



数据密集型环境下的 科学数据管理服务

王丹丹 著



科学出版社

数据密集型环境下的 科学数据管理服务

王丹丹 著

本书受 2015 年度河南省高校科技创新



资助

科学出版社

北京

内 容 简 介

本书阐释了数据、数据服务相关的重要问题——数据出版与数据引用；总结图书馆数据管理服务的构建问题，包括基本构建要素、用户需求识别方法、服务的实施过程与要点，并辅以典型案例进行说明；以现阶段数据管理服务中最普遍的、具有滚雪球效应的一个点——科学数据管理计划支持服务为例，进行深入分析；介绍科学数据服务及其平台可持续发展的问题。本书将科学数据管理服务的理论研究成果和图书馆科学数据管理服务的实践经验相结合，系统梳理科学数据管理服务的相关问题，提供有价值的参考信息源并提供具有可操作性的具体方案。

本书可以帮助图书馆员快速了解这一领域，同时也可供图书情报专业的学生，以及科学数据管理方向的科研人员阅读。

图书在版编目(CIP)数据

数据密集型环境下的科学数据管理服务/王丹丹著. —北京: 科学出版社, 2018.2

ISBN 978-7-03-054154-3

I. ①数… II. ①王… III. ①数据管理-研究 IV. ①TP274

中国版本图书馆CIP数据核字(2017)第196249号

责任编辑: 马 跃 李 嘉 / 责任校对: 刘文娟

责任印制: 吴兆东 / 封面设计: 无极书装

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>

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2018年2月第一版 开本: 720×1000 B5

2018年2月第一次印刷 印张: 15

字数: 303 000

定价: 106.00元

(如有印装质量问题, 我社负责调换)

作者简介

王丹丹，女，中国科学院文献情报中心图书馆学博士、中国科学技术信息研究所博士后、新加坡南洋理工大学访问学者、河南省优秀青年社科专家、河南省高校科技创新人才、河南科技大学青年学术带头人。主要研究方向是网络化信息服务、科学数据管理。在CSSCI来源期刊发表论文20余篇，主持国家社会科学基金青年项目1项、省部级项目3项，出版学术专著1部，获省部级奖3项。

前 言

近年来，随着数据采集设备逐渐普及、数据采集成本逐年降低，科学原始数据呈爆发式增长，科研活动围绕数据进行，科学研究在经历第一范式实验科学、第二范式理论科学、第三范式模拟科学之后进入了以数据为核心的数据密集型科学研究范式时代^[1]。在这一科研范式下，数据不仅是科学研究的成果，同时也是科学研究的对象和工具，科研人员基于数据进行研究。科学数据根据来源可分为实验室采集数据、个人观察数据、互联网数据等。虽然数据一直以来都被认为是科学研究活动的重要构成要素，但大多被用于生成期刊论文，当前环境下的学术交流仍然以科学期刊等文献为核心。随着进入数据密集型科学研究范式时代，科研活动的设计和实施将围绕数据进行，数据的采集、过滤、计算、存储和共享成为科学研究的主题，数据成为科研机构、科研人员学术交流的基本单元^[2]。

在这种大环境下，社会知识服务机构的图书馆，特别是为用户提供科研、教学支撑的学术研究型图书馆向用户提供科学数据服务，能有效保护数据使其免于丢失、提高数据曝光度、传播和出版成果、实现数据共享、公开科学质疑、鼓励观点多样性、节约科研成本、达到研究资助机构的要求等^[3]。大学与研究图书馆联盟（Association of College and Research Libraries, ACRL）不断强调科学数据管理（research data management, RDM）已成为一种重要趋势^[4]，新媒体联盟《地平线报告》也指出这是一种长期趋势，是学术研究型图书馆必须适应的一种趋势^[5]。

鉴于此，本书对过去几年有关科学数据管理服务的理论研究和实践进展进行系统梳理和客观总结，从四个方面介绍数据密集型科学环境下科学数据管理服务的相关问题：首先，回顾现状，阐释与科学数据服务相关的数据出版与数据引用等关键问题，对应本书第一章、第二章、第三章、第四章的内容；其次，总结学术研究型图书馆数据服务的构建问题，具体包括第五章的基本构建要素，第六章的用户需求识别方法，第七章的实施过程与要点，第八章、第九章和第十章的典型案例分析；再次，在上述两部分研究的基础上，以现阶段科学数据服务中最普遍的、具有滚雪球效应的科学数据管理服务的一种类型——科学数据管理计划

