

国家社会科学基金青年项目研究成果

大数据环境下 政府数据开放研究

迪莉娅 著

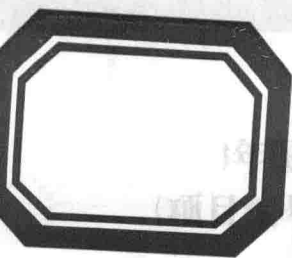
DASHUJU HUANJINGXIA

ZHENGFU SHUJU KAIFANG YANJIU



知识产权出版社

全国百佳图书出版单位



国家社会科学基金青年项目研究成果
号: 13CTQ037)

大数据环境下 政府数据开放研究

迪莉娅 著



DASHUJU HUANJINGXIA

ZHENGFU SHUJU KAIFANG YANJIU



知识产权出版社

全国百佳图书出版单位

图书在版编目(CIP)数据

大数据环境下政府数据开放研究/迪莉娅著. —北京:知识产权出版社, 2014. 7

ISBN 978-7-5130-2806-6

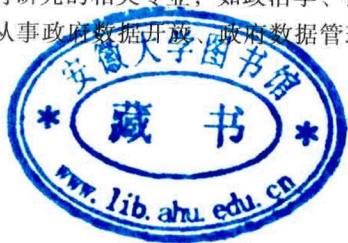
I. ①大… II. ①迪… III. ①电子政务—研究—中国 IV. ①D630.1

中国版本图书馆CIP数据核字(2014)第142972号

内容提要

本书探讨了大数据环境下政府数据开放的渊源,分析了政府数据开放产生的重要理论基础和价值,解读了国外政府数据开放的特点,剖析了我国政府数据开放的制约因素,从政策、法律、数据开放许可和数据管理方面建构了我国政府数据开放的路径和策略。

本书是政府研究的相关专业,如政治学、公共管理学、公共经济学、信息资源管理学等专业的学者从事政府数据开放、政府数据管理、内容产业、大数据产业研究的重要参考书目。



责任编辑:吴晓涛

责任出版:谷洋

大数据环境下政府数据开放研究

迪莉娅 著

出版发行:知识产权出版社有限责任公司

电 话:010-82004826

社 址:北京市海淀区马甸南村1号

责编电话:010-82000860转8533

发行电话:010-82000860转8101/8029

印 刷:北京中献拓方科技发展有限公司

开 本:720mm×1000mm 1/16

版 次:2014年8月第1版

字 数:134千字

网 址:<http://www.ipph.cn>

<http://www.laichushu.com>

邮 编:100088

责编邮箱:sherrywt@126.com

发行传真:010-82000893/82003279

经 销:各大网上书店、新华书店及相关专业书店

印 张:8.25

印 次:2014年8月第1次印刷

定 价:28.00元

ISBN 978-7-5130-2806-6

出版版权专有 侵权必究

如有印装质量问题,本社负责调换。

前 言

大数据是指海量且类型复杂的数据，具有种类多、流量大、容量大、价值高的特点。大数据被称为信息化时代的“石油”。未来国家的核心竞争力将很大程度上依赖将数据转化为信息和知识的速度与能力，而这种转化的速度和能力，实际上取决于数据方面的技术和管理应用到政府、企业决策方面的文化的形成。

大数据时代的来临，对政府机构的管理和发展提出了新的挑战。政府数据是指一切产生于政府内部或虽然产生于政府外部，但对政府活动、公共事务和普通民众有影响、有意义的数据资源的统称。政府拥有整个社会数据资源的80%以上，是数据最大的生产者和拥有者。在2009年美国奥巴马政府颁布《开放政府指令》之后，世界范围内掀起了政府数据开放的浪潮，政府数据的开放加深和拓宽了政府信息公开的力度，也预示着政府开放和透明迎来了新的里程碑。

政府数据开放运动的前身是信息获取权运动。在世界范围内，信息获取权运动的发展为政府数据开放运动奠定了重要基础。信息获取权运动和政府数据开放运动在应用的方法和策略上虽有所不同，但是其目的都是为了促进政府更加开放和透明。开放数据运动的产生与开放政府理论、数据资产理论、数据权理论的兴起和发展也息息相关，为其奠定了重要的理论基础。

