



推进数据密集科学发现 提升科技创新 能力:新模式、新方法、新挑战^{*} ——《第四范式:数据密集型科学发现》译著出版

文/梁娜 曾燕

中国科学院国家科学图书馆 北京 100190

【摘要】 海量科学数据正日益成为科学研究的有力工具及知识基础。文章从《第四范式:数据密集型科学发现》一书出发,通过地球与环境科学、生命与健康科学、数字信息基础设施和数字化学术信息交流等方面基于海量数据的科研活动的生动实例,揭示了这种数据密集型科学发现的新范式对科学研究带来的新模式、新方法,最后从科学数据的管理、互操作、数据权益管理、数据素养等方面提出了驾驭第四范式的挑战。

【关键词】 第四范式,科学数据,数据管理,知识服务

DOI: 10.3969/j.issn.1000-3045.2013.01.011

由微软公司副总裁 Tong Hey 等撰写、中科院潘教峰等翻译的《第四范式:数据密集型科学发现》(*The Fourth Paradigm: Data-intensive Scientific Discovery*, 以下简称《第四范式》)已由科学出版社出版。这是国际上第一本系统描述大数据现象、深刻揭示其对科学研究的革命性影响的著作,对我们如何理解和组织科学研究、科研管理以及科研知识服务,具有重要意义。

1 科学数据全面成为科学研究的基础设施和工具

随着信息网络和信息技术的发展,科学

研究的知识基础发生了革命性的变化。通过各类观察、感知、计算、仿真、模拟、传播等,海量的科学数据(被称为“大数据”——Big Data)正迅速产生、广泛传播和有效组织保存,成为科学研究的有力工具、甚至新的基础。

科技界一直就很重视科学数据,很早就建立了大规模的科学数据中心,例如由国际科学联合会理事会建立的世界数据中心(World Data System)和它的遍及全球的数据中心体系,英国的 JISC 国家数据中心网络(JISC National Data Centers),以及美国校际社会科学数据共享联盟(ICPSR)等等。随着技术和网络的发展,科学数据的数

^{*} 修改稿收到日期:2012年11月2日

量迅速增加,科技界开始高度关注这种大数据趋势对科技以及社会的影响。2005年美国科学理事会(NSB)发表《支持21世纪科学和教育的可持续数据集合》(*Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*),分析组织和维护大规模数据来支持科学研究的需要;2006年微软公司出版了《科学2020》(*Towards Science 2020*),探讨了网络环境下如何运用科学数据开展科学研究的趋势和要求;2007年美国国家科学基金会(NSF)提出了《21世纪科学的网络技术设施》(*Cyberinfrastructure Vision for the 21st Century Science*),探讨如何通过海量数据及其网络化计算来支持科学研究;2009年美国国家科学技术理事会(NSTC)发布了《为21世纪而驾驭科学数据潜力》(*Harnessing the Power of Digital Data for the 21st Century*);欧盟委员会科学数据高级专家组也在2010年发布《驾驭趋势:欧洲如何从科学数据的迅速涨潮中获益》(*Riding the Wave:How Europe Can Gain from the Rising Tide of Scientific Data*)。科学媒体和大众媒体也关注到大数据的兴起和影响。《经济学家》在2010年2月27日当期专门以“数据洪水”(Data Deluge)为封面文章,探讨了科学数据对科学和社会的影响。著名的*Wired*杂志甚至在2008年还提出,数据洪水可能意味着“理论的终结”(End of Theory)^[1]。

各国科技界已积极行动起来。例如,美国NSF在2007年启动了Data Net项目^[2],汇集网络基础设施、计算机与信息科学、图书馆学、档案学和领域学科专家的能力和 innovation,探索建立支持科学与工程数据保存、获取、集成、建立和分析的长期能力,建立可根据科研需要和技术发展动态调整这些能力的机制,示范可互操作的数据保存与获取网络。NSF于2008年启动Cluster Exploratory-CluE项目^[3],支持若干大学探索数据密集型计算的创新方法,不仅要促进计算机科学在数据挖掘利用上推出新理论、新模型、新方法,而且要促进这

