　94607）

摘　要：美国高等教育大数据建设已经有很长的历史，形成了包括组织运行、学生学习、教师教学与科研、服务、毕业生就业和毕业生社会贡献等社会、经济、教育领域的多种类型的大数据体系。其中有4个非常重要且相辅相成的发展路径：大学治理制度的建设、持续性收集和积累数据的过程、各级组织积极分享数据的行动以及整合来源于多元资源数据的复杂程序。缺少这4个过程中的任何一个都难以建设高等教育大数据，也不可能实现通过大数据赋能提升高等教育治理能力的目标。研究建议：中国高等教育大数据建设要加强和完善高等教育大数据治理体系，尤其是与大数据相关的立法、制度和组织建设；制定长远的高等教育大数据发展战略规划，确保数据收集和整合的持续性；加强高校内部运行数据系统的建设，提升数据收集能力，彻底改变数据“孤岛”现象；采取有力措施，加大数据共享力度，建立由政府、高等教育学会、高校联盟、社会组织共同协调的、多层面的高等教育大数据共享平台。

关键词：高等教育；大数据；建设路径；数据共享

[中图分类号]G646　[文献标志码]A　[文章编号]1673-8012(2022)04-0020-11

美国高等教育数据的发展和建设有很多年的历史，早期的数据在量和种类方面都难以展示大数据的特征。随着高校的决策模式从官僚主义、学院型等传统决策模式向更加科学的循证决策模式转变，以及高等教育发展的复杂化和竞争的日趋激烈，高校对大数据的需求持续升级。与此同时，计算机、网络、云端、人工智能、数字化等技术的快速发展及其在高校的普遍应用，为高校收集、整合、共享数据提供了便捷且能支付的技术。进入21世纪以来，美国高等教育大数据的发展和建设速度非常快。

以加州大学为例，目前收集的数据已经远远超过传统的大学运行系统的数据。新的数据来源包括广泛用于学习分析(learning analytics)的学习管理系统(learning management system)、从加州劳动局

修回日期：2022-05-19

作者简介：常桐善，男，甘肃山丹人，华中科技大学教育科学研究院兼职教授，加州大学校长办公室院校研究与规划主任，博士，主要从事院校研究。

引用格式：常桐善. 高等教育大数据建设路径：美国的经验及其对中国的启示[J]. 重庆高教研究，2022，10(4)：20-30.

Citation format：CHANG Tongshan. Toward the development of big data in higher education：the USA's experience and its implications for China[J]. Chongqing higher education research，2022，10(4)：20-30.

获得的用于学生学习成果评价(learning outcomes assessment)的毕业生就业数据、从税务部门获得的用于评价校友社会贡献度的纳税和捐赠等数据、从领英(Linkedin)等社交媒体获得的用于评价毕业生职业发展和成就的数据,以及通过数字化技术转换的加州大学从建校到20世纪80年代建立数据系统之前的在校生和课程设置数据等。总体上,到目前为止,美国高等教育大数据建设仍然领先世界各国,在一定程度上也引领世界高等教育大数据建设的发展方向。

中国高等教育领域的大数据建设也经历了相当长一段时间的发展。尤其是最近几年,国家相继出台有关文件,要求高校加快高等教育数字化转型和大数据建设。教育部2018年印发的《教育信息化2.0行动计划》明确强调,要完善教育管理信息化顶层设计,全面提高利用大数据支撑保障教育管理、决策和公共服务的能力。但不可否认,目前中国教育大数据建设仍然存在许多问题,"教育数据分散,教育信息孤岛现象较为严重","教育数据的收集和分析手段需要改进"[1]。这些问题也是高等教育大数据建设面临的挑战和亟待解决的问题。从表面上看,这些问题是数据收集和整合的问题,但实质上是大数据治理和应用的问题——缺乏大数据治理的法律和规章制度方面的保障以及在高校决策中广泛实践的理念。

虽然中美两国的高等教育体制不尽相同,但在高等教育治理和大数据建设目标方面应该有诸多相似之处。例如,高校运行数据是大数据的基础来源,循证科学决策模式、循证教学改革等都是中美高等教育体制大数据建设和应用的目标,所以美国高等教育大数据建设的经验对中国相同领域的建设应该具有借鉴价值。本研究首先简要介绍高等教育大数据的相关概念、特征以及类型,然后基于美国高校尤其是加州大学系统大数据建设的经验,阐述美国高等教育大数据的建设路径,最后讨论美国的经验对中国高等教育大数据建设的启示。

需要强调两个方面的问题:一是高等教育大数据是指高等教育领域的大数据,本质上是高校大数据的总称,所以下面阐述的高等教育大数据建设路径是基于高校大数据建设的实践经验,既反映高校层面的大数据建设路径,也适用于高等教育宏观层面的大数据建设。二是学者们为了开展高等教育学学术研究而收集和积累的数据也是高等教育大数据的重要组成部分,其特征和建设路径与高校层面通过运行和调研系统以及共享渠道获得的数据非常相似;加之美国很多学者使用的数据是从高校或相关机构获得的高等教育整合数据,所以本研究所阐述的高等教育大数据治理和数据收集、整合、共享的建设路径也适用于高等教育学学术研究领域的大数据建设。

