国际研究数据联盟及对我国科学数据共享的启示

王卷乐[1] 祝学衍[2] 石 蕾[3] 刘 鹏[1,4] 祝俊祥[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. 科技部基础研究司，北京 100862；3. 科技部国家科技基础条件平台中心，北京 100862；
4. 中国矿业大学（北京）地球科学与测绘工程学院，北京 100083；5. 中国科学院大学，
北京 100049）

摘 要：介绍了国际研究数据联盟（Research Data Alliance，RDA）的目标、主要任务、组织架构及运行机制。当前该组织已建立了约30个工作组和兴趣组，在数据引用、永久标识、元数据、数据分类编码、数据互操作等领域进展显著。基于调研和分析，提出了RDA对于我国科学数据共享在运行机制、组织机构、研究领域、参与机构、科学问题及信息化手段等方面的启示。

关键词：研究数据联盟；数据共享；研究基础设施；大数据；启示

中图分类号：TP399 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2014.02.003

Introduction of International Research Data Alliance and Its Enlightenments on Scientific Data Sharing in China

Wang Juanle[1], Zhu Xueyan[2], Shi Lei[3], Liu Peng[1,4], Zhu Junxiang[1,5]

(1. Institute of Geographic Sciences and Natural Resources Research Chinese Academy of Sciences, State Key Laboratory of Resources and Environmental Information System, Beijing 100101; 2. The Department of Basic Research, MOST, Beijing 100862; 3. National Science & Technology Infrastructure Center, MOST, Beijing 100862; 4. China University of Mining & Technology, College of Geoscience and Surveying Engineering, Beijing 100083; 5. University of Chinese Academy of Sciences, Beijing 100049)

Abstract: The paper introduced the objectiveness, mission, organization architecture, and operational mechanism of Research Data Alliance (RDA). About 30 work groups and interesting groups have been initiated in RDA, and it has made obvious progresses in data citation, permanent identification, metadata, data classification and coding, data interoperability etc. Based on the investigation and analysis, serials of inspiration obtained from RDA were proposed finally, including data sharing operational mechanism, organizational architecture, research domain, participant, science issue and informatization.

Keywords: Research Data Alliance, data sharing, research infrastructure, Big Data, enlightenments

作者简介：王卷乐*（1976- ），男，博士，副研究员，研究方向：资源环境信息集成与共享。**祝学衍**（1982- ），男，主要从事科技管理工作。**石蕾**（1982- ），女，副研究员，研究方向：科技资源管理。**刘鹏**（1991- ），男，硕士研究生，研究方向：遥感与地理信息系统。**祝俊祥**（1989- ），男，硕士研究生，研究方向：遥感与地理信息系统。

基金项目：国家科技基础条件平台专项，国家地球系统科学数据共享平台、中科院信息化专项项目“资源学科领域基础科学数据整合与集成应用”（XXH12504-1-01）。

收稿日期：2014年1月7日。

1　引言

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规范标准长期采集和管理的业务型科学数据；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、试验等站点采集的研究型科学数据（以下简称研究数据）。这两类数据都是科学数据共享中必须进行筛选、整合、集成，并为科技创新提供支撑服务的数据资源。然而，由于研究数据分散性大、分布面广、标准化程度低，且资源量大，给科学数据资源整合和共享带来了更大的难度[1]。美国国家科学基金会于2005年9月发布了《推动21世纪研究与教育的长期数字数据库》，指出研究型数据是一个或者若干个固定的研究项目产生的数据集，这些数据集中的数据只经过有限的处理与管理，一般只为特定的研究群体服务，标准化程度低[2]。这使得研究数据很难被共享和利用[3]。可见，尽管研究数据广泛存在于各科研团体中，但很难被共享和利用，如何推动此类数据的共享是国际关注的热点和难题。

针对研究数据共享，在大数据时代来临之际，美国、欧盟和澳大利亚于2012年8月共同发起了国际研究数据联盟（Research Data Alliance，RDA）。RDA第一次全体大会于2013年3月在瑞典哥德堡召开，宣布了RDA的正式启动。RDA的启动是数据共享领域的一件大事，但是RDA到底是一个什么样的组织，其对全球科学数据共享带来什么样的影响，其对发展中国家。尤其是中国会带来什么样的启示等是一系列值得思考的问题。笔者根据RDA官方网站（http：//www.rd-alliance.org/）公开的资料及实际参加RDA工作组的体会，系统介绍RDA的目标、主要任务、组织架构及运行机制，并初步分析其对我国科学数据共享的启示。

