DOI:CNKI:11-2746/G2.20121213.1730.002

数据引证研究:进展与展望*

侯经川　方静怡

摘　要　随着大数据时代的来临以及数据密集型科学研究范式的兴起,“数据引证”问题日益受到关注。本文对该领域的国际研究现状进行了梳理总结,研究发现:①对数据引证的知识计量研究,将推动文献计量学、信息计量学和科学计量学三者的合流,形成一个统一的新学科——知识计量学;②数据引证实践的现状不尽如人意,但已在诸如数据集标识系统建立等问题上取得了重要进展,统一规范化数据引证格式的趋势日益清晰;③数据引证现状评估与研究进展的追踪,数据引证索引的编纂、指标体系研究以及数据引证数据库的建立,基于数据引证行为、记录以及索引的分析,应是未来需重点突破的方向。图1。表2。参考文献14。

关键词　数据引证　知识计量学　大数据　数据密集型科学

分类号　G350　G301

Review on Data Citation in the Context of Big Data

Hou Jingchuan & Fang Jingyi

ABSTRACT　In the context of big data and data-intensive science, data citation now calls for special attention of many organizations and workshops worldwide. This paper summarizes international research on “data citation” in recent years, and gives some conclusions as follows: 1) “data citation” studies will bridge the gap among bibliometrics, informetics and scientometrics, and lead to the forming of knowmetrics; 2) “data citation” practices are far from satisfactory, but with some important advancement such as establishment of data set identifier systems, and the normalization of data citation practice is just on the way; 3) the most important research topics about data citation in the future include evaluating the progress of data citation, compiling data citation indices, building data reference database, and extending analysis based on data citation records and indices. 1 fig. 2 tabs. 14 refs.

KEY WORDS　Data citation.　Knowmetrics.　Big data.　Data intensive science.

1　引言

科学研究的科学性与可考证性,有赖于其所使用方法与数据的透明化,以及参考来源的明确化。数据引证的必要性,早在1982年就由著名的科学计量学家Howard D. White指出:“社会科学学者们应该在他们的著作中,引用他们所使用的那些数据文件(可被机器处理的数据,MRDF),并以区别于正文的规范化的参考格式列出,正如他们引用书籍、论文和报告一样,这绝对不是一个新的话题”[1]。

随着海量数据获取、存储与处理方法与技术的飞速发展,“大数据”时代已经来临,并对每个领域都造成了影响[2]。2007年计算机图灵奖得主Jim Gray在NRC-CSTB的演讲报告中提出了科学研究

*　本文系国家自然科学基金青年项目“基于信息主权的国家核心竞争力保护与提升策略研究”(项目编号:70973037)的研究成果之一。

通讯作者:侯经川,Email:jchou@infor.ecnu.edu.cn

的第四范式[3]——数据密集型科学研究(以协同化、网络化与数据驱动为其主要特征),在学界引起了巨大反响,数据在科学研究中的重要性更甚从前。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。然而,诸如数据所有权与知识产权保护[4]、数据使用的溯源[5]、数据再利用价值的评价[6]等问题也开始显现,进一步说明了规范数据参考与引用行为的紧迫性。

自2011年起,众多国际组织纷纷开展以"数据引证"为主题的研讨会与相关活动,包括DataCite、DCC(The Digital Curation Centre)、ESIP(The Federation for Earth Science Information Partners)、BRDI(The National Academy's Board of Research Data and Information)以及CODATA(The International Council for Science's Committee on Data for Science and Technology)等。奥巴马政府于2012年3月29日发布的"大数据研究与发展计划"("Big Data Research and Development Initiative")中也将"数据引证"特别列出,反映出NSF(National Science Foundation,美国国家科学基金会)致力于实现负责任的数据管理和数据可持续性的承诺[7]。

