数据期刊中科学数据的同行评议方法研究*

孔丽华[1,2,3] 习 妍[3] 郎杨琴[3] 汪 洋[3]

1)中国科学院文献情报中心,100190;2)中国科学院大学经济与管理学院图书情报与档案管理系,100049;
3)中国科学院计算机网络信息中心,100190;北京

摘 要 数据是科学假设、科学分析以及科学理论形成的基础证据,是同行科学家数据评估和检测科研结果的重要证据。数据出版是数据开放共享的重要手段之一。在数据出版的过程中,对数据的评审是重要的环节之一。文章通过对现有数据出版平台中数据评审要求进行调研,并基于FAIR数据共享原则,旨在提出一种数据期刊中科学数据的同行评议指标体系,从而提升数据质量审核,促进科学数据的可发现(findable)、可访问(accessible)、可理解(intelligible)、可重用(reusable),进而推进科学数据的开放、共享和引用。

关键词 数据出版;数据期刊;数据论文;同行评议

Study of peer review criteria for scientific data in data journals//KONG Lihua, XI Yan, LANG Yangqin, WANG Yang

Abstract Data is the basis for scientific hypothesis and analysis, as well as the formation of scientific theories. It provides important evidence for scientists to evaluate and validate scientific findings of their peers. As an importance mode of data sharing, data publishing should in the first place be premised on data review. This study looks at the data review requirements of some major data publishing platforms, in order to come up with a peer review index system for data publishing based on the FAIR principles (findable, accessible, intelligible, reusable). In doing so, we aim to apply the FAIR principles in data quality review for seeking more effective means of data sharing and citation.

Keywords data publication; data journal; data paper; peer review

First-author's address National Science Library, Chinese Academy of Sciences, 100190, Beijing, China

DOI:10.16811/j.cnki.1001-4314.2019.03.007

在信息技术高度发展的今天,科学数据不仅仅只是科学研究和科技文献的重要产出,已成为科学研究和促进社会进步的重要内容、工具和新的科学基础设施,是科技界的"一等公民"[1]。数据是科学假设、科学分析以及科学理论形成的基础证据,是同行科学家数据评估和检测科研结果的重要证据。科技界对于高质量数据的共享需求越来越强烈。英国皇家学会在2012年发表的《科学是开放事业》[2]中提出"为保持科学的自我纠错能力、支持科研结果的可验证和新知识的发现,为支持社会利用科研成果进行创新和教育,必须实现科学数据的开放共享,需要科学数据可获取、可理解、可评估、可利用"。

1 出版界对数据出版的政策要求

对于数据的发布与共享,出版界采取了一系列举措,各大出版集团和期刊均积极研究制定数据政策(DataPolicy),在论文投稿时提交和发表时公布支撑数据,已成为越来越多学术期刊的基本要求。例如,《PLoS One》2008年发布了数据共享政策,并在2014年进一步修订提升相关要求:除特殊情况外,作者需共享关于文章内容的数据,如不提供,将会撤销文章的发表[3]。而Springer Nature按照对研究数据共享要求的不同等级制定了4种类型的数据共享政策[4],从鼓励不强制,到鼓励共享且须提供数据可用性声明,再到强制性要求共享数据,以及明确要求对数据进行评审等方面提出了不同层级的要求说明。该政策适用于Springer Nature旗下所有的期刊,目前大多数期刊会酌情采用其中一项政策。如《Photosynthesis Research》采用了Ⅰ型政策,而《Plant and Soil》采用了Ⅱ型政策,BioMed Central旗下期刊、《Palgrave Communications》和越来越多的Nature期刊都支持Ⅲ型政策,《Scientific Data》(SD)和《Genome Biology》则采用了具有最严格开放数据政策的第Ⅳ类政策要求。类似地,Elsevier、Wiley、Taylor & Francis等也发布了分级数据政策。

关于这部分内容已有很多研究[5],这里不做更多研究。相比而言,我国期刊在数据方面的政策要求起步较晚,但已经开始尝试这方面的工作,例如《现代图书情报技术》从2016年起实施"支撑数据提交计划"[6],要求作者在提交论文的时候,需要提交对其论文结果有重要支撑作用的数据。中华医学会部分期刊也对作者提交数据提出了新的要求。

