科学数据引用规范的研制

王卫华 胡良霖 沈志宏
（中国科学院计算机网络信息中心，北京 100190）

摘 要：科学数据的引用可以明确数据归属，通过引用量的分析可以对数据的科研价值进行客观评价，这对促进数据的传播和再利用将起到很好的推动作用。本文首先介绍了国内外数据引用方面的研究进展；然后借鉴国际上已有的研究成果，针对中科院“数据应用环境建设与服务”项目的数据管理特点，制定了科学数据引用规范；最后对规范的实施情况进行了总结并提出了进一步研究方向。

关键词：科学数据；数据引用；引用规范；数据管理
中图分类号：G35 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2013.01.007

Development of Citation Specification for Scientific Data

Wang Weihua, Hu Lianglin, Shen Zhihong
(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The scientific data citation can clear data vesting and objective evaluate the scientific value of the data by analyzing the citation amount, which will play a good role in promoting the spreading and reuse of the data. This paper first introduces the domestic and international data citation research progress; then draw on existing international research results and consider data management features of the “CAS environment construction of data applications and services "project to develop the scientific data citation specification; Finally, summarize the specification implementation and further research directions.

Keywords: Scientific data, data citation, citation specification, data management

1 引言

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[1]。其中很大部分的科学数据是通过长期观测或工作积累而获得的，具有极高的科研价值，为相关领域的科研工作提供重要参考和支持。在科学数据共享过程中，由于缺少类似于传统文献的引用机制，科学数据的使用情况和使用效果无法衡量，导致科学数据的创建者、管理者或保存者的活动绩效无法评估，他们的科研价值得不到很好体现。这在很大程度上阻碍了科学数据生产者和管理者的工作热情[2-3]。科学数据引用课题的研究正是在这种需求下产生的。通过建立类似于传统文献的引用规则，可以明确科学数据的归属；通过引用量的分析，可以对科学数据的科研价值进行客观评价。这对促进科学数据的传播和再利用将起到很好的推动作用。

国外大型科学数据机构与科研组织对数据发布和引用模式以及基础环境建设进行了积极的探索，并取得了初步成果，对于科学数据的引用已经具有规范的格式。而国内在该领域的研究相对落后。因此，针对国内科学数据的管理特点，本文将重点对国内外科学数据引用格式和要求进行研究和探讨，提出适合我国科学数据管理的科学数据引用规范，以推进科学数据规范化引用。

第一作者简介：王卫华（1978- ），女，工程师，硕士，主要研究方向：数据库技术与标准规范、数据应用和服务。
收稿日期：2012年9月24日。

2 国内外科学数据引用格式研究状况

国外科学数据引用格式在不同研究领域和项目中有不同规定。下面重点介绍STD−DOI、DataCite、PANGAEA、ICPSR、Dataverse Network等项目或机构的引用格式。

（1）STD−DOI引用格式

在STD−DOI项目[4]中，其引用元素包括Creator(s)、Publication year、Dataset name、Publisher、Persistent identifier几个基本元素。其引用格式如下：

Creator(s)(publication year): Data set name, Publisher. Persistent identifier.

引用示例如下：

Kamm, H; Machon, L; Donner, S (2004): Gas Chromatography (KTB Field Lab), GFZ Potsdam. doi:10.1594/GFZ/ICDP/KTB/ktb−geoch−gaschr−p.

（2）DataCite引用格式

DataCite项目的数据引用基本格式包括Creator、Publication Year、Title、Publisher、Identifier 5个必选元素，另外还有Version、ResourceType两个可选元素。完整的引用格式为：

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

其中，Identifier可以以原有的格式出现，或者是以可访问的http方式出现。引用示例如下：

Geofon operator (2009): GEFON event gfz2009 kciu (NW Balkan Region).

GeoForschungsZentrum Potsdam (GFZ).

doi:10.1594/GFG.GEOFON.gfz2009kciu.

http://dx.doi.org/10.1594/GFZ.GEOFON.gfz2009 kciu

（3）PANGAEA引用格式

PANGAEA[5]的引用格式为：

Creator (PublicationYear): Title, Publisher, Identifier.

引用示例如下：

Stein, R.; Fahl, K. (2003): Distribution of grain size and clay minerals in surface sediments of the Kara Sea, PANGAEA, doi:10.1594/PANGAEA.119754.

（4）ICPSR引用格式

ICPSR的引用格式[6]较为复杂。其引用格式继承了“社会科学数据标准Data Documentation Initiative”中的一些元素规定。引用格式如下：

Author(s). Title [computer file]. City and state of the producer. Name of the data producer [producer]. Year. Unique identifier. City and state of the data distributor, distributor [distributor].date.

