数据监管：图书馆科研数据共享新服务

于 霜 姚占雷
（华东师范大学商学院，上海 200241）

摘 要：首先详细描绘了科研数据共享的现状，然后通过论述数据监管与图书馆的关系以厘清图书馆应用数据监管开展科研数据共享服务活动的必要性，最后系统梳理出图书馆开展数据监管所扮演的角色与采取的应对策略。

关键词：科研数据；数据共享；图书馆；数据监管；学科服务
中图分类号：G352 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2013.06.010

Data Wardship: Library S&R Data Sharing New Activities

Yu Shuang, Yao Zhanlei
(Business School of East China Normal University, Shanghai 200241)

Abstract: Nowadays, the research data sharing has gained widely attention, reuse and depth mining problem about research data should be solved. Firstly, this paper describes the status of research data sharing, and then clarifies the necessity of research data sharing service activities that hosting by Library using data curation, through discussing the relationship between data curation and library. Finally, we systematically summaries the roles of which library plays, difficulties and countermeasures in research data sharing service activities.

Keywords: research data, data sharing, library, data curation, subject service

1 引言

早在2000年，英国提出E-Science概念即利用新一代的网络技术和广域分布式高性能计算环境建立的一种全新科学研究模式，从而应对各学科研究领域所面临的空前复杂化问题，而随着E-Science的深入发展，全球性、跨学科的大规模科研合作资源共享与协同工作正在成为现实。2007年，计算机图灵奖得主Jim Gray在NRC-CSTB 的演讲报告中提出的“科学研究的第四范式——数据密集型科学发现”[1]，在学界引起了巨大反响，数据在科学研究中的重要性更甚从前，数据的可用性和重用性备受关注。与此同时，国内也有机构开始探索数据共享平台的价值与应用。特别是科技部组织实施的科学数据共享工程。科学数据共享工程是在国家科技基础条件平台统一规划、政策调控和相应法规的保障下，应用现代信息技术，整合离散的科学数据资源，构建面向全社会的网络化、智能化的管理与共享服务体系，实现对科学数据资源的规范化管理及其高效利用（详情参阅：http://www.acca21.org.cn/kssjgxgc/index.html）。

面对与日俱增的科研数据共享需求及其复杂性，越来越多的研究人员开始转向与其联系紧密的图书馆寻求帮助。而随着2007年科学研究的第四范式兴起，有关数据监管（Data Curation）的研究在图书馆界蓬勃发展起来。数据监管是以科研数据的长期保存、组织、维护、管理和再利用为重点任务，最终实现数据共享的新兴研究领域，有时也表述为Digital Curation。在本文讨论中，将Data Curation译成“数据监管”，一是体现其数据保存和管理的作用，二是突出其对数据的加工处理过程。

第一作者简介：于霜（1992- ），女，华东师范大学商学院信息学系本科生，研究方向：计算机信息系统。
收稿日期：2013年8月2日。

那么，数据监管是什么呢？在“数据监管”的英文名称Data Curation中的“Curation”一词源于拉丁语，本意为照顾，后来发展到博物馆领域，译为策展，其得益于博物馆、图书馆和生物学，是E-Science环境下科学数据共享和大规模科学计算的产物，并由微软研究所首席研究员Jim Gray[2]于2002年在文献中正式提及。

作为较早研究数据监管的机构，英国数据监管中心（Data Curation Centre，简称DCC）把数据监管定义为“在数字化的科研数据的整个研究周期中对其进行维护、保存以及增值的过程”[3]。伊利诺伊大学科学信息和学术中心（CIRSS）图书馆与信息科学研究生院[4]认为，“数据监管是凭借数据的生命周期在自然科学、社会科学和人文科学领域对其学术和教育活动的利益和效用对数据进行的积极和持续的管理活动”；维基百科[5]将其定义为“用来保存长期保持其可重用性的研究数据的保存管理活动”。从以上定义可以看出，虽然相关定义尚未统一，但可提炼出共同之处，即：数据监管不仅是要保存现在使用的数据，而且要存储能用于未来再利用的数据，它贯穿于数据的产生、组织、存储、加工利用的全过程。

