美国普渡大学图书馆的科学数据管理服务

胡雪环 屈宝强
（中国科学技术信息研究所，北京 100038）

摘要：对普渡大学的科学数据管理政策进行解读，在此基础上对其主要政策内容进行整合并将之归纳为数据收集政策、数据保存政策、数据使用规定、数据管理与维护四大部分，进一步总结出高校在制定数据管理政策时的注意事项，从微观层面为我国高校图书馆制定科学数据管理政策内容提供具体参考和借鉴。

关键词：高校图书馆；科学数据管理；收集政策；保存政策；使用规定
中图分类号：G350 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2015.06.011

Research on Scientific Data Management Policy in Purdue University Library

Hu Xuehuan, Qu Baoqiang
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: This article mainly introduces the scientific data management policy in Purdue University. On this basis, summarizes the contents into four parts :data collection policy, data preservation policy, data use rules and data management and maintenance , Further summarize the considerations needed to be aware of when universities make data management policies ,in order to provide content references for the research data management policy of university library in our country.

Keywords: academic library, scientific data management, collection policy, preservation policy, use rules

1 引言

作为高校教学和科研信息资源保障中心，图书馆应充分利用自身优势，积极探索符合本校的科学数据管理计划和完善的数据管理政策，主动承担起高校科学数据管理的重要职能，并将计划的制定、政策的维护、版本的更新、数据管理与服务等工作纳入正常业务范围，以满足科研人员、资助机构以及学生对本校科研数据发现、使用、管理、保存和开放共享的长期需求。高校是科学研究的主要阵地之一，其科学数据的有效管理和保存对于追踪高校科研数据来源、避免重复研究、促进数据再利用和加快科学研究进程等方面发挥着重要的作用。在E-science环境下，科学数据管理、保存和共享的需求越来越高，已引起各国政府和学校的高度重视。2007年，美国国家科学基金会（NSF）启动了DataNet计划，明确提出以图书馆为主体实施科学数据管理[1]。英、美等国多所高校也相继制定了科学数据管理政策，比如：爱丁堡大学制定了研究数据管理政策，牛津大学制定了研究数据及记录管理政策，斯坦福大学制定了研究数据保存、获取政策，杜克大学制定了研究保存、共享、所有权政策等，从政策上确保了科研数据管理与服务的长期有效

作者简介：胡雪环*（1990-），女，中国科学技术信息研究所硕士研究生，研究方向：数字图书馆与数字出版；**屈宝强**（1980-），男，中国科学技术信息研究所副研究员，研究方向：文献共享、数据共享。

收稿时间：2015年7月13日。

开展。在我国，尽管部分高校图书馆已经开展了科学数据管理的实践和探索，但是没有像英、美等国诸多高校那样制定了成熟的数据管理政策和指南。这种缺乏完善政策指导的实践不利于高校科研数据的长久管理与共享利用。

面对英、美等国高校比较成熟的科学数据管理政策和国内高校系统相关政策规定的缺失，有必要对一些有代表性的高校科学数据管理政策进行深入的解读和探究，从政策内容的具体制定层面为我国高校图书馆科学数据管理政策的制定提供一定的范本和参考。普渡大学是美国典型的研究型大学，并且较早开展了科学数据管理与服务。其图书馆承担了主要的数据管理工作且制定了完善的科学数据管理政策，并在实践中不断地得到更新与修订。本文将对美国普渡大学图书馆的科学数据管理政策进行分析，重点阐述普渡大学的数据收集政策、数据保存政策、研究仓储使用规定、数据管理与维护以及面临的挑战与风险，并通过普渡大学图书馆数据管理政策的内容，探知在政策制定中需要注意的若干关键问题，从而形成几点启示，以供我国高校图书馆制定具体政策内容时参考。