一、高等教育大数据概念界定、特征及类型

(一)大数据的概念界定及特征

大数据概念的存在已经有很长时间,最早由迈克尔·考克斯(Michael Cox)和大卫·埃尔斯沃思(David Ellsworth)于1997年在电气电子工程师学会(The Institute of Electrical and Electronics Engineers)的报告中提出。他们强调:"可视化给计算机系统提供了非常有意义的挑战:数据表格太大,导致内外部驱动器的储存容量不够。我们把这些问题统称为'大数据'问题。"[2]根据吉尔·普利斯(Gil Press)对"大数据"发展历史的梳理研究,这篇研究文章是美国计算机协会(Association for Computing Machinery)收藏的所有文章中最早使用"大数据"这个术语的文章[3]。后来,陆续出现的"信息爆炸"(information explosion)和"商业智能"(business intelligence)等概念对理解大数据以及推动大数据的建设具有重要意义。

那么,高等教育领域是从什么时候开始使用"大数据"这个术语的?笔者查看了各类文献,没有找到具体的答案。史蒂文·伯勒尔(Steven Burrell)认为,20世纪90年代,伴随着"商业智能"和"分析科学"(analytics)等的广泛使用,美国大多数高校开始制定长远的信息发展战略规划,开发数据系

统,向高等教育大数据发展方向迈进[4]。显然,高等教育大数据的发展与其他领域的发展几乎是同步的。这个结论与笔者在 2002 年撰写的计算机科学硕士学位论文的研究结果基本一致。这篇论文分析了美国 100 所高校的信息技术发展战略规划,当时有些学校的信息技术战略发展规划已经过多次更新,进入第三、第四个 5~10 年的规划周期。这些高校提出的很多目标虽然没有使用“大数据”这个术语,但都聚焦于数据建设和实施措施。

20 年后的今天,高等教育大数据已成为高等教育领域最流行的话题之一。采用“big data in higher education”(高等教育大数据)和“big data and higher education”(大数据与高等教育)等关键词在谷歌网站搜索(关键词加双引号搜索,获取包含与关键词完全匹配的信息),可以找到上百万条相关信息,内容分布在“高等教育大数据建设”“高等教育大数据科学”“高等教育大数据应用”和“高等教育大数据革新”等诸多领域。

数据究竟“大”到什么程度可称之为大数据?似乎没有具体的界定。学者通常通过对特征的描述来界定大数据。网络上广泛流传的大数据特征描述是“5V”特征:量大(volume)、种类多(variety)、变化速度快(velocity)、真实性(veracity)和价值(value)。量大顾名思义就是指数据的数量大。以美国高校学生注册数据为例,学校通常记录每学期开学 3 周后的注册学生数以及到期末时仍然在校的学生数,也就是说每学期记录两次。可能会有学生中途离开再重新注册等情况,这样每学期的记录就有近 10 万条,一年就有近 20 万条,10 年后就可以收集约 200 万条数据。这些数据的量足够供学校分析学生的注册和保留行为。当然,这些数据的量与商业系统的数据量相比,确实不大,但对高校研究学生的注册行为和保留率来说,这些数据已足够,所以称之为大数据亦不为过。大数据种类多的特征可以从两个方面解读:一是同一组数据中的变量类别多,如学生招生数据可能包括学生个人特征、家庭背景、学校特征和社区特征等数据;二是数据的来源种类多,跟踪了解大学生成长特征的数据可能包括入学前的学习表现、入学后的课程学习结果、社会活动的参与程度以及毕业后的研究生教育或就业情况等数据。变化速度快是指数据更新速度快、单位时间内数据积累的频度高。在商业领域,大数据更多地反映某种交易量的变化,如商品的交易等,所以可能在短暂的几秒钟内,就有上万条新数据载入。这种情况在高校比较少见,高校常见的变化速度快的数据包括课程注册系统的数据,可能在系统开放后,短时间内的载入量也会达到上万条,但这种情况一年也就几次。校园智慧卡是另一个数据记录变化较快的领域。有 4 万名学生的大学可能每天的数据载入量就有上百万条。考虑到学生的隐私问题,美国高校很少记录和使用校园卡的数据。大数据的真实性和价值顾名思义就是指大数据的准确性、真实程度以及服务于决策、研究等的价值。从严格意义上说,真实性和价值并非大数据的特征,因为数据是否真实、是否能展示有价值的信息,在很大程度上还取决于数据挖掘的方法和能力,以及在数据结果解读时,有效结合机构事务原则的程度。有的数据可能对今天的发展来说没有价值,但一段时间后,当学校的战略发展方向发生变化,可能就有价值了。所以,如果我们一定要说真实性和价值也是大数据的特征,那么我们只能说它们是大数据的“隐性特征”,需要挖掘方可显示,量大、种类多、变化速度快则可称之为大数据的“显性特征”。

(二)高等教育大数据的类型

理解高等教育大数据的类型有助于合理设计数据收集、整合和共享的方法。根据数据所反映的内涵意义,可以将大数据分为事实数据(fact data)和行为数据(behavioral data)。事实数据记录事情发生的频度,如注册学习某一门课程的学生人数、在顶尖期刊发表论文的教师人数以及教师的工资等。可以说,高校运行系统记载的数据大多数是事实数据。当然也有例外,如学生学习管理系统记录的数据就是混合型数据,既包括事实数据,如完成作业的情况,也包括行为数据,如先阅读资料还是先完成作业的行为。行为数据是通过对研究对象的观察、试验、人工智能工具的跟踪记录以及研究对象

