生物医学期刊开放代码政策调研

■冯昌扬[1)]　陈雨雪[2)]

收稿日期:2018-09-06
修回日期:2018-10-29

1)武汉大学信息管理学院,湖北省武汉市武昌区八一路 299 号　430072
2)福建船政交通职业学院,福建省福州市仓山区首山路 80 号　350007

摘　要　**【目的】**系统梳理生物医学期刊开放代码政策,了解期刊政策在出版结果的代码公开可用性方面发挥的作用。**【方法】**对 152 种生物医学期刊的作者说明和编辑政策进行人工审查,以分析期刊对开放代码的要求。**【结果】**63.82%的期刊不同程度上要求作者开放代码,61.86%的期刊明确提到版权或许可 67.01%的期刊并未提及开放代码的处理方式。**【结论】**开放获取期刊的开放代码政策比传统订阅制期刊更普遍,公共在线存储库是大多数期刊推荐的公开代码存储平台,高影响因子期刊出台开放代码政策的概率比低影响因子期刊高。

关键词　开放代码;期刊政策;生物医学期刊
DOI:10.11946/cjstp.201809060788

在学术交流活动中,越来越多的学术期刊意识到开放代码的好处,开始鼓励作者共享代码,并要求他们在每篇论文中包含一份关于代码可用性的声明。开放代码指的是可供免费分发和重复使用的计算机代码(软件),其源代码不受限制[1]。结合 Easterbrook 的观点[2],本研究将"开放代码"定义为在某些平台上自由发布代码、模型和算法的过程,因此,其他学科的研究人员可以对这些代码进行分析,并可以重新运行代码来验证结果。例如,为了解决日益复杂的数据和分析问题,《科学》(*Science*)扩展了数据访问要求,包括涉及数据创建或分析的代码[3];《自然》(*Nature*)、《自然方法》(*Nature Methods*)、《自然生物技术》(*Nature Biotechnology*)和《自然神经科学》(*Nature Neuroscience*)等期刊鼓励作者提供源代码、安装指南和样本数据集,以供审稿人检查[4];《生物统计学》(*Biostatistics*)创建了"再现性副主编"(Associate Editor for Reproducibility)这一职位,致力于根据收到的数据和代码重复论文;《生物信息学》(*Bioinformatics*)要求作者在提交论文时描述如何访问其软件,并在文章的标题页中指明能够访问源代码的统一资源定位符(Uniform Resource Locator, URL)[5];《内科医学年鉴》(*Annals of Internal Medicine*)则要求作者说明他们是否愿意在论文出版后分享他们在研究中开发和使用的源代码、数据和协议[6]。

鉴于期刊对科学传播的重要作用和对研究人员具有较大的影响,国外一些学者已经展开期刊开放代码政策的相关研究,如 Stodden 等[7-8]通过评估样本期刊的数据共享政策、代码共享政策、补充材料政策和开放获取状态,建立了期刊采用开放数据和代码政策的预测模型;通过向作者请求数据和代码并尝试复现已发表的结果来评估开放代码政策的有效性,研究发现,作者的数据和代码发布情况比没有推出开放代码政策时有所改进,但目前还不足以复现实验结果。此外,Rowhani-Farid 等[9]发现《生物统计学》采用再现性政策奖励带有数据和代码共享徽章的文章,他们通过样本期刊文章的提交日期绘制代码共享概率,并进行贝叶斯逻辑回归建模,发现生物统计学期刊的徽章并没有影响代码共享。如前所述,生物医学领域的研究人员更多地关注代码可用性及可重复性,然而笔者通过文献调研发现我国生物医学领域期刊鲜有对开放代码提出要求,目前也鲜有关于期刊开放代码政策的研究,因此,本研究试图采用内容分析法调研生物医学期刊开放代码政策的相关特征。

基金项目:国家留学基金委资助项目(201706270042);国家自然科学基金重大研究计划培育项目"面向多主体共享需求的国家大数据资源治理机制设计"(91546124)。

作者简介:冯昌扬(ORCID:0000-0002-2480-9889),博士研究生,E-mail:fengchangyang@whu.edu.cn;陈雨雪,助理馆员,硕士。

