科学大数据集成共享进展及面临的挑战

诸云强[1,6,7] 潘 鹏[2,3] 石 蕾[4] 孙 凯[1,5] 王筱萱[1]，杨 杰[1,5]

（1.中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室，北京 100101；2.环境保护部环境工程评估中心，北京 100012；3.国家环境保护环境影响评价数值模拟重点实验室，北京 100012；4. 国家科技基础条件平台中心，北京 100862；5.中国科学院大学，北京 100049；6.江苏省地理信息协同创新中心,江苏南京 210023；7.白洋淀流域生态保护与京津冀可持续发展协同创新中心，河北保定 071002）

摘要：科学大数据集成共享既是数据密集型现代科学研究获取数据的重要途径，也是科学数据自身价值发掘和提升的必然选择，更是国家政策的顶层要求。在分析科学大数据内涵和特征的基础上，总结科学数据集成共享主要进展，指出科学大数据集成共享面临整合集成机制、集成共享质量控制、关联集成与语义搜索、数据产权与共享安全、数据高效利用等5个方面的问题并分别提出应对策略。

关键词：科学大数据；数据共享；关联集成；数据质量；开放安全

中图分类号：TP391.7 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.05.001

Progress and Challenge of Scientific Big Data Integration and Sharing

ZHU Yunqiang[1,6,7], PAN Peng[2,3], SHI Lei[4], SUN Kai[1,5], WANG Xiaoxuan[1], YANG Jie[1,5]

(1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101;2. Appraisal Center for Environment and Engineering, Ministry of Environmental Protection, Beijing 100012;3. State Environmental Protection Key Laboratory of Numerical Modeling for Environment Impact Assessment, Beijing 100012; 4.Center of National Science and Technology Infrastructure, Beijing 100862;5.University of Chinese Academy of Sciences, Beijing 100049;6. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing 210023; 7. Collaborative Innovation Centre for Baiyangdian Basin Ecological Protection and Jingjinji Regional Sustainable Development, Hebei University, Baoding 071002)

作者简介：诸云强（1977—），男，中国科学院地理科学与资源研究所研究员，主要研究方向：地学数据本体与共享，资源环境信息系统；**潘鹏**（1985—），男，环境保护部环境工程评估中心助理研究员，主要研究方向：地学数据集成共享理论与技术（通讯作者）；**石蕾**（1982—），女，国家科技基础条件平台中心副研究员，主要研究方向：科技资源管理；**孙凯**（1990—），男，中国科学院地理科学与资源研究所博士研究生，主要研究方向：地学本体及数据关联；**王筱萱**（1983-），女，中国科学院地理科学与资源研究所工程师，主要研究方向：研究方向：科学数据共享与集成；**杨杰**（1990—），男，中国科学院地理科学与资源研究所硕士研究生，主要研究方向：地学模型数据匹配方法。

基金项目：科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”（2013FY110900）；国家自然科学基金重点项目“网络文本蕴含地理信息理解与知识图构建”（41631177）；贵州省公益性基础性地质工作项目“贵州省岩溶地下水系统功能可持续利用性研究”（黔国土资地环函〔2014〕23号）；贵州省公益性基础性地质工作项目“贵州省国土资源可持续发展战略研究”（黔国土资源函〔2016〕269号）。

收稿时间：2017年7月14日。

Abstract: On the basis of analyzing the connotation and characteristics of scientific big data, this paper summarizes the main research progress of scientific big data integration and sharing, points out 5 aspects of the problem that are integrated integration mechanism, integrated shared quality control, associated integration and semantic search, data property rights and shared security, efficient data utilization, and puts forward coping strategies.
Keywords: scientific big data, data sharing, associative integration, data quality, open and safe

科学数据是指人类在认识世界、改造世界的科技活动中所产生的原始性、基础性数据，以及按照不同需求系统加工的数据产品和相关信息[1]。科学数据是关系到科技进步与创新能力、社会经济发展与管理决策的宝贵国家财富和重要战略资源[2−3]。随着地基监测、对地观测、深地（空）探测，特别是移动互联网、云计算、物联网和社交网络等技术的迅猛发展和深入应用，科学数据的采集、处理、传输变得越来越容易和快捷，科学大数据的新纪元已经到来。科学研究已经从几千年前的直接观察、几百年前的理论方法、几十年前的计算仿真，进入到第四阶段“数据密集型研究”[4]。这一阶段的特点就是依靠海量的科学数据，从表面上看起来毫无关联的大数据中发现在传统小规模数据中无法发现的隐含在科学大数据背后的规律和知识。

尽管科学大数据的采集、处理、传输变得越来越容易和快捷，但大部分的数据仍然由少数权威行业部门、科研机构以及科研项目所拥有，因此，科学大数据的集成共享仍然是当前科学大数据挖掘利用的基础和前提。为了促进科学数据的集成共享，早在上世纪 50 年代，美国、英国等发达国家，国际科联科技数据委员会、世界数据中心等国际组织就启动了一系列的行动和计划。2014 年内罗毕发展中国家科学数据共享国际研讨会发布了“发展中国家数据共享原则”（又称“内罗毕数据共享原则”）。2015 年国务院印发的《促进大数据发展行动纲要》明确要求，构建科学大数据国家重大基础设施，实现对国家重要科技数据的权威汇集、长期保存、集成管理和全面共享。科学大数据集成共享既是数据密集型现代科学研究获取数据的重要途径，也是科学数据自身价值发掘和提升的必然选择，更是国家政策的顶层要求。

