区域生态与环境过程系列丛书

DIQIU KEXUE SHUJU GONGXIANG
DE TIAOZHAN YU SHIJIAN

地球科学数据共享的挑战与实践

——以中国西部生态与环境科学数据中心为例

王亮绪 李新 著

YI ZHONGGUO XIBU SHENGTAI
YU HUANJING KEXUE SHUJU ZHONGXIN WEILI

科学出版社

目（2016YFC0502726）
国家自然科学基金重点项目（91025001、90502010）　　资助出版
中国科学院西部行动计划项目（KZCX2-XB2-09、KZCX2-XB3-15）

区域生态与环境过程系列丛书

地球科学数据共享的挑战与实践——以中国西部生态与环境科学数据中心为例

王亮绪　李　新　著

科学出版社
北京

内 容 简 介

本书以地球科学数据共享为核心，以研究科学数据共享的概念、内容与技术为起点，阐述科学数据共享的理念，设计了一套用于科学数据中心业务的共享流程，包括数据共享流程、数据申请流程、离线申请管理流程以及数据发布流程，定义了元数据评审和文献支持两种科学数据共享平台需要的扩展功能。结合地球科学数据的汇交与共享，以3个具体的共享实践为例，阐述了地球科学数据共享的具体实践。并基于黑河综合遥感联合试验数据共享服务的实际成效来分析和评价地球科学数据共享，分析用户申请数据的时间热点、空间热点以及主题热点，对数据文献进行引证分析。

本书适合科学数据共享研究人员以及高等院校相关专业的师生阅读，有助于读者初步理解科学数据共享的内涵和技术。

图书在版编目（CIP）数据

地球科学数据共享的挑战与实践：以中国西部生态与环境科学数据中心为例 / 王亮绪，李新著. —北京：科学出版社，2019.1
（区域生态与环境过程系列丛书）
ISBN 978-7-03-059332-0

Ⅰ.①地… Ⅱ.①王… ②李… Ⅲ.①地球科学－数据共享－研究 Ⅳ.①P ②G253

中国版本图书馆CIP数据核字（2018）第249457号

责任编辑：许 健 / 责任校对：谭宏宇
责任印制：黄晓鸣 / 封面设计：殷 靓

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

南京展望文化发展有限公司排版
当纳利（上海）信息技术有限公司印刷
科学出版社发行 各地新华书店经销

*

2019年1月第 一 版 开本：B5（720×1000）
2019年1月第一次印刷 印张：7 3/4
字数：152 000

定价：70.00元

（如有印装质量问题，我社负责调换）

区域生态与环境过程系列丛书
序言

十八大以来，党中央高度重视生态文明建设。中共十八届五中全会强调，实现“十三五”时期发展目标，破解发展难题，厚植发展优势，必须牢固树立并切实贯彻创新、协调、绿色、开放、共享的发展理念。同时提出：坚持绿色发展，必须坚持可持续发展，推进美丽中国建设，为全球生态安全做出新贡献；构建科学合理的城市化格局、农业发展格局、生态安全格局、自然岸线格局，推动建立绿色低碳循环发展产业体系；推动低碳循环发展，建设清洁低碳、安全高效的现代能源体系，实施近零碳排放区示范工程；加大环境治理力度，深入实施大气、水、土壤污染防治行动计划，实行省以下环保机构监测监察执法垂直管理制度；筑牢生态安全屏障，坚持保护优先、自然恢复为主，实施山水林田湖生态保护和修复工程，开展大规模国土绿化行动，完善天然林保护制度，开展蓝色海湾整治行动。

作为我国经济最发达、城市化速度最快的地区，长江三角洲（简称“长三角”）城市群也面临着快速城市化所带来的一系列环境问题。快速城市化的过程常伴随着土地覆被、景观格局的变化而改变了固有下垫面特征，在城市中形成了特有的局地气候，导致城市热岛及极端天气的频繁发生，严重危害人们的生命财产安全。此外，工业化过程所引起的大量化学物质的使用和排放更对区域生态环境造成了莫大的威胁。快速城市化过程中所出现的环境问题，其核心还是没有很好地尊重自然，没有协调人地关系，没有把可持续发展作为区域发展的最核心问题来对待。因此，我们需要在可持续发展思想的指导下，进一步加强城市生态环境研究，以促进上海及长三角区域的可持续发展。

