



# 大数据 大科学 大发现

## ——大数据与科学发现

### 国际研讨会综述\*

文 / 郭华东  
中国科学院遥感与数字地球研究所 北京 100094

**【摘要】**“大数据时代”的到来以及数据密集型知识发现方法论为科学研究提供了全新的机遇与挑战。基于此,国际科技数据委员会(CODATA)联合全球6个大型国际学术组织以及中科院遥感与数字地球所于2014年6月在北京举办了大数据与科学发现国际研讨会。本次研讨会会对大数据及科学大数据的本质特征进行了分析,对大数据予大科学研究的知识发现开展了研讨,对大数据予大科学计划的应用提出了建议,并针对大数据服务科学计划使命提出了未来行动纲领。

**【关键词】** 大数据,科学大数据,科学发现,CODATA

DOI 10.3969/j.issn.1000-3045.2014.04.014

## 1 学术背景

随着科学技术的飞速发展和社会需求的强大驱动,并随着数据生产方式的演化及数据的产生成本急速下降,人类产生的数据量正在呈指数级增长。由于数据规模的急剧膨胀,各行各业累积的数据量越来越大,数据类型也越来越繁多、越来越复杂,已经超越了传统数据管理系统和处理模式的能力范围,“大数据”概念近年开始广泛传播。2014年4月,国际数据公司(IDC)发布的第7份数字宇宙研究报告中指出,数据量将以超过每两年翻一番的速度持续增长,2013年全球创建和复制的数据总量已达

4.4ZB,预计到2020年将增长至44ZB<sup>[1]</sup>(图1)。我国拥有的全球数据量比例预计也将由2012年的

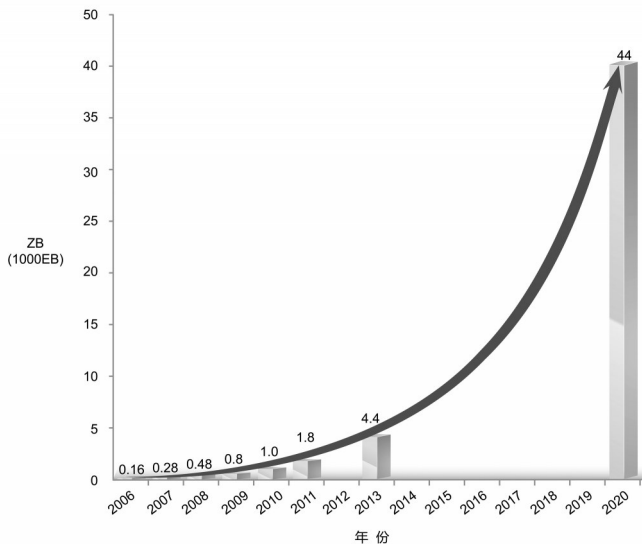


图1 2006—2020年全球数据量增长趋势

\* 修改稿收到日期:2014年6月24日

不过,与其他新兴技术领域所面临的主要问题一样,大数据的基本概念及特点、大数据要解决的核心问题等,目前尚无统一的认识;大数据的获取、存储、处理、分析等诸多方面仍存在一定的争议。此外,人们了解大数据的根源还是因其在云计算、互联网和

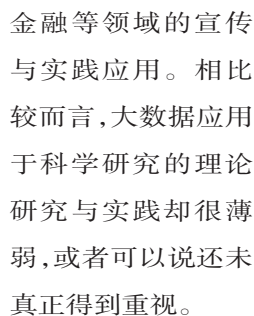
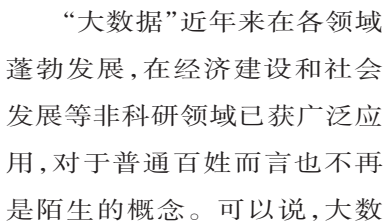


图 鉴于此,国际科学理事会(ICSU)下属跨学科主体、全球最大的科技数据学术组织国际科技数据委员会(Committee on Data for Science and Technology, CODATA)在其第 59 届执委会会议上决定组织召开“大数据与科学发现国际研讨会”,期望国际科技界共同挖掘科学大数据的能量与潜力,探索大数据服务大科