2　RDA的框架

2.1　目标和原则

RDA的目标是致力于推动全球数据驱动创新与发现，促进全球研究数据共享与交换，加强数据重复利用与开发，完善全球数据标准化。RDA希望通过在各学科领域间开展国际合作与研究，解决数据基础设施建设、政策、管理、标准化等数据热点问题。

RDA的指导原则是：（1）开放性（openness）。RDA吸纳所有对研究数据共享感兴趣并遵守RDA原则的个人和团体加盟。RDA的会议和工作过程都是开放的，其工作成果也将公开发布。（2）一致性（consensus）。RDA在工作过程中将采用适当的机制来避免和解决各方可能的分歧，并在达成一致意见的过程中推动研究数据共享。（3）平衡性（balance）。RDA努力寻求能够代表其成员和众多利益相关者间诉求的平衡。（4）协调性（harmonization）。RDA通过数据标准、政策、技术、基础设施和团体的协调和合作来推动进展。（5）团体驱动（community driven）。RDA是一个在秘书处协调下公众、团体驱动的机构，其由众多志愿者和组织者共同组成。（6）非盈利（non profit）。RDA不推销、宣扬和销售任何商业产品、技术或服务。

2.2　组织结构

RDA由理事会、秘书处、技术指导委员会、组织指导委员会、决策组、工作组、兴趣组等组成，如图1所示。

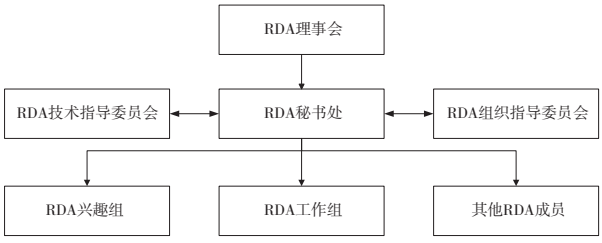

图1　RDA组织结构

理事会：全面负责RDA的日常监管、可持续发展并实现发展目标，审核批准成立符合RDA联盟目标的工作组。目前理事会有7人，未来两年内计划将增至9人。

秘书处：执行RDA日常运行和管理。多名秘书处工作人员分工负责协调不同的工作组和兴趣组。

技术指导委员会（TAB）：向理事会提供咨询和建议，协助监督和检查RDA各工作组的进展，负责完善和改进RDA发展的技术路线。技术指导委员会分两个阶段共吸纳了12名成员，第一阶段通过RDA理事会提名产生6名委员，第二阶段通过RDA大会推选产生其他6名委员。

组织指导委员会（OAB）：向理事会提供组织建议，其成员由参加机构的代表组成，主要对RDA的方向、过程和机制提供咨询。OAB负责RDA组织、规划和计划的改进和完善。OAB主席是理事会成员。

工作组：致力于推动驱动创新的开放研究数据共享。工作组的预期任务目标是促进研究数据的共享和交换、数据利用和重复使用、数据发现、数据服务和标准化的保藏、数据互操作等。

兴趣组：吸纳来自不同学科的研究人员和数据科学家共同工作，以确定共同的科学问题、研究目标和具体研究活动。兴趣组通过创建申请报告的形式把他们的想法表达为明确的行动计划，进而吸引更多对此感兴趣的一个或多个团队共同开展工作。

2.3 运行机制

RDA是一个依靠加盟团体驱动的组织，其运行机制的核心是多层次、多角度的广泛合作。这种机制是任何单一的研究领域、学科或者国家都无法创建的。RDA的核心运行机制是提供一个框架，在该框架内通过兴趣组、工作组和论坛来推动研究数据的自由、有效流通和共享服务。

RDA工作组和兴趣组是联盟的基础和核心力量，其组建需要RDA技术委员会审核、RDA理事会批准。工作组成立之前要先提交申请报告，描述工作组的目标、预期产出、受益者以及具体实现途径。该申请报告要首先经过RDA成员和技术委员会反馈意见，审核通过后，再由理事会复查，确保其与RDA的宗旨以及原则相一致。而对于兴趣小组的审核更为严格，首先要将兴趣小组的申请报告公示，充分汲取各方意见，技术委员会指定一名人员负责帮助兴趣小组复查，同时考量各方面在技术方面的意见，确保其与RDA技术发展路线一致。在公示环节至少需要4周以上时间，由公众评论审核该小组的申请报告。通过后再由理事会经过大约4周的复查，最终才能正式成立兴趣小组[4]。通常RDA给每个工作组或兴趣组的工作时间为一年到一年半。