面向科学大数据集成共享的迫切需求，本文首先阐述科学大数据的内涵与特征和科学大数据集成共享的主要进展，然后分析提出科学大数据集成共享面临的主要问题，最后提出相应的应对策略，以期在为后续科学大数据集成共享研究与应用提供参考和借鉴。

1　科学大数据内涵与特征

大数据是指无法在可容忍的时间内用传统IT技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合[5]。科学大数据则是与科学研究相关，反映和表征自然和社会科学现象及其关系的大数据。它既是支撑科学研究的重要基础，也是科学研究的重要产物和成果[6]；既有一般科学数据和大数据的特征，也有其自身独有的特征（图 1）。

科学大数据具有一般科学数据的所有特征，包括客观性、分离性、长效性、不对称性、非排他性、可传递性、增值性等[7]。然而，作为大

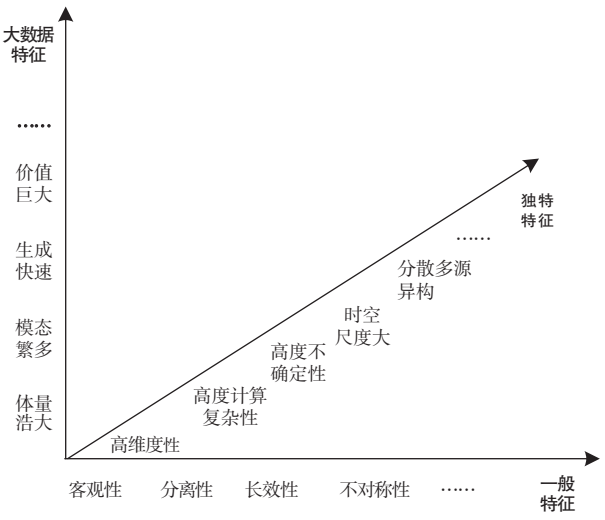

图 1　科学大数据特征

数据的一种，科学大数据还具有通用大数据具有的4V特征，即体量浩大（Volume）、模态繁多（Variety）、生成快速（Velocity）和价值巨大（Value），但密度很低[5, 8]；科学大数据的独特特征表现为：高维度性、高度计算复杂性、高度不确定性和时空尺度大、分散多源异构等[9]。高维度性是指科学大数据反映和表征着复杂的自然和社会科学现象与关系，而这些自然现象或科学过程的外部表征一般具有高度数据相关性和多重数据属性[10]；高度计算复杂性是指科学大数据应用的场景大多属于非线性复杂系统，具有高度复杂的数据模型，因而科学大数据计算问题不仅仅是一个数据处理与分析的问题，还是一个复杂系统与数据共同建模及计算的问题[11]；高度不确定性是指科学大数据的来源一般包括对自然过程的感知和科学实验数据的获取，这两种数据来源的特点决定了科学大数据普遍具有高度不确定性[10]；时空尺度大是指科学大数据由于研究对象的不同，其覆盖的时间和空间范围往往较大，在时间尺度上包含有从瞬间的地震暴发数据到上百万年的地质演变数据，在空间尺度上包含有从单点的水质监测数据到全球范围的气候变化数据等；分散多源异构是指科学大数据往往分散在从事科学研究的科研院所、高等学校的科研团队、科学家个人手中，具有不同来源、不同类型格式等特征。从数据管理和利用的视角来看，科学大数据具有不同的投资方式、产生方式、数据内容、数据类型、管理主体和服务定位。

在投资方式方面，科学大数据可以是由国家和地方财政、单位自主经费，也可以是由企业或个人经费等方式进行投资；在产生方式方面，科学大数据可以是由地面观测（监测）、考察调查、对地观测、对空探测，也可以是由统计分析、实验试验、计算模拟，甚至是由互联网挖掘、志愿数据采集等方式产生；在数据内容方面，科学大数据包括科学数据集、图集、志书/典籍、标本资源（样品、标本）和标准物质等内容；在数据类型方面，科学大数据分为空间数据（矢量、栅格等）、非空间数据（数据库表、数值文本、统计图等）或多媒体数据（文档、图片、音频、视频等）等类型；在管理主体方面，科学大数据可以是由专业机构（数据中心）、科研团队或科学家个人等不同主体管理；在服务定位方面，科学大数据可以是研究型数据（研究项目产生的数据）、资源型数据（特定领域公共的数据库）或参考型数据（长期积累的基础性数据）[12]。科学大数据的上述特征和属性，决定了科学大数据集成共享的复杂性、困难性和长期性。

2　科学大数据集成共享主要进展

2.1　科学大数据共享计划/规划

美国政府认为，全国范围的大数据创新生态系统能够帮助美国充分利用大而繁杂的数据集所创造的新机遇[13-14]。为此，2012年美国公布了“大数据研发计划”（Big Data Research and Development Initiative，BDRDI），开发大数据收集、存储、维护、管理、分析和共享核心技术，并将提高和改进人们从海量和复杂的数据中获取知识的能力作为BDRDI的重要目标[15-16]。BDRDI得到了美国卫生研究院（NIH）、国防部（DOD）、能源部（DOE）等15个不同领域的联邦部门和机构共同参与，并将鼓励数据分享和管理的相关政策以提高数据价值，正确处理大数据收集、共享和使用过程中的隐私问题、安全问题和伦理问题等科学大数据共享相关内容列为工作重点[14-15]。

