

中文翻译版

研究数据的管理与共享 最佳实践指南

Managing and Sharing Research Data

A Guide to Good Practice

Louise Corti
Veerle Van den Eynden
Libby Bishop &
Matthew Woollard

主编

殷沈琴 伏安娜
薛 崧 张计龙

译



科学出版社

中文翻译版

研究数据的管理与共享 最佳实践指南

Managing and Sharing Research Data
A Guide to Good Practice

Louise Corti Veerle Van den Eynden
Libby Bishop Matthew Woollard

主 编

殷沈琴 伏安娜 薛 崧 张计龙 译

科 学 出 版 社

北 京

图字：01-2017-7079

内 容 简 介

人文社会科学和自然科学研究过程中积累了大量的研究数据，这些数据如何管理、共享、出版和引证？如何再次利用他人的数据？过程中牵涉到哪些知识产权？本书的几位作者不仅熟悉全球尤其欧美的研究数据管理的实践应用，而且在英国数据档案馆拥有多年数据相关专业工作的管理与实践经验，他们在本书中将多年积累的最佳数据管理实践建议、指导和培训倾囊呈现。本书前三章对研究数据管理和共享进行总括介绍，包括管理和共享的重要性、研究数据生命周期以及如何制订研究数据管理计划。第四章至第六章介绍了研究数据管理的具体方法，包括数据文档编制、数据格式与组织以及数据存储与传输。第七章至第十一章讨论了数据管理与共享所涉及的问题，包括法律和伦理、知识产权、合作研究、利用他人数据以及出版和引证。

本书涵盖所有重要的研究数据技能、大量翔实的案例研究和实际操作解析，架起数据管理理论和实践之间的桥梁，为不同学科领域的数据相关专业教学和普及培训提供了一本通用教材，具有较强的实践指导性和可操作性。

图书在版编目（CIP）数据

研究数据的管理与共享：最佳实践指南/（英）科尔蒂（Louise Corti）等主编；殷沈琴等译. —北京：科学出版社，2018.3

书名原文：Managing and Sharing Research Data: A Guide to Good Practice
ISBN 978-7-03-056510-5

I. ①研… II. ①科… ②殷… III. ①数据管理—研究 IV. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 022583 号

责任编辑：丁慧颖 / 责任校对：韩 杨

责任印制：肖 兴 / 封面设计：陈 敬

This simplified-Chinese-character edition is translated from the English edition of *Managing and Sharing Research Data—A Guide to Good Practice* first published in 2014 by Sage Publications (the Proprietor) based in the USA, the UK, India and Singapore, and the translation is published by arrangement with the Proprietor.

本简体中文版仅限在中国大陆地区（不包括台湾、香港特区和澳门特区）发行。

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>

天津市新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 3 月第 一 版 开本：720×1000 1/16

2018 年 3 月第一次印刷 印张：15 1/4

字数：278 000

定价：68.00 元

（如有印装质量问题，我社负责调换）

致 谢

我们首先感谢英国国家经济和社会研究理事会（ESRC）自 1967 年以来，为英国数据服务中心和它的前身提供核心资金。正是因为这种持续性的资金支持，我们建立了世界上顶级的国家数据档案馆和基础设施，并且吸引了高技能的研究人员和技术支持人员加入我们中心。ESRC 的资助还使我们能够在开发前瞻性的能力建设方法方面保持领先地位，而这又帮助研究人员获得了更好的数据管理实践经验。我们在此同样感谢英国联合信息系统委员会（JISC）近期所提供的大量科研数据管理方面的资助机会，使我们从中受益匪浅。这些激励着我们在地方科研数据管理实践、政策和基础设施的认识和实施中做得更好。

同时我们感谢 Camille Corti-Georgiou、Sue Wood 和 Anne Etheridge 在审校和准备这份书稿方面给予的帮助。英国数据档案馆的一些工作人员亦为提供适合这本书的具体练习部分做出了贡献，他们分别是 Laurence Horton、Bethany Morgan-Brett 和 Mus Ahmet。

我们衷心感谢已故的 Alasdair Crockett 博士，他于 2006 年参与了编写第一本英国数据档案馆《管理和共享数据》指南。

最后，感谢我们所有的合作伙伴：Rob Lenart、Nick Snow、Allen Radtke 和 Penny Woollard，我们常常一起进行高强度的夜以继日的写作、修订和编辑工作。

