

# 数据论文:数据集 独立出版与共享模式研究

王丹丹 (河南科技大学管理学院 洛阳 471003)

**摘要** 文章对数据论文这一新型数据出版与共享模式进行了探讨,揭示了数据论文出版的本质特征,并基于现有的出版实践,总结了数据论文出版需要解决的问题以及推进过程中遇到的困难,提出了相应的建议。

**关键词** 数据论文 数据出版 数据共享 开放数据

Data Papers:Independent Publishing and Sharing Mode of Data Set

Wang Dandan ( College of Management, He'nan University of Science and Technology, Luoyang, 471003)

**Abstract** The paper discussed the data paper which is a new mode of data publishing and sharing, revealed the essential characteristics of data published mechanism, then started from the existing practice, summarized the basic problems that data paper publishing need to solve, difficulties faced in the process of advancing it and gave some suggestions.

**Keywords** data paper, data publishing, data sharing, open data

## 1 引言

在科学技术飞速发展的今天,科学数据迅速积累并在科学研究中发挥着越来越重要的作用。促进科学数据有效利用的前提条件是实现分散在不同国家、科研机构、研究项目以及科研人员手中的科学数据的共享。在早期,科学数据共享强调的是数据收集与整合,国家经费的支持与相关政策的约束是保证其数据共享能够长期可持续发展的必要条件。如何吸纳科研人员参与科学数据的管理与发现,让科研人员与数据中心一起推动数据共享,成为科学数据共享工作发展的新目标<sup>[1]</sup>。目前,数据出版作为有望解决这一问题的有效方法,成为出版界和数据共享界共同提出并积极探索的新领域。

数据未出版的一个主要原因是缺乏激励机制,科研人员参与创造和管理数据的积极性不高<sup>[2][3]</sup>。准备用于出版的数据是一个非常耗时的活动,如果不能得

到同行的认可,几乎不会有人愿意去出版数据。因此,基于目前成熟的科学声誉系统出现了三种促进数据共享的出版模式<sup>[4]</sup>:一是将研究数据作为一个独立的信息对象存储在研究数据知识库中;二是以所谓数据论文的形式,将研究数据作为文本性文档进行出版;三是将研究数据作为论文的附录和论文一起出版(作为注释文本内容的材料,以丰富出版物内容的用途,作为一种说明文件得以发表,“使得出版物丰富化”)<sup>[1][5]</sup>。本文重点讨论数据论文出版模式,总结数据论文机制,揭示数据论文的基本要素,并分析数据论文模式实施所面临的问题和潜在的障碍。

## 2 数据出版概述

数据出版是指将数据作为一种重要的科研成果,从科学研究的角度对数据进行同行审议和公开发布,创建标准且永久的数据引用信息,供其他研究性文章引证<sup>[6]</sup>。数据出版从搜索和浏览数据开始,科研人员

本文系国家社会科学基金“数字图书馆用户数据资源化管理研究”(编号:2013CTQ012)与河南省教育厅人文社会科学研究项目“面向智慧教育的开放关联数据应用与启示”(编号:2015-GH-412)的阶段性成果之一,并受河南省高等学校人文社会科学重点研究地“高等教育与区域经济发展研究中心”的资助。

获取数据后会熟悉、学习、审核并处理数据。其次,通过开展新型实验或从不同角度处理数据获得新数据,并展开新研究。科研人员编写数据文件对这些简单的数据进行解释或注释,以吸引其他科研人员,同时增加元数据,以使数据具备可检索和再利用的能力。第三,类似于学术论文的质量保障,对于要出版的数据文件来说,也需要对其数据及其元数据的质量进行相应的控制,使其遵守格式要求并符合科学的质量标准。最后,在数据文件及其元数据和附加文件的质量得到保证后,对数据进行出版和存储。从数据出版的整个流程来看,科研人员既是科学数据共享的重要参与者,其态度在较大程度上也决定着科学数据共享的进程和发展。因此,任何数据出版形式,只有当科研人员认为有价值的时候,才可能长期可持续发展。

传统的数据出版模式是将科学数据作为一个独立的信息对象存储在科学数据知识库中,论文出版时必须把相关数据提交到数据知识库中,并为其分配一个可长期使用的标识符号,如DOI、URN等。常见的数据知识库主要分为机构数据知识库、学科数据知识库、多学科数据知识库以及特定项目数据知识库4类<sup>1</sup>。存储在数据知识库中的数据,通过数据描述符或引用建立与论文的关联。如《自然》(Nature)给出了一个建议的数据知识库列表,要求作者将数据存储在建议的知识库中。《自然》中发表的论文必须明确标识数据集的访问控制号、链接或DOI号,并将数据集列入参考文献列表。

