数据集著录规则探讨与实例分析*

周 俊 李德华
《四川师范大学学报(自然科学版)》编辑部,610066,成都

摘 要 探讨GB/T 7714—2015《信息与文献 参考文献著录规则》新增的文献类型标识数据集(DS)的著录规则,并给出了一些中文著录实例。

关键词 数据集;文献著录规则;实例

Discussions on documenting formats of data sets in references//ZHOU Jun, LI Dehua

Abstract In this paper, we discuss the documenting formats of data sets, which is a new identifiers of document types added in Version 2015 of *Bibliographic Rules for Information and Documentation* (GB/T 7714 - 2015). Moreover, we give some Chinese examples.

Keywords data sets; bibliographic rules of documentation; example

Authors' address Editorial Board of Journal of Sichuan Normal University, 610066, Chengdu, China

DOI:10.16811/j.cnki.1001-4314.2020.01.010

GB/T 7714—2015《信息与文献 参考文献著录规则》[1]比2005版新增了4种文献类型标识:1)档案(A),分类保存以备查考的文件和材料,如人事档案、科技档案、法律法规、政府文件等;2)舆图(CM),世界、国家、区域的地图;3)数据集(DS),一种由数据所组成的集合,又称为资料集、数据集合或资料集合;4)其他(Z),凡是不属于其他15个类型的文献,均可归于"Z"中[2]。其中档案、舆图和其他较好理解,而数据集的概念较为抽象,无论是理解还是应用都有一定困难;但新标准只给出了数据集这一文献类型而未给出定义和著录示例。

张倩等[3]探讨了新增的4种文献类型标识的意义及其应用,认为微博、微信等新媒体作者原创文章的引用应归为数据集,并给出了参考著录格式。陈庆等[4]详细研究了数据集的著录格式,指出表是最简单的一种数据集,并给出了2个英文的引用实例。本文不赞同文献[3]的著录分类,因为微博、微信等文章可归为数据库这一类,而赞同文献[4]的表是最简单的一种数据集的观点;但文献[4]只给出2个英文引用实例,而且都是题名包含数据集这个名称的特殊情况,不够典型。

本文从数据集的名词解释出发,探讨其分类的定义和涵盖的范围,并给出一些常见的中文引用实例。

*国家自然科学基金项目(11701400);四川省教育厅自然科学一般项目(16ZB0063)

1 数据集的定义分析

数据集文献类型的提出是顺应大数据时代的需求,也是紧跟数字化出版的趋势;但是,在已有编辑出版类文献中并没有给出其准确的定义,只能从字面上理解为数据集合。那么,哪种文献属于数据集?参考百度百科所采用的解释,数据集即数据集合,通常以表格形式出现。每一列代表一个特定变量,每一行都对应于某一成员的数据集的问题。每个数值被称为数据资料。对应于行数,该数据集的数据可能包括一个或多个成员[5]。这个解释依然比较抽象,但有一个关键点就是相似于表格形式,具有行和列的结构。在某种程度上,我们可以认为对表的引用是最简单的数据集引用。除此之外,还有哪些文献可归为数据集类?我们认为:具有表结构的数据集合作为参考文献都应归为数据集。因为都是数据集合,数据集与常见的数据库和大数据之间该如何区分?

参考百度百科的解释:数据库是"按照数据结构来组织、存储和管理数据的仓库",是一个长期存储在计算机内的、有组织的、有共享的、统一管理的数据集合[6];大数据是IT行业术语,是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合,是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产[7]。因此,同样作为文献标识的数据库的关键特征是存储在计算机内,即电子的,常见的例如网上的一篇文章或一则新闻都可以归为数据库类型。需要注意的是,数据集并没有要求存储在计算机内,即电子或纸质的都可以。而大数据不是文献标识,是IT术语,其关键特征是"大",具体表现为海量、高增长率和多样化,常规软件难处理的数据集合。具有表结构的大数据也可归为数据集。

2 著录规则与实例

我们给出4个中文数据集文献的著录实例。首先,我们赞同文献[4]所提出的数据集著录格式,即

[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1 数据堂科技股份有限公司.5 000人婴幼儿人脸采集数据[DS/OL].北京:数据堂科技股份有限公司[2019-12-04].https://www.datatang.com/dataset/info/image/1035.