种新的计算机科学模式更好地支持科学研究和教育。欧盟、英国等也通过e-Science和e-Infrastructure等计划在积极推进科学数据的建设与利用。

尽管科学界一直在与数据打交道,但“大数据”的洪流也在改变着人们对数据及其作用的认识。当数据海量、泛在化、开放化、网络化和计算化时,它的作用也在发生革命性的变化。正如SUN公司的Martien Quwens指出,“‘多’不仅仅是数量增多,‘多’会带来本质的不同”(More is different)^[4]。在大数据时代,科学家不仅通过对广泛的数据实时、动态地监测与分析来解决难以解决或不可触及的科学问题,更可把数据作为科学研究的对象和工具,基于数据来思考、设计和实施科学研究。数据不再仅仅是科学研究的结果,而且变成科学研究的活的基础和工具;人们不仅关心数据建模、描述、组织、保存、访问、分析、复用和建立科学数据基础设施,更关心如何利用泛在网络及其内在的交互性、开放性,利用海量数据的知识对象化、可计算化,构造基于数据的知识发现和协同研究,因此诞生了数据密集型的知识发现,即科学研究的第四范式。对这一趋势的系统揭示,正是《第四范式》的意义所在。

2 第四范式:大数据带来的科学研究范式革命性变化

《第四范式》一书以吉姆·格雷提出科学研究第四范式的著名演讲开篇,邀请国际多个领域的著名科学家对数据密集型科学发现的理念、应用和影响进行深入探讨,明确提出了数据密集型科学发现范式,并将其作为与实验科学、理论推演、计算机仿真3种科研范式平行的科学研究即第四范式。该书通过地球与环境科学、生命与健康科学、数字信息基础设施和数字化学术信息交流等方面基于海量数据的科研活动的生动实例,揭示了这种新范式带来的革命性变化。

该书第一部分中,Dan Fay等人介绍了地球、环境、海洋、空间等领域的大数据环境与科学应用。在这些领域中,新型传感器实时获得上天入

海、随时随地的海量数据,各种网络传输系统和数据保存与分析设施帮助科学家获得和分析这些数据,从而具有了超越传统科研环境很多倍的“史上最强”观察能力、分析能力甚至实验能力。例如,海洋科学研究中的“全球生命维持系统——海洋”项目,通过有缆海洋观测站,超过1 200公里的海底电缆将为众多海底观察节点提供电力保障和数据传输能力,这些节点以 $24 \times 7 \times 365$ 的不间断方式持续20—30年观察海洋,并以每秒25亿—100亿字节的数据带宽与互联网连接,通过可信赖的数据库加以保存和整理,形成全时制全球化的观察数据集,而且通过将岸基设备与海底高清影像设施、捕获机器人甚至现场分析设施集合起来,可构造起实时的海洋观察网和交互的海洋实验室,供全球的科学家、学生、教师和公众进行观察和开展实验。该方式将彻底改变科学家、公众以及政策制定者与动态的全球海洋之间相互作用的方式,使自然界、互联网和海量数据集共同构成最强大的海洋研究工具。作者们提出,传感器件的微型化、嵌入化、智能化以及与互联网的充分链接,使我们对自然界的感知速度、广度和精度大幅度提高,但如何有效挖掘这种感知能力带来的海量数据中的科学现象和规律,如何支持科学家大范围地、交互协同地对这些现象进行研究,如何提高基于充分信息、可靠证据和科学方法上的科研洞察力和科学决策力,还需要改进认识科学研究的视角和科学工具与方法,才能保证“科研智能”能跟上“感知能力”。

该书第二部分中,Simon Mercer等人分析了医学、认知科学、生物系统、医疗服务等领域的数据密集型科学发现。书中举例,一个在发展中国家部署的医疗知识和数据的采集、分析与传播系统,通过手机界面支持

