国外科学数据影响力研究进展*

王毅萍[1,2] 马建玲[1]
[1] 中国科学院兰州文献情报中心 兰州 730000 [2] 中国科学院大学 北京 100049

摘要: [目的/意义] 旨在分析国外科学数据影响力的研究内容与存在的问题,为我国在该领域的研究提供参考。[方法/过程] 以国外相关重要机构研究及项目成果调研为主,以 Web of Sciences、Google Scholar 数据库文献调研为辅,采用文献追溯方法,试图对国外科学数据影响力研究状况进行调研、梳理与总结。[结果/结论] 目前,科学数据影响力研究已经得到国外学术界的关注,其研究内容大致包括科学数据影响力的内涵、类型、关系、相关主体、评价方法五类,但其整体研究仍处于初级阶段。在借鉴国外相关经验的基础上,对我国科学数据影响力未来的研究发展提出培养数据引用意识与文化、加强基础理论研究与特征指标研究的建议。

关键词: 科学数据 影响力 评价
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2017.07.017

1 引言

随着计算机技术、观测探测等科学技术的进步,科学数据规模呈指数增长,科学研究已经从计算科学向数据密集型知识发现的第四研究范式转变[1]。科学家不仅通过对大量数据进行实时、动态的监测与分析,以解决难以解决或不可触及的科学问题,更是把数据作为科学研究的对象和工具,基于数据来思考、设计和实施科学研究[2]。可以说在大数据时代下,科学数据是推动学术研究发展与进步最为基础与重要的内容。在这样的背景下,学者们开展了大量关于科学数据共享、出版、引用、质量等方面的研究,为更广泛、更深入地利用科学数据创造有利条件。但目前存在科研人员数据共享积极性不高、引用意识薄弱等问题,严重制约了科学数据的共享与交流。除此之外,科研工作者对于高质量高水平科学数据的需求,资助机构为未来的资助资金流向进行评估判定的需求,数据存储与出版组织对于提高其组织影响力等的需求都激发出对科学数据影响力进行研究的必要性。

目前国内暂无机构对科学数据影响力问题进行专门研究,只有个别学者对相关内容进行了少量研究,例如中国科学院文献情报中心的顾立平对数据级别计量的概念、发展、应用、局限及前景进行了概要介绍[3];浙江大学图书馆丁楠等人试图构建基于引用的科学数据评价体系,主要选取包括数据发布量、数据被引量、数据平均被引频次及 h 指数 4 项引文指标进行分析[4];西华大学图书馆彭国莉等人利用数据引文索引(Data Citation Index,简称 DCI),选取社会科学数据的记录数、被引记录数、被引次数、h 指数作为评估指标[5],同馆邢红梅等人也进行了类似研究,选取有效性引用频次、二次引用数量、最近引用时间作为统计指标,来分析社会科学数据的学术影响力及价值[6]。从整体上看,国内对于科学数据影响力的关注度不足、研究维度单一、研究深度过浅,评价指标运用时大都套用论文评价指标,对科学数据自身特色考虑不足。

与国内相比,国外相关的机构学者已经从多个方面开展了关于科学数据影响力的研究,但由于科学数据影响力整体研究时间较短,相关基础理论还处于摸索阶段,而且对科学数据影响力的深入探究需要建立在较为完善的数据出版、引用机制之上,目前国外关于科学数据影响力的研究仍处于初级阶段,尚未形成系统有效的评价体系,但仍对我国的相关研究有重要借鉴意义。

* 本文系中国科学院文献情报能力建设专项"全球变化知识资源中心建设"(项目编号: Y6ZG431001)研究成果之一。
作者简介: 王毅萍(ORCID: 0000-0003-0299-7476),硕士研究生;马建玲,信息系统部副主任,研究馆员,硕士生导师,通讯作者: E-mail: majl@lzb.ac.cn。
收稿日期: 2016-12-30 修回日期: 2017-03-19 本文起止页码: 118-126 本文责任编辑: 王善军

2 文献调研来源及研究方法

由于该领域研究发展时间较短，可在 Web of Science 数据库中检索到的相关文章较少，因而以国外相关重要机构及其项目的文献调研为主，这些机构和项目主要包括科学数据联盟(Research Data Alliance，简称 RDA)、世界数据系统(World Data System，简称 WDS)、科研管理信息标准推进委员会(The Consortia Advancing Standards in Research Administration Information，简称 CASRAI)、全球生物多样性信息系统网络(Global Biodiversity Information Facility(GBIF) network)等国际性组织，英国的数字管理中心(Digital Curation Centre，简称 DCC)、卓越研究框架(Research Excellence Framework，简称 REF)，美国的国家信息标准委员会(National Information Standards Organization，简称 NISO)、数据可计量项目(Make Data Count，简称 MDC)、国际政治与社会研究联盟(Inter－University Consortium for Political and Social Research，简称 ICPSR)，欧洲知识交流(Knowlege Exchange，简称 KE)，加拿大汤森路透公司的数据引文索引(DCI)，澳大利亚国家数据服务(Australian National Data Service，简称 ANDS)等。同时以 Web of Sciences、Google Scholar 数据库文献为辅，采用文献追溯方法，试图对国外科学数据影响力研究状况进行调研、梳理与总结。全文主要从科学数据影响力的内涵、类型、关系、相关主体、评价方法五方面介绍了国外的研究状况，并对未来国内的相关研究提出建议，旨在为我国后续科学数据影响力研究的开展提供参考。