中文版序

首先，我们想借此机会赞扬复旦大学大数据研究院人文社科数据研究所和社会科学数据研究中心团队所付出的努力，他们为促进中国研究数据管理与共享的能力建设开了先河。复旦大学研究团队仍在不断完善他们的数据平台，扩充研究数据和数据服务，它属于国际数据档案联盟的一部分。殷沈琴老师和她的同事具有管理和共享数据、运行数据仓储平台的专业知识和实践经验。此外，他们还倡导了中国高校重要的研究数据计划。2014 年，他们联合其他八所大学共同发起成立了“中国高校研究数据管理推进工作组”。因此，他们具有很大的优势来进一步推动中国的数据共享议程。

殷老师在复旦大学开设一门研究生课程——数据监护研究与应用，这门课程使用了我们的英文版书籍作为教材，为此我们感到非常荣幸。这种能力培养的进步，我们乐见其成，同时也热烈地欢迎我们的书（英文版）在他们培训课程中的使用。中文译本将有助于指导学术研究、研究机构和智库的研究数据应用和管理实践。

2012 年，我们意识到，应该将我们在英国数据服务中心（UK Data Service）多年的数据专业工作经验记录下来，这些经验包括向英国的研究人员提供数据管理、监护和分享方面的建议和长期专业的培训，以供研究人员、数据管理和监护专业人士参考和使用。Sage 出版社向我们介绍了市场的需求，在之后 18 个月的时间里，我们撰写了一本融愿景、知识和建议于一体的指南，涵盖我们的研究数据管理和二次分析实践经验和相关建议，旨在提供培训和学习场景下的数据实践经验和技术的指导。此书在 2014 年春季出版，获得广泛好评，已在 46 个国家售出了近 2000 本，并已在全球范围内被 40 门课程采用。

近年来，我们都有在中国讲解数据课程的机会。2015 年，Louise 在哈尔滨工业大学开办了一个为期 5 天的暑期班，讲解关于数据科学方法的课程。“生物大数据”课程是针对国际生命科学学院的研究生开设的，它是年度暑期学术研讨会——“芯片、计算机、作物”（C3）的一部分，C3 是由浙江大学陈铭教授和埃塞克斯大学 Andrew Harrison 博士共同组织举办的，在中国各大城市已经举办了 7 年。Louise 所授课程涉及研究数据的管理、数据归档和出版等领域，强调研究复制的必要性，开展讲座和学生们亲自动手练习。该课程结束之时，学生们将自己的数据发布到一个实时测试库中。此外，2015 年，Veerle 与美国密歇根大学校际政治与社会研

究联盟（ICPSR）的同仁一起在北京大学讲授了 3 天课程——“研究数据的监护与管理”，殷老师也参加了该课程。

我们不仅发现了参与者教学经历的积极之处，还观察到学生们如何超出预期热切地参与和学习共享研究数据的新实践。我们希望这本书可以为中国各学科未来的课程和暑期班提供支撑，并且非常乐意未来有机会为中国的数据共享使命提供支持。

我们为殷老师和她的同事以及中国社会科学领域的同仁献上最美好的祝愿，希望此书能够帮助大家更好地欣赏和领悟数据共享的艺术之美。

Louise Corti, Veerle Van den Eynden

2017 年 10 月 2 日

序 言

随着各个领域科学的进步，研究者对研究数据职责的认识开始发生变化。而且越来越多的研究资助者强制要求研究数据的开放获取；各国政府也对科学研究的透明性提出了更高的要求；而经济环境对数据重用的需求也越来越大；同时因害怕数据丢失，也需要更多稳健的信息安全实践。所有这些因素都意味着研究人员将需要提高、增强他们的数据管理技能并使之专业化，以应对如下挑战：使用一种可靠而有效的方式来提供最高质量的可共享、可重复利用的研究成果。研究人员的数据管理技能的提高为英国和其他国家的研究能力建设计划提供了一种战略性的支持。

