国家科学数据资源发展报告 2016

国家科技基础条件平台中心　编著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

国家科学数据资源发展报告2016

国家科技基础条件平台中心　编著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

·北京·

图书在版编目（CIP）数据

国家科学数据资源发展报告. 2016 / 国家科技基础条件平台中心编著. —北京：科学技术文献出版社，2016. 12
ISBN 978-7-5189-2161-4

Ⅰ. ①国… Ⅱ. ①国… Ⅲ. ①科学技术—数据管理—研究报告—中国—2016 Ⅳ. ① G203

中国版本图书馆 CIP 数据核字（2016）第 291158 号

国家科学数据资源发展报告2016

策划编辑：周国臻　责任编辑：周国臻　于东霞　责任校对：赵　瑗　责任出版：张志平

出　版　者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编 100038
编　务　部　(010) 58882938，58882087（传真）
发　行　部　(010) 58882868，58882874（传真）
邮　购　部　(010) 58882873
官方网址　www.stdp.com.cn
发　行　者　科学技术文献出版社发行　全国各地新华书店经销
印　刷　者　北京九州迅驰传媒文化有限公司
版　　次　2016 年 12 月第 1 版　2016 年 12 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　50千
印　　张　5
书　　号　ISBN 978-7-5189-2161-4
定　　价　38.00元

《国家科学数据资源发展报告2016》
编 写 组

主　　编：叶玉江

副 主 编：王瑞丹　苏　靖

主　　笔：石　蕾　王卷乐

编写组成员：（按姓氏笔画排列）

马向南　马俊才　王　末　王　正　王　戎
王　祎　王　晋　王　健　王　超　王　辉
王雨华　尹　岭　尹海清　卢　凡　卢　琦
田奋民　冯爱霞　吕　鑫　朱　琦　刘　清
刘艾琳　闫保平　许哲平　孙功星　纪　平
纪力强　李一凡　李军莲　李丽亚　李园白
李国庆　李俊清　杨思维　杨啸林　杨雅萍
肖景发　吴　健　吴佳妍　吴俊升　何洪林
邹自明　张英杰　张晓宇　陈　刚　陈广飞
陈志辉　陈源泉　尚雷明　周　伟　周国民
赵　强　赵国峰　赵春江　柏永青　袁　伟
夏　雪　高孟绪　崔辰州　董明媚　温　浩
谢景林　赫运涛　黎建辉

前　言

科学数据是人类在科技活动过程中产生的基本科学技术数据、资料及按照不同需求而系统加工的数据产品和相关信息，具有明显的潜在价值和可开发价值，科学数据工作贯穿于科技创新活动的全过程，并在广泛应用过程中增值，是信息时代传播速度最快、影响面最广、开发利用潜力最大的科技资源。

改革开放以来，我国科技创新能力快速提升和发展，科技投入强度不断增加，通过各级各类科技计划项目实施、科研基地建设、国际科技合作等促进了科学数据的快速积累与发展。国家科技计划、行业部门专项及地方各级科研项目都在不同程度上支持了科学数据生产与加工，以大科学装置、国家野外科学研究台站、国家重点实验室等为代表的科研基础设施和研究实验基地也积累了大量科学数据。在科学数据快速增长的同时，科学数据采集、加工、整理、存储和利用的方式和手段不断丰富，从事科学数据建设、管理和共享服务的人才队伍也不断壮大。

2004 年 7 月，国务院办公厅转发的科技部、国家发展改革委、财政部、教育部《2004—2010 年国家科技基础条件平台建设纲要》，是我国科学数据管理和共享服务工作的重要里程碑，科技部、财政部共同设立国家科技基础条件平台建设专项。自此，我国在加强科学数据资源整合，促进科学数据开放共享，推动科学数据服务方面迈出了新的步伐，促进了多部门、多单位科学数

