

高校科研数据管理 理论与实践

刘桂锋

著

Theory and Practice of Research Data
Management in University

苏大学出版社
SUZHOU UNIVERSITY PRESS

高校科研数据管理 理论与实践

刘桂锋——著



江苏大学出版社
JIANGSU UNIVERSITY PRESS

镇江

图书在版编目(CIP)数据

高校科研数据管理理论与实践 / 刘桂锋著. — 镇江:
江苏大学出版社, 2017.8

ISBN 978-7-5684-0567-6

I. ①高… II. ①刘… III. ①高等学校—科研管理—
数据管理—研究 IV. ①G644

中国版本图书馆 CIP 数据核字(2017)第 193133 号

高校科研数据管理理论与实践

Gaoxiao Keyan Shuju Guanli Lilun Yu Shijian

著 者/刘桂锋

责任编辑/郑晨晖

出版发行/江苏大学出版社

地 址/江苏省镇江市梦溪园巷 30 号(邮编: 212003)

电 话/0511-84446464(传真)

网 址/http://press. ujs. edu. cn

排 版/镇江华翔票证印务有限公司

印 刷/虎彩印艺股份有限公司

开 本/890 mm×1 240 mm 1/32

印 张/7.5

字 数/215 千字

版 次/2017 年 8 月第 1 版 2017 年 8 月第 1 次印刷

书 号/ISBN 978-7-5684-0567-6

定 价/38.00 元

如有印装质量问题请与本社营销部联系(电话:0511-84440882)

序

风起云涌的大数据时代,不仅给人类带来了一场席卷全球的技术革命风暴,而且为学术界带来了一次科研范式的重大变革。传统研究型数据服务框架与模式难以满足数字化环境下的科研工作需求。在此背景下,全球图书馆学科服务与机构知识库正迎来一场全新的转型升级,高校科研数据管理理论研究已引起图书情报学界的高度重视。目前,国外相关领域研究已取得初步理论与实践成果,国内研究正强劲起步、蓄势待发。

由江苏大学刘桂锋博士撰写的《高校科研数据管理理论与实践》正是在把握时代脉搏的背景下,涉足高校科研数据管理领域多年来研究心血与宝贵成果的辛勤结晶,非常荣幸,受作者之邀为本书作序。该书详细阐述了科研数据管理研究领域的发展脉络与体系架构,跟踪并研究了该领域的最新理论成果与议题,结合诸多国际名校的优秀科研数据管理实例,向科研数据管理研究者们提供了富有价值的研究视角与广泛的学术视野。作者对于科研数据管理领域研究之匠心独运,学术功底之扎实,可窥一斑,全书可读性与内容亮点主要体现在以下方面:

一是体系架构方面,思路清晰,逻辑性强。作者开篇介绍了高校科研数据管理研究的发展背景、研究进展,以及基本研究思路、内容和方法。之后又整合基本知识概念、科研数据生命周期理论、高校科研数据管理政策、服务内容模型及服务体系构建与策略五大模块,形成第2~6章内容。基本概念介绍由浅入深,透彻解析科研数据、数据管理、数据科学3大知识点。从各国科研数据生命周期理论与模型的汇总、高校科研数据管理政策体系构建、服务内容模型研究及服务体系构建与策略研究4个角度剖析国际科研数

据管理领域的研究体系与知识架构。

二是成果跟踪方面,实事求是,持续关注。作者结合自身在美国高校的访学经历,或亲自造访交流,或网上开展数据搜集工作,长期跟踪调研包括斯坦福大学、伊利诺伊大学、哥伦比亚大学、华盛顿大学、加州大学在内的欧美高校与科研机构在科研数据管理领域的研究成果,为我国该领域学者提供了良好的学术追踪平台。

三是案例分析方面,独具匠心,具体翔实。本书广泛搜集、精心筛选了全球顶级名校的科研数据管理实务操作案例,涉及第7~9章的国内外高校科研数据管理服务、科研数据管理教育、科研数据管理实践等层面,其实践案例项目框架介绍之全面、研究工作操作流程叙述之详细、调研数据获取之完备,皆可见作者研究态度之端正,研究精神之负责,研究功底之扎实,研究水平之优良。

诚然,作为作者的首部科研数据管理研究专著,本书在研究与撰写过程中尚存在一些不足,比如第7章国外高校科研数据管理服务实践研究,仅选取美国伊利诺伊大学厄巴纳-香槟分校科研数据管理服务一个案例,借鉴意义略显不足。