2　数据引证与知识计量学的成形

虽然"数据引证"引发了科学界的热烈关注,但在文献计量学与信息计量学相关的会议和文献中还鲜见其身影,这反映了目前文献计量学与信息计量学的研究依然停留在文献单元的层次,对于深入知识单元的"数据引证"缺乏足够的重视。然而,在当今大数据时代以及数据密集型研究范式兴起的背景下,数据引证的重要性不言而喻。可以预见,以数据引证为核心的知识单元的计量分析必将迅速崛起,从而打破文献计量学、信息计量学与科学计量学之间的鸿沟,推动三者融合统一于一门新的学科——知识计量学。

这种必然性体现在三个方面:

首先,数据引证将成为文献信息计量分析领域新的研究对象。作为科学记录①的组成部分之一,"数据引证"具有信息计量分析价值。信息计量学,诞生于信息爆炸式增长以及"大科学"的背景之下,包含了与信息以及信息的存储、检索、利用过程相关的所有定量研究[9]。数据引证,是信息利用行为的一种,也是科学交流过程中的重要一环,具有潜在的可计量性,在此基础上可进一步追踪数据集的使用情况,以及评估数据集对科学研究、科学交流的影响。这为信息计量学及科学计量学打开了一个新的视角——数据使用的视角,去揭示信息运动与科学活动的内在规律。同时,科研人员以及数据存储机构为保证数据的可获得性与可用性所付出的努力,可以通过数据的规范化参考与引用被公之于世,这也为科研评价与创新激励提供了一个新的维度。

其次,数据引证将使文献信息计量分析从文献单元深入到知识单元。对文献资料(包括专利、网络信息等)以及相关对象(如作者、期刊、研究机构、基金等)进行定量分析,特别是引文分析,是文献计量学、信息计量学与科学计量学中常见的追踪科学发展轨迹的方法,也积累了长期的经验并取得了丰硕的成果。然而,正如Jim Gray所指出的,科学研究的素材实际包含呈金字塔形的三个层面[3](见图1):文献、派生和重组数据、原始数据。金字塔的下两层占据更多的比例,特别是在数据密集型科学的背景下,这一点尤其突出。因此,将科研投入/产出研究的对象,从以往的文献资料,拓展至更为基础的科研素材——原始数据、派生和重组数据,进行基于数据引证行为与记录的分析,是极有必要并且非常自然的。

第三,以数据引证为核心的知识计量分析具有诱人的前景。数据引证的规范化以及相关研究,能为科学研究中的信息查询提供便利,从而提升信息

① 科学记录(scientific records),是包含独立的科学期刊、会议展示与文集收录,以及支撑这些出版物的数据与其他佐证的集合体[8]。

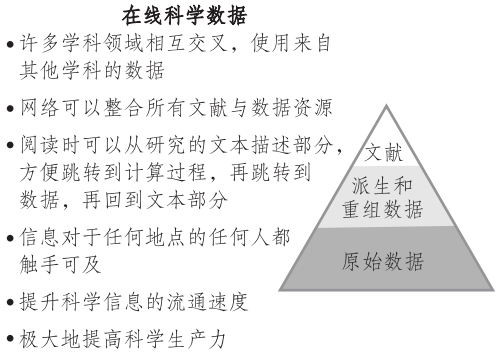

图1 科学研究的素材

（来源：The Fourth Paradigm：Data-intensive Scientific Discovery. Microsoft Research，2009）

利用的效率与收益、加快科学发现与创新的步伐，这也是信息计量学与科学计量学的宗旨所在。Jim Gray 在 NRC－CSTB 的演讲报告[3]中，呼吁联合所有的科学数据与文献形成一个互操作的世界：读者在阅读文献的时候，可以访问论文的原始数据甚至重复研究的过程，或者能够从数据开始找到与之相关的所有文献。这种数据与文献的互操作可以提升“信息速度”（information velocity），从而提高科学生产力。显然，数据引证的规范化，以及对数据与引证文献之间关系的研究，是实现这种诱人设想的第一步。