据资源集聚，在多个重点领域建成了科学数据资源开放门户系统，也形成了一支专门从事科学数据共享服务的技术人才队伍。

随着大数据时代的到来，以“数据密集型科学研究”为显著特征的科学研究“第四范式”已逐渐成为科技发现最重要的手段之一，越来越多的科学研究和发现依赖于全面、完整、准确的科学数据的收集和利用，这为科学数据的快速积累和发展带来了前所未有的机遇。但与此同时，大数据的热潮也给科学数据的有效集聚、科学管理、深入挖掘和共享知识产权保护等方面带来了巨大挑战。就我国科学数据工作而言，尽管我国在科学数据资源管理与共享方面开展了大量工作，但数据多头生产、分散管理、开放不足等问题仍然没有从根本上解决，大量科学数据资源仍然没有实现有效集成和充分利用，支撑科学数据管理和共享的技术手段有限，与美、英等发达国家差距仍十分明显。2015 年 8 月 31 日国务院印发《促进大数据发展行动纲要》，也明确指出发展科学大数据，将科学数据发展提升为国家战略之一，为科学数据管理和共享服务发展带来了新的契机。为进一步促进科学数据资源的共享与应用，各部门和单位积极研究出台科学数据管理与共享策略，并积极推进科学数据开放共享工作，取得了显著成效。

本报告面向科学数据资源管理与应用整体发展趋势，对我国科学数据资源发展状况进行了梳理和总结，并对我国科学数据管理和共享服务进展和成效进行了初步判断，提出深入推进我国科学数据管理和共享工作发展的建议。

“第一章　科学数据资源发展概述”，对科学数据的特征与范围进行了归纳和总结，结合当前科学数据发展的前沿与态势，指出了科学数据在支撑科技创新、社会民生和经济发展等方面的作用和重要意义。

“第二章　我国科学数据资源发展现状”，从战略部署、数据规模、数据内容、数据质量、数据中心建设及数据科学发展等方面展示了我国科学数据资源发展的全景图，从宏观层面多角度呈现了各学科领域数据的共性与差异。

“第三章　我国科学数据共享服务情况”，在对科学数据共享服务环境发展进行总结的基础上，结合典型案例介绍了科学数据整合，以及在服务国家重大战略、科技创新、重大工程、社会民生、科普教育、国际影响等方面的服务成效，并查找我国在科学数据管理共享方面与国外的差距。

“第四章　我国科学数据资源发展展望”，结合我国科学数据管理和发展现状，对比国内外科学数据资源管理和利用差异，从加快推进政策机制建设、推动国家级科学数据中心建设、促进科学数据积累与共享，以及加强科学数据安全保障4个方面提出推动我国科学数据发展的建议。

目　录

第一章　科学数据资源发展概述

科学数据是信息时代传播速度最快、影响面最广、开发利用潜力最大的科技资源。当代科学技术的发展已经使得科学数据成为科技创新、管理决策、经济发展等活动不可缺失的基础支撑条件，被公认为继物质和能量之后的第三类资源，成为重要的国家战略资源和各国科技实力竞争的重要资本。科学数据对现代科学可持续创新研究的作用犹如水资源对人类的生活、耕地资源对农业的发展及石油对于工业的发展，正逐渐成为国际科技竞争的重要科技基础条件，对于提升科技创新能力、推动经济社会发展及维护国家安全具有越来越重要的意义。

本报告对科学数据资源的特征与范围，以及当前科学数据资源发展的前沿与态势进行归纳，对我国科学数据资源的阶段性发展现状进行了总结，对我国科学数据的共享和利用情况进行了梳理，并在此基础上提出进一步推进我国科学数据资源建设与发展展望。

一、科学数据的特征与范围

科学数据是人类社会科技活动积累的或通过其他方式获取的反映客观事物的本质、特征、变化规律等的原始性、基础性数据，以及根据不同科技活动需要进行系统加工整理的各类数据的集合。

科学数据作为最基本、最活跃的一类科技资源，既是科技创新活动的重要产出，也是新一轮创新活动和经济社会发展的重要基础和工具。较其他科技资源而言，科学数据具有客观性、共享性、时效性、分散性、多结构性、再创造性、非排他性、传递性等多种特点。

