关于参考文献中数据集著录格式的研究*

陈 庆 陆炳新
南京师范大学学报编辑部,210097,南京

摘 要 随着大数据时代的到来,在现代化计算机科学应用技术环境下,数据集已成为众多科学研究课题项目中必不可少的重要组成部分。尤其是1980年以来数据集越来越多地受到关注、研究和引用。本文介绍了数据集的发展,分析了数据集对科技期刊论文的写作作用,然后通过具体数据表明数据集在科技期刊中的引用情况,最后针对当前数据集在参考文献中的著录格式进行了探讨和分析,建议网络资源中数据集在参考文献中的著录格式可表示为:[序号] 主要责任者. 题名:其他题名信息[DS/OL]. 制作地: 制作单位, 制作年份(更新或修改日期)[引用日期].获取和访问路径.

关键词 数据集;参考文献;著录格式;大数据;GB/T 7714—2015

Research of documenting formats of date set in references // CHEN Qing, LU Bingxin

Abstract By the era of big data, with the application of modern computer science and technology, the data set has become an indispensable component of numerous scientific projects. Particularly, the data set has been concerned, studied and referenced since the 1980s. Firstly, we describe the development of data set, analyze the role of writing of data set in scientific journals, then demonstrate the status of referencing of the data set in sci-tech journals. Finally, we suggest the documenting format for the data set in the web: [Number] Primary authority. Title: Additional title information[DS/OL]. Locality of Production: Production department, Production date (Update or modification date)[Date of reference]. Access path.

Keywords date set, reference, documenting format, big date, GB/T 7714-2015

Author's address Editorial Board of Journal of Nanjing Normal University, 210097, Nanjing, China

DOI:10.16811/j.cnki.1001-4314.2017.01.014

进入21世纪以来,随着各种信息技术的飞速扩散,各领域的生产规模、数据种类、数据规模都以前所未有的速度飞速增长[1]。从1980年美国未来学家托夫勒预测“大数据”在“第三次浪潮”中的重要角色,到2009年“大数据”作为继云计算后的又一科技热点而引起各界瞩目。现实表明,“大数据”不仅是学术界公认的下一个创新前沿,更是商业界的另一个热门市场,还是各国政府关注的战略领域,许多外国媒体和专家将2013年称为“大数据元年”[2-4]。随着大数据时代的到来,各学科产生的数据集(Date set,DS)的数量与日俱增。数据集,又称为资料集、数据集合或资料集合,是一种由数据所组成的集合。在GB/T 7714—2015[5]中,新增了4个文献类型及其标识:档案(A),舆图(CM),数据集(DS),其他(Z)。而关于数据集(DS)在参考文献中的著录格式尚未给出,本文将对此进行探讨。

1 数据集对科技期刊论文写作的作用

数据集基本上由数据集名称、数据项目、存取方式和数据构成[6]。从中国知网数据看,在1978年《羊毛工业研究协会纺织数据集(一)》[7]中最早出现了“数据集”字样。当时数据集以表格形式出现,每一列代表一个特定的变量,每一行都对应于某一成员的数据集的问题。如今,由于数据集中数据元庞大,数据集无法在某一篇文章中呈现出来。例如《中国考古学中碳14年代数据集(1965—1981)》[8]中指出,中国社会科学院考古研究所将数据汇编成集,以书籍的形式反映数据,以便于研究工作的检索和引用。

在大数据时代,数据集存在形式早已不像早期那样只是一个简单的表格。在数据库中,数据集的结构类似于关系数据库的结构,它包含了为数据集定义的约束和关系等。在统计学中,数据集一般来自实际观测得到的抽样统计人口,每一行对应于观测的一个组成部分。另外,可能会嵌入产生子数据集的软件。例如,存在数据中的经典数据集PSPP、人口人力资源数据统计学数据集等。在科技期刊论文写作中,数据集主要有如下作用。

1.1 统一了分歧数据 例如考古学中碳14年代数据。在全国建立了30多个碳14实验室,每个实验室在不同时间测定并分批发表数据,共产生了20来批1 000多个年代数据。这些数据既不便于相互对照,又难于收集备用,对研究工作的参考和碳14年代数据的引用带来了相当的不便;因而中国科学院考古研究所将所有散见于各处而数据又不一致的碳14年代数据汇编成集,以便于研究工作的检索和引用。