在大数据环境下的政府数据开放对于提高政府效能，促进政府开放、透明，大数据产业的发展、政府基于数据决策的文化的形成，以及让更多的公众参与数据的开发和利用将产生积极的功用。为了应对大数据的挑战，很多国家和组织在国家层面上制定政府数据开放的政策和路径、颁布新的法律或者修订原有的法律以适应大数据环境下的政府数据的开放，同时采用政府数据开放许可和应用先进的技术和方法管理数据等措施，不断促进政府数据的使用、重复性使用和创新性使用。

政府数据的开放成为越来越多民众的诉求和政治开明以及经济、文化发

展的重要标志。本书追溯了政府数据开放的渊源和产生的理论基础，阐释了其重要的政治、经济、文化和社会方面的价值，解析了国外政府数据开放的发展规律，深刻探讨了我国政府数据开放的制约因素，有针对性地提出了我国政府数据开放在政策、法律、许可以及数据管理方面的策略，希望对我国政府的数据开放提供具有借鉴价值的意见和建议。

在本书的写作过程中，得到了中国人民大学信息资源管理学院赵国俊教授的悉心指导，在此一并表示感谢。本书必然存在许多不足之处，敬请读者批评指正。

迪莉娅

2014年6月

目 录

第1章 导 言	1
1.1 研究的背景	1
1.2 研究的目的	2
1.3 国内外研究简述	2
1.3.1 国内研究现状	2
1.3.2 国外研究现状	3
1.4 相关概念界定	5
1.4.1 数据	5
1.4.2 大数据	6
1.4.3 政府数据开放	9
参考文献	11
第2章 大数据环境下政府数据开放的渊源和理论基础	15
2.1 大数据环境下政府数据开放的渊源	15
2.1.1 信息获取权运动的发展	15
2.1.2 信息获取权的基本要素	16
2.1.3 政府数据开放运动的兴起	17
2.1.4 政府数据开放运动与信息获取权运动的区别	19
2.2 大数据环境下政府数据开放的理论基础	20
2.2.1 开放政府理论	20
2.2.2 数据资产理论	21
2.2.3 数据权理论	21
参考文献	22

第3章 大数据环境下政府数据开放的价值	24
3.1 政治价值	24
3.2 经济价值	27
3.3 文化价值	29
3.4 社会价值	29
参考文献	30
第4章 大数据环境下国外政府数据的开放	32
4.1 国外政府数据开放政策	32
4.1.1 美国的政府数据开放政策	32
4.1.2 英国的政府数据开放政策	37
4.1.3 加拿大的政府数据开放政策	42
4.1.4 澳大利亚的政府数据开放政策	47
4.1.5 欧盟的政府数据开放政策	50
4.2 国外政府数据开放法律	51
4.2.1 美国政府数据开放相关的法律	51
4.2.2 英国政府数据开放相关的法律	55
4.2.3 加拿大政府数据开放相关的法律	57
4.2.4 澳大利亚政府数据开放相关的法律	61
4.2.5 欧盟政府数据开放相关的法律	63
4.3 国外政府数据开放许可	66
4.3.1 开放数据许可的类型	66
4.3.2 国外政府数据开放许可的应用	68
4.4 国外政府数据开放管理模式	73
4.4.1 数据门户网站的建设	73
4.4.2 数据的管理模式	74
4.4.3 用户的参与途径	77
参考文献	78

第5章 大数据环境下我国政府数据开放的制约因素分析	83
5.1 政策层面	83
5.2 法律层面	86
5.2.1 信息公开相关的法律	86
5.2.2 隐私与安全相关的法律和政策	87
5.3 许可层面	88
5.4 政府网站信息管理层面	90
参考文献	91
第6章 大数据环境下我国政府数据开放的策略	93
6.1 构筑我国政府数据开放的政策体系	93
6.1.1 我国政府数据开放的目标	93
6.1.2 我国政府数据开放政策的制定	94
6.1.3 我国政府数据开放政策的实施	98
6.1.4 我国政府数据开放政策的评估	103
6.2 制定和修改我国政府数据开放的相关法律	108
6.2.1 我国政府信息公开相关法律法规的制定和修改	108
6.2.2 我国网络安全相关法律法规的制定和修改	108
6.3 制定我国政府数据开放许可	109
6.3.1 知识共享协议的要素	110
6.3.2 知识共享的类型	110
6.3.3 我国政府数据开放许可协议的建立	112
6.4 建立我国政府数据开放的门户网站	113
参考文献	117
第7章 结 语	120
参考文献	121