本书作为我国高校科研数据管理领域为数不多的理论类著作,在保证创新性的同时,兼具极强的学术参考价值与研究指导意义,作者取材于国内外高校的成功实践经验,以理论研究为基石终又回归实践,基础理论丰富,体系设计完善,方法运用得当,应用前景广阔,非常适合高校科研数据项目设计与执行人员、研究生和学者们学习参考之用。

武汉大学 司莉

2017年5月22日

前 言

20 世纪 90 年代末,在国际学术界大规模兴起的开放存取运动旨在推动科研成果自由传播,提升其公共利用程度,保障其长期保存。互联网、物联网、大数据、云计算等新技术的发展为开放存取运动提供强劲的技术支撑,开放与共享成为开放存取运动核心理念。开放存取运动的不断发展导致开放科学运动的兴起,核心理念拓展为“自由、开放、合作、共享”。开放数据在开放科学广泛主体中占据重要位置,科研数据共享又是开放数据的重要表现形式。

国内外学者对科研数据或科学数据的概念及两者之间的异同尚未达成完全一致意见。从语言角度来看,科研数据译自英文“Research Data”,科学数据译自英文“Scientific Data”。而早在 2002 年,我国提出实施科学数据共享工程,自此,科学数据的用法更为官方。科研数据的产生主要源自科研过程,科学数据则来自各类科学研究与实践过程,因此,科学数据的外延比科研数据相对宽泛。本书虽然采用的是学界的用法,但兼具两者概念的内涵,同时保留原始资料中科学数据的用法。

随着大数据时代科学研究范式向数据密集型的转变、各国研究基金对科研数据管理重视程度的不断提高,以及科学研究人员科研数据管理的内在迫切需求,科研数据服务逐渐成为图书馆学科服务转型升级不可或缺的重要组成部分。

本书基于大数据理论、数据管理理论、数据生命周期理论等基本原理,综合运用文献调研、网络调查、现场考察、案例研究、内容

分析、比较研究等方法,围绕“研究设计—政策研究—理论研究—实证研究”的研究思路展开,从理论与实践相结合的角度力求揭示高校科研数据管理服务的基本规律。主要研究内容包括:(1)依据数据生命周期理论、网络调查与内容分析法,对英国、美国和澳大利亚三国 30 所高校的科研数据管理政策进行比较,并选取 3 所高校作为典型案例,构建高校科研数据管理政策体系框架模型。(2)采用网络调查方法,对美国 50 所高校图书馆开展的科研数据管理服务实践进行比较和分析,并选取 4 所高校作为典型案例,根据数据生命周期理论构建高校图书馆科研数据管理服务内容模型框架。(3)在上述调研基础上,构建了以服务基础为保障、服务内容为核心、教育培训为关键的“361 型”高校图书馆科研数据管理体系,并从战略规划、用户需求、内容构建、学科团队、多方合作和教育培训 6 个方面提出了高校图书馆开展数据管理服务的对策。(4)选取美国伊利诺伊大学厄巴纳-香槟分校为个案,采用网络调查和实地考察相结合的方法,构建了“以用户为中心”的科研数据管理体系,并从长期积累、多方合作、体系构建 3 个方面总结启示。(5)以美国伊利诺伊大学厄巴纳-香槟分校 iSchool 与黄石公园合作开展的 Site - Based Data Curation 项目为例,在对项目概况、实施背景和项目成果进行详细解读的基础上,从现实背景、实施方式、内容构建和运行机制 4 个角度总结特点与启示。(6)采用文献调研法和网络调查法,从背景、政策、技术、资源 4 个方面对我国 3 所典型高校科学数据管理实践的典型案例进行深入剖析,总结理论研究与实践建设经验。

高校科研数据开放与共享是一个复杂系统,需要多方协作、共同努力。本书重点探讨了高校科研数据管理的内容模型、服务体系与实践案例,促进了理论与实践的和谐统一,深化了科研数据理论,丰富了数据开放内容,拓展了数据科学内涵与外延,推动了开

放科学运动的前进脚步。

本研究成果受到江苏省博士后科研资助计划资助,得到江苏大学图书馆和江苏大学出版社的大力支持,特别是责任编辑在本书的出版过程中付出了艰辛的劳动和辛勤的汗水,在此表示诚挚的谢意。