(data management plan, DMP) 为例, 进行深入分析, 包括第十一章的科学数据管理计划的应用价值、第十二章的科学数据管理计划评价量表、第十三章的科学数据管理计划支持服务案例; 最后, 介绍了科学数据管理服务平台及科学数据管理服务的可持续发展, 对应本书第十四章和第十五章。

本书各章节的主要内容如下。

第一章介绍科学数据管理服务产生的背景, 回顾和总结研究现状。

第二章从总结数据出版的三种模式入手, 分析数据论文出版机制产生的背景, 揭示数据论文出版机制的本质特征; 从现有实践出发, 阐述数据论文的基本构成要素, 结合数据出版的基本要素, 分析数据论文出版模式的关键问题及其在推进过程中面临的难题。

第三章从数据出版模式入手, 选取每一种出版模式下的典型案例, 归纳总结其数据质量控制的实践和标准, 进行不同模式间的对比分析, 总结异同并发现问题; 总结数据质量控制目前面临的挑战和相关责任者现状, 揭示数据质量控制的发展趋势与特征。

第四章借鉴国内外相关研究及实践经验, 从数据规范引用价值认同入手, 从如何引用、引用什么和何时引用三方面阐述数据规范引用的基本问题与难点, 并以此为基础, 分析数据规范引用给相关利益群体带来的机遇和提出的要求。

第五章介绍构建数据服务的基本要素, 从政策、策略和业务规划, 科学数据管理计划, 管理项目过程中的科学数据, 科学数据选择和转移, 共享和保存科学数据, 指南、培训和支持六个方面对要素的构建要点进行阐述, 并从起步和发展两个阶段介绍科学数据管理服务构建过程中可以参考的资源。

第六章从用户研究方法的角度对近几年国外有关数据管理实践和服务需求研究的文献进行分析, 总结基于数据管理计划内容分析挖掘识别用户需求、使用数据监管档案工具进行结构化访谈捕获需求信息以及基于大规模问卷调查收集用户需求信息三种主要方法的应用场景以及优势和局限性, 并以新加坡南洋理工大学 (Nanyang Technological University, NTU) 为例, 总结图书馆开展科学数据管理服务用户需求研究的经验与体会。

第七章选取七所科研密集型大学, 梳理其图书馆实施数据管理服务的历程, 总结科学数据管理服务建设的六个要点, 即营造科学数据管理服务动力、构建科学数据管理服务合作网络、评价科学数据管理服务需求、建设科学数据管理服务能力、规划科学数据管理服务战略、打造科学数据管理服务特色。

第八章、第九章和第十章分别选取了一个典型案例, 以案例分析的方式介绍科学数据管理服务推进的协同合作策略、全面规划策略和国家策略。

第十一章总结了数据管理计划内容挖掘与应用的三个方向, 即作为一种信息资源, 帮助深入了解所在机构科研人员的数据管理实践与行为特征; 作为一种反

馈渠道，为图书馆当下服务的改进、潜在服务的开发提供启发；作为一个根植于实践的培训平台，使图书馆员的数据服务知识和技能得以有效提升。

第十二章收集面向美国和英国主要科研资助机构的科学数据管理计划评价量表，从评价量表的设计依据和应用目的、各项评价要素的选择以及评价等级与等级描述三个主要设计要素入手，分析不同评价量表的共性和差异，为相关机构设计评价量表提供参考。

第十三章选取不同学科领域开展科学数据管理计划支持服务的案例进行分析，引导图书馆思考如何更好地利用科学数据管理计划支持服务的滚雪球效应来全面促进和提升本机构的科学数据管理服务。

第十四章对科学数据管理服务现有技术平台进行调研，结合新加坡南洋理工大学科研人员的情景化访谈和对 Dataverse 平台的使用测试分析，总结数据管理、出版平台的基本功能要求和用户体验要求，为相关机构设计平台或选用现有平台提供参考。