3 科学数据影响力的内涵

科学数据作为科研成果的一种，是指在科研活动中，通过观测、探测、实验、调查、实践等活动产生的原始事实记录，以及按照不同需求，进行系统加工整理形成的数据集[7-8]。

影响力是一种改变支配他人思想和行为的能力，《牛津词典》将影响力定义为“Marked effect or influence”[9]，即标志着产生效果与影响。影响力的产生是一种传播交流的过程，其产生效果可分为正面影响与负面影响，针对同一事件不同主体对其影响力关注的方面有所不同，因而，影响力也是一个相对的概念，此外，不同领域影响力的具体内涵也存在差异。

在科研成果评价领域，REF 将影响力定义为对学术界、经济、社会、文化、公共政策与服务、生活环境与质量产生的改变与提高[10]；澳大利亚的研究质量框架(Research Quality Framework，简称 RQF)同样采用了类似的定义，其更多关注科学数据的广泛影响力，而不局限于学术影响力，认为影响力是应用研究成果对社会、经济、环境、文化所产生的有益影响[11]。

因而，从广义角度看，科学数据影响力是指应用科学数据，对学术、社会、经济、文化、公共政策、环境等多方面产生的有益影响。具体而言，科学数据影响力主要可分为学术影响力与社会影响力，科学数据的学术影响力主要是科学数据在学术领域传播交流的广度与深度，是科学数据对相关领域研究所产生积极影响范围和深度的度量；科学数据的社会影响力主要是科学数据作用于社会，对社会、经济、文化、环境、生产实践发展等带来的有益影响及改变。

4 科学数据影响力的类型

目前科学数据影响力主要分为学术影响力和社会影响力两大类，如表 1 所示：

表 1 科学数据影响力类型

科学数据影响力	类型
学术影响力	数据重用频次
	重用数据后出版物质量
	重用数据后出版物多样性
	源于单个数据集的相关网络规模
	数据集下载量
	科学新发现新创造
社会影响力	全球经济绩效的促进
	公共服务以及政策有效性的增强
	健康、生活质量以及创造性产出的提高

学术领域通常是各项科研成果最早传播流通的领域，学术影响力通常是其影响力的最初表现。目前科研成果评价研究，更多地也是围绕科研成果的学术影响力展开的，相关领域已经较为成熟。科学数据作为一种新兴的评价对象，对其学术影响力的研究也受到了很多学者的关注。

美国学者 K. Fear 在对科学数据影响力进行评价时，从科学数据重用的广度和深度角度出发，认为科学数据影响力可分为 5 种类型，包括数据重用频次、重用数据后出版物的质量、重用数据后出版物的多样性、源于单个数据集的相关网络规模以及数据集的下载数量[12]。这 5 项内容主要为科学数据学术影响力的表征形式，其中数据重用频次可以通过数据的引用来计量；重用数据后出版物的质量及多样性可以通过出版物被引状况以及出版物涉及领域范围来体现，同时应注意不同学科领域出版物被引状况存在的差异；科学

数据的下载数量是数据受欢迎程度的体现;单个数据集相关网络规模是对科学数据传播范围、广度的评估,可运用图论、数据挖掘的方法来分析计量。此外,也有学者认为科学数据学术影响力还包括基于科学数据分析所产生的科学新知识、发现、成果、产品与服务等[10]。

随着对科研成果影响力研究的不断深入,很多机构已经注意到科研成果在社会领域产生的广泛影响。英国研究理事会(Research Councils UK,简称RUCK)将科学研究的社会影响力归纳为3种,包括对全球经济绩效的促进,对公共服务以及政策有效性的增强,对健康、生活质量以及创造性产出的提高。生物科技与生物科学研究委员会(Biotechnology and Biological Sciences Research Council,简称BBSRC)在其基础上认为科学研究影响力覆盖范围更加广泛,涉及社会领域的方方面面,并将科学研究的社会影响力归纳为图1[13],科学数据作为科学研究中重要的组成部分,其产生的社会影响力同样如此。