稳健的科研数据管理技术赋予了研究人员和数据专家能够应对数据管理环境中快速而不均衡发展的技能。英国、美国和欧洲其他国家的研究资助者正在逐步实施数据管理（和共享）政策，这一政策旨在最大限度地提高数据的开放性、透明度和明确受他们资助的研究对象的职责。因为在出版物出版前要对其进行同行评议，所以越来越多的期刊出版机构也逐步要求研究人员提交其研究数据。研究资助者和数据使用者也越发认识到准备充分的数据的长期价值。而研究机构也需要优质的研究信息基础设施来应对其数据资产可能面临的伦理问题和安全威胁。

基于以上情况，本书包含了有关研究数据管理和共享方面最新的且易于理解的信息。它涵盖了研究数据生命周期中所有重要的研究数据技能，还包括翔实的案例研究和实际操作讲解，这将有助于读者更好地理解研究数据管理中的核心概念及其应用。这本详尽的指南将作者多年积累的最佳数据管理实践建议、指导和培训呈现给广大的研究人员。这涉及对专题研究项目的深入讨论，并且通过与研究人员紧密合作，开发出易于应用在标准科研活动中的解决方案和工具。我们将研究数据定义为：任何由原始数据收集或生成的、定性或定量的，及在研究项目过程中通过分析现存数据资源所得到的研究材料。其范围涵盖了数值数据、文本数据、数字化材料、图像、记录及建模的脚本。

本书适用于各级水平的研究人员，因此，无论是刚入学的硕士还是博士，抑或经验丰富的教授和研究团队负责人，都可以从本书中有所收获。对初级研究者而言，本书能够帮助其打下坚实的研究基础。对于有经验的研究人员，本书能够帮助其填补当前知识体系中的某些缺陷，并且更新他们在快速发展迅速变革的技术领域的知识，与此同时，还能帮助他们了解研究数据管理和研究伦理方面立法变化的最新动态。此外，本书也为那些专业的研究人员的实际操作提供了最好的

实践材料。尽管这本书是用社会科学研究者的思维来撰写的，但书中涉及的绝大多数研究问题和研究方法仍然能够对大多数自然科学和社会科学研究人员在数据方面的工作有所裨益。

本书的目标对象还包括当前快速增长的研究机构和政府部门中负责数据管理和共享的研究支持人员、研究资助机构、研究管理者、研究伦理监管机构、IT 服务人员或图书馆数据支持人员，他们均可以通过阅读本书获得帮助。

尽管本书并非按照既定的顺序来展开，但我们仍然推荐读者按照本书的编写顺序来阅读各章节，这样可以循序渐进地掌握其中的内容。本书中的练习能够帮助读者更加牢固地掌握本书的知识，读者通过完成练习，可以得心应手地将在本书中学到的研究技巧应用于自己的实际研究中。

接下来各章节主要介绍了数据管理中的关键要素，而这些要素对于安全地处理和共享数据具有极为重要的作用。第一章主要介绍了哪些关键因素推动了数据共享的出现，我们通过共享数据能够获得哪些好处，以及从实践角度介绍如何进行数据共享。第二章介绍了研究数据生命周期的概念及其如何延长典型的研究周期。第三章主要解决如何针对研究数据制订管理计划，包括如何使用数据管理清单、如何分配角色和职责、如何预估研究项目的数据管理成本及研究中数据管理的资金支持等问题。

第四章描述了如何编制定性和定量数据的文档、提供数据背景和信息出处及在研究数据集中如何描述课题和数据文件等细节问题。第五章涵盖了对数据进行格式的编排和组织，具体包括文件格式、数据转换、文件和文件夹的组织形式、数据质量保障、版本控制和真实性保证、数据转录和数据的数字化。第六章讨论如何储存和传输数据，主要包括了以下内容：如何才能最好地进行数据备份、确保信息安全、数据传送与加密、数据处理、数据长期存储和保存及如何进行文件共享和协作。

