数据生命周期视角下人文社会科学数据特征研究

Characteristics of Research Data in Humanities and Social Sciences from the Perspective of Data Lifecycle

孟祥保　钱　鹏
(东南大学图书馆,南京,211189)

[摘要]　科研数据是人文社会科学研究的基石与创新的保障,也是重要的科研成果产出。文章利用数据引文索引收录的历史学、教育学、人口统计学、政府与法律、商业与经济领域的科研数据,从数据生产、组织、存储、出版与利用的生命周期环节揭示人文社会科学领域的科研数据结构特征。研究发现:①科研数据生产是一个系统过程;②科研数据组织具有生命周期过程性,但规范性有待提高;③科研数据资源建设具有累积性和长期性;④科研数据分布呈现"集中—离散"特点;⑤89%以上的科研数据零引用,高被引科研数据种数极少,存在权威认同现象。进而,本文从数据共享、数据服务和数据利用三个方面提出实践建议。

[关键词]　数据引用　数据生命周期　数据管理服务　图书馆　数据引文索引

[中图分类号]G250　[文献标识码]A　[文章编号]1003-2797(2017)01-0076-13　DOI:10.13366/j.dik.2017.01.076

[Abstract]　Research data is both the fundamental resource and the outcome of research process in humanities and social sciences. Based on Thomson Reuters' Data Citation Index, the paper selects five disciplines which include History, Education & Educational research, Demography, Government & Law, Business & Economics as the research objective. Then, it describes and analyzes the distribution characteristics of research data from the perspective of data creation, organization, deposit, publication and citation by statistical analysis, citation analysis and data visualization. The findings of this study show that: 1) data produce is a systematic process. 2) Data organization is a lifecycle process, but the normative level remains to be raised. 3) Data collection development is a cumulative and long-term process. 4) data distribution has a tendency of centralization-decentralization. 5) 89% of the records have received no citation, and the number of highly cited data is few. Data citation behavior has the authoritative identity phenomenon. At last, it presents some suggestions to promote the development of data management service from the view of data sharing, library data service and data usage.

[Key words]　Data citation　Data lifecycle　Data management service　Library　Data Citation Index

1　引言

数据资源是科研的基石和创新的源泉。随着科学研究的定量化、规范化与国际化的发展,人文社会科学对数据资源的依赖性与日俱增。进入数字化时代以来,美国综合社会调查(General Social Survey, GSS)、中国综合社会调查(Chinese General Social Survey, CGSS)等大型调查项目的实施,以及政府开放数据的推进,社会化媒体数据的日益增长,科研数据

[基金项目]　本文系教育部人文社会科学研究青年基金项目"基于学术交流的高校图书馆科研服务模式与保障研究"(15YJC870017)的成果之一。

[作者简介]　孟祥保,男,硕士,馆员,研究方向:科研数据管理,Email: meng_xb@163.com;钱鹏,男,博士,研究馆员,研究方向:科研数据管理。

呈现指数级增长。人文社会科学科学数据资源建设与管理引起相关高等学校、学术组织、高校图书馆等机构的重视，纷纷不遗余力地开展人文社会科学数据的管理、组织、服务与共享等工作。

数据是科研过程的基本要素，为实现科研的透明性、可证实性和可重复性，数据成为学术交流的“一等公民”[1]。对于人文社会科学而言，“社会科学定量分析亟待建立一个透明和开源的学术机制，让研究数据和模型公开共享，使研究成果可以得到他人的验证和进一步拓展”[2]，数据的开放共享、规范组织是促进人文社会科学科学性、规范性的重要领域。2014 年，国际科学透明与开放促进委员会(The Transparency and Openness Promotion Committee)提出开放科学指南，其中就包括数据的透明性[3]。如何促进科研数据开放与共享成为亟待解决的问题。

2 相关研究述评

早在 1982 年，美国《图书馆趋势》(Library Trends)在其第 30 卷第 3 期就推出“社会科学数据图书馆”(Data Libraries for the Social Sciences)专题，内容包括数据服务、数据组织与整合、数据获取、数据图书馆发展与案例、数据服务职业培训、数据引用、数据保密性、二手数据利用等。近年来，国内外对人文社会科学领域的数据研究主要集中在三个方面：一是共享视角，着力介绍了国内外数据管理与共享的典范，如美国社会科学数据管理联盟(Data-PASS)[4]、英国社会科学数据存储(UK Data Archive，UKDA)和美国高校校际政治与社会研究联盟(Inter-university Consortium for Political and Social Research，ICPSR)的介绍[5]等，其中数据存储平台[6]、数据整合[7]、元数据[8]等是数据共享的关键因素，但是对数据的生产、存储与出版的状况鲜有论及。二是服务视角，作为数据服务主体的图书馆可以从资源建设、馆员设置和合作平台三个方面发展数据服务[9]，目前研究主要采用案例法对武汉大学图书馆[10]、复旦大学图书馆[11]以及北京大学图书馆[12]的数据服务的实践探索进行了研究，并对社会科学数据馆员进行了阐述[13]。进而言之，所有的数据服务活动都需要建立在数据资源建设基础之上，而数据资源建设又要求馆员对数据属性具有清晰、深刻的认识。三是利用视角，通过对 140 种社会学期刊调查研究发现，数据是否能够实际获取、引用与期刊的数据政策、影响因素密切相关[14]。我国人文社会科学学者利用的数据主要是数值型数据，但是在数据来源选择上存在学科差异性[15]。在图书情报学中“大量的数据没有得到有效应用，已有数据的可获得性较差”[16]。究其原因，社会科学数据家数据共享行为因素是个体动机、制度压力和数据知识库三个层面因素的综合作用[17]。社会科学家越是感知到数据再利用的实践和社会效益将会促使他们再利用数据，同行与学科因素对社会科学数据再利用数据也就越具有正向影响[18]。社会科学家的数据再利用满意度与数据的完整性、可获取性、易操作性和可信度正向相关[19]。从中不难发现，数据利用与数据的学科、类型、出版、开放性等性质密不可分。

综上可知，图书馆学界与业界已认识到科研数据对人文社会科学研究的重要性，并从管理与共享、服务、引用等角度进行了探索。然而，一个不容忽视的事实是，数据是一个生命周期运动过程，目前研究对于人文社会科学数据的基本属性与生命周期特征等还缺乏足够的认识，这在一定程度上制约了人文社会科学的数据管理服务。因此，本文以动态、关联、整体的视角来认识人文社会科学的数据创建、组织、存储、出版与利用等特征，为数据管理服务、共享与利用提供参考。

