美国Dryad数据库共享政策及启示

林芳芳 赵 辉
（中国科学技术信息研究所，北京 100038）

摘要：国外开放数据库数据共享政策对国内数据库建设和运维具有一定借鉴意义，有利于促进国内数据共享的发展。以Dryad数据库为例，从数据收集、数据发布、数据保存、收费和使用4个角度进行分析，采用对比分析法将Dryad与GenBank、Figshare进行比较，提出我国应制定细致、可操作的数据共享政策以及出台相关政策，加强期刊出版商与数据库合作的相关数据共享政策建议。

关键词：Dryad数据库；数据共享政策；开放数据库；GenBank数据库；Figshare数据库
中图分类号：G350 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2015.06.009

Dryad Data Sharing Policy and Its Implications

Lin Fangfang, Zhao Hui
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: It has certain significance for us to study the data repository's data sharing policy, which may promote the development of domestic data sharing. We take the Dryad repository as an example, from the perspective of data collection, data publication, data preservation, charging and using to analysis its data sharing policy. Then we compared with Dryad repository, GenBank and figshare's data sharing policy by contrast method, and put forward some related suggestions, such as developing detailed, actionable data sharing policies and developing relevant policies to strengthen the cooperation between journal publishers and data warehousing.

Keywords: Dryad repository, data sharing policy, open database, GenBank, Figshare

1 引言

科学数据不仅是科学研究的投入，也可作为科学研究的重要成果，是科技创新的基础。实现科学数据共享对科学研究和经济社会发展有推动作用。为更好地对数据进行管理，需要相关政策支持。期刊出版商、科研机构、基金资助机构等为保证机构有效运行以及数据开放访问陆续出台了科学数据管理政策和共享政策。OECD（经济合作与发展组织）2003年提倡所有获得公共财政资金支持的研究数据应能被公众获取、共享，并于2007年发表《公共资助可续数据开放获取的原则和指南》[1]。为便于科学数据的公开获取和共享，期刊出版商如PLOS One在2008年制定了数据共享政策，建议研究人员在发表论文时，提交相关的附加数据，并建议将数据保存在开放获取的机构数据库或多学科数据库（如GenBank、Dryad）中[2]。《Nature》[3]则要求作者将期刊论文的数据或附加信息存储于可公开访问的数据库中

作者简介：林芳芳*（1992-），女，中国科学技术信息研究所信息资源管理专业硕士研究生，研究方向：信息资源管理；**赵辉**（1971-），女，中国科学技术信息研究所副研究馆员，研究方向：信息资源管理、科技资源管理。

收稿时间：2015年7月13日。

（如GenBank、Protein DataBank等），对于有些类型的数据集（如核酸序列、蛋白质序列等）需强制存储在公共数据库中并提供各类型数据集适合的数据库。公共数据库区别于其他数据库在于其面向科研群体广、开放性强、有完善的数据管理政策等特点。Dryad作为公共数据库典型代表，分析其数据共享政策，对我国科学数据共享管理有着重要借鉴作用。

2 Dryad数据库概述

Dryad数据库于2008年9月由美国国家科学基金会资助建立。最初目的是参与北卡罗来纳大学教堂山分校元数据研究中心和国家进化中心联合项目。Dryad是由董事会管理的非营利性会员制组织，会员资格开放给所有利益相关机构，其中包括杂志、期刊&出版商、研究机构、图书馆和资助机构等。到2013年，Dryad从最初的靠基金资助状态过渡到为实现可持续发展将非盈利性和实行数据公开收费融合的状态[4]。Dryad数据库是存放优质数据资源的场所，使科学出版物背后的数据可被发现、可重复使用、可引用[5]。其目标是与学术团体、出版、研究和教育机构、基金资助机构和其他利益相关机构构成学术交流体系来协同、维持和促进学术文献中基本数据的保护和再利用。Dryad提出的联合数据存档策略被许多主流期刊采纳，并推荐Dryad作为存放数据场所。

作为数据库，Dryad支持存放各种类型数据，包括：文本、图像、表格、音频、视频等，提交的数据文件拥有永久可解析的DOI标识。Dryad数据库中数据可提供下载和重新利用，排除经期刊编辑部允许，在暂时限制时间内的数据。利用Dryad数据库平台，研究人员可获取数据，研究和验证公布的数据是否合理，或利用已有的数据解决新问题。出版商可免除出版和维护数据成本，并通过鼓励数据重用来增加出版商的影响和威信。研究机构和图书馆可将数据保存到Dryad中，并获取来自其他机构有价值的数据。截至2015年5月5日，与Dryad合作的期刊超过400多种，数据文件达26256件，作者超过3万个，下载次数超过80万次[6]。