第十五章从图书馆开展科学数据管理服务的类型、整体规划、资金来源，以及赢得科研人员 and 高层管理者关注和支持的策略等方面总结图书馆服务的可持续发展问题。

通过这样的内容组织，本书希望帮助科学数据管理服务相关利益群体理解实现数据规范引用需要解决的基本问题和重点、难点问题，以及相关利益群体在这一过程中需要承担的责任和未来努力的方向；理解不同数据出版过程中数据质量控制的特点、数据质量控制当前的挑战和相关责任者现状，以及数据质量控制的发展趋势与特征；了解数据管理计划评价量表的制定意义及设计要点。为构建科学数据管理服务，图书馆应有效开展数据管理服务用户需求研究，更好地利用数据管理计划来提升其服务，遵循最佳实践构建适合自身需求的数据管理服务模式，开发数据管理计划评价量表，合理选择或开发数据服务平台，从而有效、持续地推进科学数据管理服务的发展。

王丹丹

2017年8月30日

目 录

第一章 数据密集型环境与科学数据管理服务	1
第一节 科学数据管理服务产生的背景	1
第二节 科学数据管理服务相关研究回顾	4
第三节 本章小结	7
第二章 数据集的独立出版与共享模式	8
第一节 数据出版概述	8
第二节 数据论文出版模式	10
第三节 数据论文基本内容及出版要素解析	12
第四节 本章小结	14
第三章 数据出版过程中的数据质量控制	16
第一节 数据质量控制概述	16
第二节 不同出版模式下的数据质量控制	17
第三节 比较与启示	21
第四节 本章小结	23
第四章 科学数据规范引用关键问题	25
第一节 数据规范引用的意义	25
第二节 数据规范引用的基本问题与难点	26
第三节 数据规范引用对相关利益群体的要求	31
第四节 本章小结	33
第五章 科学数据管理服务构建要素与过程	35
第一节 科学数据管理服务构建要素概述	35
第二节 科学数据管理服务构建要素解析	36
第三节 科学数据管理服务构建过程	50
第四节 本章小结	65
第六章 科学数据管理服务用户需求识别方法	66
第一节 图书馆科学数据管理服务用户需求识别	66

第二节	三种用户需求识别方法的优势及局限性分析	71
第三节	图书馆开展 RDM 服务用户需求识别研究的体会与思考——以 NTU 为例	73
第四节	本章小结	74
第七章	大学图书馆科学数据管理服务实施状况与建设要点分析	76
第一节	大学图书馆科学数据管理服务实施状况	76
第二节	大学图书馆科学数据管理服务建设要点分析	82
第三节	本章小结	85
第八章	科学数据管理服务推进策略——协同合作	86
第一节	案例概况	86
第二节	机构数据管理蓝图项目	87
第三节	启动蓝图项目的第一阶段	95
第四节	本章小结	101
第九章	科学数据管理服务推进策略——全面规划	103
第一节	案例概况	103
第二节	实施背景	104
第三节	科学数据管理组织层面的支持	107
第四节	科学数据管理的软环境建设	108
第五节	科学数据管理的硬环境建设	111
第六节	数据管理技术性基础设施	118
第七节	本章小结	121
第十章	科学数据管理服务推进策略——国家战略	123
第一节	案例概况	123
第二节	数据服务的结构和功能	123
第三节	本章小结	130
第十一章	科学数据管理计划的应用价值	132
第一节	研究过程与数据活动	132
第二节	支持科学数据管理计划价值挖掘的工具	133
第三节	数据管理计划的价值挖掘及其应用方向	135
第四节	本章小结	138
第十二章	科学数据管理计划评价量表	140
第一节	科学数据管理计划评价量表的设计背景	140
第二节	科学数据管理计划评价量表的设计要素分析	141
第三节	本章小结	147

第十三章 科学数据管理计划支持服务案例	149
第一节 伦敦经济与政治学院	149
第二节 伦敦卫生和热带医学学院	150
第三节 加利福尼亚洛杉矶分校	151
第四节 俄勒冈大学	153
第五节 格拉斯哥大学	154
第六节 哥伦比亚大学	155
第七节 圭尔夫大学	156
第八节 巴斯大学	157
第九节 本章小结	158
第十四章 科学数据管理服务平台	160
第一节 数据平台的基本功能与利益相关群体	161
第二节 数据平台的用户访谈与使用测试	164
第三节 本章小结	168
第十五章 科学数据管理服务的可持续发展	170
第一节 科学数据管理服务类型	170
第二节 科学数据管理服务整体规划	173
第三节 科学数据管理服务的资金来源	177
第四节 赢得科研人员和高层管理者关注和支持的策略	181
第五节 本章小结	185
参考文献	188
附录 主要资助机构的科学数据管理计划指南与评价标准	207
后记	223