RDA接纳加盟机构和个人一般通过3种途径：一是对于个人可以直接在RDA官网上注册成为成员。二是对于机构或个人可以申请加入和组建新的RDA工作组或兴趣组开展工作。三是直接注册参加RDA一年两次的全体大会。

加盟RDA的成员享有以下权利：可以与RDA各工作组交互；可以在RDA网站或大会上为人所识并可能成为全球研究数据共享领导者之一；可以为你所在的部门、领域或地理区域就数据互操作或RDA发表意见和观点；可以把你的需求和有关数据交换中的具体问题反映给RDA以得到咨询和建议；有机会作为测试站点使用RDA新制定的标准和协议；可以在工作进程中接受RDA正规的管理与指导。

加盟RDA的机构成员需要缴纳适当的会费。人数小于50名员工的机构缴纳1000美元/年；人数在50到250名员工的机构缴纳2000美元/年；人数大于250名员工的机构缴纳10000美元/年。

3 工作组与兴趣组的职责

RDA的核心工作内容由各工作组和兴趣组实施。RDA的兴趣组和工作组有固定的工作周期，其小组研究主题和小组的数量都是在动态发展变化的。截至2013年12月，RDA成立有8个工作组和21个兴趣组，各组列表见表1、表2所示。

以下仅针对RDA现有的工作组，简要概述其主要目标和职责。

（1）数据引用工作组（Data Citation）。旨在汇集一批专家来讨论目前有效地引用数据子集的科学问题、需求和现存方法的优点和不足。该工作组集中在一个较窄的领域，有利于实现有效的、机器可操作的数据引用，并通过实现参考原型，开展透明化的应用。该工作组意图同CODATA、OpenAire、datacite、W3C、开放联盟

注释等相关组织合作[5]。

（2）数据基础与术语工作组（Data Foundation and Terminology）。描述一个基础的、摘要的数据组织模型，用以派生相关参考数据术语。该数据基础与术语可用于使各团体和利益相关者更好的推动研究数据的概念同步化，促进团体内部和团体之间更容易理解的交流，促进支持数据服务的模型工具制造[6]。

（3）数据类型注册工作组（Data Type Registries）。为便于自动处理大量的科学数据、优化数据生产者和使用者的沟通方式，需要将不同的数据类型进行定义、注册，并永久关联到它们的描述数据。该工作组将编译一组代码用于数据类型注册和管理，制定数据类型注册表的数据模型和设计注册表的功能，并提出与现有的数据类型注册表联盟策略[7]。

（4）元数据标准目录工作组（Metadata Standards Directory）。主要任务是发展一个协调的、开放式的标准目录，解决科学数据和相关基础设施描述的问题。短期目标是基于维基百科建立一个针对科学数据的元数据标准目录；长期目标是在支持科学数据资源的流通和互操作的情况下，将元数据集实现最小化，形成一个全球认可的元数据标准[8]。

（5）永久标识信息类型工作组（PID Information Types）。在复杂的数据领域，永久标识符PID可以用于给每个数字对象赋予身份标识，使其指向数据资源和元数据，此外还能证明其完整性、真实性和其他属性。PID工作组致力于制定一套信息类型和组织结构连续的标识符系统，并且对跨团体的各个学科都适用[9]。

（6）实践政策工作组（Practical Policy）。此处政策表示为计算机可操作的规则，是工作组的研究重点。计算机实践政策被用于加强管理、自动任务维护、验证评估标准和自动科学分析等。本工作组将要组合配置一组生产和研究政策，分析已提交政策的实践作用，促进基于政策的数据管理系统构建[10]。

表1　RDA现有工作组

序	名称	状态	序	名称	状态
1	数据引用	待定	5	永久标识信息类型	通过
2	数据基础与术语	通过	6	实践政策	通过
3	数据类型注册	通过	7	数据分类和编码标准化	通过
4	元数据标准目录	通过	8	小麦数据互操作	待定