医护人员在现场记载了每位患者的各类病症、治疗、效果以及身份、历史、位置、家族等信息,并及时传输到中心存储与分析系统。该中央系统又与海量医学文献系统链接,利用大量诊治实例和科研结果进行循证医学的实时诊断,而且还与医学专家网络链接,在必要时引入专家智能提高分析能力。该系统不仅能对紧急或特殊病患做出快速反应,指导乡村医护人员准确诊治,还能根据疾病爆发、治病因素分布、有效疗法所需时间等,提高应对速度和预见力,针对性地调配医疗资源,更准更快地消灭疾病。进一步地,这种机制可基于广泛、持续、实时的数据来挖掘引起疾病发生发展的因素与规律,充实和调整医学知识库,形成虚拟的以证据为基础的医学专家大脑,大幅度提高疾病诊断与治疗的准确性,更不用说,这种全球性的诊断和数据系统,对于加强发展中国家医疗防治能力,保证生命和生存的公平,具有非常重要的意义。当然,仅仅“数据多”还不够,这些医学数据还可加载到那些编码了生物系统基本逻辑的可执行模型上,不仅可用以描述有机体的行为,还能预测其未来状况或对刺激的反应。比如,在神经科学方面,这些模型能帮助我们了解大脑如何思考;在医疗记录方面,能帮助我们了解疾病机理和治疗方法。不过,这需要深化对各种可能关联因素以及关联机制的理解、揭示和描述,才能让海量数据真正发挥作用。

该书第三部分中,Daron Green等人分析了适应大数据时代的科学信息与科学计算基础设施面临的挑战。数据密集型计算不仅仅提供更大规模的数据传输、保存的能力,而且能迅速提供普遍的个人化的低成本、高容量、高效率的存储与计算能力,使得在可预见的不久,个人有可能拥有几年前只有超级计算中心才可能的计算能力、存储能



中国科学院

力甚至个性化的计算云。不仅如此,计算机领域正在开发新的能力,从互联网开源信息、海量科学数据和隐藏在社群交互交流信息中进行知识的发现、获取、组织、分析、关联、解释和推理。科技界也正在迅速建立传播、管理和处理全球知识的基础设施,构建将知识的交换、共享和处理作为所有应用和服务的核心的“知识即服务”(Knowledge as a Service)机制。这样的知识基础设施需要提供恰当的服务集合,不仅要支持知识内容的丰富语义化,支持语义化访问,还要提供对全球知识进行操作的计算服务,并要提供支持科学家们从领域问题出发,发现假说、探查解决路径、“试验”解决方案、预测解决方案对其他因素或在其他应用环境下的可能影响等。数据变成实验室,而且是更真实、更全面的实验室,“数据实验”变成科学的必要部分。

该书第四部分中, Lee Dirks 等人对数据密集型科学发现对学术信息交流带来的深刻变化做了描述。这时,人们以计算可操作的方式创造和传播学术记录,把数据集整合进学术记录,把学术信息交流中产生的各类知识整合起来形成新的“超级”学术记录。原来仅能依赖个人一篇一篇阅读科学文章或者分析科学数据的情况被迅速打破,人们可以一次就对数千篇文章进行“阅读”分析,找出其中的结构、演变与疑难,支持发现那些隐藏在大量结果中的现象关系和科学规律。还可以将阅读、分析与对科学知识的注释、讨论、检验、扩展结合起来,把个人“阅读”与群体“阅读”结合起来,打破知识在微观上的静态局限和个人或小组的认知限制。这样,单篇文章向由数字文献库与科学数据库组合成的“超级”科学记录转移,将实现该书第三部分提到的知识实验室。而且,科技知识的出版、传播走向开放获取,科学家乃至社会公众能迅速获取全社会产生的科学知识,还能参与到协同创新中去创造知识,支持跨学科领域、跨知识创新价值链各环节、跨科研与政策、社会行为等各层次的协同化知识发现和应用。