传统数据出版模式的核心是帮助作者出版内容以促进科学价值的体现和数据集的再利用,因此忽视了对数据的检索,必须将数据上传到搜索引擎或者知识库目录才能实现其可检索性。出版后,数据将被锁定(不可改变),只能通过发布新版本的时候修订数据,这就带来了数据维护的问题,在缺乏约束的情况下,很多作者并不会对数据进行更新。在实际操作过程中,数据公开程度主要依靠期刊编辑们的提醒和约束,一些作者虽然承诺数据公开,但论文发表后,往往采取种种借口不进行实际的数据公开。

传统数据出版模式在生命科学领域已经相当完善,该领域的学科数据知识库GenBank、多学科数据知识库Dryad已经具备一定的影响力。使用和传播GenBank的数据没有限制条件,而在Creative Commons License CC0下也可以实现Dryad数据的获取和使用。将科学数据作为论文的附录和论文一起出版直接解决了论文和数据的关联问题。这里数据是作为论文的支撑材料提交的,其目标是构建并维持一种技术环境以围绕一篇论文关联所有的信息对象,以便于创建一种知识空间,在这一空间中作为论文基础的研究数据可以免费获取。以数据论文的形式将科学数据作为文本性文档进行出版,作为一种鼓励作者主动更新数据的机制,开始获得了越来越多的关注。它允许作者出版数据,并通过传统的引用过程得到认可。作为一种期刊出版物,数据论文的主要目的是描述数据,而不是报告研究调查。因此,它包含有关数据的事实,而不包含基于这些数据产生的假设和论证,正如在传统论文

中所看到的那样<sup>8</sup>。在地球和生态学领域已经出现了数据论文出版的实践<sup>9,10</sup>。目前,这一模式已经扩展到数据期刊,主要集中在生物学、地球科学、化学化工和物理学领域。

《地球系统科学数据》(Earth System Science Data, ESSD)2008年发布了地球科学数据的描述规则,2009年开始专门发表数据论文,要求将相关的数据集存储在其他数据知识库中,其定位是出版有关原始数据集的论文,推动对地球科学有益的高质量数据的重用。Wiley公司与皇家气象学会合作推出开放存取期刊《地理科学数据期刊》(Geoscience Data Journal, GDJ),用以在线出版简短的地理科学数据论文,并关联已存储在数据中心的数据以及授权DOI的数据集等。

### 3 数据论文出版模式

相对于前两种模式,有关数据论文的界定目前还比较模糊,研究较少。是否可以将数据论文作为促进数据有效利用,保证数据共享能长期可持续发展的重要机制和有效手段,尚无定论。邱春艳认为数据论文应该从数据收集、数据处理过程、所用软件工具以及数据文件格式等方面对一个数据集进行描述。而数据期刊则是一类以描述一个或一组数据集为首要目标的出版物,既可以只出版短的数据论文,如《地球科学期刊》(Geoscience Data Journal),又可以创建一种 workflow与框架,借助导航式的自动出版过程把写作、审稿、出版、存储、分发、互操作、收集与数据再利用集成完成,如《生物多样性数据期刊》(Biodiversity Data Journal)<sup>11</sup>。Chavan认为数据论文的目的有三:提供一种可引用的期刊出版物为数据出版商带来学术认可;以结构化的人可读的形式描述数据;使数据的存在引起学术社区的注意力<sup>8</sup>。Jonathan分析了数据重用需要解决的一系列难题,认为数据论文如果设置足够合理,可以解决这些问题,比如保持作者发布数据的热情<sup>12</sup>。将科学数据与期刊论文进行对比(如表1),发现在当前的学术体系中科学数据的地位类似于二等公民,因此迫切需要一种机制来体现科学数据在学术系统中的价值。

表1 科学数据与期刊论文的对比分析

笔者认为,出版数据论文的主要目的是实现了劳动的分工,使那些拥有资源和技能,将能够完成实验和观察以收集潜在感兴趣的数据集的人的劳动分离出来,使每一个有独立背景和分析数据能力的研究者或机构在看到合适的数据时,都可以使用它。最重要的是,通过数据论文的发表以及其他人对数据论文的引用,能够使生产数据的那些人得到单独的认可(如图1)。类似于传统研究论文研究方法部分的描述,数据论文描述数据获取的过程,但是描述程度更为详