自我反馈等方式收集到的反映研究对象行为的数据。事实数据通常回答“是什么”和“如何”等问题,而行为数据则有利于解释“为什么”和“未来怎样”等问题。例如,学生的学习成绩是事实数据,可以用来回答“学生的学习成绩如何”等问题;学生学习投入反映了学生的学习行为,可以用来解释学习成绩差异的原因,回答“为什么学生的学习成绩有差异”和“如何帮助学生提升他们的学习成绩”等问题。通常情况下,相比于收集事实数据,收集行为数据的难度更大,所以研究人员在研究设计时搞清楚数据的类型对保证数据信效度极为重要。

数据也可以根据其来源分为客观数据和主观数据。客观数据是通过记录客观事件的发生情况获得的数据,可能是事实数据,也可能是行为数据。例如,教师通过记录学生在课堂上的提问次数,对学生的学习参与行为进行记录而收集到的数据就是客观数据。相反,如果我们通过调查问卷,让学生通过回顾式方法判断自己课堂参与的频度(如“经常”“有时”“偶尔”和“从不”等)而收集到的数据,是学生根据自己的回忆和判断提供的数据,就属于主观数据。尤其是当学生对“经常”和“有时”等术语的主观判断有差异时,对自己参与程度的判断就可能产生误差。显然,客观数据比主观数据具有更加可靠的信度,但要求教师长期记录学生的课堂参与程度是不现实的,所以从时间和经济效益的角度来说,在很多领域收集主观数据的可行性更强,而且在有些领域也只能收集到主观数据,如学生对学校服务的满意度,这也是美国高校目前收集类似数据的主要方法。当然,随着技术的发展,高校也逐步使用课堂反馈仪记录学生的参与情况。例如,学生提问或回答问题时用“回答仪”代替传统的举手方式告知老师。“回答仪”与学习管理系统链接,可以客观记录学生课堂参与程度数据,同时也起到了考勤的作用。

根据时间跨度,大数据又可分为片段性数据(snapshot data)与持续性数据(longitudinal data)。片段性数据顾名思义就是反映一项活动的某一个或者某几个片段的数据,而持续性数据则反映一项活动的完整过程。事实上,在绝大多数情况下,高等教育领域和商业领域的数据都是片段性数据,即使是我们常常称之为持续性数据的数据也只能是“相对持续性”的,而不是“绝对持续性”的。例如,本科生某一学期的学习过程数据对本科教育的整体过程来说是片段性数据,但每学期的成绩可能是由期中考试、平时测验、作业、参与等具有一定持续性的数据组成的。对这一学期来说,这样的数据反映了学生学习的整体情况,在这个特定的时间段内,称之为持续性数据亦可。在高等教育领域,片段性数据对战略规划进展监测、学生学习过程评估和增值评估等都具有非常重要的意义,但持续性数据更有利于大学评价治理效能、办学效益和教育质量改变的全过程。

从数据的储存形式和结构特征来说,大数据包括结构性数据(structured data)与非结构性数据(unstructured data)。结构性数据就是我们常说的储存在数据库里的行数据,可以用二维表格来展示。例如,我们常见的学生注册统计表格,就是由学生姓名、课程名称构成的纵横二维的结构化数据表格。目前高等教育大数据中的绝大多数数据都是结构性数据。非结构性数据是指字段长度不等并且每个字段的记录可以由可重复或不可重复的字段构成的数据,包括文本、图像、声音、影视和超媒体等信息。随着网络技术的发展,高等教育大数据中的非结构性数据不断增多。美国高等教育目前最常见的非结构性数据来源于社交媒体、大学申请的个人陈述(personal statement)和学生对调查问卷中开放性问题(open-ended questions)的回答。了解数据组成的结构性和非结构性特征对合理设计和开发大数据储存系统、正确选择数据挖掘工具、提升数据应用效度均有助益。

随着数据的增多,数据安全越来越成为大数据治理的难题。从数据类型来说,正确界定和处理隐私数据(privacy data)与非隐私数据成为大数据建设必须认真解决的问题。基于大数据的犯罪行为时有发生,如诈骗、欺凌、广告骚扰等,电话、邮件诈骗就是由个人信息被盗或泄漏而导致的后果。通常

情况下,与个人信息(如姓名、社会安全号、学/工号、邮件地址、住址、电话号码和社交媒体号码等)链接的所有数据都是隐私数据,不能公开,而且要加密储存。在数据汇总报告时,与个人背景特征同时报告的汇总数据都必须考虑群体人数的问题(group size issue)。例如,在美国的很多大学,如果某个族裔群体人数少于 10 人,就会在汇总报告中删除这个群体的数据,如平均成绩、课程成绩达标比例等。这样做一方面是为了保护学生的隐私,另一方面也是为了保证数据分析结果的信度。如果某个群体的人数太少,任何小的变化都可能会改变研究和分析的结论。美国高等教育需要普遍遵循的两项法律是《家庭教育权与隐私法》(Family Education Rights and Privacy Act, FERPA)[5] 和《健康保险隐私及责任法》(Health Insurance Portability and Accountability Act, HIPAA)[6]。