中国科学院院刊 501

学,创造大发现的价值,向全球科技界传递科学大数据对全面推动科技、经济和社会发展的意义。

## 2 大数据与科学发现国际研讨会

2014年6月8—9日,大数据与科学发现国际研讨会在北京举行。该会议由CODATA发起,联合世界数据系统(World Data System, WDS)、未来地球计划(Future Earth)、灾害风险综合研究计划(Integrated Research on Disaster Risk, IRDR)、研究数据联盟(Research Data Alliance, RDA)、地球观测组织(Group on Earth Observations, GEO)和国际数字地球学会(International Society for Digital Earth, ISDE)6个大型国际组织以及中科院遥感与数字地球所共同举办。

研讨会主、协办机构汇聚了国际上数据研究和管理领域最重要的国际组织:CODATA是以发展数据科学、推动科技数据应用、促进科学研究的全球最大的科技数据国际学术组织;WDS的总体目标为确保对科学数据、数据服务、产品和信息的普遍和平等访问;Future Earth的宗旨是研究全球环境可持续发展,是由ICSU等6个国际组织共同发起的为期10年的大型科学计划;IRDR是由ICSU、国际社会科学理事会(ISSC)和联合国减灾战略(UNISDR)发起的一项为期10年的灾害风险综合研究计划;RDA是由美国、欧盟和澳大利亚组建的国际数据组织;GEO是目前国际地球观测领域规模最大、最具权威和影响力的政府间国际组织;ISDE是总部设在中国的数字地球国际学会,是全球唯一的数字地球组织。以上机构的联合,既体现了大数据的独特魅力,也彰显了本次会议的号召力和影响力。

研讨会共设置6场学术报告分会。来自中国、美国、欧洲、日本、澳大利亚、印度等研究单位的学者分别做了各领域科学报告,报告内容涉及计算机科学、地球科学、生命科学及化学科学等。与会专家充分肯定了大数据在全球变化、数字地球、高能物理、计算生物学、环境保护、灾害风险等热点

问题中发挥的重要作用,肯定了CODATA在推动数据科技发展中的积极贡献以及该研讨会的重要性与必要性,认为该研讨会将成为大数据科学发展史上的一座里程碑。同时,专家一致认为大数据是人类共有的资源,也是科技发展的重要财富,是科学研究战略高地。作为大数据的重要组成部分,科学大数据正在使科学世界发生变化,驱动着科学研究进入数据密集型科学发现范式这一全新阶段,为此呼吁国际学术组织、各国政府政策制定与管理者、广大科技工作者共同努力,推动科学大数据在各科学领域的应用与发展。

## 3 大数据与科学大发现

研讨会认为科学大数据是科学发现与知识创新的新引擎,它将改变人类生活及对世界的深层理解。从更为广泛的角度来看,“大数据”及其研究代表着一个信息时代、一个思维方式、一个技术潮流。科学大数据是与科学研究和工程实践相关的“大数据”。

为了更好地研究科学大数据,本次研讨会的数据科学家梳理了其商业大数据、互联网大数据等区别的本质属性和特点。整体看来,科学大数据具有如下的外部特征:(1)从数据内容来讲,科学大数据一般表征自然客观对象和过程;(2)从数据体量来讲,科学大数据在不同学科中存在较大的差异;(3)从数据速率来讲,科学大数据依学科不同,数据速率变化较大,包括高能物理、对地观测等领域的“快”数据和天体演变、地质过程、人类进化等领域的“慢”数据;(4)从获取手段来讲,科学大数据一般来自观测和实验的记录以及后续加工;(5)从分析手段来讲,科学大数据一般是与科学原理模型相结合,形成知识发现的方法,而完全依赖数据分析,抛开科学原理模型的领域与方法并不多见。

通过归纳科学大数据的外部特征,其内部特征也变得相对清晰,主要概括为:(1)数据内容的不可重复性。正如哲学家赫拉克利特的名言“人不能两次踏进同一条河流”,对于一般自然与物理

的客观过程的观测内容具有一定的不可重复性;(2)数据的高度不确定性。由于采用观测和记录等获取手段以及非直接的观测方式和采样手段,导致科学大数据存在高度不确定性<sup>[5]</sup>;(3)数据的高维特性。由于科学大数据面临数据源种类繁多、数学分析手段困难等原因,一般具有高维特性,导致维数灾难的形成;(4)数据分析的高度计算复杂性。由于数据的不确定性、高维特性,以及与科学数据分析相伴随的原理模型的复杂性,导致了科学数据处理的计算复杂性。因此可以说,科学大数据具有不同于一般大数据的显著特征,其内在机理及如何应用于知识发现值得深入研究。