其中，[computer file]指该数据集为电子格式的数据，Year指数据产生的时间，date指数据发表日期（能够从网上获得的时间）。引用示例如下：

Earls, Felton J., Jeanne Brooks−Gunn, Stephen W. Raudenbush, and Robert J. Sampson. PROJECT ON HUMAN DEVELOPMENT IN CHICAGO NEIGHBORHOODS (PHDCN): CHILD BEHAVIOR CHECKLIST, WAVE 1, 1994−1997 [Computer file]. ICPSR13582−v1. Boston, MA: Harvard Medical School [producer], 2002. Ann Arbor, MI: Inter−university Consortium for Political and Social Research [distributor], 2005−07−22.

（5）Dataverse Network引用格式

The Dataverse network project规定，在网络应用环境下，基本格式包括6个必须元素：Author、Date、Title、Unique global identifier、Universal Numeric Fingerprint(UNF)、Bridge service和一个可选元素Value[fieldname]。其中，Bridge service为URL格式，通过Unique global identifier进行链接。fieldname默认来源于DC的type−vocabulary，如果来源于其他的元数据scheme,则需要特别注明。如果为打印格式，Bridge service为可选。Dataverse Network引用格式[7−8]为：

Author, Date, “Title”, Unique global identifier Universal Numeric Fingerprint(UNF) Value[fieldname]

引用示例如下：

Gary King; Langche Zeng, 2006, “Replication Data Set for ‘When Can History be Our Guide? The Pitfalls of Counterfactual Inference’” hdl:1902.1/DXRXCFAWPK

UNF:3:DaYlT6QSX9r0D50ye+tXpA==

Murray Research Archive [distributor]

国内科研领域已经意识到科学数据引用的重要性，也具有科学数据引用的需求。世界数据中心冰川（雪冰）冻土学科中心[9]要求在使用该中心数据时注明“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”。中国西部环境与生态科学数据中心[10]要求用户在使用此中心数据集时在致谢栏里注明“致谢：数据下载于国家自然科学基金委员会中国西部环境与生态科学数

据中心，http://westdc.westgis.ac.cn”。中科院计算机网络信息中心国际科学数据镜像网站[11]要求使用者在发表论文成果时注明“数据来源于中国科学院计算机网络信息中心国际科学数据镜像网站(http://datamirror.csdb.cn)”。

从上述科学数据引用格式研究状况分析中可以看出，国际上大型组织机构和科研项目制定的引用规范已经具有规范的格式，引用内容已经细化到科学数据的名称、作者、时间等属性。同国外相比，国内科研领域虽然意识到科学数据引用的重要性，并且在各自的数据资源网站规定了引用注意事项和要求，但没有形成统一的引用格式，引用内容相对宽泛，不能反映科学数据本身的特性。另外，国内的数据引用规范要求中很少体现数据生产者个人的价值和贡献，仅标明了科学数据的发布机构。

国外科学数据引用规范元素的设置对比如表1所示。从表1中可以看出，这些规范的设置虽然繁简不同，但都包含数据创建者、数据发表时间、数据集名称、数据发表机构、唯一标识符这几个必选元素。这反映了国际社会对这几个元素的设置具有高度认同。这几个元素回答了科学数据引用中最核心的问题，即有关科学数据的标识、归属、来源和时间的问题。

3　中国科学院的数据管理

多年来，中国科学院“数据应用环境建设与服务”项目形成了院内跨所联合，共建共享的组织管理模式。中国科学院计算机网络信息中心是项目总承担单位，负责协调实施数据资源中心建设和科学数据库建设，为数据应用环境提供技术支撑服务，同时负责整个项目的门户系统建设，为数据的发布和访问提供总平台，但不拥有数据所有权。中科院各研究所负责数据资源建设，对各自承建的数据资源拥有所有权和使用权，同时可以独立对外提供数据服务。

这样的项目组织方式是一种多层次、权限交

表1　国外科学数据引用元素对比

	STD−DOI	DataCite	PANGAEA	ICPSR	Roper Center	Dataverse Network
作者（数据创建者）	Creator(s)	Creator (s)	Creator (s)	Author(s)	Authorship	Author(s)
数据发表（创建）年份	Publication year	Publication Year	Publication Year	Year the data was produced	Production date	Date the data set was published or otherwise made public.
数据（集）名称	Dataset name	Title	Title	Title [material designator]	Title[material designator]	Title
数据版本		Version			Edition/Version	
数据发表（创建）机构	Publisher	Publisher	Publisher	City and state of the producer	Location of the producer	Value[fieldname]
				Name of the data producer [producer]	Name of the data producer[producer]	
数据类型		ResourceType		[material designator]	[material designator]	
永久标识（唯一标识符）	Persistent identifier	Identifier	Identifier	Unique identifier		Unique global identifier bridge service
						Universal Numeric Fingerprint(UNF)
数据传播机构				City and state of the data distributor	Location for the distributor	
				Distributor [distributor]	Distributor [distributor]	Value[fieldname]
数据传播时间				Date the data was made available in the repository	Distribution date	
						Bridge service

