公共科学图书馆·综合期刊的数据共享措施

吴 思 赵 伟 屈宝强
（中国科学技术信息研究所，北京 100038）

摘要：学术期刊作为科学数据共享的重要媒介之一，其数据共享相关的政策极大地促进了科学数据的开放访问。为全面了解学术期刊科学数据存档与数据共享相关的政策和规定，文章以公共科学图书馆·综合期刊为研究对象，从数据的提交、存储、访问到利用及数据的隐私保护等方面的数据共享政策进行详细分析。通过对其数据政策及共享措施的探索，为我国学术期刊制定相关的数据政策及推进科学数据共享实践提供有效借鉴和启示。

关键词：学术期刊；科学数据；PLOS ONE；数据共享政策；数据共享措施
中图分类号：G350 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2015.06.010

Data Sharing Measures of PLOS ONE

Wu Si, Zhao Wei, Qu Baoqiang
(Institute of Scientific & Technical Information of China, Beijing 100038)

Abstract: Academic journals, as one of the important medias, it is their relevant policies on data sharing that promote the open access to scientific data. To fully understand the relevant regulations and policies on data archiving and data sharing of academic journals, take the example of PLOS ONE, the paper analyzed it's data sharing policies in detail from submission, storage, access, to utilization and privacy protection. Through the exploration of data sharing policies and measures, to provide effective references and inspirations for making related data policy for our academic journals and promoting advance of scientific data sharing practices.

Keywords: academic journals, scientific data, PLOS ONE, data sharing policies, data sharing measures

1. 引言

20世纪中期以来，科学数据管理与共享逐渐引起科学界与学术界的关注。当今，科学研究正进入第四范式——数据密集型的科学发现，科学数据的管理与共享在科学发现与创新中的作用尤为凸显。科学数据共享已经成为全球各国政府和国际组织的共识，全面公开地获取数据不但能使其在应用过程中增值，也是推动科学认识突破的重要条件，而数据政策是数据管理和共享活动的基本保障[1]。期刊是国内外学术交流和传播的主要媒介，良好的科学数据管理可以保证期刊的持续发展，促进科学数据的开放访问。目前，国外诸多期刊已制定了相关的数据共享政策。例如：《Science》[1]的编辑和出版政策包括数据库存储政策、许可协议、再版和访问政策等；《Nature》[2]

作者简介：吴思*（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、公共科研机构信息公开；**赵伟**（1975-），女，中国科学技术信息研究所副研究员，研究方向：科技资源管理；**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。

的编辑政策包括生物伦理、数据、材料和方法的共享及数据开放时间（Embargo）等相关数据共享政策；BMC[3]（BioMed Central)编辑政策包括数据及材料的发布、重复发布及引用等相关数据共享政策；公共科学图书馆·综合（Public Library of Science，简称PLOS ONE[4]，是一个由科学家和医生组成的非营利机构，致力于把世界上科学和医学的文献作为免费资源向公众开放）的数据编辑和出版政策包括数据从提交、存储、访问到利用等过程中的数据共享政策。这些期刊的数据共享规定和政策的制定极大地促进了科研数据的开放访问，使科学家可以对公布的调查结果进行验证，探索新的分析方法。我国对学术期刊科研数据的管理与共享起步较晚，且期刊数据共享政策相关的研究较少。唐义[5]等探究了国际科学数据共享政策法规体系，其中包括了期刊的数据共享政策；吴蓉[6]等对《Science》、《Nature》及PNAS（Proceedings of the National Academy of Sciences of the United States of America，美国科学院院报）等国外学术期刊的数据共享政策进行了探析。然而，PLOS ONE作为为数不多进行开放共享并较早提出科学数据共享政策的学术期刊之一，其科学数据共享政策影响较广、实践较早且数据共享较为规范。可是，关于PLOS ONE的研究多为针对其开放存取的出版模式的探讨，缺乏对其数据共享政策的研究。本文将选取PLOS ONE作为研究对象，对其数据共享相关政策进行调查分析，将详细地描述其完整的数据生命周期各阶段相关的数据共享，以期为我国科技期刊数据共享政策的制定提供参考。

2　PLOS ONE的数据共享政策分析

PLOS ONE数据共享政策的要求是：作者在发表论文的同时公开相关的原始数据，将数据存储到公共数据库或以论文附件形式提交，实现论文与数据的实时关联和集成出版。作者提交论文时，必须同时提供一份数据可用性声明，表示符合期刊的数据共享政策。如果文章一旦被期刊接受，那么该数据可用性声明将作为论文的一部分同时发表。鉴于PLOS ONE数据共享政策的要求，本文拟对其数据提交、存储、使用和保护等做进一步的探讨。