图1 BBSRC关于科研成果社会影响力的图示

目前,已经有机构开始对科学数据的社会影响力进行研究。例如,ANDS十分重视对科学数据社会价值的研究,在其两个重要报告《数据供给的成本与收益》与《开放科学数据报告》中,都对科学数据的社会价值进行了探讨[14];同时,ANDS目前正在开展科学数据影响力运动,面向社会征集那些可被证实的、能展现科学数据研究推动澳大利亚进步、益于其发展的故事,旨在对科学数据的社会影响力进行更为深入的了解[15]。

除此之外,科学数据影响力类型按照不同的特点、性质还有很多种分类方法,对科学数据影响力类型的研究有助于对科学数据影响力的本质、规律进行把握,为后续相关研究的开展打下基础。

5 科学数据影响力与相关要素间关系研究

科学数据的影响力研究是科学数据整体研究中的一部分,它与科学数据内部的其他研究有着紧密的联系。数据共享与数据引用是对科学数据影响力进行评价的基础;数据出版方式影响了科学数据影响力评价方式的选取;科学数据影响力的大小在很大程度上又与数据质量相关。下面笔者将对这几组关系的国外研究进展进行介绍。

5.1 科学数据影响力与数据共享间的关系

数据共享是指个体或机构自愿给其他具备合法科研目标的个体或机构提供数据信息的行为[16],它是科学数据领域包括数据质量、引用等在内研究的基础,更是科学数据发挥其影响力的前提和必要条件,同时科学数影响力研究对数据共享的进一步深化有着促进作用。一方面科学数据只有在广泛意义上达成共享、开放的共识,才能让其他专家学者发现、重用该数据,才能促进数据间的交流,进一步发挥科学数据的价值,提高科学数据的影响力。另一方面科学数据影响力的研究与评价,有助于对数据责任者产生激励,进而提高其共享数据的热情,加快科学数据的流动,进一步推动科学事业的发展。

K. M. Fear在其研究科学数据重用影响力计量与预测的博士论文中,认为数据共享与数据影响力之间有着紧密关系,并对当前数据共享的激励与抑制因素、数据共享后科学数据进行数据重用的方式与效果进行了分析,试图对更可能得到重用的数据集进行识别,以便相关学者共享更多有重用需求的科学数据[17]。

KE(Knowledge Exchange)是欧洲五个重要国家组织的合作组织,旨在开发基础设施与服务以促进数字技术的使用,进而提高高等教育与研究。在其《科学数据价值》报告中指出科学数据计量发展是数据共享的潜在刺激,可以将科学数据的相关计量整合为一个合理、专业的奖励框架,进而促进数据共享。同时报告中考虑了科学数据计量、数据共享发展中各利益相关者的不同观点、问题、挑战以及需要考虑的多种因素[18]。

相关机构学者普遍认识到了科学数据影响力与数据共享间的密切关系,达成科学数据影响力研究需要数据共享支持的共识,但在实际情况中数据共享程度并不高。以基因组学为例,基因表达微阵列数据集在数据仓储中的存储率很低,不足整体的50%[19]。在未来发展中需要不断提高科研工作者的数据共享热情,

为科学数据影响力评价研究打下基础，同时也需要通过科学数据影响力研究的发展为数据共享提供有效激励。

5.2 科学数据影响力与数据出版间的关系

数据出版是实现数据重用的必经途径，科学数据出版形式的不同会影响到其影响力的评价方式。尽管当前学者就数据出版的定义、模式还存在很多争议，但从主流观点来看，目前最主要的数据出版模式包括两种，一种是数据在互联网上出版，另一种是数据作为出版物类似于论文进行出版[20]。第一种出版模式是非正式出版模式，其数据质量与数据的长期存储及可读性无法得到有效保障，数据位置可能随着时间的推移而出现变化，此外，该种出版模式下数据元数据的完整程度低，非数据创建人可能无法准确理解数据的内容。第二种出版模式是数据的正式出版模式，其经过正式的出版程序，经同行评议，能够确保数据的完整性、位置的固定性以及元数据的完备性，其数据质量在一定程度上有所保障。

为了确保数据的质量和长期可用性，更多的学者较为推崇数据的正式出版方式。数据正式出版方式共有5种，包括数据独立出版、数据论文出版、附录数据出版、期刊驱动的数据出版、Overlay 数据出版[21]。这几种出版方式的承担主体各不相同，且有着不同的出版及评议政策。采取不同的出版方式在一定程度上会对数据引用造成影响，在进行数据影响力评价时要考虑数据出版方式的因素，对不同数据出版方式下的数据影响力评价方案进行考量，做到具体问题具体分析。