上海师范大学是上海市重点建设的高校，环境科学是上海师范大学重点发展领域之一。1978年，上海师范大学成立环境保护研究室，开展了长江三峡大坝环境影响评价、上海市72个工业小区环境调查、太湖流域环境本底调查和崇明东滩鸟类自然保护区生态环境调查等工作，拥有一批知名的环境保护研究专家。经过三十多年的发展，上海师范大学现在拥有环境工程本科专业、环境科学硕士点专业、环境科学博士点专业和环境科学博士后流动站，设立有杭州湾生态定位观测站等。2013年，上海师范大学为了进一步加强城市生态环境研究，成立城市发展研究院。城市发展研究院将根据国家战略需求和上海社会经济发展要求，秉承“开放、流动、竞争、合作”原则，进一步凝练目标，整合上海师范大学学

科优势，以前沿科学问题为导向，以社会需求和国家任务带动学科发展，构建创新型研究平台，开拓新的学科发展方向，建立国际一流的研究团队，加强国际科研合作，更好地为上海建设现代化国际大都市提供智力支撑。城市发展研究院将重点在城市遥感与环境模拟、城市生态与景观过程、城市生态经济耦合分析等领域开展研究工作。通过城市发展研究院的建立，充分发挥上海师范大学在地理、环境和生态等领域的学科优势，将学科发展与上海城市经济建设和社会发展紧密结合，进一步凝练学科专业优势和特色，通过集成多学科力量，提升上海师范大学在城市发展研究中的综合实力，力争使上海师范大学成为我国城市研究的重镇和政府决策咨询的智库。

该丛书集中展现了近年来城市发展研究院中青年科研人员的研究成果，既涵盖了城市污泥资源化的先进技术、新兴污染物的迁移转化机制及科学数据应用于地球科学的挑战，也透过中高分辨率遥感与卫星遥感降水数据，分析极端天气的变化趋势及变化区域，通过反演地表温度，揭示城市化过程中地表温度的时间维、空间维、分形维的格局特征，定量分析了地表温度与土地覆被、景观格局、降水和人口的相关关系。同时从环境变化和区域时空过程的视角，对城市环境系统的要素、结构、功能和协调度进行分析评价，探讨人类活动影响对区域生态安全的影响及其响应机制，促进区域环境的可持续发展。该丛书有助于我们对城市化过程中的区域生态、城市污泥资源化、新兴污染物的迁移转化、滑坡灾害防治、景观格局变化、科学数据共享、环境恢复力以及城市热岛效应等方面有更深入的认识，期望为政府及相关部门解决城市化过程中的生态环境问题和制定相关决策提供科学依据，为城市可持续发展提供基础性、前瞻性和战略性的理论及技术支撑。

上海师范大学城市发展研究院院长

院士

2016年6月于上海

前　言

科学数据已渗透到科学研究的方方面面，科学数据已进入“大数据”时代。科学数据共享现今具有重要的意义和迫切的实际需求。进行科学数据共享，可以支持科学研究的再现和验证，也有助于研究者利用现有数据研究新的科学问题，更有助于提升科学研究和创新水平，促进科学的可持续发展。但由于科学数据共享的复杂性，尤其是在地球科学领域，科学数据的收集、汇交、发布、共享、激励、评价等仍面临诸多挑战。

本书以地球科学数据共享为核心，在回顾科学数据共享发展历史的基础上，从研究科学数据共享的概念、内容与技术为起点，阐述科学数据共享的理念，以科学数据共享平台的设计与实现为工具，结合科学数据的汇交与共享具体实践，分析和评价科学数据共享的实际成效。

科学数据共享的本质就是科学数据的开放和共用，并使其在共享过程中发挥出更大的价值。为实现共享的本质目标，科学数据共享平台包括三个核心部分，即数据汇交和整理、数据管理以及数据发现和获取，分别针对数据提供者、科学数据中心以及数据用户，科学数据共享平台的核心任务就是满足这三类用户的实际需求。科学数据共享的动力主要有两个方面，即推动科学的发展以及寻求科学研究的再现和验证。而科学数据共享面临的挑战是多方面的，包括数据量的持续增长和数据存储的挑战、保证数据可持续共享的挑战、科学数据汇交的挑战、科学数据质量的挑战等。国内在科学数据共享方面也存在对应的挑战，同时存在其他问题，包括没有明确的科学数据汇交政策、不完善的科学数据共享体系、对数据的产权还不够重视等。

