

科学数据引用流程中相关主体角色定位探析*

王思明 宁子晨

(东北师范大学信息科学与技术学院, 长春 130117)

摘要: 科学数据作为一种重要资源, 其引用问题得到广泛关注。本文尝试厘清科学数据引用流程中相关主体的角色定位, 为我国规范数据引用实践提供借鉴和参考。基于文献调研, 总结归纳出国际社会对科学数据引用过程中相关主体的定位, 即研究资助机构、数据中心、研究仓储、出版机构等。在此基础上对四大主体机构的典型代表网站进行网络调研, 分析其在实际活动中的具体引用行为。结合科研全生命周期, 最终将研究者、研究资助机构、数据中心、数据存贮库、出版商、图书馆界定为科学数据引用流程中的六大主体, 并阐释各主体在数据引用流程中的角色定位。通过明确引用流程中相关主体及其角色定位, 减少数据引用活动中的角色混乱和引用不规范等问题。

关键词: 科学数据; 数据引用; 数据中心; 数据存储库; 出版商; 图书馆

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2019.04.005

引用是知识传播的基石, 是评估研究质量和指导科学投入的主要手段, 也是学术架构的支柱之一^[1]。目前, 我们正在迅速转向科学的第四范式(即数据密集型科学研究)。数据和学术越来越多地交织在一起, 诸如数据学术(data scholarship)^[2]和数据密集型研究(data-intensive research)^[3]等新概念日渐成为学术界和科学界的中心。数据, 已经渗透到当今每一个行业和业务领域, 成为重要的生产要素^[4]。此外, 国际社会对数据透明化与数据共享日益强烈的需求, 以及全球范围内广泛兴起的关联数据运动、政府数据开放运动, 增加了数据资源的可获得性与可用性。科学数据引用越来越引起大家的关注, 然而在实际的学术活动中数据很少被正式引用。造成这种差异的部分原因是各个利益主体的角色定位不明确, 缺乏一致的建议及构建规范的工作流程。因此, 我们需要一致的指导方针, 指导各个相关主体之间紧密合作, 定义引用规范, 明确引用标准, 向用户提供引用数据集的明确指示。

1 国内外数据引用相关研究

以Data Citation和数据引用为关键词分别对Web of Science和CNKI进行文献检索。对Web of Science进行检索时将时间限定在2015—2018年, 因国内研究较少所以并未对时间进行限定。将得到的文献经过去重, 最终得到外文文献53篇, 中文文献33篇。对相关文献进行筛选阅读, 国内外对于数据引用的研究总结为以下方面。

1.1 数据引用动机

关于数据引用的动机, 国内外研究人员主要将其归纳为数据共享和研究成果的重现。典型的代表作者是Silvello^[1], 他在分析研究不同学科领域文献资源的基础上, 最终将数据引用的动机总结为数据归因、研究成果与研究数据连接、数据发现、数据共享、数据影响及数据重现6个方面。而国内学者邱均平等^[5]通过对科学数据被引情况的分析, 发现科学数据共享对引用频次

*本研究得到国家社会科学基金项目“图书馆书目数据社会化应用研究”(编号: 15BTQ018)和中央高校基本科研业务专项资金项目“科学数据引用行为及其影响因素研究”(编号: 130021822)资助。

以及成果影响力的提升均具有明显的促进作用,数据共享还有利于科研人员提升自身的影响力,进而推动科学的发展与进步。

1.2 数据引用规范

数据引用规范成为国内学者的主要关注点。其中,张静蓓等^[6]从引用对象界定、元数据构成、引用位置、关键问题(引用粒度、数据版本、唯一标识符)等方面重点探讨了数据引用规范的研究现状,指出科学数据引用规范是推动数据引用实践的核心组成部分。黄如花等^[7]选取国外关于科学数据的相关组织机构,调研了这些机构的科学数据引用规范,并从引用原则、引用元素、引用格式、引用对象、相关主体5个方面进行分析,同时倡导对于科学数据引用文化的发展,强调在整个生命周期创建科学数据引用文化的重要性。而邸弘阳等^[8]在对科学数据引用内容分析后,得出科学数据引用规范的内容特点包括:多遵循科学数据引用原则;受传统规范影响较大;不同的学科领域在引用对象、引用元素与引用格式等方面存在差异。