2 数据出版的同行评审

出版已经被认为是保证出版物质量的重要过程。同行评议作为一种学术成果审查程序,为帮助出版而进行的审稿活动已有近300年历史。论文评审的同行评议最开始的雏形可以追溯到17世纪中叶,英国皇家

*中国科协科技期刊青年编辑业务研究择优支持项目(castqk2017-qnkt-06);中国科学院自然科学期刊编辑研究会2018年研究课题(YJH-2018008)

学会刊物《The Philosophical Transactions of the Royal Society》(Phil. Trans)创刊时期[7]。目前,数据出版的基本流程主要参考学术期刊出版流程[8],对数据的评审也成为数据出版过程的重要步骤之一。

2.1　数据出版的主要模式

科学数据的出版是一种全新的数据共享模式,是指通过"数据提交、同行审议、数据发布、数据永久存储、数据引用和数据影响评价"等基本环节发布高质量的数据,力求科学数据资源的最大化使用。根据数据的依托关系和主要存储形式,研究人员将科学数据出版划分为以下 3 种主要模式[9]。

1)独立数据出版,即数据直接在数据中心或数据知识库存储发布。如各国政府的数据中心(.gov),以及专业数据中心(如 NASA[10]),还有一些公共存储库(如 Dryad[11]、FigShare[12] 等)由中国科学院计算机网络信息中心研发支持的 ScienceDB 也提供相关服务。

2)作为论文辅助资料的数据出版,主要指发布论文的同时,按照期刊相关政策要求,将数据作为附件提交至期刊指定的位置(如期刊网站或其他指定的存储库),如《PLoS One》、Springer Nature 旗下期刊等。

3)以数据论文形式出版,即将数据通过数据论文的形式进行出版,如 SD、《Earth System Science Data》(ESSD)等。

不同数据出版模式下对数据的出版形式和质量评审要求也有所不同。独立数据出版主要关注于提供数据的长期保存和数据行业标准;作为论文辅助资料的数据出版则关注提交数据对论文研究内容的支持程度;以专业发表数据论文的数据期刊的形式,更关注于共享研究数据集的完整发表,这些数据期刊在"发表"数据之前都会进行较为严格的同行评议。因此,本文选取了对数据评审要求最高的第 3 种模式下的数据评审作为主要研究内容。

2.2　数据期刊的同行评议标准

以数据论文形式出版的数据出版就是数据生产者将之前为了某个目的采集和处理的数据整理后,编写成数据论文,将数据论文和数据同时提交至指定的地方,数据论文和数据通过同行评议后发表出来的过程。在这个过程中,数据论文和数据都会被分配一个唯一的永久标识符(如 DOI),并将两者相互关联起来。数据经过正式出版,使得其他使用者能便捷地发展、获取、理解和再分析利用,且可在科研论文及其他相关科研成果中引用。数据出版的基本流程如图 1 所示。

文献[13-14]的调查显示,几乎所有的数据期刊都采用了同行评议机制来保证质量,而同行评议在传统学术出版中的至高地位也使这一特征成为数据论文和数据期刊在质量控制方面优势的集中体现[15]。尽管如此,当前各期刊采用的评审方法和评估标准各不相同。我们针对数据期刊对同行评议标准的制定,调研了 6 种具有代表性的数据期刊 ESSD[16]、《Geoscience Data Journal》(GDJ)[17]、《Biodiversity Data Journal》(BDJ)[18]、SD[19]、《Brief in Data》(BD)[20]、《GigaScience》(GS)[21],对其在网站上公布的评审要点进行了统计,如表 1 所示。