为整合各成员国的科研力量，提升欧洲总体研究水平，欧盟1984—2013年实施了7期框架计划。最近一期的第七框架计划（7th Framework Programme，FP7）实施周期为2007—2013年，包括了合作计划、原始创新计划、人力资源计划、研究能力创新计划4个专项计划。其中的研究能力创新计划主要包括加强基础学科研究、建设知识区域、提高欧洲的研究潜力、加强国际合作等7项内容，该计划将科学大数据集成共享纳入到计划范围内，并启动了包含科学大数据集成共享内容的全球科学数据基础设施建设项目GRDI 2020（Global Research Data

Infrastructures）。继GRDI 2020之后，欧盟还在2014年正式编制并启动了新的研究与创新框架计划“地平线2020”（Horizon 2020），该计划旨在帮助科研人员实现科研设想，获得科研上新的发现、突破和创新，促进新技术从实验室到市场的转化。Horizon 2020确定了基础科学、工业技术和社会调整3个共同的战略优先领域。其中，基础科学领域下属的欧洲基础研究设施建设行动计划将e-基础设施建设作为重点内容，e-基础设施建设通过整合不同的设备、服务、数据源以及广泛的跨国合作，促进欧洲的研究与创新潜力的发展[17]。Horizon 2020对整合欧盟各国的科研资源、推进科学大数据共享、提高科研效率、促进科技创新发挥着积极作用[18]。

2.2 科学大数据共享典型项目

2015年，美国商务部宣布启动国家海洋与大气管理局（National Oceanic and Atmospheric Administration，NOAA）大数据项目。NOAA每天收集来源多样、内容多元的数据超过20Tb，数据主要来自多普勒雷达系统、气候卫星、浮标网络和浮标站、验潮仪、实时气候站、船舶、飞行器以及超级计算机等，包括气候变化、海上情况、潮汐变化等内容。NOAA通过大数据项目创建开放平台，使决策者和行业人员快速、有效地获取到相关数据，并帮助私有行业、学术界和个体创新者通过云服务访问前所未有的大规模数据[3, 19]。

GRDI 2020项目是由欧盟第七研发框架计划FP7资助的构建科学数据基础设施项目。该项目旨在于2020年实现全球科学数据基础设施建设的战略愿景[20]。2012年3月，GRDI 2020发布《全球科研数据基础设施：大数据的挑战》报告，指出科学是一项全球性事业，而科研数据是全球的资产，因此，需要全球科研数据基础设施来克服语言、政策和社会的障碍，并减少地理时空和国家间的壁垒，从而更加方便地发现、访问和利用数据[3]。同时，GRDI 2020提出了构建全球科学数据基础设施所面临的主要挑战和必须解决的问题，包括跨学科的开放科研和开放数据的原则，科学组织各方面可能面临的冲突，科学数据生态系统、统一规范定义的数据模型和查询语言，科学数据和文件之间的互操作，海量数据的管理、集成、发现和传输工具等。

目前，我国也有科学大数据共享相关项目。2001年，科学数据共享工程启动气象科学数据共享试点，在资源环境、农业、人口与健康、基础与前沿等领域共24个部门开展了科学数据共享工作，启动了9个科学数据共享试点，开展了科学数据共享政策法规、技术标准体系的调研和编制工作，整合了跨部门跨领域国家投入产生的数据资源，并开展了科学数据共享服务[21]。2003年，科技部、财政部共同设立了科技基础条件专项建设平台（简称“科技平台”），科学数据共享工程作为重要组成部分纳入科技基础条件平台建设。科技平台由研究实验基地和大型科学仪器设备共享平台、自然科技资源共享平台、科学数据共享平台、科技文献共享平台等23家国家科技平台构成，其宗旨是充分运用现代技术，推动科技资源共享，促进全社会科技资源优化配置和高效利用，提高我国科技创新能力。其中，科学数据共享平台以政府资助获取与积累的科学数据资源为重点，整合相关的主体数据库，构建集中与分布相结合的国家科学数据中心群，形成国家科学数据分级分类共享服务体系[22]。2013年，为应对大数据时代的挑战，国家发展改革委员会和中国科学院联合启动了“基础研究大数据服务平台应用示范”项目，构建基础研究大数据服务平台，实现基础研究大数据汇聚融合、开放共享与高效处理，为科研工程技术人员和社会公众提供在线基础研究大数据的集成共享与知识发现服务，并在天文、材料领域开展应用示范[23]。

2.3 科学大数据共享政策机制

为提高所有科技领域内重要数据的质量，增进数据的可靠性，改进数据的管理，扩大数据的可获取性，国际科学联合会（International Council for Science，ICSU）牵头成立了国际科技数据委员会（Committee on Data for Science and Technology，CODATA）[24]。CODATA通过任务