此外，为进一步理解数据监管，还需要厘清与它相关的两个概念：数据归档（Data Archiving）和数据保存（Data Preservation）。本文在英国联合信息系统委员会（JISC）[6]对三者概念界定的基础上进行梳理。具体表述如下。

（1）数据监管（Data Curation，简称DC）能够满足当前使用的需要并被用于未来的发现和利用，从数据产生开始就对数据进行管理和完善的活动，它涉及政策咨询层面和系统工具层面。

（2）数据归档（Data Archiving，简称DA）合理地对数据进行选择、储存，以确保其物理上或概念上的完整性，并具备可获取性、安全性和可靠性的活动，它是从内容层面确保数据的可用性。

（3）数据保存（Data Preservation，简称DP）对具体数据对象进行持续维护，以确保其在硬件技术变革后仍能被读取和理解的活动，它是从技术层面确保数据的可持续性。

从上述定义来看，三者之间存有差异，但也有关联，即：数据保存是数据归档的一部分，数据归档需要数据监管的支持，三者相互作用、相互影响（图1）。

由三者的定义来看，数据归档对数据的选择和存储包涵了数据保存对具体数据的维护，而数据归档和数据保存又可以影响数据监管未来对数据的使用和发现。从某种程度来讲，数据监管在对数据进行管理的活动中对数据归档的选择和存储活动起到支持作用。

随着E-Science时代的到来，科学研究对数据的需求日益扩大，数据监管的作用与价值越来越得以彰显并且不可替代。

2　科研数据共享之惑

大数据时代，各类数据呈爆发式增长，且多以非结构化的形式堆放在数据库中。随之而来的科研数据重用与深度挖掘问题亟待解决，并开始受到关注，如：面对时下微博研究甚多且数据集尚未统一，林鸿飞在2013年第十九届全国信息检索学术会议期间呼吁可以像TREC和NCTIR一样建立标准的数据集，并据此共享或开放数据继而推进相关研究[7]；牛登科[8]认为，论文图形背后的数据往往是

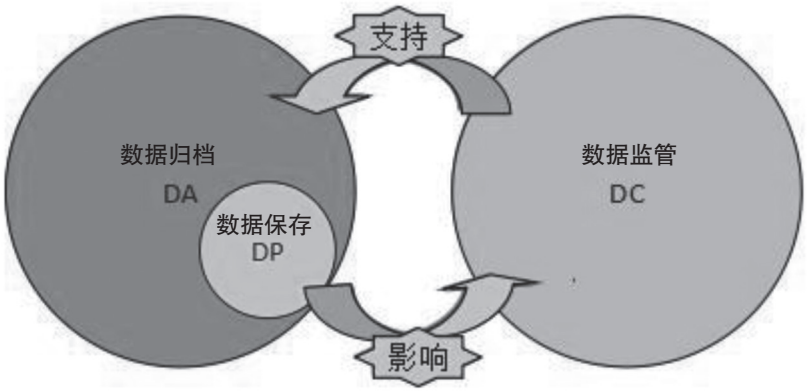

图1　数据监管、数据归档、数据保存三者的关系图

对原始数据进行的分析、提炼或总结且非常便于学者再使用、再分析，它的公开与利用将会推动科研工作的新一轮革新[3]；李亦学坦言科研数据难以共享已成为国内生命科学研究的一大障碍，而在大数据时代，其负面效应还可能被继续放大[9]。