1.3 数据引用行为研究

国内外对于数据引用行为的研究较多针对各个学科领域展开。学者Berez-Kroeker等^[9]对41位语言学专家召开的三次研讨会进行研究,进一步探讨了关于语言学领域的科学数据引用和科学数据学术信用标准问题,指出学术信用的建立不仅仅存在于引用行为上,同时也要贯穿科学数据的创建、管理、存储等各个环节。学者Park等^[10]重点调查了遗传学中的数据共享与数据重用,尤其是在数据引用行为特征上,指出目前数据引用的挑战依然是正式引用的缺乏。国内对于数据引用行为的研究主要包括丁楠等^[11]对我国图书情报领域权威期刊2003—2013年的数据引用行为进行调查分析,指出数据引用在图书情报研究领域中的重要性正在逐步提升,但总体而言,数据行为较少、不规范引用问题大量存在。并说明造成这样问题的原因有:大量数据没有得到有效应用,已有数据的可获得性较差,以及缺乏规范的数据引用格式。随后屈亚杰等^[12]从被引科学数据内容的视角分析社会科学领域科学数据的引用现状与特点,发现在社会科学数据的被引行为中,调查数据类被引用次数最多。大部分被引用的社会科学数据存在时间

跨度短、规模小、更新次数少等特点。

1.4 相关研究述评

综上所述,目前国内外关于科学数据引用的研究主要集中在两个方面:一是通过引用动机分析以说明科学数据引用的重要性;二是运用内容分析、调查、访谈等方式调研不同学科研究人员、研究机构的数据引用行为与引用标准,为规范数据引用标准,建立最佳数据引用实践。这两方面的研究主要定位在对于微观的具体引用实践的研究。而从宏观角度出发,整合数据引用流程各个相关主体并明确主体角色定位的研究则较少,这导致引用标准难以规范化、统一化。因此,本文首先从明确数据引用流程中相关主体出发,确定相关主体之后,进一步对各个主体机构进行网络调研,依据其自身角色定位及其对数据引用标准的建设情况,尝试构建基于整个数据引用流程的相关主体角色定位体系,以此来保证数据引用活动的规范化、体系化,进而促进科学数据的使用及其价值的评估。

2 数据引用活动中的相关主体

Ball等^[13]于2015年指出,应该倡导研究中的数据与论文、专著存在同样的价值。然而,目前对于如何引用数据集以及如何在实际中构建正确的数据引用缺乏一致性的建议。在整个数据引用的活动中存在不同主体,他们因为各自的利益要求不同,创建了不同的数据引用准则。影响研究人员对数据进行规范化引用,进而阻碍了数据的共享与重用,同时也限制了对于数据引用行为的评估。因此,需要明确各个利益相关者的角色,协调各个主体密切合作;继而定义引文内容,向用户提供有关如何引用数据的明确指示。

2.1 ODE中的相关主体

ODE (Opportunities for Data Exchange) 是由APARSEN (Alliance Permanent Access to the Records of Science in Europe Network) 于2010年开展的关于推动数据共享项目。其在2012年发表的有关数据引用最佳实践报告中,指出了对于不同利益相关者对数据引用的机会与挑战^[14]。报告中,明确指出了数据引用活动中的相关主体有研究者、出版商、数据中心、图

书馆。同时总结了各个利益主体在具体的数据引用活动中的角色定位，如表1所示。

表1 ODE各个利益主体的数据引用角色

利益主体	角色定位
研究者	接受数据引用的协议
	遵循数据集的元数据标准
	使用持久标识符如DOIs
出版商	建立统一的数据引用标准
	遵循数据集的元数据标准
	使用持久标识符如DOIs
	数据出版物
数据中心	鼓励建立统一的数据引用标准
	支持并促进持久标识符的开发
图书馆	降低研究人员对数据再利用的障碍
	将数据集集成到检索服务中

