

DATA

R E S O U R C E



系统介绍大数据资源的权威著作

◎ 数据、技术、应用，大数据三要素。没有数据，大数据就是无米之炊。数据在哪儿？《大数据资源》告诉你，包括政府大数据资源、科学大数据资源、农业大数据资源等。《大数据资源》还阐述数据治理以便合理合法使用数据，阐述数据质量管理以便形成高质量的数据资源。

◎ 主编
——
朱扬勇

大数据 资源

 上海科学技术出版社

大数据资源

朱扬勇 主编

上海科学技术出版社

图书在版编目(CIP)数据

大数据资源/朱扬勇主编. —上海:上海科学技术出版社,2018.1

ISBN 978-7-5478-3426-8

I. ①大… II. ①朱… III. ①数据处理—研究
IV. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 277426 号

大数据资源

朱扬勇 主编

上海世纪出版(集团)有限公司 出版、发行
上海科学技术出版社

(上海钦州南路71号 邮政编码200235 www.sstp.cn)

上海盛通时代印刷有限公司印刷

开本 787×1092 1/16 印张 21

字数:500千字

2018年1月第1版 2018年1月第1次印刷

ISBN 978-7-5478-3426-8/TP·57

定价:80.00元

本书如有缺页、错装或坏损等严重质量问题,
请向工厂联系调换

内容提要

《大数据资源》基于大数据行业的发展情况,选择应用比较热门的行业,对该行业的大数据资源分类、特点和获取方法等进行深入介绍,主要包括金融、能源、农业、制造业、交通、医疗、科学研究等领域。本书共分11章:第1章,绪论;第2章,政府数据资源;第3章,科学数据与资源共享;第4章,农业领域数据资源;第5章,制造业大数据资源;第6章,金融数据资源;第7章,交通数据资源;第8章,能源大数据资源;第9章,医疗数据资源;第10章,数据质量;第11章,大数据治理。

《大数据资源》的读者对象包括计算机学科和数据科学学科的高等院校师生,政务、农业、金融、制造业、医疗、交通、能源、智慧城市等领域应用大数据资源的工程技术人员,以及广大大数据相关专业的管理、决策人员。

《大数据资源》编写人员名单

第1章 朱扬勇

第2章 叶雅珍

第3章 石蕾 王卷乐 高孟绪 王超

第4章 宋长青 李俊清

第5章 张洁 秦威 吕佑龙 汪俊亮

第6章 陈云 张超 俞立 刘可伋

第7章 翟希 何承 顾承华 张扬

第8章 任庚坡 葛志松 毛俊鹏

第9章 汤春蕾

第10章 蔡莉 朱扬勇

第11章 杨琳 高洪美 宋俊典 张绍华

序

2008年,熊贲教授和我发表了一篇题为“加强数据资源保护和开发利用”的文章。我们提出了“数据资源是重要的现代战略资源,其重要程度将越来越显现,在本世纪有可能超过石油、煤炭、矿产,成为最重要的人类资源之一”;“数据资源开发利用滞后于网络基础设施和应用系统的建设,制约了国家信息化的综合效益”;“数据资源保护不利、开发不足、利用不够的现象将长期存在”;“提高数据资源开发利用水平、保护国家的战略资源是增强我国综合国力和国际竞争力的必然选择”。也是在这篇文章中,我们第一次提出“数据界(DataNature)”、“数据科学(DataScience)”和“数据学(Dataology)”。我们还建议“从信息化转向数据资源开发利用”、“政府政务公开数据要有限度”、“加强国家、企业和公民隐私数据保护”。很欣慰,这些观点现在都已经被大家所接受。2012年,我提出“大数据包括数据、技术和应用三个要素”,2015年熊贲教授和我在《大数据》创刊号上以“大数据是数据、技术还是应用”为题对此进行了系统论述。也很欣慰,大数据包含数据、技术和应用三个要素的观点正被广泛接受。在大数据发展开始进入理性并逐步落地的阶段,我想是时候将数据、技术和应用三个大数据要素更详细地进行阐述,为国家的大数据发展尽绵薄之力。于是在和上海科学技术出版社商量之后,我开始组织编写《大数据资源》、《大数据技术》和《大数据应用》,也可以算“大数据三部曲”吧。