第1章 导言

1.1 研究的背景

据中国互联网数据中心的报告显示：2010年，全球数据量已达1.2ZB，到2020年将暴增到35ZB；2011年，全球被创建和复制的数据总量为1.8ZB；2013年，生成这样规模的信息量只需10min。而在如此庞大的数据中，只有10%的数据是存储在数据库中的结构化数据，其余的则是由邮件、视频、微博、页面点击等产生的大量的半结构化数据和非结构化数据。2011年5月全球知名咨询公司麦肯锡发布的《大数据：创新、竞争和生产力的下一个前沿领域》，使人们深刻意识到大数据时代已经来临。该报告指出：“大数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。”

面对大数据的挑战，很多国家和政府将政府数据的管理提升到国家战略层面，从国家政策或者制度层面规划大数据的发展、挖掘大数据的潜力。例如，2012年3月，美国奥巴马政府推出“大数据研究与开发计划”，提出“通过收集、处理庞大而复杂的数据信息，从中获得知识和洞见，提升能力，加快科学、工程领域的创新步伐，强化美国国土安全，转变教育和学习模式”。而英国、法国、日本等国政府也已经关注大数据和大数据技术，积极开展和推动政府、企业大数据的管理及大数据技术的应用。

政府是数据最大的生产者和拥有者，政府数据约占整个社会数据的80%以上，政府数据的开放对于政府利用大数据技术和方法促进政府信息和数据资源的利用、政府的管理向精细化方向转变以及提高政府科学决策的水平具有重要的作用。

1.2 研究的目的

(1) 在大数据环境下，通过对政府数据开放产生的实际效果进行分析，探索国内外政府数据开放的内在规律，尤其是政府数据开放对政治、经济、文化、社会等方面产生的实际功用；

(2) 通过对国外政府数据开放所产生的实际价值，从分析国外政府推行数据开放所采取的方法和策略中获得启示，为我国政府的数据开放提供借鉴和参考；

(3) 通过对目前我国政府信息公开的现状和制约因素的分析，探索在大数据环境下我国政府推行数据开放行之有效的策略。

1.3 国内外研究简述

从世界范围来看，2009年由美国引领的政府数据开放运动在世界范围内掀起了开放数据的浪潮，国外的政府数据开放在研究和实践领域都已经全面展开，而我国的政府数据开放还处于学者的研究阶段，在实践方面还没有完全开始。

1.3.1 国内研究现状

随着大数据时代的来临，政府信息公开也进入了新的阶段，即数据开放的阶段。国内学者关于数据开放的研究基本上还处于起步阶段，在中国期刊网上输入检索词，政府数据开放的论文并不多，主要集中在以下几个方面：

(1) 介绍国外政府数据开放，如美国、英国、加拿大等国家的经验与启示方面的论文较多，对我国政府数据开放的研究较少（见表1-1）。

表 1-1 政府数据开放研究代表性论文一览表

作者	论文名称	论文摘要	发表期刊与时间
谭健	开放数据及其应用现状	论述了开放数据在科学学术交流、Web2.0时代信息组织中的愿景，重点介绍了开放数据在当今美国、加拿大、英国等国家的应用现状，以及非营利性组织、企业对开放数据的推动情况	图书与情报 2011（4）

续表

作者	论文名称	论文摘要	发表期刊与时间
侯人华、徐少同	美国政府数据开放的管理与利用分析——以www.data.gov为例	以www.data.gov为例,从数据管理的视角,分别对其数据的特点、数据采集方式、数据管理策略、数据利用方式以及优势五个方面进行分析,揭示美国数据网站在信息公开过程中对数据信息资源的管理与利用方式、方法和水平,期望对我国政府信息公开网站建设提供借鉴	图书情报工作 2011(2)
陆健英、郑磊、Sharon S. Dawes	美国政府数据开放、历史进展与启示	论述了美国政府数据开放的发展历程,特别是政府通过网站开放数据的实践;重点分析了美国政府数据开放的最新进展和www.data.gov网站的特点与不足,进而探析了美国政府数据开放对中国的启示意义	电子政务 2013(6)