1.2 实现了资源共享 数据集的建立为科技论文提供了全面、可靠的参考数据。一方面,数据集的共享可以避免重复劳动。据统计,我国科研项目重复率高达40%,而另外60%中部分重复率在20%以上,已造成了人力、物力、财力的严重浪费[9-10]。如果我们能对同一项目建立数据集,进行研究汇总甚至合作研究,那么,既

* 2015年度江苏省期刊协会研究课题(2015JSQKA006)

可避免不必要的重复劳动,还可聚集更多的资源,得到更好、更多的研究成果。另一方面,数据集更好地保护了知识产权。一个数据集的建立,可以是某几个作者、科研单位、组织机构。数据集的公开有其著作权,它的产生受时代背景、地理位置、研究设备、关键技术等条件的影响,当某个研究项目出现一种新的成果,需要做出新的数据集时,旧的数据集将为其提供参考数据。当数据集被直接引用时,我们应尊重作者的著作权。

1.3 提高了工作效率 数据集更能全方位地反映某个项目的众多要素。科技论文写作时,可以从著作中或网络资源中的数据集中查找。在大数据时代,数据元更新速度快,数据集多以电子格式存在,这样我们可以从计算机中直接检索即可,避免逐一查找纸质目录,省时省力,提高了工作效率。此外,通过大量数据集的整合、分析,能够得出一套行为预测模板,更准确地提供我们的意向行为。

2 数据集在科技期刊中的引用情况

随着大数据时代的到来,在现代化计算机科学应用技术环境中,数据集已成为众多科学研究课题项目必不可少的组成部分。数据集是科技期刊的热点词汇之一,也是当前国内计算机软件及计算机应用、自动化技术、互联网技术、数学、生物学、自然地理学和测绘学等领域研究的重要内容。笔者以"数据集"从关键词角度进行检索,就 2011—2015 年间,以"数据集"为关键词文献的学科主要是计算机软件及计算机应用、自动化技术,分别占总文献量的 55.7%和 18.2%。

另一方面,就数据集在科技期刊中的引用和研究情况做以下研究。在具体数据来源及选取上主要以 CNKI 中国学术期刊网络出版总库为主要数据来源[11]。以"数据集"从关键词途径进行检索,就 1915 年以来以"数据集"为关键词的文献数量增长情况进行分析。在 5 个时间段即 1915—1975 年、1976—2000 年、2001—2005 年、2006—2010 年、2011—2015 年内文献的增长量分别为 6、675、1 168、3 397、2 565。可见作为"第三次浪潮"的"大数据"时期的到来对数据集的文献数量逐年增加,且增长得越来越快,其中 2006—2010 年间增长最多,而 2011 年以来增长逐渐缓慢。数据集增长量与日俱增充分说明了它在科技期刊中的不可小觑的地位和作用。

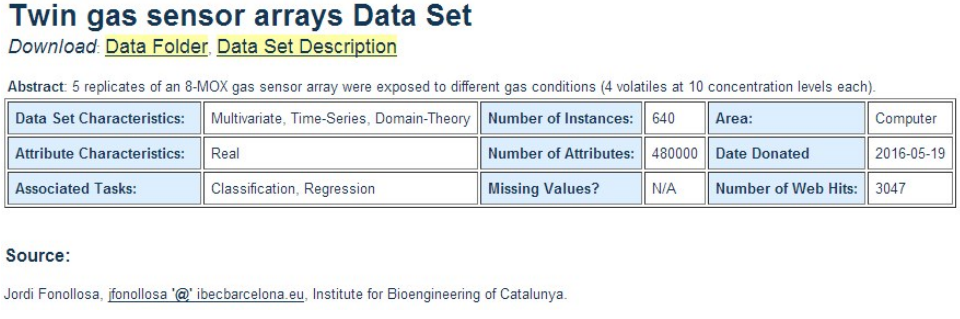
3 数据集在参考文献中的著录格式

目前,数据集大致以 3 种形式出现:期刊中析出的数据集、普通图书中的数据集和网络资源中的数据集。下面对上述形式的著录格式加以分析。

3.1 期刊及普通图书中析出的数据集 期刊及普通图书中析出的数据集的表达形式已有明确的著录格式,分别按期刊和普通图书的著录格式著录即可,即无须出现文献类型标识"DS"。在此不再举例细述,可见文献[12-15]。