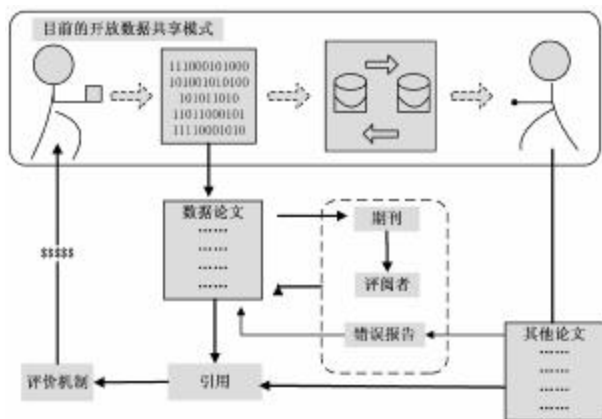


图1 传统数据共享模式与基于数据论文的数据共享模式

尽,同时包含有关试验设计的原理、动机和相关考虑的讨论,但不提供任何数据分析和分析结果;类似于论文的发表,数据论文也需要经历同行评议的过程,以确保数据获取所使用方法得当、数据质量可靠、有关数据的描述准确且完整。数据期刊出版专门针对科学数据的数据论文,不以科学数据的详细分析为内容,只对科学数据的题名、数据创建日期、数据创建者、摘要、永久识别符、存档资源的链接或者实验条件、设施、环境要求等元素进行描述。

#### 4 数据论文基本内容及关键问题

##### 4.1 数据论文基本内容

全面且细致的数据规范是数据论文效用得以实现的先决条件,是防范错误理解和误用数据集风险的有效手段。数据论文应该提供给数据用户在更详细的程度上理解数据集所需要的信息。为了实现数据集的有效共享和再利用,需要对研究动机和设计思路进行清晰而综合的描述,同样需要说明所认为的重要的备选设计方案以及他们被拒绝的理由。缺乏这些内容,会导致用户分析和解释产生偏见(例如招聘和采样),限制结果的有效性和普遍性。用户需要充分了解实验过程中所做出的决策,以充分估计他们产生的影响。因此,每一篇数据论文都应该明确指出研究者收集数据的动机。即使发布的只是数据的一个子集,在数据论文中也应包括没有发布的数据的信息(数据论文中应该包含所有的情景信息)。另外,截至发表论文日期,使用目前数据的所有论文都应列为参考文献,以使未来的数据用户了解使用这些数据可以预期获得哪些结果。

Krzysztof 等研究了神经影像学社群的特点,认为其领域的数据论文应包括:研究概述,指出创建数据样本的明确目标以及研究设计指导原则;参与者(研究对象)描述,包括样本规模、招募策略、入选和排除标准、样本的人口统计学特征、知情同意的等方法;实验设计,包括研究类型(横向的或纵向的)、研究时间表、研究 workflow、扫描会话提纲、任务和激励描述、陈述法则、针对被研究者的指令、不包含在共享数据样本中的数据描述;表型评估协议,包括人口统计学、诊断评估协议、科研人员收集数据的资质(包括测度

信度的措施);扫描会话详细信息,包括 MR 协议指南描述每次扫描的顺序、类型、目的和采集参数,每次扫描的条件(例如休息时眼镜的开闭、参见实验设计)。数据发布方式,包括发布站点、发布类型(数据库、机构库、本地 FTP)、图像数据格式(如 NIFTI、DICOM、MINQ)、成像数据公约(神经的或放射性的)、表型数据关键、丢失的数据和授权协议<sup>[3]</sup>。

##### 4.2 数据论文出版要素解析

**数据类型。**数据共享和开放科学是两个相关但是不同的现象。个人可以选择在有限的一组合作者之间或者在更广泛的社区里共享数据。即使对于打算开放获取的数据,在同意使用数据前也必须保护参与者的隐私。重要的是,对于出版严格受限数据的数据论文应该分享有关试验设计的见解,或者提供一种潜在渠道鼓励科学社区的成员联系数据生产者并寻求合作。从引用的角度,选择发表数据论文的作者若不与人分享的话被引用的可能性就比较低,除非他们共享有潜在价值的研究设计;相比之下,开放获取数据集被频繁引用的可能性更大。因此,数据论文出版的内容不应该仅仅局限于开放获取数据集。然而,对作者而言最重要的是必须清楚指明他们所需要遵从的数据共享政策。