1 数据来源与概况

本研究使用的数据集来自 2017 年《期刊引证报告》(*Journal Citation Reports*,JCR)中的生物医学研究期刊,这些期刊广泛分布于 Web of Science(WoS)数据库的生物化学和分子生物学、生物学、细胞生物学、晶体学、发育生物学、生物医学工程、免疫学、医学信息学、微生物学、显微术、多学科科学和神经科学等分类中。由于生物医学期刊数量较多,本研究将研究样本限制为 Q1 区的期刊。具体检索式为:Select Categories:Biochemistry & Molecular Biology, Biology, Cell Biology, Crystallography, Developmental Biology, Engineering、Biomedical, Immunology, Medical Informatics, Microbiology, Microscopy,Multidisciplinary Sciences,Neurosciences;Select JCR Year:2017;Select Edition:SCIE;Category Scheme:WoS;JIF Quartile:Q1。

初始数据集包括 299 种期刊。经过人工审核,排除非英语语种期刊、简短报告和评论期刊、基础医学或临床研究期刊,最终确定的研究样本包括 152 种期刊,占 JCR 该领域 Q1 区期刊总量的 50.84%。从 JCR 报告中获取的数据,包括期刊标题(Full Journal Title)、期刊的总被引频次(Total Cites)、影响因子(Impact Factor)和特征因子分值(Eigenfactor Score)。2017 年各影响因子区间的期刊数量和相应的占比如表 1 所示。提取这些期刊的期刊政策(Information for Authors 和 Editorial Policies)作为编码文本。

表 1 期刊影响因子分布

影响因子区间	期刊数量/种	占比/%
2~3.99	17	11.18
4~5.99	71	46.71
6~7.99	24	15.79
8~9.99	11	7.24
10~29.99	25	16.45
≥30	4	2.63

2 研究方法

采用内容分析法研究国外期刊开放代码政策。内容分析法是从文本语料库中发现定量模式的有效方法。在内容分析法中,编码是数据收集和数据解释之间的关键联系,它可以为研究人员提供一套系统的指导方针(即编码方案)来解释数据。

编码的第一步是确定研究目标并创建一个明确的编码方案。本研究结合 Stodden 的编码方案[7],创建一个编码草案。为了补充在拟定草案时未注意到的元素,笔者采用基础理论方法将草案应用于数据集的一个子集,以帮助进一步改进编码方案,得到最终版编码方案(表 2)。

两位具有编程背景,并具有文本编码经验的信息科学专业的研究生作为编码人员,对 50 个随机抽取的期刊样本进行编码。本研究使用 Cohen′s kappa 系数来测量编码人员之间的可信度(Interrater Reliability,IRR),其 IRR 为 0.8,这为一位编码人员提供了足够的可靠性来编码所有样本。

表 2 最终版编码方案

代码分类	编码代码	代码描述
A 开放代码政策	3	作为论文发表的必要条件
	2	明确鼓励,可能会被审查或托管
	1	没有提及,或隐晦提出
B 期刊获取	O	开放获取
	S	订阅
C 推荐公开方式	G	公共在线存储库,如 GitHub
	J	期刊托管
	R	读者向作者发出请求
	N	没有明确说明
D 代码处理方式	V	期刊审核
	H	期刊仅托管
	N	没有明确说明
E 版权声明	Y	明确提及
	N	没有明确说明

3 研究结果

3.1 开放代码政策概况

在 152 种样本期刊中,40 种(26.32%)期刊将开放代码作为发表条件,57 种(37.5%)期刊明确鼓励开放代码,但并未强制要求,55 种(36.18%)期刊未提及任何有关开放代码的内容(表 3)。

表 3 样本期刊开放代码政策的概况

政策	期刊数量/种	占比/%
3 作为论文发表的必要条件	40	26.32
2 明确鼓励,可能会被审查或托管	57	37.50
1 没有提及,或隐晦提出	55	36.18

3.2 期刊对公开代码的处理方式

65 种(67.01%)期刊没有明确提及公开代码的处理方式,2 种(2.06%)期刊愿意托管作者提交的代码,30 种(30.92%)期刊会对代码进行审核。与仅鼓励作者开放代码的期刊相比,将开放代码作为发表必要条件的期刊审核代码的概率更大(表 4)。