3 数据引证的研究现状

数据引证规范化的重要性，早在几十年前就被认识到，然而当前的状况却不尽如人意。一项发表于2000—2010年间、覆盖6种期刊500篇文献的研究显示[10]，数据引证行为在学界中实际上是严重缺乏的：198篇涉及数据再利用的文献中，只有14%的文献在他们所使用的数据集中标出了数据集的唯一标识；只有12%的文献（当中存在部分重叠），提到了数据集作者和数据存储地的名称；很少有文章将其对数据集的引用纳入正文后的参考来源部分。与此同时，也鲜有政策涉及推荐或要求采纳正确的规范化的数据引证方式：估计只有三分之一的数据存储机构（n＝26），6%的期刊（n＝307），1/53的科研资助方（n＝53）给出了对数据使用行为的要求或建议[10]。

建立一种统一的、严谨的、规范化的数据引证格式，对后续的基于数据引证行为与记录的分析至关重要。在这之前，一些关键性的问题需要得到解决，例如：数据著作权与所有权的界定、数据保存与可持续性维护、数据特征描述的指导规范、数据集标识系统的建立、科学记录中引用数据的规范、数据集与文献之间的有效链接等。国际上众多致力于“数据引证”相关研究的组织与研讨小组（见表1、表2），已经在某些问题上取得了重要进展。正如 Joseph A. Hourclé 总结的那样[11]：“我们现在已经有了一些用于指导数据引证中应当标明哪些属性的规范、对不同的标识系统（identifier system）的分析、通过 EZID 生成价格合理的 DOI（Digital Object Identifier）、OAI－ORE（Open Archives Initiative－Object Reuse and Exchange，开放存取先导计划之数字对象再利用和交换计划）以及用于描述合集与替代集（aggregates and alternatives）的元链接（Metalink），我们现在已经具备了建立数据引证框架体系所必需的构成元素。”

4 亟待研究的若干问题

就目前来说，规范的数据引证还未得到广泛的实践，也还未在科学出版中被正式要求，但是这种趋势是显而易见的。对数据引证行为与记录的知识计量研究，亟待推进。相比等一切条件都完善时才采取行动，未雨绸缪是更佳的选择。其中，以下几个问题亟待解决：

（1）数据引证现状评估与研究进展的追踪

在西方国家已经有了一些针对数据引证行为的调查研究[10,12－13]，这些调查结果为掌握数据引证实践及相关政策的现状提供了切实的证据，暴露出现存的问题并引起了必要的关注。但在中国，目前还没有发现有关“数据引证”的调查性研究，相关的理论性探讨也很鲜见。中国的科学记录格式与西方国家存在着一定的差异，并具有自身独特的特点。因此，国内学者们应该对此给予足够重视，

表 1　以“数据引证”为议题的研讨会

时间	活动	主办方
2011.3	“地理数据信息学：探索地理数据的生命周期、引用以及整合”研讨会 “Geo-data Informatics：Exploring the Life cycle，Citation and Integration of Geo-data” Workshop	NSF Directorate for Geosciences
2011.5	“数据引证的原则”研讨会 “Data Citation Principles” Workshop	IQSS（Institute for Quantitative Social Sciences at Harvard University）
2011.8	“发展数据属性及引用行为规范”研讨会 Developing Data Attribution and Citation Practices and Standards ”An International Symposium and Workshop	US CODATA and the Board on Research Data and Information in collaboration with CODATA – ICSTI Task Group on Data Citation Standards and Practices
2011.11	“建立科学数据引证的文化”研讨会 “Building a Culture of Research Data Citation” Workshop	ANDS
2012.3	“大数据研究与发展计划”中特别提出数据引证的“致同僚的一封信” “Dear Colleague Letter” specifically addressing data citation as part of the “Big Data Research and Development Initiative”	NSF Directorate for Geosciences
2012.3	“科学数据访问与保存”峰会中的“数据引证”专家组 Data Citation Panel in Research Data Access and Preservation Summit（RDAP12）	ASIS&T
2012.4	“数据生命周期：通过数据引用追踪数据使用”研讨会 “Bridging Data Lifecycles：Tracking Data Use via Data Citations” Workshop	University Corporation for Atmospheric Research（UCAR）
2012.5	探索数据引证最佳实践的“别忘了数据”研讨会 “Don’t Forget the Data” Workshop Explores Best Practice for Data Citation	University Corporation for Atmospheric Research（UCAR）
2012.5	“数据引用与 DataCite 介绍”研讨会 “An Introduction to Data Citation and DataCite” Workshop	JISC，DCC，DataCite
2012.7	“描述、传播、发现：实现有效数据引证的元数据”研讨会 “Describe，Disseminate，Discover：Metadata for Effective Data Citation” Workshop	JISC，DCC，DataCite