本书在分析科学数据共享需求的基础上，探讨了科学数据共享涉及的各利益相关者的实际利益需求，设计了一套用于实际业务的科学数据共享流程，包括数据共享流程、数据申请流程、离线申请管理流程以及数据发布流程，并定义了元数据评审和文献支持两种科学数据共享平台需要的扩展功能。以元数据为核心，对内联系科学数据实体，对外实现科学数据共享功能。以现有开源技术为基础实现了一套功能可扩展的科学数据共享平台，并在科学数据共享实践过程中逐步优化、调整相应的功能，核心功能覆盖了科学数据的发现、获取、发布和服务等，并以此数据共享平台为核心，扩展了科学数据共享的辅助功能，包括科学数据的知识挖掘功能以及科学数据库的对外共享功能，以应对科学数据共享的诸多挑战。

在进行科学数据共享实践过程中，本书介绍了三个具体的共享实践，即中国西部环境与生态科学数据中心的数据共享实践、黑河综合遥感联合试验的数据共享

实践以及黑河计划数据管理中心的数据共享实践。在中国西部环境与生态科学数据中心的实践过程中，形成了一套适合科学数据共享的方法体系，包括保证科学数据的有效信息传递、保证科学数据共享中利益相关者的权益、保证科学数据共享的可持续发展等。在黑河综合遥感联合试验数据共享实践过程中，对共享系统进行了拓展，对科学数据及期刊文献进行了对应管理，并采用元数据分析方法介绍了黑河综合遥感联合试验的数据概貌。在黑河计划数据管理中心数据共享实践过程中，从制度和技术上实现了科学数据汇交支持，更加注重保护数据作者的权益，同时集成了黑河流域生态水文观测数据库，解决了黑河计划数据共享的实际挑战。

基于黑河综合遥感联合试验数据共享服务的实际成效，本书对科学数据共享的服务成效进行了探讨。在服务成效分析上，从用户申请的数据结果分析了用户申请数据的时间热点、空间热点以及主题热点，从申请数据的用户行为分析了用户的时间、单位、项目及其与黑河综合遥感联合试验项目组的关系。在文献分析方面，对黑河综合遥感联合试验数据产出的文献进行了分类，认为其包括专题文献、引证文献以及数据作者推荐文献三个类别，并从期刊来源、文献作者、研究主题三个方面进行了分析。同时针对黑河综合遥感联合试验专题分析了科学数据共享在提升文献引用方面的积极作用。利用定量化的访问、下载、引用数据，分析了黑河综合遥感联合试验数据集在共享时这三类定量数据之间的关系，并分析了从访问到下载、从下载到引用的统计规律。

最后，基于科学数据共享的实践结果，我们尝试总结科学数据共享的成效并评价科学数据共享过程。利用科学数据中心的成效分析数据，发现科学数据中心也存在马太效应和长尾效应。利用网络信息计量学方法和网站访问记录方法，评价了西部数据中心的共享成效。在这两个科学数据共享评价的实例基础上，我们参照期刊文献评价方法提出了一种面向数据的科学数据共享评价框架，定义了单一科学数据评价模型以及科学数据集合评价模型，并以西部数据中心的数据集和黑河综合遥感联合试验数据集为例探讨了科学数据共享评价实例。

感谢国家自然科学基金重点项目中国西部生态环境科学数据中心（90502010）与面向黑河流域生态—水文过程集成研究的数据整理与服务（91025001）、中国科学院西部行动计划黑河流域生态水文遥感试验项目（KZCX2-XB2-09、KZCX2-XB3-15）在研究过程中给予的支持，感谢科技部重点研发项目（2016YFC0502726）和上海市高峰高原学科在经费上的资助。也感谢共同参与相关项目研究的各位课题组成员。

作者在科学数据共享领域研究尚浅，书中不免有纰漏之处，望读者批评指正。

作　者

2018年2月

目　录

区域生态与环境过程系列丛书序言
前言

第 1 章　地球科学数据共享现状 …… 1
1.1　地球科学数据 …… 1
1.2　国内外研究现状 …… 2
1.2.1　国际数据共享研究现状 …… 2
1.2.2　国内数据共享研究现状 …… 6
1.3　本书章节安排 …… 8
1.3.1　本书主要内容 …… 8
1.3.2　本书结构 …… 9