组、工作组、委员会或其他针对特定数据问题小组开展国际共享合作，利用互联网构建了全球范围内的科学数据交换体系，面向科学家和工程师提供数据共享服务[25]。CODATA确定了包括科技数据应用与共享中心在内的6个科学技术数据领域的前沿问题，并先后确定了亚洲—太平洋国家数据资源共享、发展中国家科技数据保护与共享、全球物种数据共享等11项任务作为国际合作的共同行动计划，并且这些计划已经获得了联合国教科文组织（UNESCO）、国际科学技术信息委员会（ICSTI）、国际科学联合会（ICSU）、世界知识产权组织（WIPO）等相关国际组织的支持[26]。CODATA在解决当前科学数据共享的主要问题以及协调参与国家与组织的行动方面发挥了重要的作用。

世界经济合作与发展组织（Organization for Economic Cooperation and Development，OECD）也认为政府和研究机构应该对数据、信息和知识的获取条件予以更多的关注，倡导要建立公共资金资助的研究数据获取机制[3]。2006年，OECD颁布了《关于公共资金资助的研究数据获取的原则与指南》，提出了指导成员国制定、完善科学数据共享政策的13项原则，包括开放性、灵活性、透明性、法律一致性、保护知识产权、正式性、专业性、协作性、保障质量、安全性、效率、评价和持续性[27-28]。

美国首先是通过国家法律法规政策的强制驱动来推进科学数据的共建共享，相应出台了《信息自由法》和《版权法》，公布了以“完全与开放”科学数据共享政策为核心的“全球变化研究数据管理政策”，从而为美国科学数据共享活动提供了法律依据和政策保障[29]，并在此基础上开展了BDRDI、RDA、NOAA等一系列数据共享计划和项目。

为了满足全球对数据基础设施日益增长的需求，美国国家科学基金会于2013年资助推出了国际研究数据联盟（Research Data Alliance，RDA），RDA着眼于研究者和创新者们跨技术、跨学科以及跨国界公开共享数据，旨在建立使数据实现共享的社会桥梁和技术桥梁[30-31]。在构成成员方面，RDA由志愿者和个人合作组成，任何个人或机构只要愿意遵守RDA的开放、协商决策、技术中立、均衡代表各方利益等基本原则，就可以加入联盟。在工作任务方面，RDA主要开展5个方面的工作：生物学、农业、社会科学、工程等领域的科学数据共享、数据归档和出版、科研和教育数据共享与重复利用、数据引用参考、数据管理、集成、共享等基础设施建设。在任务实施方面，RDA由工作组、兴趣组和合作组组成。其中，工作组致力于短期内实现特定的工具、代码、最佳实践、标准等；兴趣组负责在更宽广的范围和更长时间尺度上确定常见问题及兴趣，而这些工作最终导致更多焦点合作组的创立[31-32]。RDA与日本科学技术振兴机构于2016年在日本东京举办议题为“在开放科学时代促进数据共享”的第七届大会，会议认为开放科学从基础面上来说是由个别团体主导的，需要依赖免费知识共享以及获取工具和服务[33]。RDA符合全世界对数据共享的需求，加速了基于大数据的创新[31]。

2.4　科学大数据共享数据中心

国际科学联合会（ICSU）成立了世界数据中心（World Data Center，WDC）专门从事数据收集、交换、服务和共享等工作[26]。WDC重点在地球科学、空间科学和环境科学领域推进数据集成和共享。自国际地球物理年（1957—1958）创立以来，WDC在全球已经建立了50余个学科数据中心，各中心之间的数据交换和共享建立在互惠互利的基础上，每个数据中心整合集成了本国该领域中的权威数据资源，并以不同的形式提供给各国科学家使用[34-35]。WDC倡导的科学数据开放共享的理念和做法影响巨大，为地球科学和相关学科的发展提供了大量的数据支撑服务[36]。

作为领域内权威的政府间国际组织，地球观测组织（Group on Earth Observations，GEO）制定并通过了十年执行计划，旨在建立一个综合、协调和持续的全球地球综合观测系统（Global Earth Observation System of Systems，GEOSS），

在灾害、健康、能源、气候、天气、水、生态系统、农业和生物多样性等 9 个社会发展领域，为各国决策者提供数据产品和应用服务[3, 37−38]。GEOSS对于推进全球观测数据集成共享发挥了重要作用。总体来看，当前科学大数据集成共享具有一些共同特点：在组织管理方面，通常由一个国家（组织）牵头（发起），其他成员在接受一些基本原则或协议后可以共同参与，一般由政府或组织自上而下启动而很少由科技工作者自发组织；在投资方式方面，通常由政府或基金投资建设并维护；在数据内容方面：共享的数据通常具有公有性（由国家资助的科研产生）、公益性（对大多数相关成员均有利）、基础性（对社会经济有基础支撑作用）和领域性（通常是特定领域的数据共享）[26]；在开放共享模式方面：通常以促进区域科技创新为宗旨，不以盈利为目的，带有公益性质；在数据共享服务方面：主要以提供源数据为主，对数据关联集成、挖掘利用还不够深入和广泛。

3 面临的挑战与应对策略

3.1 科学大数据整合集成机制

由于科学大数据主要分散在科研机构、高等院校、科研项目团队和科学家个人手中，因此采用何种机制来整合集成科学大数据以持续为科学大数据注入“新鲜血液”是首先要解决的关键问题。

