BIG DATA GOVERNANCE

AN EMERGING IMPERATIVE

大数据治理

[美] 桑尼尔·索雷斯（SUNIL SOARES）著

匡 斌 译

重磅推荐！

工业和信息化部软件服务业司司长 陈伟

11带资本董事长 田溯宁

IBM全球副总裁兼IBM软件集团大中华区总经理 胡世忠

清华大学出版社

BIG DATA GOVERNANCE

AN EMERGING IMPERATIVE

大数据治理

[美] 桑尼尔·索雷斯（SUNIL SOARES）著
匡 斌 译

清华大学出版社
北 京

内 容 简 介

大数据将打开各行各业的数据“潘多拉魔盒”。社交网站、电商巨头、电信运营商乃至金融、医疗、教育等行业，都将加入大数据的“淘金”热潮，政府部门同样会从大数据中获益匪浅。如何将海量数据应用于决策、营销和产品创新？如何利用大数据平台优化产品、流程和服务？如何利用大数据更科学地制定公共政策、实现社会治理？所有这一切，都离不开大数据治理。可以说，在大数据战略从顶层设计到底层实现的“落地”过程中，治理是基础，技术是承载，分析是手段，应用是目的。《大数据治理》一书的翻译出版，正当其时。

《大数据治理》一书较好地满足了理解大数据治理框架的需要，系统地阐述了大数据治理的各个版块，分析了五大类大数据的治理，考察了大数据治理在典型行业的实践，并深入浅出地介绍了当今主流的大数据技术与平台。该书具有一定的可参照性、可操作性和可读性，是大数据治理领域值得一读的参考书。

图书在版编目(CIP)数据

大数据治理/(美)索雷斯(Soares,S.)著；匡斌译．—北京：清华大学出版社，2014
书名原文：Big data governance:an emerging imperative
ISBN 978-7-302-36406-1

Ⅰ．①大… Ⅱ．①索… ②匡… Ⅲ．①数据管理—研究 Ⅳ．①TP274

中国版本图书馆 CIP 数据核字(2014)第 098469 号

责任编辑：刘志彬
封面设计：汉风唐韵
责任校对：宋玉莲
责任印制：宋 林

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社总机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经 销：全国新华书店
开 本：170mm×240mm **印 张**：19.75 **字 数**：262 千字
版 次：2014 年 6 月第 1 版 **印 次**：2014 年 6 月第1 次印刷
定 价：49.00 元

产品编号：056084-01

谨以此书献给我漂亮的女儿丽兹和玛雅

感谢我的妻子卡莉娜在本书写作中给予的支持

感谢我的父母赛西利亚和胡蓓特的祈祷和指导

对本书的赞誉

大数据涉及不同来源的复杂数据。倘若缺乏得当的数据治理，那就很难正确地整合数据。《大数据治理》一书为您提供了制订大数据治理计划所必需的信息和见识，而大数据治理计划是支持大数据整合项目不可或缺的。好样的，桑尼尔！

Symphony IRI Group 技术研究副总裁
杰·犹斯科博士

本书是一个信息治理专家奉献的鸿篇巨制，作者以极其实用和通俗易懂的风格，倾心向读者解读大数据治理这一复杂主题。

作为一家大公司的资深 IT 专家，我本人在面对数据窘境时，感觉不知所措。对数据领域的从业人员来说，面临的问题多过答案。我所在的组织是南非的主导电信运营商，我们拥有海量的电话详单、位置数据和社交媒体生成的数据。要明智地使用数据，就必须管理所有数据。

本书匠心独运，揭开了大数据的迷人景致，为我们应对大数据领域的挑战，提供了必要的智力成果。

本书的字里行间，流淌着丰富的信息。如今，我终于有机会将本书所述的理念和知识融会贯通。我更有信心应对公司面临的大数据挑战，对此，我满怀热忱，决心已定。

拜桑尼尔在本书中提供的指南所赐，我们所有数据从业人员都将获得成功！

南非电信数据治理办公室主任
柯马林·伽迪

中文版推荐序之一

数据如涓涓细流，自人类结绳记事起，流淌到今。随着移动互联网、云计算、物联网等新技术浪潮的兴起，数据的体量、类型、速度和价值，在很短时间内达到前所未有的程度。IDC 预计，全球数据总量将以每两年翻一番的速度增长，2020 年将达到 40ZB，是 2011 年的 22 倍。以传感器数据为例，“万亿传感器社会”(Trillion Sensors Universe)终将从憧憬与倡议走进现实，传感器数据的爆炸式增长及与其他数据的整合，将极大地提升数据的价值。