3 研究设计

3.1 数据来源

(1)数据引文索引概述。2012 年，美国汤森路透(Thomson Reuters)发布了数据引文索引(Data Citation Index，DCI)，旨在促进数据的发现、获取与引用，解决数据存缴与引用意愿缺乏、数据认可与信誉度低等问题[20]。DCI 收割第三方数据知识库的元数据，整合至统一平台 Web of Science 之中并深度标引，目前 DCI 收录约 6442073 种数据。数据的质量、持续性与稳定性、数据环境、语种是 DCI 收录的主要考虑因素[21]，严格的筛选标准保障了 DCI 数据质量。DCI 自发布以来，就成为数据发现、引用与评价的重要工具，国内丁楠等利用 DCI 分析了人口调查数据[22]，国外 Peters 等

分析了 DCI 数据的引用特征与替代计量指标特征[23]。Robinson-García 等从整体角度评价了 DCI 的功能与特点[24]。基于上述 DCI 特点及应用情况，本文数据来源是 DCI 收录的人文社会领域的科研数据。

(2)数据获取过程。本文研究对象界定为历史学、教育学、人口统计学、政府与法律、商业与经济 5 个学科的科研数据，主要基于三个方面的考虑：一是根据研究目标与范围，这 5 个学科是人文社会科学的典型代表；二是经初步的数据检索和分析，发现人文社会科学数据的元数据较为完整丰富、引用次数较高，具有较强的可操作性；三是 WOS 平台最多显示 100000 条检索结果，在数据导出实现上，数据量太大或者太小都不具有可行性。本文的数据检索式是“DT＝(data study OR data set OR repository) AND SU＝学科名称”，分别检索 5 个学科的数据，时间跨度是 1900—2016 年，检索范围为数据引文索引科学库(DCI-S)，数据引文索引社会科学库(DCI-SSH)，检索时间是 2016 年 6 月 17 日，检索结果如表 1 所示。

表 1　数据检索结果

学科中文名	学科英文名	数据量	%DCI	%DCI-SSH
历史学	History	2087	0.03	0.13
教育学	Education & Educational Research	2735	0.04	0.17
人口统计学	Demography	9032	0.14	0.57
政府与法律	Government & Law	9132	0.14	0.57
商业与经济	Business & Economics	16360	0.25	1.03

3.2　研究方法

数据生命周期是指科学数据自身在生命周期各阶段的状态、特征与规律。英国数字管理中心(Digital Curation Centre，DCC)的数据生命周期模型包括概念化、创建、获取与利用、评价与选择、处理、摄入、保存行动、再评价、存储、获取与再利用、转换[25]。UKDA 针对社会科学提出的数据生命周期包括数据创建、数据处理、数据分析、数据保存、数据获取和数据再利用环节[26]。ICPSR 的数据生命周期则包括提出建议与制订数据管理计划、项目启动、数据收集与文件创建、数据分析、数据共享准备、数据存档六个阶段[27]。社会科学数据管理包括数据选择、数据评价、数据保留(retention)与保存等环节，涉及的主体包括数据生产者、数据拥有者、数据存储者[28]。综上，结合人文社会科学研究特点及 DCI 元数据特征，本文按照数据生产、数据组织、数据存储、数据出版与数据利用的生命周期环节分析数据的结构特征，具体如表 2 所示。

表 2　本文分析框架

分析维度	分析单元	定义
数据创建	创建主体	与数据创建直接相关的数据创建者、资助机构
	创建方法	数据采集过程中的所运用的方法
	数据类型	数据物理载体的呈现形态
数据组织	数据类别	数据集、数据研究、数据知识库三种组织形式
	主题词	揭示数据内容的关键词
	DOI 号	科研数据的唯一标识符
数据存储	地理分布	数据的国家或地区分布特征
	存储平台	数据物理存储平台分布特征
数据出版	年份	出版年代
	语种	数据语种
数据引用	引用次数	被 WOS 核心合集文献引用的次数
	高被引数据	被引 100 次以上的科研数据

数据分析方法主要是：①统计分析方法，利用 bibexcel 抽取数据的元数据字段，清洗数据和统计分析数据的属性值。②引文分析方法，分析科学数据主题词的词频特征、数据引用次数等属性。③数据可视化方法，运用云图制作软件 tagxedo 等可视化展示科研数据的特征。

4　研究结果

4.1　数据的创建

(1)创建主体。主要包括数据创建者和资助机构，数据创建者是科研过程中的数据的直接生产者和利益相关者，资助机构是科研项目或数据调查项目的资助机构，也是数据的重要利益相关者。

①数据创建者。包括个体或团体作者，从表 3 中可见，历史学、教育学的科研数据主要是由个体完成，而人口统计学、政府与法律、商业与经济的科研数据主要由团体创建，这与学科特征密切相关，历史学等人文学科一般是由个人完成，而社会科学研究尤其是一些综合性调查需要多个单位相互合作，如美国 GSS 调查等。

从分结果来看,商业与经济数据的团体创建者主要是欧盟统计局(Eurostat)、路易斯·哈利斯调查公司(Louis Harris and Associates, Inc)、美国人口统计局(U.S. Census Bureau)、世界经济合作与发展组织(Organization for Economic Co-operation and Development, OECD)、美国卫生保健质量和研究署(Agency for Healthcare Research and Quality, AHRQ)。政府与法律数据的团体作者主要是领域研究公司(Field Research Corporation)、路易斯·哈利斯调查公司。人口统计学主要是领域研究公司、美国人口统计局、美国商业部(United States Department of Commerce)、澳大利亚统计局(Australian Bureau of Statistics, ABS)。教育学数据的团体作者主要是美国教育部(United States Department of Education, ED)和美国国家教育统计中心(National Center for Education Statistics, NCES)。因此,政府机构、专业调查公司是人文社会科学大型数据项目的主要推动者和实施者。

表 3 数据创建主体

学科	个体创建者		团体创建者	
	数据种数	比例(%)	数据种数	比例(%)
历史学	1413	67.70	852	40.82
教育学	1631	59.63	1157	42.30
人口统计学	3697	40.93	5585	61.84
政府与法律	3873	42.41	5378	58.89
商业与经济	4945	30.23	11482	70.18