我们已经看到,科学研究第四范式,确实将带来科学研究的革命。当科研人员可以方便地从宏观到微观、从自然到社会获得海量的实时的观察或实验数据,当这些海量数据普遍地可网络获取、可计算、可开放关联,当对这些数据进行分析、更新、扩展的方法和技术成为科学家和公众的普惠的工具,知识成为可以被及时更新、广泛连接、灵活计算的活的生命体,可个性化地、动态地、交互地、智能化地嵌入到我们的研究、学习、管理和生活中。许多激动人心的潜力将被进一步开发,许多未知的领域和方向将呈现在我们面前。我们期待着科学研究第四范式继续对科学研究和社会发展做出巨大贡献。

3 大数据还在“爆炸”,驾驭第四范式还需加倍努力

第四范式展示的能力和潜力已得到科技界和社会的高度认可,许多国家正在启动各种相关的计划和措施,例如美国白宫2012年宣布大数据资助计划(White House Big Data Initiative)^[5],欧盟委员会提出驾驭大数据浪潮的战略思路(Riding the Wave: How Europe can gain from the rising of scientific data)^[6],日本刚发布《面向2020的ICT综合战略》,提出要构造丰富的数据基础。但我们也要清醒地看到,充分利用大数据机会、驾驭第四范式提供的能力,还存在许多挑战。除了《第四范式》已提出的海量数据计算方法、分布式数据存储与管理等挑战外,这里着重讨论几个常被忽略的挑战。

第一是如何可靠地管理科学数据^[7]。数据是否被系统、客观、准确、精确地采集,数据以及采集技术、方法与环境是否准确描述,数据是否被全面、准确记载,数据在处理环节之间计算或转换时是否失真、是否能可靠溯源,数据在整个项目生命周期中的处理、转换、修改、保存、发布、删除等的规则和责任体系是否建立,数据及其各个“版本”在项目结束后是否可公共获取,等等,这些都直接决定数据最后成为垃圾还是科学知识。目前,相

当多科研项目所产生的数据(尤其是占项目绝大多数的“小项目”产生的科学数据)没有得到详细描述和完整保存,甚至很快就踪影全无。因此,必须将科学数据的管理作为科研管理的重要环节。自2011年,NSF提出数据管理与共享要求^[8],要求项目申请者必须提出数据管理与共享计划,作为项目审查内容之一。英国经济与社会科学研究理事会在2010年制定了数据管理政策^[9],要求申请者说明项目数据的质量控制、共享与保存、知识产权管理等,并由英国政府资助数据管理中心(Data Curation Center)^[10]来系统地研究和提出科学数据管理的政策、指南和最佳实践。

第二是科学数据的互操作^[11]。我们现在有太多的科学数据格式,甚至在同一领域中对同一类数据,也因为种种原因有若干不同的数据格式;但我们对科学数据的描述又往往缺乏细致的元数据,尤其是对科学数据采集、处理、转换、转移过程,几乎没有可靠的元数据进行描述,对科学数据的权属、权利转让、管理要求、使用许可等就更缺乏规范的计算机可读的元数据;对于数据单元和数据集的标识和引用,缺乏广泛认可和可互操作的唯一标识符体系和引用规范;对于科学数据的各种复杂使用的规范的规则才开始建设……。要实现第四范式的潜力,上述障碍必须打破。《科学》杂志2012年刊文^[12],提出健全支持科学数据广泛共享和利用的开放标准体系;英国联合信息系统委员会要求重建数据驱动的基础设施体系^[13],支持数据的可靠保存、交换和利用;有关科研团体提出了开放数据协议^[14],希望提供科学数据库的开放检索标准接口;CODATA专门建立了数据引用标准和实践工作组^[15],希望建立数据集标识和引用的公认规则。