由于时间仓促,水平有限,书中难免会存在错误和不足,恳请广大专家、学者、同行和读者不吝赐教。

刘桂锋

2017年4月20日于镇江

目 录

第1章 绪 论 001

1.1 科研数据管理研究背景 001

1.1.1 大数据时代科学研究范式的转变 001

1.1.2 科学研究基金组织对数据管理的基本要求 002

1.1.3 科学研究人员数据管理的内在需求 003

1.1.4 图书馆学科服务与机构知识库的转型升级 003

1.2 科研数据管理研究进展 004

1.2.1 国外科研数据管理研究进展 004

1.2.2 国内科研数据管理研究进展 011

1.3 研究思路、内容与方法 016

1.3.1 研究思路 016

1.3.2 研究内容 017

1.3.3 研究方法 019

1.4 本章小结 020

第2章 科研数据管理基本概念 021

2.1 科研数据相关概念 021

2.1.1 数据 021

2.1.2 科研数据 022

2.1.3 科学数据 025

2.1.4 开放数据 026

2.2 数据管理相关概念 028

2.2.1 科研数据管理 028

2.2.2 数据治理 030

2.3 数据科学相关概念 034

2.3.1 数据科学 034

2.3.2 开放科学 036

2.3.3 公众科学 038

2.4 本章小结 039

第3章 科研数据生命周期理论 041

3.1 专业机构的数据生命周期 041

3.1.1 英国数据管理中心 041

3.1.2 美国校际政治与社会科学数据共享联盟 043

3.1.3 英国数据档案馆 045

3.1.4 美国 DataONE 项目 046

3.2 高校的数据生命周期 048

3.2.1 牛津大学科研数据管理流程 048

3.2.2 弗吉尼亚大学数据生命周期 049

3.2.3 加州大学圣地亚哥分校数据生命周期 049

3.2.4 密歇根州立大学数据生命周期 051

3.3 本章小结 052

第4章 高校科研数据管理政策体系构建研究 053

4.1 高校科研数据管理政策调查研究 053

4.1.1 英国高校科研数据管理政策 053

4.1.2 美国高校科研数据管理政策 060

4.1.3 澳大利亚高校科研数据管理政策 065

4.2 高校科研数据管理政策典型案例研究 072

4.2.1 利兹大学 074

4.2.2 明尼苏达大学 076

4.2.3 澳大利亚天主教大学 077

4.3 高校科研数据管理政策体系构建研究 080

4.4 本章小结 082

第 5 章	高校科研数据管理服务内容模型研究	084
5.1	高校科研数据管理服务文献调研	084
5.2	高校科研数据管理服务网络调查	087
5.2.1	斯坦福大学:数据管理服务	093
5.2.2	哥伦比亚大学:科研数据管理	095
5.2.3	华盛顿大学:数据管理	097
5.2.4	加州大学圣克鲁兹分校:数据管理服务	100
5.3	科研数据管理服务内容模型构建研究	103
5.4	本章小结	107
第 6 章	高校科研数据管理服务体系构建与策略研究	109
6.1	高校科研数据管理服务现状	109
6.2	高校科研数据管理服务体系案例	110
6.2.1	加州大学圣地亚哥分校	110
6.2.2	威斯康星大学麦迪逊分校	113
6.2.3	德州大学奥斯汀分校	115
6.2.4	埃默里大学	118
6.3	高校科研数据管理服务体系构建	122
6.3.1	技术平台	123
6.3.2	机构人员	123
6.3.3	资源工具	123
6.3.4	服务内容	124
6.3.5	教育培训	124
6.4	高校科研数据管理服务策略	125
6.4.1	以战略规划为根本点,全面营造服务氛围	125
6.4.2	以用户需求为切入点,着力创造服务效益	126
6.4.3	以内容构建为着重点,切实保障服务运行	126
6.4.4	以学科团队为突破点,顺利引导服务实施	127
6.4.5	以多方合作为支撑点,确保提高服务质量	127
6.4.6	以教育培训为结合点,努力搭建服务桥梁	128

