

中国科研信息化蓝皮书

China's e-Science Blue Book 2011

中国科学院

中华人民共和国教育部

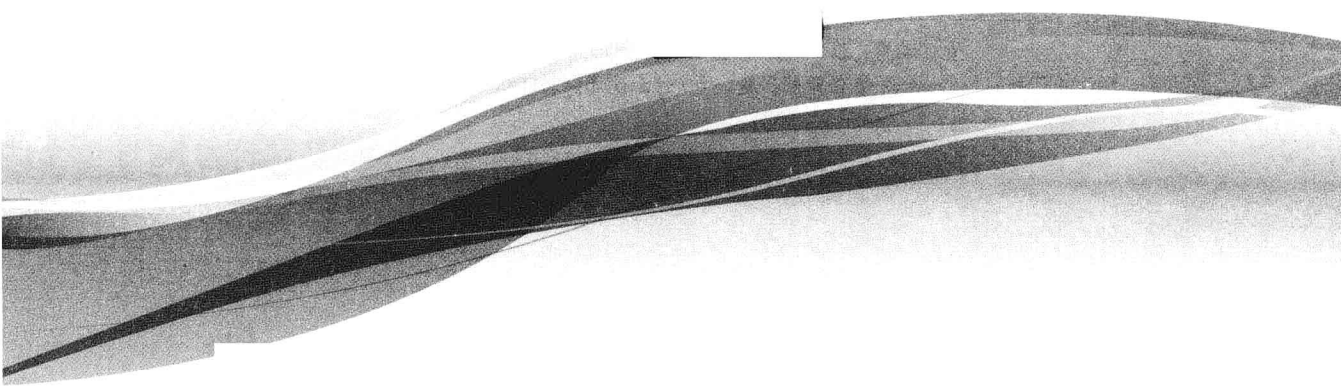
国家自然科学基金委员会



科学出版社

中国科研信息化蓝皮书 2011

中国科学院
中华人民共和国教育部
国家自然科学基金委员会



科学出版社

北京

内 容 简 介

本书是由中国科学院联合中华人民共和国教育部、国家自然科学基金委员会共同编著而成的报告,旨在系统地展示中国科研信息化的整体发展情况,推动中国科研信息化的进程。

本书邀请了国内外科研信息化领域的专家、学者和企业家,针对科研信息化涉及的网络环境、超级计算环境、数据环境,以及科研信息化的技术发展和应用实践的现状与趋势进行了客观阐述,对科研信息化的战略态势进行了深入的分析与探讨,力求推动科技创新与创新模式的转变,为中国未来科技创新提供全局性、战略性的参考,向国内外读者展示中国科研信息化的全貌和前沿成果。

本书可作为政府部门、科研机构、高等院校和相关企业进行科技战略决策的参考书,也可供国内外专家、学者研究和参考。

图书在版编目(CIP)数据

中国科研信息化蓝皮书2011/中国科学院,中华人民共和国教育部,国家自然科学基金委员会. —北京:科学出版社,2011.12

ISBN 978-7-03-032831-1

I. ①中… II. ①中… ②中… ③国… III. ①信息技术—应用—科学研究工作—研究报告—中国—2011 IV. ①G322-39

中国版本图书馆CIP数据核字(2011)第238887号

责任编辑:张 濮 朱雪玲 王 哲/责任校对:钟 洋

责任印制:赵 博/封面设计:刘可红

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

中国科学院印刷厂印刷

科学出版社发行 各地新华书店经销

*

2011年12月第 一 版 开本:787×1092 1/16

2011年12月第一次印刷 印张:16 3/4

字数:390 000

定价:95.00元(含光盘)

(如有印装质量问题,我社负责调换)