(2) 政府数据开放环境下数据管理方面的研究。①关联数据在政府数据开放中的应用。吴玥、李占羽论述了关联数据适应于国内发布开放的政府数据到 Web 的工作流程,提高政府数据的透明度,增强公民对政府的监督和问责。^[1]②开放数据许可的探讨。李佳佳在《国外开放数据许可及相关机制研究》^[2]一文中,论述了国外开放数据许可机制的重要意义及其相关机制和实践状况,对我国数据开放许可机制的研究和发展具有一定的参考作用。③关联数据协议。张春景、刘炜、夏翠娟^[3]指出:关联数据发布、消费和再创造的过程中,必然涉及参与者的利益问题,需明确不同类型和归属的数据(或数据集)的所有权、发布权、使用权、收益权等。关联数据的发展迫切需要相关的许可协议,并在文中阐述了开放数据共用、开放政府许可协议、知识共享协议等关联数据相关的协议。

1.3.2 国外研究现状

国外数据开放方面的研究从理论和实践方面都比我国先行一步,研究所涉足的领域包括以下几个方面:①开放数据的价值。Thorhildur Jetzek 从效率、透明、参与、创新的角度阐述了政府数据开放的价值。在研究中她认为开放数据的价值产生于更有效的利用数据;通过私营部门利用政府数据将其转化为产品和服务;增加政府工作的透明度,鼓励正当程序的公平性和合法

性；增加公众的参与度和合作，提高维护社会公平正义和解决社会复杂问题的能力。^[4] ②开放数据法律方面的研究。Katleen Janssen 回顾了开放政府的环境下欧洲再利用政府信息指令的角色，总结了欧盟成员国开放数据的成效，提出了在数据开放的环境中现存的信息自由权需要得到保证等建议。^[5] ③加强电子政务社会性网络的建设，促进数据的再利用。Enrique Bonsón、Lourdes Torres、Sonia Royo 等对欧盟范围内地方电子政务应用社会性网络、促进政府透明度和电子参与方面进行了调查。联合国经济与社会部专门出版了在开放政府环境中，公民参与和利用开放数据的指导书，指出政府开放数据的路径以及加强公民利用的方法和措施。^[6] ④开放数据的再利用研究。例如 Fabricio F. Costa 论述了在大数据环境下，生物医药数据的应用对于医药行业的促进作用，指出生命科学中的数字科学提高了医疗的质量并降低了成本，更重要的是可以让研究者充分地利用资源。而以数据为中心的方法使研究者可以充分利用信息，走向个性化医药的发明。^[7] ⑤数据开放在决策方面的应用。Philip M. Napoli、Joe Karaganis 认为公共政策的决策必须要有公开的可利用的数据成为根本的原则。二位学者以美国通信政策为例，阐述了美国通信政策决策的现状，指出数据的公开利用对决策的重要性，并提出了以下对策：为所有的政策研究提供受托数据；提高利用的模式和程序；为商业敏感数据提供现实的利用标准；除审议过程中的数据受信息自由法豁免等。^[8] Haluk Demirkan、Dursun Delen^[9] 在研究中指出在云环境中应用面向服务的决策支持系统使组织机构的决策更加灵活并且提高了决策的质量，分析了面向服务的决策支持系统所应具备的条件，并提出了数据即服务、信息即服务和分析即服务的模式。⑥数据管理方面的研究。Clément Turbelina、Pierre-Yves Boëllea 认为公共健康监控机构产生的数据与日俱增，开放这些数据是数据共享的核心，但是在互联网上目前还没有出台相关的标准采纳这些信息。以法国的健康监控系统（FSN）为例，介绍、比较了 SDMX-HD 和 OData 两种开放数据标准优缺点，指出在公共健康预警系统中这两个标准的整合是共享系统建设的重要性。^[10] Eleni Galiotoua、Pavlina Fragkou 则提出了将关联数据应用到希腊开放数据的发布和共享中，指出在网络上提高开放数据的发布、共享以及关联的功能，为数据的整合和互操作提供了新的路径。Li Ding、Timothy Lebo、John S. Erickson 等论述了政府开放数据语义门户网站（TWC LOGD）创建的基础设施和工作流，在开放政府数据领域中已取得成果，被美国联邦政府第一数据网 www.data.gov 所采用。^[11] ⑦政府开放数据成效方面的研究。