Susan Reilly的一份调查研究也生动形象地揭示了科研数据的共享与重用现状，认为当前科研数据的载体分布呈金字塔状（图2）[10]。自下而上可以看到，金字塔底层代表科研人员存储在本地磁盘中的原始数据和数据集；第二层代表存储在数据库中的文章参考的数据或文章中描述的可用数据；第三层代表对文章数据的补充数据，是经处理后的数据代表；金字塔顶层代表在出版文章中包含或解释的数据。当前科研数据载体的分布情况在一定程度上限制了科研数据的最大化共享，居于多数、最为原始或翔实的科研数据难以被后人验证、引用或者重组。然而，塔尖上文章或刊物中内嵌的数据，限于出版篇幅或易理解考虑，往往倾向于用散点图、柱状图等图形将研究结果直观地显示出来[8]，但数据图形化表示后，读者就只能接受结论，而研究过程中数据的后续分析与利用，就显得十分困难。

通过探究图书馆应用数据监管开展科研数据共享服务的可行性，本文意在倡导开放“底层”的科研（实验）数据，提高科学研究活动中原始数据和数据集的重用性与深度挖掘的可能性，真正发挥数据“载入一次，多次使用”的作用，由此为各类研究人员提供数据支持，以推进科学研究活动的深入发展。

3　数据监管在科研数据共享中的作用

3.1　数据监管解决科研数据存储问题

小型科学研究具有科研经费少、团队人员少的特点。目前，我国大多数的科学研究都属于这种小型的科学研究，数据存储、设备故障等问题常常增大了研究人员的负担，而数据监管以其高级的存储设备和科学有效的管理为研究人员高效地解决了这些问题。

3.2　解决跨域的科研数据共享问题

随着科学技术的发展和网络的普及，科学研究跨域合作现象增多，但是这样的跨域合作将因数据的传输和共享问题而影响科研进展。数据监管服务同样可以很好地解决这一问题。

3.3　提升科研数据的重用能力

开放塔底（图2）数据，提升科研数据的重用性和深度挖掘的可能性，这一方面为研究人员的研究提供数据支持，另一方面也为数据的后续开发利用创造条件。除此之外，科研数据的集成共享与管理，也为相关科研辅助分析挖掘工具的研制提供丰富的源材料，这有助于优化科研活动。

3.4　推动科研教学有机统一

魏红等认为，教师的科研成果和教师的教学效果呈现较为显著的正相关，教师的科研对其教学是有促进作用的[11]。作为科学研究活动中的重要组成部分，科研数据是重复科学研究的必要素材，这对于后人重复前人研究、培育学生研究兴趣是极其重要的。同时，科研成果在实验教学活动中也能得以

图2　Susan Reilly的数据发布金字塔

进一步的检验、优化和完善。

4　国内外图书馆对数据监管的应用探索

近年来，国外图书馆界有关数据监管的研究持续高涨，并积累了不少研究成果。笔者于2013年7月28日在WOS（Web of Science）上检索有关数据监管的研究论文（不包括2013年发表的论文），经汇总统计后其发文量如图3所示。

由此可见，数据监管被提出的2001至2004年间，并没有引起学者的重视，而2005年以后数据监管逐步发展起来，越来越多的学者开始关注它，且在2008年有一次正向跳跃，2011年相关数据监管的发文数飞跃到24篇。自2007年起，数据监管发展极为迅速，这或许得益于美国于2007年9月启动的DataNet计划。为了进一步厘清数据监管在国外的发展脉络，笔者汇总了一些重要的时间节点。

（1）2001年10月“Digital Curation：digital archives，libraries and E-Science seminar”国际研讨会在伦敦举行，由此奠定了数据监管基础。

（2）2002年6月微软研究所首席研究员Jim Gray在文献中首次提及数据监管，但没有明确的定义。

（3）2004年3月联合信息系统委员会和电子科学核心项目联合组建英国National Digital Curation Center（DCC），并为其做了明确的定义。

