

科学数据引用的探讨

□ 彭洁 涂勇 / 中国科学技术信息研究所 北京 100038

摘要: 科学数据引用问题一直是困扰科学数据领域的难题之一。文章从科技文献引用的概念入手, 在分析科学数据特点的基础上, 提出科学数据引用的概念、引用对象和存在的问题, 对国内外科学数据引用情况进行了实证分析, 论述了建立科学数据引用规范的必要性, 总结出科学数据引用规范的内容, 从数据生产者、使用者、管理者和政府的角度分析了科学数据引用规范的推广价值, 最后给出了在中国推广科学数据引用规范的建议。

关键词: 科学数据, 引用规范, 数字对象唯一标识, DOI, 科技文献, 数字图书馆

DOI:10.3772/j.issn.1673-2286.2008.10.002

1 引言

科技文献的被引用情况通常作为评价科技文献影响力和权威性的事实型标准, 参考文献已经是科技文献的重要组成部分, 科技文献发表或出版后, 被其它作者学习、参考、借鉴和引用, 然后发表新的文献, 从而促进学科体系循序渐进地发展。文献是否被引用和引用量的多少, 说明文献对后继工作的影响和作用。文献的引用量和被引量能在一定程度上客观地反映文献和被引文献学术水平的高低。利用引文分析法对期刊、被引作者、引文主题等进行测定和评价, 可以判断一种专业期刊的学术水平、办刊质量, 有利于发现人才, 预测本学科的发展趋势^[1]。

与文献类似, 科学数据的引用也存在大量的需求。科学数据作为科研活动中产生的结果, 具有很大的参考价值, 在进行科研成果发布的时候有必要对所引用的科学数据进行标记, 同时, 科学数据被引用次数也是衡量科学数据价值的重要指标。规范科学数据引用是开展科学研究的重要组成部分, 有利于快速查找到原始的科学数据, 系统地反映科学研究的继承性和关联性, 体现对科学数据版权所有者的成果的认可和尊重。

由于目前客观上还没有形成科学数据引用规范, 而且尽管科研人员意识到科学数据引用的重要性, 但他们尚未形成标识数据引用的习惯, 或者说没有规范要求其对引用的科学数据进行标识, 造成了目前科学数据引用比较混乱的局面。文章拟

从科学数据引用的相关概念入手, 探讨规范科学数据引用的可行性, 并引入数字对象唯一标识技术(DOI)来构建科学数据引用规范, 从而解决科学数据引用规范化和量化的问题。

2 科学数据的特点

科学数据能够被引用, 主要是由科学数据本身的特点所决定的:

(1) 科学数据具有权威性

科学数据是人类社会科技活动所产生的基本数据、资料, 以及按照不同需求而系统加工的数据产品和相关信息, 具有明显的潜在价值和可开发价值, 并在应用过程中得以增值, 是信息时代最基本、最活跃、影响面最宽的科技资源^[2]。其中大部分的科学数据是通过长期观测获得的, 或者是通过长期艰苦工作积累的具有极高价值的科学数据, 具有不容置疑的正规性和权威性, 这种权威的信息资源有很强的证明力和说服力, 它与公开出版、发表的权威资料、著作、网上期刊等有着相同的引用价值, 可辅助研究者开展相关领域的研究, 它成为科技领域甚至人文领域所引证的材料是十分自然和必要的。

(2) 科学数据具有较高的认知度

科学数据所涵盖的领域比较广泛, 主要包括资源环境、基础科学、农业科学、区域综合、人口健康、工程技术等领域, 科学数据的潜在价值和可开发价值已经为人们所关注, 不仅被相关学科研究人

DLF

特别关注

Special Focus

员所关注,也容易被边缘学科、交叉学科研究人员采用,具有较高的引用概率。科学数据引用的受众比较广泛,涵盖科学研究的诸多领域,因此科学数据作为一种认知度高、又有很高利用率和引用价值的信息资源,可以成为引用源之一。

(3) 科学数据具有时序性

部分科学数据(如地震数据、天文观测数据)是连续性的观测数据,许多珍贵的数据是独一无二的,并且也是不可再生的资源,对于解释目前的一些自然现象具有很好的参考价值,其中某些关键的信息点往往能够被有关研究人员敏锐地捕捉到并成为课题研究新的着眼点,有利于得出结论,形成创新性的科研成果。从这个角度讲,科学数据作为连续的珍贵的数据财富,应该成为可以引用的资源。