第 2 章　科学数据共享的理念、挑战与关键技术 …… 11
2.1　科学数据共享的研究内容 …… 11
2.1.1　什么是科学数据共享 …… 11
2.1.2　为什么要共享科学数据 …… 12
2.1.3　科学数据共享的挑战 …… 13
2.1.4　我国科学数据共享面临的主要挑战 …… 15
2.1.5　科学数据共享流程 …… 16
2.1.6　科学数据共享的研究趋势 …… 17
2.2　科学数据共享的主要技术 …… 17
2.2.1　元数据技术 …… 17
2.2.2　数据标识与引用 …… 20
2.2.3　数据访问与获取 …… 21
2.3　本章小结 …… 22

第 3 章　科学数据共享平台的设计与实现 …… 23
3.1　科学数据共享的流程设计 …… 23
3.1.1　科学数据共享涉及的各利益相关者 …… 23
3.1.2　科学数据共享的核心需求 …… 24

3.1.3 科学数据共享的核心业务流程设计 ……25
3.1.4 科学数据共享的扩展功能设计 ……26
3.2 科学数据共享平台的实现 ……27
3.2.1 整体框架 ……27
3.2.2 基于用户群的功能实现 ……28
3.2.3 关键问题 ……30
3.3 科学数据共享平台的扩展功能实现 ……32
3.3.1 元数据发布与评审 ……32
3.3.2 科学数据的知识挖掘 ……33
3.3.3 关系型数据库的共享 ……34
3.4 本章小结 ……35

第 4 章 科学数据中心的建立与共享实践 ……36
4.1 实践1：中国西部环境与生态科学数据中心 ……36
4.1.1 西部数据中心的数据分析 ……36
4.1.2 西部数据中心的组织架构 ……37
4.1.3 西部数据中心的数据服务 ……38
4.1.4 西部数据中心的共享平台 ……38
4.1.5 问题与讨论 ……40
4.2 实践2：黑河综合遥感联合试验数据的共享 ……41
4.2.1 试验介绍 ……41
4.2.2 数据整理及共享政策 ……42
4.2.3 数据共享系统 ……43
4.2.4 元数据分析 ……45
4.2.5 问题与讨论 ……51
4.3 实践3：黑河计划数据管理中心 ……53
4.3.1 数据政策 ……53
4.3.2 汇交与共享系统 ……54
4.3.3 与黑河流域生态水文观测数据库的共享集成 ……56
4.3.4 元数据分析 ……57
4.3.5 问题与讨论 ……60
4.4 本章小结 ……62

第 5 章 科学数据共享的成效分析：以WATER数据共享为例 ……63
5.1 WATER数据的服务分析 ……63

5.1.1 用户申请的数据分析……63
5.1.2 申请数据的用户行为分析……67
5.2 WATER数据的文献分析……69
5.2.1 文献分析……69
5.2.2 文献与数据……72
5.3 WATER数据共享的定量化分析……76
5.3.1 WATER数据的查看统计……76
5.3.2 WATER数据的下载统计……77
5.3.3 WATER数据的引用统计……77
5.3.4 从查看到下载的追踪分析……79
5.3.5 从下载到引用的追踪分析……80
5.4 本章小结……81

第6章 科学数据共享的评价探讨……82
6.1 科学数据共享的认知……82
6.1.1 马太效应……82
6.1.2 长尾效应……84
6.2 西部数据中心的共享评价……84
6.2.1 西部数据中心的网络影响力评价……85
6.2.2 基于网站访问日志的数据影响力评价……87
6.3 面向数据的科学数据共享评价框架……89
6.3.1 科学数据共享评价的可用数据……91
6.3.2 单一科学数据共享评价……92
6.3.3 科学数据集合共享评价……93
6.3.4 科学数据共享评价方法试验……94
6.3.5 问题和讨论……95
6.4 本章小结……95

第7章 结论与展望……96
7.1 结论……96
7.1.1 以元数据为核心的科学数据共享平台……96
7.1.2 科学数据汇交实践……96
7.1.3 科学数据共享实践……97
7.1.4 科学数据共享评价……97
7.1.5 本书创新之处……97