（4）2007年4月北卡罗莱纳大学教堂山分校的信息与图书馆学学院主持召开的国际性会议DigCurr 2007探讨从事数据监管的具体做法。

（5）2007年9月美国国家自然科学基金委员会（NSF）启动DataNet计划，明确以图书馆为主体，预算1亿美元，用5年时间资助5项数据监护重点研究课题。

（6）2009年8月新墨西哥大学主持的DataOne（Data Observation Network for Earth）项目获得NSF的全额资助。

（7）2009年10月约翰霍普金斯大学图书馆系统研发的Data Conservancy项目获得了NSF的2000万美元的资助。

（8）2011年9月数字图书馆理论与实践国际会议在柏林召开，进一步推进了数据监管的实践。

虽然，国外图书馆界已经在数据监管的研究和实践中取得了一定的成果，但是在国内图书馆界有关数据监管的研究始于2011年，才刚刚起步。鉴于国内相关学者对“Data Curation”的表述不一，笔者分别用“数据监管”、“数据监护”、“数据策展”等词语在CNKI数据库中按主题检索发现，到目前为止CNKI共收录了11篇相关文献，如夏姚璜的《欧美Data curation的实践及启示》、高红文和陈清文的《国外数据监管研究综述及启示》、杨鹤林的《从数据监护看美国高校图书馆的机构库建设新思路——来自DataStaR的启示》、刘雄洲和王菲的《国外数据存管实施现状及其对国内高校图书馆的启示》等4篇论文分别介绍了国外数据监管的发展现状及其对国内图书馆（尤其是高校图书馆）的启示；丁培的《数据策展与图书馆》、时婉璐和任树怀的《数据策管：图书馆服务的新创举》、杨鹤林的《数据监护：美国高校图书馆的新探索》、张秋彦的《高校科学数据监护》、沈婷婷和卢志国的《数据监管在我国高校图书馆的应用展望》等5篇论文从不同角度阐释了数据监管和图书馆的关系；崔宇红的《E-

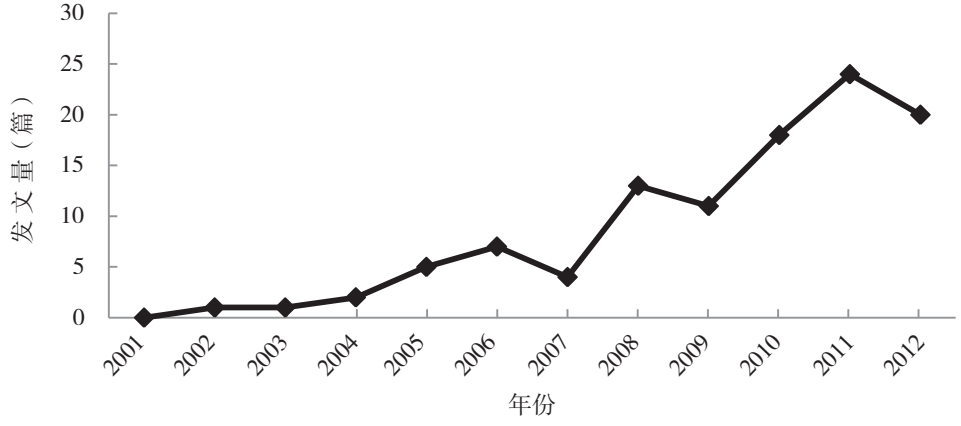

图3　WOS中有关数据监管研究的发文量（2001-2012年）

Science环境中研究图书馆的新角色：科学数据管理》、沈婷婷和卢志国的《科研项目不同阶段的科学数据监管的方法》则对数据监管的管理方法进行了深入的探究。

5　图书馆在数据监管服务中的定位

随着信息技术的发展，人们对信息的需求量越来越大，获取途径也越来越多，且随着出版、印刷技术的提高、电子媒体和通讯技术的发展，信息的传播速度加快，特别是电子计算机和互联网的普及，为信息的传递和利用提供了更为广阔的空间，图书馆在信息服务中作用正在逐渐淡化[12]。不过，数据监管的兴起及其深入发展，为图书馆提供了难得的发展契机。