下面将对Dryad数据共享政策进行初步分析。

3 Dryad数据共享

Joyce在《Research Data Management》中指出，产生的大量数据经常储存在研究者个人的电脑、实验室服务器中，只有少量的数据存储在固定地方，有明确归档位置[7]。数据存储位置分散使数据共享难度增加。期刊出版商与数据库合作，要求论文中相关数据提交到指定数据库中，使论文和数据同时出版。这样使数据存储相对集中，便于数据共享。为更好地管理和运行，大型数据库如Dryad会制定详细的数据政策。本文以Dryad数据库为例，从数据收集、数据发布、数据保存以及收费和使用角度对Dryad数据库共享政策进行分析。

（1）数据收集

数据是数据库的基础，制定收集政策可以明确和规范数据收集的范围和内容。Dryad数据库的数据来源可分为合作期刊上发表的数据、非合作期刊上发表的数据以及存储在其他数据库中的数据。Dryad不接受未发表的数据，除非是由国家进化综合中心（NESCent）的科学家创造的数据。Dryad数据库参与国家进化综合中心（NESCent）和元数据研究中心（MRC）的联合项目，且相关元数据研究工作由NESCent和MRC工作人员共同开发。按照规定，国家进化综合中心（NESCent）所产生的数据必须要立即释放给公众，没有限制期间[8]。

对于存储在其他数据库中的数据，Dryad的策略是拷贝数据并保存到辅助库中使数据可用，或提供链接给研究人员使研究人员能够在辅助库中找到数据。对于合作期刊数据，Dryad数据库收集政策是尽可能收集出现在其合作期刊上论文所有数据。虽然该数据库专注服务于合作期刊，但对于非合作期刊数据，其采取的数据收集政策是将那些没有合适的数据库存储了其他领域科学论文数据纳入到自己的数据库中，并不仅仅包括

生物、生态和医学领域的数据。非合作期刊数据须经Dryad管理委员会同意才能发表。数据被收集后，如被认为缺乏足够的科学价值，Dryad则将数据从资源库中删除。

（2）数据发布

数据库的发布政策为数据提交者提供详细信息，使提交者了解数据发布条件和流程。对于数据发布，Dryad数据库规定必须要满足以下7条内容标准[9]：内容必须与已发布的科学、医疗或其他学术研究论文相关联；数据库中的大多数数据包要有同行评议的文档，来源于学术论文或图书但无同行评议文件的相关数据也可接受；保证发布者是数据的创作者或有足够权限对数据进行发布；内容发布符合期刊出版商政策；符合期刊出版商要求的报告准则和格式；数据包的容量不超过10兆，超过则要收费；语言要求是英语。

Dryad数据库是集成一体化的，提交到Dryad中的相关数据默认情况下要与论文同时发表。但如果作者不同意将数据和论文同时公开出版，可选择在论文出版后一年内公开数据。若被出版商编辑部批准，可一年后再公开数据，但要有明确的数据发布日期，且最长不超过10年。对于公布之前的论文或数据，当发现内容有问题时，Dryad会推迟授权给出版商，推迟数据发布。当公布后发现数据有问题时，Dryad在保持内容可用性的情况下，会提供链接来提醒出版商。原始数据被更正时，Dryad也会发类似链接来提醒出版商。

（3）数据保存

数据的长期保存对于数据库发展至关重要，若无完善的保存策略，易造成数据丢失，失去数据提供者对数据库的信赖。数据长期保存政策更是衡量数据库是否值得信赖的指标。国际空间数据系统咨询委员会（Consultative Committee for Space Data Systems，CCSDS）在2011年发布的审计和认证可值得信赖的数据库文件中指出，数据库应具有适当的继任计划、应急预案和到位的代管安排，以防数据库停止运行或资助机构、管理部门在其范围内大幅度地变更[10]。虽然过去几年制定数据库数据保存政策越来越普遍，但由于许多小型的科学数据库是服务于自身需求的，很少制定数据的保存政策。总体来说，这样的数据库是落后于社会科学数据库的。社会科学数据库往往有更稳健的政策，如美国校际社会科学数据共享联盟（Inter−University Consortium for Political and Social Research，简称ICPSR）就有完备的数字保存政策和计划。它还有一个政策模型，任何组织可以用它来开发自己的数字保存政策框架[11]。