6.5 本章小结 128

第7章 国外高校科研数据管理服务实践研究 130

7.1 高校科研数据管理服务案例 130

7.2 UIUC 科研数据管理服务阶段 134

7.3 UIUC 科研数据管理服务体系 137

7.3.1 技术平台 138

7.3.2 机构人员 140

7.3.3 资源工具 141

7.3.4 教育培训 146

7.3.5 服务内容 146

7.4 UIUC 科研数据管理服务启示 151

7.4.1 持续不断的积累是成功开展科研数据管理服务实践的雄厚根基 151

7.4.2 寻求多方的合作是顺利开展科研数据管理服务实践的有效途径 151

7.4.3 运营体系的构建是成功开展科研数据管理服务实践的重要保障 152

7.5 本章小结 153

第8章 国外高校科研数据管理教育实践研究 154

8.1 Data Curation 研究现状 154

8.2 UIUC 科研数据管理教育实践基本概况 156

8.3 UIUC 科研数据管理教育课程案例项目 163

8.4 UIUC 科研数据管理教育实践案例项目 166

8.4.1 项目概况 166

8.4.2 实施背景 167

8.4.3 项目成果 167

8.4.4 SBDC 项目特点与启示 176

8.5 本章小结 179

第9章 我国高校科学数据管理实践研究 181

9.1 我国科学数据管理实践 181

9.2 武汉大学科学数据管理实践 184

9.3 复旦大学科学数据管理实践 187

9.4 北京大学科学数据管理实践 189

9.5 我国高校科学数据管理实践比较 192

9.6 本章小结 195

参考文献 196**附录 221****后记 224**

第1章 绪论

数据既是科学研究的成果,又是科学研究的基础。随着数据密集型科学研究范式的快速兴起,科研数据管理成为高校、科研机构、科研人员面临的一项重要时代课题。科研数据管理贯穿于整个数据生命周期全过程,是一项复杂性的系统工程。数据生命周期是重要的理论基础,数据管理政策是推动数据管理活动的核心动力,数据管理内容模型是数据管理活动的认知基础,数据管理体系是数据管理活动的重要保障,数据管理实践是数据管理活动的应用场景。

1.1 科研数据管理研究背景

1.1.1 大数据时代科学研究范式的转变

随着云时代的来临,大数据技术引起了全球学者的广泛关注。早在2008年,*Nature*就推出了Big Data专刊。“大数据”一词真正成为热点是在2011年5月。之后,随着麦肯锡报告的发布、全球互联网巨头的加入、美国政府的重视,大数据迅速引起了学界与业界的高度重视。大数据时代的特征有3个^[1]:一是数据更大,借助先进的计算机技术,可以获得研究对象的总体数据。二是数据更复杂,一些不可计量、存储、分析和共享的很多东西都可以被数据化而形成非结构化数据,如图像、声音、影视、超媒体等。三是数据更好,通过数据挖掘等技术,可以为人们展现从未想过的因果关系,为决策提供依据。大数据时代的这些特点正在促使人们的思维方式发生转变。

大数据时代的来临,引起学术交流环境和科研环境发生变化,科学研究经历实验型科研、理论推演、计算机仿真 3 种科研范式之后,正在向数据密集(Data Intensive)型的“第四范式”转变。随着 e-Science 模式发生巨大的变革,微软公司于 2009 年 10 月发布了《e-Science:科学研究的第四种范式》^[2],首次全面描述了快速兴起的数据密集型科学研究,该论文集分为地球与环境、健康与生活、科学基础设施、学术交流 4 大部分。显然,数据已经成为科学研究的重要组成部分。

科学数据成为研究人员的主要研究对象,对其搜集、分析、管理、保存与共享成为数据管理、服务机构等部门的工作拓展领域。大数据技术不仅给科研人员带来巨大的挑战,更给高校科学研究带来广阔的发展空间和机会。大数据时代科研过程的方式、途径、模式等都将发生改变,提供满足科研人员需求的科研数据管理服务将成为高校图书馆新的重要工作之一。

1.1.2 科学研究基金组织对数据管理的基本要求

科学研究基金组织对数据管理的要求使得数据管理服务首先在美国、英国、澳大利亚等许多国家陆续开展。早在 2010 年,美国国家科学基金会(National Science Foundation,简称 NSF)推出了科研数据共享政策,之后要求被提交到 NSF 的申报书,必须包括一项“科研数据管理规划”的补充文档^[3]。在英国,研究基金组织是科研数据管理的最主要驱动者,他们认为“好的研究需要好的数据支撑”。大多数重要的研究基金组织都提出了相关的数据管理政策,来促使更好地开展科研数据的管理、保存和出版。澳大利亚各界联合发布的《澳大利亚负责任的科研行为规范》(Australian Code for the Responsible Conduct of Research)中对于科研数据和重要资料的管理提出了明确的要求。

