国外科学数据仓储的数据出版流程研究

王舒[1] 黄国彬[2]
(1. 山西财经大学图书馆，太原 030006；2. 北京师范大学政府管理学院，北京 100875)

摘要：科学数据仓储是未来科学数据出版的主导性媒介之一。本文以数据出版流程为切入点，从数据提交、数据存储、数据审核和数据发布4个方面对科学数据仓储的出版功能进行分析，试图为规范科学数据仓储的出版功能提供建议：建立以自助提交为主的提交模式，制定本仓储科学数据质量审核标准，施行自动审核与人工审核并行的质量审核方式，采用多渠道发布数据集，为数据集提供数字对象唯一标识符。

关键词：科学数据仓储；出版流程；科学数据；数据出版
中图分类号：G250 DOI：10.3772/j.issn.1673-2286.2021.01.009
引文格式：王舒，黄国彬. 国外科学数据仓储的数据出版流程研究[J]. 数字图书馆论坛，2021 (1)：60-66.

随着计算机技术和互联网技术在科技活动中的广泛应用，使得科学数据以惊人的速度增长，已成为当下增速最快的资源。与此同时，随着数据密集型科研范式的兴起，科学数据已经由科学研究的起点和基础发展成为科研活动的牵引力之一。面对上述情况，如何对科学数据进行有效的组织、共享和利用，成为科学界共同关注的热点问题。而科学数据出版被认为是有效解决这一问题的重要手段。2018年国务院办公厅出台的《科学数据管理办法》指出“主管部门和法人单位应积极推动科学数据出版和传播工作，支持科研人员整理发表产权清晰、准确完整、共享价值高的科学数据”。但截至目前，学术界对科学数据出版的定义还没有统一。笔者认为，科学数据出版是学术出版的一种，在将科学数据公之于众之前，需要对其质量进行审核，使发布的科学数据达到可发现、可获取、可理解和可追溯的状态。但与学术出版不同的是，科学数据只能通过网络出版，因此，科学数据仓储作为依托数字技术和网络技术建立的采集、保存、管理与发布科学数据的平台，在科学数据出版中显得尤为重要。基于此，本文以科学数据仓储为研究对象，对其出版流程进行剖析，总结科学数据仓储在出版科学数据中的经验，为科学数据仓储的功能设计者提供参考。

1 研究现状

近年来，国内外学者对科学数据出版的研究，可以归纳为3个方面。①对科学数据出版模式的研究。黄国彬等[1]从科学数据的产生情形，将科学数据出版模式归纳为科学数据集成出版与独立出版两种模式；张静蓓等[2]基于科学数据出版的国内外实践与研究现状，提出4种出版模式，包括数据独立出版、数据论文出版、期刊与指定数据仓储合作出版及期刊自行出版；涂志芳[3]认为虽然划分维度不同、模式名称表述存在差异，但仍然在一定程度上达成了共识，即作为论文附件的数据出版、独立的数据出版和数据论文3种模式。②对科学数据仓储的研究。科学数据仓储即科学数据的发布平台，大多数学者选取国内外典型的科学数据出版平台对其功能进行研究，国外学者多以某一个具体的科学数据仓储为例开展研究，如Roman等[4]介绍了科学数据仓储Data Graft数据转换、发布和托管功能等功能；Brase等[5]研讨了以世界数据中心（World Data Centers）的数据出版实践。而国内学者多选择国内外