传统科学数据整合集成主要采用自上而下，由国家统一规划、出台政策，并通过强制性数据汇交或者奖补经费的方式进行。如国家科技基础条件平台在仪器设备、自然资源、科学数据、科技文献、实验基地和检测资源等领域，部署认定了 23 家国家平台，每年通过考核评估，进行经费后补助，支持推动分散科技资源的集成与共享服务；科技基础性工作专项通过数据汇交管理制度，要求各项目验收前必须完成数据汇交工作等。该模式通过政策的约束和稳定经费的支持，可以有效保障国家财政投资形成的科学数据的持续性、系统性集成共享，然而也存在难以激发科研人员积极性、评价数据提供者贡献困难等问题。在大数据时代，必须在现有的自上而下数据整合集成机制的基础上，探索一条自下而上的能够充分调动科学家个人积极性的数据整合集成机制，形成“人人都是数据使用者和贡献者”的志愿共享数据的氛围。

科学数据出版作为一种新的数据集成与开放共享机制，正在引起全球科研人员的广泛关注[2, 39]。类似于论文出版，科学数据出版通过数据投稿、同行评审、发表出版、共享引用等，可以明确数据成果的署名，让科学数据也能够被正式引用，并最终纳入科研考核体系中，从而有效保障共享数据的科研人员的根本权益，激发科研人员志愿参与数据共享。因此，科学大数据的整合集成在保持现有国家科技条件平台和科技计划项目数据汇交的基础上，应进一步推进科学数据出版工作。

3.2 科学大数据集成共享质量控制

数据质量是影响数据共享利用效果的重要因素，低质量的数据可能无法使用甚至影响研究结论的正确性。科学大数据的高维性、复杂性、不确定性以及分散异构、来源多样、时空尺度较大等特点，给科学大数据的质量控制带来了巨大的挑战。

科学数据质量本质上需要在数据生产源头环节，通过选用合格的仪器设备，采用正确规范的采集、处理方法，符合精度要求的模型算法以及严格的数据质量控制规范，努力提高操作人员责任心等方法进行控制。从集成共享的角度，首先，要准确掌握数据来源和数据质量信息，包括数据源及其处理方法、属性字段语义、数据精度和不确定性、数据适用范围等，这就要求数据生产者在开放数据的同时提供数据来源和质量元数据。为了能够自动识别这些元数据，需要采用具有明确语法和语义定义并公开发布的元数据标准，如：DC、DIF、ISO 19115 等，进行元数据的编写。其次，要大力发展基于领域知识和机器学习的大数据质量自动检测工具，从数据的完整性、规范性、一致性、正确性等角度，对不同来

源和领域的数据质量进行甄别。同时，可以采用互联网众包模式，鼓励数据用户参与数据质量的评估、标识和修订。

3.3　关联集成与语义搜索

当前，大部分科学数据集成共享主要通过元数据形式实现[40]，即利用元数据描述、发布、查询、定位和访问数据资源。该模式下描述每个数据集的元数据独立发布，仅仅通过主题分类或关键词匹配将元数据进行简单的归类和链接，相互之间缺乏有机的语义关联，很难从一个数据发现另一个高度相关的数据。同时，由于受限于元数据的质量，一旦元数据描述不准确或者与用户的查询关键词不一致，将极大地影响数据共享的效果。因此，如何高效智能地发现与用户需求最相关的数据，甚至实现数据的主动推送，是促进和提升科学大数据共享的重要因素。

解决上述问题可以采用以下两种对策：一是在数据集成阶段，利用关联数据（Linked Data）技术实现数据的关联集成；二是在数据发现阶段，利用语义推理技术，实现数据的语义搜索。关联数据是指通过明确的语义表达发布数据资源，使数据之间能够相互关联和连接。作为语义网的一种实现，关联数据为构建一个富含语义、人机都可理解的、互联互通的全球数据网络奠定了基础。在科学大数据集成过程中，可以从时空范围、内容属性、主题分类、类型格式等多个维度，通过相关度的计算，建立起科学大数据之间定量化的语义关联[41]，从而通过精准的关联关系，实现相关数据资源的智能搜索和主动推荐。利用关联数据技术，还可以根据应用需求，实现不同学科领域、地理位置、时间阶段数据资源的关联聚合，形成具有高度关联、能够满足特定需求的“块数据”[42]。关联集成可以是语义搜索的基础，也可以通过图搜索，利用关联边，根据关联关系和关联度，从一个数据发现另一个相关度高的数据（图2）。

3.4　数据产权与共享安全

数据产权和共享安全问题始终是数据集成共享的重要问题，尤其在大数据时代，数据面临不断被集成、共享、再生产、再集成、再共享的过程，数据产权保护和共享安全保障遇到巨大挑战。

数据产权保护可以利用数字对象标识符（DOI）对数据资源进行全球唯一标识、解析与发现，并利用关联的元数据对数据产权进行详细说明。同时，可以通过数据出版，将数据正式出版。当其他应用研究或论文作者利用该数据时，可以在应用成果或论文中对数据来源DOI进