《中国科研信息化蓝皮书2011》

指导委员会

施尔畏 杜占元 何鸣鸿

编写委员会

主 任： 谭铁牛

副主任： 娄 晶 曾 明

成 员： 廖方宇 黄向阳 陈明奇
 吴丽辉 谭 华 肖人毅
 李望平 许海燕 刘 冰
 王常青 刘 锋

序 言

进入21世纪以来,信息化在世界范围内得到迅猛发展,正在引发当今世界的深刻变革,重塑着世界政治、经济、社会、科技、文化和军事发展的新格局。

随着高性能计算机、互联网、物联网等信息技术的飞速发展,以及强子对撞机、空间遥感等大型科学装置和设施的建立与运行,人类科技活动以崭新的姿态步入了信息化发展的快速轨道,科学研究活动进入了科研信息化(e-Science)阶段,人类科技创新能力提高到了前所未有的水平。

以互联网为代表的信息传输革命,奠定了科学研究跨国家、跨地域、跨领域合作、交流与协同的基础,通过电子邮件、带有语音通话和图像传输功能的即时通信系统以及基于网络而形成的虚拟实验室,全球科学家能在任何时间、任何地域方便、快捷地无障碍交换数据和资料,远程操控科学试验,共同分析和分享研究结果,“面对面”地进行科学思想的碰撞与交流。

随着高性能计算机、数据存储技术、高效能计算方法、可视化技术和成熟的模拟仿真软件的发展,产生了各类海量数据,当代科技进入了“大数据”时代。2011年6月,日本的超级计算机“京”实现了每秒8 612万亿次的运算速度,EMC和IDC公司的研究显示2010年全球范围内所产生的数字信息为1.2 ZB,并预测2012年将产生1.8 ZB的数字信息。可以说,人类对自身各类活动以及地球(乃至地外)各类信息的感知、处理和分析,对未来各类活动和过程的模拟、预测,达到了前所未有的境界和高度。

目前,世界许多科技大国都把科研信息化作为国家间科技竞争的重要手段和科技发展战略的主要部分,投入巨大的人力物力来支持和发展。美国从20世纪80年代起,建设了以互联网为基础、以高性能计算为核心的新一代科研基础设施,启动了下一代互联网和网格两个方向的国家研究计划,提升了科研网格的资源能力和覆盖范围,并从2011年开始着手启动新一轮科研信息基础设施计划。英国在世界上首次提出科研信息化的概念,并以国家计划支持网格中间件、网格工作流、网格门户等一系列公共支撑软件的开发,取得了举世瞩目的成绩,使其在这一领域走在世界前列。欧盟框架计划专门设立了“e-基础设施”研究方向,构建欧洲科研教育高速网络,连接欧盟各国的科研教育网络,促进科学数据共享与联合,构建面向各学科领域的虚拟组织,并通过实施大型强子对撞机(LHC)运行的计算网格(LCG),引领了全球协同的科

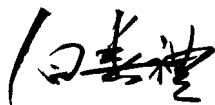
学研究新模式。

我国科研信息化工作始于20世纪90年代中期,十几年来取得了巨大的进展。中国科技网(CSTNET)和中国教育和科研计算机网(CERNET)相继建成并与国际互联网连通,奠定了我国科研信息化的网络基础。我国千万亿次高效能计算机的研制成功,为科学研究提供了强有力的计算工具。在自主开发的网格软件支撑下,中国国家网络集成了每秒430万亿次以上的计算能力,2 200万亿字节的存储能力和200多个应用与工具软件,支持了资源共享和协同工作,支撑了700多项国家各类科技计划项目的研究工作。以科技资源整合和共享为重点的国家科技基础条件平台建设,在自然资源、科学数据和文献等方面,为科技信息共享提供了丰富的资源。海量数据处理、网络虚拟计算环境、大规模并行计算方法与支撑框架、下一代互联网等一批重要项目的部署,为解决科研信息化所面临的关键应用提供了技术保障。在环境、流体、高能物理、计算化学、生物信息等许多领域开发了一批应用示范系统,在探索网络环境下科研新模式方面进行了有益的尝试。