2.2 CODATA-ICSTI中的相关主体

CODATA-ICSTI数据引用标准和规范任务组在2012年的数据引用实践、政策及技术现状分析的报告中指出，与数据引用关系密切的主体有国际科学组织、研究人员和研究机构、出版商、学术研究图书馆、研究资助机构^[15]。同时探讨了各个主体的角色定位，如表2所示。

表2 CODATA-ICSTI中各个主体的角色定位

利益主体	角色定位
国际科学组织	发布资助科学数据的原则和指南，以支持数据的共享与重用
	开发数据重用的通用标准
	开发和使用新的软件工具来自动化和简化数据集的创建和开发
研究人员和研究机构	接受数据引用的相关协议
	研究机构制定明确的研究数据管理政策辅助研究人员记录研究数据的完整及可用
	参与正式数据引用标准的创建
出版商	将期刊出版物与研究数据双向链接
	使用唯一标识符
学术研究图书馆	部分提供DOI登记
	支持并辅助研究人员将研究文章与数据存储空间之间建立联系
	作为数据引用的倡导者
研究资助机构	运用数据引用行为评估研究人员的价值
	制定强制数据管理计划
	将数据输出与相关论文的结果交叉连接起来

2.3 澳大利亚国家数据服务中心中的相关主体

澳大利亚国家数据服务中心（ANDS）在数据引用方面针对研究者建立相对完整的数据引用文化^[16]。通过查找并分析其发布的《数据引用指南》文件，发现其数据引用文化仅围绕统一标识符展开，共分为创建、使用、测量、奖励四个阶段。在四个阶段中，分别对起主导作用的利益相关者的主体进行了划分，包括科研人员、数据存储库、数据中心、研究社区、研究资助者及研究团体。5个主体在数据引用中的具体活动定位如表3所示。

表3 ANDS引用文化中的相关主体定位

利益主体	角色定位
科研人员	创建研究数据集及与数据集相关的出版物
	获取一个持久标识符（如DOI）
数据存储库	存储可访问的数据集
数据中心	给予数据持久标识符（如DOI）
研究社区	使用DOI访问数据集并开展相关研究
	使用DOI来引用数据集生成的新的出版物
研究资助者及研究团体	审查出版物和数据集引用指标
	确定研究人员未来的资助和推广数据引用指标的影响

通过调研ODE、CODATA-ICSTI、ANDS对于数据引用文化中的相关主体定位发现，3个机构所界定的主体各不相同，并未形成清晰的角色定位。同时，对于相关主体的探讨与研究，仅是分散谈论，并未将各个主体的角色明确定位在整个数据引用的流程中。这导致主体之间协作性较差，难以建立统一的规范标准，将无形中加大研究人员的工作量。所以，建立明确的数据引用文化，给予清晰的数据引用流程以及流程中各个阶段的相关主体角色定位，才能保证数据引用活动的展开。

3 数据引用流程及相关主体的角色定位

研究在前期不同组织对于相关主体的划分基础上，初步将与数据引用活动密切相关的主体确定为研究资助机构、数据中心、研究人员、数据存储库、出版机构、图书馆。如表4所示，调研这6个主体的典型代表官方网站，查找其具体的要求及实践的发展。最终构建数据引用流程及在流程中各个阶段的利益主体的角色定位，创建体系化的科学数据引用文化。

表4 相关主体网站调研

相关主体	代表机构	
研究资助机构	ODCE (Organization for Economic Co-operation and Development) ; RDA (Research Data Alliance)	
数据中心	DataCite; CoData (the International Council for Science's Committee on Data for Science and Technology) ; DCC (Digital Curation Centre) ; WDC (World Data Center) ; 加拿大汤森路透公司的数据引文索引 (DCI)	
数据存储库	地理科学	PANGAEA (Data Publisher for Earth & Environmental Science) ; ESIP (The Federation For Earth Science Information Partners)
	生命科学	GBIF (Global Biodiversity Information Facility) ; Dryad
	社会科学	Dataverse; 大学间的政治和社会研究联盟 (ICPSR)
出版机构	数据出版商	国际科学、技术和医学出版商协会; 开放存取科学出版商协会; Clarivate Analytics
	数据期刊	Nature集团出版的 <i>Scientific Data</i> ; Wiley集团出版的 <i>Geoscience Data of Journal</i>