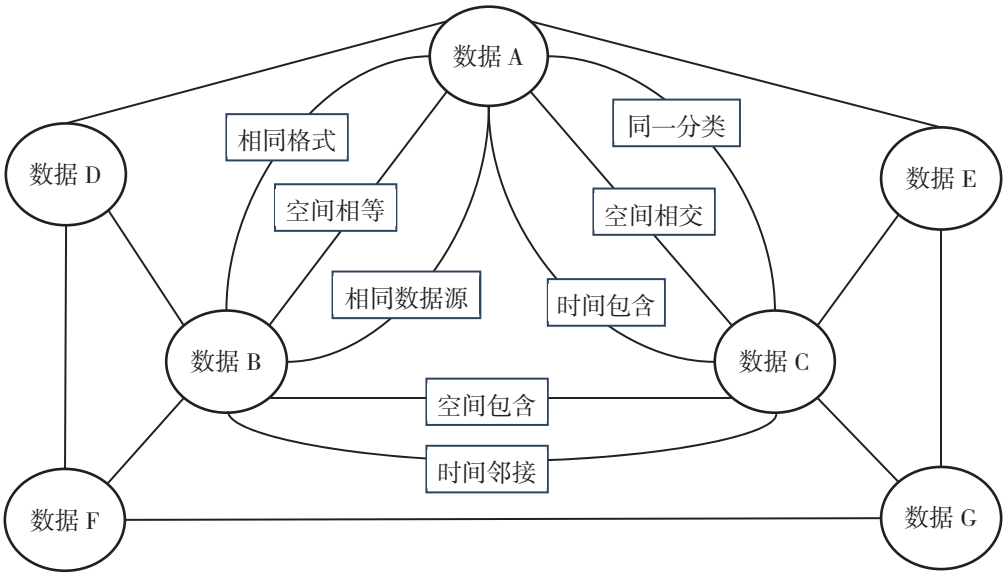

图2　基于多维特征的数据关联

行标注或在参考文献中进行正式引用。通过数据DOI，可以对该数据集全球引用情况进行统计，从而有效保护数据产权。

共享安全可以利用传统的数字水印技术，或最新的区块链技术，防止在数据共享过程中数据的篡改和伪造。区块链就是把加密数据（区块）按照时间顺序进行叠加（链）生成的永久、不可逆向修改的记录[43]。区块链技术将所有数据都储存在一个个数据区块中，数据共享交换信息形成一个完整链条包含在区块链里，所有数据由计算机加密生成。利用区块链技术可以生成一套按时间先后记录的、不可篡改的、可信任的全网统一的数据库，并且这套数据库具有去中心化、数据无法伪造、不可撤销、不可逆转的特点，在没有任何可信第三方存在的时候，能够使参与者对全网数据交换共享记录的事件顺序和当前状态建立共识，解决数据共享的安全问题[40, 44]。

3.5 科学大数据高效利用

现代科学研究范式（第四科研范式）的特点就是利用海量科学数据，通过挖掘分析、模拟预测等寻找、发现数据背后隐含的科学规律和问题[45]。因此，科学大数据的高效利用除了数据集成共享外，还需要处理、利用数据的模型工具、文献资料，以及支撑数据处理、利用的高性能计算能力。由于传统数据共享、处理分析模型和高性能计算等相互独立，并未有机地耦合在一起，严重制约了科学大数据的高效利用。

为了解决上述问题，应在数据共享的基础上，大力发展集数据、模型、文献、计算资源共享为一体的协同信息化科研环境（e−Science）（图3）。e−Science的概念早在2000年就由英国研究理事会提出，是指在重要科学领域中的全球性合作，以及使这种合作成为可能的下一代基础设施，主要包括：计算资源、数据资源、科技文献、模型工具、网络通信资源以及科学仪器设备等。为了实现这些科学研究基础设施的共享，需要在云计算的支撑下，实现硬件资源虚拟化、数据资源和软件资源服务化；利用服务链和语义对