2　普渡大学图书馆科学数据管理概述

美国普渡大学是一所典型的研究型大学，比较早地开展了科学数据管理与服务。 在制定科学数据管理政策之初，普渡大学便明确规定了开展数据管理工作的主要目的是向该校科研人员和学生提供数据资源和服务，促进科研数据管理、传播和保存。具体目标是：收集、发布和保存隶属于普渡大学或者与普渡大学研究项目相关的数据集和数据文档；帮助普渡大学的研究人员满足基金资助机构对管理、共享和保存研究数据的需求；为研究者、政策制定者及其他人发现和获取研究数据集提供一种渠道和方式；提供可持续的保存环境，使存放的研究数据能够支持研究的历史记录，并且方便获取使用最新科研知识[2]。

当前，普渡大学图书馆主要开展8项数据管理服务：（1）在线普渡大学研究资源库（Purdue University Research Repository，简称PURR）；（2）帮助科研人员制定数据管理计划；（3）元数据处理及数据保存；（4）提供分布式管理中心，进行专业的数据引用和高效科研培训；（5）开展课堂教学和实验数据管理指导和教育；（6）组织、获取、保存和记录大数据；（7）数据咨询服务；（8）数字对象标识符（DOI）服务[3]。PURR是由普渡大学图书馆和普渡大学信息技术部以及研究副校长办公室合作研发与提供支持，是开展上述8项系列服务的重要平台，各项服务的开展和相关说明均被纳入到该研究资源库政策中进行解释和规定。

3　普渡大学图书馆科学数据管理相关政策

3.1　数据收集

（1）数据收集的特定对象

在收集政策中指出，PURR是一个提供服务和虚拟研究环境的机构知识库，主要为普渡大学的研究人员和其直接相关的合作者提供数据管理支持，凡是普渡大学的教师、职员、学生和任何从普渡大学毕业的研究生，都可以创建项目和提交数据。非普渡大学的合作者则需要至少与一个有效的普渡大学注册用户有关（比如合作开展科研工作等）才有资格提交数据集[4]。

（2）数据收集的学科范围和条件

收集政策的目的是加强普渡大学所有学科领域研究数据的收集和管理，从各个领域和学科收集来的数据在PURR进行发布或存档需要满足以下条件：数据的提供者是PURR指定的成员；被提交的数据必须是普渡大学、图书馆、PURR以及他们各自的政策和规定所允许的；被提交的数据必须是法律部门和法规条例所允许的；当一个用户提交一个数据集，他或她便授予大学非独占许可权，使普渡大学对数据集有永久管理、发布的权利[4]。

（3）数据收集的格式推荐

PURR的数据收集包括研究数据集和相关信息。其中，研究数据集通常是文件和元数据的集合，包括与研究数据相关的保存和显示信息以及注释和辅助内容。所有的提交内容将以数字化形

式存在。提交的数据必须与普渡大学的科研项目和研究教学任务有关，且接收的数据范围更侧重于原始数据的输入输出，如电子表格、传感器和仪表数据、调查、记录、图片、视频和软件源代码等[3]。另外，在数据格式推荐部分，不仅列出了PURR可接受的数据格式，而且针对不同的数据格式，指明了哪些适合于长期持续存储，哪些是PURR支持但不利于长期存储的以及哪些是无法进行长期持续保存的数据格式，以指导用户根据特定需求选择合理的数据格式[5]（表1）。

在进行数据收集时，由PURR的工作人员进行文件格式识别，并验证文件的原始格式，未来任何有关数据的转换和迁移工作都将包括原始文件格式信息和任何有关该对象数据集的历次修改信息。

3.2 数据保存

PURR作为普渡大学的研究数据资源库，必须用来支持教学、科研等活动，因此，图书馆的一个关键使命就是要保存该校学术信息资源，负责识别和保存不断增加的数字格式资源，使这些数据能够支持大学的研究、教学和学习需要。

（1）数据保存参与者角色与职责分配

虽然图书馆员和档案员在保存和提供学术资源访问方面担任着主要责任，但是从当前来看，数字资源的保存已经成为所有利益相关者的共同责任。PURR指导委员会（包括图书馆馆长、研究副校长以及信息技术部副校长和首席信息官）要负责评估和批准有关提交给PURR的资源内容的相关政策和规程。档案员和图书馆学科专家负责甄别和筛选长期保存的数字内容。而对于从事科学数据长期保存的教职员工以及其他研究人员则有义务将其科研数据转换成符合PURR要求的数据保存格式、元数据和相关配合行动[5]。