《大数据资源》主要阐述什么是数据资源;如何进行数据治理;在建设数据资源过程中,如何控制数据质量,以便将来形成高质量的数据资源。还给出一些典型的大数据资源,包括政府数据资源、科学大数据资源、农业数据资源、金融数据资源、交通大数据资源、制造业大数据资源、能源大数据资源、医疗数据资源等,一个领域的大数据资源包括本领域生产的数据、领域外部生产的和本领域数据分析相关的数据。作者尽可能列出这些数据资源的出处,便于读者在实际应用中能够找到这些数据资源。参与编写的有叶雅珍、石蕾、王卷乐、高孟绪、王超、宋长青、李俊清、张洁、秦威、吕佑龙、汪俊亮、陈云、张超、俞立、刘可伋、翟希、何承、顾承华、张扬、任庚坡、葛志松、毛俊鹏、汤春蕾、蔡莉、杨琳、高洪美、宋俊典、张绍华等,感谢

这些作者的辛勤劳动。

《大数据技术》主要阐述大数据技术。信息化是生产数据的,大数据是开发数据的,开发数据的技术称为数据技术;更重要的是,信息化是“技术进步促进数据增长”,而大数据是“数据增长促进技术进步”。面对日益增长的数据规模,大数据技术对人类社会发展意义重大。希望2018年完成《大数据技术》。

《大数据应用》主要给出了一些大数据应用的案例。我计划最后来写《大数据应用》。主要原因是目前很多关于大数据的美丽故事,离我们理解的大数据还有差距。现在还没有让我满意的大数据应用案例,希望在未来两年能够收集到足够好的大数据应用案例。希望2019年完成《大数据应用》。

想写好大数据三部曲的心情不言而喻。《大数据资源》已成,所有参与的作者都非常努力和认真,表现出高水平,但由于我本人知识水平和组织能力的限制,书稿还是有许多不满意和遗憾,在此我向读者表示歉意,向参与的作者表示歉意。有位编辑说,写本书能引起大家批评也不错,说明大家在关注。所以,等待读者的批判,感谢。

《大数据资源》即将交付印刷了,不能再修改和等待了。刚好今天航程将近6小时,让我在白天能够有这么长时间的宁静,思绪到远方,写下这些文字,作为序。

朱扬勇

2017年11月8日

目 录

第1章 绪论	1
1.1 基本概念	1
1.1.1 数据	1
1.1.2 数据界	2
1.1.3 大数据	4
1.2 数据资源	6
1.2.1 数据资源的形成	6
1.2.2 数据矿床	7
1.2.3 数据资源的战略性	8
1.3 数据资源建设	8
1.3.1 面临的问题	9
1.3.2 数据权属	9
1.3.3 国有数据资源和市场数据资源	10
1.4 数据资源开发	11
1.4.1 大数据与信息化	11
1.4.2 数据开发的“6用”问题	12
1.4.3 数据流通	12
1.4.4 数据产业	14
1.5 小结	15
参考文献	16
第2章 政府数据资源	17
2.1 政府数据开放	17
2.1.1 政府数据的特点与类型	17
2.1.2 政府数据开放与共享	18

2.1.3	政府数据开放的基本做法	19
2.1.4	政府数据管理与治理	19
2.2	中国政府数据资源与开放	20
2.2.1	政府数据资源	20
2.2.2	公共资源数据库	21
2.2.3	中国政府数据开放	23
2.3	国外政府数据开放	29
2.3.1	美国政府数据开放	30
2.3.2	英国政府数据开放	32
2.3.3	新加坡政府数据开放	34
2.3.4	其他国家政府数据开放	36
2.4	国际组织数据开放	39
2.4.1	欧盟	39
2.4.2	世界银行	41
2.4.3	经济合作与发展组织(OECD)	42
2.5	小结	43
	参考文献	46

第3章 科学数据与资源共享 48

3.1	科学数据的特征、机遇与挑战	48
3.1.1	特征与范围	48
3.1.2	机遇与挑战	49
3.2	科学数据的全生命周期	52
3.2.1	全生命周期概述	52
3.2.2	科学数据的采集与生产	53
3.2.3	科学数据的加工与保存	54
3.2.4	科学数据的共享服务	55
3.3	我国科学数据的管理与开放共享	56
3.3.1	科学数据的总体规模	56
3.3.2	科学数据的管理	58
3.3.3	科学数据的开放共享	60
3.3.4	科学数据目前存在的主要问题	61
3.4	我国科学数据的发展建议	63
3.4.1	科学数据发展的政策机制与标准规范	63
3.4.2	科学数据的整合与产业化发展	63
3.4.3	科学数据的管理和知识挖掘	64
3.4.4	科学数据共享服务	65