3.1 数据引用流程及各个利益主体的角色定位

研究人员对科学数据的提交、科学数据的获取和数据引用行为评估三个阶段。明确各个阶段中相关主体的角色定位,如图1所示。

基于以上研究初步将整个科学数据引用活动分为

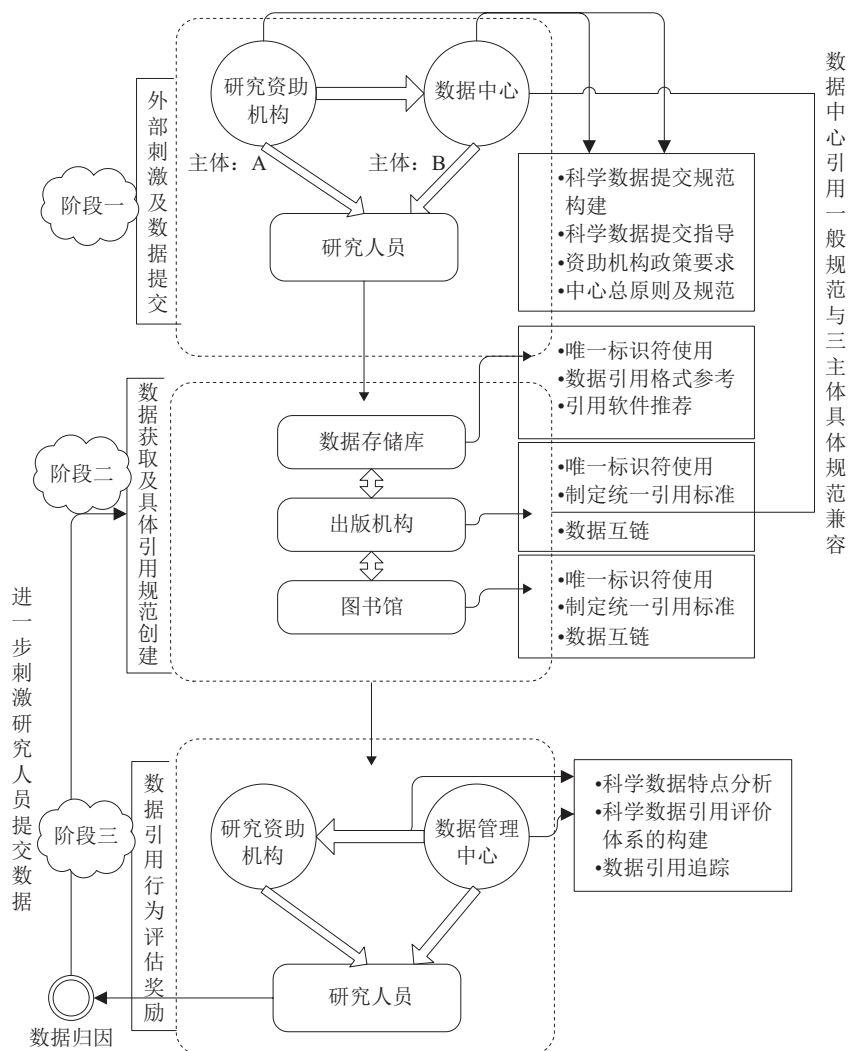


图1 科学数据引用流程及相关主体角色定位

3.2 具体阶段分析

3.2.1 科学数据提交阶段

在科学数据的提交阶段,主要责任主体有研究人员、研究资助机构、数据中心。研究资助机构作为科学数据生产者的外部刺激动力,在政策上要求并指导研究人员的整个科学研究活动。数据中心在此阶段的作用是与研究资助机构的政策保持一致,制定数据引用的总原则及一般规范,同时协调、引导、辅助、规范第二环节中的3个主体相关活动。

研究资助机构与数据中心合作,共同创建适用于科研人员的科学数据提交规范,如图2所示。但在数据

中心行为方面,本文将存储机构及出版机构的数据引用规范放在其子集部分。目的是说明数据中心所提供的仅是一般性的规范与指导,它们并不完全满足不同科学领域的数据库及出版商等机构结合,确保所制定的指导方针与更广泛的社区实践保持一致,为研究人员提供兼容与通用的数据引用规范。