注：来源于https：//www.rd-alliance.org/working-and-interest-groups.html，2013.12.30。

表2　RDA现有兴趣组

序	名称	状态	序		状态
1	农业数据互操作	通过	12	合法互操作	通过
2	大数据分析	待定	13	研究数据的长尾效应	通过
3	生物多样性数据集成	通过	14	海洋数据协调	待定
4	数据共享服务中介	通过	15	材料数据、平台及其互操作	通过
5	数字资源库认证	通过	16	元数据	通过
6	团队能力模型	待定	17	电子基础设施保护体系	通过
7	数据上下文语义	待定	18	数据发布	通过
8	定义城市科学数据交换	待定	19	研究数据来源	通过
9	对发展中国家的云计算能力和教育发展研究	待定	20	结构生物学	通过
10	数字化在历史和民族志中的应用	通过	21	毒理基因组学的互操作性	通过
11	宣传动员组	通过			

注：来源于https：//www.rd-alliance.org/working-and-interest-groups.html，2013.12.30。

（7）数据分类和编码标准化工作组（Standardization of data categories and codes）。该工作组与ISO639合作，面向数据共享、数据发现以及数据库互操作制定分类编码，这将使得跨学科和本领域的研究人员受益，有利于提高数据采集和存储及共享利用能力。预期的产出成果包括：参与ISO639标准化TC37/SC2；建立TC37澳大利亚镜像委员会等[11]。

（8）小麦数据互操作工作组（Wheat Data Interoperability）。本工作组致力于提供一个遵从开放标准的，描述、表达和发布小麦数据的通用框架。该框架将持续推动小麦数据共享、重复利用和可操作。预期将研制成为一套“食谱”（cookbook），指导人们如何生产“小麦数据”，使其更容易地被共享、重用和互操作[12]。

4 对我国科学数据共享的启示

（1）运行机制符合大数据时代的要求。“大数据”的观念已经渗透到各个行业领域。在这一背景下，来自科学研究领域的科学数据如何面临全球共享机制下的挑战。RDA提出以“促进研究数据无障碍的全球共享和利用”为使命，以增强研究数据的开放和互操作为核心，广泛动员各个机构、科学家个人以加盟的方式开展工作，这种依靠团体驱动的运行机制适合于大数据的时代需求。其实，数据联盟的思想，在地球系统科学领域早在1990年代就建立了“地球系统科学联盟”，国家地球系统科学数据共享平台也创建了“地球系统科学数据共享联盟”[13]，但这些联盟成员和机构还只是局限于地学领域，并未建立跨越如此多学科领域的研究数据联盟。RDA创建的这一广泛合作的联盟机制降低了机构、科学家个人加盟的门槛，有助于实现其推动全球研究数据共享的理念。RDA首先由美国、欧盟和澳大利带头发起也符合欧盟关于全球基础设施共享[14]的发展思路。

（2）组织机构建设具有活力。RDA采用以理事会为核心，以技术咨询小组和组织咨询小组为指导，以工作小组和兴趣小组为生力军的组织结构，可以保障整个联盟的高效、科学运作。其组织机构在发展中，首先建立兴趣小组凝练科学问题、探索性地开展工作，进而形成任务更具体的工作小组具体实施，在这个过程中使得工作目标更科学、具体、清晰，能够解决实际问题。各小组的工作周期固定在12～18个月，便于目标考核和过程监督。各小组工作方案完全透明、公开，研讨充分。这种在组织机构保障下的“化整为零”的工作模式，能够克服RDA自身领域过大、难于顶层设计的弊端，反而通过自底向上、以问题为导向的小组式工作模式，循环推进，短期内易形成阶段性成果。RDA自2013年3月启动时的不到20个工作和兴趣小组，能够在不到一年的时间内发展到30多个工作和兴趣小组，显示了这一组织机构的活力。

（3）数据共享领域广泛。RDA在总体围绕提高研究数据互操作性的总目标下，广泛吸纳各领域参加，各小组的设置不拘一格。目前已有农业、城市发展、历史和民族文化、海洋、生物、基因等领域的参与，已建立了农业数据互操作、小麦数据互操作、城市数据交换、历史和民族志数字实现、结构生物学、毒理基因组学互操作、海洋数据协调等实体工作组和兴趣组。在这种模式和趋势下，可以预见，RDA的工作和兴趣小组还将在更多领域拓展，例如大气科学、地理科学等。