2.1　提交数据的类型、方式及规定

PLOS ONE将提交的数据类型定义为“最小数据集”，主要包括形成论文结论的相关元数据和方法及其重现完整报告研究结果所需要的其他数据。无论是否提交“最小数据集”，论文的主体部分都必须包括核心描述数据、方法和研究结果等内容。这样不仅保证了数据的可重用性，而且最大程度地保护了作者的权益。

PLOS ONE接受的数据提交方式主要有两种：一是存储在合适的公共数据库，二是以论文附件形式提交到网站服务器。将数据作为论文的主要内容（附件）可以更加便于期刊审核者、同行评议者及读者的访问和搜索；将数据集存储在公共数据库中可以优化其存储及检索。因此，作者需要考虑选择以下哪种方式上传数据。

第一种是以论文附件的形式上传数据。这种方式主要是针对小的数据集和数据类型确定的数据。以下从文件类型、格式及规模3个方面加以说明。（1）文件类型：附件是论文主要内容的补充，PLOS ONE把这些文件存储在自己的网站服务器上，因此可以接受并处理多种类型的文件，包括图像、表格、音频、视频、软件及数据库等。（2）文件格式：作者在提交附件时需选择最佳文件格式，使其可用性和重用性最大化，如EXCEL比PDF更适合于提交表格数据。（3）文件规模：为了方便读者获取，文件大小最好控制在10MB以内。若文件过大，可以压缩文件、转换格式或将文件打包成压缩文件夹后再上传。

第二种是将数据存储到公共数据库。这种方式是将论文基础结论中的所有相关的数据及元数据存储在一个合适的公共数据库中。PLOS ONE没有指定的数据库，作者需要根据论文的学科标准和数据类型选择合适的数据库，可以将数据存储在特定学科及特定数据类型的专有数据库或多种数据类型的通用数据库。如将微阵列数据提交到ArrayExpress或GEO（Gene Expression

Omnibus），将基因序列数据提交到GenBank、ENA（EMBL Nucleotide Sequence Database）或DDBJ（DNA DataBank of Japan），将生态学数据提交到Dryad等。PLOS ONE推荐的存储特定数据类的具体数据库如表1所示。此外，作者必须在提交的数据可用性声明中详细说明数据已经公开存储，并列出数据库的名称、DOI（数字对象标识符）及相关数据集的登录号。作者在论文尚未被期刊接受无法提供DOI或者登录号时，需在论文被接受之后提供这些信息。

以“软件”类型数据的提交为例，PLOS ONE支持开源软件的发展，所提交的论文如果是对一种新方法、软件或数据库的描述，并符合实用性、可验证性及可用性的标准，那么期刊将予以发表。该“软件”必须满足以下要求：“软件”基于开源标准；“软件”符合开源定义；“软件”存储在一个开放的软件数据库；“软件”包括在附件中并与论文一起上传；论文中提供了“软件”的链接。尽管期刊优先考虑完全的开源软件，但并不排除接受Mathematica和MATLAB等商业软件。当将软件存储到开放的软件数据库时必须提交以下信息。

（1）软件相关的源代码。应尽可能遵循公认的社会标准和适当的许可协议，如BSD（Berkeley Software Distribution）协议、LGPL（Lesser General Public License—GNU宽通用公共许可证）或MIT（开源软件许可协议）。而且，代码应该很容易被找到和下载，而无需创建用户账号登录或输入其他个人信息。

（2）软件安装和运行的文档。作为终端用户的应用程序，必须有指导安装和使用软件的说明；对于软件数据库还应有使用应用程序接口的说明。

（3）有相关控制参数集的测试数据集。标准测试集的测试结果应尽可能地包括在数据集内；待测试数据之间应尽可能地相互独立而不应该有任何依赖关系。

此外，存储软件的公共数据库必须建成5年以上，且存放1000多个软件项目，例如：SourceForge、Bioinformatics.Org、Open Bioin-