多个典型的数据仓储为样本进行分析。如秦顺等[6]选取欧美地区14个科学数据出版平台，从科学数据出版政策或愿景，科学数据整合、标识与交互，科学数据出版与分发，科学数据引用，数据生命周期管理与出版质量控制5个方面进行分析；张玲玲等[7]选取中美具有代表性的8个地理科学数据仓储，从数据提交、同行评审、数据发布和永久存储、数据引用以及影响评价5个基本环节进行调研分析。屈宝强等[8]探讨了当前科学数据发布平台中存在的用户黏合度不高等问题。③对科学数据出版流程中的具体环节进行研究。如王丹丹等[9]对不同出版模式下的科学数据质量审核的实践、标准进行对比分析；李晓蕾等[10]对地质领域的科学数据的质量控制措施和公开化审查进行了分析。涂志芳等[11-12]认为科学数据仓储在数据出版过程中的质量控制实践还未成熟，我国数据知识库仍存在高度依赖计算机的辅助，可持续发展机制尚不成熟等问题。此外，有学者认为为科学数据分配数字对象唯一标识符（DOI）是科学数据出版的重要环节。吴立宗等[13]总结了DOI在数据出版领域的意义，并讨论它在数据出版与引用方面的不足。

综上所述，现有研究已取得了一定的进展，学者从不同角度对科学数据出版模式进行划分与研究，充分承认科学数据仓储在数据出版过程中的重要性，同时剖析科学数据仓储的功能、服务与存在的问题，对科学数据出版流程中的质量审核环节进行深入研究，但目前还没有学者对科学数据仓储的出版流程进行深入分析，尤其是没有涉及存储过程、发布时间、发布渠道等细节。因此，本文从数据出版流程的角度，对科学数据仓储的出版功能进行调研与分析。

2 科学数据出版流程剖析

科学数据仓储的数据出版功能是其面对数据生产者而设计的，实现该功能的内在逻辑是科学数据出版的流程，包括数据提交、数据存储、数据审核和数据发布。

2.1 数据提交

科学数据的提交方式主要有两种。一种是数据生产者自助提交。在开放获取潮流和数据共享理念的影响下，该模式成为科学数据出版中数据来源的主流渠道。另一种是工作人员协助提交。如美国高校社会科学联合会数据仓储（Inter-university Consortium for Political and Social Research，ICPSR）通过定期审查联邦资助机构数据库、学术期刊，关注专业的科学会议、参考会员机构和本机构工作人员建议等渠道收集数据。

2.1.1 数据生产者自助提交

数据生产者自助提交是由数据生产者本人将数据集存入科学数据仓储。数据提交的具体操作由数据生产者独立完成，但需要科学数据仓储提供完成数据提交所须的基础设施——在线提交平台和提交指南。

目前科学数据仓储提供的在线存储平台主要有两种。一是基于开源软件开发的存储平台，一部分是依托现有开源软件开发而成，如Dryad数据仓储、爱丁堡大学的DataShare等是基于开源软件DSpace开发而成；哈佛大学的Harvard Dataverse是基于开源软件Dataverse开发而成；另一部分是自建形成的开源软件平台，如Figshare均允许科研机构和出版机构在其基础上进行二次开发。二是由科学数据仓储自主开发的存储平台，他人无法在此基础上进行二次开发，如社会科学领域的英国数据存档（UK Data Archive，UKDA）、ICPSR、英国考古数据服务（Archaeology Data Service，ADS），地理环境科学领域的地球与环境数据出版平台（PANGAEA Data Publisher for Earth & Environmental Science，PANGAEA）、澳大利亚海洋数据网（Australian Ocean Data NetworkPortal，AODN Portal）、美国冰雪数据中心（National Snow & Ice Data Center，NSIDC），生物医学领域的ArrayExpress，化学物理领域的PubChem、剑桥晶体数据中心（Cambridge Crystallographic Data Centre，CCDC）等都根据本仓储的实际需求自主开发而成。然而，无论是自主开发的存储平台，还是基于开源软件二次开发的存储平台，都需要在提交指南的指导下使用。

编制提交指南，是科学数据仓储为数据生产者提供的另一个基础设施，通常与在线提交平台配合使用；是为了使数据生产者在自助提交数据时更好地使用在线提交平台。指南通常包括4个部分，即提交原因、提交准备、提交流程以及提交后对数据集的处理。其中，“提交原因”是帮助用户理解为什么使用该仓储，以及将数据集存储入该仓储的益处；“提交准备”旨在帮助用户在提交前准备数据集，包括描述数据集、规范数据集格式、剔除数据集中隐私数据等；“提交流程”是存储指南的核心内容，旨在帮助用户使用在线提交平