**发布时间。**过去一些期刊比如《科学》和《美国国家科学院院刊》等出版物尽管没有明确的数据出版机制和数据共享时间表,但有一段声明指出论文中所使用的数据将会开放共享。对期刊而言,一旦承诺后未能及时地用合适的方式共享数据的话,会破坏用户对期刊的信任。同样数据论文也应该明确说明在什么时候会以什么方式共享数据。例如,在数据论文出版时实现数据共享,或在一个具体的时滞期后开放获取或者限制获取。总之,无论共享政策是什么,评阅者均有权获取数据以验证其一致性,也为数据共享做好准备。

**共享规范。**关于共享的标准尽管伦理委员会号召达成共识,但各领域仍存在显著差异。如 FCB/INDI 要求共享数据遵从 HIPAA 法案的隐私规则<sup>[4]</sup>。遵从 HIPAA 规则不需要事先征得参与者的同意,但是并非所有的数据都可以满足 HIPAA 的隐私规则。即使完全去除身份信息,一些综合信息或者潜在的判别信息仍然会增加数据共享的风险。另外,不同国家之间的隐私规定可能会存在差异。因此,在数据论文中必须明确说明征得共享同意的过程(所获取数据的国家的立法背景),证明共享的合适性,证明数据共享符合当地伦理委员会的要求,并提供一项违反隐私的风险评估内容。

**作者或声誉。**数据论文的一个主要激励机制是确保参与研究的所有人都得到合适的认可。数据论文提供给那些最直接参与设计和生产数据的人一个机会使他们可以得到合适的认可,避免真正的贡献者淹没在冗长的作者列表中。此外,也减少了分析导向型论文的认知压力,使研究论文的方法部分所占的篇幅可以大幅压缩,不需要再充分描述整个获取过程。

**同行评议。**与传统论文相类似,数据论文质量保证也需要进行高质量的同行评议。如果评阅者不能直接检查作者所准备发布的数据,证明所发布的数据具有可读性,将快速破坏数据共享的过程和数据论文的

可信性。需要说明的是评阅者不是人工检查每一个共享的数据项,评阅者的工作是评价研究设计和特定样本的收集程序、选择共享的位置、评价共享报告的可读性。应该通过期刊将数据论文和清晰的检查明细提供给评阅者,以确保数据论文达到基本标准。

错误纠正。数据发布后需要对数据进行更新和更正。数据共享允许外部的调查者使用最初数据创建者没有考虑到的方式评审数据,并发现数据中的错误。应该提供简单快捷的错误报告和纠正渠道,比如小的错误通过系统的评论实现,更实质性的修订由编辑者来操作。此外,应通过期刊对数据的错误纠正进行报导,并提供便利。使用勘误表(首选)或资料误差修正更新补充机制以避免对作者产生的消极影响,使其愿意报告错误情况。

## 5 结语

出版数据论文的目的是为了获得相关专业的认可,并经过同行评议使数据质量得到控制。数据论文对作者荣誉的划分更为详细,使数据生产者在研究设计、执行和维护中所付出的努力得到适当的认可,也创造了一种与他人合作的机会。数据论文实现了独立于分析的数据质量评价,使科研人员通过基于出版物影响力的计量指标得到共享数据的认可,有助于实现数据共享的可持续性,促进数据的可获取性。