第七章我们探讨了研究的伦理和隐私问题，介绍如何处理与数据共享有关的法律和伦理问题，以及通过何种途径获得数据。这些问题主要包括知情同意书、统计披露控制、数据的匿名化及访问控制措施的方法。第八章介绍了数据知识产权的相关知识，主要包括版权、数据库和其他权利及如何在现行权利框架下对现存的数据资源进行利用或再利用。第九章介绍了协同研究项目中的数据策略，包括如何制订标准的操作协议、操作流程和共享资源，以及在团队中如何协调数据记录、分配数据角色和职责。第十章通过分析六个真实的数据再利用的案例，来介绍如何利用他人的研究数据，寻找再利用数据的机会和应对其中所产生的挑战。第十一章我们探讨了发布和引证数据，这一部分主要包括：寻找发布数据的场所及如何通过使用永久数字标识符技术来创建可供引证的数据资源。

最后，读者可以通过访问英国数据服务的网站 <http://ukdataservice.ac.uk/manage-data/handbook>，来获取本书其他的配套练习资料。

目 录

第一章 管理和共享研究数据的重要性	1
第二章 研究数据的生命周期	17
第三章 研究数据管理计划	23
第四章 编制数据文档和提供数据的背景信息	36
第五章 数据格式和数据组织	56
第六章 数据存储与数据传输	87
第七章 数据共享中的法律和伦理问题	108
第八章 与研究数据有关的权利	147
第九章 合作研究：研究团队和研究管理者的数据管理策略	162
第十章 利用他人的研究数据：机遇和局限	173
第十一章 研究数据的出版和引证	201
结语	219
缩略词表	221
英中名词对应表	223
译者后记	231

第一章

管理和共享研究数据的重要性

研究数据不仅是科学知识、学习和创新的基石，而且也是我们寻求理解、解释和发展整个人类和社会的基石。在数字时代，研究数据不仅呈指数级增长，而且现时代的数据很容易存储、保存和在世界各地交换。对技术进步带来的优势的需求在日益增长，这个优势使我们处理和利用研究数据的方式更加现代化。

在 1953 年，Watson 和 Crick 在《自然》(Nature) 上发表了关于 DNA 结构的单页论文，当时没有任何原始数据来支持他们的这一发现。在近期，千人基因组计划 (the 1000 Genomes Project Consortium, 2010) 的研究人员在《自然》上发表了他们的最新研究成果，伴随着 4.9T 的 DNA 序列数据在其项目网站上公布，同时也存储在单核苷酸多态性数据库 (dbSNP) 中 (Kiermer, 2011)。以上的遗传学研究只是一个例子，用来证明开放和交换信息 (包括研究数据) 如何加速我们的研究和探索进程。我们就能够认识到随着遗传学研究的逐步深入所带来的医学进步。

自 2000 年以来，数据共享的驱动力及进行数据共享的人力和物力都呈现出爆炸性增加。研究资助者正逐步促进研究数据和数据计划被便捷开放地获取，以确保研究数据的质量、可持续性、可获取性和开放程度的最大化。学术研究成果的出版商要求其能够访问学术研究成果背后的支持性数据，用于审校或进一步探索。各国政府正在提出对研究透明度的要求，此外当前的经济环境也使得数据再利用变得更加广泛，而这种数据重复利用可以使科学投资回报最大化。许多研究人员认为，数据获取不足阻碍了科学的进步。

数据可获取意味着，如果需要，可以对科研成果进行验证和审查。社会需要获取数据的目的：企业能够利用新知识开发工具和应用程序；允许组织质疑政府的政策和决定；让成千上万的公民参与到研究过程或“公民科学”中，以推动我们公共科学知识的发展。

因此，在不同的科学领域，研究人员对他们研究数据的管理职责也在不断地变化。研究人员的研究数据管理能力需要完善、提高和专业化，以应对如下挑战：

使用一种可靠且有效的方式来提供最高质量的研究成果和可持续性数据，并且有能力共享和再利用这些科研成果。

通过数据管理，即为使研究数据达到最高质量而进行的数据实践、操作、改进和处理过程，我们必须要对其进行良好的组织、记录、保存，以保证数据的可获取、可持续和可再利用。数据管理技能的提升为英国和其他国家的研究能力、建设计划能力提供了一种战略性的支持。而科研机构也需要高质量的科研数据管理，用以解决其数据资产可能面临的伦理问题和安全风险。稳健的科研数据管理技术赋予了研究人员、数据专家和科研支持人员相应的技能，该技能能够应对数据管理环境中快速而不均衡的发展状况。