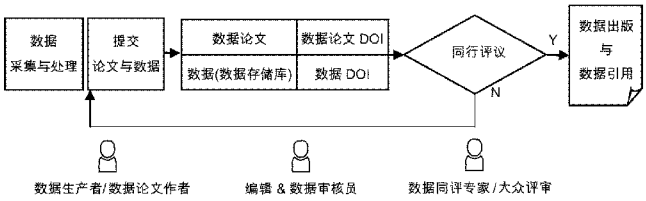

图 1　数据(论文)出版的一般流程

表 1　数据期刊的评审标准要点统计

	评审内容	ESSD	GDJ	BDJ	SD	BD	GS
论文	内容和格式审查	●	●	●	●	●	●
	对其他数据集和论文的引用完整有效性	●	●	●			
	符合出版平台其他相关要求						●
数据	可访问性	●	●	●	●	●	●
	数据格式与标准	●	●	●	●	●	●
	方法描述详尽程度	●	●		●	●	●
	数据完整性	●	●	●	●	●	●
	数据科学研究内容的完整性(数据丰度)	●	●	●	●		●
	质量控制或评估	●	●	●	●		●
	元数据质量	●			●		
	数据与论文、元数据的一致性	●		●	●		
	存储库的匹配度			●	●		●
	数据资料的完整性(工具和软件环境等)		●	●			●
	数据的科学重要性和独特性			●			
	数据处理方法的科学性和有效性		●		●		●
	数据重用性	●	●	●	●		
其他	背景信息(收集和分析数据的理由)						●
	知识产权和共享许可(包括对加工数据的原始数据的使用许可)	●	●	●			●
	遵循科学/医学研究的伦理标准(第三方协议引用)			●	●	●	●
	潜在用途		●		●		●
	对学科的贡献度		●				

注:需要说明的是,上述的调研是基于各刊在网站公布的评审要求统计而来,可能会存在因为描述不够详尽而没有明确说明,但在具体评审时还会考虑的内容,这些内容不体现在这里。

从调研可见,大多数数据期刊都延用了传统期刊的同行评议模式,制定了自己的标准,但审查数据和数据论文与审查传统学术论文有很大不同。数据期刊将

论文和数据集的同行评议整合到一个过程中,其评估对象包括数据论文、数据集以及两者间的关系。另外,数据集的质量评议对于评审人员是一个巨大的挑战,因为数据集本身带有复杂性,往往数量庞大,结构复杂,且与传统评审内容截然不同,所以在对数据和数据论文的同行评议中就会包含更多的问题,使过程变得更加复杂。目前对于数据论文尤其是数据本身的同行评议还处在一个初步阶段,且尚无通用的数据质量评判标准,针对数据的同行评议也没有公认的理解。

3 《中国科学数据》的数据质量评价指标设计

中国科学院计算机网络信息中心于2015年创办了数据期刊《中国科学数据(中英文网络版)》(本文简称《中国科学数据》),并在办刊过程中积极探索关于数据出版的各种问题。本选题基于该工作展开研究,目的是提出一种数据出版的数据质量评价指标体系,以提高同行评审过程中数据的可见性,改进编辑和同行评审服务,促进数据的开放和共享,实现数据的可获取、可理解、可评估以及可重用的高质量共享目标。

3.1 指导原则

指标的设计以FAIR“科学数据管理的指导原则”[22]为基本原则(表2)。该原则是一个用以促进开放科学的基础性原则,推动在可发现、可存储、可互操作、可重用的原则下,提高获取公共财政资助的研究成果的便利性,适用于各个领域包括数据论文的数据出版。

表2 FAIR指导原则[23]

基本原则	具体含义和说明
可发现性	F1(元)数据具有全球唯一并且持续稳定的标识符; F2 数据有丰富的元数据来描述; F3 元数据清晰、明确地包括所描述数据的标识符; F4(元)数据在可检索资源中得到注册或标引。
可获取性	A1(元)数据可以通过自身的标识符、采用标准通信协议被获取; A1.1 上述协议是开放的、免费的并且在全球都能普遍实施的; A1.2 上述协议在必要时允许身份验证和授权; A2 即使数据已经无法获得时,元数据也是可获得的。
互操作性	I1(元)数据在知识表达上使用正式的、可获得的、可共享的、并广泛应用的形式; I2(元)数据采用的词汇体系也遵循FAIR原则 I3(元)数据对其他高质量的(元)数据进行规范的引用。
可重用性	R1 数据通过大量准确且相关的属性予以丰富的描述; R1.1(元)数据发布时有清晰的、可获得的数据使用许可; R1.2(元)数据有详细的出处; R1.3(元)数据符合相关科研领域的标准。