这个数据集包含5 000人婴幼儿人脸采集数据,每个人5~10张彩色生活照,涵盖多种场景、多年龄段、多角度,数据可用于婴幼儿人脸识别等任务。从以上描述可以看出,这些数据按人可分为5 000行,又按场景、年龄段、角度分为若干列,虽然每个单元格中数据不是数字而是图像,但它们具有典型的表结构,因此可归为数据集类。类似地,声音采集数据集合也可归为数据集类。

例2 数据堂科技股份有限公司.794小时四川方言手机采集语音数据_朗读[DS/OL].北京:数据堂科技股份有限公司[2019-12-04].https://www.datatang.com/dataset/info/speech/52.

该数据包括2 507名来自四川盆地发音人,在安静的室内环境下的录音数据。录音内容广泛,覆盖日常短信及多领域客户咨询。句子平均重复次数1.3次,平均句长12.5字。由四川本地人参与质检校对,文本转写更精准。由此可见,数据集的引用对于计算机图形图像处理和语音识别的研究和应用具有广泛而深远的意义。最后,我们给出2个最简单的数据集——表的著录实例。

例3 数据堂科技股份有限公司.160万条大型超市商品数据[DS/OL].北京:数据堂科技股份有限公司[2019-12-04].https://www.datatang.com/dataset/info/text/202.

该文献为某大型超市的全国所有门店的商品数据。数据包括某一年的商品及交易数据(采样见图1)。存储格式为excel表格。

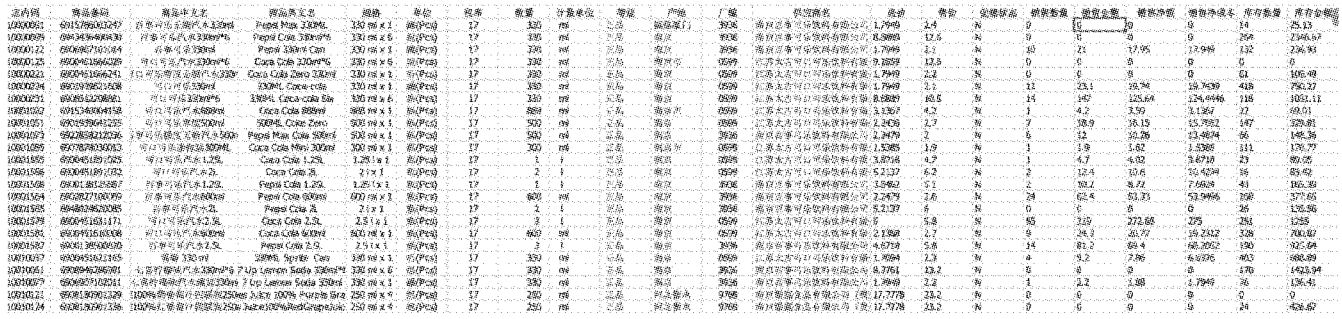

图1 160万条大型超市商品数据采样

例4 中国气象数据网.中国高空规定层累年月值(1981—2010年)[DS/OL].北京:中国气象信息中心,2015[2019-12-04].https://data.cma.cn/data/cdcdetail/dataCode/B.0021.0002.html.

该数据集是根据中国气象局气象资料发展与改革专项工作的高空基础气象资料专项工作研制的全国高空月报信息化数据文件,并基于《高空气候资料整编统计方法(1981—2010)(发布版)》统计加工而得。该数据集为中国87个高空气象站1981—2011年月气候值。数据集包含气压、位势高度、温度、温度露点差、比湿、大气密度、风向和风速等要素的旬、月、年气候值。

3 参考文献

[1] 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

[2] 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):341

[3] 张倩,曹健,姚实林,等.参考文献著录规则新增4种文献类型标识意义及其应用[J].安徽理工大学学报(社会科学版),2018,20(2):105

[4] 陈庆,陆炳新.关于参考文献中数据集著录格式的研究[J].编辑学报,2017,29(1):43

[5] SILBERSCHATZ A, KORTH H F, SUDARSHAN S. 数据库系统概念[M].杨冬青,李红燕,唐世渭,等译.6版.北京:机械工业出版社,2012

[6] 王珊,萨师煊.数据库系统概论[M].5版.北京:高等教育出版社,2014:4

[7] 世纪乐知(北京)网络技术有限公司.大数据概述[EB/OL].(2019-12-03)[2019-12-07].https://blog.csdn.net/marunwei679/article/details/103376703

(2019-11-22收稿;2019-12-08修回)