表 2 致力于“数据引证”相关研究的国际组织

名 称	贡 献
◇ DataCite Helping you to find, access, and reuse research data	➢ 提供诸如数据集 DOI 生成与注册等多种服务与资源,以指导科研人员寻找、识别与正确引用数据,为数据中心提供数据集标识以及数据发布的流程与标准,为期刊出版商提供数据与相关文献的有效链接
◇ ANDS Australian National Data Service	➢ 联合 DataCite 发展元数据标准,提供为数据集特别定制的 DOI 生成与注册服务,携手汤森路透社、Elsevier 研究通过 DOIs 对数据使用进行追踪与记录的可行性,推进数据发布并提倡将数据维度纳入科研评价中
◇ IASSIST International Association for Social Science Information Services & Technology	➢ 传播数据引证手册,提供数据引证资源,建立 SIGDC(Special Interest Group on Data Citation),发布对如何正确引用数据集的指导规范与建议
◇ DCC The Digital Curation Centre	➢ 提供如何引用数据集并将数据集与出版物相关联的详尽指导
◇ Thomson Reuters Web of Knowledge SM	➢ 开启数据引证索引(Data Citation Index)的编纂,联合诸如 ICPSR 等大型数据存储机构,收集数据资源的参考引用记录,并建立数据资源与文献之间的有效链接,方便数据的发现、使用与分配,为基于数据的研究提供支持
◇ Data ONE ◇ OpenWetWare ◇ ESIP, the Federation for Earth Science Information Partners ◇ BRDI, the National Academy's Board of Research Data and Information ◇ CODATA, the International Council for Science's Committee on Data for Science and Technology	➢ 赞助与举办与“数据引证”相关的研讨会及相关活动,包括对一些诸如数据引证的文化以及可持续性等更广泛议题的探讨

并填补国内这一研究的空白,展现中国数据引证实践的真实状况,为后续研究以及相关规范、政策的制定提供事实依据。同时,也需要对“数据引证”领域国际上的研究前沿与动态进行实时跟踪,以便及时了解国内研究的不足并加以补充与追赶。

(2)数据引证索引的编纂、指标体系研究以及数据引证数据库的建立

与数据引证有关的指标体系、索引体系的建立与编纂,有助于更好地评估数据对于科学研究与科学交流的影响。SCI、SSCI、CSCI 以及 CSSCI 各种引文数据库的存在,为信息计量研究(例如引文分析)提供了大量的基础数据,极大地促进了信息计量学的发展。同样,数据引证索引以及相关引用数据库的建立,将对数据引证的相关研究产生巨大的推动作用。2012 年 6 月 22 日,汤森路透社已经发出通告,其知识产权与科学部(The Intellectual Property & Science Division of Thomson Reuters)将在美国图书馆协会会议(ALA)上开始预览“数据引

证索引”(the Data Citation Index ™),其正式版本预期于2012年年底在Web of Knowledge ℠平台上推出,这将极大地方便国际范围内对“数据引证”的研究[14]。同时也提醒中国科学界应该立即开启相关研究。

(3)基于数据引证行为、记录以及索引的分析

数据集与文献资料均为科学记录的组成部分,因此将目前对文献资料(包括专利、网络信息等)以及相关对象(如作者、期刊、研究机构、基金等)进行的定量分析,引申至数据集的分析甚为自然。未来可供探索的方向包括:通过数据引证与再利用分析数据共享效率;数据集的共现、共引分析;数据与文献之间的链接分析;数据溯源分析、数据集的质量与再利用价值的评估;通过追踪数据使用与再利用研究数据生命周期,提高科研投入/产出效率;通过识别潜在的科学数据共同体促进科研合作、优化数据资源配置;从数据使用角度跟踪科学的发展脉络等。