2012年，柯平提出构建第四代图书馆——网络图书馆的三维空间模型（资源维、服务维和管理维）[13]，为图书馆的发展建构了一个立体模型（图4）。第一维是资源维，没有资源不可能成为图书馆，图书馆的资源对应文献（Document）、信息（Information）和知识（Knowledge）；第二维是服务维，服务维面向读者或用户，没有服务可能是藏书楼，不是真正的图书馆；第三维是管理维，管理出效益，管理依靠图书馆员，管理也是发挥资源和服务作用的动力机制。在此基础上，数据监管不仅为这三维的拓展提供了更加广阔的平台，而且也为图书馆明确了自己的定位。

从目前国内外研讨探索中得知，图书馆应用数据监管所扮演的角色主要有如下3种。

（1）从资源维的角度来讲，图书馆为资源保存机构。由于长时间的积累，图书馆不仅在资源保存类型上多种多样，而且在资源保存方面有着丰富的经验。数据监管服务为图书馆尤其是高校图书馆带来了更多有价值、高质量的科研数据资源，丰富了图书馆的馆藏。

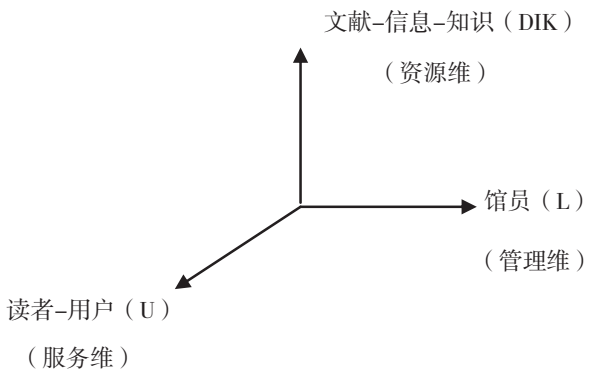

图4　图书馆的三维空间模型

（2）从服务维的角度来讲，图书馆是为学科服务的提供机构。图书馆可以为研究人员提供一对一的学科服务，并为其科研项目提供数据监管服务。同时，数据监管也帮助图书馆将目标由科研成果转为科研支持，从而摆脱简单的“收藏者”角色。

（3）从管理维的角度来讲，图书馆为数据监管的素质教育培训机构。数据监管包含了数据的筛选、甄别、收集、存储和更新等管理活动，为图书馆培养了高素质的专业人员，提高了图书馆的服务质量和管理效率。同时，还能够在激发学生科研兴趣、培养学术科研动手能力等方面起到一定的推动作用。

6　图书馆应用数据监管的困境与对策

尽管国内外对有关数据监管在图书馆中的研究探索取得了一定成果，但是总体来看数据监管从提出到现在也只有短短的10余年时间，图书馆应用数据监管开展科研数据共享活动来完成资源保存到数据管理的角色转换依旧面临着巨大的挑战。

参考图书馆三维空间模型（图4），在先前的研究成果基础上，笔者进行了系统的汇总梳理，认为图书馆应用数据监管开展科研数据共享活动面临七大挑战，并给出了相关的解决方案（表1）。

7　结语

作为科学研究活动中的重要组成，科研数据是后人重复前人研究工作、孵育科研新人、激发学生科研兴趣等的重要素材。它的公开共享，在提高科研成果的透明度和重复性同时，也能够推动科研工作新一轮革新[8]。当前科研数据共享需求强烈，科研数据的重用与深度挖掘问题亟待解决，且已经得到了学者关注与广泛讨论[7–9]，如，数据堂正是在这样的背景下应运而生。数据堂是数据堂（北京）科技有限公司的简称，由国家科技部大力支持，并通过与国内外著名科研机构、高等院校、研发企业通力合作，积累丰富的科研数据资源，借助统一的平台提供服务，使得科研机构、企业、高校和个人之间实现充分的数据共享。