为进一步推进我国科研信息化工作的深入发展,中国科学院联合中华人民共和国教育部、国家自然科学基金委员会编撰出版了首部反映我国科研信息化发展的综合性研究报告——《中国科研信息化蓝皮书2011》。该书系统阐述了科研信息化的理念与内涵,介绍了国内外科研信息化发展态势,汇总展示了我国科研信息化建设与应用现状,论述了我国科研信息化涉及的主要技术问题,展望了我国科研信息化发展的总体趋势。该书内容丰富、数据翔实、分析透彻,具有重要的参考价值,将成为国内外了解和研究中国科研领域信息化的权威著作。

工欲善其事,必先利其器。随着人类对客观世界的探索与认识不断向新的深度和广度拓展,科研信息化对支撑科技创新及创新模式的转变将发挥越来越重要的作用。让我们插上信息化的翅膀,释放出无尽的创新能量,为我国科技创新和跨越发展提供强有力的支撑,推动中华民族在世界科技发展的历史上做出新的、更大的贡献。

是为序。



中国科学院院长

2011年11月28日

目 录

序言	白春礼 (i)
第一篇 态势战略篇	(1)
一、大力发展科研信息化, 服务国家科技创新	江绵恒 (2)
二、e-Science and Data-Intensive Research	Tony Hey, Jim Pinkelman (6)
三、加强信息资源开发利用与共享, 提高信息化建设效益	孙九林 王卷乐 (27)
四、数据密集时代的科研创新	郭华东 (42)
五、以创新技术支持中国科研信息化发展	方之熙 (48)
第二篇 技术发展篇	(55)
一、科研信息化2.0:从构造为科学研究的服务到服务化的科学	郭毅可 (56)
二、中国高性能计算研究与应用发展状况	孙凝晖 (65)
三、多相复杂系统的多尺度并行计算	葛 蔚 李静海 (72)
四、以网络为基础的科学活动环境	胡春明 沃天宇 怀进鹏 等 (80)
五、云计算及其在科学研究中的应用	程耀东 陈 刚 (96)
第三篇 基础设施篇	(107)
一、下一代科教互联网与应用展望	吴建平 李 星 (108)
二、中国科技网及其应用	陈 炜 (114)
三、科研网络新进展	葛敬国 任勇毛 (120)
四、中美俄环球科教网络及其应用	陈江宁 (125)

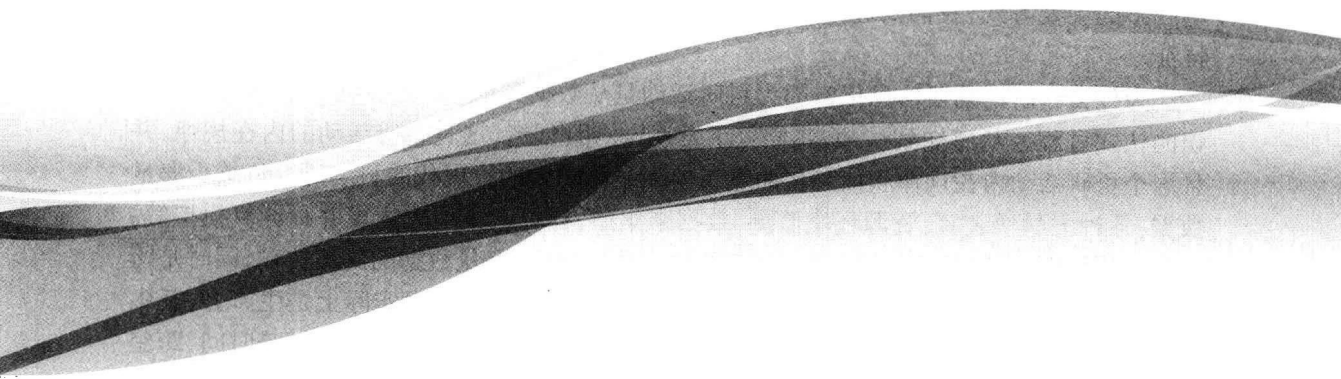
五、国家高性能计算环境及其应用	钱德沛 (130)
六、中国教育科研网格及其应用	金 海 吴 松 (136)
七、中国科学院超级计算环境及其应用	金 钟 朱 鹏 肖海力 等 (144)
八、高性能计算机系统	廖湘科 孙凝晖 肖利民 等 (156)
九、中国科学院数据环境及其应用	黎建辉 虞路清 (168)
十、面向复杂科学的数据密集型网格计算平台	陈 刚 (173)