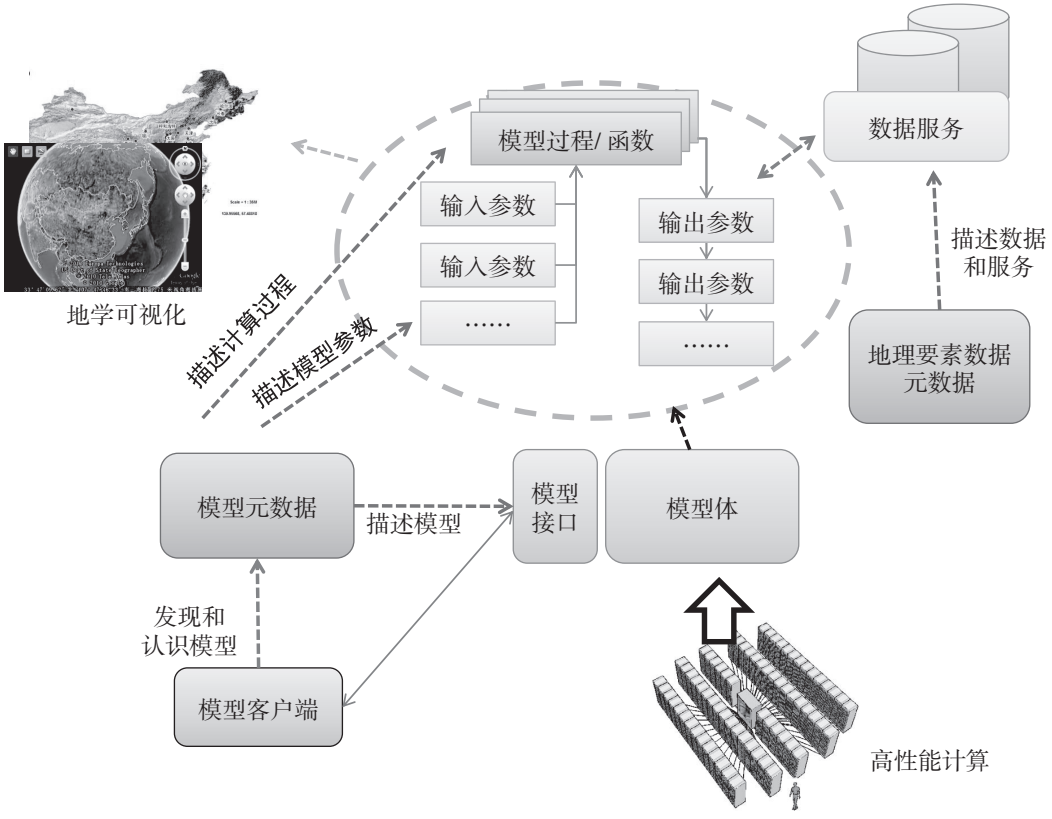

图3 数据—模型—计算资源共享一体化的信息化科研环境

齐等，实现科学数据、模型、计算资源等的有机集成与一体化共享。e-Science以科学数据为核心，自动为数据处理与模拟分析，模型工具对应匹配提供需要的计算资源，从而大力提升科学大数据的利用效率。

4　结语

随着地基监测、对地观测、深地（空）探测，特别是移动互联网、云计算、物联网和社交网络等技术的迅猛发展和深入应用，科学数据的采集、处理、传输变得越来越便捷，科学大数据的时代已经到来。现代数据密集型科学研究范式的特点就是利用海量科学数据，通过挖掘分析、模拟预测等寻找、发现数据背后隐含的科学规律和问题，因此，愈发依靠科学大数据。本文分析科学大数据内涵和特征，总结科学数据集成共享国内外主要进展，指出科学大数据集成共享面临的主要挑战，提出具体的应对策略。

（1）科学大数据的集成在保持现有国家科技条件平台和科技计划项目数据汇交的基础上，应进一步推进科学数据出版，充分激发广大科研人员的积极性，促进形成“人人都是数据使用者和贡献者”的志愿数据共享氛围。

（2）在控制好科学大数据生产环节质量的同时，应重视数据来源和数据质量元数据，大力发展基于领域知识和机器学习的大数据质量自动检测工具，采用互联网众包模式，鼓励数据用户参与数据质量的评估、标识和修订。

（3）充分利用关联数据技术，通过明确的语义表达发布数据资源，实现数据的关联集成，构建一个富含语义、人机都可理解的、互联互通的数据网络，从而支撑数据的智能发现与主动推荐。

（4）通过数字对象标识和数据出版引用以及区块链技术等，利用全球唯一标识和不可篡改的全网数据交换共享记录，有效保障数据产权和共享安全。

（5）在数据集成共享的基础上，应大力发展集数据、模型、文献、计算资源共享为一体的协同信息化科研环境，大力提升科学大数据的利用效率。

参考文献

[1] 中华人民共和国科学技术部. SDS/T1003.2-2004.科学数据共享工程技术标准,科学数据共享概念与术语第2部分:术语[S].2004.

[2] 诸云强, 朱琦, 冯卓,等. 科学大数据开放共享机制研究及其对环境信息共享的启示[J]. 中国环境管理, 2015, 7(6):38-45.DOI: 10.16868/j.cnki. 1674-6252. 2015.06.008.

[3] 陈明奇, 黎建辉, 郑晓欢,等. 科学大数据的发展态势及建议[J]. 中国教育信息化, 2016(21):5-9.

[4] 李德伟.大数据改变世界[M].北京:电子工业出版社, 2013:8.

[5] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域: 大数据的研究现状与科学思考[J].中国科学院院刊, 2012, 27(6): 647-657.

[6] 诸云强, 朱琦, 冯卓,等. 科学大数据开放共享机制研究及其对环境信息共享的启示[J]. 中国环境管理, 2015, 7(6):38-45.DOI: 10.16868/j.cnki.1674-6252. 2015. 06.008.

[7] 孙九林, 林海. 地球系统研究与科学数据[M]. 北京: 科学出版社, 2009.

[8] MATTMANN C A. A vision for data science[J]. Nature,2013, 493: 473-475.

[9] 郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球[J]. 科学通报, 2014, 59(12): 1047-1054.

[10] ABARBANEL H D I, BROWN R, SIDOROWICH J J, et al. The analysis of observed chaotic data in physical systems[J]. Rev Mod Phys, 1993, 65:1331–1392

[11] ROCHA L M. Complex systems modeling: using metaphors from nature in simulation and scientific models[R]. Los Alamos: Los Alamos National Laboratory, 1999.

[12] NATIONAL SCIENCE FOUNDATION. Long-lived digital data collections enabling research and education in the 21st century[EB/OL].[2015-10-26]. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

[13] 美国当局发布大数据研发战略计划[EB/OL].[2016-06-27]. https://zhuanlan.zhihu.com/p/21431164.