尽管数据论文方法理论上非常有效,但推广这一机制仍然有较大困难。首先是数据出版者方面的困难。比如在生物多样性领域有三类收集和共享数据的组织:一是以自然历史博物馆为代表公共资助机构,共享数据是他们的基本责任,他们主要关注如何使数据直接在线可利用<sup>[5]</sup>;二是科研人员,他们对数据论文很感兴趣,但是多数不熟悉出版机制;三是问题驱动型的专业调研人员,他们大多数不愿意出版未经分析的原始数据,而且只有在相关的研究论文发表之后才愿意共享数据。对于第三类用户,只有先解决了将数据集作为研究论文的一部分出版的问题后,才能是数据论文。其次,设置数据论文基本内容标准,以确保数据论文的完整性和数据效用仍存在较大分歧,即什么样的数据论文才是合格的?谁该为数据文件的发布买单是有争议的。第三,实施集中、联合还是独立的数据存储平台也是未解的难题。目前没有集中式的数据托管机制,正在努力实现的数据资源联合共享方案(例如,INCF数据空间<sup>[16]</sup>)正在开发阶段,不清楚领域是否愿意接受这样一个实体。愿意共享数据的科研人员可以将其数据放入现有的任何一个数据知识库中,或者自己保存数据。就自己保存数据而言,维护对数据集的获取既不容易又不便宜;另外,无法保证数据的持续获取性,当两个或多个平台管理同一数据集时,保持数据的一致性和同步性也存在风险。而且外部的数据知识库依赖于资助机构的资助程序或维护服务持续性的基金。最后保证不给评阅者带来额外的负担比较困难。尽管出版数据论文有助于简化和澄清知识产权问题,比如数据所有权和引用权,但是对科

学社区有限的评阅者而言,面向数据论文的同行评议过程给评阅者增加了额外的负担。因此,需要基于现有的信息计量学工具,开发半自动化的工具以减轻评阅者的工作量,帮助其快速地进行评价。

总之,通过对数据论文进行出版并计算共享数据产生的生产力将会吸引对数据共享产生价值的关注。因此,资助机构和高校必须努力出台政策,增加对产生高质量的数据科学价值及其需求的认识,以激励和奖励产生数据供科学社区使用的调查者。通过开发和优化现有的计量学工具,进一步优化数据论文的过程。未来,可以考虑使用共享因子而不仅仅是影响因子来评价科研人员的学术贡献,以构建起一种共享的文化。不仅仅评价科研成果的引用次数,也评价科研人员对社会知识和信息共享所做出的贡献程度。

## 参考文献

- [1] 吴立宗,王亮绪,南卓铜.科学数据出版现状及其体系框架[J].遥感技术与应用,2013(3):383-389.
- [2] Costello M J. Motivating online publication of data[J]. BioScience, 2009, 59(5):418-427.
- [3] Duke C, Porter J. The ethics of data and reuse in biology[J]. Bioscience, 2013, 63:483-489.
- [4] Heinz Pampel, Stijn Dallmeier Tiessen. Open Research Data: From Vision to Practice[M]. Springer: Opening Science, 2014:213-224.
- [5] 顾立平.科学数据权益分析的基本框架[J].图书情报知识,2014(1):34-52.
- [6] 杜伟,张静.科学研究数据的出版与获取[J].出版科学,2013(6):86-90.
- [7] 刘峰,张晓林,孔丽华.科研数据知识库研究述评[J].现代图书情报技术,2014(2):25-30.
- [8] Chavan V, Penev L. The data paper: A mechanism to incentivize data publishing in biodiversity science[J]. BMC Bioinformatics, 2011(15):12.
- [9] The AGU Journals Have Published Data Papers for Many Years [EB/OL]. [2014-12-23]. [http://www.agu.org/pubs/authors/policies/data\\_policy.shtml](http://www.agu.org/pubs/authors/policies/data_policy.shtml).
- [10] Instruction for Data Papers [EB/OL]. [2014-12-23]. <http://www.esapubs.org/archive/>.
- [11] 邱春艳.期刊文献与科学数据的关联服务研究[J].情报资料工作,2014(2):63-66.
- [12] Jonathan R. Recommendations for Independent Scholarly Publication of Data Sets [EB/OL]. [2014-12-23]. <http://sciencecommons.org/wp-content/uploads/datapaperpaper.pdf>.
- [13] Gorgolewski K J, Margulies D S, Milham M P. Making Data Sharing Count: A Publication-based Solution [EB/OL]. [2014-12-23]. <http://www.ncbi.nlm.nih.gov/pubmed/23390412/>.
- [14] Mennes M, Biswal B B, Castellanos F X. Making Data Sharing Work: The FCP/INDI Experience [EB/OL]. [2014-12-23]. <http://www.ncbi.nlm.nih.gov/pubmed/23123682>.
- [15] Xiaolei Huang, Bradford A. Hawkins, et al. Biodiversity data sharing: Will peer-reviewed data papers work? [J]. BioScience, 2013(1):5-7.
- [16] INCF Dataspace [EB/OL]. [2014-12-23]. <http://incf.org/resources/data-space/>.

[作者简介] 王丹丹,女,1980年生,河南科技大学管理学院副教授。  
收稿日期:2015-05-19