当前大科学、大数据时代科技创新越来越依赖大量、系统、高可信度的科学数据基础，对科学数据的综合分析已经成为科技创新的一种方式。海量科学数据的产生已经对生命科学、天文学、空间科学、地球科学、物理学等多个学科领域的科研活动带来了冲击性的影响，科学研究的方法发生了重要的变革。科学数据工作已贯穿于科技创新活动的全过程，通过基础性、专题性、功能性服务等多种方式支撑科技创新活动。在科研资料收集阶段，科学数据直接启发创新思路和方向；在科研资料组织阶段，针对应用需求对科学数据进行集成、分析和挖掘，直接为科技创新服务；在创新实施阶段，需要使用科学数据对创新成果进行反复验证，从而实现更完美的创新成果。

国内外机构从不同视角对科学数据提出了多种定义，但观其本质和特点是趋于一致的。例如，美国国家科学基金会（NSF）在其出版的报告中将数据描述为以数字化形式存储的数据资源。美国国家航空航天局（NASA）在对美国政府资助的科学数据进行管理时，指出数据应该包括观测数据、元数据、数据产品、信息、算法，包括相关的源代码、文档、模型、图片和研究结果。开放数据（Open Data）把数据分成两种相关的类型：一类是开放的政府数据，另一类是在大学或研究机构通过政府资助的研究项目中的科研活动所产生的数据。在我国，根据科学数据产生的渠道不同，有学者认为科学数据资源通常可分为两大类型，一类是

行业部门按照统一的规范标准长期采集和管理并用于科学研究的数据（业务数据）；另一类是国家各类科技计划项目在研究过程和结果中产生的，以及为支持科学研究而通过观测、监测、试验等站点采集的科学数据（研究型数据）。根据科学数据资源的生产方式不同，也有学者提出科学数据包括观测数据、研究数据、监测数据、调查数据、统计数据、模拟数据等。

综合科学数据的相关特征和不同定义分类，本报告中科学数据的范围主要包括以下几类：①通过长期观测、试验所获取的描述事物、现象分布格局与变化过程数据；②通过科学考察、调查所获取的描述事物、现象的情景数据；③通过实验、测试分析所获取的描述物体（含生命体）特征的数据；④科技应用、研究活动中所产生和积累的具有科学价值的数据和相关信息。

二、科学数据的前沿与发展态势

1. 科学数据对科技创新的支撑作用日益显著

在科学发展的进程中，数据作为对科学研究的记录，一直是科学研究的重要基础。随着大数据的出现，科学发现越来越依赖于对海量数据的收集、管理和分析，科学研究水平也越来越多地取决于对数据的积累，以及将数据转换为信息和知识的能力。对生命科学、天文学、空间科学、地球科学、物理学等多个学科领域的科研活动更是带来了冲击性的影响，科学研究方法发生了重要的变革，以数据密集型为特征的科学研究模式不断涌现。如生命科学领域的科学家们利用海量 DNA 数据重新认识生命，高能物理科学家通过对海量实验数据的处理和分析发现希格斯粒子。

欧美等发达国家已经将科学数据的持续积累和开放利用能力

提高到国家科技战略高度进行部署，并投入了大量的人力、物力和财力。通过多年持续积累，形成了一批权威、长序列、多尺度的科学数据库，并在科研过程中发挥了重要作用。例如，英国著名的洛桑农业实验站，积累了长达 160 多年的土壤样品和生态试验数据，成为全世界研究人类耕作制度、施肥方式和土壤酸化演变等方面不可多得的宝贵科学财富；加拿大在全国定期开展格网化的资源调查，持续积累和提高其对自然资源的管理与利用的能力；Argo 在全球范围内部署海洋浮标，用于大尺度全球气候变化观测；美国国家航空航天局（NASA）、大气和海洋局（NOAA）及卫生研究院（NIH）等机构支持建立的多个国家数据中心，长期开展基础科学数据的积累和共享服务，为美国及全球航天、大气、海洋和生命科学研究提供了重要数据资料；由发达国家主导形成的全球碳监测网络，在应对全球气候变化国际合作中发挥了关键作用等。