②资助机构。科研数据的基金资助数量在一定程度上反映出一个学科对基础研究重视程度。各学科数据基金资助比重如图 1 所示。

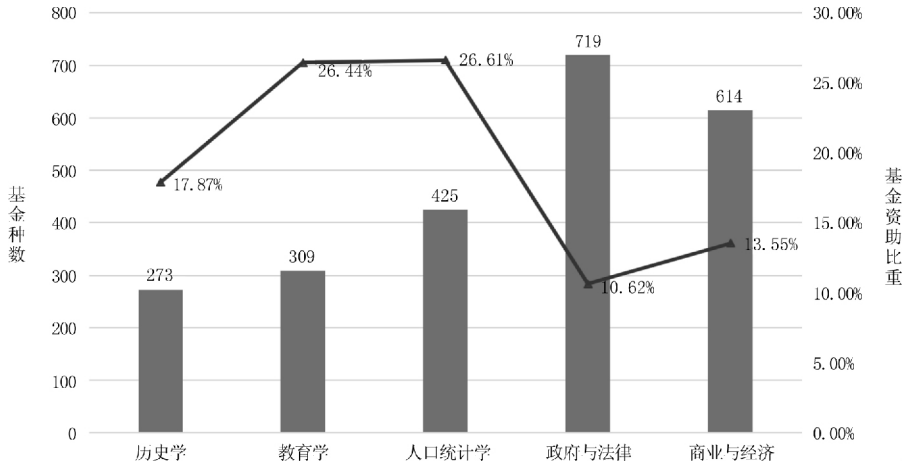


图 1 基金资助

从图 1 发现:①在资助力度上,人口统计学、教育学数据的基金资助比重较高,其中人口统计学有 2403 种数据受到各种类型基金资助,表明人口统计数据采集得到较高重视,而政府与法律、历史学、商业与经济数据的基金资助比例相对较低。②在资助层面上,国家基金起到主导作用,如英国经济与社会研究理事会(Economic and Social Research Council, ESRC)、美国国家自然科学基金(National Science Foundation, NSF)、英国艺术与人文研究委员会(Arts and Humanities Research Council, AHRC)等是各学科数据的主要资助力量。诸如世界银行(World Bank)、国际农业发展基金(International Fund for Agricultural Development, IFAD)等国际基金,推动了国际性的跨地区大型课题的数据采集。③在资助类别上,国家基金是基础性、战略性、全局性数据资源生产的支持者,而专业基金是学科数据的有生资助力量,如人口统计学以及商业与经济领域的英国国家统计局(Office for National Statistics)、历史学领域的利华休姆信托基金(Leverhulme Trust)、教育学领域的麦克阿瑟基金会(McArthur Foundation)、政府与法律领域的公共宗教研

究所(Public Religion Research Institute)以及美国司法部(United States Department of Justice)。

(2)数据创建方法。数据创建方法是数据采集中所利用的方法,从图2可见,具有数据创建方法字段的科研数据比例由高到低依次是教育学、人口统计学、历史学、商业与经济、政府与法律。

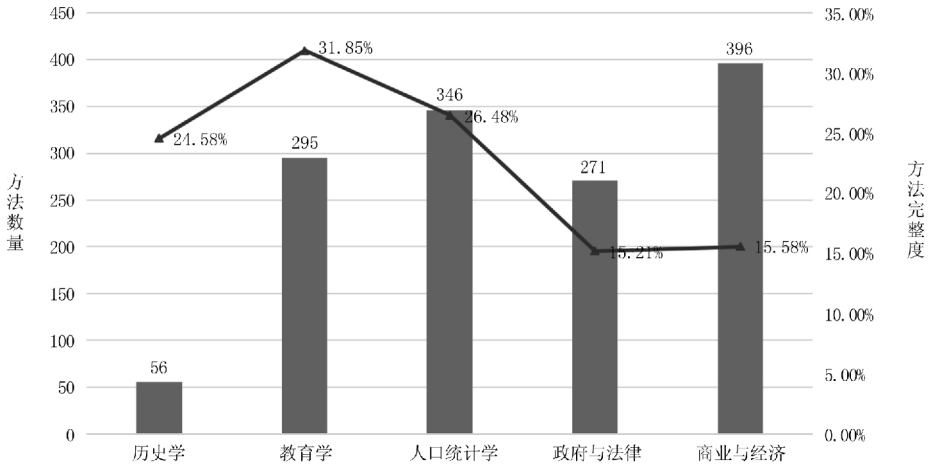


图2　数据创建方法

分析发现,一方面人文社会科学数据创建方法存在着共性,面对面访谈、资料编纂或汇编、档案研究、邮寄调查、电话访谈、个人访谈、问卷调查等是5个学科数据的主要收集方法。另一方面数据创建方法也存在着学科差异性,历史学研究主要依赖文献资料,数据创建方法主要是汇编资料、档案研究。教育学数据创建方法主要包括教育测量、心理测量、观察方法等,体现出教育学与心理学的交叉性。人口统计学数据创建方法还主要包括计算机辅助电话访谈、社会调查法等,体现出人口统计学数据的规模性、社会性。政府与法律的数据创建方法主要是面对面访谈、档案研究、资料编纂或汇编。

(3)数据类型。数据类型是科研数据的物理表现形态,如文本、数据集、图片、音频等。具有数据格式字段的数据种数占所在学科的比例由低到高依次是教育学、政府与法律、人口统计学、商业与经济、历史学。见图3。

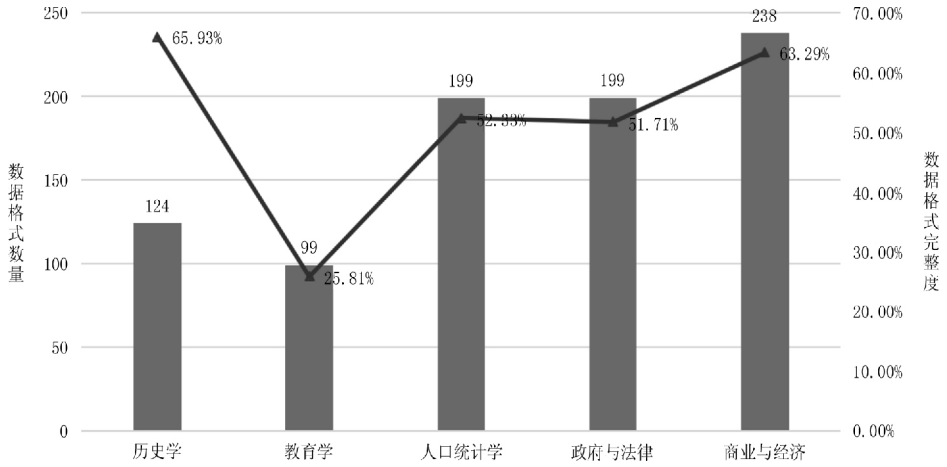


图3　数据类型

分析结果显示，人文社会科学数据类型复杂多样，并非都是数值型数据，还包括文本数据、档案数据、汇编数据、PDF 格式等，还包括微观尺度数据和宏观尺度数据，体现出人文社会科学研究的多样性和不确定性。数据类型也基本反映出学科研究之间的差异性，历史学以文本数据、数值数据、照片为主，教育学、人口统计学和政府与法律则以调查数据、数值数据、个体或微观层面数据为主，商业与经济以列表数据、数值数据、汇编或宏观数据、数值调查数据为主。