表1　PLOS ONE推荐的存储特定数据类型的数据库

数据类型	PLOS ONE推荐的数据库
非结构化数据库/大型数据库	Dryad、figshare、GigaDB、Havard Dataverse Network、Zenodo等
序列分析、序列测定	dbSNP、DDBJ、ENA、GenBank、SRA(NCBI Sequence Read Archive)等
组学	ArrayExpress、DIP(Database of Interacting Proteins)、EGA(The European Genome−phenome Archive)、GEO、PRIDE(Proteomics Identifications)等
结构数据库	BMRB(Biological Magnetic Resonance Data Bank)、COD(Crystallography Open Database)、EMDB(Electron Microscopy Data Bank)、PCDDB(Protein Circular Dichroism Data Bank)、FlowRepository等
生物模型	BioGRID(Biological General Repository for Interaction Datasets)、EuPathDB(Eukaryotic Pathogen Database Resources)、FLYBase、MGI(Mouse Genome Informatics)、RGD(Rat Genome Database)等
分类与物种多样性	ITIS(Integrated Taxonomic Information System)、GBIF(Global Biodiversity Information Facility)、NCBI Taxonomy、The Knowledge Network for Biocomplexity等
生物医学科学	Influenza Research Database、NAHDAP(National Addiction & HIV Data Archive Program)、NDAR(National Database for Autism Research)、Cancer Imagine Archive、PhysioBank等
生物化学	caNanolab、KiMoSys(Kinetic Models of Biological Systems)、PubChem等
自然科学	AADC(Australian Antarctic Data Centre)、CARD(Cold and Arid Regions Science Data Center)、NCDC(National Climatic Data Center)、NERC(National Environmental Research Council Data Centres)、WDCC(World Data Center for Climate at DRKZ)等
社会科学	ICPSR(Inter−university Consortium for Political and Social Research)、Qualitative Data Repository等

formatics Foundation、Google Code、Savannah、GitHub及Codehaus等开放的软件数据库。

2.2　数据存储的位置与费用

PLOS ONE存储提交数据的位置主要分为两种：一是将以论文附件形式上传的较小数据集存储在自己的网站服务器上，读者可以通过附件链接准确获取数据信息。二是将数据集及相关的元数据等存储在公共的第三方数据库上，其中包括一些PLOS ONE集成合作伙伴计划成员数据库。当前数据库集成合作伙伴有figshare[7]（保存和共享科研产出的在线数字资源库）、FlowRepository[8]（流式细胞仪实验的公共数据库）、GenBank[9]（DNA序列数据库）及Dryad[10]（致力于科学出版物基础数据的发现、重用及引用，主要存储基础科学及其医学出版物中的多种类型的研究数据）等。PLOS ONE并未指定数据库，但鼓励作者在考虑数据学科标准和数据类型的同时，选择达到接受标准的、可信赖的、比小型本地数据库更可能长久维护和存在的大型国际数据库。

以PLOS ONE集成合作伙伴计划的首位成员Dryad为例：作者将数据集上传到数据库之后，Dryad会提供一个数据集的数字对象唯一标识符（DOI）及一个私有的用于审核者获取数据的URL。作者在提交数据可用性声明时提供数据库的名称、DOI及相关数据集的登录号等，并提供URL密码以便于审核者获取作者提交的数据集。PLOS ONE与Dryad等公共数据库合作，集成论文的存储，以确保文章及其基础数据发表在一起，联系在一起。数据的集成有助于文章和数据的存储，也促进了论文的同行评议。PLOS ONE正在扩大当前的合作伙伴，将集成更多的数据存储中心。

PLOS ONE的自存储是免费的，而数据的第三方存储需要作者承担数据的存储费用。存储费用没有统一的标准，因具体的数据库而异。下面以PLOS ONE的集成合作伙伴计划成员数据库Dryad为例加以说明。由于Dryad的原则是使存储的内容能免费用于科研和教育，个人用户和机构用户没有任何的访问费用。在Dryad中存储数据时，要求数据提交者承担出版费用，除非相关期刊或其他组织已经承担了数据出版费用，或提交者所在的国家被世界银行划分为低收入和中等收入经济国家类别。如果投稿期刊没有付费计划或赞助没有到位，作者需承担数据出版费用。数据库集成合作伙伴期刊的作者需支付80美元存储10GB以内的数据（非数据库集成合作伙伴期刊的作者需支付90美元）；若超过10GB，除超出的第一个10GB需支付15美元外，之后每超出10GB需支付10美元。具体收费标准如表2所示。不同的数据库收费机制不一样，如figshare的存储费用不仅与存储文件的大小有关，而且与私人存储空间、合作者人数等有关。