此外，普渡大学图书馆、信息技术部和研究副校长室等相关责任部门也在致力于通过合作发展的途径使PURR成为一个持久可信的数据仓储，积极倡导与其他大学、图书馆、机构和组织的合作，进一步就数字内容保存的技术等方面进行探讨，以共同实现数字资源有效保存共享的愿景[5]。

（2）数据保存期限

图书馆的数字保存工作虽然有识别、保护、提供数据保存方法的责任以确保可持续访问选定

表1 数据格式推荐

数据文件类型	利于持续保存的数据格式	被支持的数据格式	不利于持续保存的数据格式
文字处理	PDF/A, Open Document Text	PDF/B, Microsoft Word, Microsoft Open XML, Rich Text Format	Corel Word Perfect, Lotus WordPro, PDF
纯文本	Plain Text, Comma−separated file, Tab−delimited file		
结构化标记	SGML w/DTD, XML w/DTD		SGML w/o DTD, XML w/o DTD
电子表格	Open Document Spreadsheet, Comma−separated file, Tab−delimited file, PDF/A	Microsoft Excel, Microsoft Excel Open XML	
数据库	Delimited Flat File w/DDL	Microsoft Access, dBase Format	
音频	WAVE	AIFF (uncompressed), Standard MIDI, Windows Media Audio, MPEG, MP2AAC	Audio CD, DVD−Audio, RealAudio, Shorten, RIFF−RMID, Extended MIDI, Module Music Formats
视频	Archival format not currently established.	AVI, MPEG−1, MPEG−2, MPEG−4, Quicktime	Windows Media Video
图像	TIFF, JPG 2000	JPEG, PMG, PDF/A, GIF	RAW, Adobe Photoshop, Kodak Photo CD, Encapsulated PostScript, Flash Pix, PDF

的数字资产。但是并不意味着所有存储在PURR里的数据资源将被永久保存和收录。对于上传到PURR的数据会有一个10年的基本保存期限；超出了10年的数据将根据长期保存项目的相关标准由专业馆员进行筛选保留，而这一行动的有效实施还要依赖于相关的待批准的预算以及其他资源的保障[5]。因此，普渡大学图书馆必须考虑对这些数据进行合理分类，并设定合理的优先保存规则，目前，针对以下几种数据将优先享有保存资格[5]：与出版发表的论文或著作相关的数据集；独立的数据出版物；有较高的研究教学价值的数据集；其他经过筛选的数据文件和材料。

（3）数据保存原则

该部分详细列出了普渡大学数据保存具体遵循的原则，比如：遵守开放档案信息系统（OAIS）参考模型标准和其他数字保护标准；寻求符合ISO 16363标准认证要求的可信数字仓储；遵守一切知识产权、版权和所有权的保护规定；建立安全的和充分的数据备份和灾难恢复保障等，以力求PURR能实现长久地发展和持续访问的目标[5]。

（4）数据保存标准

在开展数据保存行动时，对于需要保存的数据进行筛选和收录的原则主要是基于定期检查和更新的通用标准规范。同时，对于筛选出来的用于持续保存的内容要求能够支持普渡大学的教学和学术研究，除此之外，要符合国家和国际馆藏的发展和维护标准。其他有关数字保存形式和功能的标准，由于对数字资源的持久有效保存有重要影响，在保存数据筛选原则和标准的制定中而被考虑。此外，增强普渡大学图书馆馆藏的内容范围也是图书馆开展数据保存行动的指导方针之一。因此，针对保存在其他数据仓储中但对该校教学科研有重要影响的数据也会择优被PURR收录保存[5]。

（5）数据保存级别

进入PURR的每一类数据对象都将根据数据集自身特征和保存目标等按照某一特定类型的保存策略进行保存。PURR提供了3种级别的保存策略：位级保存（Bit-level Preservation），即基础的保存级别；有限保存（Limited Preservation），即较高级别的保存；完全保存（Full Preservation），即最高级别的保存。每一个保存级别下都对应着具体的保存行动，见表2所示。