叉、约束松散的形式，项目最终成果（主要是数据）的所有权和使用权与项目管理中的权责角色严重脱离。这种情况使得在科学数据引用中所涉及的角色较多，除了对数据本身属性的描述外，还需要理清各方对数据的干预。

4 科学数据引用规范的制定

本文学习借鉴了国际科研组织制定的数据引用规范，在设计上继承了那些必选元素，保持与国际社会一致。同时在分析“数据应用环境建设与服务”项目数据管理特点的基础上，补充选取了其他元素，以便明确不同角色对数据的影响和贡献。

4.1 元素设置

对科学数据的引用包括作者、名称、发布机构[发布机构]、发布年份、传播机构[传播机构]、传播时间、唯一标识符和解析地址等8个必选元素以及版本等1个可选元素。

前4个元素与传统文献引用元素含义相同，作者表示数据（集）的创建者，可以是单位或者个人；名称即该数据集的名称；发布机构类似于文献引用中的发表刊物，即公开该数据（集）的机构，该元素后面跟[发布机构]限定词；发布年份表示该数据（集）公开的年份。这4个元素可以使读者快速了解该数据集的特征，明确数据的归属。

传播机构即该数据集的存储和传播机构，用来说明该数据集从何处获得，该元素后面跟[传播机构]限定词；传播时间即该数据集可以从存储传播机构获取该数据的起始时间。 这两个元素对数据的来源进行了明确。

唯一标识符是一个全球唯一的字符串。该字符串可以独立于科学数据的存储位置而唯一永久地标识一条科学数据，从而建立起科学数据引用信息与真实数据之间的永久联系。

最后一个必选元素是解析网址。由于只有少数唯一标识符（比如DOI）得到部分浏览器的支持，大多数唯一标识符，尤其是领域或机构内部的唯一标识符不能被浏览器直接识别和解析，因此需要明确指定唯一标识符的解析网址。解析网址为URL格式，由解析服务网址和唯一标识符两部分组成，例如http://citation.csdb.cn/是用于唯一标识符解析服务的网址，cn. csdb. datamirror. LE71230392011343EDC00是“科学数据库及其信息系统”项目（缩写：CSDB）分配给该数据（集）的唯一标识符，那么该条数据（集）完整的解析网址为：http://citation.csdb.cn/CSDB:cn. csdb. datamirror. LE71230392011343EDC00。在浏览器中输入该解析网址，就可以获得所标识的数据对象。如果唯一标识符中出现需要转义的字符集，需要将这些字符转义为浏览器支持的字符。当引用信息在网页中显示时，可以将解析网址作为唯一标识符的超链接。

除必选元素外，该标准定义了一个可选元素，即版本。版本用来标识同一数据集的不同版本。本规范建议将同一数据集的变动视为不同的版本，但在新版本的元数据应包含指向最初版本的链接。对于一些海量的数据集，新版本可以仅包含原始数据的变化。

4.2 引用格式

一条由必选元素组成的引用信息格式为：

作者.名称.发布机构[发布机构],发布年份.传播机构[传播机构],传播时间.唯一标识符;解析网址.

示例：

NASA.500m地表反射率8天合成产品.NASA[发布机构],2007.中科院计算机网络信息中心[传播机构],2009−03−02.csdb:cn.csdb.datamirror.LE71230392011343EDC00;http://citation.csdb.cn/csdb:cn.csdb.datamirror.LE71230392011343EDC00.

一条由全部元素组成的引用信息格式为：

作者.名称(版本).发布机构[发布机构],发布年份.传播机构[传播机构],传播时间.唯一标识符;解析网址.

示例：

中国科学院地理科学与资源研究所. 人地系统主题数据库元数据标准(V2.0). 中国科学院地理科学与资源研究所[发布机构],2010.中科院计算机网络信息中心中心[传播机构], 2011−01−19.csdb:cn.csdb.TR−REC−015−01;http://citation.csdb.cn/csdb:cn.csdb. datamirror. LE71230392011343EDC00.

5 总结

科学数据引用研究仍处于探索阶段，国际上一些大型的组织机构和科研项目对该课题进行了积极探索并取得了初步的研究成果，为进一步深入研究提供了经验和参考。本文借鉴国外研究成果，同时分析了“数据应用环境建设与服务”项目数据管理的特点，制定了科学数据引用规范。

目前，该引用规范已应用于中国科学院“数据

应用环境建设与服务”项目中，通过系统自动生成了项目内500多个数据库（数据集）的引用信息。用户进行数据检索时，在每一条检索结果中同时提供了该条数据的引用信息，以方便用户进行引用标注。

数据应用规范的推广和应用，离不开数据发布和引用系统的支持。首先必须保证数据能够被永久地标识，另外还需要有一个可靠的、可以长久定位的地址。其次，在完善科学数据发布和引用技术的基础上，需要研究和建立科学数据引用机制。通过出台鼓励政策，倡导引用科学数据的习惯，增强科学数据知识产权保护的意识。可以通过推进数据中心与学术期刊的合作，在学术文章发表的同时提交相关研究数据，建立文献与数据之间的交叉引用。

参考文献

[1] 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.