数据共享议程

研究人员一直以来都认识到了共享的重要性：在科学出版物中分享研究成果；通过同行网络分享专业知识；并与学习型社会进行协作。技术的进步使这种共享达到一个新的水平，并以不同的方式加以应用，包括通过开放渠道访问的研究出版物及研究数据、工具、软件和教育资源。20 世纪 90 年代初期，人们呼吁在线开放已发表的研究文章，后来逐步扩展到更多的原始研究材料。

加速开放研究数据的主要驱动因素是经济合作与发展组织（OECD）获取公共资金研究数据的原则和准则，以及关于开放获取科学和人文科学知识的“柏林宣言”。

OECD 准则宣称：由于公共资金资助的研究数据是一种基于公共利益而生产出来的公共物品，因此应当在不危害知识产权基础上，以及时和可靠的方式，极少限制地对公众开放（OECD，2007）。《柏林宣言》（Berlin Declaration，2003）呼吁通过互联网的开放获取范式来促进知识的传播，而这一方式需要国际互联网的可持续、可交互和透明化支持，以及易获得且可兼容的内容和工具的开放。

欧盟科学数据高级专家组在报告中指出，在整个欧洲层面数据量存在不断增长的趋势，并提议我们需要一个科学基础设施来支持数据的无缝访问、使用、重复利用和可信任性，来应对即将由数据所激发的科研能力的巨大飞跃（European Commission，2010）。这份报告描述了加速发展科学数据基础设施的收益和成本。开放的基础设施、开放的文化和开放的内容需要齐头并进。

英国皇家学会（The Royal Society，2012）认为：为了从巨量的数据中获得潜在的利益，研究人员和研究机构应该加大对同行及公众开放数据程度，并加强对研究数据的认识。通用数据标准的采用和数据出版的授权是同等重要的。在技术层面，报告强调了对科学领域数据管理专家的需求，以及对使用工具来分析大数

据流的需要，以便最大限度地发挥科学数据的潜力。

数据共享的重要性体现在数据储存库的指数型增长。目前，数据库检索工具——Databib 登记表中列出了 518 个数据库，其中 42 个是具体的社会科学数据库，有将近 200 个数据库是 2000 年之后创建的，如图 1.1 所示(Databib, 2013)。

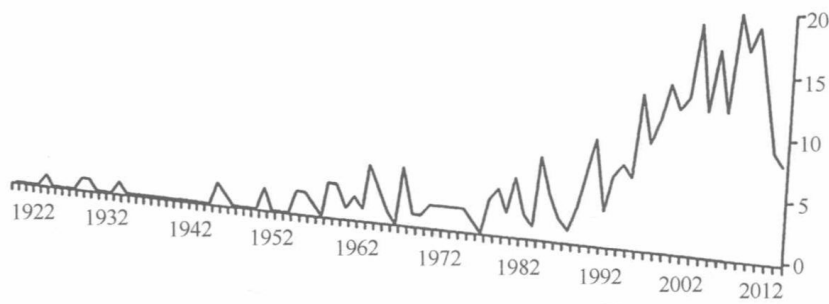


图 1.1 每年新创建的数据库数量增长情况图

来源：Databib, 2013

各国政府和组织也乐于采用开放数据来提高其各项活动的透明度。开放他们的信息供任何人进行访问和使用，增强了公共部门信息的经济和创新潜力。自 2009 年以来，全球内各国政府、地区和国际组织启动了 200 多项开放数据举措(CTIC, 2013)。高度开放的数据门户有世界银行目录、联合国目录、联合国数据中心和欧盟开放数据平台。这些门户向个人、企业和学术界开放组织信息。欧盟委员会还与成员国就现有网站之间的数据格式和互操作性开展合作 (European Commission, 2011)。

在英国政府的《开放数据白皮书》上清楚地解释了公民、企业和公共部门有望从政府和公共服务中获得开放数据的益处 (Cabinet Office, 2012)。该文件以标准化、机器可读和开放的格式快速发布数据并且制订了明确的标准，使数据可以自由和便捷地获取，并且发布的是默认格式的政府数据。政府的数据门户网站 (data.gov.uk) 拥有 9500 多个数据集，还展示了如何将政府数据用于创新应用、报告、地图、政策和服务。