二、高等教育大数据的建设路径

(一)完善大数据治理制度

数据治理(data governance)是大数据时代新的管理实践理念。传统的数据管理模式强调从技术层面进行数据的收集、储存、管理等事宜,数据治理则彰显了数据管理的综合性和整体性特征,强调大学决策、利益人和数据安全等,涉及大学战略发展的整体过程,其目的是通过综合治理模式有效控制和管理大学的数据资产,提升其质量、价值和利用率。从数字化转型的视角讨论,这种从数据管理到数据治理的变化除了需要强大的数字化技术支撑外,更需要数字化战略的引导,包括对大数据的认知和持有的价值观、国家和地方层面的大数据治理法律保障和指导性政策,以及高校层面的数据资产管理和数据共享的规章制度等[7]。

从 1965 年开始,美国联邦政府先后出台的涉及数据治理的法律包括《美国高等教育法》(Higher Education Act of 1965)、《国家研究法》(The National Research Act of 1974)和《高等教育透明度法》(Higher Education Transparency Act)等。这些法律包括数据收集、学生隐私保护和数据公开等多方面的条款。由于美国高等教育运行隶属州政府管辖,各州也出台了支持本州高等教育大数据发展的法律和相关政策。例如,加州的《公共信息法》(California Public Records Act)要求在保护个人隐私的前提下,包括公立高校在内的州立公共机构有义务对外分享数据,提升公众对教育的知情权。2021 年,加州又通过了旨在建立加州“摇篮到职场”数据系统(The California Cradle-to-Career Data System)的法律,要求公立教育系统以及教育局、卫生局、劳动局等州政府机构共同参与,建立加州学生层面的教育数据系统。这个系统建成后,将包括加州所有基础教育和高等教育在校学生的背景、课程学习、就读经历和学业完成情况等学生成长数据,而且会每年更新。在高校层面,大数据治理的规章制度更多。加州大学校长办公室在过去十多年已逐步完成了从数据管理向数据治理的过渡,相继出台了一系列数据治理政策,包括部门之间协调运行数据的合作要求,数据的收集、储存和整合制度、报告平台开发和工具选用指南、院校研究数据服务指南、数据资产管理制度、数据隐私保密制度、用户培训措施以及数据结果公开和报告制度等[8-9]。

另外,建立健全大数据建设组织机构也是提升大数据治理能力的重要组成部分。美国高校大数据建设通常由学校的信息中心与院校研究部门合作完成。院校研究部门从大学运行规则和数据分析的角度提出大数据需求报告,包括数据变量、标准、使用工具和需要解决的问题等。当然,院校研究部门通常与职能部门(教务、财务、学生工作、科研和人事等部门)合作完成这样的报告。信息中心会根据大数据需求报告,提出技术需求及实施报告。显然,美国高校的院校研究机构立足院校运行规则对大数据建设提出需求,有效推动了大数据建设的进程[10]。

(二)持续收集数据

高等教育大数据的主要来源是高校的运行数据系统,也就是用于大学日常业务运行的数据系统,

包括工资发放、财务报销、学生注册、学生学习和教师科研等各个领域。在此基础上,大学通常建立大学层面的数据仓储,与运行数据系统对接,定期导入运行数据系统的数据,供大学数据报告、分析和相关学术研究使用。由于美国高校在20世纪80年代就开发了类似的运行数据系统,所以很多高校在部分领域的数据积累已经有40年的历史,基本形成了本校的大数据系统。例如,加州大学总校目前的部分学生系统就包括过去30多年的400多万名学生的申请资料和入学数据。另外,加州大学总校通过数字化技术,将加州大学从1869年建校开始到建立加州大学数据系统之前的学生信息进行数字化处理,形成了包括几千万条信息的学生数据库。所以说,高校持续性地收集数据是高等教育大数据建设的主要路径。

为了进一步了解学生的学习和活动行为、教师的教学、科研和服务行为以及他们对学校服务、校园校风的满意度等,美国高校从20世纪80年代开始,通过调查问卷收集这些领域的行为数据。可以说,美国几乎所有的高校都定期开展这方面的调查研究。加州大学从2002年开始,每两年对本科学生进行一次调查研究,截至目前已经开展了11次调研,积累了60多万条调查数据,每一条数据包括300~400个变量[11]。此外,加州大学还对研究生的就读经历、在校学生的经费开支以及教师的教学等情况进行调查。所有这些数据都可以与运行系统的数据对接,形成了集事实数据与行为数据、客观数据与主观数据、片段性数据与持续性数据为一体的数据系统。虽然调查数据本身的量不算大,但与运行数据系统的学生入学、课程学习、毕业、资助、就业以及教师教学等数据整合后,其价值就会成千上万倍地增加。

另外,随着高校对个性化教育的日趋重视,学校希望通过学习分析来解决学生学习差异的问题,所以很多学校开始通过学习管理系统收集学生个性化学习数据,尤其是针对网络课程教学,通过学习管理系统收集数据更加可行。但由于涉及很多教师和学生隐私方面的问题,学习管理系统对数据的收集还停留在供授课教师使用的层面,还没有广泛导入大学整合型的大数据系统。相信随着大数据治理体系和技术的成熟,学生学习数据将是高等教育大数据最有价值的数据组成。