自此,科研数据管理便逐渐成为各种基金项目申请的基本要求。除了一些基金组织的要求外,一些知名期刊(如 Nature 等)^[4]在数据管理方面的要求为其受到科研人员的重视起到了推波助澜的作用。我国在这方面进行了一些卓有成效的实践探索,旨在为

科研人员的数据管理与共享搭建交流的平台,如在2002年科技部提出的我国将实施科学数据共享工程^[5],以及最近创办的关于科学数据出版的学术期刊《中国科学数据》^[6]、发布科研成果的在线通用存储库 SciDB (Science Data Bank)^[7]。除了创办新的数据期刊外,一些传统期刊对学术论文提出了数据共享的要求,如《现代图书情报技术》(2017年起改为《数据分析与知识发现》)规定从2016年起,所有投稿论文必须提交支撑科学数据,以便将来进行数据共享^[8]。

1.1.3 科研人员数据管理的内在需求

大数据时代的到来正在悄然改变着传统生物医学科研模式。国外医学科研人员正在利用大量的科研数据代替真实的试验。Service^[9]开启了大数据的研究模式“干试验”(Dry Lab Biology),在海量的数据库中搜集公开的信息,并对这些生物医学科研数据进行深度挖掘。

科研人员虽然意识到了数据管理的重要性,但对数据管理问题认识不足,除了在科研过程中得到有价值的结论之外,在数据处理和分析方面往往束手无策。瑞士洛桑联邦理工学院曾经对科研团体关于科研数据的管理问题做过一个历时18个月的调研^[10],对于科学数据的管理,研究人员普遍存在的问题是:“我们有太多的数据,但我们不知道如何处理这些数据,我们害怕面对大量的数据处理工作,我们认为应该把数据发布到网络上。”根据有关研究,科研人员在科研数据管理方面面临着许多困难,例如,缺乏专业的数据描述组织方法、基础设施环境与软件工具,亟需良好的方法和机制来管理科研数据等^[11]。

1.1.4 图书馆学科服务与机构知识库的转型升级

科学研究范式的转变使图书馆面临新的挑战,即图书馆如何获取新技能?如何调整角色为 e-Science 和 e-Research 提供数据支撑?如何对颇具科研价值的数据进行管理和服务?美国、英国、澳大利亚的一些高校已把科研数据管理作为图书馆学科服务的一项重要内容,其中包括世界名校剑桥大学、麻省理工学院、耶

鲁大学、哥伦比亚大学、康奈尔大学等,这说明有部分高水平的研究型高校图书馆已经开始关注科研数据管理的重要性,并且开始尝试参与科研数据管理,从而提高对学术研究和教学的支撑服务能力。

大数据时代科学研究范式向数据密集型的转变、科学研究基金和学术期刊对数据管理重视程度的不断提高,以及科研人员对数据管理的内在迫切需求等多方面的因素促使数据管理服务成为图书馆学科服务转型升级不可或缺的重要组成部分。图书馆在数据管理服务方面面临着巨大的机遇,但是要为科研人员提供优质的服务也面临着诸多挑战。图书馆需要深刻理解和把握数据管理服务的本质含义,准确掌握服务流程,探索和优化服务模式。同时还应管理和利用科研数据,让其实现保值与增值,努力在信息挖掘、情报分析和知识服务等方面进行艰苦卓绝的探索。

1.2 科研数据管理研究进展

1.2.1 国外科研数据管理研究进展

在 Web of ScienceTM (WOS) Core Collection (核心集合) 中的 Science Citation Index Expanded (SCI-EXPANDED)、Social Sciences Citation Index (SSCI)、Conference Proceedings Citation Index - Science (CPCI-S) 数据库中,输入检索式:标题 = “research data” OR “scientific data” OR “data management” OR “data curation” OR “digital curation” OR “digital management”,不限时间跨度。检索日期为 2016 年 6 月 3 日,共获得检索结果 2849 条,有 193 条数据属于图书馆信息学学科 (Library and Information Science),其中期刊论文 114 篇。

采用文献调查和内容分析法对 114 篇期刊论文进行梳理,发现国外科研数据管理研究主要分为 3 个方面:科研数据管理理论研究、科研数据管理服务研究和科研数据管理案例研究。下面对这 3 个方面逐一进行介绍。