表 4 样本期刊对公开代码的处理方式

开放代码政策	H 期刊仅托管		R 期刊审核		N 没有明确说明	
	期刊数量 /种	占比 /%	期刊数量 /种	占比 /%	期刊数量 /种	占比 /%
2 明确鼓励,可能会被审查或托管	2	2.06	5	5.15	50	51.55
3 作为论文发表的必要条件	0	0.00	25	25.77	15	15.46
合计	2	2.06	30	30.92	65	67.01

3.3 期刊推荐的代码公开方式

75 种(49.34%)期刊建议作者通过公共在线存储库开放代码,2 种(1.32%)期刊推荐通过期刊托管方法,2 种(1.32%)期刊建议读者请求作者共享,73 种(48.03%)期刊没有指定代码公开方式(表 5)。

在要求开放代码(编码为 A-3)的 40 种期刊中,36 种期刊建议通过公共存储库开放代码,4 种期刊没有指定代码公开方式。

表 5 样本期刊推荐的代码公开方式

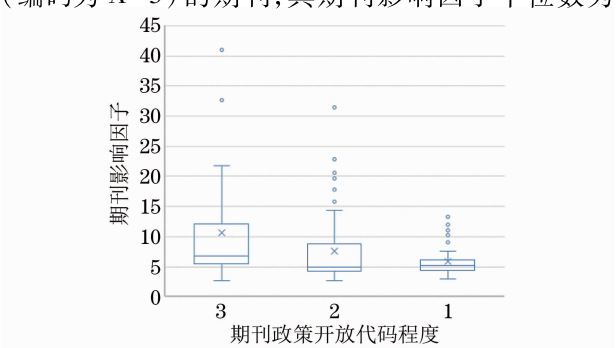
推荐公开方式	3 作为论文发表的必要条件		2 明确鼓励,可能会被审查或托管		1 没有提及,或隐晦提出	
	期刊数量 /种	占比 /%	期刊数量 /种	占比 /%	期刊数量 /种	占比 /%
G 公共在线存储库	36	23.68	38	25.00	1	0.66
J 期刊托管	0	0.00	2	1.32	0	0.00
R 读者向作者发出请求	0	0.00	2	1.32	0	0.00
N 没有明确说明	4	2.64	15	9.87	54	35.53

3.4 开放获取期刊的开放代码概况

对开放获取期刊的开放代码概况进行 Fisher 精确检验(Fisher′s Exact Test),结果得出 P 值为 0.007,表明开放代码政策强度与期刊访问模式之间差异有统计学意义。笔者进一步将编码 A 分为两大类,期刊政策将开放代码列为必需条件(编码为 A-3的期刊)和期刊政策认为开放代码是非必需的(编码为 A-2 或 A-1 的期刊),并使用卡方检验(Chi-Square Test)来测试此关联,发现两类期刊开放代码政策的差异有统计学意义(P=0.009)。两个结果都表明,开放获取期刊更有可能要求开放代码。

3.5 不同影响因子期刊的开放代码要求

图 1 所示为 2017 年每个开放代码级别期刊的影响因子中位数。2017 年具有最强开放代码政策(编码为 A-3)的期刊,其期刊影响因子中位数为 6.772;然而,没有提及开放代码(编码为 A-1)的期刊的影响因子中位数为 5.186。



注:长方形方块的上边缘所对应的纵坐标刻度值为影响因子上四分位数,下边缘所对应的纵坐标刻度值为影响因子下四分位数,长方形方块里的横线所对应的纵坐标刻度值为影响因子中位数,×所对应的纵坐标刻度值为影响因子平均数;长方形方块两端的竖线的上限所对应的纵坐标刻度值为影响因子上四分位数加上1.5倍四分位距的取值,下限所对应的纵坐标刻度值为影响因子下四分位数减去1.5四分位距的取值;超出此范围的离群样本表示为点。

图 1 不同影响因子期刊的开放代码要求

影响因子与编码 A 的差异具有统计学意义(Kruskal-Wallis 检验,P<0.0001)。通过检查编码 A 中各类别之间的成对差异,笔者发现将开放代码政策作为论文发表必要条件(编码为 A-3)的期刊具有显著高于期刊政策认为开放代码是非必需(编码为 A-2 或 A-1)的期刊影响因子(Wilcoxon 检验,均为 P <0.0001)。笔者进一步将期刊政策分为两类,即期刊政策将开放代码列为必需条件(编码为 A-3 的期刊)和期刊政策认为开放代码是非必需的(编码为 A-2 和 A-1 的期刊),发现需要开放代码期刊的影响因子显著增加(Wilcoxon 检验,P<0.0001)。