4.2 数据的组织

(1)数据类别。数据类别是 DCI 所收录科研数据的组织层次，主要分为数据知识库(Repository)、数据研究(Data Study)和数据集(Data Set)三种类别。数据知识库由数据研究和数据集构成，主要是存储和提供获取原始数据。数据研究是指在数据研究过程中的科研描述或者是实验相关数据，包括长时间的系列研究或者纵贯研究。作为数据集合、数据知识库或者实验研究的有机组成部分，数据集则是单一的或者具有内在一致性的系列数据或者数据文件[29]。5 个学科各自的数据集、数据研究、数据知识库的种数占所在学科比例如图 4 所示。

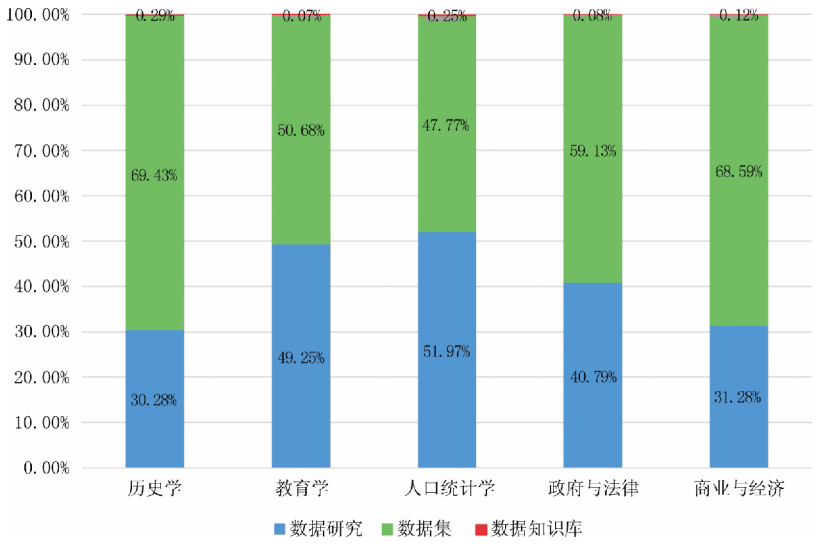


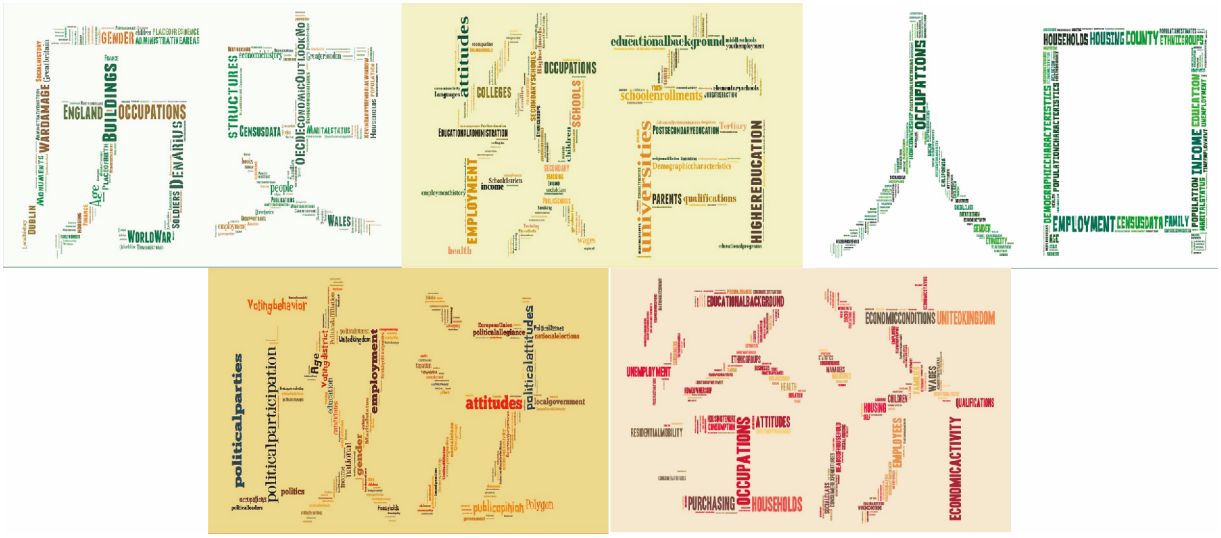
图 4 数据类别

由图 4 可知，历史学、教育学、人口统计学、政府与法律、商业与经济均以数据集和数据研究为主，但在比例上存在一定的差异，历史学的数据集 1449 种，在 5 个学科中所占比例最高，反之，其数据研究占所在学科比例最低。人口统计学的数据研究 4694 种，在 5 个学科中占所在学科比例最高，反之，数据集的比例最低。从中反映出学科的属性差异性，历史学以文献研究、档案研究为基础，而人口统计学的数据多来自于国家层面的统计数据汇编、具有连贯性的调查数据等，具有突出的社会科学实证研究特质。

(2)主题词。具有主题词完整字段的数据种数所占学科数据总数的比例由高到低依次是：历史学 76.90%、教育学 51.48%，人口统计学 38.52%、政府与法律 36.68%、商业与经济 28.61%。主题词具有表征科学数据内容、属性的重要作用。图 5 展示了各学科主题词词频分布状况。

从中不难发现，主题词具有两个方面特点：一是主题词可以表征科研数据的学科属性。历史(History)、英格兰(England)、威尔士(Wales)、第一次世界大战(World War，1914—1918)、婚姻地位(Marital status)、经济史(economic history)等高频词，揭示出历史学的专业术语、地名、历史事件、专业历史等学科形态。教育学中的大学(universities)、高等教育(higher education)、学校招生(school enrollments)、教师