3 科学数据引用的概念及存在的问题

科学数据引用(scientific data citation)通过一定的标识技术和机制,对所使用的科学数据资源进行描述,标识数据的来源,从而在一定程度上促进科学数据的知识产权保护,便于进行科学数据引用情况的统计和分析。科学研究工作具有协作性和继承性,以文献和数据的引用为基础,通过参考文献和参考数据来构建一整套学术促进、监督、积累以及规范体例的研究系统。

科学数据类型比较广,从数据管理的角度上说,其引用的对象包括:

数据库:人们为解决特定的任务,以一定的组织方式存储在一起的相关的数据集合^[3];

主题(体)数据库:按照特定应用组织起来的数据库,体现了数据库综合性的特点;

专业数据库:通常是主题数据库中的各专业领域数据库;

特色数据库(集):能体现数据中心特色,且数据质量较好的,从事该领域研究必备的数据资源集合;

数据集:由一个或者多个数据记录组成的数据集合,如中国西部环境与生态科学数据中心提供的长时间序列中国植被指数数据集等;

数据产品:由原始数据为满足特定需要而加工

的产品,如遥感数据中的0-4级数据产品;

数据记录:科学数据组织的最小单元,通常以一条记录来描述事物的特征,如自然科技资源数据。

科学数据引用的研究还处于初步阶段,其机制、规范都需要进行研究并在实践中不断地完善,具体体现在:

(1) 缺乏对科学数据引用的意识:在发表科技文献时科研人员已经形成了列出参考文献的习惯,但是对于文中所引用的数据却并没有留意,在进行科学数据共享的时候只是在网站上说明了标注引用情况,但是数据使用者并没有去实现它,同时数据生产者缺乏对数据引用的约束。

(2) 科学数据引用缺乏规范:科学数据中心普遍意识到科学数据引用的重要性,但是对科学数据引用情况缺乏标识规范,没有统一的格式进行说明。

(3) 科学数据引用无法量化:数据生产者在生产数据后并没有对数据的引用情况进行统计,数据在流通过程中所产生的二次数据,以及通过二次数据产生的科研成果都没有对原始来源资料进行说明,也就没有办法来衡量该数据(集)被引用的次数。

(4) 缺乏对数据引用信息的标识符:科学数据引用从一定程度上亟需使用一种简单易用、能够实现引用内容和被引内容之间链接、参考文献相对应的一种机制来实现对科学数据引用的标识。科学数据唯一标识符与该数据的链接地址是联系在一起的,能链接到被引科学数据的详细信息,从而将科学数据的标识和科学数据的获取联系起来,使科学数据的引用“活”起来。

(5) 科技文献的载体比较固定,如期刊、图书等,而科学数据出版缺乏一个稳定的载体。科学数据的出版虽然有些大的数据中心管理,但是整体上缺乏权威固定的载体,从而限制了科学数据的访问和使用。

4 国内外科学数据引用的实证分析

国外大型数据机构与组织都意识到科学数据引用的重要性,也在积极探索科学数据引用的问题,希望能够有效地促进科学数据的知识产权保护。德国国家科技图书馆(TIB)在这方面做出了有益的探索,并开始在实践中应用,由德国研究基金资助的

“科学基础数据引用与出版”项目中将数字对象标识(DOI)引入到数据标识中来,基于DOI给出了数据引用的基础格式,并在文献中对其引用的科学数据进行标识,是值得借鉴的科学数据引用范例。

Storm博士在他的《汉诺威与迈阿密气候对比》一文中引用了Weather博士注册的汉诺威地区的气象基础数据,并使用DOI作为数据的唯一标识^[4]。

引用标识如下:

Weather, 2003: “Weather in Hannover for 2003”

(DOI: 10.1594/WDCC/W_Han_2003_MMB_2)

目前国内各科学数据共享中心、数据生产商都对数据引用提出了需求,希望数据使用者在使用该中心数据资料取得的研究成果或论文中对该数据中心致以谢意,有利于更多人关注数据生产单位所提供的服务,便于确定数据的使用级别,并提供了基本的数据引用说明,如:

(1) 世界数据中心冰川(雪冰)冻土学科中心(兰州)数据引用格式^[5]

中文:致谢:本研究使用的***资料由世界数据中心兰州冰川与冻土学科中心提供。英文:Acknowledgement: The *** data in this paper were provided by the WDC for glaciology and geocryology, at Lanzhou. 其中***为作者所引用的数据名称。

(2) 中国西部环境与生态科学数据中心数据引用格式^[6]

用户在使用此数据集时,请在致谢栏里注明:致谢:数据下载于国家自然科学基金委员“中国西部环境与生态科学数据中心, <http://westdc.westgis.ac.cn>”;

用户在使用SPOT VGT-NDVI数据时,须在相关成果的显著位置上明确注明数据来源,我们建议:“Source for this dataset was the VITO (Flemish Inst. Technological Research, Belgium), <http://www.vgt.vito.be>”。

在科学数据现有的元数据中也存在对数据引用情况的描述,但是没有统一的格式,并没有在公开发表物中揭示出来,地质科学元数据标准中有两项表达了数据引用的信息^[7],引用(citation)项是关于数据集名称、日期、版本等的说明资料,提供了该数据集被引用的基本信息,数据志(lineage)项是对从数据源到数据集当前状态的演变过程的说明,包括获取或生产数据使用的数据源(原始资

料)的说明、数据处理过程中的事件、参数、步骤的情况以及负责单位的有关信息等,数据志项说明了该数据集引用其它数据的情况说明。从数据志的内容中可以抽取相关的标识信息,比如数据的来源等。但是这些信息都包含在详细元数据中,一般读者或者使用者无法获取这些信息,因此需要使用一种显性的手段来表达。

5 建立科学数据引用规范的必要性

科学数据引用作为一个基础性的问题由来已久。文献资源的引用比较规范,科学数据资源作为基础性科技资源,其引用机制、规范并没有建立起来,极大地影响了基础性科学数据资源的有效利用。从上述对国内科学数据引用的分析来看,一方面科学数据共享工程由建设阶段转向服务阶段,该工程中汇集的大量的科学数据正在为科学家和科学研究服务,产生了一系列的研究成果,促进了科学数据资源有效的管理和利用;另一方面,大量的科研成果并没有对所使用的科学数据资源进行引用标识,使科学数据资源的使用变得无序,大量有价值的科学数据资源并没有揭示出来,科学数据生产者的知识产权得不到尊重和保护,极大地影响了科学数据机构进行数据共享的积极性。

建立科学数据引用规范,将便于进行科学数据引用统计分析,揭示科学结论和所引数据内容的关联程度,其作用主要体现在:

(1) 通过建立科学数据引用规范,可以反映出论文作者的科学态度和科学结论(成果)具有真实、广泛的科学依据,也反映出该结论的起点和深度。科学技术以及科学技术研究工作都有继承性,现时的研究都是在对数据进行分析的基础上进行的,结论的正确程度通常取决于其引用的数据的情况;因此,在涉及到与数据相关的结论,必然要对数据进行分析和计算,标记数据引用信息即能表明言之有据,提高结论或者成果的可信度乃至深度。这在一定程度上为评估科学结论(成果)的价值和水平提供了客观依据。

(2) 通过建立科学数据引用规范,将规范科学数据引用标记,不仅表明了对数据生产者的尊重,而且保护了数据的知识产权,提高了数据的知名度。

(3) 建立科学数据引用规范将有助于科技情报人员进行情报研究和计量学研究。

6 科学数据引用规范的内容

6.1 数字对象标识技术在科学数据引用规范中的应用

数字对象唯一标识 (Digital Object Identifier, 简称DOI) 为网上数字物体提供独一无二的标识, 提供了指向数字物体本身的链接, 方便了计算机的互操作并可以实现永久指向、动态维护。DOI技术是表达引用最简洁的技术手段, 它将文献和科学数据的引用标识统一起来, 很方便地进行文献和数据的关联, 同时通过链接构建两者之间的关联关系, 揭示引用背后的知识。