2013 年夏天，美国奥巴马政府通过启动开放数据项目，表明了对非分类联邦政府数据透明度的承诺。其目的是给公众和私营部门公司提供更多的数据，用于创新和商业发展 (Higgins, 2013)。虽然开放获取议程一直促进学术界发布和获取更多的公共数据，我们同时也看到了“公民科学”的兴起。公民科学是指普通公众凭借自发的努力和热情在业余参与到数据的收集和分析 (CSA, 2012) 中。星系动物园 (Galaxy Zoo) 可能是其中最著名的，如下面的案例研究所示。

案例研究

公民科学联盟

公民科学联盟（CSA）是由科学家、软件开发人员和教育家合作创立的一个机构，他们通过共同开发、管理和利用基于互联网的公民科学项目，以推动科学本身的发展，并加强公众对科学和科研过程的理解。

公民科学项目的理念是让公民用他们的业余时间、兴趣和能力参与到科学项目中。动物宇宙（Zooniverse）中的各个项目都是由提供初步想法的科学家团队、一些有效的资助及用户导向的受众给予其创作灵感和支持。项目研究范围广泛，从古典科学到气候科学，从生态学到行星科学。

最有名的星系动物园（<http://www.galaxyzoo.org/>），要求其公民研究人员根据星系的形状来将其分类。星系动物园始于2007年，是由斯隆数字巡天调查项目（Sloan Digital Sky Survey）捕捉到的上百万个星系所组成的数据集。令人惊讶的是，在项目发起后的24小时内，平均每小时接收到近7万次分类。第一年就收到了来自15万人发来的约5000万份分类。

最近的挑战仍是分析来自哈勃空间望远镜最远的“超深”图像，望远镜使用的是新的广角摄像机3，它在最后一班穿梭飞行任务期间安装成功。

研究资助机构数据管理和共享政策

在英国、美国和欧盟，资助者和出版商的数据共享政策已成为数据共享和加强数据管理的重要驱动因素。它们是基于OECD的原则，即公共资助的研究数据是一种公共物品，应尽可能减少限制地提供给公众（OECD，2007）。

其实，早在20世纪90年代中期的英国，一些公共研究资助者就已经采用了数据共享政策，其标志是在2011年所有的英国研究委员会成员一致通过了关于数据政策的共同原则（RCUK，2011），之后又通过了研究成果的获取政策（RCUK，2012）。这要求从2013年4月起，所有由理事会资助的、经同行评议后的研究论文发表在一些指定的期刊上，如果这些期刊对当前数据公开政策不是很满意，它们刊登一则关于如何获取潜在的研究材料，如数据、样本和模型的声明。要知道在全英国高等教育机构中，55%的研究是由研究理事会所资助的，每年的资助金额高达30亿英镑，这些政策对研究实践产生了深远的影响（HESA，2012）。

在本书写作之时，以下的英国研究资助者已经公布了其数据共享政策。

- 艺术和人文研究理事会（AHRC）
- 生物技术和生物科学研究理事会（BBSRC）
- 英国科学院

- 英国癌症研究机构（CRUK）
- 国际发展部（DFID）
- 健康部
- 经济和社会研究理事会（ESRC）
- 工程和物理科学研究理事会（EPSRC）
- 医学研究理事会（MRC）
- 自然环境研究理事会（NERC）
- 努菲尔德基金会
- 科学与技术设施委员会（STFC）
- 惠康基金会