第四篇 应用实践篇 (181)

一、建设服务天文学的信息化应用系统	刘 梁 郑裕民 高 娜 (182)
二、高性能计算生物信息处理平台	徐姚晨 徐书华 杨家亮 等 (189)
三、基于下一代互联网技术的自然灾害应急数据传输和数据共享平台	李国庆 于文洋 周 旭 (196)
四、地学科研信息化环境及其在科学考察中的应用	诸云强 孙九林 宋 佳 等 (204)
五、面向青海湖区域重要野生鸟类集成研究的e-Science应用	阎保平 罗 泽 周园春 等 (211)
六、面向空间天气学研究的信息化环境	王 赤 黄朝晖 邹自明 (221)
七、工业生物技术知识环境的构建及其应用	刘 斌 邢雪荣 马俊才 等 (227)
八、计算化学网格研究与应用	张瑞生 杨 裔 胡荣静 等 (233)
九、碳汇碳排放与环境监测大规模无线传感网典型应用	刘云浩 (240)
十、构建社会科学研究中的信息技术应用平台	王国成 (247)
后记	(259)

第一篇

态势战略篇



一、大力发展科研信息化，服务国家科技创新

江绵恒
(中国科学院)

(一) 把握科研信息化的发展内涵

信息化是当今时代的大背景，是社会发展的一个重要主题。科研信息化是当今时代科技活动最鲜明的特征之一，是科研模式的重大变革，也是迈向新一轮科技革命的必经之路。

所谓信息化，是指信息技术在材料、器件、系统方面的发展使得信息的产生、获取、传输、存储、处理、应用形成了一个系统。早在古代我们的祖先就用烽火台来传输信息；马拉松则是用跑步来传输信息；到了今天，信息的传输则是以光速来表征。

当信息技术的发展催生了以信息化为主要特征的新产业，包括信息制造业和信息服务业，且当这样的新产业成为社会生产力的主要方面，或者说信息化的过程成为提高劳动生产率的主要方式，我们则进入了信息社会。信息化是人类文明下一个发展阶段的主要特征。

为什么科研活动要信息化？因为科研活动是从数据到信息的感知过程和从信息到知识的认知过程。从感知到认知，我们有了发现、发明、创造，这是科研活动的内在规律，并且这个科研活动内在规律和信息化的逻辑内涵是相吻合的。现代的科学研究的依赖于海量数据、各种大科学装置、资源环境监测、数据密集型、计算密集型大规模并行计算模拟等。比如高能物理、生物信息等，依赖于高速的科研数据网络，海量数据的高速传输，同时是跨领域、跨地域的科研合作，这都需要有一个协同的环境。科学技术实际上是生产力，其中的生产方式在进行转变。因此科研活动要信息化，其意义在于使得跨学科、跨时间、跨空间的大规模科研合作的资源共享与协同工作成为可能，同时改变科学家们从事科研活动的方式和模式，极大地提高科研活动的劳动生产率。

(二) 发展科研信息化要面向国家科技创新战略

十七届五中全会指出，目前我国发展仍处于可以大有作为的重要战略机遇期，要以加快转变经济发展方式为主线，坚持把经济结构战略性调整作为主攻方向，把科技进步和创新作为重要支撑，提高产业核心竞争力，培育发展战略性新兴产业。这些都要求我们要高度重视科技创新工作，切实发挥科技是第一生产力的重要作用。