3.3　数据使用

普渡大学图书馆致力于其科研数据的开放获取，认可开放存取的柏林宣言的核心理念。可访问的用户包括普渡大学的教师、研究人员、研究生，与普渡大学有合作的相关研究者；隶属于其他研究机构的教师、学生和研究人员；独立学者和公众等。虽然普渡大学致力于科研数据的开放共享，但是，仍要遵守相关的知识产权法律法规，部分机密性数据不会给予公开，部分不适合向大众开放的数据只能在小范围内公开等。具体的使用条款、保证条款等简介如下。

（1）使用条款

在PURR使用条款部分，主要包括协议的修改说明；注册，访问和终止用户使用PURR的条件说明；用户在上传数据时的隐私和保密选择说明；使用PURR必须遵循的行为规范和禁止行为说明；普渡大学对于PURR用户数据损坏或丢失等情况的免责声明；在发生法律纠纷时所遵循的法律选择和律师费等情况说明[8]。

（2）保证条款

使用普渡大学研究仓储必须同意其保证条款，该部分规定了数据发布者在发布数据前必须同意的系列条款，其中包括授予普渡大学在一定范围内的复制权、分发权等；同时要求数据发布者保证上传的科研数据不侵犯他人知识产权，不违反相关法律条款；在上传的具体内容里不应包含任何软件病毒或任何其他有可能破坏数据管理系统的程序代码，也不能含有高风险的机密信息；涉及人类敏感问题的科研数据需征得IRB部门批准等[9]。

（3）侵权说明

由于数据的开放共享，侵权行为在所难免。在侵权说明部分，主要是指导PURR使用者如何按照合理的程序方式解决侵权问题，比如：当使

表2　不同级别的保存策略及其具体保存行动[7]

保存策略	保存行动
位级保存	DOI对象标识符
	保存元数据
	安全存储和备份
	定期病毒检测
	定期固定性检查
	比特流维护
	转换和标准化
有限保存	针对格式变化实施战略监控
	迁移到更适合保存的格式
全保存	迁移到适合持续保存的格式

注：较高级别的保存策略中涵盖较低级别的所有保存行动。

用者发现版权侵犯行为时，需要提交相应的材料和收集足够的证据，最好是有完整的URL等事实证据，以帮助快速定位内容，维护相应权益[10]。

（4）访问声明

为了尽可能地提高网站的可访问性和可用性，该部分主要列出了网站遵循的一系列指导方针和标准规范、技术测试规定等[11]。

3.4　数据管理与维护

PURR有责任保护其用户提交的数据完整性，以确保数据能够持续访问。因此，制定了完备的数据管理与维护措施[12]。

（1）元数据保存：每一个提交到PURR的数据集都将全面实现其元数据保存，PURR采用多个元数据标准以确保充分描述不同数据集的特殊格式和独特性质。Dublin Core Metadata Initiative用于提供发现和引用数据；MODS（Metadata Object Description Schema）用于记录数据集的创造者和访问权限；PREMIS（Preservation Metadata Maintenance Activity）用于记录每个数据集所经历的保存事件和法律权利分配；METS（Metadata Encoding and Transmission Standard）用于表示数据文件的结构和层次体系结构。

（2）数据格式识别：对获取的每个数据集进行文件格式分析。采用技术注册表，PRONOM和格式识别工具、DROID来验证每个数据集的格式，该信息用于记录对象的整个生命周期中潜在的数据转换、迁移和固定性检查等。

（3）安全存储和备份：所有PURR保存的数据都将被完整的复制，并备份到另外的网站，以预防灾难性的信息损失或者定期检查时造成的数据丢失。

（4）固定性检查：所有的PURR保存的数据都将定期进行固定性检查，以确保没有数据丢失，并对已经损坏的无用数据进行定期清除。

（5）变换/标准化：由于提交到PURR的数据集在一开始不是结构化的数据，所以必须对数据进行相应的格式转换和标准化处理，使之符合保存的基本格式要求。当然，在可能的情况下，PURR会尽量保存数据的原生格式；即使因为长期保存和标准化的需要必须转换格式，转化数据也将被记录在其整个生命周期的数据元数据中。