Sukumar Ganapati、Christopher G. Reddick 针对数据开放的情况对美国州政府的信息主管进行了调查。依据调查显示,将近三分之二的信息主管认为电子政务的开放已经达到了很高的水平,在信息公开透明方面有了很大的进步,但在公民参与和协作的方面所采用的先进方法并不乐观。^[12] Tim Davies 对英国数据网——www.data.gov.uk 开放数据使用情况进行了调查,从理论和实证的角度审视谁在使用政府开放数据和如何使用这些数据以及在什么方面可能会影响到政府民主改革的模式等。^[13]

综上所述,国外政府开放数据的实践比较早,其研究也较成熟,已经涉及数据开放的众多层面。在我国,自2008年《中华人民共和国信息公开条例》公布以来,我国的信息公开进入了实质性的发展阶段,但是我国政府信息公开信息的粒度还没有达到数据的层次。当前,我国学者的研究更多地是在制度、法律、许可层面,以介绍国外政府数据开放的经验为主。同时在技术层面,我国数据开放的本土研究也已经开始,如关联数据技术、语义技术、Web2.0 技术的研究等,但需要不断地学习、应用如何更好地将先进的技术 with 政府信息公开和数据开放结合起来。

1.4 相关概念界定

大数据和政府数据开放都是目前研究的热点问题,对数据、大数据、政府数据开放等概念的厘清和界定对研究的进一步发展具有重要的意义。

1.4.1 数据

数据的概念从古至今也在不断的演变过程中。查阅各种词典,关于数据的定义有很大的不同。汉语词典的定义为:科学实验、检验、统计等所获得的和用于科学研究、技术设计、查证、决策等的数值。^[14] 新华字典中的定义为:电子计算机加工处理的对象。早期的计算机主要用于科学计算,故加工的对象主要是表示数值的数字。现代计算机的应用越来越广,能加工处理的对象包括数字、文字、字母、符号、文件、图像等。^[15] 在百度知道中的定义为:关于自然、社会现象和科学实验的定量或定性的记录。^[16] 在维基百科中定义为:关于事件的一组离散且客观的事实描述,是构成信息和知识的原始材料,是计算机加工的“原料”,如图形、声音、文字、数、字符和符号等。数据可分为模拟数据和数字数据两大类。^[17]

数据(data)起源于拉丁字母的复数形式,是信息科学领域的核心概念,

随着网络技术和信息社会的发展，学者们对数据的概念有了新的认识。

早在 1975 年 Wersig 和 Neveling 在《文档的术语：1200 个基础词汇的选择》一书中认为：数据是被量化的或者合格的数据集。^[18] 1996 年 Wellisch 在《摘要、索引、分类、词库建设：词汇表》一文中认为：在计算机系统中，数据是指编码的不变性。数据是实体或概念的代表，由人类或自动化系统内置，作为交流、解释或者处理的格式。^[19] Chaim Zins 在 2007 年发表的《定义数据、信息、知识的概念方法》一文中又指出：数据是指在计算机使用过程中的记录或者编码中的记录，但是更广范围应用于统计记录和其他记录或者证据的集合。^[20]

随着信息科学中知识管理的兴起，关于数据的定义学者们更多地从数据、信息的关系中开始界定。Chaim Zins 在 2006 年发表的《重新界定信息科学——从信息科学到知识科学》指出：数据是信息的原材料，也是知识的原材料。^[21] Davis 和 Olson 在 1985 年出版的《管理信息系统》一书中指出：信息是被加工后对接收者有意义的数。据。^[22] Checkland 和 Scholes 在 1990 年出版的《在运行当中的软件系统方法论》一书中认为：信息等同于数据+意义。^[23] Lucey 1991 年在《管理信息系统》一书中指出：数据需要被解释和加工才能成为信息。^[24] Warner 在 1996 年的著作《信息系统的交流技巧》中对信息的定义界定为：信息是已经被解释和被接收者所理解的数据。^[25] G. A. Silver 和 M. L. Silver 在 1989 年出版的《系统分析和设计》一书中认为：数据是被加工和提炼成信息的原材料等（见图 1-1）。^[26]