KE 针对此问题进行了一定的研究，在《科学数据价值》报告中认为数据正式出版方式是最适用于提取及发展科学数据影响力计量体系的。在数据正式出版方式中，数据论文出版类似于传统科学出版物出版，因而可以运用现存的引用和 altmetrics 指标；数据独立出版方式可能也会运用到相似的指标，但是需要解决包括数据仓储类型和可用性、数据出版场所选择、数据元数据可用性等问题，以提取相关可用指标。目前，altmetrics 以及其他一些社会网络指标适用于所有数据出版模型下数据影响力的计量。此外，KE 还指出在未来数据出版模型中可能存在的并会对数据影响力计量产生影响的问题，例如，数据集的识别，使用标识符的选取；数据集的粒度与版本问题，数据集与传统出版物不同，会随观测、探测、实验等研究实践的进行而持续增加，同时也会存在为增加数据出版量而将数据集切分为多个数据集出版的现象，因而应对数据集单元的确认进行研究；数据自引现象；数据原创作者问题（名誉、代笔数据作者）等，所有出现于科学出版物模型中的问题都有可能出现在数据出版中，需要在后续研究中多加注意[18]。

5.3 科学数据影响力与数据引用间关系

数据引用是指包含在已出版论文参考文献中的数据资源的正式引用行为，这些数据资源通常支持论文结论的得出[22]。可见数据引用在其内涵上标志着数据的重用与影响力。如参考文献首席专家 E. Moss 所说“如果数据的使用是可识别的，那么它的影响力可以更好地去计量”[23]，数据引用便是使数据的使用可识别的关键，它是进行科学数据影响力评价的地基与桥梁，只有当科学数据经过统一、标准化格式的引用，科学数据的影响力才有迹可循。因而很多机构和学者在进行科学数据影响力研究时都会提及数据引用的重要性。

2012 年，大气研究高校联盟（University Corporation for Atmospheric Research，简称 UCAR）召开的“构建数据生命周期：通过数据引用跟踪数据使用状况研讨会”中指出数据引用是追寻数据使用状况时所面临的严峻挑战，并对数据引用的相关活动、工具、挑战、障碍、措施进行了探讨，着重分析了如何为科学数据分配标识符，旨在促进数据管理与引用，为数据使用状况的追踪奠定基础[24]。

ICPSR 十分注重对数据引用文化的培养，鼓励良好的数据引用实践及规范化的引用标准，并通过数据引用情况来追寻科学数据影响力状况；曾通过数据关联文献目录对在其存储库存储的数据引用量进行追踪，发现有 60 000 多篇文章引用了 ICPSR 数据；同时 ICPSR 还对下载量、数据重用地点进行跟踪，旨在帮助研究人员了解谁使用了数据，数据被用于哪里[25]。

DCC 学者 A. Ball 和 M. Duck[26] 在其报告中提到当科学数据具有完备的元数据时，其他研究人员才能更好地发现与重用科学数据，其引用数量将得到提升，进而扩大影响力。同时，他们指出对数据论文的引用可以算作对原科学数据的引用，可纳入科学数据影响力中。最后他们还提到并不是所有科学数据的引用都可以计入影响力，存在虚假数据引用的情况，其并不能导致论文相关结论的得出，因而在实际计量中应考虑到这种状况的存在，同时借助其他指标来客观评价科学数据的影响力。

目前，数据引用领域已经公布了一定的规范与标准，未来科研交流与电子学术组织（The Future of Re-

search Communication and e-Scholarship，简称 FORCE11）推出了《数据引用原则》，为后续相关实践以及工具开发提供了指南，原则包括数据引用的重要性、信誉与归属性、证据性、可识别性与永久性（机器可操作的唯一永久性标识）、便于访问获取性（包括对数据本身及元数据等相关材料的获取）、可验证性、互操作性与灵活性[27]；Dataverse 项目在 FORCE11 的原则上对数据引用进行了规范，并建议利用唯一永久性标识符取代 URL，同时采用通用数字指纹（Universal Numerical Fingerprint，简称 UNF）方法来确保数字资源的唯一性；并将数据引用确定为 7 部分，包括 5 个人类可读项：作者、标题、年份、数据仓储（出版商）和版本号，两个机器可读项：HDL 或 DOL 以及 UNF，推动了数据引用的标准化与规范化[28]。

但在实践中发现，研究人员的数据引用意识较低，有待进一步加强。2013 年 4 月至 6 月，格林纳达大学 EC3 文献计量小组对 DCI 进行了首次数据引用分析，发现 DCI 中 88% 的数据没有被引用，但 2007 年后数据引用水平呈逐年递增趋势[29]，可见，研究人员的数据引用意识普遍不足，但逐渐增强，数据引用标准化、常规化是未来的发展趋势。通常而言，当科学数据引用数量高时，其影响力也高，反之亦然。但同样也存在自引、互引、负面引用等传统文献存在的问题，需要在评估中考虑这些因素。同时数据引用是一个复杂的活动，需要多方利益主体共同参与，推进数据引用的发展，并在此基础上完善对数据影响力的评估。