第三是科学数据本身及其共享的权益

管理。科学数据的权益管理涉及两个问题,一是科技界和社会对科学数据的获取、使用和保存的权利。OECD早在2007年就发布了《关于公共资助研究获得科学数据的开放获取的原则与指南》^[16],指出多数由公共资金资助的科学研究,应促进整个科技界、企业和公众对其数据的获取;英国皇家学会2012年发表了《科学是一个开放的事业》报告^[17],提出科学结果的可靠审查、重复验证、可靠利用、扩展应用等都依赖科学数据的开放获取。美国白宫2012年连续推出数据开放利用计划,支持企业和社会创新^[18]。但是,科学数据又是通过大量投入和复杂过程才获得的,有时甚至是长期工作积累起来的重要科研成果,涉及到研究者和投资者在科技竞争、经济利用、长期研究需要等方面的利益,可能还涉及到隐私和保密问题,涉及到谁有权利决定在什么范围以什么形式共享和使用数据。另一个常被遗忘的问题是,数据采集者、处理者、保存维护者是否在科学论文和科研评价中得到认可和激励。在前述OECD和英国皇家学会的报告中,对如何保护公共利益和保护研究者利益提出了原则性的规则,NSF和英国研究理事会等已提出了科学数据管理与共享的原则和规则,欧盟等也在组织对科学数据的著作权和使用权的研究。人们还提出开放数据共用许可^[19]的建议,借鉴创作共用协议(Creative Commons Licenses)提出了一系列可采用的规范权益管理协议。美国科学院召开研讨会讨论科学数据的贡献形式和署名机制^[20]。其实,国际科学出版规范从来就要求,将在研究项目的数据采集和处理中做出实质贡献者作为共同作者,而且《自然》、《科学》等著名期刊已将研究项目专门负责采集、处理、分析数据的人员作为作者,并在作者贡献(Author contribution)中予以说明。当然,如



中国科学院

何承认、评价和激励数据长期监护和公共数据服务,仍是一个常被忽视的问题。

第四是数据素养(Data Literacy)问题。在数据密集型科学发现时代,要能可靠和有效地设计数据采集、管理和共享计划,要掌握好数据采集、处理、转换、传播、保存等的方法、政策与工具,显然是个严峻的挑战。《哈佛商业评论》2012年发文指出^[21],仅仅有好的数据不一定就导致好的决策。多数人或者对数据盲目信任,或者习惯性忽略数据及其构成的证据链,只有少数“知情批判主义者”(informed skeptics)才有可能充分受益于大数据的潜力。英国《卫报》甚至对 *Wired* 杂志提出的大数据导致“理论的终结”提出质疑^[22],认为仅仅依靠数据还不能帮助我们深刻理解自然和社会现象的本质规律,而且仍严重存在的数字鸿沟、不同人群的行为习惯、不同领域或国家的信息习惯与政策等等都会扭曲数据。不过,无论我们相信还是反对“理论的终结”,要在大数据时代可靠地开展科学研究,必须具备可靠的数据素养。欧盟委员会联合研究中心对数字素养进行了分析^[23],认为其覆盖了技术、工具、媒介、内容创作、知识管理、社交网络、对信息与知识的批判性认知、沟通、协作、法律等方面的了解和能力。这虽然不是直接针对数据素养,但它所揭示的技能与知识的广度显然提示我们要重视数据素养,并将其作为科研素养和科学教育的有机组成部分。《哈佛商业评论》甚至提出,数据科学家将拥有 21 世纪“最性感”的工作^[24],因为他们将是实现大数据和第四范式潜力的关键,但培养足够的和足够水平的数据科学家(或者数据图书馆员)并不断激励和支持他们,也是一个重大的挑战。

感谢《第四范式》让我们看到了科学发现和社会进步的巨大机遇,而且科学革命还在提速,大数据及其革命化影响还在深化。我们也充分意识到必须融汇来自科技领域、计算机与网络、图书馆与知识管理等各方面的力量协同创新,突破面临的重要障碍,才能抓住机遇、乘风而上,让第四范式