表2　Dryad存储费用

数据规模X/G	费用/$
X≤10	80
10≤X≤20	95
X>20	95+(X−20)

2.3　数据的访问及使用

作者提交论文的同时应将数据提交到公共数据存储库，论文被期刊接受并发表时，数据库中存储的数据集自然公开，读者可以免费下载和重用。而在论文发表之前，仅同行评议者及期刊审核者可以根据作者提交数据时提供的URL获取数据。此外，PLOS ONE的文章在美国东部时间的下午2点之前是禁止访问的。

学术期刊的数据使用的Creative Commons（知识共享）协议主要有4种：CC0（放弃一切权利，直接让作品进入公有领域）、CC−BY（Creative Commons Attribution License，知识共享署名许可）、CC−BY−NC（Creative Commons Attribution−NonCommercial License，知识共享署名—非商业用途许可）、CC−BY−NC−ND（Creative Commons Attribution−NonCommercial−No Derivative Work−sLicense，知识共享署名—非商业用途：禁止演绎许可）。PLOS ONE数据政策并未明确数据库中数据的版权，但指出数据库中的数据使用许可协议

不应比CC-BY更严格。CC-BY允许任何用户下载、打印、摘录、重用、归档和发布信息，只需正确引用作者及文献来源，确保数据尽可能广泛地被传播和利用。

PLOS ONE要求论文结论中相关的基础数据不受限制的完全公开，除了少数情况例外（详见数据隐私保护与公开豁免一节内容）。期刊不接受出于个人利益的原因（如专利及未来潜在的出版物等）限制数据的公开，且拒绝接受研究结论取决于专有数据（如商业数据等）的论文，除非同时提交一份基于公共数据的分析，以便验证和重现研究结论。如果论文出版后存在作者访问数据受限的情况，期刊有权联系作者的机构和资助者并要求其修正，或者在极端情况下撤销论文的出版。

针对数据存储及附件数据因涉及道德、法律等问题无法公开数据的情况，PLOS ONE提供了两种可选的方法进行评估。

一是依申请公开。当数据涉及道德、法律问题时，作者可以基于研究者的合理请求向其公开数据。在这种情况下，数据可用性声明中必须详细说明“数据基于请求可用”，并明确研究者应将请求提交给的团体，如数据访问委员会或道德规范委员会。同时，作者必须在数据可用声明中详细说明限制数据公开存储的原因，并且不能接受论文作者作为唯一确保数据能被免费获取的责任人。

二是正确引用第三方数据。若论文中的主数据集并非由作者自己生成，PLOS ONE则要求有需求的研究人员可以从指定的原始来源中独立地获取第三方数据。在这种情况下，作者提交的数据可用性声明必须说明完整的引文数据来源，同时也必须详细说明限制数据公共存储的原因。

2.4　数据隐私保护与公开豁免

隐私保护是科学数据共享中极重要的环节，有利于保护利益相关者的权利，规范并促进数据共享的有效实施，更好地发挥数据共享的价值。PLOS ONE坚决不能侵犯病人的隐私，并鼓励研究者遵守已有的指导及当地适用的法律。例如：美国NIH的Protecting the Rights and Privacy of Human Subjects、英国Data Archive的Anonymisation Overview及澳大利亚National Data Service的Ethics Consent and Data Sharing等。针对人类受试者研究的临床数据、临床试验等病人隐私保护，PLOS ONE提出了明确的要求：作者必须避免公开能识别病人身份的信息，除非有严格的提交要求。当有严格的提交要求时，作者必须确保病人签署了《PLOS ONE期刊出版同意书》。PLOS ONE详细地列出了禁止公开的个体患者信息，如表3所示。

表3　禁止公开及不宜公开的个体患者信息

	序号	患者信息
禁止公开	1	姓名，姓名缩写、地址、邮编或部分邮编
	2	电话或传真等联系方式、邮箱
	3	独特识别号码、车牌号、医疗设备识别码
	4	网络或互联网协议地址
	5	生物特征数据、面部照片或类似的图像、录音带
	6	亲属姓名
	7	与个人相关的日期，包括出生日期
不宜公开	1	治疗或保健专业负责人
	2	性别
	3	罕见病或治疗
	4	敏感数据，如职业、工作地点、收入或家庭教育
	5	家庭成员
	6	人体测量
	7	妊娠指数
	8	种族
	9	出生年份或年龄
	10	逐字记录或反应