在出版的各个环节中,不同角色对于数据的质量具有不同的责任和权利。数据作者主要负责收集、处理和分析数据,是数据质量的基础。在此过程中,作者还需严格遵守相关法律与伦理规范,并应遵守科研机构与资助机构的政策、数据标准也应符合学科社群与科研机构的标准。数据论文作者则负责根据研究内容整理数据和撰写(数据)论文,并按期刊要求,在符合有关法规、保护个人隐私和遵循研究项目的相关政策规定下将论文和数据提交至指定的地方,并同意将数据按照承诺的公开范围进行共享,为数据的访问和下载做好准备。在数据论文的描述方面,应尽量详尽,以保证第三方能够理解和重用相关数据。出版商为数据的出版提供平台,包括明确的数据出版政策,对论文和数据进行审核,实现论文与数据的关联出版,提供数据引用说明等。数据存储库为数据的长期保存提供平台和工具,与出版平台实现对接,提供数据访问服务,并监管数据。读者在下载和使用数据时,应尊重数据知识产权,合理引用和使用数据。

3.2 数据出版中数据质量评价指标设计

我们对数据出版中科学数据质量评价指标体系进行了分级设计,逐层深入和细化各项指标。

1)一级指标。依据前述指导原则,我们认为对于出版数据的质量问题评价维度主要有获取、评估、理解和重用4个方面。故定义本研究的数据质量一级评价指标为可获取、可评估、可理解和可重用。可获取性描述数据是否处于易发现、易获取的状态;可评估性描述数据或信息的可靠性能被判断或评估;可理解性描述数据是否易于被审查者(或读者、使用者)所理解;可重用性则关注数据是否处于他人能够使用的格式和环境,以便被再次用于不同的研究目的。

2)二级指标。数据质量问题产生于数据整个生命周期的各个环节。一个完整的数据生命周期包括信息获取、整合分析、加工处理、存储、共享发布与应用等多个阶段。我们在二级指标的设计中加入了部分数据生命周期中的质量元素,并考虑到数据重用性是数据出版的重点关注内容,以及专家在同行评议时的可行性,在一级指标的基础上扩充各指标参数内容,得到二级指标体系如图2所示。

其中,可访问性是指该数据(集)可以通过提供的访问地址或唯一标识符(如DOI)访问到。可获取性则是指数据(集)不仅可以访问到,还可以下载获取查看;如果数据不能下载或有限下载,则还需判断是否提供了权限说明。可信性是指数据是否准确、真实反映实际信息,包括数据来源信息;数据丰度则指从研究意义上而言,数据(集)是否覆盖了一定的(时序或空间)

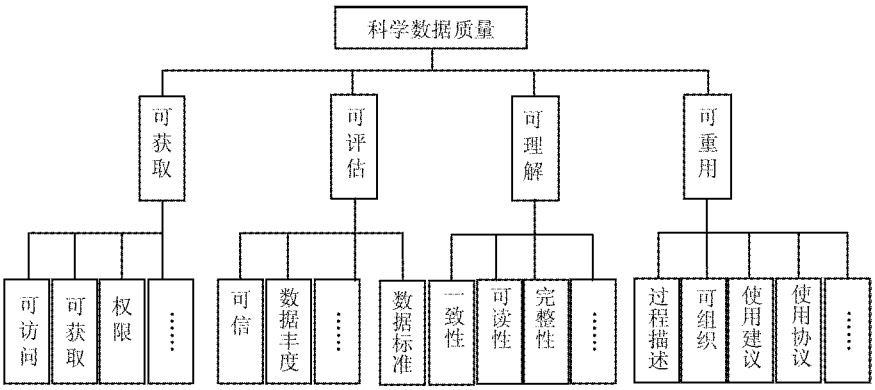

图 2　科学数据质量评价指标体系

范围,足以构成一个有意义的数据集,且数据是否符合相关领域数据标准的要求,达到标准性,而在数据处理过程中,是否对数据进行校验和质量控制,也是数据的质量判读的一个重要元素。一致性则表示实体数据及其属性是否具有一致的定义和含义,元数据信息以及论文描述数据与实体数据是否保持一致;完整性则主要考察该论文描述问题所需要的数据是否完备,且在数据格式方面描述清晰,以便他人阅读和获取相关数据结构和内容信息。而在数据重用方面,则重点考量对数据生产和处理过程的描述是否清晰明了,如该数据在特殊的软硬件环境方面产生和处理,则需考虑是否提供相关软硬件内容,从而可以与其他数据重新组织和发现。而在数据使用协议和相关建议上,也建议应该进行清晰地说明,适当描述数据是否在可控、安全的范围内发布和使用,以便他人重用。