科学数据是科学研究的基础支撑与保障，科学数据的爆发式增长，已把科学研究各个领域和环节推到了一个前所未有的“大数据”时代，由数据驱动的科研创新模式正逐步形成。利用已有数据、产生科技成果、形成新的数据、继续共享利用已成为科学数据支撑科技创新的可行模式，并得到各国政府和科技界的普遍认同。在科学数据开放共享促进科技创新成果产出的同时，开展科学数据管理工作对节约科技创新成本也将做出巨大贡献。特别是积极推动公共财政支持产生的科学数据的汇交集成与规范管理，将有效提高国家财政科技投入的效益。

2. 科学数据的全生命周期管理取得高度共识

科学数据同其他科技资源一样，具有形成、成长、成熟、衰

亡的生命过程。欧美等国家及世界知名科学数据中心通常都按照科学数据的生命周期对科学数据进行科学生产、规范管理和合理利用。科学数据的全生命周期一般包括规划设计、采集生产、加工保存、共享服务和资源处置5个主要环节，在每个环节中又有一系列具体流程和方法。其中，根据科学数据管理现状，以及科学数据的生命周期与科研活动周期的关系，科学数据的采集生产、加工保存、共享服务是当前科学数据生命周期管理的重点环节（如图1.1）。

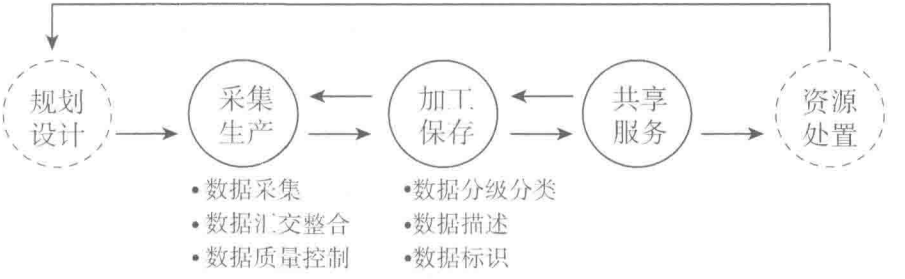

图1.1　科学数据全生命周期

宏观的科学数据管理（data management）贯穿整个科学数据生命周期，通常包括数据收集、数据归档、数据认证、数据加工、数据保存、数据发布、数据共享等。目前已开展科学数据管理的国家和机构多以全生命周期为主要轨迹进行科学数据管理。例如，美国就针对科学数据生命周期各环节制定出台了相关法律制度，以及相应管理细则予以规定和保障。在科学数据收集方面，美国国家科学基金会（NSF）要求在项目建议书中必须包括不超过2页的“数据管理”计划。英国也根据科学数据生命周期建立了完整的数据管理流程和相应的法律制度，明确了部门对递交数据管理计划的要求，以及对数据管理的经费支持。同样，澳大利亚在制定的科学数据管理法律及法规中，明确要求通过澳大利亚国家

数据服务中心对外提供科学数据共享服务。

微观的科学数据管理（Data Curation）侧重于科学数据的加工保存环节，这一层面的数据管理正迅速成为专业学科领域、信息科学和图书馆学的关注热点。在 Web of Science 数据库中以“Data Curation”为主题进行检索，发现1900—2000年仅有44篇相关文献，从2001—2016年则增加到1615篇。图1.2展示了2001—2015年15年间相关文章数目的变化。