3.4.5	科学数据基础设施建设	65
3.4.6	科学数据资源保护和知识产权	66
3.4.7	科学数据发展的人才队伍和科技投入	67
	参考文献	67
第4章 农业数据资源		68
4.1	农业积累的数据资源	68
4.1.1	种植业数据资源	68
4.1.2	林业数据资源	71
4.1.3	畜牧业数据资源	74
4.1.4	渔业数据资源	76
4.1.5	农业水利数据资源	78
4.1.6	农产品加工数据资源	80
4.2	农业相关领域的的数据资源	81
4.2.1	生物基因数据资源	82
4.2.2	气候气象数据资源	83
4.2.3	地理数据资源	84
4.2.4	农业生产资料数据资源	86
4.2.5	农产品物流与市场数据资源	87
4.2.6	国际农业数据资源	89
4.3	农业领域相关的数据资源机构	91
4.3.1	国内农业大数据及相关领域科学数据资源所在机构列表	91
4.3.2	国外农业大数据及相关领域科学数据资源所在机构列表	92
4.4	农业领域数据资源的获取途径和方法	94
4.4.1	农业领域数据资源的获取要求	94
4.4.2	农业领域数据资源的获取途径	95
4.4.3	农业领域数据资源的主要获取方法	96
4.5	小结	97
	参考文献	97
第5章 制造业大数据资源		98
5.1	大数据：制造业的新资源	98
5.1.1	大数据与新一轮制造业革命	99
5.1.2	制造业的资源组成体系	107
5.1.3	大数据与传统制造资源的关系	108

5.1.4	制造业大数据资源的构成	111
5.2	企业内部制造业大数据资源	113
5.2.1	产品数据资源	113
5.2.2	工艺数据资源	119
5.2.3	生产运行数据资源	120
5.3	企业外部制造业大数据资源	125
5.3.1	设计相关外部数据	126
5.3.2	工艺相关外部数据	126
5.3.3	生产运行相关外部数据	127
5.4	制造业大数据资源机构与获取途径	131
5.4.1	行业大数据资源机构	131
5.4.2	企业运营大数据资源机构	133
5.4.3	物流大数据资源机构	135
5.4.4	工商大数据资源机构	137
5.5	小结	139
第6章 金融数据资源		140
6.1	金融行业数据资源	140
6.1.1	证券期货数据资源	140
6.1.2	银行数据资源	142
6.1.3	保险数据资源	145
6.1.4	跨行业互联网金融数据	148
6.1.5	外汇数据资源	150
6.2	与金融业相关的数据资源	152
6.2.1	国内相关数据资源	152
6.2.2	国外相关数据资源	162
6.3	金融数据资源的主要来源	167
6.3.1	金融相关数据库简介	167
6.3.2	金融相关网站简介	169
	参考文献	171
第7章 交通数据资源		172
7.1	城市交通数据资源	172
7.1.1	城市交通数据资源的分类与组成	172
7.1.2	道路交通行业数据	172
7.1.3	公交行业数据	176

7.1.4	轨道行业数据	178
7.1.5	出租车和停车行业数据	179
7.2	与交通相关的行业数据资源	180
7.2.1	支撑交通管理决策的相关行业数据	180
7.2.2	与交通互为影响的相关行业数据	183
7.3	交通数据资源所有机构	186
7.3.1	政府交通主管部门	186
7.3.2	交通运输相关企业	188
7.3.3	运营商及其他来源	189
7.4	交通数据资源获取的途径	189
7.4.1	源数据获取的方式	189
7.4.2	数据获取的媒介	191
7.4.3	数据获取的途径	191
7.5	典型交通大数据资源机构情况介绍——上海交通大数据资源中心	192
	参考文献	195
第8章	能源大数据资源	196
<hr/>		
8.1	能源大数据积累的数据资源	196
8.1.1	能源大数据信息简介	196
8.1.2	能源大数据信息基础数据的采集	201
8.2	能源大数据的信息特征与价值	204
8.3	能源大数据的采集、传输、存储和分析处理	207
8.3.1	能源大数据采集技术	207
8.3.2	能源大数据传输技术	209
8.3.3	能源大数据存储技术	215
8.3.4	能源大数据分析处理平台	220
8.4	能源大数据资源机构与获取途径	223
8.4.1	能源领域相关的数据资源机构	223
8.4.2	能源大数据资源的获取途径和方法	224
第9章	医疗数据资源	225
<hr/>		
9.1	医疗数据的特征、问题与挑战	225
9.1.1	数据壁垒、隐私和安全	225
9.1.2	医疗数据的国际差异	226
9.2	临床医疗数据资源	227
9.2.1	电子病历数据	227