我们面临着一个历史性的时代机遇，那就是要通过城市化、工业化、信息化三化协同发展，齐头并进，来解决我们发展中遇到的问题。我们正处在一个信息社会的初始阶段，

信息社会的生产工具是信息网络。在城市化、工业化、信息化的进程中,土地、能源、矿藏、信息都是稀缺资源,要解决资源的稀缺问题,就必须依靠科技创新。

中国要完全走出一条自己的发展道路,我们的科技创新就要着力于资源的生产、分配和消费三大环节,而这些科技的创新可以催生出战略性新兴产业,就会产生新的就业机会。信息是信息时代新的生产力的一个关键要素,是信息社会的稀缺资源,在信息的生产、分配和使用方面,给我们提供了很多的机会,催生了包括硬件、软件、服务、电信、传媒等战略性新兴产业。

科研活动是科技创新的重要组成部分,在国家科技创新体系中占有重要地位。一方面,关系国计民生的很多创新成果、战略性新兴产业的很多源泉,都来自于能源、材料、信息、空间、生物医药等众多领域的科研活动。科研信息化直接服务于科研活动,以新的生产力形式提高科研活动的水平和效率,对于经济社会发展也就有着重要的意义。另一方面,科研信息化也是国家信息化战略的重要组成部分,是整个社会信息化的先导。它引领了国内信息化领域的发展,经过多年建设的科研信息化基础设施环境也是国家科技创新基础设施的重要组成部分,在信息化进程中所培养和锻炼起来的一大批人才队伍也成为了国家信息化建设的重要力量。

因此,我们要认清科研信息化的重要意义,明确科研信息化工作的出发点和落脚点,加强为国家科技创新战略服务的责任感与使命感。我们现在面对的最大问题就是如何将信息化的过程应用到经济社会发展的过程当中,实际上这方面的机会是巨大的。从事科研信息化工作的同志不仅要抓住机会,还要发挥引领作用。

(三) 科研信息化的发展趋势

随着信息技术的发展、社会需求的进步,科研活动向广度和深度延伸,科研信息化的理念与形态也在不断地发展之中。把握好发展趋势,将有力地指导科研信息化今后的工作。概括来说,当前科研信息化工作呈现出以下几个趋势。

1. 信息化与信息技术的融合互动

信息技术与信息化是相辅相成、融合互动的。信息技术是基础,在材料、器件、系统上的创新为信息化提供了支撑条件;信息化是手段,通过将信息技术应用于信息的产生、传输、计算、处理、应用等各个环节,改变人的科研、生产和生活方式,最终将信息技术形成新的社会生产力。

信息技术需要信息化提供出口,信息化也需要新的信息技术来提高水平。很多信息技术,正是有了信息化的应用,才有了真正的用武之地,并通过应用来检验技术的实效和水平,进而更好地提升技术,产生创新。

也有很多信息技术产生于科研信息化的过程中,并通过科研信息化的检验后,推广到生产生活和社会信息化过程当中。比如现在社会上很热门的物联网,很大程度上就源于无线传感网的研发。经过十余年的历程,无线传感网从实验室走进特定行业领域,最终成为广泛影响众多行业的新兴产业。又比如用于北京奥运大气环境监测系统的光学仪器、雷达技术、遥感技术等,之前都是分别研发的信息、光学、材料等技术,有了以任务为牵引

的科研信息化项目后,将各个相关的技术创新集成起来形成了更有成果的新的创新,并在任务结束后能够继续为社会的环境保护工作服务,产生持续的社会效益。

信息化与信息技术的融合互动,首先要从理念上加深认识,明确方向;其次要打破体制上的障碍,破除不同单位、不同部门之间的壁垒;最后就是要在开放合作的评价机制、利益分配机制上进行革新,探索可持续发展之路。

2. 海云体系

最近两年,大家关注的一个热点是云计算。云计算的概念及产生,打一个形象一点的比方,就是原始社会人类是把湖泊作为共同的水源,后来每户人家将自家的井作为水源,到了现代社会自来水成为公用设施。信息的处理和存储也经历了一个从集中到分散到再集中的过程——云计算是信息存储和处理的工业化过程,是信息服务业的基础设施和服务平台。

这里提出一个“海计算”的概念。在人与社会、人与自然、自然与社会的“新三网融合”应用中,有许许多多的应用并不一定都要到云端解决,特别是当信息的获取量大到足以使传输成为瓶颈时,也许很多信息的应用处理,可以在底层的“海”里加以解决。我们可以想象,其应用的领域是无穷无尽的,当信息化渗透到社会的各个方面,包括科研、商务、娱乐、社交、医疗、教育乃至其他社会生活,其技术创新和商业模式创新无外乎围绕三个方面展开:信息化的服务端、应用端以及两者相互联系的空间。未来科技创新和信息产业的兴起,海阔天空!