我们对上述指标进行了解释,并在数据论文和数据本身中获取相关数据,以对相关数据质量进行评价。同时,根据理想模型,依托文献调研和对数据论文数据质量评价指标设计结果,我们设计了一般意义上的数据论文评审指标,其中含 4 个一级指标,20 个二级指标,主要包含论文质量和数据质量 2 方面的评审,并在《中国科学数据》上试行(图 3)。

其中对于数据质量评价方面的评价内容及指标对应情况如表 3 所示。

4　结束语

对数据的评审是数据出版过程中的一个重要环节。本文基于数据论文的出版模式,对数据论文及数据的评审模式及内容进行了分析,设计了一般意义上的数据论文评审指标,主要包含论文质量和数据质量 2 方面的评审,并在《中国科学数据》上试行。但目前还处在初步阶段,尤其在数据本身的质量判定方法上有很多复杂性问题需要解决,包括数据本身的复杂性带来的机器和人工判读等,有待深入研究和探讨。

评价项目	主要评价指标
一、论文质量	
题目	能准确反映文章内容，简明扼要。
摘要&关键词	简明扼要，涵盖主要信息。
论文的可读性	论文行文流畅，包含了数据的处理方法、数据的描述及规范性等主要内容，有助于对数据的理解。
数据的可重用性描述	数据价值高，方法和数据处理的每一步的步骤描述都足够详细清晰，足以使他人复制这些步骤；提供了让他人重用数据集或者与其他数据整合所需的所有信息。
数据加工处理方法描述	数据加工处理方法严谨合理，有新意和借鉴意义。
论文价值	数据论文具有较高价值，如学术价值及社会价值等。
二、数据质量	
数据存储	数据存储在可靠且适合的存储库中；作者存放的数据文件齐全，并与数据论文描述相符。
数据查看	该数据集可以访问和查看（便于读者快速阅览，且包含主要元数据信息的预览展示页面）；用于查看数据的软件应提供了包括版本信息在内的相关信息等。
数据质量与丰度(完整性)	数据的生产方法严谨、合理； 数据格式恰当合理，符合业内标准； 根据作者的研究内容，数据的深度、范围、大小及（或）完整性应充分覆盖，数据值应落在预期范围内； 该数据不应含有明细的错误； （如果需要）提供了关于数据质量方面可信的技术验证实验、数据质量统计分析及误差分析。
数据一致性	数据集与数据论文中的描述一致

图 3　《中国科学数据》同行评议表

5　参考文献

[1] BOLIKOWSKI L, HOUSSOS N, MANGHI P, et al. Data as "Firstclass Citizens" [J/OL]. [2016-09-21]. D-Lib Magazine, 2015, 21(1/2). http://www.dlib.org/dlib/january15/01guest_editorial.html

表3 《中国科学数据》数据质量同行评议主要内容及指标对应

评价主要内容	评价参照内容	对应评价体系考察指标	
		一级	二级
数据访问	DOI(或URL等唯一标识符)可访问或提供相关说明	可获取	可访问
数据查看	可下载;若不能提供需提供相关数据获取声明	可获取	可获取、权限说明
数据存储	可靠的存储库(符合出版的数据政策)以及与论文描述相符并齐全	可获取	可获取
数据加工背景	论文相关描述(项目或数据生产加工的背景信息描述)	可评估	可信
数据标准	符合相关科研领域的标准	可评估	标准
数据质量控制	数据论文相关描述	可评估	质量控制
数据丰度	数据集元数据及数据实体描述的数据范围符合作者研究内容,从研究意义上数据(集)覆盖了一定的(时序或空间)范围,足以构成一个有意义的数据集	可评估	数据丰度
数据一致性	论文元数据、数据集元数据、数据实体一致	可理解	一致性
数据完整性	提交数据的元数据信息(数据集元数据,论文中数据集表)及数据实体完整	可理解	完整性
可读性	论文样例数据描述清晰且与数据实体一致	可理解	可读性
数据处理方法合理	数据论文相关描述严谨合理,或有创新可借鉴	可重用	数据处理描述
对科研成果的支持验证	数据论文相关描述及数据实体与研究的逻辑一致性	可重用	数据处理描述
数据的可重用价值	数据论文相关描述,包括数据处理方法的详尽程度,并提供用于重用或组织数据所需的所有信息	可重用	可组织
数据使用	论文相关描述:数据使用建议及使用协议声明	可重用	使用建议使用协议