5.4 科学数据影响力与数据质量间的关系

科学数据影响力与数据质量间既有着千丝万缕的联系，又存在着本质的区别。数据质量主要是数据自身完整性、科学性状况的反映，而影响力主要是一种改变、支配他人思想和行为的能力，科学数据影响力的显现必须经过传播交流，同时也会随着传播交流的深化而不断提升，但在这一过程中科学数据本身的质量是不会发生任何变化的。

通常情况下，科学数据自身质量高时，其引用量会随之增加，数据影响力也会随之提高，这种情况与高质量论文引用量高的情况是一致的。但在实际中存在很多特殊状况，不能一概而论，例如当一篇文章引用某科学数据集，指出其中存在的错误及问题时，并不是因为数据质量高而进行引用，相反因为数据本身存在问题而以反例的形式对其进行引用，其影响力也是在提高的。此外，一些科学数据内容与人们关注的热点问题相关时，包括饮食、健康等，其在社交媒体中被提及的频率较高，影响范围也更加广泛，但这与数据自身质量并没有太大的关系[30]。因而学术界也出现了一些反对声音，他们认为不应该过多的去关注影响力的计量，而应该将更多地注意力放在数据质量的控制上[31]。笔者认为这种观点存在偏颇，影响力与质量并不是相互对立的，相反，对两者中任何一种的研究都会对另一种研究产生促进作用。两者是科学评价的不同方面，它们的侧重点不同，影响力是对需求的反映，对科学数据影响力的研究可以促使研究人员产出更多有价值、为人所用的科学数据；质量是对科学数据本身的反映，对数据质量进行严格控制，是科学研究科学性、严谨性和科学事业可持续发展的保障。因此应对两者保持同等的重视，同时在研究中需要关注两者的关联性，以促进对两者更为深入的认识。

6 科学数据影响力相关主体研究

科学数据影响力评估是一个复杂的过程，涉及的相关利益主体众多。

RDA 和 WDS 联合成立了数据出版兴趣小组，在该小组的一份报告中提及了科学数据影响力评价中受益的群体，包括数据生产者、数据中心、数据管理者、研究设施和学术机构、科学出版商以及资助机构，在这一过程中他们可以定量且正式地展示数据对社会进步与发展的重要意义[32]。

KE 针对科学数据共享与评价中涉及的主要利益相关主体进行研究，并对各主体间关系进行展示。其中，主要相关利益主体有资助者、研究基础设施、科研工作者、数据中心、出版商、图书馆、出版物数据库，各主体间关系如图 2 所示[18]：

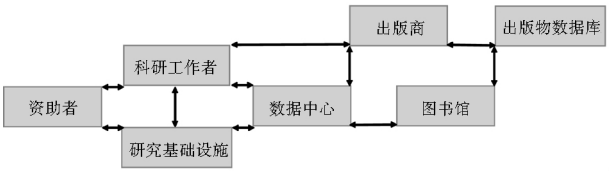

图 2 科学数据影响力相关主体关系示意图

在数据共享与评价中各相关主体的关注点为：

（1）资助者：资助科学数据的产出，并使公共科研资助效益最大化；鼓励数据重用；促进数据共享与重用的意识与奖励。

（2）研究基础设施：从资助角度，支持成果的长期保存与获取；从提供者角度，促进数据的使用。

（3）科研工作者：将数据成果纳入信誉评价体系中；在出版物中对科学数据进行引用，并使数据可用。

(4) 数据中心：以全球统一的方式规范存储和创建元数据；追踪数据重用，促进良好的科学实践；提供数据保管并推荐科学数据的引用格式。

(5) 出版商：处理已提交出版物的数据；执行和促进数据引用与数据计量政策与标准。

(6) 图书馆：使数据可识别、可获取；与数据中心、学者进行协作。

(7) 出版物数据库：出版物与数据引用之间的关联；促进引用计数以及相关指标的运用。

在对科学数据影响力进行研究时，一定要鼓励各方的参与，发挥各利益相关主体的作用，了解各方优势与兴趣点，明确各方责任，协调各方利益，共同致力于科学数据影响力研究，也只有各方相互配合，才能不断深化对科学数据影响力的认识。

7 科学数据影响力评价研究

科学数据影响力评价是科学数据影响力研究的主体内容，也是科学数据影响力研究中最重要的目标之一。目前国外关于科学数据影响力评价的研究主要围绕科学数据影响力评价标准、方法、指标以及工具开展。

在标准方面：CASRAI 成立了数据集级别计量课题小组(Dataset Level Metrics Subject Group)，旨在增强不同倡议间的互操作性以及利益相关主体间的沟通，促进相关标准的统一。2013 年，NISO 启动了 altmetrics 评估计量行动项目，其任务之一便是对数据集、软件等非传统科研成果的 altmetrics 计量进行研究，该项目在 2016 年公布的成果中，对数据集计量的相关概念及相关组织机构开展的研究进行了简要介绍，同时强调了数据引用规范对数据计量发展的重要性[33]。从整体来看，目前关于科学数据影响力标准的研究较少，现有研究也主要集中于概念上的统一，评价流程及方法的标准构建依赖于数据引用及影响力研究的不断进步，有待进一步加强。