大数据服务大科学研究是一个重要方向。大科学一般是指多学科交叉的大型的基础科学研究项目,具有投资巨大、项目科研人员数目众多、拥有大型科研基础设施以及实验环境的特点。国际上较为著名的大科学计划包括大型强子对撞机、人类基因组计划、地球观测系统、全球变化研究等。大科学计划被认为是现代科学研究的一个成功组织模式,已在若干重大关键科学领域发挥了重要作用。

大科学中的研究是与大数据紧密联系的。这是因为在一般意义上讲,大科学计划能够产生海量的实验数据或者观测数据。在高能物理领域,大型强子对撞机一年可产生15PB的数据。在人类基因测序方面,到2013年,全球范围内至少有30万个人类个体基因组被全部或部分测序,这意味着将产生30PB的序列数据,并需要至少150PB的相应存储和分析的计算能力。全球变化研究作为地球科学、环境科学、生命科学、社会科学和计算科学等多学科交叉的研究,其数据类型更是多种多样,且时间序列超长。预计到2020年,基于地球系统数值模式的全

球变化预测资料的数据量将达50PB,遥感卫星数据也将达50PB,其他类型数据将达到2PB。这一数字预计到2030年将分别上升为185PB、150PB和5PB。

在未来地球计划的八大交叉能力中,观测(Observing)、数据系统(Data systems)、地球系统建模(Earth System Modeling)同大数据密切相关。观测能力中,由卫星、航空、地面、深海等观测网络组成的地球观测系统提供大量的观测数据,具有体量大的特点;数据系统能力不但需要快速获取大量数据,并进行实时处理和分析能力,还需要通过元数据管理和合理的数据政策减少数据质量不确定性;地球系统建模涉及社会、科学模型、对地观测、经济等数据,类型极为丰富。

IRDR通过四大核心项目,促进其科学目标的实现,在其实施中也与大数据紧密相关,如AIRDR项目涉及综合减灾研究的文献库,其具有较长的时间跨度,数据量巨大。RIA项目也包括海量个体和政府的决策行为和灾害信息传播相关的数据,其涵盖社会、经济、心理、减灾技术等方面,类型多样。FORIN项目将开展多区域、多灾种的案例经验分析,并通过快速建模实现对未来灾害的情景模拟。DATA项目在灾害数据的获取、存储、共享政策、标准制定的角度展开工作,有助于数据真实性的提升。

地球观测组织的全球综合地球观测系统(Global Earth Observation System of Systems, GEOSS)中,数据包括长时间跨度的各类卫星观测数据、地面实测数据、各类应用产品库,体量巨大;其类型含有不同应用目的和地表目标的观测数据、服务产品、科学模型、文档材料等,类型丰富。GEOSS也特别注重元数据、数据质量控制、数据共享政策的工作,以保证数据真实性。同时GEOSS在快速获取对地观测数据的同时,



中国科学院



通过GEONETcast、GCI等系统实现各类数据的快速分析和应用。

中科院遥感与数字地球所具有长时间系列的大量对地观测数据存档,数据总量超过450TB;3个遥感卫星地面站及两架遥感飞机可以快速获取不同时间、空间、地物对象的多源对地观测数据;数字地球科学平台具有快速处理分析海量空间数据和知识发现能力;对地观测数据共享平台开展共享元数据、数据质量控制、数据共享政策的工作。

值得注意的是,大科学计划中有相当大的的一部分科学研究属于反问题框架:其对象机理模型极其复杂,模型不确定性和计算复杂性极大,实验数据或者观测数据对于模型发展(即知识发现)具有明显的推动作用。大科学计划的本质决定了其科学发现的意义具有广泛而深远的影响。科学大数据已经并将继续在数字地球、全球变化、高能物理、人类基因组计划、深空探测等大科学领域中发挥重要作用,未来必将在大科学领域为科学发现做出更加重大的贡献。