3. 以数据为中心

信息社会中信息是稀缺资源,信息的主要来源就是数据,因此要把科学数据作为科技创新的战略性资源,在很大意义上要建立起以数据为中心的科研思维,这不仅是指气象环境、海洋、生物物种、地质资源等国家资源类的数据,而且还包括基础研究、前沿探索和高技术开发过程中所产生的广泛的科学数据。

伴随着大科学工程装置、物联网技术在科研活动中的应用,海量科学数据的获取、传输、存储、处理、应用成为新的挑战。数据是信息的“原材料”,谁掌握了这一资源,谁就掌握了发展的主动权。因此要加强对科学数据的管理,建立有效的工具、体系和机制。

除了上述所讲的三个趋势以外,可视化、虚拟化、智能化、开放合作共享、安全、大科学等都是科研信息化的重要方向,需要我们在实践过程中不断凝练和探索。

(四) 中国科学院的科研信息化工作

中国科学院始终高度重视科研信息化工作,经过多年的建设,不断取得新的进展。“十一五”期间,中国科学院坚持以应用需求为出发点,在天文e-VLBI观测、宇宙起源、气候模拟、野外考察、环境监测等多方面开展了一系列的e-Science应用探索,并集中在基础、资源、生物、高技术四大领域部署实施了14个科研信息化示范应用项目。多个学科领域开展了信息化环境下新型科研方法的探索和实践,利用遥感数据采集、高速网络传输、超级计算环境、协同工作平台等信息技术手段,使得一些以前无法开展的工作成为可能,提高

了科学研究的效率和水平,加快了科学研究成果的转化速度。信息化对传统科研行为方式的变革和对科技创新跨越式发展的推动作用开始显现。

中国科学院在未来实施“创新2020”和落实“十二五”规划的工作进程当中,将进一步突出其科技创新的着力点,也就是要解决关系国家全局和长远发展的基础性、战略性、前瞻性重大科技问题。中国科学院的科研信息化工作,也将紧密围绕这一战略定位,以科技创新的需求为牵引,把握信息技术发展态势,进一步夯实科研信息化基础设施,提升基础设施的综合应用和服务能力,实现基础设施之间的互联互通与协同服务,完善科研信息化应用环境,部署一批直接服务于重大科研活动需求的应用平台,推进信息化与科研活动的深度融合,有力支撑中国科学院创新跨越、持续发展,为服务国家科技创新体系建设、发挥科技是第一生产力的作用作出新的贡献。

二、e-Science and Data-Intensive Research

以数据密集型科学为焦点的全新科学研究方法

Tony Hey, Jim Pinkelman
(Microsoft Research)

摘 要

本章重点介绍e-Science的发展历程并阐述e-Science技术在支持数据密集型科学方面的必要性。首先,简要介绍了海量科学数据的来源,并通过两个例子引入“第四范式”的概念。然后,提出数据的爆炸式增长给数据采集、管理和分析带来的新挑战,并着重讨论科学数据共享的优势和面临的困难。接着,指出在数据密集型科学作为“第四范式”出现的同时,Web和e-Science技术也正推动着一场科学交流的革命。在这场科技变革中,随着开放式存取的持续发展,大学图书馆和机构知识库即将在新的科学交流领域发挥核心作用。此外,本章还提及在最近两个关于数据密集型科学的政府倡议中,云计算、自然用户界面和语义技术等体现出日益增长的重要性。