在方法方面：目前科学数据影响力评价主要包括定性评价与定量评价两种方法。学者较多关注的是科学数据的定量评价方法，包括引文分析方法、altmetrics 方法、数据仓储附加影响计量方法、管理影响力计量方法[34]等，其中引文分析方法仍然是数据影响力评价中最重要的方法，主要是对科学数据的引用情况进行统计分析；altmetrics 方法主要是统计分析科学数据在网络、社交媒体中传播交流状况的手段，是数据引文分析外重要的补充方法；数据仓储附加影响力计量方法主要是针对数据仓储为科学数据增加的价值及影响进行计量的方法，其代表为生物领域中基于 GBIF 的数据使用索引(Data Usage Index，简称 DUI)；管理影响力计量方法是不同组织管理方式对科学数据传播、交流、重用产生影响的计量方法，美国国家大气研究中心的科学数据资料库(Research Data Archive，简称 RDA)使用了该方法。后两种方法从不同的角度出发对数据影响力进行评价，为数据影响力计量开辟了新的思路。但这些方法的研究深度普遍较浅，现有研究中，前两种方法过度借鉴传统出版物的相关方法，没有充分考虑科学数据的特殊性；后两种方法目前适用于特定的领域或机构，其科学性、普适性有待进一步研究。

当前关于数据影响力评价的定性方法研究较少，主要集中于科学数据同行评议方法的研究，包括出版前同行评议以及出版后同行评议两种方法。英国联合信息系统委员会(Joint Information Systems Committee，简称 JISC)资助的 PREPARDE 项目[35]对不同出版模式下科学数据的同行评议方法进行了研究。无论是定性评价还是定量评价都存在自身的优势与不足，定量评价可以量化评价科学数据的影响力，在给定标准后可以较为客观的反映数据影响力状况，但存在标准合理性、数据不完善等问题，其科学性、严谨性、全面性仍有待研究；定性评价由领域权威专家学者进行评价，其评价结果更为权威，但存在评议者选取、周期长、主观片面性等问题，因而在对科学数据影响力进行评价时要同等重视定性评价与定量评价方法，运用定性与定量相结合的方法对科学数据影响力进行科学、系统、全面的评估。

在指标方面：评价指标是科学数据影响力评价研究中的重要内容，是科学数据影响力评价时的标尺与准则，在当前研究中，学者较多关注科学数据影响力的引文指标与 altmetrics 指标，包括对于两种指标在科学数据中的定义、适用性等探索。引文指标主要包括被引频次、施引文献的 G 指数、施引文献的 Rao－Stirling 多样性指标以及数据场所来源因子，其中被引频次是最基础与最重要的指标，在一定程度上代表着数据的重用状况；施引文献的 G 指数是对科学数据二次影响力的衡量；施引文献 Rao－Stirling 多样性指标是对数据重用范围与广度的测量；数据场所来源因子类似于

期刊的影响因子，这些指标对于科学数据影响力评价都有着重要意义。altmetrics 指标主要包括下载量、浏览量以及分享、保存、推荐、评论、标签等指标，在 MDC 的项目调查[6]中发现学者们普遍认为下载量对于科学数据影响力具有重要价值，而浏览量对于科学数据影响力的影响并不大，其他社交媒体指标的应用还在探索阶段，尽管当前应用范围并不广泛，但仍具有潜在价值。

另外，有相关机构初步构建了科学数据影响力评价指标体系，包括 KE 的数据影响力计量概念化模型、基于 GBIF 的 DUI 指标体系等。可以说对科学数据影响力指标方面的研究已经进入初级阶段，但仍存在大量问题，首先在对科学数据影响力进行评价时过多参考单篇论文的评价系统，相关研究不够系统深入，没有理清楚各项指标对于数据对象的切实意义，在未来研究中需要进一步对各项指标的适用性与应用范围进行界定；其次缺乏特征指标，常将某些指标的计量混淆等同于论文计量，需深化对科学数据自身属性的研究，并根据科学数据自身的特质设定特征指标，以使对科学数据影响力的评价更加合理科学。