(teachers)等高频词,人口统计学中的统计局数据(Census data)、人口统计学特征(Demographic characteristics)、人口(population)、家庭(family)等,政府与法律中的选举(elections)、政治党派(political parties)、投票行为(Voting behavior)、政治态度(political attitudes)、基层政府(local government)等,商业与经济中的经济活动(economic activity)、购买(purchasing)、职工(employees)、工资(wages)等,也同样揭示出研究对象、研究热点等学科属性。二是主题词具有揭示科研数据内容属性的作用,如性别(Gender)、年龄(Age)、收入(income)、职业(occupations)、住户(Households)、态度(attitude)等高频词透视出数据的测量单元,统计学(Statistics)、工作满意度(job satisfaction)、社会价值观(social values)、政府绩效(Government performance)等高频词又可以看出科研数据的方法论特征。因此,从科研数据主题词基本可以判读出人文社会科学的研究特点。



注:图中字母大小表示词频的高低

图5　数据主题词

(3) DOI号分配。数字对象唯一标识符DOI号是科研数据唯一、持久的标识符号,是追溯、引用、集成、关联科研数据的重要手段。由图6不难发现,人口统计学、商业与经济、教育学、政府与法律具有DOI号的科研数据比例较低,历史学数据相对较高。大量缺失DOI号的科研数据影响了数据组织水平,因为通过DOI号可以实现科研数据的集成与发现、实现科研数据与科学文献的关联。DOI号的缺失也会给数据的规范引用、数据出版造成一定负面影响。

4.3　数据的存储

(1)地理分布。5个学科的科研数据如图7所示,从图中我们可以形象地看出科研数据集中分布在美国(2021种)、英国(9210种)、欧盟(5954种)和澳大利亚(1707种),占到了数据总量的93.74%。究其原因:一是由于欧盟等国人文社会科学发展较为成熟,实证研究占据主流,产生了大量的数据。二是对人文社会科学数据管理与共享意识较高,建立了诸多数据中心或数据知识库。三是DCI主要收割的是英语语种数据,以英文为主的国家占据优势。

(2)存储知识库。科研数据知识库是数据存储、组织与发布的物理平台,5个学科的数据知识库数量分别是:人口统计学52个、商业与经济50个、政府与法律35个、历史学32个、教育学26个。统计分析发现,各学科的科研数据主要集中在少数几个专业性的科研数据知识库,具体分布如表4所示。

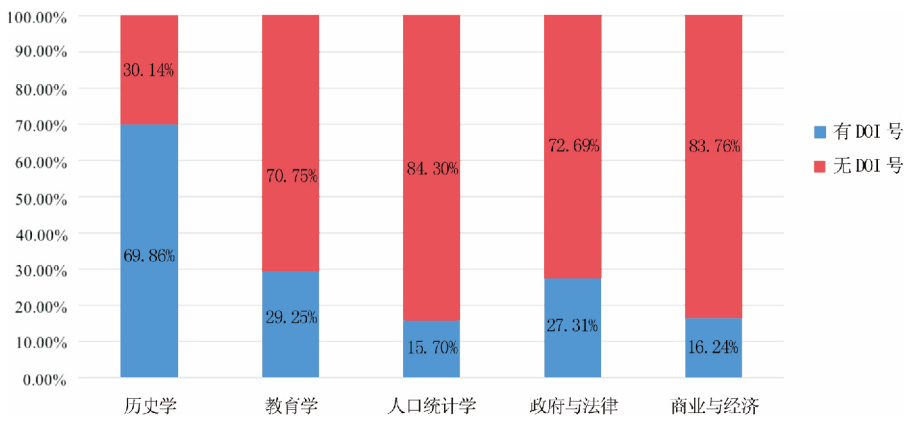


图 6　DOI 号比例

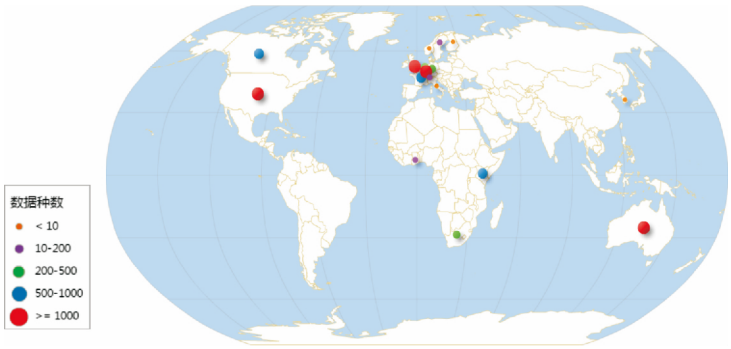


图 7　人文社会科学数据地理分布

表 4　主要存储知识库

知识库名称	历史学	教育学	人口统计学	政府与法律	商业与经济
UCD Digital Library	993				
UK Data Archive	453	425	832	838	1587
GESIS Data Archive for the Social Sciences	360	334	353	1572	886
OECD iLibrary	97				621
U.S. National Archives and Records Administration Dataverse	63		154	170	274
Inter-university Consortium for Political and Social Research	56	397	897	863	868
The Abdul Latif Jameel Poverty Action Lab		554	480		360
Australian Data Archive		286	628	344	410
Odum Institute Data Archive		283		1544	1081
Institute for Quantitative Social Science		108	262		131
UCLA Social Science Data Archive			1312	1495	
International Food Policy Research Institute			1290	1293	2267
Dataweb			1028		72
Washington State University Data Center			493		
The Association of Religion Data Archives			449	154	
Medical Expenditure Panel Survey			284		285
Canadian Opinion Research Archive				391	
U.S. Census Bureau TIGER/Line Shapefiles				240	
Eurostat					5954

从表4中可知，科研数据知识库在分布上具有两个方面的特征：一方面科研数据在科研数据存储呈现“集中—分散”特征，各学科领域的20%科研数据知识库集中了该学科的70%以上的科研数据。例如，加州大学数字图书馆(UCD Digital Library)、UKDA、GESIS社会科学数据存储(Data Archive for the Social Sciences，GESIS-DASS)、OECD iLibrary、美国国家档案与文件管理平台(U.S. National Archives and Records Administration Dataverse)、ICPSR存储了历史学94.01%的数据资源。再如，麻省理工学院的阿卜杜勒·拉蒂夫·贾米尔贫困行动实验室(The Abdul Latif Jameel Poverty Action Lab，J-PAL)、UKDA、ICPSR、GESIS-DASS、澳大利亚数据存储(Australian Data Archive，ADA)集中了教育学72.98%的科学数据。另一方面是科研数据知识库既有综合性又有学科专业性，如UKDA、GESIS、DASS、ICPSR、ADA为多个学科所共有，而加拿大民意研究存储(Canadian Opinion Research Archive)、欧盟统计局(Eurostat)则具有学科的专业性与单一性。造成这种现象的原因：一是上述科研数据知识库建立时间较早、数据管理经验丰富，形成了一套较为成熟的科研数据采集、存储、组织、分析、评价与服务体系，数据资源丰富、特色鲜明、影响力较高，正因为具有这些优势所以才被DCI筛选进来。二是西方社会科学研究的实证主义传统，对科研数据比较重视、定量研究方法较为纯熟，因此数据存储、共享和再利用意识较高。