除了高校从管理的角度持续性地收集数据外,高等教育领域的学者为了开展学术研究也在不断收集相关数据。虽然这些数据可能由于缺少个体识别信息,难以与高校大数据系统的数据整合,但仍然是高等教育大数据的重要组成部分。学者通过大数据研究,从不同的角度为高等教育发展提供了科学研究依据。

(三)交换与共享数据

在国家和州层面的众多法律和规章制度的保护和授权之下,高校以及相关机构展开了多层次、多维度的数据交换和共享。这种数据共享是高等教育领域大数据建设的主要途径之一,在高等教育大数据建设中发挥的作用与高校内部、学者进行的数据收集同等重要。从大数据的特征和类型来说,共享数据可以大幅度增加数据的种类,提升数据的量和价值,可以拓展持续性数据的时间跨度和数据变量维度,有利于增强高等教育大数据的完整性。以学生发展数据为例,我们需要学生就读大学前的学业与成长环境相关数据、学生入学后的学习过程数据以及学生离校后的继续深造和职业行为数据,才能够全面展示学生接受本科教育的成就,也就是增值结果。更重要的是,这样的具有持续性、连贯性的大数据也有利于探索和研究学生的成长规律,为改进教学提供依据。要获得完整的数据,单纯地依靠自己收集永远也无法达到要求,必须依靠数据共享。目前,加州大学共享数据的部门除了高校外还包括教育行政部门、劳动部门、税务部门、出版集团、社交媒体、高等教育协会、高校联盟和考试机构等。

美国高校共享数据最为典型的方法是通过政府投资建立数据共享平台、大学联盟以及非营利机构。表1展示了5个较为普遍的数据共享案例。

表 1　美国政府以及高校联盟教育数据共享系统

联盟/大数据系统	隶属关系	大数据建设目的	大数据描述
整合型高等教育数据系统	联邦政府	为教育实践和政策制定提供科学证据,并与教育人员、家长、决策者、研究人员和社会公众分享高等教育信息	大学层面的整合数据,包括高校公私立等基本特征、大学规模、录取率、毕业率、学生组成等数据;所有数据在网络公开发布
加州“摇篮到职场”数据系统	州政府	加州正在筹建的包括全州学生从入读幼儿园到就业的数据跟踪系统,为大学、学者、家长、学生以及社会各界提供加州教育信息,提升教育透明度、公平、机会和质量,帮助学生实现他们的教育目标	包括加州各级公立学校入学、大学招生、就业、学校背景、社区背景、学生学业成绩和学生家庭背景等领域的数据;系统通过统一的识别信息跟踪记录学生成长的全过程
美国大学协会数据交换联盟	高校联盟	美国大学协会的数据交换机构,其目的是提升协会成员的信息质量和使用力度,为大学决策提供支持	数据交换包括公开和保密数据;保密数据主要是成员大学根据各自的需求而交换的各自的数据;公开的报告显示,交换数据包括财务、研究生教育、教师、本科教育等数据
研究型大学本科生就读经历调研联盟	高校联盟	为联盟大学探讨本科教育发展所面临的问题以及分享学生就读经验调研数据提供平台,帮助联盟大学改进本科教育	通过问卷收集学生学习参与、时间分配、满意度、就业计划等领域的数据;联盟大学收集调查数据后,与学生学业发展方面的数据整合,并与联盟成员大学共享学生层面的数据
全国学生数据中心	民间组织	非营利民间组织,提供教育分析报告、数据交换平台、学位验证和研究服务的机构	大学自愿参加的数据交换机构,数据包括学生的入学和毕业信息

注:资料来源于各机构的官方网站:整合型高等教育数据系统(https://nces.ed.gov/ipeds/),加州“摇篮到职场”数据系统(https://c2c.ca.gov/),美国大学协会数据交换联盟(https://www.aaude.org/),研究型大学本科生就读经历调研联盟(https://cshe.berkeley.edu/seru),全国学生数据中心(https://www.studentclearinghouse.org/)

美国教育统计中心建立的“整合型高等教育数据系统”(Integrated Postsecondary Education Data System, IPEDS)是高校分享学校层面数据的主要平台。这个平台除了在线公开数据外,也为高校提供网络下载全部数据的工具,高校可以将下载数据与内部数据整合,是标杆研究数据的主要来源。加州正在开发包括全州学生从入读幼儿园到就业的“摇篮到职场”学生发展数据跟踪系统。这个系统将包括加州各级公立学校入学、大学招生、就业、学校背景、社区背景、学生学业成绩以及学生家庭背景等领域的数据。系统通过统一的识别信息跟踪记录学生成长的全过程,数据系统建成后对提升教育决策绩效、推进教育改革会有实质性的作用。事实上,包括美国俄勒冈和康涅狄格等在内的十多个州也已经启动了类似的教育数据共享系统建设项目。美国大学协会数据交换联盟(American Association of University Data Exchange,AAUDE)是该协会成员大学数据交换和共享的平台。AAU 包括 65 所北美顶尖大学,其中 6 所大学是加州大学的分校。笔者所在的院校研究部门也经常使用这个平台开展各项研究,为学校提供决策支持。研究型大学本科生就读经历(Student Experience at the Research University,SERU)调研联盟是由加州大学伯克利分校倡议并创建的,先后有近 50 所大学参与,并利用统一的调查问卷每两年收集一次数据,同时共享学生层面的调研数据,帮助联盟大学进行本科生就读经历的参照比较研究,为学校制定本科生教育政策提供有效依据。全国学生数据中心(National Student Clearinghouse,NSC)是一个非营利机构,是高校追踪了解学生入学去向的数据平台。这个机构的运行模式是高校作为会员加入组织,并分享在校生的有关数据,数据需求大学提交学生姓名等可以识别的学生信息,中心从其数据平台将搜索到的信息反馈给需求大学。例如,加州大学每年汇报的