Dryad作为大型且稳健的数据库，有长期保存政策。在数据备份方面，Dryad备份发表和未发表的内容，内容被备份到独立的远程服务器进行长期存储。当服务器出现故障时，如数据没有及时备份到远程服务器，可能会导致数据的丢失。在数据存储位置方面，当Dryad判断其保存的数据被访问的可能性提高时，Dryad将自行对文件进行迁移，添加到已发布数据包中。它不能保证迁移的文件会和原始文件一样，但同时会尽量减少由文件迁移而导致原有文件中数据缺失。数据维护和故障方面，Dryad是以参与出版商的身份与CLOCKSS进行合作，如果Dryad数据库出现故障不能维持其服务，在Dryad上注册过的数据将会被更新，以使用备份在CLOCKSS的数据。使得在相同的条款下，用户可以继续获取数据。Dryad可自行决定从CLOCKSS更换到另一个故障切换服务的供应商。

（4）收费和使用

Dryad数据库是非营利性机构，经费最初源自基金资助。为维持数据库管理和保存数据的基本成本和数据库的长期发展，Dryad数据库于2013年制定数据发布收费政策（DPCs），对数据提交者收费，实现非盈利性和数据公开收费融合。目前，Dryad数据库收费政策分3个方面：一是免费，研究者、教育工作者、学生可免费下载和使用数据；二是基本收费，除豁免者外，对内容被接受且符合内容标准的数据提交者实行收费。豁免者是按世界银行划分的低收入或中低收入国家的研究者；三是额外收费，对文件容量超过10GB或使用外部传输服务将外部大文件转移到Dryad中

或论文和数据再次提交的情况要额外收费。

使用政策是为了规范和明确用户使用的行为和条件，Dryad数据库使用政策总体分为两点：一是允许并鼓励用户在合法合理的情况下以任何方式对数据进行使用。合理合法包括不侵犯法律或损害Dryad对他人的服务以及损害Dryad的安全性。二是希望数据使用者遵守数据引用规范，给予数据提交者认同。数据引用规范最理想的格式是要有完全解析的DOI的URL。

4 Dryad与GenBank、Figshare数据库共享政策的比较

随着世界各地科研能力的不断增强，科研产生的数据越来越多，科研数据管理难度随之增加。为解决科研数据归档位置分散问题，研究者建立数据库的目录或注册系统来帮助人们认识和查找在线信息库信息，如OAD、re3data.org、Databib等[12]。re3data.org是全球研究数据库的注册系统，截至2015年5月6日，在该系统中注册的数据库数量为1223个，其中开放共享的数据库有1029个[13]。科研人员需求不同，数据存放地点有所差别，有些因系统化、专业化服务能力需求而存放在学科数据库，有些因开放性、共享性需求而存放在公共数据库，见图1。总体来说，Dryad与GenBank、Figshare数据库在国际上使用频率较高，且运作方式类似，故选择这3个数据库比较他们数据共享政策的异同点，见表1。

GenBank是美国国立卫生研究院基因序列数据库，面向生物信息学领域，是所有公开可用的DNA序列集合数据[15]。Figshare是一个让研究者所有科研成果可被引用、被分享、被发现的数据库[16]，面向学科领域广泛。Dryad与GenBank、Figershare数据库在国际上使用频率较高，且运作方式类似，对它们的数据共享政策的比较结果见表1。

Dryad与GenBank、Figershare数据库相类似，不同之处在于，其面向的学科领域既不像GenBank那样只面向生物信息学，也不像Figshare那样面向领域很宽泛。在数据类型方面，GenBank需要特定的上传格式，Dryad和Figshare收集各种类型的数据。总体来说，与Figshare较为类似，但其特色在于与采取JDAP策略的期刊出版商合作，重点收集论文中数据，有明确定位。且Dryad数据库对于数据的收集、出版、保存、收费和使用有明确细致的条款。

5 启示

Dryad数据库在出版物相关数据共享方面具有良好的实践性。作者在提交论文时，有些出版商要求将数据提交到Dryad中。Dryad数据库实现了论文与科学数据的互联[17]。在我国，一些科研机构也已制定相应的数据共享与管理政策，如：中国科学院资源环境科学数据中心制定并实施《中国科学院资源环境科学数据中心数据共享