7.2 问题与展望……98
7.2.1 提升用户的共享参与度……98
7.2.2 重视科学数据资源的长期建设 ……98
7.2.3 进一步提升知识挖掘能力 ……99
7.2.4 面向机器的科学数据共享服务 ……99
7.2.5 持续推动科学数据的出版与引用 ……99

参考文献…… 101
缩略词表…… 113

第1章　地球科学数据共享现状

1.1　地球科学数据

数据是数和量的组合，它既表示事物的多少、大小，也是判断事物或分析事物的客观依据（刘红和胡新和，2012）。科学数据在科学活动中产生，是科学研究的基础。

人们对科学数据的理解经过了不同阶段，在实验科学和理论科学阶段，数据代表了一系列的观测结果，是理论抽象和知识形成的基础，因此被认为是信息和知识的原始材料。对于很多复杂问题，理论知识或实验难以解决，人们开始寻求计算模拟的方法，数据在这一阶段表现为存储在计算机中的数字化原材料，其模拟结果表现为智力劳动和科学研究的成果，是知识的一种表现。

进入信息化时代，全球数据量迅速增长，大数据时代已经到来（Hey & Trefethen, 2003），催生了以大数据为研究对象的数据密集型科学研究，“科学就是数据，数据就是科学”已渗透到科学活动的方方面面（Hey et al., 2009）。但是，很多用户仍然面临找不到所需数据的问题（孙九林，2003），一方面，大量投入所产出的科学数据得不到充分利用和有效管理（Ailamaki et al., 2010）；另一方面，许多科学数据需求得不到满足，科学数据的供需矛盾依然突出，并已成为当前科学界亟待解决的尖锐问题（Ran et al., 2007）。科学数据的有效管理和共享能推动科学的发展（Christensen et al., 2011; Arie et al., 2007; Mooney & Winstanley, 2007）。

科学数据具有两种基本属性，即社会属性和自然属性。科学数据的社会属性表现为数据的观测和分析是科研经费支持的结果，是全社会的共同财产，理应为全社会服务（程津培，2004）。科学数据的自然属性也是科学数据的科研价值，首先，科学数据不仅能为原始观测目标服务，还能为其他研究服务，即数据具有再利用价值（黄鼎成，2003）；其次，科学成果和研究结论是在科学数据的基础上产生的，科学数据决定了科学结论的真伪，因此科学数据的公开应该是学术成果发表的必要环节。科学数据必须进行共享才能充分发挥其价值（黄鼎成，2003；李集明，2003；刘闯和王正兴，2002）。

地球科学的研究范围涉及水、土、气、生、人各个圈层。地球科学数据具有自身的特点，如空间性（所描述的客观实体是地球系统的一个部分，具有明显的空间位

置特征）、时间性（多数数据带有明显的时间范围信息）、类型多样性（数据有各种各样的表现方式，如空间数据、栅格数据、属性数据等）、多源性（有不同的数据来源，如遥感、航空、地面测量、计算机模拟等）、海量（随时空分辨率的变化，数据量会急剧上升）、明显尺度特征（所描述的客观实体在不同时空尺度上具有不同的表现特征）等（廖顺宝等，2005）。因此，研究地球科学数据的共享方法更具有挑战性。

随着大数据概念的提出，数据（信息）已经与能源、材料一起被认为是三大战略资源。2012年，美国政府投资2亿美元专门从事大数据研究及发展计划。*Synthese*（Bonilla, 2014; Grasswick, 2010; Whyte & Crease, 2010）、*Science Engineering Ethics*（Timmermann, 2014; Bezuidenhout, 2013; Macrina, 2011; Hackett & Rhoten, 2011; Fischer & Zigmond, 2010; Frugoli et al., 2010; Giffels, 2010; Giffels et al., 2010; Joshi & Krag, 2010; Taylor, 2009; Pascal, 2006）、*Sciences*（Akil et al., 2011; Baraniuk, 2011; Carpenter, 2011; Curry, 2011; Evans & Foster, 2011; Fox & Hendler, 2011; Greve & Svenning, 2011; Jasny et al., 2011; Kahn, 2011; King, 2011; Lang, 2011; Los & Wood, 2011; Mathews et al., 2011; Overpeck et al., 2011; Reed, 2011; Reichman et al., 2011; Rowe & Frank, 2011; Severin, 2011; Field et al., 2009; Kaiser, 2009; Kleppner & Sharp, 2009; Hahn, 2008）和*Nature*（Butler, 2013, 2007; Monastersky, 2013; Nature, 2013a, 2013b, 2009a, 2009b; Priem, 2013; Russell, 2013; Van Noorden, 2013; Wilbanks, 2013; Nelson, 2009; Schofield et al., 2009; Roberts & Chavan, 2008; Dittert et al., 2001）等知名期刊相继采用撰文或开设专刊等形式讨论科学数据的管理和共享等问题。