台；而“提交后对数据集的处理”是存储服务的后续工作，通常是指人工质量审核等。此外，存储指南的格式包括HTML、PDF、Video等。

2.1.2 工作人员协助提交

协助提交也是科学数据仓储常用的方式之一，即由科学数据仓储的工作人员协助数据生产者将科学数据存入仓储中。工作人员通常需要对科学数据进行评估以判断是否适合本仓储，对科学数据进行格式化调整以利于提交或保存，帮助数据生产者将数据上传至仓储。可将协助的环节分为评估环节、准备环节和提交环节。

评估环节是指工作人员依据一定的标准评估数据集是否适合或值得纳入该仓储。如英国环境数据分析中心数据仓储（Centre for Environmental Data Analysis-archive，CEDA）的评估环节由仓储工作人员依据“NERC数据价值清单”对科学数据的存储价值进行评估，包括科学数据的质量、完整性、原创性等，以评估数据集是否适合存储入该仓储中；若不适合，还会给出其他推荐的存储位置，如英国国家环境理事会（the Natural Environment Research Council，NERC）资助的其他科学数据仓储等[14]。又如癌症图片数据仓储（the Cancer Imaging Archive，TCIA），要求数据贡献者向TCIA提交数据存储申请，由其顾问小组（TCIA Advisory Group）进行审查，该小组由癌症成像和相关技术专家组成，每月审查一次数据提交申请，TCIA顾问小组依据审查标准和资源的可用性审查每个候选集合，并决定是否接受/拒绝或要求重新提交申请[15]。

准备环节是指数据集提交前所做的准备，包括制订数据提交计划、对数据集进行描述、规范数据集格式、确定数据集获取级别和使用条件、确定传递方式等。提供数据准备方面协助的科学数据仓储较多，如UKDA工作人员协助制订数据提交计划、确定数据获取级别和使用条件[16]。澳大利亚数据存档（the Australian Data Archive，ADA）由工作人员根据用户填写的数据集存储表和提供的相关文档（问卷、技术报告、相关出版物，以及其他有助于研究人员分析和理解数据的材料），对数据集进行描述[17]。ADS要求用户在提交数据前通过邮件或电话联系ADS数字存储管理员以确定数据传递方式等[18]。TCIA的审核人员协助用户对数据集进行去标识化处理与描述，确保数据使用者无法通过数据中包含的信息识别出被试人员，并与数据提交者一起创建数据集摘要。

①直接由工作人员完成科学数据提交，即要求科学数据贡献者通过一定的方式将数据集传递给仓储工作人员，再由工作人员将科学数据集存入科学数据仓储。如UKDA由数据贡献者通过埃塞克斯大学ZendTo服务（邮件）、邮递或者亲自递送的方式传送数据，由工作人员存入仓储；ADA要求数据贡献者通过邮寄、邮件等方式将数据集传递给工作人员后，再由工作人员将其存入仓储；ADS要求通过CD-ROM、便携式硬盘、电子邮件和云服务等方式传递数据，最终由工作人员存储数据集；CEDA根据数据集大小和复杂程度向用户提供不同的传递数据集的方式，最终由工作人员将数据集存储到仓储中；BioGRID[19]要求数据贡献者通过邮件向仓储工作人员发送一个包含科学数据的表格或纯文本文件，之后由工作人员将数据纳入BioGRIO；GenBank要求数据贡献者使用提交工具（Sequin、tbl2asn）对数据集进行格式化后，再由数据存储者通过邮件（或SequinMacroSend）将数据集发送给工作人员，由工作人员将数据集存储至仓储[20]。由上述案例可知，虽然每个仓储要求的传递数据集的方式不同，但最终数据集的提交均由科学数据仓储的工作人员完成。②工作人员帮助数据贡献者提交数据集。如ICPSR通过可移动介质（CD-ROM或DVD）将数据携带至物理提交场所，在工作人员帮助下将数据集复制到安全位置[21]。dbGaP要求数据提交者通过邮件与仓储工作人员联系，工作人员将提交链接发送给数据存储者，由数据提交者上传数据集[22]。③根据数据集大小、类型等因素提供不同的存储服务。如UKDA根据数据集大小来确定存储方式。科研人员的科学数据集，通常数据集较小，需采用自助存储方式，通过在线提交平台为ReShare存储数据；而大型调查项目或系列调查项目产生的数据集，通常数据集较大，因此需要仓储工作人员协助存储，仓储工作人员会依据相关政策对数据集进行评估，通过后，将其存入仓储[23]。