传统数据已完成了质变，大数据破土而出。从严格意义上说，大数据并非颠覆式创新，但它将引发翻天覆地的产业变革，却是政、产、学、研、用等各界的共识。

大数据蕴藏的巨大价值，引起了各国政府的高度重视。2012 年，奥巴马政府出台了“大数据研究和发展计划”，将大数据提升到国家战略和国家竞争力的高度。2013 年，澳大利亚政府发布了《公共服务大数据战略》。2013 年，中国国务院出台了《推进物联网有序健康发展的指导意见》，正式将促进大数据发展纳入国家产业政策。未来，中国的大数据国家战略，有望在关键技术的突破、人才的培养、行业标准的制定、产业政策的支持、数据保护和安全防范、试点与示范，以及法律法规的完善等方面取得重要进展。

大数据将打开各行各业的数据“潘多拉魔盒”。社交网站、电商巨头、电信运营商乃至金融、医疗、教育等行业，都将加入大数据的“淘金”热潮，政府部门同样会从大数据中获益匪浅。如何将海量数据应用于决策、营销和产品创新？如何利用大数据平台优化产品、流程和服务？如何利用大数据更科学地制定公共政策、实现社会治理？所有这一切，都离不开大数据治理。可以说，在大数据战略从顶层设计到底层实现的“落地”过程中，治理是基础，技术是承载，分析是手段，应用是目的。《大数据治理》一书的翻译出版，正当其时。

大数据治理关乎组织的所有部门和全部流程，贯穿数据的整个生命周期。要打破数据分割的窘境，实现大数据价值的最大化，就要聚合分散孤立、类型各异的大数据集；要提高数据的质量，最大限度地使用可信的数据，就需要清洗和净化各种类型的数据；要消除数据风险、保护用户隐私，就需要对大数据进行加密和屏蔽……

毋庸置疑，大数据治理已不单是一个软件问题，还涉及人员和流程。大数据的价值变现，依然存在着安全问题、隐私问题等诸多制约因素。但从长远看，大数据所蕴藏的极大正能量，一定会得到释放。

《大数据治理》一书较好地满足了理解大数据治理框架的需要，系统地阐述了大数据治理的各个版块，分析了五大类大数据的治理，考察了大数据治理在典型行业的实践，并深入浅出地介绍了当今主流的大数据技术与平台。该书具有一定的可参照性、可操作性和可读性，是大数据治理领域值得一读的参考书。

中国是人口大国和全球第二大经济体，是名副其实的大数据大国。中国的移动电话用户数、固定电话用户数、互联网/移动互联网用户数、社交媒体用户数、电子商务交易额、软件产业规模等指标，均令全球其他国家和地区瞩目。与此同时，华为、东软、曙光、浪潮等软件业的民族企业，百度、腾讯、阿里巴巴和网易等本土IT翘楚，在大数据技术研发、平台部署、方案推广和分析应用等重要领域，已取得初步成果。

相信在可预见的将来，经过循序渐进的治理，大数据将成为重要的国家资源和企业的核心生产要素。大数据将给中国的政府、企业和其他组织，带来切切实实的收益。

工业和信息化部软件服务业司司长　陈伟　教授

2014年1月16日

中文版推荐序之二

在不到两年时间中，大数据迅速成为热门词，但对其的解读，却见仁见智。数据科学家醉心于前沿的数据技术开发，经济学家关注大数据的产业价值，企业家期盼大数据的阳光照进日常的经营现实，法学家强调隐私保护……

欣慰的是，拥抱大数据成为各方的共识，且思且行的大数据“淘金”之旅，已然启动。大数据的“淘金”之旅，需要脚踏实地的努力。大数据治理是连接大数据科学和应用的桥梁，若要到达风光无限的大数据彼岸，大数据治理一定是“必修课”之一。要实现大数据的变现，就离不开科学的大数据治理，离不开与时俱进的管理革新。因此，桑尼尔的《大数据治理》一书，可谓应运而生。中国联通研究院的匡斌先生将该书翻译成中文，相信对中国读者会有所助益。

大数据治理是传统信息治理的延续和扩展。它不可能与传统的信息治理切割，延续性既是保护历史投资的需要，也体现了信息治理准则的一脉相承。

不同类型数据的整合，结构化数据与非结构化数据、准结构化数据的整合，主数据与社交媒体等其他类型数据的整合，不同部门乃至不同行业数据的整合，都需要大量细致的工作。大数据治理涉及人员、流程和软件，大数据需要去伪存真，需要删繁就简，需要化大为小。凡此种种，不胜枚举。