1.2 国内外研究现状

1.2.1 国际数据共享研究现状

国际上，科学数据共享大致可分为以下三个研究阶段。

1. 第一阶段（20世纪初至20世纪80年代）：科研需求导向的初级数据共享

人们对数据的价值并不是一开始就如此重视的。在理论科学阶段，实验数据仅被用作发现和验证科学理论的根据，在同一理论指导下，实验数据可以被重复获取，数据不具有再利用价值。即使在现代科学研究中，很多研究者依然不需要数据或者研究结束后不需要保存数据。生物统计学创始人Francis Galton在1901年

介绍了统计学在处理原始数据中的方法和意义，并强调了原始数据共享的理念，成为首位提出原始数据共享的科学家（Galton, 1901）。Fienberg等出版了*Sharing Research Data*一书，首次较为系统性地阐述了科学数据共享的好处、利益相关方、数据共享耗费、数据共享方式等内容（Fienberg & Martin, 1985）。

为满足国际极地年（International Polar Year, IPY）的数据需求，国际科学联合会理事会（International Council of Scientific Unions, ICSU）在1957年成立了世界数据中心（World Data Center, WDC），WDC成立的最初目的是协调IPY从各个国家获得研究所需要的数据，以收集、存档和发布多种地球物理和太阳数据为主要任务（Ruttenberg, 1992）。在随后半个多世纪的发展过程中，WDC所涉及的学科和领域越来越广，逐渐成为国际数据共享领域的一支重要力量。目前，WDC在全世界共有52个学科中心，分布在美国、欧洲、中国、日本和印度等国家和地区（王卷乐和孙九林，2007）。

1966年，国际科学联合会理事会又成立了另外一个重要的国际性数据组织——国际科技数据委员会（Committee on Data for Science and Technology, CODATA），致力于提高对整个科技领域有重要意义的数据的质量、可靠性、管理与可访问性，旨在推动和鼓励对科学技术有重要意义的、可靠的数值数据进行编辑、评价、传播。其目标是：① 增加数据的质量与可访问性；② 促进数据专家和研究者之间的国际合作；③ 不断提高国际社会对数据共享重要性的认识；④ 考虑数据存取和知识产权问题。与WDC重视数据管理和数据共享相比，CODATA更加重视数据的科学价值，如数据质量、研究合作和知识产权。

WDC和CODATA的成立是国际数据共享第一阶段的两个重要标志性事件，WDC的成功数据管理经验为以后成立数据中心和研究计划数据管理提供了必不可少的基础。WDC所提出的“Full and Open”共享政策被数据共享界普遍接受，并成为美国的数据共享政策。之后，各国纷纷参照WDC模式建立了很多国家级科学数据中心，并在规模上已远远超过了WDC系统本身。

2. 第二阶段（20世纪80年代至20世纪末）：宏观政策主导的国家科学数据共享体系

随着人们对科学数据的重视，各个国家都投入了大量的经费用于数据采集和分析，数据得到不断的积累，但科学家并不能及时获得他们所需要的数据。科学界对于数据共享的呼声越来越高，呼吁政府采取措施解决这一问题。

美国是首先建立国家数据共享体系的国家。1990年NASA决定建设分布式活跃数据档案中心群（Distributed Active Archive Centers, DAAC）。DAAC的建立标志着美国国家层面上的科学数据共享工作的开始（刘闯和王正兴，2002）。1994年，将9个DAAC的数据资源建成为一体的数据信息网络（Earth Observing System