#### 4 大数据服务于国际科学计划

研讨会认为跨学科国际科学计划为大数据发展提供了显著的发展机遇。这些跨学科国际科学计划旨在通过产生研究成果和数据的方式以完善在人类和环境等关键问题上的决策。为实现此目标,需要集成日益变化的复杂数据集,设计数据收集的新形式,利用复杂统计方法、复杂仿真模型和其他密集型计算方法提取和解译知识。因此可以说,大数据在跨学科国际科学计划中具有极其重要的地位并赋予这些计划以新的内涵。

鉴于此,CODATA联合7个协办组织在研讨会上共同发布了“大数据服务国际科学计划声明”,提出了7项建议:

(1)积极响应大数据服务于国际科学计划的重要性。大数据为知识发现尤其是跨学科的研究方面提供了有目共睹的机遇。应当充分认识其重要性,并积极采取行动。要共同面对并解决因数

据体量、复杂性和不均匀性所形成的挑战,使大数据为国际科学计划带来显著效益;

(2)开发利用大数据的优势用于服务社会。作为国际科学计划的主持机构,需要推进利用大数据促进应用研究的各项活动的开展。大数据的利用与开发需要协调与合作,研究资助者、国家科研院所、高校、数据服务商和其他研究执行机构应制定协调策略以鼓励大数据的发展与应用,从而满足高优先级的科学需求;

(3)通过国际合作提升对大数据的理解和认识。应进一步加强大数据在国际科学计划实践应用中的方法论、理论基础和技术研究。这有赖于汇集众多领域、学科的众多专家的力量,保持强有力的国际合作也将至关重要;

(4)通过全球研究基础设施推动大数据的普及。在国际科学计划中利用大数据产生的知识发现可使全人类受益。然而,实现这一优势将取决于大数据的可参与度和可访问度。对于可进行大数据处理工作的全球研究基础设施,其人人皆可访问的特性所展示的对科学的重大贡献,应当得到重视并使其得到可持续发展,并通过各国及国际组织间的协作得到有力支持;

(5)探索并解决大数据管理的挑战。大数据可重复使用等特性为挑选、保存、记录以及促进有针对性数据产品的传播,提供了重要的激励作用。为了使大数据得到最有效的应用,需要加强对相关数据源的管理、质量评估和不确定性量化。要进一步审视大数据在长期保存和管理中所面临的挑战及巨额成本;

(6)鼓励能力建设和技能培养。在大数据和数据分析的商业潜力已得到国际上广泛共识的形势下,呼吁并倡议合作伙伴与合适的国家及国际组织开展合作,共同促进大数据科学的能力建设和技能培养,包括大数据科学在教育制度和青年科学家职业道路发展的优先考虑;

(7)促进政策发展,最大化开发利用大数据。大数据的出现不断导致在管理、访问和重复使用

中更复杂的新问题出现。应当进一步发展政策、指南、国际条约和协议,最大化完成大数据的收集、共享和潜能开发,更好地服务科学研究。这一行动的推动,需要以国际合作和多学科交叉为基础。CODATA 联合利益相关者可以共同发挥重要作用以支持这种政策发展,从而解决大数据的具体挑战。

同时,该声明也针对 CODATA 在大数据服务科学计划的下一阶段提出了行动纲领,即组建 CODATA 大数据服务科学计划工作组。声明提出 CODATA 应召集和协调一批有坚实基础的学者和专家组成国际工作组以检验大数据问题和促进大数据发展,具体包括:

(1)建立大数据服务国际科学计划的案例研究。更清晰地阐述大数据挑战与机遇的确切性质,与相关国际研究计划建立合作关系,建立一系列的可再生研究案例,关注以数据使用为导向解决科学问题的方法,从而有效帮助提高对大数据的理解及促进其共享;

(2)推动跨学科间的大数据应对举措的共享。研讨会对一些可行的技术、基础设施和分析解决方案进行了认定,其中包括借鉴和应用参考模型及在研究背景与不同学科间的解决方案转换。工作组应与国际科学计划和其他合作伙伴一同促进机制的完善,讨论、宣传、共享和采用适当的解决方案。在跨学科间的和社会相关研究的大数据发展路线图应是此工作组的主要产出;

(3)大数据的研究政策、伦理道德和法律问题。在从事大数据政策议程制定过程中的特定问题必须详尽阐述,包括利用观察或传感器网络、社交媒体和涉及人类活动、伦理道德和法律问题等其他数据获取手段。大数据不仅在体量上有别于小规模数据集,在特性上也有特定的区别,其所提出