2.2 数据存储

2.2.1 数据存储格式

安全、可靠、高效的科学数据存储环境是科学数据仓储稳定运行和持续服务的前提。经数据生产者自助提交或由科学数据仓储工作人员协助提交后，科学数据仓储需要对各类科学数据，通过相关的科学数据元

数据框架，对科学数据进行描述、标引、分类和存储，以便为后续的科学数据检索与发现、科学数据的发布与引用、科学数据的分析与挖掘提供支撑。

科学数据仓储会以主题进行聚类，而主题聚类的维度，主要包括基于学科专业领域（物理、天文、地理等）、基于实验环境与科学数据创建方式（如实验获得、观测获得等）、基于科学数据的表现形式（如文本型、数据型等）等；同时，在存储格式上进行统一部署，包括：①针对以文本/电子表格格式呈现的科学数据，其可选用的存储格式有doc、docx、dot、rtf、txt、pdf、xls、xlsx；②针对以图形格式呈现的科学数据，其可选用的存储格式有bmp、jpg、jpeg、png、gif；③针对以结构绘图数据格式呈现的科学数据，其可选用的存储格式有cdx、c3d、cwg、csml、skc、xyz；④针对以音频格式呈现的科学数据，其可选用的存储格式有wav、pcm、tta、flac、au、ape、tak、wv、mp3、wma、ogg、aac；⑤针对以动画格式呈现的科学数据，其可选用的存储格式有avi、rmvb、rm、asf、divx、mpg、mpeg、mpe、wmv、mp4、mkv、vob、mov、flv、swf。

而在科学数据的存储方面，目前的科学数据仓储主要采用两种存储模式，即基于云端的科学数据存储以及基于本地的科学数据存储。从安全性来看，这两种存储模式各有利弊，需要科学数据仓储运行者根据自身的技术条件、服务对象、资金支持等进行综合权衡。对于科学数据生产者而言，如何选择一家可靠的科学数据仓储提供机构，对其后续在科学数据的管理与维护、传播与利用等方面，也是较为关键的一个问题。

2.2.2 数据唯一标识符

科学数据的科学合理组织与存储是实现科学数据被高效检索发现、进而被广泛引用与重用的基础。通过可靠的规则，赋予科学数据DOI，是实现科学数据后续开发与利用的关键一环。

数据唯一标识符是科学数据仓储赋予即将发布的科学数据的数字资源唯一标识，用于科学数据引用和重用，主要包括但不限于：DataCite为所有数字资源提供的DOI、个别科学数据仓储提供的入库编号、统一资源定位符（URL）等。其中DOI是最重要和常见用于科学数据引用的唯一标识，不仅能够唯一标识数字资源，还能形成链接形式，直接链接到数据集内容页面。而个别科学数据仓储提供的入库编号，虽然能唯一标识科学数据，但使用范围局限于本仓储内部，且不能形成链接的形式。这种唯一标识符常出现于学科科学数据仓储中，如dbGaP、ArrayExpress、CCDC。以CCDC为例对仓储编号进行说明：在数据提交3个工作日内，一个7位数的编号（CCDC4367857）会通过邮件发送给提交者，并确保通过这一编号，实现该科学数据与对其加以引用的期刊论文建立关联；该编号也可用于仓储中数据查询。URL是对可以从互联网上得到的资源位置和访问方法的一种简洁的表示，是互联网上标准资源的地址[24]。对于科学数据引用，该URL通常指向科学数据的内容页面。虽然点击URL，页面能直接跳转至数据集内容页面，但其长期稳定性远不如DOI。