被录取学生入读其他高校的数据、从加州大学转入其他高校的学生学业完成动态数据以及加州大学本科毕业生攻读其他高校研究生学位的学业动态数据都来自这个平台。

除了上面阐述的数据共享和交换行动,美国高校之间也通过合同形式互相交换学校层面的数据。笔者工作的院校研究部门开展了很多类似的数据交换工作。例如,我们通过合约不定期地从加州教育局获取中学生参加加州课程教育评价考试的数据,每年从大学董事会(college board)获取加州所有参加类似于中国高考的SAT考试和先修课程(advanced placement courses)考试成绩的数据,从加州社区学院获取就读加州社区学院学生的修课和学业成绩的数据,通过合约从加州劳动局定期获取所有在加州工作的加州大学毕业生的就业数据等。这些数据包括学生个人的识别信息,我们可以与大学内部收集的其他数据整合,有效实现推动大数据建设和数据赋能的目的。

(四)整合数据

整合数据(integrating data)是将技术和组织运行规则(business rules)相结合,清理、合并从不同渠道收集到的数据,通过统一数据标准、提升数据质量、增加数据价值,为开展有意义的数据挖掘、产生有价值的信息奠定基础。数据清理较为容易理解,即对原始数据中存在的问题进行处理,确保数据的准确性和真实性。但如前所述,大数据是否具有真实性是一个复杂的问题,主要取决于数据收集设计的合理与否。在数据整合过程中很难完全确定和解决数据的真实性问题,但至少可以核实数据变量的界定是否符合高校的事务运行原则。整合数据是一项非常重要且花费时间的工作。根据笔者从事院校研究的经验,如果我们需要从整合数据开始来开展一项研究工作,那么花费在数据整合方面的时间远远超过花费在数据分析上的时间。但其好处是,整合好的数据可以重复使用,这也是在建立大数据系统时必须对数据进行整合的原因。

整合数据通常包括3个基本步骤。一是将从不同渠道获得的原始数据(包括从运行系统导入的数据、外部数据等)装入大数据仓储的暂存数据(staging data)空间。通常情况下,从原始数据到暂存空间之间不进行任何数据清理和整合工作,而是保留原始数据,以备后面开展数据清理和整合出现错误需要重新装载原始数据时使用。二是按照数据设计结构以及组织商业运行规则,清理储存在暂存空间的数据,合并数据,建立数据表格之间的关联,然后将数据导入基本数据(base data)空间。这时就可以让数据用户测试数据。三是在基本数据的基础上进一步清理和整合,以供数据报告和分析使用。整合好的数据通常被称为数据产品(data production)或商业智能数据,也叫终端数据。

上面阐述的从暂存数据到基本数据再到终端数据之间都要经过抽取、转换和加载(extract,transform and load,ETL)3个清理和整合的过程,抽取是把收集到的或者由运行系统导入的数据进行筛选,清除垃圾,保留有用的东西。转换过程是数据纠错和规范化的过程,常见的转换包括统一变量格式,如统一不同表格中出生年月、性别、日期等的格式,在原始数据的基础上建立新的能够反映学校运行规则的变量,如根据专业信息建立用于数据分析和报告的学科类别变量等。

图1展示了加州大学整合型数据系统中的学生大数据建设路径。每一个方框代表了大数据中的一个维度,每个维度可能包含众多数据变量,如"大学申请"数据维度含有超过3 000个数据变量,分布在学生背景、家庭背景、高中学业完成情况、大学入学考试成绩和课外活动参与情况等10多个领域。横向维度展示了学生从大学准备(学前情况)、申请大学、进入大学(入学、修习课程)一直到毕业、就业的发展路径。这些数据来源于州教育部门、考试机构、大学申请系统、学生注册系统、调研数据、学习管理系统、非营利性机构、校友调查数据、州劳动局、税务部门以及社交媒体。所有这些数据维度和变量都通过学号、学校等识别变量链接成一张网,构成上万个数据变量和上百万个数据变量组合(如学习成绩与性别就是一个组合)。显然,任何单项数据的价值都无法与整合后的大数据价值相比。整合后的平台实际上已经形成了集基础教育、加州大学教育、就业以及加州经济、人口特征、社会

贡献为一体的学生发展路径大数据平台。利用这个平台,我们不仅可以全面系统研究和总结加州大学本科生的成长特征和规律,探究加州大学本科教育的优劣势,也可以研究整个加州高等教育的公平性和对社会经济发展的影响,以及加州大学学士学位的价值和毕业生对社会的贡献力。所以研究结果既可以为州政府制定基础教育和高等教育政策提供依据,也是加州大学制定本科教育政策、大学预算的主要依据之一。