大数据治理的约束条件构成一个三层结构的金字塔，最底层无疑是特定的文化背景和规制环境。根深蒂固的隐私文化，动态演进的隐私规制，是发掘大数据价值面临的最大挑战。第二层则是技术。大数据技术是治理大数据的基础，前向兼容、后向扩展、简便易用的大数据平台和解决方案，自然语言处理、人脸识别等非结构化数据处理等技术，形成“物”的制约。第三层则是人的因素。大数据治理呼唤大批熟稔大数据技术的人才，也需要更多的大数据管理者和应用开发者，他

们可以得心应手甚至出神入化地将技术、行业、流程、功能等进行整合。

说到底，大数据治理的核心是人。人既是大数据价值的追求者，又是大数据隐私的主体和捍卫者。就这个意义而言，人的因素是大数据治理的最大制约。人类历史上每一个技术发明与创造，均有“善”与“恶”两面，文明的进步就是发挥技术“善”的一面，治理控制“恶”的一面。

《大数据治理》一书以实用性为导向，通过教科书式的体例安排，对大数据治理进行了全方位的解构，并将大数据治理规程化。对于尚处于大数据战略起飞阶段的组织，本书是一本很好的大数据治理参考蓝本。作者举重若轻，以朴素无华的语言，微言大义的案例，为致力于大数据治理的实操者，奉献了一本有价值的通俗读物。

纵观当今的大数据技术、平台和解决方案，海外厂商仍占据了绝对主流地位。有关大数据的研究和著述，同样如此。现阶段，“拿来主义”尤有必要。从大数据的体量看，中国在大数据领域的潜在地位，无异于中东地区在石油业的地位。相信在不远的将来，在大数据领域，中国将异军突起。

大数据的思想启蒙运动正在进行。从大数据治理起步，不断探索这个领域的产权、法律和交易等问题，才能成为进入大数据世界的先行者。

宽带资本董事长　田溯宁

2014 年 1 月 10 日

英文版推荐序之一

当下，我们生活在一个数据看似无穷无尽的年代。数据已经润物细无声般浸淫到我们的生活中，这是真实的一幕。我们仰仗数据完成各色各样的任务，从治理经济和推动科学进步，到保存记录健康信息的电子病历，不一而足。我们已逐渐意识到，必须获取大数据分析所呈现的洞察力，将洞察力转化为信息、知识和最终行动，真正了解和解密大数据的价值。近年来，随着社交媒体、传感器网络、流数据技术和类似技术的涌现，需要揭开大数据价值的大量工作，已超出传统数据库和数据仓库技术的能力范围。换言之，我们已进入大数据时代。

何为大数据？何为大数据治理？又如何通过大数据创造商业价值？本书回答了这些问题以及有关大数据的更多问题。如果您还不能回答上述问题，那么您有必要阅读此书，否则，您将会错过利用大数据实现业务差异化和增长的机会。

通过详尽的案例研究，本书作者桑尼尔做了出色的工作，帮助读者了解一个复杂并尚处于演变中的主题。譬如，他分享了这样一个故事：零售商店使用社交媒体数据，分析消费者对各种不同产品的意识和情绪。通过此种方式，零售商可以优化产品折扣，从而极大地提高公司的利润。

在本书中，桑尼尔分享了二十多个案例研究。在阅读这些案例研究时，我鼓励读者进行“二步法”的思想训练，就像我所做的这样。第一步，考虑怎样可以尝试达到案例研究中的目的，但必须使用传统的数据技术和流程。回到上述的零售商实例，通过市场调查，可以了解消费者对各种产品的意识和市场情绪，但这个流程要耗费数个月时间。在研究结果可用时，对实际设定价格折扣已不再有效，因为客户的意识和情绪早已时过境迁。我们如今处于一个必须利用社交媒体数据的时代，使用大数据分析法，可以提供比传统方法更及时的商业

智能。第二步，考虑存在于既有数据管理流程和系统间的差距。不妨再次以零售商为例，毋庸置疑的是，我们必须能够实现社交媒体数据与产品的匹配，但是，传统主数据管理系统中的匹配规则，可能无法胜任此项工作。比如，新的匹配规则要求具备自然语言处理能力，这就是差距；然后，我们需要提供针对此类规则的治理，这又是一个差距。如此，等等。