9.2.2	临床笔记数据	228
9.2.3	医学影像数据	228
9.2.4	临床试验数据	229
9.3	非临床医疗数据资源	229
9.3.1	队列研究数据	229
9.3.2	生物组学数据	230
9.3.3	文献典籍数据	231
9.3.4	药学数据	232
9.3.5	医疗事务数据	232
9.3.6	医保索赔数据	233
9.4	医疗相关领域数据资源	233
9.4.1	环境医学数据	233
9.4.2	互联网数据	234
9.4.3	社交媒体数据	234
9.4.4	物联网数据	235
9.4.5	移动互联数据	235
9.5	医疗数据的产业化发展	236
9.5.1	数据创新转化医学	236
9.5.2	跨境医疗中的数据共享	237
9.5.3	区域医疗中的结果共享	237
9.6	小结	238
	参考文献	238

第10章 数据质量 240

10.1	数据质量概述	240
10.1.1	数据质量带来的影响	240
10.1.2	影响数据质量的因素	241
10.1.3	数据质量定义	243
10.1.4	大数据时代数据质量面临的挑战	244
10.2	数据质量标准	245
10.2.1	ISO 8000 国际标准	245
10.2.2	地理信息质量标准 ISO 19100	247
10.2.3	统计数据质量标准	249
10.3	数据质量相关技术	250
10.3.1	数据集成	250
10.3.2	数据剖析	252
10.3.3	数据清洁	254

10.3.4	数据溯源	257
10.4	数据质量评估	258
10.4.1	数据质量维度	258
10.4.2	数据质量评估框架	259
10.4.3	数据质量评估方法	263
10.5	数据质量管理	264
10.5.1	数据质量管理方法	264
10.5.2	数据质量管理团队建设	266
10.5.3	质量管理成熟度模型	267
10.6	小结	269
	参考文献	270

第 11 章 大数据治理 274

11.1	大数据治理概述	274
11.1.1	国内外数据治理研究成果	274
11.1.2	大数据治理定义	278
11.1.3	大数据治理的重要性	279
11.1.4	大数据治理的范围	280
11.2	大数据战略和组织	282
11.2.1	大数据战略指明企业转型的方向	282
11.2.2	企业制定大数据战略的要点	283
11.2.3	大数据战略对组织的影响	284
11.3	大数据架构	286
11.3.1	大数据架构参考模型	286
11.3.2	大数据架构的实现	289
11.4	大数据安全和隐私保护	292
11.4.1	大数据安全和隐私的问题与挑战	293
11.4.2	大数据安全防护	295
11.4.3	大数据隐私保护	298
11.5	大数据质量管理的重要性和复杂性	301
11.5.1	大数据质量管理重要性	301
11.5.2	大数据质量管理复杂性	302
11.6	大数据生命周期管理	302
11.6.1	大数据生命周期概述	303
11.6.2	大数据采集	303
11.6.3	大数据存储	305
11.6.4	大数据整合	306

11.6.5	大数据呈现与使用	308
11.6.6	大数据分析与应用	309
11.6.7	大数据归档与销毁	310
11.7	大数据治理实施	311
11.7.1	大数据治理实施的目标和动力	311
11.7.2	大数据治理实施关键要素	313
11.7.3	大数据治理实施过程	315
11.7.4	大数据治理实施路线图	317
	参考文献	317

第1章

Chapter 1

绪 论

网络空间(cyber space)是指计算机网络、广电网络、通信网络、物联网、卫星网等所有人造网络和设备构成的空间,这个空间真实存在。信息化的本质是将现实世界中的事物转化成数据并存储到网络空间中,即信息化是一个生产数据的过程。随着信息化的普及、深入和持续发展,生产的数据越来越多并积累了下来,形成一个个大规模的数据集。其中,具有开发价值的数据集就是数据资源,而网络空间中的所有数据则构成数据界(DataNature)^[1,2]。本章介绍有关数据的基本概念,包括数据、信息化、大数据、数据界等,介绍了数据资源、数据资源开发、数据产业等内容,作为本书的一个导引。