4.4　数据的出版

(1)出版年份。5个学科的科研数据出版年份如图8所示。

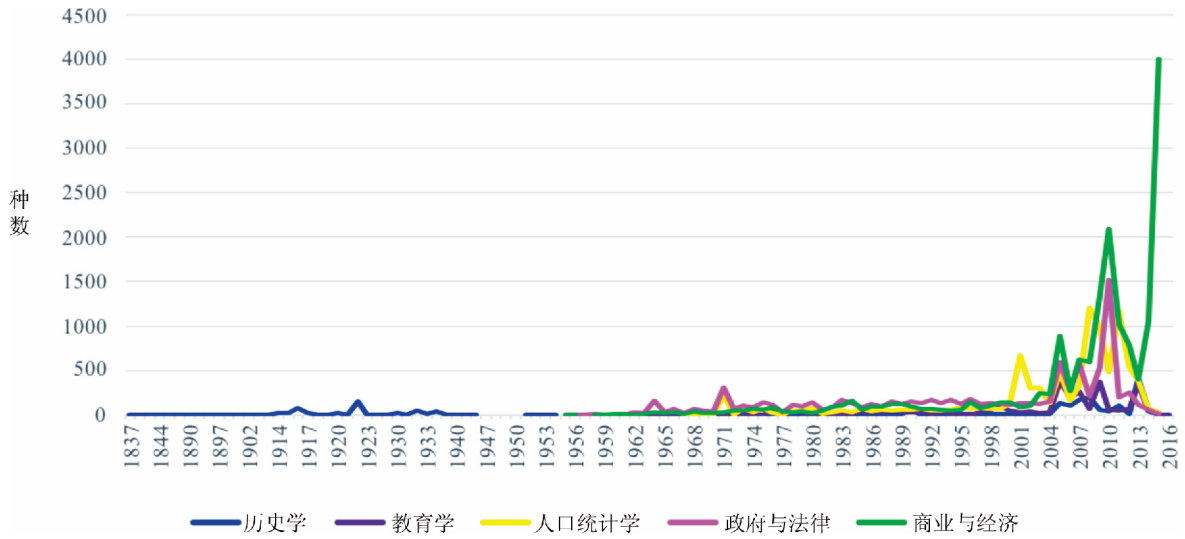


图8　数据出版年

历史学数据的时间跨度为1837—2015年之间，时间跨度较大，而其他4个学科数据时间区间主要是在20世纪90年代以后。信息与通信技术的发展、科研数据知识库的不断建设、大规模社会调查项目的实施，以及学术界科学数据共享意识的不断提升，学术交流日益紧密与学科发展的进度，这些因素促使了人文社会科学的数据快速增长。

(2)语种。5个学科的数据98%以上为英语，其中历史学均为英语。教育学、人口统计学、政府与法律、商业与经济中有极少数德语数据。

4.5　数据的引用

(1)引用次数。引用次数是科学数据利用的重要指标。分别将5个学科所有科学数据按照引用频次从高到低排序，得到表5。

从表5中可以看出，在被引用的数据中绝大多数数据引用次数仅为1—2次，而89%以上的数据没有

被引用，这一比例在历史学、人口统计学、政府与法律、商业与经济中高达92%以上，教育学零引用数据比例也高达89.06%。本文认为其成因是：一是科学研究的内在属性与传播交流机制，任何研究都是以原创性和新颖性为追求目标，命题决定数据，因此科研人员会去采集新的数据，即使引用二手数据也会采用新的分析方法或者多种来源。现代科研成果的交流和评价是以学术论文为主要形式的，科研数据出版还未被广泛接受，因此数据的公开出版、共享与再利用还有待科学成果交流机制的深入发展。二是科研数据出版、数据平台建设、馆藏资源发展、科研数据组织与揭示都远远落后于科技文献管理，对科研数据资源分布与建设的认识不足，从而制约了数据共享、数据引用，如前文所述DOI号的缺失可能会导致科研数据与科学文献无法关联。三是人文社会科学研究人员的数据素养，数据共享意识、数据引用规范、数据出版、二手数据利用等知识与技能有待进一步加强。

被引用100次以上的数据极少，其中人口统计学99种、商业与经济41种、政府与法律25种，教育学6种，历史学仅为2种，具体分布如表6所示。

(2)高被引科研数据。表6列出了被引次数100次及以上的172种科研数据。

表5　被引次数分布

被引次数	历史学		教育学		人口统计学		政府与法律		商业与经济	
	种数	比例%	种数	比例%	种数	比例%	种数	比例%	种数	比例%
0	1971	94.44	2442	89.29	8044	89.06	8423	92.24	15192	92.86
1	84	4.02	162	5.92	260	2.88	232	2.54	631	3.86
2	10	0.48	49	1.79	109	1.21	102	1.12	144	0.88
3-10	16	0.77	51	1.86	364	4.03	205	2.24	171	1.05
11-20	1	0.05	6	0.22	61	0.68	94	1.03	58	0.35
21-50	2	0.1	6	0.22	79	0.87	43	0.47	45	0.28
51-100	1	0.05	13	0.48	16	0.18	8	0.09	78	0.48
＞100	2	0.1	6	0.22	99	1.1	25	0.27	41	0.25

表6　≥100次高被引数据(部分)

题名	被引次数	出版年	数据类别	学科
National Longitudinal Surveys of Labor Market Experience, 1966—1992	1236	1995	数据研究	劳资关系与劳工
Panel Study of Income Dynamics 1968—1999: Annual Core Data	1138	2002	数据研究	人口统计学
American National Election Study 1972	604	1999	数据研究	政治学
American National Election Study, 1980	528	1999	数据研究	政治学
American National Election Study, 1968	502	1999	数据研究	政治学
Household Income and Labour Dynamics in Australia (HILDA) Survey	493	2000	数据知识库	社会学、人口统计学、经济学
American National Election Study 1976	490	2000	数据研究	政治学
American National Election Study, 1964	472	1999	数据研究	政治学
American National Election Study, 1984	391	2000	数据研究	政治学
American National Election Study 1960	390	1999	数据研究	政治学
American National Election Study 1956	348	1999	数据研究	政治学
Study of Health in Pomerania	322	1997	数据知识库	健康护理与服务、人口统计学、社会学
American National Election Study, 1988: Pre- and Post-Election Survey	313	2000	数据研究	政治学
American National Election Study, 1992: Pre- and Post-Election Survey [Enhanced with 1990 and 1991 Data]	305	1999	数据研究	政治学
American National Election Study 1974	303	2000	数据研究	政治学