研究人员在数据提交阶段的主要行为有:为自己的数据选择合适的存贮路径,保证数据的可获取;根据存贮机构的要求运用元数据对数据进行科学合理的描述,保证数据的再利用、可再生;确保数据获得唯一标识符。

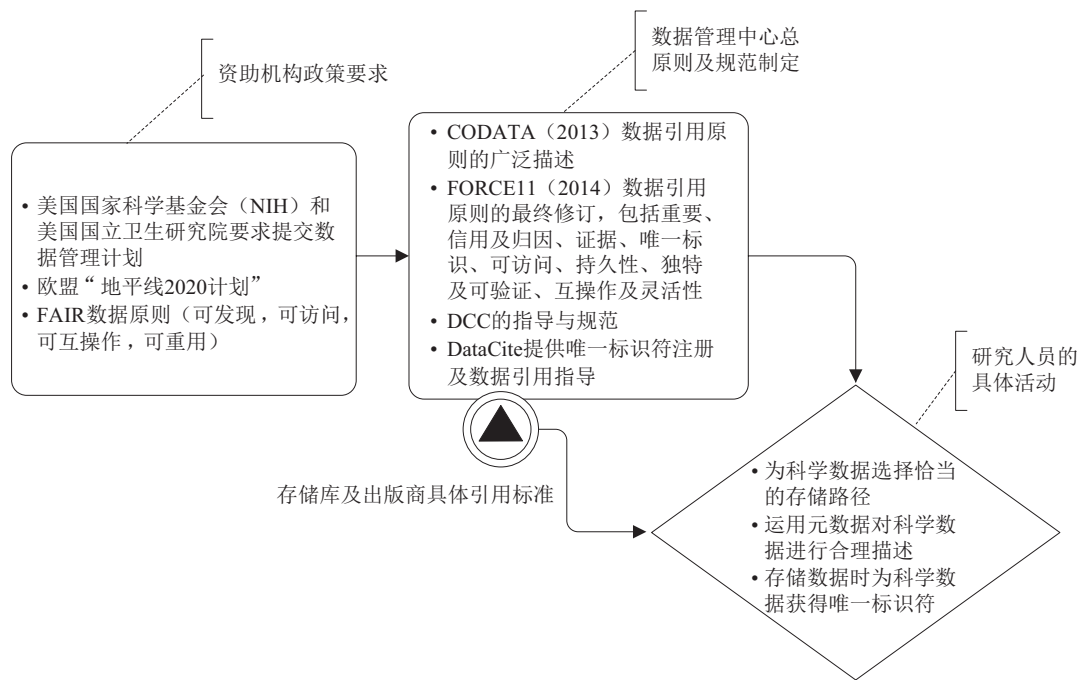


图2 国外科学数据提交阶段相关主体及提交规范

3.2.2 科学数据获取阶段

数据获取位于整个数据引用流程的第二个阶段,可谓流程的主要环节,也是与研究人员直接相关的活动。本文将数据获取主体限定在科学数据存储库发布、出版机构发布、图书馆。研究人员在受到科研机构激励及数据中心的原规范下,将科学数据提交到阶段二的3个主体中。图3给出了数据获取阶段相关主体及其引用行为。

在数据获取阶段,科学数据库及出版机构常作为

科学数据发布的主要主体,图书馆只是偶尔充当数据存储的角色。更多的情况下,图书馆只是位于数据存储或出版机构与研究人员之间的桥梁,充分利用与研究人员紧密关系这一巨大优势,排除研究人员与存储机构和出版机构之间相关联的障碍。数据获取阶段各个相关主体的引用行为主要包括以下4个方面。

(1) 在元数据描述方案制定方面,数据存储库或出版机构提供元数据在一定程度上可以说是数据中心现有方法的扩展或详细的概要。所以,在为用户提供科学数据的元数据描述、帮助用户提交符合规范的数据