插图目录

图 2-1	传统数据共享模式与基于数据论文的数据共享模式	11
图 6-1	NTU RDM 服务用户需求识别方案	73
图 8-1	针对研究数据的三层元数据模型及其例子	93
图 8-2	合作提供咨询服务的方法	100
图 9-1	莫纳什大学的管理研究数据介绍页面	110
图 9-2	科学数据管理基础设施与研究过程的关联	112
图 9-3	数据管理基础设施概览	114
图 11-1	贯穿整个研究过程的数据相关的活动及其关系	133
图 14-1	上传数据集页面中导致用户产生疑问的关键词选项	166
图 14-2	悬停文本	167
图 15-1	爱丁堡大学基于数据生命周期的数据管理培训内容设计	174
图 15-2	爱丁堡大学数据管理服务模块	175
图 15-3	风险得分的计算方法	182

列表目录

表 2-1	科学数据与期刊论文的对比分析	11
表 3-1	数据期刊数据评议指南比较	20
表 6-1	DCPT 的访谈模块及对应的 DCP 文档部分	68
表 6-2	相关研究的问卷调查要点汇总	70
表 6-3	DCPT 的使用对象与使用目的	72
表 11-1	DDC 主题元素及发布日期	134
表 12-1	评价量表的设计依据与应用目的	142
表 12-2	约翰·霍普金斯大学的评价量表	143
表 12-3	DCC 的 14 个科学数据管理计划主题	143
表 12-4	不同数据管理计划评价量表涵盖的主题	144
表 12-5	DRAT rubric 的一部分	146
表 14-1	数据出版平台的功能需求列表	161
表 14-2	数据管理平台的功能需求列表	162
表 14-3	数据出版平台的用户体验需求列表	168
表 15-1	不同资金获取渠道的优势、劣势对比分析	178
表 15-2	项目的过去、现在或未来	181
表 15-3	所有项目的风险评分	181
表 15-4	不同风险等级的行动计划	182
表 15-5	针对某项目的数据管理的个人行动计划	182
附表 1	ESRC 的科学数据管理计划质量评价标准	207
附表 2	地平线 2020 的科学数据管理计划指南	208
附表 3	地平线 2020 的科学数据管理计划评价量表	208
附表 4	艺术与人文研究理事会的科学数据管理计划评价量表	209
附表 5	英国工程与自然科学研究委员会的科学数据管理计划评价量表	211
附表 6	英国生物技术与生物科学研究委员会的科学数据管理计划评价量表	212

附表 7 英国医学研究理事会的科学数据管理计划评价量表	214
附表 8 英国自然环境研究理事会的科学数据管理计划评价量表	217
附表 9 英国惠康基金会的科学数据管理计划评价量表	219
附表 10 英国癌症研究中心的科学数据管理计划量表	221

录 目 录 表

第一章 数据密集型环境与科学 数据管理服务

第一节 科学数据管理服务产生的背景

一、数据密集型科研范式

随着科学研究工作向数据密集型科学研究范式转变，科学研究已不再独立观察某一实验或领域的的数据，学科间交叉合作研究和学科内继续研究成为科研发展的新趋势。科研人员开始认识到数据成为连接人和思想的新载体，它取代了媒体曾经承担的某些任务^[6]。科学数据是可以以数字形式存储的任何信息，包括文本、数字、图像、视频、电影、音频、软件代码、算法、方程式或化学反应式、模型、动画等^[7]。从来源看，科学数据既包括科研、实验过程中的实验数据，计算机中的存储数据等数字化数据，如实验记录数据等，又包括原始的非数字形式数据，如神经图像等；从数据组成类型与格式看，科学数据既包括传统的结构化数据，如数值型数据、多媒体数据，也包括大数据时代的非结构化数据、半结构化数据，如文本数据、HTML 数据、社交数据等^[8, 9]。