DOI将作为科学数据引用规范的重要组成部分, 德国国家图书馆已经开展了这方面的尝试 (见本文第4节), 它将作为表达科学数据引用的重要技术手段, 将规范科学数据引用情况的标注, 为访问原始科学数据提供了稳定的渠道, 从而提供了一种核查结论真实性的手段, 还能方便地链接到与该数据关联的其他数据, 丰富了数据展示的内容。

6.2 科学数据引用规范的基本要求

根据科学数据特点, 科学数据引用的基本要求主要包括:

- (1) **明确数据归属**: 应标明所引用科学数据的发布单位、时间等信息;
- (2) **提供数据来源**: 标明所使用的数据来自于哪个数据库, 采用规范的数据集名称;
- (3) **数据真实准确**: 保证所引用的数据与发表成果 (论点) 匹配, 其数据真实准确;
- (4) **数据权威性**: 最好引用具有权威性的数据, 如国内该行业权威机构产生的数据, 提高科研成果的可信程度;
- (5) **重视科学性**: 应选择具有科学研究性质的数据进行引用, 这样才能体现数据的价值。

6.3 科学数据引用规范的基本格式

科学数据引用格式将列出针对各种科学数据都通用的字段信息, 通过这些信息来唯一确定该数

据, 并能够快速地链接和定位, 为查询、核查该数据的准确性以及数据与观点的一致性提供一个可靠的渠道。

在描述科学数据的信息过程中, 数据的生产者、生产时间、数据及名称、数据发布单位, 以及该数据的唯一标识符是满足科学数据引用要求的信息。因此在科学数据的参考格式中应注明这些信息。

科学数据引用的基本格式描述如下:

数据生产者+数据生产时间+数据库 (集) 名称+数据发布单位+DOI标识码

西南大学 (2006): 海岛棉, 国家自然资源e平台,

doi:10.3416/db.ninr.2151C0001A00014341

各字段的含义表示为:

西南大学: 代表数据生产单位, 该数据产权所有者;

(2006): 该数据集创建年份, 标识到年即可;

海岛棉: 该数据集的名称;

国家自然资源e平台: 该数据集发布者名称
doi:10.3416/db.ninr.2151C0001A00014341: 该数据集的DOI标识码, 在该字段上附有链接, 能查看该数据集的详细信息。

6.4 建立参考数据列表

参考文献是一篇规范的学术文献的重要组成部分, 随着科学数据的大量使用, 本文认为建立在科学数据引用格式基础上的参考数据列表也理应成为学术文献的有机组成部分, 参考数据将列出在文献的研究和写作中直接利用的或有直接帮助和启发作用的科学数据, 而不是作者用过的全部科学数据。所列的科学数据包括数据集、数据库、数据产品、数据记录等内容, 如果列出的是数据产品还应能追溯到数据的来源 (data source), 能反映出数据变化的历史, 提供核查数据真实性和准确性的重要途径。如果是数据库, 最好能列出该数据库包含的主要数据集的名称、类型等信息, 方便了解整个数据库的全貌。

7 科学数据引用规范的推广价值及建议

科学数据引用是促进科学数据规范使用的重要手段之一,虽然目前并没有有效地建立这种机制,但是从数据生产者、使用者、管理者和政府的角度进行分析,已经突显其推广价值。

(1) 从数据生产者的角度

通过对其数据的引用情况进行标识,可以提高数据的引用程度,提高该机构在国内甚至国际上的知名度和权威地位。

(2) 从数据使用者的角度

通过对数据引用情况的标识和说明,能够方便地链接到相关的数据,对结论和数据之间的联系进行核查,是一种很有效的研究手段和途径。

(3) 从科学数据共享与管理者的角度

科学数据引用成为科学数据共享管理的重要内容,科学数据有效引用量是衡量科学数据共享效率的重要指标,

(4) 从政府的角度

在数字对象标识技术的基础上,需要不断完善科学数据引用机制,通过出台鼓励政策,在使用数据的时候倡导一种标识所引用科学数据的习惯,增强科学数据知识产权保护的意识,根据上述参考数据基本格式规范,形成文献(数据)的参考数据体系,同时能通过技术的手段对引用情况进行跟踪和反馈,更有效地提升科学数据的价值,有利于构建科学数据的学术规范和科研诚信体系。