注：因篇幅所限，表中列举被引次数大于300次的数据。

高被引数据在内容和形式上具有高度集中性，具体表现是：①在学科类别上，主要是人口统计学 101 种、社会学 83 种、经济学 40 种、商业 37 种、政治学 25 种、家庭研究 16 种、健康政策与服务 11 种、教育与教育研究 8 种、健康护理与服务 6 种、社会工作 3 种、历史学 2 种、劳资关系与劳工 1 种。②主要来源是 UKDA(93 种)、ICPSR(60 种)。③数据类别是数据研究(167 种)和数据知识库(5 种)，换言之，高被引科学数据主要是汇编数据或派生数据。④内容主题主要是大型调查项目或者系列研究的数据，如美国家庭综合调查(General Household Survey)、家庭支出调查(Family Expenditure Survey)、美国国家大选调查研究(American National Election Study)、国家儿童发展研究(National Child Development Study)。⑤在时间上，高被引数据出版时间主要是 20 世纪 90 年代以后。究其原因：一是科研数据的真实性和完整性，从高被引数据分布特征来看，往往是来自权威机构的、可信科研数据知识库的系列数据更容易被科学文献引用，这些数据具有较高的可信度，在内容主题上具有全局性和普遍意义，数据的格式、时间、地址、采集方法与范围等属性详尽规范，具有较高的可操作性和重复性。二是数据的可获取性，毋庸置疑，开放的、数字化的科研数据更易于扩散与再利用，上述高被引科研数据均是如此。三是学科研究范式决定了对科研数据的依赖程度，在被引 100 次以上的科研数据中，170 种来自社会科学领域，而历史学仅为 2 种，经济学、社会学、教育学等学科具有较强的实证倾向，主要是将二手数据与数学模型有机结合起来解决问题，而历史学研究一般是以文献资料为基础、以逻辑思辨为主要研究方法，对二手定量数据依赖程度较低。

5　研究结论及启示

5.1　主要结论

通过对历史学、教育学、人口统计学、政府与法律、商业与经济 5 个学科科研数据的分析，本文初步揭示了人文社会科学数据在数据生命周期各个环节的基本特征。

(1)科研数据创建与生产是一个系统过程。如前所述，人文社会科学的数据生产是一个由创建主体、数据创建方法、数据对象组成的系统，数据创建主体包括数据直接创建者和基金资助机构。数据生产过程中所利用的方法具有人文社会科学研究的鲜明特征。所产生的数据类型复杂多样，以教育学、人口统计学、政府与法律、商业与经济为代表的社会科学数据类型主要是定量调查数据，以历史学为代表的人文学科数据主要是文本数据。同时，文本、音频、图片等非结构化数据也占了相当比例，数据格式复杂多样。进而言之，人文社会科学的数据类型更为丰富，数值型数据、文本资料、口述资料、图片、照片、录像都可以视为科研数据，数据的生产过程也是科研人员利用科研方法作用于研究对象的过程，具有科研生命周期和数据生命周期的二重性，从这个意义上来说，文献即数据，数据即科研。

(2)科研数据组织具有生命周期过程性。DCI 的元数据集较好地诠释出科研数据的数据生命周期过程，从数据生产到数据组织、从数据存储到数据出版、从数据共享到数据引用，在这一过程中，既体现出数据生产者、资助者、存储者、出版者、引用者等利益相关主体的相互协同性，又具有感知科研数据及其运动的认识论意义，科研数据组织的层次和水平是对科研数据特征认识高度的集中体现，例如，对科研数据与科学文献关联特征的认识，科研数据在生命周期运动过程中，通过内容引用实现科研数据的扩散与增值，在资源组织层面则是通过 DOI 号揭示二者的关联关系。再者，数据类型、地理空间信息、时间、方法等叙词又能较好地展现科研数据的科研生命周期过程。此外，DCI 中的数据质量与组织规范程度也有待进一步提高，主题词、基金资助、DOI 号等值存在大量缺失，这就给数据检索、发现和利用造成巨大的障碍。

(3)科研数据资源建设的累积性和长期性。与科学文献一样，科研数据资源建设也具有累积性。数据是伴随着科研过程而产生的，因此需要较长时间，尤其是一些大型调查项目和纵贯研究，往往需要历时多年乃至几十年。从存储与组织上来看，数量丰富、质量较高、知名度较高的数据知识库一般都建立时间较

早,因此,科研数据的规范组织、科学管理和共享服务都需要长时间的努力积累与科学管理。

(4)科研数据分布呈现"集中—离散"趋势。在数据存储上具有高度的集中性、专业性,5 个学科各自的 20%科研数据知识库集中了该学科的 70%以上的数据资源,5 个学科的科研数据合并去重后计算发现,以 Eurostat、国际食品政策研究所(International Food Policy Research Institute)、UKDA、GESIS-DASS、ICPSR 等为代表的 11%科研数据知识库集中了 5 个学科的 82%科研数据,相对而言,其余的 18%科研数据分布更为零散,广泛分布于知名度较低、数据种数较少的其他 80 个科研数据知识库之中。语种上,98%以上数据为英语。出版时间上,5 个学科的 50%以上数据是在 2000 年以后出版,数字化、开放化促进了科研数据的快速流动和价值增值、促使科研数据来源分布也更为广泛。

(5)科研数据引用次数较低,存在大量零被引数据,高被引数据具有集中性,存在权威认同现象。如前文所述,5 个学科约 90%科研数据存在零被引现象,其余的 10%科研数据被引次数也多是 1—2 次引用,高被引科研数据凤毛麟角,主要是来自权威机构的大型调查项目的数据,在学科、类型、时间上高度集中。因此,数据扩散广度和深度极为有限,严重制约了科研数据资源的价值发掘,这既需要我们建设国家级人文社会科学数据基础设施,促进科研数据开放共享,努力实现科研数据与科学文献的关联实现,也需要不断提高科研人员的数据素养,规范数据引用,更需要改进现有科研评价机制,将科研数据纳入科研成果的有机组成部分。

