浦江创新论坛 研究报告

（2021 年第 4 期，总第 139 期）

上海浦江创新论坛中心　　2021 年 6 月 16 日

2021 浦江创新论坛专题简报之四

科学数据的管理、共享与应用

编者按：2021 浦江创新论坛新兴技术论坛——“数生万物”科学数据创新大会上，来自国内外的知名专家学者立足科学数据管理共享和应用服务的创新理念和实践，分享了科学数据开放、共享、应用、交互、协作各方面的优秀成果，并共同围绕数据开放的政策、原则和体系等展开深入研讨。本文由嘉宾[1]报告整理而成，供参考。

[1] 嘉宾包括：中国科学院院士、博士生导师，俄罗斯科学院外籍院士、芬兰科学与人文院外籍院士，发展中国家科学院院士、国际欧亚科学院院士，中国科学院空天信息创新研究院学术委员会主任、研究员郭华东，复旦大学大数据学院院长、类脑智能科学与技术研究院院长冯建峰，中国极地研究中心副主任、国家极地科学数据中心主任徐韧，中国科学院脑科学与智能技术卓越创新中心副主任孙衍刚，中国科学院计算机网络信息中心科技云部主任、CODATA 副主席黎建辉，中国科学院计算机网络信息中心工程师姜璐璐，CODATA 国际数据委员会主席、Fair 原则创始人、荷兰莱顿大学教授 Barend Mons，EGI（European Grid Infrastructure)欧洲网格基础设施解决方案总监 Gergely Sipos 等。

2021 浦江创新论坛专题简报之四

科学数据的管理、共享与应用

近年来，科学技术发展呈现出明显的大科学、定量化研究特点，科技创新越来越依赖于大量系统、高可信度的科学数据以及对科学数据的综合分析和挖掘。**与会嘉宾指出，全球科学数据共享网络与机制建设是大数据时代十分紧迫而重要的任务，是实现数据更深层次的价值及构建科学数据创新生态的关键。**

一、科学数据已成为科技创新发展新引擎

一是数据是解决复杂问题的科技钥匙。科学研究步入数据密集型的“第四范式时代”，没有数据很多问题难以解决。据**中国科学院院士、中国科学院空天信息创新研究院学术委员会主任、研究员郭华东**介绍，联合国曾提出的变革世界的 17 个可持续发展目标中有 41%处于“有方法、无数据”状态。**中国极地研究中心副主任、国家极地科学数据中心主任徐韧**指出，从北极航道缔结国际条约、南极罗斯海新站立项到“雪龙 2”号极地科考船的研制，无不需要以庞大的科考数据为基础。

二是数据赋能已在多个场景释放巨大潜力。大数据时代，数据赋能正在加速科技创新，形成显著的拉动效应、放大效应和乘数效应。**复旦大学大数据学院、类脑智能科学与技术研究院院长冯建峰**指出，生物大数据是智能医疗的基础，智能医疗的终极目标是精准预测个体的身体与精神健康状况。目前，其研究团队已可通过步态识别判断抑郁症（准确率超过 70%），并可基于所研发的软件系统，突破传统脑卒中患者发病时间难准确判断的诊疗障碍，通过脑影像精准判断病人可否进行溶栓手术。据**郭华东院士**介绍，其研究团队用数据证明了

中国对全球土地退化零增长做出了最大贡献，并发现 1999 年至 2018 年全球冰川储量减少了 6%，等效于海平面高度上升了 12mm。

二、数据生态打造催生科技成果产出新动能

一是人和机器都能理解的数据生态是未来发展的关键。国际数据委员会（CODATA）副主席、中科院计算机网络信息中心科技云部主任黎建辉强调，未来随着技术的发展，机器对数据的自动获取与理解将成为提升数据应用效能的核心关键，需跨越学科边界，构建人和机器都能理解和操作的数据生态。**CODATA 国际数据委员会主席、Fair 原则创始人 Barend Mons 教授**认为，现阶段 Fair（FAIR data principles）除了原先所包含的四项原则（数据可找寻、可访问、可交互、可利用）外，还应涵盖数据完全能够为 AI 所应用。