[2] Chavan VS, Ingwersen P. Towards a Data Publishing Framework for Primary Biodiversity Data: Challenges and Potentials for the Biodiversity Informatics Community[J/OL]. [2012−08−20]。 http://www.biomedcentral.com/1471−2105−10/s14/s2.

[3] 彭洁,涂勇.科学数据引用的探讨[J].数字图书馆论坛,2008(10):14−18.

[4] The German National Library of Science and Technology (TIB). Publication and Citation of Scientific Primary Data [EB/OL]. [2012−08−20] www.std−doi.de.

[5] Alfred Wegener Institute for Polar and Marine Research (AWI). PANGAEA [DB/OL]. [2012−08−20] http://www.pangaea.de/.

[6] NSCU LIBRARIES. How to Cite a Date Set[EB/OL]. [2012−08−20]. http://www.lib.ncsu.edu/data/citingdatasets.html.

[7] King Gary. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing[J]. Sociological Methods and Research, 2007(36):173−199.

[8] Micah Altman, Gary King. A Proposed Standard for the Scholarly Citation of Quantitative Data. D−Lib Magazine [J/OL]. [2012−08−20] http://www.dlib.org/dlib/march07/altman/03altman.html.

[9] 中国科学院寒旱所遥感室.世界数据中心冰川(雪冰)冻土学科中心[DB/OL]. [−−]. http://wdcdgg.westgis.ac.cn/chinese/index.htm.

[10] 中国科学院寒区旱区环境与工程研究所.中国西部环境与生态科学数据中心[DB/OL]. [2012−08−20]. http://westdc.westgis.ac.cn/data/fecdec71−77d1−43c2−b472−8b1c729874cb.

[11] 中国科学院计算机网络信息中心. 国际科学数据服务平　台[DB/OL]. [2012−08−20]. http://datamirror.csdb.cn/dem/files/gc.jsp.

（上接第35页）

[8] 未来互联网报告[EB/OL].[2011−05−06].https://connect.innovateuk.org/c/document_library/get_file?folderId=861750&name=DLFE−33761.pdf.

[9] 全球大型强子对撞机网格(WLCG)[EB/OL].[2012−09−26].http://wlcg.web.cern.ch/.

[10] 欧洲数据保护指令[EB/OL].[2012−09−26].http://ec.europa.eu/justice/data−protection/index_en.htm.

[11] Koopa, David, et al.基于数据溯源的基础设施支持可执行文件的生命周期[EB/OL].[2012−09−26].http://vgc.poly.edu/~juliana/pub/vistrails−executable−paper.pdf.

[12] 联合国教科文组织欧洲委员会.开放访问:机遇与挑战[EB/OL].[2012−09−26].http://ec.europa.eu/research/science−society/document_library/pdf_06/open−access−handbook_en.pdf.

[13] OpenAIR – 开放访问欧洲研究基础设施[EB/OL].[2012−09−26].http://www.openaire.eu/.

[14] 开放性研究者与贡献者ID[EB/OL].[2012−09−26].http://about.orcid.org/.

[15] 数据生命周期模型与概念[EB/OL].[2012−09−26].http://wgiss.ceos.org/dsig/whitepapers/Data%20Life−cycle%20Models%20and%20Concepts%20v8.docx.

[16] EGI 联合云任务组[EB/OL].[2012−09−26].http://www.egi.eu/infrastructure/cloud/cloudtaskforce.html.

[17] eduGAIN – 网络服务和应用的联合访问[EB/OL].[2012−09−26].http://www.edugain.org.

[18] Demchenko Y, Ngo C, Makkes M, et al.定义互联云架构的互用性和集成性[C]. 2012年第三届国际云计算、网格和虚拟化云计算大会,法国尼斯,2012年7月22−27日.

[19] 云参考框架[EB/OL].[2012−06−27].http://www.ietf.org/id/draft−khasnabish−cloud−reference−framework−03.txt.

[20] eduroam[EB/OL].[2012−06−27].http://www.eduroam.org.

[21] Shibboleth – 开源联合身份管理系统[EB/OL].[2012−06−27].http://shibboleth.net/.

[22] CILogon Service[EB/OL].[2012−06−27].http://www.cilogon.org/.