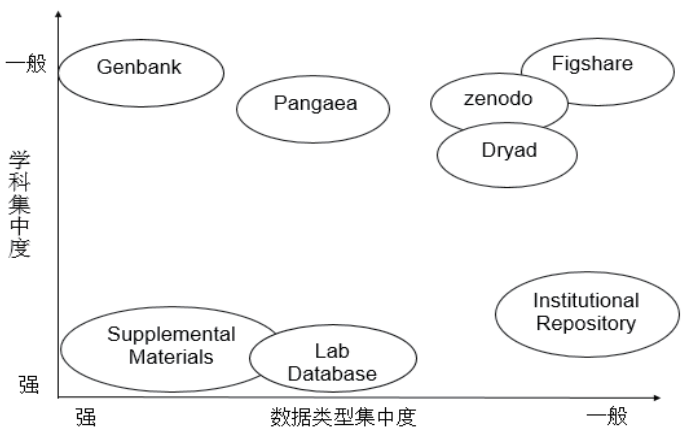

图1 数据归档位置[14]

表1　三个开放数据库比较

名称	Genbank	Dryad	Figshare
成立年份	1983年	2008年	2012年
国家	美国	美国	英国
领域	生物信息学	生态学、医学、环境科学、自然科学	生命科学、工程科学、自然科学和社会科学
数据唯一标识符	序列标识	DOI标识	DOI标识
数据类型	有限	任何	任何
收集政策	收集公开可用的DNA序列数据	收集公开的研究数据，重点是期刊论文中的数据	收集公开的任何数据
使用政策	开放获取，有限重用和存储	开放获取、重用与存储	开放获取、重用与存储
费用政策	免费上传、使用、下载	下载和使用免费，数据上传实行收费	上传和访问等基础服务免费，若要增加私人存储空间、扩充上传文件容量等服务需要收费
长期保存政策	自己负责	与CLOCKSS合作	与CLOCKSS合作

管理暂行条例》[18]，也有类似Dryad的数据库，只是表现形式不一样，如中国科学院数据云、数据堂等。但我国目前普遍存在的是机构数据库、学科数据库以及少数学科领域之间的数据库，不利于数据的共享与利用，可查阅的共享数据偏少且不全面。通过对Dryad数据共享政策的分析，我们得到以下两个方面的启示。

（1）制定细致、可操作的数据共享管理政策。我国针对气象、地震、水利、农业等领域建立各自的数据库平台，如中国气象科学数据共享服务网、国家农业科学数据共享中心等。针对多学科领域间的数据库包括中国科学院数据云、数据堂等。然而，由于我国数据政策整体缺乏国家层面的立法保障和宏观指导、导致科研机构只是制定项目层面的数据管理与共享细则，很难做到学科领域内或多学科的数据管理与共享[19]。大部分可查阅数据库数据共享政策只是对数据共享做了原则性规定，具体的操作细节还有待增强。借鉴Dryad数据共享政策，我国大型数据共享平台包括机构数据库、学科数据库、项目数据库等均可从收集、发布、长期保存以及收费和使用等角度建立或完善可操作的数据共享管理政策。如数据堂虽提供数据质量管理和数据安全管理说明，但可从数据收集范围以及类型、数据发布原则、长期保存策略、收费和使用条款等角度对该数据库数据共享政策进行说明，以进一步促进科研数据共享。

（2）出台相关政策，加强期刊出版商与数据库合作。许多主流期刊认可和采纳了Dryad提出的联合数据存档策略（JDAP），并将Dryad作为存放论文数据的公共库，促进科学数据在研究者间的共享。我国对于论文所涉及的科学数据并没有明确要求存储到合适的数据库中，部分原因在于我国期刊出版商和科研人员对科学数据共享意识不够强烈以及缺乏相应数据共享政策。我国期刊出版商可借鉴国外期刊出版商措施，出台相关数据共享政策，要求将支持论文观点和结论的数据或附加数据存放到类似Dryad等公共数据库中，便于科学数据集中管理以及实现科学数据共享。

6　结语

本文通过数据收集、数据发布、数据保存、收费和使用角度分析发现Dryad数据库有良好的数据共享政策。在数据收集方面，Dryad数据库数据来源广泛，主要收录合作期刊数据；在数据发布方面，数据发布需符合一定内容标准，在默认情况下数据与论文同时发表；在数据保存方面，Dryad数据库会将内容备份，且与CLOCKSS合作，以保证出现故障时也可提供服务；在收费和使用方面，实行非盈利性和数据公开收费相结合政策。研究Dryad数据库对我国数据共享政策

（下转第94页）

16(4): 471−479.