3.2 网络资源中的数据集 网络资源中的数据集尚无明确的著录格式,下面通过实例进行探讨。研究发现,网络资源中外文数据集的共享资源比较丰富,结构形式相对统一。其基本结构为:数据集名称、数据来源介绍、数据摘要、关键词、数据格式、数据详细信息、数据预览等。在"数据详细信息"中一般可以准确了解到该数据集的基本情况。参照 GB/T 7714—2015[5] 中电子文献的著录要求,电子文献的著录项目主要包括:主要责任者,题名项(题名,其他题名信息,文献类型标志(含文献载体标志)),出版项(出版地,出版者,出版年,更新或修改日期,引用日期),获取和访问路径。建议网络资源中的数据集的著录格式可表示为:[序号] 主要责任者.题名:其他题名信息[DS/OL].制作地:制作单位,制作年份(更新或修改日期)[引用日期].获取和访问路径。

例如,图 1 为"Twin gas sensor arrays Data Set"中的部分截图。



Twin gas sensor arrays Data Set

Download: Data Folder, Data Set Description

Abstract: 5 replicates of an 8-MOX gas sensor array were exposed to different gas conditions (4 volatiles at 10 concentration levels each).

Data Set Characteristics:	Multivariate, Time-Series, Domain-Theory	**Number of Instances:**	640	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	480000	**Date Donated**	2016-05-19
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	3047

Source:

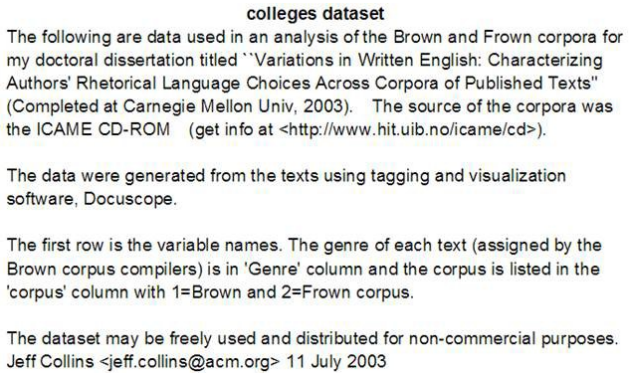
Jordi Fonollosa, jfonollosa '@' ibecbarcelona.eu, Institute for Bioengineering of Catalunya.

图 1 "Twin gas sensor arrays Data Set"中"abstract"和"source"

由图 1 可知数据集的基本信息,建议将该数据集表示为:

[1] Fonollosa J. Twin gas sensor arrays dataset [DS/OL]. Catalunya: Institute for Bioengineering of Catalunya, 2016[2016-06-06]. http://archive.ics.uci.edu/ml/datasets/Twin+gas+sensor+arrays.

例如,图 2 为某“学院数据集”中的“数据详细信息”的部分截图。



colleges dataset

The following are data used in an analysis of the Brown and Frown corpora for my doctoral dissertation titled ``Variations in Written English: Characterizing Authors' Rhetorical Language Choices Across Corpora of Published Texts" (Completed at Carnegie Mellon Univ, 2003). The source of the corpora was the ICAME CD-ROM (get info at <http://www.hit.uib.no/icame/cd>).

The data were generated from the texts using tagging and visualization software, Docuscope.

The first row is the variable names. The genre of each text (assigned by the Brown corpus compilers) is in 'Genre' column and the corpus is listed in the 'corpus' column with 1=Brown and 2=Frown corpus.

The dataset may be freely used and distributed for non-commercial purposes.
Jeff Collins <jeff.collins@acm.org> 11 July 2003

图 2　某“学院数据集”的“数据详细信息”

由图 2 可知数据集的基本信息,建议将该数据集表示为:

[1] JEFF C. Colleges dataset[DS/OL]. Pittsburgh: Carnegie Mellon University,2003[2016-06-06].http://wenku.baidu.com/link? url = v4j0daeP6tBALCWgmsv Eucbn-GrlsFbYBf7f0IXG0bTHAg HShTIm WrA074vpexx3Fpo AWa6Y3Xdsb0yCO XusTg2KB23jN2BTLibxAgjKqXNG.