成为普遍而成功的科学春天。

参考文献

- 1 The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.
- 2 Chris Anderson, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory Sustainable Digital Data Preservation and Access Network Partners (DataNet)http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141.
- 3 NSF. Cluster Exploratory CluE 2008. <http://www.nsf.gov/pubs/2008/nsf08560/nsf08560.htm>.
- 4 Martien Ouwens. The Petabyte Age: More Isn't Just More — More Is Different http://www.ncg.knaw.nl/Studiedagen/09/PointClouds/presentations/PointCloud_15_MartienOuwens.pdf.
- 5 Obama. Administration Unveils "Big Data" Initiative: Announces \$200 Million in New R&D Investments. March 29, 2012. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.
- 6 Riding the Wave: How Europe can gain from the rising tide of scientific data. 2010. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>.
- 7 Christine L Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 2012, 63(6): 1 059-1 078.
- 8 NSF. Data management plan requirements (2011) <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.
- 9 ESRC Research Data Policy, September 2010. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf.
- 10 Data Curation Center. <http://www.dcc.ac.uk/>.
- 11 Pasquale Pagano. Data Interoperability. GRDI 2020 Conference. October, 2011. <http://www.grdi2020.eu/Repository/FileScaricati/c4fb6ab0-d83b-49ae-ab14-6d8030fc2422.pdf>.
- 12 Haak L L et al. Standards and Infrastructure for Innovation Data Exchange. *Science*, 2012, 338:(6 104): 196-197.
- 13 Max Hammond. Preparing for Data-driven Infrastructure. JISC Observatory TechWatch Report. September 2012. <http://observatory.jisc.ac.uk/docs/data-driven-infrastructure.pdf>.
- 14 Open Data Protocol. <http://www.odata.org/>.

- 15 CODATA. Data Citation Standards and Practices Task Group. <http://www.codata.org/taskgroups/TGdatacitation/>.
- 16 OECD. Principles and Guidelines for Access to Research Data from Public Funding. 2007. <http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf>.
- 17 The Royal Society. The Science as an open enterprise. 21 June 2012. <http://royalsociety.org/policy/projects/science-public-enterprise/report/>.
- 18 White House Fueling American Entrepreneurship With Open Data <http://www.whitehouse.gov/blog/2012/10/03/fueling-american-entrepreneurship-open-data>.
- 19 Open Data Commons. Legal tools for open data. <http://opendatacommons.org/>.
- 20 The Board on Research Data and Information. National Academies of Science. For Attribution: Developing Data Attribution and Citation Practices and Standards, 2011. http://sites.nationalacademies.org/PGA/brdi/PGA_064019.
- 21 Good Data Won't Guarantee Good Decisions. by Shvetank Shah, Andrew Horne, and Jaime Capellá. Harvard Business Review. April 2012.
- 22 Mark Graham. Big data and the end of theory? The Guardian. March 9, 2012. <http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory>.
- 23 Joint Research Centre. Digital Competence in Practice: An Analysis of Frameworks. <http://ftp.jrc.es/EURdoc/JRC68116.pdf>.
- 24 Thomas H Davenport and Patil D J. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, October, 2012.

Promote Data-intensive Scientific Discovery, Enhance Scientific and Technological Innovation Capability: New Model, New Method, and New Challenges

Comments on “The Fourth Paradigm: Data-intensive Scientific Discovery”

Liang Na, Zeng Yan

(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract The mass data of scientific research is increasingly becoming a powerful tool of scientific research and knowledge base. Starting from the book “The Fourth Paradigm: Data-intensive Scientific Discovery”, this paper illustrates the vivid research activities based on big data in the fields of Earth and environmental science, life and health sciences, digital information infrastructure and digitization of academic information exchange, reveals the models and new methods of the data-intensive scientific discovery paradigm. Finally, the paper presents the challenges of the fourth paradigm from the scientific data management, interoperability, data rights management, and data literacy.

Keywords The fourth paradigm, scientific data, data management, knowledge service

梁 娜 中科院国家科学图书馆业务处项目主管, 博士。分别于2001、2004年获四川大学管理学学士、硕士学位, 2007年获中科院管理博士学位。多年从事数字图书馆、标准规范等方面的研究, 发表论文10余篇, 参与专著3本。E-mail: liangna@mail.las.ac.cn

曾 燕 中科院国家科学图书馆副研究馆员、业务处副处长。1994年获武汉大学图书情报学院文学学士学位, 2003年获中科院研究生院管理学硕士学位。多年从事文献资源建设、文摘/索引数据库建设, 发表论文10余篇。E-mail: zengy@mail.las.ac.cn



中国科学院