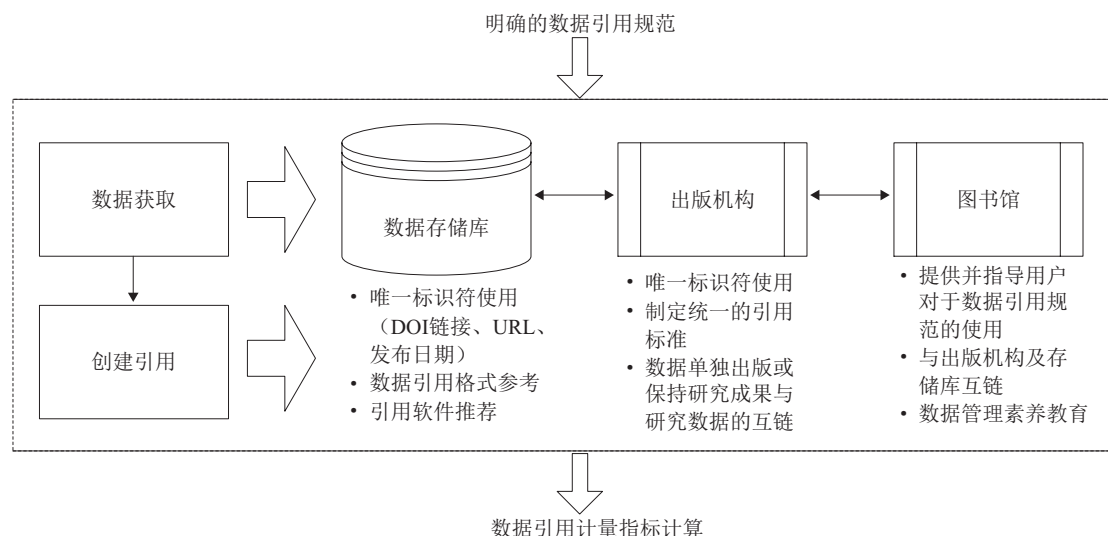


图3 数据获取阶段相关主体及其引用行为

时,要确保所使用的元数据与数据中心提供的元数据计划兼容。扩展并定制符合自身要求的数据提交方案,并向有提交科学数据意向的研究者提供方案,以此指导及规范研究人员的数据提交活动。

(2) 在统一标识符的赋予方面,无论是存储机构还是出版机构在此环节要特别注意原始科学数据与研究成果的互链,以此保证科研人员能找到完整的科学成果,真正实现科研成果的共享与重用。同时在具体的数据引用实践上,存储机构和出版商通过为研究人员提供引用示例,帮助数据使用者更好地创建数据引用格式,值得注意的是,示例要进一步明确作者归属、为科学数据配备良好的版本跟踪、数据集粒度以及提供全面的文档描述等问题。尽量为用户提供所有应该包含在数据集引用中的重要元素,以此来帮助研究人员明确清晰地开展正式数据引用活动。此外,还可以结合自动化工具创建引用,以减少用户的工作量。

(3) 在科学数据引用行为追踪方面,数据存储库及出版商要保证对于科学数据引用行为的追踪。与数据中心紧密合作,以采用一定的指标来衡量数据引用产生的影响,评估数据引用的价值及为数据的提交者建立数据归因。

(4) 在学科领域科学数据特点把控方面,关注学科领域科学数据特点的差异性,需要具体的学科数据库、期刊出版商及图书馆通力合作。采用一种合理严谨的方法,为本科学领域数据提供一个有用和精确的引用规范。

3.2.3 科学数据引用评估阶段

建立合理完善的数据引用评估,是目前亟需解决的问题。完善的数据引用评估不仅可以为提交数据的研究人员建立信用归因,以承认正式的数据贡献,从而进一步提升研究人员公开科学数据的意愿,形成良性循环;也是研究资助机构评估资金价值及考虑可持续供资的主要手段;还可以为数据中心开展相关工作提供指导与评价。但目前国内外对于科学数据引用的评价尚处于摸索阶段,并未形成系统有效的评价体系。给定一个数据集,我们仍然没有任何定量的、一致的和既定的方法来知道它是如何被使用的。考虑到数据引用评估的重要性及迫切性,目前的首要任务应该是充分结合科学数据的典型特点,明确科学数据评估中的相关主体的责任与义务,为系统化的科学数据引用评估体系奠定基础。