（6）迁移：为了确保长期保存和访问获取，对于那些已过时的格式保存的数据将进行转换，转换迁移可能包括升级数据集到一个新的版本，转换到一个新的文件格式或文件结构。当然，这些变化也会在描述数据集的元数据中有所记录和体现。

3.5　数据管理挑战和风险

数字资源的管理与保存涉及的内容远远不

只是资源的有无，还包括技术、资金、人力等各个方面的因素。在政策部分，普渡大学图书馆也列出了其在进行科学数据管理时面临的各类风险和挑战，主要包括以下几点[6]：一是技术的识别以及跟上技术的发展变化；二是成本，主要包括涉及人员、设备、软件和基础设施的费用以及其他各项费用；三是开发和维护一个成功的沟通框架，针对开发商、管理员和用户不断变化的需求、实践等，能及时充分地识别和应对；四是如何实现长久的数据保存，保存计划贯穿于整个数据管理过程，而完成整个项目数据的保存并非易事；五是适应各种各样的数据集，未来数据的存在会有更多类型、格式、大小等，充满复杂性。

PURR的发展虽然面临诸多挑战，但是，随着越来越多的基金资助机构要求科研项目承担者必须提交相应的科研数据管理计划，以描述其数据管理细节，因此，PURR在帮助科研基金申请者满足资助机构的数据管理需求上有着巨大的优势。而且，PURR对收录的数据集会分配相应的数字资源唯一标识符（DOI），以方便其他科研人员发现和引用数据，这也给普渡大学提供了一个证明其科研影响力的良好机遇。

4　几点启示

透过普渡大学图书馆数据管理政策的内容，探知在政策制定中需要注意的若干关键问题，仅供我国高校图书馆制定具体政策内容时参考。

（1）明确各利益团体的责任和义务。科研数据的管理、保存和利用涉及多方利益：科研人员或科研团体、基金资助机构、数据保存管理机构、数据使用者等。在政策制定时，一定要明确各方的责任和义务并进行具体的文字规定，比如科研人员在提交数据时需要知悉哪些事项，同意对哪些数据进行处理的规定；数据保存机构对于科研人员上传的数据有哪些处理权限，对于不同保存级别的数据有哪些具体的管理行为；使用者在使用数据时需要遵循哪些版权规定等。只有这样，才能有效避免因知识产权问题带来的系列数据管理和使用纠纷，保证科学数据从收集到保存利用的顺利进行。

（2）对数据保存中有可能导致的数据丢失或损坏问题划定责任归属。图书馆作为科研数据的具体管理和保存机构，有义务最大限度地保持数据的完整性。但是在管理保存科学数据时，由于数据的筛选、迁移和格式转换等过程中会不可避免地出现一定的数据丢失，因此，针对数据丢失或者损坏的问题，一方面要划清责任归属问题，另一方面要对数据管理保存中可能出现的问题在政策中予以明确说明，避免由于此类问题而产生的数据丢失纠纷。

（3）制定全面的数据收集、筛选、保存标准和使用规定等。政策的制定是为了更有效地指导科学数据管理工作的开展，高校图书馆在收集科学数据时主要有两方面的考虑：一是要满足数据使用者对共享利用的需求，二是要满足图书馆进行长期数据保存的需求。因此，数据收集的范围、数据筛选的标准、数据保存的标准和级别、数据的使用条款、访问声明等必须作为政策制定的重点内容加以细化。同时，考虑到数据长期保存的需要，在政策制定时，需要对科学数据的管理与维护制定详细的实施策略，比如：针对不同格式、不同类型的文件要有完整的元数据配套方案；针对不同级别的科研数据，必须制定有针对性的数据保存策略，以防止后期由于成本、人力和数据的快速增长所导致的存储容量问题，针对不断出现的新型数据格式及保存平台，要制定合理的数据迁移和固定性检查措施，尽量减少数据丢失。总之，在政策制定时，建立完善的后期维护与检查机制至关重要，是科学数据得以长期保存和持续获取访问的有效保障。