3.6 期刊开放代码版权声明概况

只有 60 种(61.86%)期刊明确提到版权或许可,即使是要求开放代码(编码为 A-3)的期刊,也只有 30 种(30.93%)期刊提到版权或许可(表 6)。

表 6 期刊开放代码版权声明

开放代码政策	Y 明确提及		N 没有明确说明	
	期刊数量 /种	占比 /%	期刊数量 /种	占比 /%
2 明确鼓励,可能会被审查或托管	30	30.93	27	27.84
3 作为论文发表的必要条件	30	30.93	10	10.31
合计	60	61.86	37	38.14

4 讨论、结论与展望

4.1 讨论

从整体来看,生物医学期刊对开放代码有不同

程度的要求,但大多数期刊没有明确提及公开代码的处理方式,这或许与这项工作背后需投入的巨大人力有关。但从另一方面考虑,期刊审核无异于一种作者代码监督机制,可以在一定程度上减少粗糙的代码。如若可行,期刊还可不定期跟踪作者共享代码的后续,如研究人员是否对代码进行更新和维护、代码的影响力如何等。

从代码公开方式来看,大多数期刊推荐作者使用 GitHub 等公共在线存储库分享代码,这与实际情况相吻合。在实践中,GitHub、FigShare、Zenodo 和 Bitbucket 已成为预选的学术交流工具,尤以 GitHub 为甚[10]。GitHub 成立于 2008 年,广泛用于存储、分享、更新数据集和软件代码。截至 2018 年 6 月 13 日,谷歌学术(Google Scholar)中有超过 22.3 万篇学术论文引用 GitHub 存储的代码[11]。GitHub 在 2018 年 7 月的美国 Alexa 网站排名中位居 32 位[12]。

从期刊获取方式来看,开放获取期刊更倾向于出台开放代码政策,这也是开放科学的题中之意。尽管开放代码与开放获取、开放数据的目的不尽相同,开放代码更多地是为了让读者重复,甚至更新模型、算法、实验步骤,但开放获取、开放数据和开放代码之间存在着紧密联系,开放科学、数据共享、软件共享都是未来的发展趋势[13]。

从影响因子来看,影响因子与开放代码政策显著相关,高影响因子的期刊如《自然》《科学》等更有可能要求作者开放代码。但反过来,影响因子计算方法中的总被引频次是根据 WoS 所收录的 SCI 期刊论文对该期刊两年内发表论文的引用情况计算出来的,那么开放代码是否与期刊被引频次存在相关关系,并对期刊影响因子产生影响,则是笔者在下一个研究中讨论的问题。

从版权声明来看,明确提及版权声明的期刊所占比例不高,知识产权是否是作者不愿分享代码的原因还有待考量。但从既往研究来看,Stodden 等[14]从用户角度对机器学习社区进行调查,并指出了开放代码没有得到广泛实践的原因,其中权属不清晰占 44%,其次是专利问题(40%);Barnes[15] 也发现公开代码与机构知识产权相悖,这是科学家没有公布其代码的原因之一 。可见,知识产权问题确是影响作者共享代码的因素之一,至于如何拟定版权声明,包括代码公开范围、使用范围、程度等需要声明的条款,则可以成为后续研究考虑的问题。

4.2 结论

普遍和大规模的计算正在改变人们对科学方法的实践。如果没有代码,就会导致所提供的信息不足,影响他人再现已发布的计算结果。在这项研究中,笔者试图了解期刊在出版结果的代码公开可用性方面发挥的作用。

本研究通过对 2017 年 JCR 中 Q1 区的 152 种生物医学期刊开放代码政策的相关特征进行分析,发现开放获取期刊比传统订阅制期刊的开放代码政策更普遍;公共在线存储库如 GitHub 等是大多数期刊推荐的公开代码存储平台;高影响因子期刊比低影响因子期刊更有可能出台开放代码政策。

4.3 局限与展望

本研究的局限性主要体现在样本的选择上:(1)由于选取的是 JCR 中生物医学 Q1 区的期刊,鉴于样本期刊已具有较高的影响力,它们选择通过开放获取来扩大影响力的可能性相对较低,因此本研究的样本期刊选取具有一定的局限性;(2)部分期刊如《自然》影响因子显著高于 Q1 区期刊影响因子的平均水平,这可能对显著性检验产生了一定影响;(3)笔者假设年轻期刊更有可能出台开放代码政策,并意图研究期刊创刊年份对开放代码政策的影响,但由于 JCR Q1 区期刊多为老牌期刊,不适合开展此研究。