数据监管，作为E–Science环境下科学数据共享和大规模科学计算的产物，自提出至今，得到了各界的广泛关注。它不仅强调要保存现在使用的数

表1　图书馆应用数据监管开展科研数据共享所面临的困境及对策

维度	挑战	详细描述	对策
资源维	数据获取问题	科研数据哪里来	加强与研究人员沟通、阐明利弊 充分利用第三方开源数据、努力争取国家教育上和政策上的支持
	数据存储问题	哪些数据需要保存	进行数据评估、质量控制
		保存期限	因地制宜，区别对待
		怎么保存	开展入库标准研究
	数据格式问题	数据格式多元化	限定数据格式种类
	数据管理问题	对数据归档、检索能力要求高	培养图书馆员专业技能
服务维	数据利用问题	谁有权使用	明确使用权限及规范
	知识产权保护问题	数据使用过程中侵害数据产生者的知识产权	引入数据引证[14]的机制与方法
管理维	管理人员配置问题	针对馆员的新要求，也会给图书馆带来人员培训负担	加强业内或与高校的交流与合作、优化资源共享机制，总结与分享好的经验和做法，培养更多的专业管理人员

据，而且强调要存储能用于未来再利用的数据。它贯穿于数据的产生、组织、存储、加工利用整个过程，为科研数据共享提供了很好的思路，而图书馆（尤其是高校图书馆），与研究人员有着天然的紧密联系，在科研数据共享与利用方面存有诸多优势。因此，应用数据监管开展科研数据共享服务，也为图书馆发展提供了一次难得的发展契机。不过作为一种新兴服务，图书馆也面临着许多挑战。同时，科研数据得以共享之后，如何让科研数据产生最大价值还是一个值得进一步研究的课题。

简言之，图书馆应用数据监管开展科研数据共享服务，任重道远，但前景看好。

参考文献

[1] Jim G. On E−Science—A Transformed Scientific Method [C]// Tony H, Stewart T, Kirstin T. The Fourth Paradigm: Data−intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009:19−33.

[2] Jim G. Online Scientific Data Curation,Publication,and Archiving[EB/OL]. [2013−06−10]. http://research.microsoft.com/pubs/64568/tr−2002−74.pdf.

[3] What Is Digital Curation?[EB/OL].(2012−11−26)[2013−07−10]. http://www.dcc.ac.uk/digital−curation/what−digital−curation.

[4] CIRSS. Data Curation Education Program[EB/OL]. (2011−10−09). [2013−06−10]. http://cirss.lis.illinois.edu/CollMeta/dcep.html.

[5] Data Curation[EB/OL]. [2013−06−10]. http://en.wikipedia.org/wiki/Data_curation.

[6] Lord P,Macdonald A. Data Curation for E−Science in the UK: An Audit to Establish Requirements for Future Curation and provition[EB/OL]. (2011−12−03). [2013−06−11]. http://www.jisc.ac.uk/upload_documents/E−Science ReportfFnal.Pdf.

[7] 林鸿飞.林鸿飞的微博[EB/OL].(2013−07−20).[2013−07−21].http://weibo.com/1937618377/A0SzqmgoU.

[8] 牛登科.Nature系列期刊促进信息传播交流的新举措[EB/OL].(2013−05−07).[2013−07−01].http://blog.sciencenet.cn/blog−61772−687399.html.

[9] 李亦学.科研数据难共享阻碍国内生物科技发展[EB/OL].(2013−07−17).[2013−07−21].http://www.biodiscover.com/news/research/105002.html.

[10] Reilly S. The Role of Libraries in Supporting Data Exchange[EB/OL]. (2012−05−24). [2013−06−10]. http://conference.ifla.org/past/ifla78/116−reilly−en.pdf.

[11] 魏红,程学竹,赵可.科研成果与大学教师教学效果的关系研究[J].心理发展与教育,2006(2):85−88.

[12] 章洁.信息时代的图书馆:挑战与变革[J].贺州学院学报,2008,24(4):121−127.

[13] 柯平.重新定义图书馆[J].图书馆,2012(5):1−5.

[14] 侯经川,方静怡.大数据时代的数据引证研究:进展与展望[J].中国图书馆学报,2012(6):1−7.