2.3 数据审核

数据审核是科学数据出版的核心环节，不同科学数据的审核方式、内容、时间各不相同。

2.3.1 审核方式

目前科学数据仓储在开展数据出版服务过程中，对出版的数据集审核方式主要有人工审核与自动审核两种。人工审核是指科学数据仓储成立专门的质量审核工作组或安排专门的质量审核工作人员，在数据集提交前后对数据质量进行审核，如ADS成立数据评估工作组（Collections Evaluation Working Group）对数据质量进行审核；而PANGAEA会安排数据编辑（Data Editorial）来开展审核工作。自动审核是指在数据提交过程中，数据存储系统或集成到系统中的校验工具对上传的数据集质量进行审核。如Harvard Dataverse在数据提交过程中，由提交系统自动对数据集的格式、元数据进行审核，以确认数据集的运行状况和元数据的完整性。

2.3.2 审核内容

质量审核的内容包括数据集本身及其元数据。数据集质量包括技术质量与科学质量。

技术质量是指数据集本身的完整性、描述的充分性，对于含有个人隐私数据的科学数据，技术质量还包括数据集是否去标识化；而科学质量是指数据集收集方法的评价、科学数据的合理性和再使用的价值。目前

科学数据仓储对数据集本身的质量审核侧重技术质量。如PANGAEA直接明确数据集的科学质量由数据提交者负责，而仓储只负责审核科学数据的技术质量，主要包括数据集格式的正确性、数据集内容的完整性等[25]。Figshare系统自动对上传数据集的完整性进行审核，数据集的科学质量由数据贡献者负责，但若数据集涉及侵权（隐私权、知识产权）行为，该仓储有权删除[26]。dbGaP[22]和CCDC[27]都由系统对上传的数据集进行审核以保证数据集正确、完整地上传至存储空间，同时检查报告可供用户下载。

科学质量的审核主要有两种情况。一种是由科学数据仓储进行审核。如UKDA、ICPSR对数据集的内部质量进行审核，如对变量名称与变量值进行审查，对随机样本、均值方差、异常值进行检测等。NSIDC对于不同资助机构资助的科研项目产生的数据集的审核内容不同，其中，对由NASA资助产生的科学数据，审查内容包括科学价值、唯一性、归档和分发的成本等[28]。另一种是邀请外部人员对科学数据集的质量进行评审，外部人员是相对于仓储的工作人员而言，具体包括期刊论文的评审专家、数据使用者。例如Dryad，其合作期刊的同行评议人员在论文质量审核过程中对数据集的科学数据质量审核；BioGRID允许数据使用者指出数据集的错误，包括科学性方面的错误，并为用户提供专门的渠道来上报错误信息[29]。

此外，由于科学数据与学术论文、科技报告、科技图书等传统的科学文献不同，从形式来看可能是一组观测数值、实验数据记录、问卷数据或者一段计算机代码。如果不对其变量含义、产生背景、获取方法等进行描述，则无法掌握科学数据的具体含义。因此，除了对数据集本身进行审核外，还需对元数据进行审核。审核内容包括以下3点：①是否符合元数据标准，如ICPSR审核其数据集的元数据是否符合DDI元数据标准[30]；②是否与数据集信息相一致，如PANGAEA对元数据内容与数据集的一致性进行审查；③元数据字段是否完整，如ArrayExpress审核元数据是否缺少公开发布日期、用于测序实验的协议等[31]。