由于数据具有复杂性、易变性、不稳定性等特征,其与典型的传统出版物有很大的区别。数据很难被解释为一个独立的实体。如果没有相关上下文信息、软件和相关对象的链接,大多数数据都没有意义。因此,对于数据引用评估首要的任务是保证数据的唯一性、稳定性。这就需要研究资助机构与数据中心的合作,通过明确数据唯一标识符及数据粒度、版本等问题,保证数据的相对稳定性。当数据处于相对稳定的状态时,随即数据存储机构、出版商、图书馆三者结合对科学数据引用状况进行追踪。现存的数据引用评价方法主要有引文分析方法及作为补充的Altmetrics方法(下载量、浏览量,

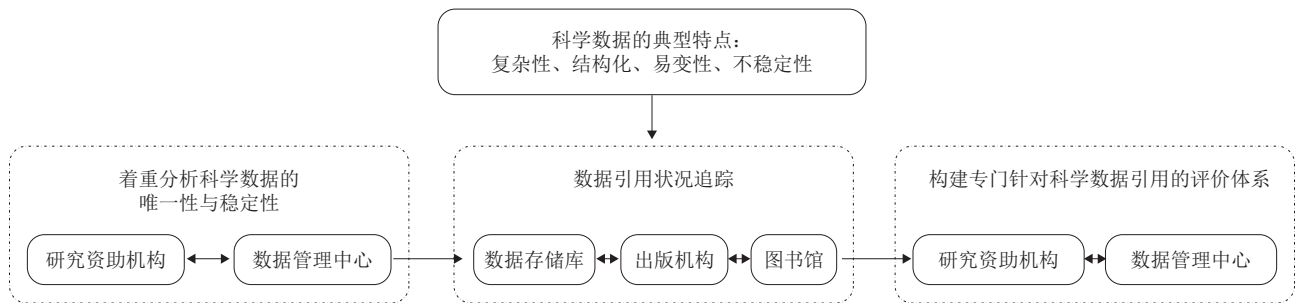


图4 引用评估阶段相关主体及评估流程

以及分享、保存、推荐、评论、标签等指标)。在对科学数据引用状况追踪环节中,研究人员要保证采用正式规范的数据引用格式。这样存储机构、出版商、图书馆才能对数据的使用情况保持持续的记录与追踪。

最后,研究资助机构及数据中心结合数据引用追踪阶段3个主体的具体活动,进一步构思与设计完善的科学数据度量标准,从而构建出专门针对科学数据引用的评价体系。使数据生产者、数据管理人员、研究机构和学术机构、科学出版商都能以可信、有价值和有意义的方式接受。进而以定量和正式地展示科学数据的意义和可行性,促进科学的进步。

4 结语

目前,我国的科学数据引用标准已经出台,无论是在唯一标识符的要求与运用,还是在科学数据引用元素的选择与描述方法,或是科学数据通用引用格式的引导,都给出了明确的解释与说明。本文基于数据引用流程的不同阶段,将不同主体的具体角色在数据引用流程的各个阶段进行定位,进一步明确其责任与定位,根据科学数据引用标准进一步设计出符合其角色的数据引用系统,从而更好地开展正式科学数据引用活动。但要将各个利益相关主体共同作用在一个统一的工作流程中,需要使用不同的渠道直接将相关主体统一起来,如采用研讨会、专题讨论会和各种学术交流活动等方式。最终形成科学数据引用文化,实现科学数据的共享与重用,促进科学及其应用的总体进展。