在工具方面：科学数据影响力评价过程中应用的工具主要有引文工具和 altmetrics 工具。现存的引文工具有 DCI、CrossRef Search、Google Scholar、Microsoft Academic Search 等，其中 DCI 是汤森路透专门针对科学数据而开发的索引数据库，通过 DCI 可以对科学数据进行检索和探索，同时能够对科学数据的引文情况进行分析[7]，这也是目前科学数据引文分析中最重要的工具；CrossRef Search 可以检索任何具备 DataCite DOI，并被 CrossRef 索引的学术文章引用的科学数据，但检索过程十分复杂[8]；Google Scholar 和 Microsoft Academic Search 可以提供数据论文的引文数据，同样能够为科学数据的引文分析提供帮助。现存的 altmetrics 工具有 PlumX、Altmetric、ImpactStory 等，以及正在研发当中专门针对于科学数据集的 DLM（Data - Level Metric）。不同工具间服务对象、覆盖数据范围、挖掘层次各有不同，在具体应用中应该根据不同需求，综合应用各种工具对各项评价指标数据进行挖掘与分析。现存工具在使用过程中遇到的最大挑战在于数据引用意识缺乏导致可追踪数据过少；其次现存工具对科学数据的覆盖范围窄，部分科学数据不在其追踪范围内；此外，针对同一科学数据，不同 altmetrics 工具的统计结果往往存在差异，需要进一步对各工具统计数据的有效性进行检验，以确定其结果的权威性。

8 我国科学数据影响力研究的发展建议

我国科学数据影响力研究处于起步阶段，无论是理论基础还是实践应用都未形成有效体系，在对国外相关领域研究与实践进行调研与分析后，笔者认为我国可以从以下几个方面来推进科学数据影响力的研究。

（1）提高科学数据共享引用意识，培养科学数据引用文化。对科学数据的规范引用是对科学数据影响力进行追踪的基础，我国应加快相关政策和制度的制定，运用强制性与鼓励性手段，加强科研人员关于科学数据的引用意识，逐步建立起科学数据引用文化，这是我国科学数据影响力研究开展的必要前提。

（2）加强科学数据影响力的基础理论研究。在我国未来的研究中，应切实把握科学数据影响力的具体内涵及特征，厘清科学数据影响力涉及的领域及主体，挖掘科学数据影响力与其他领域的深层关系等，只有对研究对象进行彻底认识，其后续研究才有据可循，后续科学数据影响力的评价才可能具备科学性和有效性。

（3）加强特征指标的构建。我国在科学数据影响力研究前期借鉴传统出版物的评价体系是十分必要的，但更应该依照科学数据的自身特征，构建适用可行且具备科学数据特色的评价体系。在构建指标体系时，应厘清各项指标对于数据对象的切实意义，切忌与论文计量混为一谈。

参考文献：

[1] 邓仲华，李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作，2013，34(4)：19－23.
[2] 方璐. 大数据时代的科学研究方法[D]. 杭州：浙江工业大学，2014.
[3] 顾立平. 数据级别计量——概念辨析与实践进展[J]. 中国图书馆学报，2015(2)：56－71.
[4] 丁楠，黎娇，李文雨泽，等. 基于引用的科学数据评价研究[J]. 图书与情报，2014(5)：95－99.
[5] 彭国莉，吕先竞，刘文君. DCI 社会科学数据分析研究[J]. 西南民族大学学报(人文社会科学版)，2015(3)：231－233.
[6] 邢红梅，吕先竞，刘文君，等. 基于 DCI 的社会学数据影响力分析[J]. 图书馆理论与实践，2016(2)：43－46.
[7] OECD principles and guidelines for access to research data from public funding[EB/OL]. [2016－11－14]. http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm.

[8] 李慧佳，马建玲，王楠，等. 国内外科学数据的组织与管理研究进展[J]. 图书情报工作，2013，57(23)：130－136.

[9] Oxford living dictionaries [EB/OL]. [2016－11－10]. https://en.oxforddictionaries.com/definition/impact.

[10] PENFIELD T，BAKER M J，SCOBLE R，et al. Assessment，evaluations，and definitions of research impact：a review [J]. Research evaluation，2013，23(1)：1－12

[11] DURYEA M，HOCHMAN M，PARFITT A. Measuring the impact of research [EO/BI]. [2016－11－10]. http://facdent.hku.hk/docs/ResGlob2007.pdf.

[12] FEAR K. The impact of data reuse：a pilot study of five measures [EB/OL]. [2016－11－12]. https://www.slideshare.net/asist_org/kfear－rdap.

[13] BBSRC policy on maximising the impact of research [EB/OL]. [2016－11－14] http://www.bbsrc.ac.uk/documents/bbsrc－impact－policy－pdf/.

[14] The value of research data [EB/OL]. [2016－11－15]. http://www.ands.org.au/working－with－data/articulating－the－value－of－open－data.

[15] Dataimpact campaign [EB/OL]. [2016－11－15]. http://www.ands.org.au/news－and－events/dataimpact.

[16] FIENBERG S E，MARTIN M E，STRAF M L. Sharing research data [M]. Washington：National Academies Press，1985.

[17] FEAR K M. Measuring and anticipating the impact of data reuse [D]. Michigan：University of Michigan，2013.

[18] COSTAS R，MEIJER I，ZAHEDI Z，et al. The value of research data－metrics for datasets from a cultural and technical point of view [EB/OL]. [2016－10－28]. http://www.knowledge－exchange.info/event/value－reseach－data－metrics.