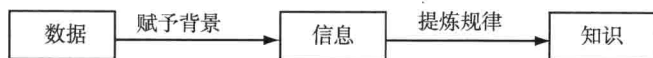


图 1-1 从数据到知识的转化过程图^[27]

数据概念的发展演变是一个循序渐进的过程，尤其是随着大数据技术的发展，对于数据的重新认识和界定加深、拓宽了数据的外延和内涵。在大数据的环境中，本书所指的数据是原始数据，即没有被分析、加工或者处理过的数据，既包括结构化数据，也包括半结构化和非结构化的数据。

1.4.2 大数据^[28]

关于大数据这个词的起源，有学者认为来自美国未来学家托夫勒的《第三次浪潮》，实际上，1997 年 Michael Cox 和 David Ellsworth 在电子电器工程协

会（IEEE）举办的第八届可视化会议上发表的论文中首次提到大数据这个概念，二位学者在文中指出：可视化为计算机系统的发展提出了有趣的挑战：数据集的数量大，加重了内存、磁盘甚至是远程磁盘的负担，将其称为“大数据的问题”。当数据不适合存储在主存储器，甚至是本地磁盘，最常用的方法是获取更多的资源方便存储。

1998年硅谷图形公司的前首席科学家 John Mashey 在题为《大数据下一代架构的压力》报告中以大数据为主题，对大数据的概念做了初步的界定。Mashey 所言的大数据主要指数据容量的快速增长，其中特指互联网等相对较新的数据源，并且介绍了它对存储系统的影响。同样在1998年 S. M. Weiss 和 N. Indurkha 出版的《预测性数据挖掘：实用指导手册》一书中也提到：非常大的数据集被集中存储在数据库当中，让分析家们充分利用强有力的工具全面地分析数据。在理论上，“大数据”在数据挖掘的应用中可以产生更有利的结论，但在实践中困难却很多。^[29]

在1999年8月 Steve Bryson、David Kenwright、Michael Cox、David Ellsworth 和 Robert Haines 在《ACM 通信》中首次以大数据为题目，以论文的形式发表了《科学可视化的大数据》。虽然之前 John Mashey 也发表了以大数据为题目的报告，但普遍认为 John Mashey 的报告属于内部报告，并非正式出版的学术文献。在同年10月 Robert van Liere 和 Sam Uselton 在 IEEE 可视化会议上发表的《自动化或者互动：大数据对谁最有利》的论文中也对大数据进行了探讨。

对大数据这个概念做出突破性界定的当属 Laney。2001年 Garter 的分析师 Laney 提出了 3V 学说，对大数据的认识有了进一步的发展。Laney 认为大数据的增长不仅仅是数量（volume）的增长，而且也是速度（velocity）和种类（variety）的增长。在电子商务环境中，关注信息管理需要更多的合作，可以使企业从信息资产中获得更大的回报。^[30]

自从2001年以来，正如 Randal E. Bryant、Randy H. Katz、Edward D. Lazowska 等在《大数据计算：在商业、科学和社会领域创造革命性的突破》一文所言：“大数据技术可能是过去十年最大的创新。我们已经开始在各行各业的数据收集、组织和处理方面看到了它的潜力。”^[31]

2011年麦肯锡发表的《大数据：下一个创新、竞争和生产率的前沿》是从经济和商业维度诠释大数据发展潜力的第一份专题研究成果，该报告系统阐述了大数据概念，详细列举了大数据的核心技术，深入分析了大数据在不同行业的应用，明确提出了政府和企业决策者应对大数据发展的策略。该报

告认为，“大数据”是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集。该定义有两方面内涵：①符合大数据标准的数据集大小是变化的，会随着时间推移和技术的进步而增长；②不同部门符合大数据标准的数据集大小会存在差别。目前，大数据的一般范围是从几TB到数PB（数千TB）。

Danah Boyd 和 Kate Crawford 在 2012 年发表《反思大数据：针对这一文化、科学和学术现象》一文，在该文中将大数据定义为：一个依赖技术（利用计算和算法的准确性实现对数据的收集、分析和对比等）、分析（依据很大的数据库挖掘类型，并做出经济、社会技术和法律判断）和神话（一种广泛传播的信念：大数据将提供一种更高形式的智慧和知识，它所生成的真实、准确和客观的洞察将是以前的时代所不可能实现的）的互动现象，它所引发的是广泛的乌托邦和反乌托邦的言辞。^[32] 2013 年可以称为大数据年，从新闻到学术机构以及政府、企业对于大数据的报道和研究空前高涨。