2.3.3 审核时间

质量审核的时间包括数据集提交前、数据集提交中与数据集提交后。具体选择在何时进行质量审核，与审核的方式密切相关。通常，自动审核发生在数据集提交过程中，这是由于在线提交系统往往自带审核功能或集成审核工具，如Harvard Dataverse的在线提交系统具有对数据集校验的功能，CCDC数据提交系统中集成了checkCIF/PLATON等工具供数据提交者对数据集进行校验。人工审核通常发生在数据提交前或数据提交后，通常数据提交前，工作人员对数据集内容是否适合该仓储、是否具有再利用价值等进行审核。例如，ADS在数据提交前，对数据集的再利用价值进行评估；而数据提交后，工作人员对数据集的格式、数据集及其元数据的一致性、完整性进行审核，如PANGAEA在数据提交后，对元数据和数据的完整性、一致性进行审核。

2.4 数据发布

2.4.1 发布渠道

数据出版的最终实现，是通过一定的渠道将其发布出来。不同科学数据仓储，数据集发布渠道不同。目前科学数据仓储的数据发布渠道包括本仓储的数据目录、相关期刊论文和集成数据目录。其中，本仓储的数据目录是主要的发布渠道，发布的信息一般包括数据集本身、元数据信息和使用许可协议。值得注意的是，不同的数据仓储其元数据的详略程度不同。通常情况下，专业型科学数据仓储的元数据信息较通用型科学数据仓储的元数据信息更加详细。

对有来源文献的科学数据，科学数据仓储通常将期刊论文作为发布数据的补充渠道。来源文献中需要注明数据集的存储地址和访问方式，以此来发布科学数据。如PANGAEA，其Web服务允许在论文页面上动态地嵌入数据信息。在这种方式下，来源文献可以帮助用户更好地理解数据集。

此外，集成目录也是科学数据仓储发布数据集的渠道之一，如CEDA允许科学数据的元数据被NERC的数据目录（NERC Data Catalogue）收割；EIDC允许科学数据的元数据被英国政府数据门户（data.gov.uk）和欧洲INSPIRE门户（EU INSPIRE portal）收割。通过集成目录发布数据集的元数据，是科学数据仓储的扩展发布渠道，增加了数据集被发现的可能性。

2.4.2 发布时间

不同科学数据仓储对科学数据的发布时间规定不

同。原则上，科学数据仓储鼓励和允许数据集在提交、审核后尽快发布。但允许在下列情况下，由科学数据提交者决定是否延迟发布，并且大多数科学数据仓储规定了延迟期限。

（1）将科学数据集的发布时间延迟至期刊论文见刊时间。通常在该情况下，数据集与其支撑的论文相伴而生，科研人员将论文提交至期刊，同时将支撑论文结论的数据提交至科学数据仓储，为保护论文作者的知识产权和期刊出版商的利益，科学数据仓储允许在论文见刊之时，再公开发布数据集。如在数据集提交至仓储中到来源文献见刊这段时间内，PANGAEA允许数据集预发布，意味着仅有作者和期刊论文的审核者通过密码访问该数据集，一旦期刊论文见刊，则数据集的状态由预发布改为正式发布。而Dryad允许数据集在期刊论文发表1年以后再发布，但前提是需要期刊编辑或出版商向本仓储提供书面协议。

（2）因包含敏感信息而延迟发布。对于以人体为研究对象的学科，其科学数据集通常会涉及被试个人信息。对于被试个人信息等敏感信息的处理，有些仓储实行匿名化处理后，即可进行发布；但有些仓储会因包含敏感信息而延迟发布，如对于因包含敏感信息而延迟发布的数据集，ADS会延迟发布时间长达70年。

（3）因资助机构要求而延迟发布。资助机构为保证研究者的利益，通常允许科学数据在产生2年后再发布。有些科学数据仓储为响应资助机构的要求，允许数据集提交至本仓储2年后再公开发布，如CEDA和EIDC，对于NERC资助项目产生的数据集，可以允许2年后公布。