Abstract

This paper sets out the e-Science agenda and explains how such e-Science technologies are needed to support data-intensive science. After briefly describing some of the sources of the scientific data deluge, the case for a “Fourth Paradigm” for scientific exploration is presented and illustrated with two examples. The explosion of data brings new challenges for data capture, curation and analysis. This discussion leads on to consideration of the benefits and difficulties of sharing scientific data. In parallel with the emergence of data-intensive science as a fourth paradigm, the Web and e-Science technologies are fuelling a revolution in scholarly communication. The Open Access movement continues to grow and it is argued that university research libraries and institutional repositories will play a central role in the new scholarly communication landscape. Two recent government initiatives on data-intensive science are reviewed before the paper is concluded with a brief mention of the growing importance of Cloud computing, natural user interfaces and semantic technologies.

Introduction

The last few decades of scientific research have been uniquely influenced and shaped by computing. Scientists have increasingly come to rely on computing technology for almost all aspects of their research—to automate and control experiments, to collect and analyze data, and to model systems and run simulations. However, in recent years scientists have been confronted with a new challenge: how to manage, manipulate, visualize, and mine data sets that are several orders of magnitude larger than they have had to work with in the past.

In this paper, we begin by briefly describing some of the sources of this data deluge. We will then introduce the concepts of the “Fourth Paradigm” and the e-Science tools and technologies required for data-intensive research^[1]. To illustrate these concepts, we give some brief examples of data-intensive science that show how e-Science computing technologies and data management techniques are being used to make use of large data sets to accelerate research. We then describe some of the challenges researchers face during the three major data-related activities in data-intensive science: capture, curation, and analysis. Next, some of the principles, policies, and motivations of sharing data are explored. This is followed by a discussion of some of the emerging trends and transformations that are occurring in the field of scholarly communication and open access. Government and policy initiatives in both the United States and Europe—with regard to data-intensive science—are covered prior to set of brief conclusions.

The Data Deluge

Experimental scientists now have a vast array of electronic devices and systems capable of producing very large volumes of data. Small, single-purpose sensors such as optical sensors, accelerometers, strain gauges, and thermal sensors can be deployed in mass arrays and used to collect data over distributed regions or a wide range of environments^[2]. Similarly, orbiting satellites collect vast quantities of images of the Earth in many parts of the electromagnetic spectrum and the raw data need considerable post-processing to generate useful data sets for scientists. After the success of the pioneering Sloan Digital Sky Survey^[3], astronomers are planning ambitious new telescopes such as the Large Synoptic Survey Telescope^[4] and the Square Kilometer Array^[5]. Both these projects will generate many hundreds of petabytes of data and require petascale or even exascale computational resources to process the data.

Particle physicists are now running experiments with enormously complex detectors to record the proton-proton collisions at the Large Hadron Collider (LHC) at CERN in Geneva^[6]. Detailed simulations of the experimental detectors are needed to model and understand their characteristics. Real-time computation is also required to reduce the raw collision rates down to manageable volumes for later offline analysis, by keeping only a rare, interesting subset of these raw events. In June 2011, CERN announced that the LHC detectors had collected data from about 70 trillion proton-proton collisions. Finally, the data have then to be distributed around the world for analysis by the hundreds of participating researchers. Each experiment at the LHC is expected to generate multiple petabytes of data annually.

In biology, we are on the threshold of a stunning explosion in the amount of genomics data being collected. This revolution is being driven by the large-scale automation of the task of gene sequencing. The new generation of gene sequencing machines is able to sequence approximately 10^{10} nucleotides per day: this is equivalent to each machine being able to

sequence more than three entire human genomes daily! The Beijing Genomics Institute, now known as BGI and located in Shenzhen, China, is one of the world's leading sequencing centers: its sequencing output is expected to exceed more than 15,000 human genomes per year^[7]. BGI's 500-node supercomputer also currently processes approximately 10 terabytes of raw sequencing data per day. In addition to DNA sequence data for a variety of animals, crops, and other organisms, there is a correspondingly large increase in gene expression, protein structures and other related data for each of these projects^[8].