1.1 基本概念

大数据是当前的一个热词。在大数据之前是信息、信息科学、信息技术和信息产业等,为什么今天叫“大数据”而不是“大信息”?一个显而易见的事实是“大数据里不一定有大信息”,这也是本书称为数据资源而不是信息资源的原因。下面介绍数据、信息化、大数据、数据界等基本概念。

1.1.1 数据

“数据”的含义很广,不仅指 1011、8084 这样一些传统意义上的数据,还指“dataology”“上海市数据科学重点实验室”“2013/09/06”等符号、字符、日期形式的数据,也包括文本、声音、图像、照片和视频等类型的数据,而微博、微信、购物记录、住宿记录、乘飞机记录、银行消费记录、政府文件等也都是数据。

数据是指能够输入到网络空间中的任何东西,是指网络空间中唯一存在的,是可度量的、可处理的、可观测的,并占有空间的。直观上,可以对数据进行如下分类:

1) 依据数据表示的含义来划分

从数据表示的含义来分,数据可以分为两类:一类是表示现实事物的数据,称为现实数据;另一类则不表示现实事物,只在网络空间中存在,称为非现实数据^[1]。

(1) 现实数据主要包括:

① 感知数据——通过感知设备(如温度传感器、天文望远镜)感知现实世界获得的数

据,包括感知生命的数据。这类数据是现实世界的直接反映。

② 行为数据——人类科学研究、劳动生产、生活行为等所产生的数据。这类数据是人类行为的直接反映。

(2) 非现实数据种类繁多,目前还不能很好地进行分类,举例如下:

① 计算机病毒——能够自我复制和传播的计算机程序,只在数据界中存在,而在自然界没有映射。

② 网络游戏——包括与自然界对应的场景映射到数据界中,也有只在数据界中的游戏场景设置。

③ 垃圾数据——没有任何含义的数据。

2) 依据数据的权属来划分

数据权属还没有法律的界定,从情理上看,数据非天然,数据理应属于数据的生产者。但实际情况往往比较复杂,从目前数据的生产和数据被占有的情况来看,数据可以分成如下类别:

(1) 私有数据:指个人或组织自己生产、自己保管、非公开的数据,这类数据权属清晰。

(2) 多方生产的数据:大部分数据是由很多方共同生产的,如电商平台、银行、电信、医院等的数据都是多方生产的。电商平台的数据是由购物者、网店卖家、支付系统、物流系统、平台等共同生产,这些数据的权属没有界定。电商数据目前基本上是电商平台占有并获取利益,购物者和卖家没有主张权利。但是,如果医院的数据被医院占有并谋取利益,民众就会强烈反对。因此,这类数据的权属有待法律界进行法律界定,以避免数据的灰色地带和数据黑产。

(3) 政府数据:主要指政务数据、政府财政投资生产的数据,以及国有企业数据。这部分数据权属属于政府。

(4) 公网数据:主要是指发布在公共网站上的数据,这些数据能够通过搜索引擎访问到。如果按照目前的物权法和知识产权法,这类数据权属属于数据的原创者,也是不能随便下载使用。但是,在公共网络上下载数据是普遍的行为,应该受到法律的保护。因此,这类数据的权属也同样有待法律界进行法律界定。

3) 依据数据的组织形式来划分

从数据的组织形式来看,数据主要有下列一些组织形式:

(1) 专用格式数据:有相当多的数据是由专用数字化设备产生的数据,如医学影像数据(X线片、MR、CT等)、遥感数据。还有一些是GIS、多媒体等数据。这些数据的处理需要专门的设备或专门的软件。

(2) 通用格式数据:在信息化早期,大多数数据库是存储在通用数据库中的,由通用的数据库管理系统来管理。这些数据库结构清楚,处理方便。

(3) 互联网数据:互联网上的数据门类和格式繁多,还包括很多垃圾数据、病毒数据,关键是找到有用的数据。由于互联网数据的形成,使得整个网络空间中的数据更加显现出自自然界的一些特征。

1.1.2 数据界

人类社会的进步发展是人类不断探索自然(宇宙和生命)的过程,当人们将探索自然界