Data and Information System, EOSDIS)。1995年美国启动全球变化数据信息系统(Global Change Data and Information System, GCDIS)项目。该项目是一个比DAAC规模更大、涉及内容更广泛、牵涉的部门更多、层次更高的一个长期项目,DAAC是GCDIS的主要组成部分。与DAAC由NASA代表国家管理不同,GCDIS由白宫直接协调管理。1999年,DAAC为了扩大其服务的领域,建立起DAAC联盟(DAAC Alliance),并全部加入ICSU组建的WDC系统中。

在欧洲,英国、法国、德国、荷兰和瑞典等国非常重视数据管理与共享。欧盟的《关于数据库法律保护的指令》、英国的《布加勒斯特宣言》和《信息自由法》等,在科学数据的产权归属、共享管理和开发利用等方面均有明确规定,以保障科学数据共享活动的有序开展(肖永英,2003)。但目前尚未出现如美国完全开放制度性的共享环境,正在朝数据更加开放共享的方向努力。与美国的数据共享不同,欧洲的数据共享多是由科研资助机构推动的。此外,美国的数据共享是在科学数据共享"大循环"模式下开展的,数据共享由国家主导,国家开展规划、投资和管理,数据共享让全社会和整个国家受益。而欧洲的数据共享是从商业化运行的角度开展的,采取有偿共享,从市场上收回一定的数据成本,科研资助单位很少单独资助数据共享项目。

这一阶段的数据共享总体上是以数据中心建设为主体,数据共享强调对数据的管理,为未来的数据再利用奠定了基础。

3. 第三阶段(21世纪初至今):"科学即数据,数据即科学"

这一阶段正处于探索阶段,没有明确的开始时间。本书以2009年*The Fourth Paradigm: Data-Intensive Scientific Discovery*一书的出版作为这一阶段开始的标志。这一阶段数据共享的主要特点是数据已经正式成为知识发现环节的重要组成部分,更多的利益体开始关注数据,将数据管理作为知识管理和知识发现的一部分。

(1)很多科研资助机构明确要求研究项目提交数据管理计划,将数据成果管理和共享视为研究成果管理和共享的一种重要方式。例如,美国自然基金会2010年发布了项目管理指南,要求所有项目申请书必须包含数据管理计划(data management plan);英国的10个主要科研资助机构中有8个要求研究人员提供数据管理计划;澳大利亚研究理事会2008年起就鼓励发现项目(Discovery Projects)的研究者将研究成果存放在合适的学科库或机构库中(司莉和邢文明,2013)。

(2)出版界将数据视为学术成果的重要组成部分,数据的公开和共享不仅可以满足学术质疑的需要,还可以促进数据的再利用。很多著名的期刊纷纷要求作者在文章正式发表前将相关数据公开,如*Evolution*、*Molecular Biology and Evolution*、*Nature*、*PLoS Biology*和*Science*等。期刊的出版要求能有效推进科学数据的共享(Anagnostou et al., 2013)。此外,越来越多的期刊将数据出版视为学术成果出版

的重要形式，如Pensoft出版集团提出了Data Paper的概念，与文章相关的辅助数据可以“Data Paper”的形式发表，学术论文通过引用的方式建立文章与数据的关系（Penev et al., 2011）。生命和生物领域国际期刊联合成立了Dryad[①]数据中心，协助期刊开展数据的统一注册、发布和管理（Isard et al., 2007）。专门针对科学数据的期刊（Data Journal）开始出现，如*Data Science Journal*、*Geoscience Data Journal*、*Earth System Science Data*、*Ecological Archives Data Papers*、*Dataset paper in Science*、*Journal of Chemical and Engineering Data*、*Journal of Physical and Chemical Research Data*、*Biodiversity Data Journal*、*Scientific Data*等，可将科学数据以论文的形式出版发布。

（3）图书馆成为科学数据管理和共享的一个重要力量。图书馆长期以来以管理科学文献为主要目标，在学术交流体系中起到了至关重要的作用。随着数据在科研活动中的重要性的增加，数据也被视为知识的一种重要载体，数据与文献成为数字资源和知识管理的重要组成，数据共享也成为图书馆的一个重要研究内容。国际上很多图书馆已经开展科学数据服务，如NSF在2007年启动了以图书馆为主体的DataNet计划（Lee et al., 2009）；新墨西哥大学图书馆实施了DataOne项目，构建了分布式的地球观测数据管理系统（Michener et al., 2012; Michener et al., 2011）；麻省理工大学图书馆实施了对地理学科数据进行管理的MIT地理数据知识库；明尼苏达大学图书馆成立了科学研究网络基础设施联盟（司莉等，2013）。与数据中心关心数据的标准、组织和挖掘不同，图书馆参与数据共享更加关注数据作为知识成果的传播、关联和知识发现。