（4）参与机构多样。通过调研和实际参加RDA工作发现，RDA的成员来源非常广泛，涉及政府机构、科研教育、商业出版等各个方面。例如，有国际知名的数据库机构，如国际蛋白质数据库（Protein Data Bank，PDB）；有国际大型数据共享组织，如CODATA、WDS；有国际知名的出版商，如Springer、Thomson Reuters等；有数据服务公司，如英国的Trust-IT Services公司等；有研究机构，如澳大利亚联邦科学与工业研究组织（CSIRO）、微软研究院等；有大学教育机构，如英国的牛津大学、美国的加利福尼亚大学等。

（5）研究命题深入数据共享难题。分析发现RDA的研究小组制定的研究主题有两个主链条，一是围绕数据本身的生命周期设计，简称数

据链，二是围绕领域的扩展应用设计，简称领域链。许多研究命题非常深入，是突破数据共享瓶颈的关键问题。在数据链上，从数据的采集、产生、加工、描述、编码、分类到出版和引用等各个环节设定研究工作小组，例如建立有数据永久标识、数据术语、数据类型注册、数据分类编码、元数据目录、数据中介、大数据分析等工作和兴趣小组。这些研究命题对于解决长期困扰数据共享的知识产权保护等难题具有重要意义。具有在领域链上，尽管涉及的领域非常广泛，但其核心问题都集中在数据互操作上，例如建立的农业数据互操作、小麦数据互操作、毒理基因组学互操作、海洋数据互操作等。通过这些不同领域的研究对比，更容易形成有关研究数据互操作的基础理论、方法和未来的产业化应用。

（6）工作方式信息化特点显著。RDA充分利用网络平台开展工作值得借鉴，例如其会议网站平台、Twitter参加模式、电视电话会议等非常先进，极大地提高了各成员间的交流效率。

5　结语

本文通过最新资料调研，介绍了RDA的目标、任务、组织机构和运行机制等总体情况。RDA以工作组和兴趣组为主要推进力量的这一运行机制是其具有活力的重要原因。当前，RDA已建立了约30个工作组和兴趣组，已在数据引用、永久标识、元数据、数据分类编码、数据互操作等领域取得进展。在大数据时代来临之际，我国科学数据共享也面临着诸多挑战。建议我国科学数据共享能够借鉴RDA在运行机制、组织架构、研究领域、参与机构等方面的做法，进一步破解我国科学数据共享难题，健全和发展既与国际数据共享接轨又符合我国国情的科学数据共享机制和模式。

参考文献

[1] 孙九林,王卷乐.探索分散科学数据资源共享之路——记“地球系统科学数据共享网”[M]//国家科技基础条件平台.国家科技基础条件平台回顾与展望.北京:中国科学技术出版社,2008.

[2] National Science Foundation. Long Lived Digital Data Collections: Enabling Research and Education in 21st Century[R]. September 2005.

[3] 王卷乐.科学数据整合集成与共享中的关键技术问题研究:以研究型、参考型数据为例[R].2007.

[4] Research Data Alliance[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/group−process−procedures.html.

[5] Research Data Alliance. Working Group On Data Citation[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/groups/data−citation/wiki/working−group−data−citation−case−statement−proposal.html.

[6] Research Data Alliance. Working Group on Data Foundation and Terminology[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/groups/data−foundation−and−terminology/wiki/rda−case−statement−data−foundation−and−terminology−dft.

[7] Research Data Alliance. Working Group on Data Type Registries[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/groups/data−type−registries/wiki/data−type−registries−proposed−case−statement−rda−working−group.html.

[8] Research Data Alliance. Working Group on Metadata[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/groups/metadata−standards/wiki/case−statement−proposal−rda−metadata−working−group.html.

[9] Research Data Alliance. Working Group on PID Information Types[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/pid−information−types−charter.html.

[10] Research Data Alliance. Working Group on Practical Policy[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/working−groups/practical−policy−wg.html.

[11] Research Data Alliance. Working Group on Standardisation of Data Categories and Codes[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/working−groups/standardsation−data−categories−and−codes−wg.html.

[12] Research Data Alliance. Working Group on Wheat Data Interoperability[EB/OL]. [2013−12−30]. https://www.rd−alliance.org/working−groups/wheat−data−interoperability−wg.html.

[13] 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.

[14] Costantino Thanos. A Vision for Global Research Data Infrastructures[J]. Data Science Journal, 2013,12(13): 71−90.