科学数据引用的研究还处在初期阶段,并需要在实践中进行应用和完善,综合考虑可行性和复杂程度的基础上,建议科学数据引用机制可以首先在如下领域中进行推广:

(1) 在国家科技基础条件平台中建立科学数据引用的机制

为863、973等科技计划项目产生的科学数据赋予DOI号,将该项目产生的数据的引用情况作为衡

量此项目共享情况的重要指标之一,同时在发表与项目相关的科技文献时必须按照科学数据的引用格式注明其参考数据的出处。

(2) 在理工科学学位论文中建立科学数据引用的机制

选择与科学数据使用密切相关的研究领域作为试点,将标注参考数据信息列入论文写作规范,在自然科学领域涉及到数据分析和应用的学位论文必须标注其所引用的数据,并将其作为论文写作规范与否的一种重要的评价指标,评审者可以通过数据的引用信息方便地链接到原始数据(适用于已共享的数据),并对所引用的数据进行核查,判断所引数据与结论之间的关联和匹配程度。

(3) 建立科学数据中心与科技期刊的联盟

在期刊中建立科学数据引用试点,明确规定与科学数据使用密切相关的论文在发表的同时必须标注其参考数据,同时建立畅通的渠道,便于作者在研究过程中产生的有价值的数据或者首次发表的科学数据在相应的科学数据机构进行登记注册,为其赋予唯一标识符,扩大该数据的影响力,方便用户进行检索。

8 结语

科学数据规范引用是一个新的议题,对该部分的研究还处于初步阶段,但是在目前科技信息广泛应用的背景下,参考文献的信息已经不能完全涵盖所引用的全部科技信息的内容,大量有引用价值的科学数据虽被引用却没有进行标注,因此建立科学数据引用的规范和标准对于规范数据的使用、提升科学数据的价值、保护科学数据知识产权具有十分重要的价值和意义。

参考文献

- [1] 苏广利,许新军.社科论著参考文献引用中的七种不良行为[J].图书馆工作与研究,2002(2):15-17.
- [2] 傅晓峰.关于促进科学数据共享管理的一些思考[J].中国基础科学,2006(6):17-19.
- [3] 安迎建.数据库原理与应用[M].北京:中国石化出版社,2004.
- [4] PASKIN N. Digital Object Identifiers for scientific data[C]//19th International CODATA Conference, Berlin, Nov 10, 2004.
- [5] 世界数据中心冰川(雪冰)冻土学科中心.数据引用说明[EB/OL]. [2008-07-01]. <http://wcdetg.westgis.ac.cn/>.
- [6] 中国西部环境与生态科学中心.长时间序列中国植被指数数据集介绍

[EB/OL]. [2007-07-01]. <http://westdb.westgis.ac.cn/>.

[7] 戴爱德,李德兴.WDC中国地质科学数据网的核心元数据[J].国土资源信息化,2004(1):8-11.

作者简介

彭洁(1965-),女,湖北武汉人,中国科学技术信息研究所资源共享促进中心主任,研究员,硕士生导师,研究方向为科技资源共享、信息组织和知识管理。通讯地址:北京市海淀区复兴路15号资源共享促进中心 100038

(下转45页)

势,从以下几个方面去实现图书馆2.0平台在用户体验方面的自我提升,增加用户对网站的忠诚度:

(1) 便捷性:传统的图书馆服务种类甚多,如:书刊检索、馆际互借、参考咨询、文献传递等,在网站中将这服务资源进行集成,用户通过网站平台的一站式入口即可完成基本的业务流程,享受图书馆服务;

(2) 功能性:图书馆的数字资源非常丰富,让用户随心所欲地对各种数据库和网页添加标签并记录用户的行为历史,帮助用户在每一次使用平台的时候都可以迅速定位所需的网站内容以及数字资源;

(3) 娱乐性:图书馆是文化传播的主要媒介,大量的资源除了具有极高的收藏价值外,也是文化娱乐的最佳形式,如上海图书馆收藏的各类历史图片库、经典电影和音乐的视听资料等。在My favorite中建立My movie、My music、My photo,让用户充分体验网络文化生活的无限乐趣。