5　结语

作为科学记录的基础性成分之一,数据引证的必要性以及数据引证的潜在可计量性已经引起国际上的关注。Howard D. White 口中“绝对不是一个新话题”的数据引证,在当今大数据时代来临以及数据密集型范式兴起的背景下,被赋予了新的内涵。对数据引证的知识计量研究,将推动文献计量学、信息计量学和科学计量学三者的合流,形成一个统一的新学科——知识计量学。得益于国际上众多致力于“数据引证”相关研究的组织和研讨小组,与“数据引证”相关的一些基础性研究已经取得了重要的进展。与文献计量学、信息计量学和科学计量学的传统研究对象相比,数据集具有其独一无二的复杂性,数据引证在实践中肯定会遇见许多未曾遭遇过的问题与障碍。但是,新问题的存在也预示着科学发展的新机会。

参考文献:

[1] White H. Citation analysis of data files use [J]. Library Trends, 1982, 31(3): 467-477.

[2] Steve L. The age of big data[N/OL]. The New York Times, [2012-2-11][2012-8-28]. http://forum.ccer.edu.cn/showtopic.aspx?topicid=124765&page=end.

[3] Jim G. On eScience—A transformed scientific method [C]// Tony H, Stewart T, Kirstin T. The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009: 19-33.

[4] Wallis J, Borgman C. Who is responsible for data? An exploratory study of data authorship, ownership and responsibility [C/OL]// Proceedings of the Annual Meeting of the American Society for Information Science and Technology, 2011, 48: 1-10. [2012-08-29]. http://dx.doi.org/10.1002/meet.2011.14504801188.

[5] UCLA Library. Bridging data lifecycles: Tracking data use via data citation [EB/OL]. (2012-04-05)[2012-08-28]. http://library.ucar.edu/data_workshop/.

[6] Palmer C, Weber N, Cragin M. The analytic potential of scientific data: Understanding re-use value[C/OL]//Proceedings of the 74th Annual Meeting of the American Society for Information Science & Technology. Silver Spring, Mary. (2011-12-10)[2012-08-28]. http://www.asis.org/asist2011/proceedings/submissions/174_FINAL_SUBMISSION.pdf.

[7] White House. Big data fact sheet [EB/OL]. (2012-03-29)[2012-08-28]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf.

[8] Clifford A. Jim Gray's fourth paradigm and the construction of the scientific record [C]// Tony H, Stewart T, Kirstin T. The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009: 177-183.

[9] Egghe L, Rousseaur R. Introduction to informetrics: Quantitative methods in library, documentation and information Sci-

ence[M]. Amsterdam,The Netherlands: Elsevier Science Publishers, 1990: 1 -2.

[10] Valerie E,Sarah W,Nicholas M,et al. Data citation in the wild [R/OL]. (2010 -9 -13)[2012 -8 -28]. http://precedings.nature.com/documents/5452/version/1.

[11] Joseph A. Advancing the practice of data citation: A to-do list [J]. Bulletin of the American Society for Information Science and Technology, 2012, 38: 20 -22.

[12] Sieber J,Trumbo B. Giving credit where credit is due: Citation of data sets [J]. Science and Engineering Ethics, 1995 (1): 11 -20.

[13] Hailey M,Mark P. The anatomy of a data citation: Discovery,reuse,and credit [J/OL]. Journal of Librarianship and Scholarly Communication, 2012, 1(1),eP1035[2012 -8 -28]. http://jlsc-pub.org/jlsc/vol1/iss1/6.

[14] Thomson Reuters. Thomson Reuters unveils data citation index for discovering global data sets [EB/OL]. Philadelphia, PA(2012 -6 -22)[2012 -8 -28]. http://thomsonreuters.com/content/press_room/science/686112.

侯经川 华东师范大学商学院信息学系教授,管理学博士,博士生导师。

通讯地址:上海市闵行区东川路500号华东师范大学商学院。邮编:200241。

方静怡 华东师范大学商学院信息学系硕士研究生。通讯作者同上。

(收稿日期:2012 -09 -13)