后期研究,笔者将集中在以下 3 个方面:(1)笔者在对期刊政策进行编码时发现,一些出版社旗下期刊存在共用该出版社同一套期刊政策的现象,出版社对开放代码政策的影响可以进行回归分析和预测分析;(2)期刊开放代码政策的有效性,即验证作者的执行情况,甚至可以对比期刊论文开放代码出台政策前后有何不同,以检验开放代码政策的影响力;(3)研究期刊创刊年份对开放代码政策的影响,是否年轻期刊出台开放代码政策的可能性更大。

参考文献

[1] Gacek C, Arief B. The many meanings of open source[J]. *IEEE Software*, 2004, 21(1): 34-40.

[2] Easterbrook S M. Open code for openscience? [J]. *Nature Geoscience*, 2014, 7(11): 779-781.

[3] Hanson B, Sugden A, Alberts B. Making data maximally available [J]. *Science*, 2011, 331(6018): 649.

[4] *Nature* editorial. Does your code stand up to scrutiny? [J]. *Nature*, 2018, 555(7695): 142.

[5] Instructions toauthors [EB/OL]. [2018-07-10]. https://academic.oup.com/bioinformatics/pages/instructions_for_authors.

[6] Laine C, Goodman S N, Griswold M E, *et al*. Reproducible

research: moving toward research the public can really trust[J]. *Annals of Internal Medicine*, 2007, 146(6): 450-453.
[7] Stodden V, Guo P, Ma Z. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals[J]. *PLoS ONE*, 2013, 8(6): e67111.
[8] Stodden V, Seiler J, Ma Z. An empirical analysis of journal policy effectiveness for computational reproducibility[J]. *Proceedings of the National Academy of Sciences*, 2018, 115(11): 2584-2589.
[9] Rowhani-Farid A, Barnett A G. Badges for sharing data and code at *Biostatistics*: an observational study[J]. *F1000 Research*, 2018, 7: 90.
[10] GitHub and more: sharing data & code[EB/OL]. [2018-07-10]. https://101innovations.wordpress.com/2016/10/09/github-and-more-sharing-data-code.
[11] Microsoft's purchase of GitHub leaves some scientists uneasy [EB/OL]. [2018-07-10]. https://www.nature.com/articles/d41586-018-05426-0.
[12] Top sites in United States[EB/OL]. [2018-07-10]. https://www.alexa.com/topsites/countries/US.
[13] Gewin V. Data sharing: an open mind on open data[J]. *Nature*, 2016, 529(7584): 117-119.
[14] Stodden V C. Policies for scientific integrity and reproducibility: data and code sharing [C]. American Association for the Advancement of Science, 2011.
[15] Barnes N. Publish your computer code: it is good enough[J]. *Nature*, 2010, 467(7317): 753.

作者贡献声明:

冯昌扬:确定选题,提出研究框架,撰写论文初稿;
陈雨雪:文献检索,图表制作,修改论文。

An empirical analysis of open code policies of biomedical journals

FENG Changyang[1], **CHEN Yuxue**[2]

1) School of Information Management, Wuhan University, 299 Bayi Road, Wuchang District, Wuhan 430072, China
2) Fujian Chuanzheng Communications College, 80 Shoushan Road, Cangshan District, Fuzhou 350007, China

Abstract: [**Purposes**] This study systematically reviews the open code policy of biomedical journals to understand the role of journal policies in the code availability of the published results. [**Methods**] The author's guidelines and editorial policies of 152 biomedical journals were manually reviewed to analyze open code requirements of journals. [**Findings**] A total of 63.82% of journals require authors to open code to varying degrees, and 61.86% of journals explicitly mention copyright or license, but 67.01% of journals do not mention how they deal with open code. [**Conclusions**] Open access journals are more likely to introduce open code policies than traditional subscription journals, and public online repository is the recommended platform for most journals. Besides, high-impact journals are more likely to introduce open code policies than low-impact journals.
Keywords: Open code; Journal policy; Biomedical journal

(本文责编:刘晶晶)