英国研究理事会关于数据政策的一般原则（RCUK，2011）

- 公众资助的研究数据是一种公共物品，它符合社会大众的利益，应以及时、可靠且不侵犯知识产权的方式尽可能不受限制地开放获取。
- 机构及研究项目具体的数据管理政策和计划应当与相关的标准和团体实践一致。具有长期研究价值的数据应该被储存起来，并使其在未来的研究中同样易于获得且便于使用。
- 为保证研究数据能够被其他研究者所发现并被有效地再次利用，元数据应描述充分且易于公开获取，以使得其他研究者能够了解这项研究并进一步发掘数据的潜力。而出版物中还应当介绍获取其研究中支持数据的方法。
- 英国研究理事会（RCUK）认为，当前在开放研究数据方面仍存在诸多法律、伦理及商业限制等问题。为保证研究进程不会被某些不恰当的数据开放所阻碍，研究机构的政策和实践应当保证这些问题在研究进程的各个阶段都能得到妥善处理。
- 为保证研究团队在搜集和分析数据方面所做出的努力能够被充分地尊重，那些承担了研究理事会资助工作的团队应当被授权在一段时间内对其所搜集的数据享有排他性以方便其发表研究成果。这一时间的长短应当根据研究领域的不同而进行调整，或在某个合适的时机，在个人研究理事会的出版政策中进行进一步探讨。
- 为分辨出那些生成、保存和共享重要数据集的研究者的贡献大小，所有数据使用者都应当在其获取的数据下方注明其研究数据的来源，并遵循与数据政策相关的条款和条件。

- 我们应当使用公共资金来支持那些对公共资金资助的研究数据所进行的管理和共享工作。为在有限的预算中获得最大的研究效益，上述行为的机制应该富有效率并且在公共资金的使用上具有成本效益。

然而，上述的一般原则只是给数据政策提供了一个框架，实际上每个研究理事会或资助者都有其自己的数据政策，其侧重点也略有差异。其中一些机构被授权或鼓励数据共享，但很多研究机构，正如第三章所谈论到的，则需要申请者递交附有经费申请的数据管理和共享计划。关于数据管理和共享的职责问题，大多数理事会选择与获得资助的研究者共同来承担，而只有英国工程和物理科学研究理事会（EPSRC）是和研究主办机构共同承担，这一内容在英国工程和物理科学研究理事会关于研究数据的政策框架中有确切的表述（EPSRC，2011）。英国工程和物理科学研究理事会（EPSRC）要求自2015年5月以来受到其资助的研究机构，在存储研究数据方面出台相应的政策和流程，并对这类数据访问请求进行回复。研究人员同时被要求发布其研究数据的元数据，并且在其最近一次访问数据日期后的10年内保存这些数据和提供安全访问这些数据的渠道。这一政策受英国《信息自由法案》中的伦理观影响很深，这一法案主张开放数据和有时间限制地公布信息。上述行为的结果还是导致越来越多的大学和研究机构出台其数据政策，尽管在实行上还有诸多限制（DCC，2013a）。英国数据监护中心（DCC）之后出台了一个研究理事会数据政策需求的清单（DCC，2012）。

研究理事会还对数据共享的支持服务机构和基础设施进行了资助，具体例子如下：

- 英国数据服务中心为经济和社会研究理事会（ESRC）资助的研究人员提供数据管理、共享指导和支持，以及对社会科学和经济研究数据的管理、保存和传播。
- 指定自然环境研究理事会（NERC）数据中心负责对自然环境研究数据进行管理、保存、传播及增值的一体化工作。
- 英国医学研究理事会的数据支持服务（MRC）是用于推动人口健康科学数据的共享。
- 阿特拉斯千兆存储（STFC）提供科学数据的中长期储存服务。
- 考古数据服务（AHRC）提供保存和传播数字化的考古数据服务。

一些研究资助机构，如英国数据服务（ESRC）、自然环境研究理事会（NERC）、阿特拉斯千兆存储中心（STFC），为其资助的研究者提供了出版物信息库，这一

信息库提供了最新的数据存储和获取功能。

英国生物技术和生物科学研究理事会 (BBSRC)、英国癌症研究机构 (Cancer Research UK)、惠康基金会 (Wellcome Trust) 则是英国 PubMed 中心 (UK PubMed Central) 的合作伙伴。

在美国, 推动数据共享议程的重要数据共享政策制订机构是美国国家科学基金会 (NSF) 和美国国立卫生研究院 (NIH)。

美国已经出台的数据政策的研究资助机构如下:

- 美国国家科学基金会 (NSF)
- 美国国立卫生研究院 (NIH)
- 美国能源部 (DOE)
- 海军研究办公室 (ONR)
- 美国教育部 (ED)
- 美国国家环境保护局 (EPA)
- 美国国际开发署 (USAID)
- 美国国家海洋和大气管理局 (NOAA)
- 美国心脏协会 (AHA)
- 阿尔弗雷德·斯隆 (Alfred P. Sloan) 基金会