(4) 参与性:与用户建立良好的沟通机制,对

用户使用过程中的满意度进行调查分析,征集用户对于网站的意见,作为网站改良优化的参考并加以实施。同时提供友好的交互界面,如Msn实时咨询机器人、QQ在线服务等,增强网站互动性,提高用户粘度和网站价值。

5 结语

将SNS模式引入图书馆2.0,把“以用户为中心”的理念真正贯彻于图书馆服务平台中,建立用户主导^[5]的服务模式,不仅顺应了Web2.0用户在信息需求和信息行为方面的变化,提高了图书馆对用户的粘度,而且极大的提升了网站在用户体验方面的核心价值。虽然目前图书馆在SNS应用方面的实践仍处于摸索阶段,但是随着Web2.0理念和技术不断发展,以用户交流为纽带、用户体验为核心的图书馆服务将会成为未来图书馆2.0的主要研究方向。

参考文献

- [1] 范并思,胡小菁.图书馆2.0:构建新的图书馆服务[J].大学图书馆学报,2006(1):2-6.
- [2] Danny.如何提高用户粘度[EB/OL]. [2008-05-01]. <http://column.iresearch.cn/u/bjdj/archives/2007/1010.shtml>.
- [3] 李宋伟. web2.0网站成功三要素 [EB/OL]. [2008-05-01]. <http://www.weamax.com/articles/17/2007-10/20071029082349.html>.
- [4] 阮京文.改善网络环境对用户体验的影响 [EB/OL]. [2008-05-01]. <http://news.iresearch.cn/0468/20080305/77281.shtml>.
- [5] 图书馆2.0工作室.图书馆2.0:升级你的服务[M].北京:北京图书馆出版社,2008.
- [6] 刘炜,葛秋妍.从Web2.0到图书馆2.0:服务因用户而变[J].现代图书馆情报技术,2006(9):8-12.
- [7] 叶鹰.开源互动理念指引图书馆2.0研究——兼论图书馆2.0与个性化图书馆的结合[J].图书馆杂志,2007(8):15-18.
- [8] BURNETT K.核心客户关系管理[M].北京:电子工业出版社,2002.
- [9] Ball State University Libraries[OL]. [2008-05-01]. <http://www.myspace.com/brackenlibrary>.
- [10] Brooklyn College Library[OL]. [2008-05-01]. <http://www.myspace.com/brooklyncollegelibrary>.
- [11] Hennepin County Library[OL]. [2008-05-01]. <http://www.Myspace.com/hennepincountylibrary>.
- [12] Barnard College Library[OL]. [2008-05-01]. <http://www.Myspace.com/barnardzinlibraryl>.

com/barnardzinlibraryl.

作者简介

金家琴(1980-),女,华东师范大学信息系毕业,硕士。通讯地址:上海图书馆文献提供中心 200031

Study on the Construction of SNS in Library Based on User Value

Jin Jiaqin / Shanghai Library, Shanghai, 200031

Abstract: Information needs and behaviors of library user have been dramatically transformed with the evolution of Web 2.0 service. It is a great challenge for the library to renovate service idea and service model by the applications of Web 2.0 technology. Considering user's viscosity and degree for meeting user demand, this article analyzed SNS's advantages in the library service renovation context as a core application of Web2.0 technology. Combining main features home of service in Library SNS application, this article explores the approach to implement SNS system and research on how to enhance user's viscosity by improving user experience.

Keywords: Web 2.0, Service mode, User viscosity, User value, Social Network Services, Digital library

(收稿日期:2008-05-07;责任编辑:贾廷霞)

Research on the Scientific Data Citation

Peng Jie, Tu Yong / Institute of Scientific & Technical Information of China, Beijing, 100038

Abstract: Scientific data citation is one of the problems that puzzled the science data field for a long time. The paper started with the definition of scientific literature citation, generalized the definition of scientific data citation, research objects and problems based on the characters of scientific

data. Then it analyzed the status of scientific data citation both at home and abroad, discussed the need of the building the scientific data citation standard, finally summarized the content of the standard, analyzed the promotion significance, and gave the suggestions for promoting the scientific data citation in China.

Keywords: Scientific data, Citation standard, Digital object identifier, DOI, Scientific literature, Digital library

(收稿日期:2008-07-02;责任编辑:贾廷霞)

(上接18页)