A major area that combines both data collection and modeling is that of weather and climate simulations. As an example of the data challenges in this area, the fifth Coupled Model Intercomparison Project (CMIP5) will involve the global production and analysis of several petabytes of data^[9]. To support this project, involving more than 20 international modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) has established the global Earth System Grid Federation (ESGF) of data producers and data archives. The ESGF will provide a set of globally synchronized views of globally distributed data—including large cache replicants, some of which will be kept for decades (at a minimum). The project will stress international networks, as well as the data archives themselves.

Another large source of data is from large-scale computer simulations, today running on petascale supercomputers and currently being planned for execution on exascale systems. This data needs to be saved for two reasons—firstly, to provide a checkpoint for restart in case of failure during long running simulations, and secondly, for visualization and analysis of the results of the simulation. A study of fusion simulations identified the need to output 2 gigabytes of data per simulated time step for each core in the parallel simulation. For “just” a million cores, this corresponds to 2 petabytes of data per time step and requires an aggregate input/output (I/O) rate of 3.5 terabytes per second for a 10-minute time step in a simulation of 1 billion cells and 1 trillion particles^[10].

Finally, although the data volumes in these “big data” science fields are impressive, there are also difficult data challenges for “small data” science. Peter Murray-Rust has used the term “long-tail” science to characterize the scientific research that takes place in tens of thousands of ordinary laboratories^[11]. This often involves “wet” science experiments, for which digital data are also collected as part of the experiment. Typically, there will be no detailed advance planning of their data requirements, as would be required for “big data” science projects. Long-tail science includes parts of biomedical science, chemistry, materials, crystallography, and other disciplines. The amount of data involved is typically only in the terabyte range but is very heterogeneous—in contrast to LHC data or gene-sequencing data. These scientists are often confronted with the challenge of how to manage hundreds of thousands—or more—files, together with the need for a reliable and persistent data repository system with the capacity to reliably store their data sets.

e-Science and the Fourth Paradigm

The term “e-Science” was introduced in 2000 by John Taylor, then director general of the UK Research Councils^[12]. Taylor had been director of HP Labs Bristol, UK and e-Science was a natural extension of HP’s concept of IT as a utility—or “computing as a service”—which is now fast becoming a reality with Cloud services from companies such as Amazon, Google, and Microsoft. In his new role with the Research Councils UK, Taylor recognized the increasingly important role that IT must play in the collaborative, multidisciplinary, and data-intensive scientific research of the twenty-first century. He therefore launched the UK e-Science Programme to develop the tools and technologies needed to support this networked, data-intensive future of scientific research. We shall therefore use the term “e-Science” to encompass the collection of these tools, technologies, and infrastructure that are required to support such research.

In parallel with this development in the UK, Jim Gray from Microsoft Research had also seen these “big data” trends and begun working with scientists to understand the problems they had working with such large data sets and what kind of tools they needed to help alleviate these problems^[11]. Gray simply defined e-Science as where “IT meets scientists.” As we have seen, researchers are using many different methods to collect or generate data—from sensors and charge-coupled devices (CCDs) to supercomputers and high-throughput sequencing machines. Gray frequently encountered scientists who were “drowning in data”—what happens when you have 10,000 Microsoft Excel spreadsheets, each with 50 workbooks in them? It was clear to him that the world of science was changing and the scientist was becoming more remote from the actual raw data.

“The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.”^[13]

Originally, there were just two methodologies for conducting research—experimental science and then theoretical science, with Newton’s Laws of Motion and his theory of gravity, with Maxwell’s equations, with the Schrödinger equation, and so on. These are the two classic paradigms of scientific research and they have underpinned all of science up to the mid-twentieth century. Then, as the systems being studied became more complex, the theoretical models became just too complicated to solve analytically, even with sophisticated approximation methods. Fortunately, electronic digital computers arrived in the 1950s and people began the numerical simulation of such models on computers. Indeed, for some problems—like climate change or galaxy formation—simulations are the only way we can explore the physics. Ken Wilson, Nobel Prize winner in Physics in 1982, called computational