[14] Administration issues strategic plan for big data research and development[EB/OL].[2016-05-23].https://obamawhitehouse.archives.gov/blog/2016/05/23/administration-issues-strategic-plan-big-data-

research−and−development.

[15] 美国政府出台大数据研发计划[EB/OL].[2012−04−24].http://www.most.gov.cn/gnwkjdt/201204/t20120424_93877.htm.

[16] KALIL Tom. Big data is a big deal[EB/OL].[2012−03−29].http://www.whitehouse.gov/blog/2012/03/29/big−data−big−deal.

[17] 地平线2020计划[EB/OL].[2014−07−01]. http://www.cstec.org.cn/ceco/zh/show/359.aspx.

[18] HORIZON 2020. The EU framework programmer for research and innovation[EB/OL].[2015−06−06].http://ec.europa.eu/programmes/horizon2020/

[19] National oceanic and atmospheric administration[EB/OL]. [2017−06−08].http://www.noaa.gov/.

[20] 欧盟GRDI 2020[EB/OL].[2014−11−20].http://www.grdi2020.eu/.

[21] 科学数据共享工程[EB/OL].[2013−07−01].http://www.most.gov.cn/ztzl/kjzg60/kjzg60hhcj/kjzg60jcyj/200909/t20090911_72832.htm.

[22] 国家科技基础条件平台建设专项简介[EB/OL].[2010−10−22].http://3y.uu456.com/bp_1tri89oyyy7b8vd53zkt_1.html

[23] 基础研究大数据服务平台应用示范[EB/OL]. [2013−05−17].http://www.cas.cn/xw/yxdt/201305/t20130521_3843550.shtml.

[24] 国际科技数据委员会[EB/OL]. [2014−07−22].http://www.bic.cas.cn/gjzz/201307/t20130719_3902696.html.

[25] 李娟, 刘德洪, 江洪. 国际科学数据共享现状研究[J]. 图书馆建设, 2009(2):25−27,31.

[26] 孙鸿烈, 刘闯. 国际科学技术数据前沿领域发展研究[J]. 中国基础科学, 2003, 18(1):329−333.

[27] Organization for economic cooperation and development[EB/OL]. [2017−04−16].http://www.oecd.org/.

[28] 李娟, 刘德洪, 江洪. 国际科学数据共享原则和政策研究[J]. 图书情报工作, 2008, 52(12):77−80.

[29] 杨友清, 陈雅. 科学大数据共享研究:基于国际科学数据服务平台[J]. 新世纪图书馆, 2014(3):24−28. DOI: 10.16810/j.cnki.1672−514x. 2014.03.005

[30] Research data sharing without barriers[EB/OL]. [2017−05−05].https://www.rd−alliance.org/node.

[31] 美国国家科学基金会支持研究数据共享[EB/OL].[2012−12−14].http://www.most.gov.cn/gnwkjdt/201212/t20121213_98503.htm.

[32] 王艳翠, 李书宁, 李爱红. 研究数据联盟:建立全球数据共享和数据交换的基础架构[J]. 图书馆理论与实 践, 2015(1):52−54. DOI:10.14064/j.cnki.issn1005−8214.2015.01.014.

[33] RDA Seventh Plenary Meeting, Tokyo, Japan[EB/OL].[2016−02−29]. https://rd−alliance.org/plenary−meetings/rda−seventh−plenary−meeting.html.

[34] WORLD DATA CENTER[EB/OL]. [2017−05−26]. http://wdc.org.ua/.

[35] 王卷乐, 孙九林. 世界数据中心(WDC)中国学科中心数据共享进展[J]. 中国基础科学, 2007, 9(2):38−42.

[36] 王卷乐, 孙九林. 世界数据中心(WDC)回顾、变革与展望[J]. 地球科学进展, 2009, 24(6):612−620.

[37] 建立全球综合地球观测系统须无间合作[EB/OL].[2010−11−08]. http://news.sciencenet.cn/sbhtml−news/2010/11/238246.html.

[38] GEOSS[EB/OL]. [2016−05−16]. http://www.earthobservations.org/geoss.php.

[39] 刘闯. 论全球变化科学研究数据出版[J]. 地理学报, 2014,69(Z): 3−11.

[40] 诸云强.地球系统科学数据共享关键技术研究[D].北京:中国科学院地理科学与资源研究所,2006.

[41] ZHU Yunqiang, ZHU A−Xing, SONG Jia, et al. Multidimensional and quantitative interlinking approach for Linked Geospatial Data[J].International Journal of Digital Earth, 2017,10(9): 923−943.

[42] 大数据战略重点实验室. 块数据3.0:秩序互联网与主权区块链[M]. 北京:中信出版社, 2017.

[43] 区块链技术[EB/OL]. [2016−06−20].http://www.jianshu. com/p/4c2754ba4766.

[44] 林小驰, 胡叶倩雯. 关于区块链技术的研究综述[J]. 金融市场研究, 2016(2):97−109.

[45] HEY Tony, TANSLEY Stewary, TOLLE Kristin. The Fourth Paradigm: Data−intensive Scientific Discovery[M].The United States of America: Microsoft Corporation, 2009.