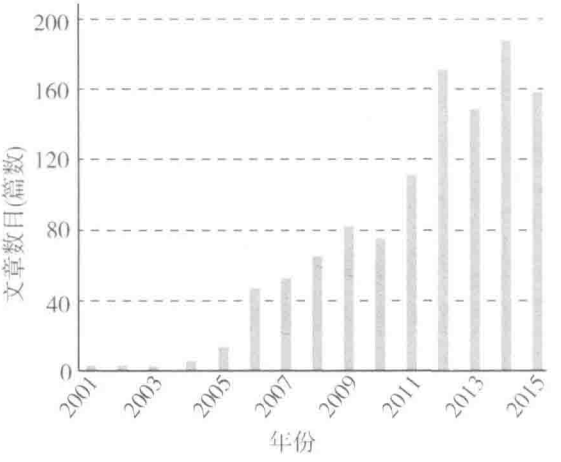

图1.2　2001—2015年以数据管理（Data Curation）为主题发表的文章数目

3. 大数据时代科学数据发展和相关研究呈现新进展

大数据时代促进科学数据急速增长。计算机技术、以互联网为代表的通信技术和以物联网为代表的传感技术的持续创新和广泛应用使人类的数据化能力和范围快速扩张，越来越多的观察、计算和传播等仪器设备正在产生着源源不断的海量复杂的数据，这使得几乎每个学科领域都在面对着空前的数据爆炸。以中、英、美、德等多个国家承担的大型国际合作项目“千人基因组计划”为例，在其实验阶段就实现了每月生成大约1万亿碱基序列数据，而最初的人类基因组项目花费了10年才产生400亿碱基的DNA

序列数据。2015 年，斯隆数字巡天（Sloan Digital Sky Survey，SDSS）发布了最新版的公开数据，超过 100TB 的第 12 版数据（DR12）包含了对近 5 亿颗恒星和星系的相关测量，这是天文学历史上规模最大、内容最为丰富的数据库之一。

大数据的来临使得数据密集型科学研究得到突破性进展。伴随着大数据时代的到来，数据密集型科学研究已成为新的科研范式，气象、天文、地球物理、基因组学等更是利用先进的科研方法实现科学数据的管理和分析，解决更新、更复杂的科学问题。例如，欧洲核子研究组织（CERN）利用网格计算和大数据技术发现“上帝粒子”，即希格斯粒子。寻找希格斯粒子的大型强子对撞机（LHC）实验是一个典型的基于大数据的科学实验，在 1 万亿个事例中才可能找出 1 个希格斯粒子。在生命科学领域，基于流感监测、基因序列及社会经济等海量数据进行分析挖掘，并结合大量的相关性分析，发明的可以快速选择流感疫苗株的技术，为更快制备疫苗、防控流感提供了新方法，该成果被《Nature Communications》选为“亮点文章”。

大数据的开发应用模式给传统科研活动新的启发。在大数据创新理念的激发下，利用网络平台推动科学数据开放共享、开发利用，采取“众包众筹”的方式加速科研进程，给传统的科研方式带来了巨大的冲击和影响。例如，Foldit 项目通过互联网发起了大规模的协同研究，以数据为纽带联合数千名科研人员共同参与研究，进行联机计算，使得该项目能够以前所未有的速度得到推进。再如，GalaxyZoo 研究项目召集了 25 万个研究者，包括专业研究人员、业余研究者及爱好者帮助共同收集星际数据，从而发现了一个星系的新类，加深了人类对宇宙的认识。又如，Polymath 项目中，各个领域的研究者及非专业数学家协作解决了一个

传统方法长期无法解决的问题，这种大众协作参与科研的方式被称为“科研众筹”（Crowd Science）。目前，我国多个科学数据库建设项目除了采用传统的项目数据采集之外，也在积极探索采用众筹方式获取本领域科学数据，以加速科学数据的整合集成和分析利用，提升科研效率。

4. 权威科学数据中心加速抢占科学数据资源

国际上欧美等发达国家依托其优势科学数据中心，持续加大科研数据的开放力度，进而吸引全球优势科学数据资源。据统计，目前国际知名科学数据库主要集中在欧美日等发达国家，表1.1列举了部分相关学科领域国际知名的科学数据库。这些科学数据中心在科学数据标准、技术、资源等方面具有权威性。发达国家依靠这些权威性科学数据中心，持续整合和汇聚全球科学数据资源，并逐渐形成标准化的科学数据收集、管理和存储解决方案。例如，全球生物多样性信息网络（GBIF）是目前全球最大的生物多样性信息服务机构，该组织通过合作和种子基金等各种途径促进生物多样性原始数据（Primary Data）的共享，已形成一个面向全世界用户的、关于全球生物多样性的综合性信息服务系统。再以GenBank（基因银行）为例，目前GenBank已经成为世界权威的基因序列登记库，并被科学共同体所接受，发表学术论文往往需要提供基因登记号。《Nature》杂志明确规定，关于基因测序数据必须汇交到指定数据库，以此作为文章发表的门槛，GenBank已整合大量世界高水平的基因序列数据。