PLOS ONE指出，涉及不能公开数据存储数据库及论文附件中的数据主要有以下3种：一是涉及道德或法律问题的数据不能公开，例如可能侵犯病人隐私的数据；二是可能带来其他威胁的数据不能公开，如化石沉积的位置、濒临灭绝的物种信息；三是从第三方获取的数据不能直接公开，如作者并未生成自己的主数据集，而是引用他人的数据。

3　启示

目前，国内数据共享活动尚处于初级阶段，

学术期刊数据共享相关的政策急需加强指导和行为规范。现受到PLOS ONE期刊的数据共享措施的启发，对我国学术期刊探索科数据的开放、访问和获取的可行方式提出以下建议。

（1）数据提交的格式要求及方式

国内学术期刊可以建议或者要求作者在投稿的同时提交论文结论涉及的基础数据。提交数据的类型时可以重点考虑提高数据的可重用性和避免数据的过度使用，可参照PLOS ONE的“最小数据集”作出格式要求。数据的提交可以采取如下3种方式：一是将提交数据作为论文发表的前提条件，并要求作者将数据存放在合适的数据库中；二是以论文附件的形式提交数据，读者通过附件链接容易的获取数据；三是论文数据依申请公开，即有需求的研究者向作者提出获取数据的合理请求。通过上述方式实现数据和文献之间的整合，公共存储的数据和文献之间可链接和可用。

（2）数据著作权归属及许可协议

目前，PLOS ONE的数据共享政策中对于公共数据库中数据的版权许可尚不明确。如果国内学术期刊建议将数据存放到第三方数据库，为了避免由于数据版权不明带来的问题，在制定相关政策时可以要求作者提供数据权益声明，包括数据产生的方式、数据著作权归属、数据使用的许可协议等。数据的著作权归属可以根据科研资助机构的项目要求以及科研教育机构的职务作品要求、科学伦理与道德的约束规范等确定。作者应在提交论文及其数据时明确署名，并允许期刊在使用许可的条件下对数据进行传播利用[6]。针对学术期刊使用的4种数据许可协议，一般情况下，将CC-BY作为默认的授权许可协议，在保护作者署名权的同时，允许读者及社会大众、科研人员重用，包括数据挖掘及文本挖掘、允许期刊及第三方团体对数据进行商业利用等。

（3）数据的隐私保护及公开豁免

我国学术期刊在制定数据共享政策时，应注重隐私保护及公开豁免。对涉及法律、隐私以及来自第三方等数据限制其开放，并详细说明理由。对一些限制公开的数据，在不违背隐私保护的前提下，给出相应的解决方案。例如：对数据开放的对象划分权限，对提出合理使用申请的科研人员经身份及申请理由鉴定通过的情况下，可由作者向其提供，而一般读者则不对其开放。相关政策和规定可参照国外一些优秀期刊制定。

良好的科学数据共享政策，不仅促进了数据的共享，而且在一定程度上节省了共享成本。我国学术期刊也应该积极参与到科研数据开放共享活动中，借鉴国外的有益经验，并结合我国学术期刊的实际情况，制定具有我国特色的学术期刊数据共享政策。

参考文献

[1] 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.

[2] Science[EB/OL].[2015-06-28].http://www.sciencemag.org/site/help/authors/policies.xhtml.

[3] Nature[EB/OL].[2015-06-28].http://www.nature.com/authors/policies/availability.html.

[4] BMCEvolutionary Biology[EB/OL].[2015-06-28]. http://www.biomedcentral.com/bmcevolbiol/about.

[5] PLOS ONE[EB/OL].[2015-06-28].http://www.plosone.org/static/policies.action.

[6] 唐义,张晓蒙,郑燃,等.国际科学数据共享政策法规体系: Linked Science制度基础[J].图书情报知识, 2013 (3):67-73.

[7] 吴蓉,顾立平,刘晶晶. 国外学术期刊数据政策的调研与分析[J]. 图书情报工作,2015,59(7):99-105.

[8] figshare[EB/OL]. [2015-06-28]. http://figshare.com/about.

[9] flowrepository[EB/OL]. [2015-06-28]. https://flowrepository.org/.

[10] GenBank[EB/OL]. [2015-06-28]. http://www.ncbi.nlm.nih.gov/genbank.

[11] Dryad[EB/OL]. [2015-06-28]. https://datadryad.org/pages/faq.