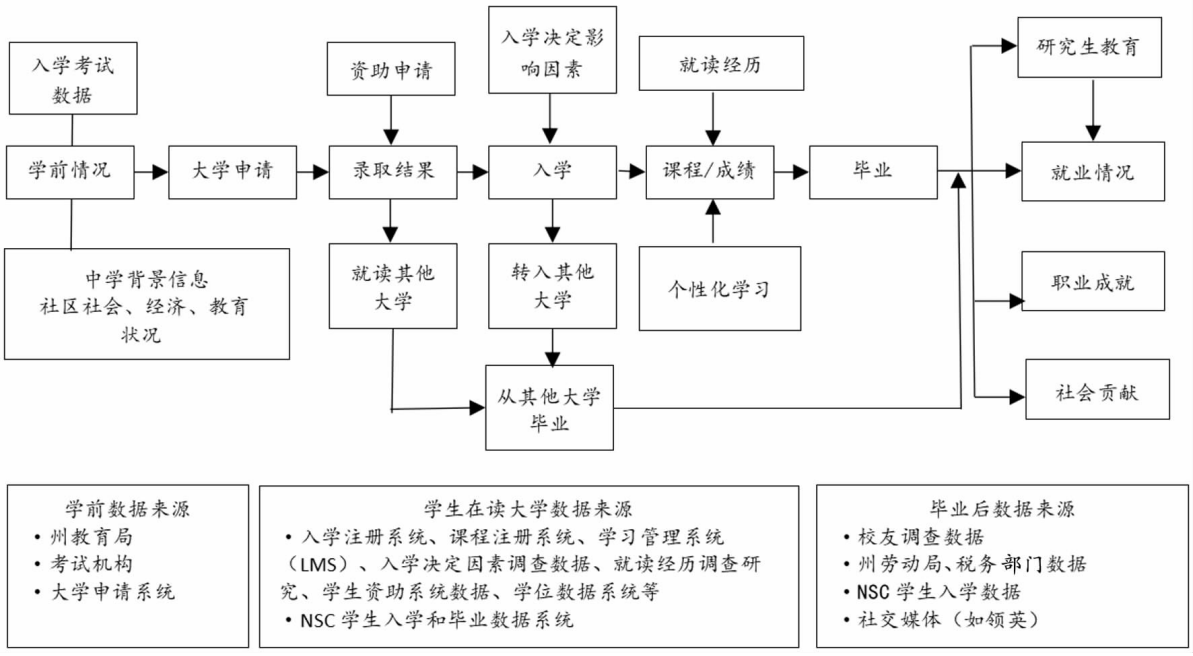

图 1　加州大学系统本科生发展大数据基本元素和整合路径

三、对中国高等教育大数据建设的启示

综上,美国高等教育数据类型众多,既有事实数据,也有反映大学教育、科研和服务的行为数据;既有反映大学运行的客观数据,也有反映大学不同群体声音的主观数据;既有展示大学某一个阶段的片段性数据,也有展示大学一个世纪以来开展的教育活动的持续性数据。虽然美国高等教育大数据仍然无法与众多商业领域的大数据相比,但其数量、种类、变化的速度以及包含的数据价值等都已今非昔比,能够满足高等教育众多领域的数据研究需求,所以称之为大数据不应该有任何异议。当然,美国高等教育大数据的建设经过了多年的努力,在投入了大量的人力和财力的基础上才取得这样的成就。其建设路径可以概括为 4 个相辅相成、同步发展的过程:大数据治理制度和文化氛围的建设过程,持续性收集和积累数据的过程,大学以及考试机构、州劳动局、非营利性组织积极分享数据的行动过程以及赋能数据的整合过程。美国经验证明,缺少这 4 个过程中的任何一个都难以建设高等教育大数据,也难以实现高等教育循证决策以及提升教育质量、消除教育不公平的战略目标。当然,美国高等教育大数据建设仍然存在很多需要改进的地方,如数据安全、学生和教师隐私保护措施不到位、教学过程中的个性化学习和线上教学数据短缺以及教师教学质量评价数据缺乏可靠性等问题。但其长期以来总结的经验仍然对中国高等教育大数据建设有一定的启示,对利用大数据提升教育治理能力有借鉴价值。

第一,要加强和完善高等教育大数据治理体系建设,尤其是与大数据相关的立法、制度和组织建设。中国高等教育大数据的建设在很大程度上仍然沿用传统的管理模式,还没有从法律、制度层面形成国家、高校层面大数据治理的完整体系。因此,还没有形成浓厚的大数据建设和应用氛围。另外,

目前高校的大数据建设基本上依靠信息部门,而职能部门尤其是类似于高教研究所等数据使用部门的参与程度很低。这样的模式非常不利于技术与大学运行规则的整合。当然,没有明确的院校研究组织机构也始终是推动数字化转型、大数据建设和应用的主要障碍。没有院校研究就难以形成大数据应用的文化氛围,没有大数据应用的文化氛围,自然很难树立全校范围内的大数据建设观念,更谈不上树立大数据治理理念。所以,加强和完善高等教育大数据治理体系建设是解决"数据分散"和"数据孤岛"问题的关键措施。