参考文献

- [1] SILVELLO G. Theory and practice of data citation [J]. Journal of the Association for Information Science and Technology, 2018, 69 (1): 6-20.
- [2] BORGMAN C L. Big Data, Little Data, No Data: Scholarship in the Networked World [M]. Boston: The MIT Press, 2015.
- [3] HEY T, TREFETHEN A E. Cyberinfrastructure for e-Science [J]. Science, 2005, 308 (5723): 817-821.
- [4] 魏来, 高希然. 大数据背景下高校数据馆员的角色定位 [J]. 情报资料工作, 2015 (5): 90-94.
- [5] 邱均平, 何文静. 科学数据共享与引用行为的相互作用关系研究 [J]. 情报理论与实践, 2015, 38 (10): 1-5.
- [6] 张静蓓, 田野, 吕俊生. 科学数据引用规范研究进展 [J]. 图书与情报, 2014 (5): 100-104.
- [7] 黄如花, 李楠. 国外科学数据引用规范调查分析与启示 [J]. 图书馆学研究, 2016 (10): 2-9.
- [8] 邸弘阳, 耿骞, 黄国彬, 等. 科学数据引用规范的内容特点分析 [J]. 数字图书馆论坛, 2017 (6): 9-15.
- [9] BEREZ-KROEGER A L, GAWME L, KUNG S S, et al. Reproducible research in linguistics: A position statement on data citation and attribution in our field [J]. Linguistics, 2018, 56 (1): 1-18.
- [10] PARK H, WOLFRAM D. An examination of research data sharing and re-use: Implications for data citation practice [J]. Scientometrics, 2017, 111 (1): 443-461.
- [11] 丁楠, 丁莹, 杨柳, 等. 我国图书情报领域数据引用行为分析 [J]. 中国图书馆学报, 2014, 40 (6): 105-114.
- [12] 屈亚杰, 王亚男. 社会科学领域科学数据的引用现状与特点分析 [J]. 数字图书馆论坛, 2017 (6): 25-31.
- [13] BALL A, DUKE M. How to cite datasets and link to publications [EB/OL]. [2019-04-22]. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_Cite_Link.pdf.
- [14] KOTARSKI R, REILLY S, SCHRIMPF S, et al. Walshe K (2012). Reports on best practices for citability of data

- and on evolving roles in scholarly communication [R/OL]. [2018-08-05]. <http://epic.awi.de/31396/>. [2019-04-22]. https://www.researchgate.net/publication/264934269_Out_of_cite_out_of_mind_The_current_state_of_practice_policy_and_technology_for_data_citation.
- [15] CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of Data [R/OL]. [2018-09-15]. <http://www.ands.org.au/guides/data-citation-awareness>.

作者简介

王思明, 女, 1989年生, 硕士研究生, 研究方向: 信息服务, E-mail: 1310141065@qq.com。
宁子晨, 女, 1994年生, 硕士研究生, 研究方向: 信息服务, E-mail: ningzc317@nenu.edu.cn。

Analysis of the Relevant Subject Roles in the Data Citation Process

WANG SiMing NING ZiChen

(School of Information Science and Technology, Northeast Normal University, Changchun 130117, China)

Abstract: According to the current research on relevant subjects in data citation activities in international organizations, the role of relevant subject roles in the scientific data citation process is clarified. Using the method of literature research, it summarizes the relevant subjects in the process of citation of scientific data at present, and conducts network research on some typical subjects. The relevant subjects in the scientific data citation process are initially identified in the researcher, research funding agencies, data centers, data repositories, publishers, libraries, and their role in the data citation process.

Keywords: Scientific Data; Data Citation; Data Center; Data Repository; Publisher; Library

(收稿日期: 2019-03-27)

■ 书 讯 ■

《汉语主题词表》

《汉语主题词表》自1980年问世以后,经1991年进行自然科学版修订,在我国图书情报界发挥了应有作用,曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要,由中国科学技术信息研究所主持,并联合全国图书情报界相关机构,自2009年开始进行重新编制工作,拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表(工程技术卷)》共收录优选词19.6万条,非优选词16.4万条,等同率0.84,在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表(自然科学卷)》共收录专业术语12.4万条,包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域,收词系统、完整,语义关系丰富、严谨,每条词汇都有相应的学科分类号表现其专业属性,并与同义英文术语对应。同时,建立《汉语主题词表》网络服务系统,提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘,是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版,分为13个分册,总定价3 880元。

《汉语主题词表(自然科学卷)》已于2018年5月由科学技术文献出版社出版,分为5个分册,总定价1 247元。两卷均可分册购买。