DATA

RESOURCE

大数据资源

⊙主编——朱扬勇

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。

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
上海科学技术出版社 出版、发行
(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

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，熊赟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年熊赟教授和我在《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

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