（4）由数据提交者决定科学数据发布时间。如Harvard Dataverse为每个数据提交者提供用户个人空间（My Data），数据提交者可以将数据提交至此空间，具体何时发布数据集，由其自主决定。而ArrayExpress会在数据集发布的前60天、30天和7天通过邮件提醒数据提交者，数据提交者可对数据发布时间进行更改。

3 结语

科学数据仓储是科学数据出版的主导性媒介之一，调研国外各领域科学数据仓储的出版功能，并从出版流程的角度进行分析，总结出最佳实践，为科学数据仓储的建设者和功能设计者提供参考。①建立以自助提交为主，协助提交为辅的提交机制。在网络环境下，受开放获取潮流的影响，科研人员更习惯以自助方式将科学数据提交至仓储以备出版。仅当数据集文件过大或遇特殊情况时，需要由专门的工作人员协助提交。该提交机制可实现全天候24小时不间断服务，减轻工作人员的重复性劳动，为科学数据仓储节约人力成本和提高服务效率。②制定科学数据及元数据质量审核标准，保证其出版科学数据的内容完整、描述充分、格式适用性强。③设置专门的质量审核岗位，搭建质量审核系统，形成人工审核与系统自动审核相结合的方式，针对科学数据及元数据不同的审核内容，灵活采用适当的审核方式。④采用多渠道发布数据。科学数据仓储应尽可能多地扩展发布渠道，以增加科学数据被发现的可能性。科学数据仓储应开放元数据，允许被各大数据库搜索进而收割元数据，或主动提供元数据。此外，数据仓储还应明确要求数据使用者，在使用本仓储的数据所产生的学术出版物中引用该数据，并注明数据集的存储地址和访问方式。⑤分配数字对象唯一标识符。科学数据仓储应为每个数据集提供DOI，使数据集实现永久追溯，同时有助于学者引用该数据集。

参考文献

[1] 黄国彬，王舒. 科学数据出版模式比较研究 [J]. 大学图书馆学报，2018（1）：33-40.

[2] 张静蓓，任树怀. 科研数据出版模式、流程及引用策略研究 [J]. 图书情报工作，2015，59（9）：21-27.

[3] 涂志芳. 科学数据出版生态系统与质量控制体系构建 [J]. 图书与情报，2019（1）：125-134.

[4] ROMAN D，DIMITROV M，NIKOLOV N，et al. Datagraft：simplifying open data publishing [C] //European Semantic Web Conference：The Semantic Web. Berlin：Springer，2016：101-106.

[5] BRASE J，SCHINDLER U. The publication of scientific data by World Data Centers and the National Library of Science and Technology in Germany [J]. Data Science Journal，2006（5）：205-208.

[6] 秦顺，汪全莉，邢文明. 欧美科学数据开放存取出版平台服务调研及启示 [J]. 图书情报工作，2019，63（13）：129-136.

[7] 张玲玲，陈媛媛. 中美地理科学数据出版平台研究 [J]. 数字图书馆论坛，2020（10）：67-72.

[8] 屈宝强，宋立荣，王健. 开放共享视角下科学数据出版的发展趋势 [J]. 中国科技期刊研究，2019，30（4）：329-335.

[9] 王丹丹. 科学数据出版过程中的数据质量控制 [J]. 图书情报工

作，2015，59（23）：124-129.

[10] 李晓蕾，齐钒宇，孟洁，等. 地质科学数据出版的质量控制及公开化审查研究 [J]. 中国矿业，2019，28（6）：65-68.

[11] 涂志芳，刘兹恒. 我国多学科领域数据出版质量控制最佳实践研究 [J]. 图书馆杂志，2020，39（9）：70-77.

[12] 涂志芳，刘兹恒. 国外数据知识库模式的数据出版质量控制实践研究 [J]. 图书馆建设，2018（3）：5-13.

[13] 吴立宗，王亮绪，南卓铜，等. 科学数据出版现状及其体系框架 [J]. 遥感技术与应用，2013，28（3）：383-390.

[14] CEDA. Steps to archiving data with CEDA [EB/OL]. [2021-01-04]. https://help.ceda.ac.uk/article/138-steps-to-archiving-data-with-ceda.