网络对大量数据的承载能力，以及今天高速发展的信息和交流技术基础设施，促使科研人员对自然科学、社会科学以及艺术人文领域数据研究产生热情^[10]。事实上，科研人员并不是对原始数据以及数据存储感兴趣，而是对数据的使用和再使用及其嵌入的情景感兴趣。科研人员需要对大量的数据进行收集、分析、管理、保存和共享。这给科研人员带来了数据管理上的困难，一方面科研人员缺少时间和经费来管理数据，另一方面担心别人滥用其共享的数据；与此同时，资助机构要求项目方案包含 DMP 的新规定，又使科研人员产生更迫切的数据管理需求。如何对数据进行合理描述？如何有效组织管理数据？

如何设置共享数据的限定条件等? 这使数据管理服务成为图书馆未来主要的服务方向之一。

二、科研人员的态度和行为

数据密集型科学的成功与否取决于参与其中的科研人员的行为。数据共享是数据密集型科学的核心概念。数据共享就是出版数据以供其他人使用。制度因素、技术因素和个人因素三个方面的因素促使科研人员共享数据。共享数据是获取他人数据的条件,是接受资助的先决条件,不同的资助机构开始对其提出不同程度的要求^[11]。更确切地说,对数据开放获取的要求,在不同的国家、机构和学科也存在差异。对科研人员数据共享实践及其观点的调查结果表明,来自不同学科的大多数科研人员对数据共享持积极的态度,虽然真实情况是只有一少部分科研人员有共享数据的经历^[12]。

Cox 等指出,在英国,数据管理的需求在最近几年涌现,因为英国科研资助机构开始关注改善数据管理的质量^[13]。Kruse 和 Thestrup 介绍了丹麦大学存储、保存和提供数据获取的情况^[14]。Vlaeminck 和 Wagner 介绍了社会科学研究可重复性低的几个问题^[15]。数据共享毫无疑问是个复杂的问题,因为科研人员可能出于多种原因,如对数据进行记录和描述需要花费很多时间和精力而不进行数据共享。然而,主要的原因是缺乏兴趣,因为在大多数领域,学术奖励不是针对数据管理,而是针对论文发表^[16]。更大程度的开放性很明显需要科研人员从永久性资产控制,转向避免误用和曲解数据。另外,每一个学科有本身的数据文化,一些学科的数据可能格式一致,因此相对于其他一些格式差异较大的学科而言,更容易实现数据共享。安全和可控问题也是一个重要因素。整体而言,资助机构和出版商对数据共享的要求可能存在冲突,同时技术和文化的障碍也阻碍这些数据的共享。然而,学术社区正在为不断改进数据获取而努力,这一进展影响着科研过程的每一个方面^[17]。

科研人员开始学习数据的管理和监管知识。然而,在数据管理实践方面,大部分科研人员尚未接受过正式的培训,尽管他们对自己数据管理方面的专业知识并不满意。只有少数的科研人员考虑对所持有的数据进行长期保存。尤其是处于职业生涯早期阶段的那些科研人员,由于出版相关的需求不断增多,所以被迫考虑长期数据监管的问题。与此同时,只有当元数据和记录文档对科研人员完成论文有帮助时,他们才会产生数据管理的意愿。然而,几乎没有科研人员意识到图书馆可以提供数据管理服务,因为在他们的观念里图书馆只是保存图书和论文的地方^[18]。

一些研究尝试使科研人员的数据管理,成为更有目的性的而不是自然发生的过程。例如,Goodman 提出数据管理的几个原则^[19],指导科研人员确保他们的数据和分析有价值,他给出的建议包括:在进行研究时要考虑数据一定程度的再使用;尽可能地建立数据与出版物的链接;支持让数据产生者从数据中获得荣誉,并描述这是如何发生的;对共享数据的科研人员进行奖励;为了促使数据的解释和再使用,出版对处理过程进行描述的文档;宣传和使用数据知识库;使用永久标识符共享数据。Buckland 指出这些被认定为激励的行为可以通过数据共享所需的不同步骤进行补充,具体包括:如果有合适的数据集的存在,确保其具有可发现性;指出它的位置;分析复本是否可用;澄清是否允许使用;确保互操作性,如是否是标准化的、是否可以通过一定的努力被再次使用;判断其描述是否足够清楚、是否表明具体的数据集中所展示的信息;确保可信性;确定对于某些目的、给定的数据集而言,是否是可用的^[20]。