科学数据的长期管理和共享利用是科研领域未来发展的必然趋势，高校作为科学数据产生的重要机构。为了将来科学数据管理活动的顺利实施，应该将政策的制定纳为科研管理实践的重要内容之一。从当前开始，国内各高校图书馆和相关负责领导应该主动增强科研人

（下转第81页）

中国标准出版社,2014.
[9] 党跃武,谭祥金. 信息资源导论[M].3版.北京:高等教育出版社,2015:1−20.
[10] Nonaka I. The Knowledge−creating Company[J].Harvard Business Review,2007,85(7/8):162.
[11] 张爱霞,沈玉兰.美国政府科技报告体系建设现状分析[J].情报学报,2007(4):496−502.
[12] Nonaka I, Chia R, Holt R, et al.Wisdom, Management and Organization[J].Management Learning, 2014,45(5):640.
[13] 贺德方.中国科技报告制度的建设方略[J].情报学报,2013,32(5):452−458.
[14] 吴春玉,苏新宇.各种“场”及其在知识创造过程中的作用[J].情报学报,2004, 23(2): 247−520.
[15] 陈卫红.科技期刊知识管理的战略思考[C].2008年第四届中国科技期刊发展论坛论文集.中国科学技术协会、新闻出版总署、中国科学技术期刊编辑学会,2008:5.
[16] 任建,刘振峰,邹新国.知识管理在政府数字信息资源建设中的应用[J].山东煤炭科技,2009(3):185−186.
[17] 贺德方,曾建勋.科技报告体系构建研究[M].北京:科学技术文献出版社,2014:125−130.
[18] 曾建勋,许燕.关于科技报告版权登记制度的思考[J].中国科技资源导刊,2015(5):20−25.
[19] 周杰.科技报告集成管理系统构建[J].情报学报,2014(8):808−812.

（上接第64页）

员对科学数据管理的重要性认识和数据共享的自觉意识。图书馆要主动承担新的服务职能，重视学科馆员和学科专家在数据筛选时的重要作用，定期开展有效的数据培训工作，培养新型学科专家，尝试牵头拟定科学数据管理政策并负责相关解释工作等，这也为图书馆在数字化环境下实现服务转型提供了一条重要途径。

5 结语

科学数据的长期管理和共享利用是科研领域未来发展的必然趋势，对于实现科学数据资源的有效增值，推动科技自主创新，减少科技领域的资源浪费具有积极作用[13]。科学数据管理政策对于科学数据管理服务与实践有着重要的指导作用。本文主要对普渡大学图书馆开展的科学数据管理服务的政策内容进行了详细的解读与归纳分析，在此基础上为我国研究型大学图书馆制定科学数据管理政策提出若干需注意的关键问题，并呼吁高校图书馆工作人员应在开展科研数据管理实践的伊始将政策问题落到实处，以此为科研数据管理的长期发展提供有效的政策保障。此外，笔者也会继续关注国内外科学数据管理政策制定的相关动态，为我国高校图书馆科学数据管理政策的制定提供参考。

参考文献

[1] 谢春枝,燕今伟.国内外高校科学数据管理和机制建设研究[J].图书情报工作,2013,57(6): 11−17,38.
[2] PURR Digital Preservation Policy[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/digitalpreservation.
[3] Researchdata[EB/OL].[2015−05−09]. https://www.lib.purdue.edu/researchdata
[4] Collection Policy[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/collection−policy.
[5] File Format Recommendations[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/file−format−recommendations.
[6] PURR Digital Preservation Policy[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/digitalpreservation
[7] Preservation Support Policy[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/preservation−support−policy.
[8] Terms of Use [EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/terms.
[9] Purdue University Research Repository (PURR) Terms of Deposit [EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/termsofdeposit.
[10] Copyrights[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/dmcapolicy.
[11] HUBzero Accessibility Statement[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/accessibility.
[12] Preservation Strategies[EB/OL].[2015−05−09]. https://purr.purdue.edu/legal/preservation−strategies.
[13] 王凯,彭洁,屈宝强,等.科学数据管理与共享领域文献计量研究[J].中国科技资源导刊,2015,47(4): 31−39.