从中国期刊网查找显示，我国学者中，最早进行大数据相关探讨的是李京基、姜兰、徐暄于 1985 年在物化探测计算技术上发表的《利用磁盘实现大数据量二维快速傅里叶变换的方法》一文，探讨了在地矿部业京计算中心的国产 150 机上调通了能使 52 万大数据量的航磁数据进行快速傅里叶变换的程序，并对 Ekludh 快速矩阵转置算法作了进一步推广，将它用于通过磁盘实现大数据量二维快速傅里叶变换并获得成功。

在此之后，我国学者关于“大数据”的探讨基本上体现在“大数据量”“大数据集”“大数据文件”方面的研究，并且这方面的研究在逐年增加。尤其在 2011 年关于“大数据”这个概念的文章开始增多，在 2013 年达到 2026 篇，但对大数据含义的认识也基本上沿用国外学者、研究报告的定义。例如：涂兰静在《专家观点：“大数据”与“海量数据”的区别》^[33] 一文中做了分析，将大数据总结为两种认识：①大数据=海量数据+复杂类型的数据；②大数据包含 ABC 三个要素，即大分析（Analytic）、大带宽（Bandwidth）和大内容（Content）。张毅菁在《大数据对我国政府信息公开立法修改的启示》^[34] 沿用的是涂子沛的观点：大数据是指一般的软件工具难以采集、提取、转化、存储和分析的大容量数据，其更大的意义在于：通过对大容量数据的交换、整合和分析，发现新的知识，创造新的价值，带来“大知识”“大科技”“大利润”和“大发展”。冯芷艳、郭迅华、曾大军等在《大数据背景下商务管理研究若干前沿课题》^[35] 一文中列举的定义有：①维基百科的定义，“无法在一定时间内用常规软件工

具对其内容进行抓取、管理和处理的数据集合”；②权威 IT 研究与顾问咨询公司 Gartner 的定义，“在一个或多个维度上超出传统信息技术的处理能力的极端信息管理和处理问题”；③美国国家科学基金会（NSF）的定义，“由科学仪器、传感设备、互联网交易、电子邮件、音视频软件、网络点击流等多种数据源生成的大规模、多元化、复杂、长期的分布式数据集”等。

综上所述，我国学者对于大数据的定义基本上沿用国外研究报告、机构以及学术文献中的定义，对大数据的含义未有定论。本书从来源、特点、来源+特点+价值+属性的角度总结和分析大数据的含义（见表 1-2）。

表 1-2 大数据定义的角度分析

来源	特点	价值	属性
科学仪器、传感设备、互联网交易、电子邮件、音视频软件、网络点击流等产生	3V、4V、4V + 1C “4V + 1C” ^[35] 的特点有：Variety，一般包括结构化、半结构化和非结构化等多类数据，而且它们处理和分析方式有区别；Volume，通过各种设备产生了大量的数据，PB 级别是常态；Velocity，要求快速处理，存在时效性；Vitality，分析和处理模型必须快速变化，因为需求在变；Complexity，处理和难度非常大	大知识、大科技、大利润、大发展等	数据集

1.4.3 政府数据开放

（1）国际组织关于政府数据开放的定义。世界银行在 2012 年发表的《如何认识开放政府数据提高政府的责任感》报告中认为：开放数据指的是非专有的、机器可读的数据，任何人都没有法律或技术的限制，可以自由使用、重复使用、操作和传播，而开放的数据可能来自任何地方。政府开放数据是开放数据的一部分，是指政府产生的、收集和拥有的数据，在知识共享许可下发布，允许共享、分发、修改，甚至对其进行商业使用的具有正当归属的数据。^[37]

（2）政府机构关于数据开放的定义。2012 年德国内政部发布的《德国数据开放》的报告中，将政府数据开放定义为公共行政机构所有的能被第三方重新利用的数据。^[38] 同年，英国政府《开放数据白皮书》中将开放政府数据定义为：公共领域的信息中已经开放给公众使用的数据。^[39]