4　著录细则补充说明

若数据集著录项目中制作地、制作单位、制作年份、更新或修改日期情况不明,则可省略此项。

5　参考文献

[1] 冯海超.大数据时代正式到来[J].互联网周刊,2012,14(24):36

[2] 杨绎.基于文献计量的“大数据”研究[J].图书馆杂志,2012,31(9):29

[3] 吴锋.“大数据时代”科技期刊的出版革命及面临挑战[J].出版发行研究,2013,27(8):66

[4] 侯经川,方静怡.大数据时代的数据引证研究:进展与展望[J].中国图书馆学报,2013,36(1):71

[5] 信息与文献 参考文献著录规则:GB/T 7714—2015[S].北京:中国标准出版社,2015

[6] 梁仲相.地方 MOS 数据集及应用程序包的建立方法介绍[J].广西气象, 1985,6(1):42

[7] 羊毛工业研究协会纺织数据集:一[J].毛纺科技,1978,6(1):51

[8] 康捷.中国考古学中碳十四年代数据集:1965—1981[J].考古,1984,29(3):286

[9] 黄尤来.建立全国查新报告数据库的作用及问题[J].河南科技,2010,35(20):4

[10] 欧阳红红.高校图书馆建设地方特色文化数据库问题探析[J].图书馆,2008,32(6):107

[11] 白娜娜.我国图书情报领域数据的引用情况[J].中国管理信息化,2015,18(16):214

[12] 曹敏.GB/T 7714—2015《信息与文献 参考文献著录规则》标准解析[J].科技与出版,2015,23(9):41

[13] 黄城烟,王春燕.参考文献新国标若干重要概念的理解和著录方法[J].编辑学报,2016,28(3):239

[14] 陈浩元.关于 GB/T 7714—2015 编校失误答同人问[J].编辑学报,2016,28(1):封二

[15] 陈浩元.GB/T 7714—2015 新标准对旧标准的主要修改及实施要点提示[J].编辑学报,2015,27(4):339

(2016-07-29 收稿;2016-11-17 修回)

请期刊编校质量评审专家慎重判错

2017 年 1 月中旬,一位朋友把他们参评政府奖的英文期刊的《报刊编校质量检查结果反馈意见表》发给我,向我咨询他们的疑惑,其中有 2 处“差错”涉及我参与修订的 GB/T 7714—2015《信息与文献 参考文献著录规则》,有的同人对此可能也有疑惑,因此有必要作一澄清,同时吁请评审专家慎重判错。

1)“et al 应改为斜体(文内其他处通改)”,因“字体差错”全刊计 1.5 个“差错”。这是评审者的误判。该刊对“et al.”统一采用正体写法是正确的,也是 GB/T 7714—2015 所倡导的。该标准 10.2.2 就有“对欧美著者只需标注第一个著者的姓,其后附‘et al.’”的写法,正文中 5 处、附录中 4 处“et al.”,全都为正体。ISO 690:2010《信息和文献 参考文献和信息资源引文指南》中,“et al.”也都采用了正体。

2)“et al 结尾处只需一个句号”,因“标点符号错误”计 0.5 个“差错”。这又是一个误判。查该刊第 972 页第 9 行,我认为“et al.[7].”的排法是符合规范的:其中第一个句点是不应省略的缩写点(评审者却将其省略了!),如果没有上角标“[7]”,结尾处确实只需要一个句点,现因插入了上角标“[7]”,理应以句点结束该句的陈述。该刊的文献序号出现在句尾时,依据 GB/T 7714—2015 给出的示例,采用了统一放在标点前面与引用信息紧密相连的科学标注方式,如第 971 页第 4 行的“... investigated[1-3].”。值得指出的是,不少期刊存在上角标序号位置混乱的情况,建议向该刊学习,参照 GB/T 7714—2015 予以规范统一。

(陈浩元)