的新的科学和数据政策问题有待鉴别和解决。例如在不同尺度内海量数据集成的实现依然存在着许可和访问的问题。政策议程制定应旨在消除不恰当的壁垒、认知或其他反面作用,从而充分地使用和重复使用大数据源;

(4)大数据研究管理和可持续性挑战。与 WDS 和其他合作伙伴联合共建工作组,帮助协调涉及大数据产品管理服务国际科学计划中面临的长期保存以及可持续数据基础设施建设的双重挑战<sup>[6]</sup>。

## 5 结语

大数据研究的重要意义毋庸置疑,“大数据时代”的到来以及数据密集型科学发现范式的确立,为现代科学提供了全新的科研方法论<sup>[7]</sup>。

本次研讨会对以下内容进行了热烈讨论并达成一定程度的共识:相对于互联网、社会、商业大数据,科学大数据需要深入的研究。从研究对象层面上讲,科学大数据尚未清晰定义和分析;从研究目的层面上讲,大数据研究一般追求“相关性”研究,这与科学研究中的“因果性”知识发现存在着巨大矛盾;从研究方法层面上讲,尽管数据密集型科学发现范式的学术思想已被广泛认可,但具体的大数据科研方法论尚未完全建立。

在研讨会提出的“大数据、大科学、大发现”的科学大数据研究思路下,一个值得注意并预期能产生显著效果的方法论是在反问题的框架下,对科学大数据综合应用智能方法、统计方法和信息论方法,优化大科学研究中的模型空间和数据空间,基于高性能计算和云计算平台,进行大科学的知识发现。这样的思路已经在国外近期进行的科学大数据研究中初步得到实施,并预期获得显著成果。

综上,一方面科学大数据具有重大的研



中国科学院

究意义,另一方面当前科学大数据研究中在研究对象、研究目的和研究方法等方面均需要深入研究。科学大数据研究亟需学科顶层设计和方法论指导。大数据与大科学研究需要一支高水平的队伍。要加强国际合作,发达国家应为发展中国家提供科技支持,同时各国要从国家层面推进中长期规划和政策的实施。

**致谢** 衷心感谢王力哲研究员、梁栋先生为本文付出的辛勤劳动。

#### 参考文献

- 1 Turner V, Gantz J F, Reinsel D et al. The digital universe of opportunities: rich data and the increasing value of the internet of things, Framingham: IDC Analyze the Future, 2014.
- 2 Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. Framingham:

IDC Analyze the Future, 2012.

- 3 Special issue: Big data. Nature, 2008, 455(7209): 1-136.
- 4 Jonathan T O, Gerald A M, Sandrine Bony et al. Special online collection: dealing with data. Science, 2011, 331(6018): 639-806.
- 5 Kennedy M C, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B(Statistical methodology), 2001, 63(3): 425-464. DOI:10.1111/1467-9868.00294.
- 6 CODATA. Big data for international scientific programmes: Challenges and opportunities A statement of recommendations and actions. Beijing: Committee on data for science and technology, 2014.
- 7 Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery. Redmond, Washington: Microsoft Research, 2009, ISBN:978-0982544204.

## Big Data, Big Science, Big Discovery

### ——Review of CODATA Workshop on Big Data for International Scientific Programmes

Guo Huadong

(Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China)

**Abstract** The arrival of the “big data era” and a data-intensive knowledge discovery methodology has provided new opportunities and challenges for scientific research. Under such circumstances, in June 2014 CODATA together with six other large international academic organizations and the Chinese Academy of Sciences Institute of Remote Sensing and Digital Earth held an international workshop in Beijing on big data and scientific discovery. The workshop produced analyses of the essential features of big data and scientific data, and reviewed the research on knowledge discovery in big data. Participants looked to the future with suggestions for applications of big data to big science, and discussed future plans of action aiming at how big data serves the scientific mission.

**Keywords** big data, scientific big data, scientific discovery, CODATA

**郭华东** 中科院院士,中科院遥感与数字地球所所长,研究员。主要从事雷达遥感信息机理、多模式遥感信息地物识别方法、空间信息前沿技术研究。发表论文300余篇,获国家及省、部级科技奖励13项。现任国际科技数据委员会主席、《国际数字地球学报》主编、“973”项目和大科学工程项目首席科学家。E-mail:hdguo@radi.ac.cn