二是数据全自动化处理是未来努力的方向。中国科学院脑科学与智能技术卓越创新中心副主任孙衍刚表示，目前神经元重建主要基于手工或半自动化模式操作，需要耗费非常大的时间和人力成本。面向数据量庞大且结构极其复杂的小鼠及猕猴全脑介观神经联接图谱绘制研究，亟需更深入地应用人工智能深度学习等新兴技术展开三维数据识别，实现神经元全自动化重建。冯建峰院长表示，基于脑科学大数据的脑血管影像分析在抑郁症、自闭症等传统疾病智能诊疗方面已展现出巨大应用潜能，有望形成高价值突破。

三是数据智能化集成服务系统是数据共享体系建设的基石。《国家科学数据管理办法》发布以来，20 个国家科学数据中心的建设联合推动了不同学科领域科学数据的汇交采集、存储管理、加工挖掘和开放共享等工作的开展。**郭华东院士**认为，数据共享亦要实现不断创新，要摆脱“拷 U 盘式”的传统模式，构建数据、计算、服务

一体化的数据智能服务系统，对不同认识程度、不同理解程度、不同领域用户都可以提供不同的服务方式和数据共享形式。**中国科学院分子细胞科学卓越创新中心研究员、生物信息学平台主任石建涛**认为，数据除了可发现、可利用、可交互操作外，存算亦要一体，如何从用户角度出发，提供给用户更好的计算便利还有一些路要走。

三、数据发展面临机遇和挑战并存的新局面

一是多领域、跨区域的数据交叉融合应用前景广阔。据**Barend Mons 教授**介绍，多学科的数据融合能在很大程度上推动更高价值的成果产出，提升原始数据价值。基于他们所建立的完全 AI 支持的全球新冠数据库，研究人员可利用算法通过机器分析，在无需做临床试验的情况下，快速判断某个药物或一个新的药物是否适合用在病人身上。据**黎建辉主任**透露，未来十年，CODATA 将以实现多领域数据交叉融合、共享、应用为目标，建立跨领域合作网络，支持未来多学科交叉研究，尤其是面向流行病、气候变化、碳达峰、碳减排、SDG 等重大问题。

二是数据开放共享的激励机制仍需进一步完善。现阶段，全球科学界还缺乏有效机制与技术实现互联互通与资源共享。**中国科学院计算机网络信息中心工程师姜璐璐**认为，目前科研一线工作者的数据共享氛围尚弱，数据伦理规范的研究及实践还不够深入，大体量数据开放共享与国际间数据交流仍存在问题，应积极探索数据开放共享的激励机制，强化科学数据引用文化建设。**中科院计算机网络信息中心大数据部副主任，国家基础学科公共科学数据中心主任，CODATA 中委会秘书长胡良霖**认为，科学数据开放共享基本上很难处于标准先行期，应呼吁大家尽可能地利用国内国际现有行业标准，促进数据汇

聚所产生的价值效应的进一步放大。**EGI 欧洲网格基础设施解决方案总监 Gergely Sipos** 表示，给科学家提供激励，让他们能够使用彼此的数据，是非常具有挑战性的，而如何准确地描述数据是目前要解决的关键问题。

三是基础设施建设是数据开放共享成功与否的关键要素。中国极地研究中心国家极地科学数据中心副研究员吴立宗表示，数据中心建设人才非常稀缺，而人才的培养基本上和大的基础设施或者研究机构紧密相关。其中，基础设施的概念绝不能仅仅等同于硬件、计算机、服务器或云平台，把数据、标准规范、硬件聚合在一起提供服务的才叫基础设施。**Gergely Sipos 总监**认为，从根本上来说，必须要有一个将不同计算及云的方式结合在一起的超级平台，通过以最合适的方式对科学应用进行集成，来满足那些复杂的、混合的现实应用需求。

整　理：郑　奕