三、科学数据管理

科研人员的态度和行为是他们管理科学数据的基础,也是参与该过程和使用科学数据管理服务的基础。数据密集型研究范式所需要的科学数据服务成为图书馆服务的一部分^[21]。科学数据服务包括数据管理(data management)和数据监管(data curation)。这两者虽然不同,但是又不能被完全割裂。可以把数据管理和数据服务比作竞技场,不同的利益相关者汇聚在这里,对新涌现的工作范围进行博弈。Verban 和 Cox 的研究表明,图书馆是这一领域中唯一声称具有管辖权的组织,把数据服务看作其现有的开放获取服务和信息素养教育的延伸^[22]。数据监管的目的是使所选择的数据可获取、可使用以及在整个数据生命周期内具备有用性。它属于数字保存的范畴,通过补充性的描述文档、描述性元数据等提供情景信息^[23]。

数据监管提出了与数据所有权相关的问题,如它的保留、维护、获取、开放性以及成本等。在这一情境下,数据监管者必须能够回答以下问题,即谁拥有数据,其他人(如资助机构和出版商)提出了哪些要求,哪些数据应该被保留,数据应该被保留多长时间,如何对数据进行保存,与其相关的道德问题有哪些,需要进行哪些风险管理,如何获取数据,数据的开放程度如何,可负担的成本是多少,本地数据管理还存在哪些选择^[24]。数据监管者参与一系列活动,具体包括:详细说明数字监管政策、程序和实践,规划、实施并监督数字监管对象和服务;选择进行长期保存的数字文档;与现有或潜在的利益相关者交流数字监管的价值;诊断并解决问题,确保数字对象的长期可获取;监测文件格式、硬件、软件的过时性,并开发相

应的新的文件格式、硬件、软件；识别确保不同的应用系统和保存技术可以互操作的方法和技术；证实并记录需要保存的数据的出处；建立并维护与不同利益相关者之间的合作关系；组织并管理元数据标准、访问控制和使用流程；组织员工向其提供教育、培训和其他支持，使其适应数据监管的新发展。与这些职责类似，他们应该熟悉不同数字对象的数据结构，评价数据对象权威程度、完整性和一段时期内准确性的方式，存储和保存政策、程序和实践，相关的质量保证标准，信息丢失的风险，对信息基础设施的要求，以确保合适的获取、存储和数据恢复^[25]。

第二节 科学数据管理服务相关研究回顾

一、科学数据管理服务类型

李慧芳将图书馆的数据服务总结为检索服务、发现服务、申请服务、获取服务、管理服务、关联服务、传递服务和存储服务 8 种类型^[26]。张凯勇、肖潇等将其划分为开发服务、存储服务、检索服务、咨询服务、分析服务^[27, 28]。Tenopir 等将其划分为信息或咨询类服务和技术型或实际动手操作型服务。信息或咨询类服务包括：给教师、员工和学生提供 DMP 的咨询；给教师、员工和学生提供有关数据和元数据标准的咨询；扩大服务范围，与校园内外的其他数据服务提供者合作；为发现和引用数据及数据集提供参考和支持；创建有关数据、数据集和数据知识库的网页指南和发现帮助；与其他的图书馆员、学院的其他工作人员以及数据服务专家讨论数据服务。技术型或实际动手操作型服务包括：为数据服务系统提供技术支持（如知识库、获取和发现系统）；删除或交换知识库中的数据；准备存储到数据知识库中的数据；为数据创建元数据；识别可能会存储到校内外知识库中的数据；作为项目团队的成员直接和科研人员合作^[21]。

二、科学数据管理服务实践

Peters 和 Dryden 发现科研人员最需要的服务主要是一些定向型的服务，如为研究资助提供支持服务^[29]。Bach 等发现大多数生物多样性知识库只对用户提供低层次的服务^[30]。就图书馆和图书馆员在提供服务方面的角色调查发现，有三分之一的被调查者认为在五年内管理来自 e-science 项目的数据集是图书馆员的主要责任^[31]。MacColl 建议图书馆员应该参与到整个研究过程中去^[32]。Peters 和 Dryden 针对来自美国和加拿大的研究图书馆协会（Association of Research