2012 年, 欧盟委员会 (European Commision, 2012a; 2012b) 发布了可以进行科学信息获取和保存的通信手段和推荐规范。这就要求在所有成员国间进行协作以推动数据的公开、长期保存以及促进开放科学的能力建设, 这不仅包括欧盟委员会资助的研究项目, 还包括了各成员国资助的研究项目。而欧盟委员会资助了所有成员国全部研究项目的 10%。在进一步的公开咨询会后, 开始进行一个关于数据公开获取的试点项目, 该项目已经为下一个框架项目“地平线 2020”做好了规划。

期刊和出版社

随着在研究资助者中数据共享政策的指数级增长, 出版社也越来越多地实施数据政策, 政策要求在已经发表的同行评审文章中支撑其研究成果的研究数据可以供读者和审稿人使用。这种趋势的形成, 一部分是由主要的研究资助者推动的, 他们促使出版社采用更加严格的数据政策来助推开数据运动; 另一部分是由出版团体本身所驱使的。自 2010 年以来, 开放获取的出版机构——生物医学数据开放中心 (BioMed Central) 通过其开放数据声明、负责开放数据出版的跨出版社工

作组,以及正在进行的出版物相关数据的版权和许可的努力,强烈推动了期刊数据政策的实施(Hrynaszkiewicz and Cockerill, 2012)。此外,许多进化领域的主流期刊在2011年共同采用的联合数据存档政策,自从被其他学科的许多期刊相继采用后,便对期刊数据政策的增长产生了极大的影响。作为数据的发布条件之一,该政策要求将支持论文中所描述结果的数据储存在合适的公共档案库中,通常推荐将Dryad数据库作为合适的数据仓储平台(Dryad, 2013)。

期刊数据政策各不相同,有的是根据请求提供研究数据,有的是作为对出版材料的补充提交数据,也有的是要求数据存储 in 合适的公共存储库中,还有作为出版材料的条件之一将数据提交到指定的存储库中。后者最有可能是出版机构的政策,这种双向关系的一个例子是在荷兰爱思唯尔(Elsevier)出版集团和盘古大陆(PANGAEA)间的地球科学和环境数据的出版网络。还有《自然》针对特定学科指派特定数据库,以及Dryad数据库也是很多期刊的数据存储地点,这会在第十一章中进一步讨论。

通过对141个有名的经济学期刊数据政策的质量和广度的分析表明,其中20%的期刊具有数据可用性政策,其中大部分是强制要求数据在论文出版之前进行提交的(Vaeminck, 2013)。《美国经济评论》(*American Economic Review*)的政策在该领域内公认是最好的,经常被其他经济学期刊所采用,该政策规定只发表这些论文,数据文件和用户文档必须在出版前提交,即在论文分析中所使用的数据可以随时供任何研究者重复使用。在其他社会科学学科,期刊还没有相关强制性的数据政策,但确实有些期刊如《社会学年鉴》(*Annual Review of Sociology*)接受数据文件作为补充材料。在政治学领域,自2012年以来,美国政治学协会(APSA)在其伦理指导手册修订版中对数据访问和研究的透明度进行了明确陈述(APSA, 2012)。美国政治学协会主办的在众多期刊中享有高度知名度的学术研究杂志——《美国政治科学评论》(*American Political Science Review*),旨在将这些数据共享原则纳入到期刊文章的提交要求中。

正如Savage、Vickers(2009)和Wicherts(2006)等在其研究中阐述的那样,如果研究人员仍然不愿意与其他研究人员共享数据,单靠政策本身并不能增加数据共享。Wicherts的研究申请了研究数据,它们来源于10个研究人员发表于开放获取期刊——《PLoS 医学》(*PLoS Medicine*)和《PLoS 临床试验》(*PLoS Clinical Trials*)中的研究成果,只有一个研究者提供了原始研究数据的公开获取。Wicherts等研究表明,在向美国心理学协会的4本期刊上发表过研究成果的研究人员申请研究数据用于二次分析时,成功率为73%,即141个研究团队中有103个同意。这些请求包含再三地尝试和全面地保证,即所请求获得的数据不会被公开发布或重复使用。