[42] Hughes B, Joshi I, Lemonde H, et al. Junior Physician’s Use of Web 2.0 for Information Seeking and Medical Education: A Qualitative Study[J]. International Journal of Medical Informatics, 2009, 78(10): 645−655.

[43] Paisley W, Hardy A, Fife M, et al. Information and Work: Research on the Improvement of Practitioner Information Systems[R]. Stanford University, Institute for Communication Research,1980.

[44] Casebeer L, Bennett N, Kristofco R, et al. Physician Internet Medical Information Seeking and On−Line Continuing Education Use Patterns[J]. Journal of Continuing Education in the Health Professions, 2002, 22(1): 33−42.

[45] Von Muhlen M, Ohno−Machado L. Reviewing Social Media Use by Clinicians[J]. Journal of the American Medical Informatics Association,2012,19(5): 777−781.

[46] 蒋佳文.医学用户网络信息检索行为的调查研究[J].医学信息学杂志,2006, 27(3): 178−180.

（上接第52页）

的启示在于我国机构数据库、学科数据库虽然很多，也制定了相应的数据管理政策，但细化以及可操作性不足，同时要加强与期刊出版商的合作。目前，较少有研究以Dryad数据库为例对国外开放数据的数据共享政策进行研究，本文则从4个角度对其共享政策进行分析，并与Genbank、Figshare进行对比，从而了解国外开放数据库的数据共享政策。但对于国外开放数据库的数据共享政策如何制定和实施，以及如何开展与期刊出版商的合作来促进数据共享还有待进一步探究。

参考文献

[1] OECD Principles and Guidelines for Access to Research Data from Public Funding[EB/OL]. [2015−10−10]. http://www.oecd.org/sti/sci−tech/38500813.pdf.

[2] PLOS One[EB/OL]. [2015−10−10]. http://www.plosone.org/static/policies.action.

[3] Nature[EB/OL]. [2015−10−10]. http://www.nature.com/authors/policies/availability.html.

[4] Mannheimer S, Yoon A, Greenberg J, et al. A Balancing Act: The Ideal and the Realistic in Developing Dryad’s Preservation Policy [J]. First Monday, 2014, 19(8). DOI: http://dx.doi.org/10.5210/fm.v20i10.5401.

[5] The organization – Dryad [EB/OL].[2015−05−05]. http://datadryad.org/pages/organization.

[6] Dryad Digital Repository [EB/OL]. [2015−05−05]. http://datadryad.org/.

[7] JM Ray. Research Data Management: Practical Strategies for Information Professionals [M].West Lafayette: Purdue University Press, 2014: 27−28.

[8] Dryad Collection Policy [EB/OL].[2015−05−05]. http://dryad.googlecode.com/svn−history/r3005/trunk/dryad/dspace/modules/xmlui/src/main/webapp/themes/Dryad/pages/collectionPolicy.html.

[9] Policies − Dryad [EB/OL].[2015−05−06]. https://datadryad.org/pages/policies.

[10] Audit and Certification of Trustworthy Digital Repositories [EB/OL].[2015−05−06].http://public.ccsds.org/publications/archive/652x0m1.pdf.

[11] Digital Preservation Policies and Planning at ICPSR [EB/OL].[2015−05−06]. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/index.html.

[12] 黄永文,张建勇,黄金霞,等.国外开放科学数据研究综述[J].现代图书情报技术,2013(5):21−27.

[13] Re3data[EB/OL].[2015−05−06]. http://www.re3data.org/.

[14] A Day in the Life of a Dryad Curator[EB/OL].[2015−05−06].http://www.ils.unc.edu/digccurr/curategar2015−talks/hull.pdf.

[15] Genbank[EB/OL].2015−05−07]. http://www.ncbi.nlm.nih.gov/genbank.

[16] Figshare[EB/OL].[2015−05−07]. http://figshare.com/about.

[17] 黄如花,邱春艳.国外科学数据共享研究综述[J].情报资料工作,2013(4):24−30.

[18] 中国科学院资源环境科学数据中心数据共享管理暂行条例[EB/OL].[2015−10−13]. http://www.lreis.ac.cn/sc/subsite/details.aspx?cid=212&v=7.

[19] 朱艳华,胡良霖,袁雅琴,等.国内外科研资助机构科学数据共享政策分析[J].中国科技资源导刊, 2015, 47(3):50−57.