（4）数据知识产权得到重视，数据成果有可能成为评价科研产出成果的重要指标。作为知识的重要载体和表现，数据具有知识产权。尊重数据的知识产权应该作为数据共享的基本前提，只有这样科学家才有意愿共享数据。欧洲议会1996年发布了“Database Directive”，首次提出对数据库进行知识产权保护，但没有对非数据库形式的其他科学数据进行保护（Powell, 1996）。著名的知识共享机构CC（Creative Commons）针对数据著作权认证机制问题提出了Science Commons项目[②]。对科学数据如何合法的授权也是知识产权保护的一个重要内容（Ball, 2011）。在与数据相关的知识产权中，数据的署名权是科研人员普遍关注的一项基本权利，以数据期刊和数据论文为主要内容的数据出版体系的建立将为数据的署名权提供有力保证。针对数据成果的评价，国际上也正在提出积极的评价方法，如汤森路透（Thomson Reuters）旗下的知识产权与科技事业部推出了数据引文索引（Data Citation Index, DCI），采用类似文章的评价方法来评价数据的价值，以提供更好的知识发现方法（Torres-Salinas et al., 2014; Thomson Reuters, 2012）。

① http://www.datadryad.org.

② http://creativecommons.org/science.

1.2.2 国内数据共享研究现状

我国的数据共享研究起步较晚，数据共享的发展历程和特征与国际差异较大，以标志性事件进行分类，大致也可分为三个阶段。此分类与王巧玲等（2008）、张翔（2013）通过文献计量分析方法的分类结果不一致。

1. 第一阶段（20世纪80年代到2002年10月）：数据共享探索阶段

其标志性事件是1984年和1988年中国先后加入了CODATA和WDC两个重要的国际数据组织，建立了9个WDC学科数据中心，成立了CODATA中国委员会，组建了10个科技数据协作组（王卷乐和孙九林，2007）。这一阶段的数据共享主力为中国科学院，其自1982年起就持续资助“中国科学院科学数据库”建设，在“九五”末期建立了130多个专业数据库，数据总量725 GB（桂文庄，2007）。中国科学院还先后对下属的5个WDC数据中心进行了资助，WDC数据中心得以持续发展，在WDC开展的评估活动中，我国的WDC数据中心得到了较高的评价。虽然这一阶段的数据共享缺乏国家固定经费支持，但数据中心在数据共享政策、元数据标准、数据管理方面取得了宝贵的经验，为国家数据中心体系的建立积累了丰富的经验。

2. 第二阶段（2002年11月至2011年）：国家级数据中心体系基本建成

2002年11月，以“中国科学数据共享”为主题的香山科学会议的举行标志着我国第二阶段数据共享工作的正式启动，徐冠华、孙鸿烈、孙枢、程津培、秦大河和张先恩等与会代表呼吁建立一套完整的科学数据管理政策法规体系，使科学数据共享工作有法可依；加大国家在科学数据及其共享方面的投入，使数据共享成为一项公益性系统工程，促进国家科技水平提高和社会经济发展；并建议设立科学数据共享工程专项，对科学数据共享进行持续支持。

由科技部支持的“国家科学数据共享工程”于2002年开始试点，并于2004年正式实施（张先恩，2004）。其强调在国家统筹规划下，整合各种数据资源，纳入国家数据共享管理的统一框架内，形成跨部门、跨地区、跨学科、多层次、分布式的数据共享服务体系，为国家整体发展和科技水平提高提供可靠的数据资源保障。一批科学数据共享项目开始研究构建（孙九林和施慧中，2003）。

2006年中国科技基础条件平台中心成立，承担国家基础条件平台建设项目的过程管理和基础性工作。2006年中国科技资源共享网正式开通，这是国家科技基础条件平台门户网站，负责对数据中心的数据共享成效进行评估。2011年，科技部正式授予了20多家数据中心“国家级数据中心”的称号，标志着我国较为完整的