[15] TCIA. Starting the submission process [EB/OL]. [2020-12-04]. http://www.cancerimagingarchive.net/primary-data/.

[16] UKDA. What you need to know to deliver a dataset [EB/OL]. [2021-01-04]. https://www.ukdataservice.ac.uk/deposit-data/how-to/regular-depositors/deposit.

[17] ADA. How To Deposit Data [EB/OL]. [2021-01-04]. https://www.ada.edu.au/ada/how-to-deposit-data.

[18] ADS. Guideline for Depositors [EB/OL]. [2021-01-04]. http://archaeologydataservice.ac.uk/advice/DepositingData.xhtml#How%20to%20Deposit.

[19] BioGRID. Contact Us/Send Us Your Data [EB/OL]. [2021-01-04]. https://wiki.thebiogrid.org/doku.php/contribute.

[20] Submitting Sequences using Specific NCBI Submission Tools [EB/OL]. [2021-01-04]. https://www.ncbi.nlm.nih.gov/books/NBK53709/.

[21] ICPSR. ICPSR：A Case Study in Repository Management [EB/OL]. [2021-01-04]. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/index.html#receipt.

[22] dbGaP submission process [EB/OL]. [2021-01-04]. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf.

[23] UKDA. How to deposit [EB/OL]. [2021-01-04]. https://www.ukdataservice.ac.uk/deposit-data/how-to.

[24] 百度百科. URL [EB/OL]. [2021-01-08]. https://baike.baidu.com/item/url/110640?fr=aladdin.

[25] PANGAEA. Benefits and Details [EB/OL]. [2021-01-04]. https://www.pangaea.de/submit/.

[26] Figshare. Data Integrity and Authenticity Policy [EB/OL]. [2021-01-04]. https://knowledge.figshare.com/articles/item/data-integrity-and-authenticity-policy.

[27] CCDC. Step 4：Validation [EB/OL]. [2021-01-06]. https://www.ccdc.cam.ac.uk/Community/depositastructure/structuredepositioninformation/.

[28] WEAVER R，DUERR R. Data Acceptance Plan [EB/OL]. [2021-01-06]. https://nsidc.org/sites/nsidc.org/files/files/data/daac/daac_data_policy_v09-1.pdf.

[29] BioGRID. Point out any Errors/Corrections to our Existing Data [EB/OL]. [2021-01-06]. https://wiki.thebiogrid.org/doku.php/contribute#point_out_any_errors_corrections_to_our_existing_data.

[30] ICPSR. Details on Appraisal Critera [EB/OL]. [2021-01-06]. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/details.html.

[31] ArrayExpress. Review by ArrayExpress curators [EB/OL]. [2021-01-06]. https://www.ebi.ac.uk/fg/annotare/help/submit_exp.html.

作者简介

王舒，女，1992年生，硕士，助理馆员，研究方向：数字资源建设，E-mail：bnuwangshu2018@163.com。
黄国彬，男，1979年生，博士，副教授，研究方向：信息法学、信息分析。

Foreign Research on Data Publishing Process of Scientific Data Repository

WANG Shu[1] HUANG GuoBin[2]
(1. Shanxi University of Finance and Economics Library, Taiyuan 030006, China;
2. The School of Government, Beijing Normal University, Beijing 100875, China)

Abstract: Scientific data repository is one of the leading media of scientific data publishing in the future. Based on the data publishing process, this paper analyzes the publishing of scientific data repository from three aspects: data submission, data storage, quality review, data release, and attempts to provide suggestions for standardizing the publishing function of scientific data repository: establishing the submission mode based on self-service delivery, formulating the quality review standard of scientific data repository, and implementing automatic audit in parallel with manual review, releasing data sets through multiple channels, and providing digital resource unique identifier for data sets.

Keywords: Scientific Data Repository; Publishing Process; Scientific Data; Data Publishing

（收稿日期：2021-01-08）