第二,要制定长远的高等教育大数据发展战略规划,确保数据收集和整合的持续性。如前所述,与其他很多领域的大数据建设相比,高等教育大数据建设周期更长,所以更需要持续性建设。也就是说,高校必须从数据收集和整合、系统开发、技术和人财物配置等方面制定长远的发展战略规划,确保各方面工作的持续性。中国教育行政部门和高校习惯于通过短期科研立项、项目开发外包等方式来建设大数据平台。这样的平台虽能解决短期数据需求问题,但从长远的大数据建设来说,弊大于利。所以,中国高等教育大数据建设的当务之急是行政部门、高校以及高等教育相关机构共同制定长远的高等教育大数据建设战略规划,提升高校内部的大数据建设能力,确保数据收集和整合的持续性。这其实也是大数据治理的重要内容。

第三,加强高校内部数据运行系统的建设,提升数据收集能力,彻底改变数据"孤岛"现象。如前所述,高校的数据运行系统是高等教育大数据建设的基础,如果没有强大的数据运行系统,高等教育大数据建设就是一句空话。前面阐述的美国大数据建设中存在的问题也主要是由个性化和线上教学运行系统(学习管理系统)、教师教学质量评价运行系统还不够完善导致的。要提升运行系统的数据收集能力,就必须将运行系统的设计与学校的运行规则和决策需求紧密结合起来,使运行系统不仅能满足大学日常运行的需要,也能够为高等教育大数据研究打好基础。同样重要的是,要通过数据整合和大学层面的数据仓储平台建设,打通各部门运行系统之间的链接渠道。否则,即使有了强大的运行系统,也难以建成整合型的数据平台,也就不可能形成大数据。从技术上来说,这项工作已经非常容易,但关键是要有高等教育的数据建设顶层设计,包括教育行政部门和高校两个层面的设计。

第四,采取有力措施,加大数据共享力度,建立由政府、高等教育学会和高校联盟等组织协调的、集基础教育和高等教育以及社会和经济为一体的、多层面的高等教育大数据共享平台。如果没有数据共享,高等教育也许能够满足大数据的量大和变化速度快的特征,也许能够积累足够的反映本校的事实数据,但很难达到大数据种类多元、齐全的要求,当然也必然导致数据价值的局限性问题。促进数据分享的另一个好处是,学者也可以从各级高等教育机构获取高等教育数据来开展大数据研究,从而增强大数据的使用价值,实现真正意义上的大数据赋能。

参考文献:

[1] 张伟.用大数据技术助力教育变革[N].光明日报,2017-12-05(16).

[2] COX M, ELLSWORTH D. Application-controlled demand paging for out-of-core visualization: proceedings of the 8th IEEE visualization 1997 conference[EB/OL].[2022-04-29]. https://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/parallel_volviz/paging_outofcore_viz97.pdf.

[3] PRESS G. A very short history of big data[EB/OL].[2022-04-29]. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#504546da55da.

[4] BURRELL S. The past, present and future of big data in higher education[EB/OL].[2022-04-29]. https://evolllution.com/technology/metrics/the-past-present-and-future-of-big-data-in-higher-ed/.

[5] Family education rights and privacy act[EB/OL].[2022-04-29]. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

[6] Health insurance portability and accountability act[EB/OL].[2022-04-29]. https://www.hhs.gov/hipaa/index.html.

[7] 常桐善.推进高等教育数字化转型,强化治理效能:美国的实践经验及其对中国的启示[J].中国教育信息化,2022,28(2):13-26.
[8] Systemwide information technology policies & guidelines[EB/OL].[2022-04-29].https://www.ucop.edu/information-technology-services/policies/index.html.
[9] Data operations hub[EB/OL].[2022-04-29].https://data.ucop.edu/index.html.
[10] 常桐善.美国院校研究的过去、现在和未来[J].高等工程教育研究,2013(2):133-142.
[11] Institutional research and academic planning survey services[EB/OL].[2022-04-29].https://www.ucop.edu/institutional-research-academic-planning//services/survey-services/index.html.

(编辑:张　腾　校对:杨慷慨)

Toward the Development of Big Data in Higher Education: The USA's Experience and Its Implications for China

CHANG Tongshan

(1. *School of Education*, *Huazhong University of Science and Technology*, *Wuhan* 430074, *China*;
2. *The University of California Office of the President*, *Oakland* 94607, *USA*)

Abstract: The USA has a long history of developing big data in higher education, which include social, economic and educational data in many different areas such as institutional operation, student learning, research, services, alumni employment and contributions to the society. This paper describes four important and yet interrelated processes for big data development. They are data governance, continuous and longitudinal data collection process, actions of data sharing across organizations, and integration of data from various sources. The study stresses that without any of these four processes, higher education cannot develop big data to achieve its goal to improve governance by data empowerment. It is suggested that Chinese higher education should enhance and improve its data governance, especially establishment of big data laws, regulations, and organizations; develop a long term strategic plan for big data development to ensure longitudinal data collection and integration; enhance operational data systems at the institutional level, improve capacity of data collection and break down data islander walls; adopt effective steps to strengthen data sharing and develop multi-level data sharing platforms coordinated by governments, higher education associations, institutions, and non-profit organizations.

Key words: higher education; big data; building pathways; data sharing