5.2 实践启示

(1)对于数据共享而言,数据共享是数据创建者、数据中心、基金资助机构、出版机构等利益相关主体的协同过程,需要各方的共同努力。各方利益相关主体需要明确自身在数据生命周期环节中的权责和利益。一是需要提供数据组织的规范性,建立科研数据质量评价体系,提高数据质量。二是对人文社会科学数据共享的必要性、可行性和相应的措施逐步凝练共识,从而不断提高数据共享水平。

(2)对于图书馆数据服务而言,数据资源是服务的基础,需要重视人文社会科学的数据资源建设,将数据资源整合至馆藏资源之中,"在资源建设规划时,图书馆应将社科数据作为整体考虑"[30]。人文社会科学数据"集中—分散"特征、数据的引用特征可为图书馆数据资源建设提供参考借鉴。数据的学科、语种、出版时间等属性也是重要的考虑因素。在数据组织方面,图书馆可充分发挥专业特长,规范数据组织,提高数据的关联度。在数据服务方面,可提高科研人员的数据共享、数据版权保护意识,为科研人员推荐合适的数据知识库。

(3)对于人文社会科学研究人员而言,应提高数据共享意识,数据的开放、共享与再利用有助于科研成果的验证,也有利于提高科研成果的影响力,促进学术交流。此外,科研人员也需要提升自身数据素养,如了解期刊的数据出版政策、选择合适的数据知识库存储科研数据、如何保护科研数据版权、如何利用现有数据避免重复劳动。

6 结语

一流的人文社会科学研究需要一流的科研数据支撑,对人文社会科学数据特征的整体把握和深化认识将有助于数据管理、数据服务、数据共享、数据再利用等问题,有助于提高数据的可获取性和可信度、促进数据的流动与增值。在后续研究中,可扩大学科范围,提高研究成果的适用范围,同时进一步探索数据特征的形成机制。此外,科研数据是一个生命周期运动过程,数据在流动过程中存在扩散、增值、变异、聚变等现象,"后置"的数据世系(data provenance)问题和"前置"的数据扩散与关联机制也是需要解决的命题。

致谢:美国汤森路透公司为本文研究提供了数据支持,在此致以诚挚谢意。

参考文献

1 Bolikowski L, Houssos N, Manghi P, et al. Data as "first-class citizens"[J/OL]. D-Lib Magazine,2015,21(1/2)[2016-07-12]. http://www.dlib.org/dlib/january15/01guest_editorial.html
2 陈云松,吴晓刚. 走向开源的社会学 定量分析中的复制性研究[J]. 社会,2012(3):1-23
3 Nosek B A, Alter G, Banks G C, et al. Promoting an open research culture[J]. Science,2015,348(6242):1422-1425
4 彭建波. 美国社会科学数据管理联盟(Data-PASS)的发展与借鉴[J]. 图书情报工作,2014(10):117-121
5 孟祥保,钱鹏. 高校社会科学数据管理的国际经验及其借鉴——以 UKDA 和 ICPSR 为例[J]. 情报资料工作,2013(2):77-80
6 覃丹. 英美社会科学数据管理与共享服务平台调查分析[J]. 图书情报工作,2014(16):67-75,142
7 Zilinski L D, Barton A, Zhang Tao, et al. Research data integration in the Purdue Libraries[J]. Bulletin of the Association for Information Science and Technology,2016,42(2):33-37
8 Mayernik M S. Research data and metadata curation as institutional issues[J]. Journal of the Association for Information Science and Technology,2016,67(4):973-993
9,30 刘澈,李桂华. 中外高校图书馆社科数据服务比较[J]. 图书馆论坛,2016(6):142-148
10 项英,赖剑菲,丁宁. 高校图书馆科学数据管理服务实践探索——以武汉大学社会科学数据管理为例[J]. 情报理论与实践,2013(12):89-93
11 张计龙,殷沈琴,张用,等. 社会科学数据的共享与服务——以复旦大学社会科学数据共享平台为例[J]. 大学图书馆学报,2015(1):74-79
12 朱玲,聂华,崔海媛,等. 北京大学开放研究数据平台建设:探索与实践[J]. 图书情报工作,2016(4):44-51
13 Xia Jingfeng, Wang Minglu. Competencies and responsibilities of social science data librarians: an analysis of job descriptions[J]. College & Research Libraries,2014,75(3):362-388
14 Zenk-Moeltgen W, Lepthien G. Data sharing in sociology journals[J]. Online Information Review,2014,38(6):709-722
15 沈婷婷. 人文社科领域科学数据使用特征分析——基于《中国社会科学》样本论文的实证研究[J]. 大学图书馆学报,2015(3):101-107
16 丁楠,丁莹,杨柳,凌晨,等. 我国图书情报领域数据引用行为分析[J]. 中国图书馆学报,2014(6):105-114
17 Kim Y, Adler M. Social scientists' data sharing behaviors: investigating the roles of individual motivations, institutional pressures, and data repositories[J]. International Journal of Information Management,2015,35(4):408-418
18 Curty R G. Beyond "data thrifting": an investigation of factors influencing research data reuse in the social sciences[D]. Syracuse: Syracuse University, 2015
19 Faniel I M, Kriesberg A, Yakel E. Social scientists' satisfaction with data reuse[J]. Journal of the Association for Information Science and Technology, 2016, 67(6):1404-1416
20,29 Force M M, Robinson N J. Encouraging data citation and discovery with the data citation index[J]. Journal of Computer-aided Molecular Design,2014,28(10):1043-1048
21 丁楠,潘有能. 数据引用索引工作机理与应用现状综析[J]. 情报理论与实践,2014(6):59-62
22 丁楠,黎娇,李文雨泽,等. 基于引用的科学数据评价研究[J]. 图书与情报,2014(5):95-99
23 Peters I, Kraker P, Lex E, et al. Research data explored: an extended analysis of citations and altmetrics[J]. Scientometrics,2016,107(2):723-744
24 Robinson-García N, Jiménez-Contreras E, Torres-Salinas D. Analyzing data citation practices using the data citation index[J]. Journal of the Association for Information Science and Technology,2015
25 DCC curation lifecycle model[EB/OL]. [2016-07-12]. http://www.dcc.ac.uk/resources/curation-lifecycle-model
26 Research data lifecycle[EB/OL].[2016-07-12]. http://www.data-archive.ac.uk/create-manage/life-cycle
27 ICPSR. Guide to social science data preparation and archiving: introduction[EB/OL].[2016-07-12]. http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/
28 The selection, appraisal, and retention of digital social science data[J/OL]. Data Science Journal,2004, 30(3)[2016-07-16]. http://doi.org/10.2481/dsj.3.209

(收稿日期:2016-10-16)