[19] OCHSNER S A，STEFFEN D L，STOECKERT C J，et al. Much room for improvement in deposition rates of expression microarray datasets [J]. Nature methods，2008，5(12)：991.

[20] CALLAGHAN S，DONEGAN S，PEPLER S，et al. Making data a first class scientific output：data citation and publication by NERC's environmental data centres [J]. International journal of digital curation，2012，7(1)：107－113.

[21] LAWRENCE B，JONES C，MATTHEWS B. Citation and peer review of data：moving towards formal data publication [J]. International journal of digital curation，2011，6(2)：4－37.

[22] MAYERNIK M S. Data citations：initiatives，issues，and firststeps [J]. Bulletin of American society for information science and technology，2012，8(5)：23－28.

[23] MOSS E. Viable data citation：expanding the impact of social science research [EB/OL]. [2016－11－17]. http://www.slideshare.net/asist_org/rdap13－moss.

[24] MAYERNIK M. Bridging data lifecycles：tracking data use via data citations data workshop [EB/OL]. [2016－11－17]. http://opensky.ucar.edu/islandora/object/technotes：505.

[25] KONKIEL S. Tracking citations and altmetrics for research data：challenges and opportunities [J]. Bulletin of the American society for information science and technology，2013，39(6)：27－32.

[26] BALL A，DUKE M. How to track the impact of research data with metrics [EB/OL]. [2016－11－05]. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_track_data_impact.pdf.

[27] Joint declaration of data citation principles [EB/OL]. [2016－11－25]. https://www.force11.org/group/joint－declaration－data－citation－principles－final.

[28] Data citation [EB/OL]. [2016－12－22]. http://best－practices.dataverse.org/data－citation/#data－citation－standard.

[29] PETERS I，KRAKER P，LEX E，et al. Research data explored：citations versus altmetrics [J]. Journal of the association for information science & technology，2015，66(10)：2003－2019.

[30] COLQUHOUN D. Should metrics be used to assess research performance? A submission to HEFCE [EB/OL]. [2016－11－20]. http://www.dcscience.net/2014/06/18/should－metrics－be－used－to－assess－research－performance－a－submission－to－hefce/.

[31] BURROWS R. Living by numbers? Metrics，algorithms and the sociology of everyday life－sydney Ideas－the university of sydney [EB/OL]. [2016－11－20]. http://sydney.edu.au/sydney_ideas/lectures/2014/professor_roger_burrows.shtml.

[32] Bibliometrics working group case statement [EB/OL]. [2016－11－20]. http://www.rd－alliance.org/sites/default/files/case_statement/RDA_WDS_WG_Publishing_Workflows.pdf.

[33] Alternative assessment metrics (altmetrics) initiative [EB/OL]. [2016－11－24]. http://www.niso.org/topics/tl/altmetrics_initiative/.

[34] KONKIEL S. Beyond citations：new metrics for measuring the impact of research data [EB/OL]. [2016－12－02]. https://scholarworks.iu.edu/dspace/handle/2022/16979.

[35] DCC. Preparde [EB/OL]. [2016－11－24]. http://www.dcc.ac.uk/preparde/preparde.

[36] STRASSER C，KRATZ J，LIN J. Make data count－unit 1 final report [EO/BL]. [2016－12－03]. https://dx.doi.org/10.6084/m9.figshare.1328291.v4.

[37] Data citation index [EB/OL]. [2016－11－15]. http://wokinfo.com/products_tools/multidisciplinary/dci/.

[38] KONKIEL S. Tracking the impacts of data－beyond citations [EB/OL]. [2016－11－17]. http://blog.impactstory.org/data－impact－metrics/.

作者贡献说明：

王毅萍：研究内容的调研与论文内容的整体撰写；

马建玲：论文题目选取的建议、论文的修改与润色。

Progress of Studies on Foreign Scientific Data Impact

Wang Yiping[1,2] Ma Jianling[1]

[1] Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000

[2] University of Chinese Academy of Sciences, Beijing 100049

Abstract: **[Purpose/significance]** This article aims at analyzing current progress and problems of scientific data impact in foreign countries, as well as providing references for domestic researches. **[Method/process]** This article was mainly for investigation of foreign institutions, aided by Web of Science Database and Google Scholar literature researching. It used the retroactive method to summarize the progress of foreign scientific data impact research. **[Result/conclusion]** Now, the evaluation of scientific data impact has attracted foreign academia. Roughly, the contents contain five aspects: connotation, type, relationship, stakeholder and evaluation method. However, the overall researches are still in early stages. Based on foreign experience, this article suggests, for future domestic research on scientific data impact, cultivating consciousness and culture of data citation, and strengthening the research of basic theory and characteristic index.

Keywords: scientific data impact evaluation