[2] 英国皇家学会:科学是开放事业[EB/OL].(2012-06-21)[2018-05-30]. http://royalsociety.org/policy/projects/science-public-enterprise/report/

[3] PLoS. Data availability[EB/OL].(2014-03-01)[2018-05-27]. http://journals.plos.org/plosone/s/data-availability

[4] Research data policy types[EB/OL]. [2018-10-22]. https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096

[5] 陈全平. 学术期刊数据政策及相关研究[J]. 图书与情报, 2015(5): 009

[6] 《现代图书情报技术》支撑数据提交要求[EB/OL]. [2017-09-21]. http://manu44.magtech.com.cn/Jwk_infotech_wk3/fileup/1003-3513/NEWS/20160408165409.pdf

[7] 李霞. 话说科技出版国际英文科技期刊的同行评议:上[EB/OL].(2008-06-12)[2018-05-27]. http://news.sciencenet.cn/html/showxwnews1.aspx?id=207779

[8] 张小强, 李欣. 数据出版理论与实践关键问题[J]. 中国科技期刊研究, 2015, 26(8): 813

[9] 张晓林, 沈志宏, 刘峰. 科学数据与文献的互操作[M]//CODATA中国全国委员会. 大数据时代的科研活动. 北京: 科学出版社, 2014: 149

[10] NASA Administrator. Data from NASA's missions, research, and activities[EB/OL].(2017-02-16)[2018-10-30]. https://www.nasa.gov/open/data.html

[11] Dryad Policies[EB/OL]. [2016-09-27]. https://datadryad.org/pages/policies

[12] Figshare Q & A[EB/OL]. [2016-09-27]. https://figshare.com/

[13] CANDELA L, CASTELLI D, MANGHI P, et al. Data journals: a survey[J]. Journal of the Association for Information Science and Technology, 2015, 66(9): 1747

[14] 欧阳峥峥, 青秀玲, 顾立平, 等. 国际数据期刊出版的案例分析及其特征[J]. 中国科技期刊研究, 2015, 26(5): 437

[15] KENALL A. An open future for ecological and evolutionary data?[J]. Bmc Ecology, 2014, 14(1): 280

[16] Earth System Science Data. Review criteria[EB/OL]. [2016-10-18]. http://www.earth-system-science-data.net/peer_review/review_criteria.html

[17] WILEY. Geoscience data journal, guidelines for reviewer[EB/OL]. [2016-10-18]. https://rmets.onlinelibrary.wiley.com/hub/journal/20496060/features/guidelines-for-reviewers

[18] PENEV L, MIETCHEN D, CHAVAN V, et al. Pensoft data publishing policies and guidelines for biodiversity data. Pensoft publishers[EB/OL]. [2016-10-18]. http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

[19] Nature. Scientific data, guide for referees[EB/OL]. [2016-10-20]. http://www.nature.com/sdata/policies/for-referees

[20] Guide for authors[EB/OL]. [2018-10-05]. https://www.elsevier.com/journals/data-in-brief/2352-3409?generatepdf=true

[21] Guide for GigaScience reviewers[EB/OL]. [2018-05-03]. https://academic.oup.com/gigascience/pages/reviewer_guidelines

[22] WILKINSON M D, DUMONTIER M, Jan AALBERSBERG I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific Data, 2016(3): 160018

[23] The fair data principles[EB/OL]. [2017-09-20]. https://www.force11.org/group/fairgroup/fairprinciples

(2018-11-14 收稿;2019-01-28 修回)