通过识别和讨论每个案例研究中的差距，桑尼尔将此种比较训练变得方便起来。最终结果是，读者能够深入了解大数据治理，同时得到上述问题的答案。

无须多说了。阅读此书吧，它将带您迈入大数据时代。

尹德帕尔·班达利

美国快捷药方公司副总裁兼首席数据官

英文版推荐序之二

“要治理，早治理”，如今已成为聪明的企业和解决方案架构者的激昂呼喊，这些企业和个人被授权设定企业数据整合计划的范围和方向——特别是当前“大数据”整合的顶级要求。了解大数据治理的范围、多样性和整合性难题，是极富挑战性的。

桑尼尔·索雷斯的新书，是IT专家规划其从传统IT实践到“大数据”最新技术趋势及“大数据”分析组件的“罗塞达石碑”[①]。具体而言，除其他益处外，该书提供了一个颇受好评的大数据参考架构，以及展现了大数据分析的业务成果的案例研究。

显然，在实践中，大数据需要有效和可持续的主数据管理（MDM）和数据治理。2013年到2014年间，大数据将回归到作为另一个数据源的MDM构造之中。特别是，大数据应用需要具备MDM意识，而MDM也可存储每个大数据源的首选项。此外，大数据分析需要了解有效的客户和产品主视图。最后，利用大数据挖掘填充社交MDM，并在大数据存储区运行实体匹配，将提供此类实体来自公开版、发行版和企业版数据反馈的全方位视图。大数据需要根据预期商业战略予以治理，这反过来又在积极的组织数据治理政策中得到反映。不然的话，由于可能的过度投资，加之缺乏合适的商业战略执行力，企业的经济和策略风险将翻番。

如果您的数据治理计划设计得当，那么您应该扩展数据治理流程，以将企业管理的数据和大数据的更广阔世界囊入其中。桑尼尔此书的看点之一是，认识到数据的规模并不像如何使用它们那样重要——深处其中的IT专家，必须学会拥抱大数据及其治理等新一代

---

① 罗塞达石碑（Rosetta Stone）制作于公元前196年，用三种文字刻有古埃及法老托勒密五世的诏书。其独特的三语对照写法，成为解码失传千余年的埃及象形文的关键。“罗塞达石碑”暗喻解决某个谜题或难题的关键线索或工具。——译者注

技术解决方案。

桑尼尔关于“大数据治理”的新书是一个丰富的源泉，是您在大数据之旅中的称职导师和指南。

阿朗·佐尼斯

MDM 协会首席研究官，“MDM 暨数据治理峰会”（伦敦、纽约、旧金山、上海、新加坡、悉尼、东京、多伦多）会议主席

作 者 自 序

我的处女作《IBM 数据治理统一流程》(MC Press,2010),列出了实施信息治理计划的 14 个主步骤和 100 个子步骤。该书聚焦于实施数据治理计划的人员、流程和软件工具。其中的 14 个主步骤如下。

1. 定义业务问题
2. 取得管理层支持
3. 实施成熟度评估
4. 建立路线图
5. 制定组织蓝图
6. 创建业务词库
7. 了解数据
8. 创建元数据库
9. 定义度量标准
10. 治理主数据
11. 对分析进行治理
12. 管理安全与隐私
13. 治理信息生命周期
14. 对结果进行测量

我的第二本书《说服企业实施信息治理:行业和工作职能最佳实践》(MC Press,2011),考察了推销信息治理价值的最佳实践。该书分为三个部分:

1. 行业最佳实践

银行和金融市场、保险、医疗保健、制造、零售、旅游与交通、政府、石油与天然气、电信和公用事业部门等行业中信息治理准则的应用。

2. 工作职能最佳实践

销售和市场营销、财务、信息技术、信息安全与隐私、人力资源、法律和合规、运营、供应链和生产管理等关键工作职能中信息治理准则

的应用。

3. 跨行业最佳实践

与角色和义务、度量标准、元数据、成熟度评估、业务案例和参考架构有关的最佳实践。在跨工作职能、行业和地域的信息治理计划中，这些主题表现出一致性。

该书旨在提供一个典型样本，而不是一个推销组织内信息治理价值的详细清单。第一本书侧重实施信息治理计划的最佳实践，而第二本书则探索从信息治理计划中产生组织收益的最佳实践。

大数据治理

大数据包括 Twitter 消息、Facebook 帖子、博客、LinkedIn 资料、蜂窝电话 GPS 信号、机器日志和无线射频识别(RFID)数据，此处仅举此数类。大数据本质上是动态的，并具有体量大、速度快和格式多样化的特征，包括结构化、非结构化和准结构化数据。正像需要治理主数据和参考数据等其他类型数据一样，组织也必须治理大数据。我们将大数据的治理称为“大数据治理(big data governance)”。

本书汇集了我前两本书的最佳实践，按照准则、数据类型和行业，讨论了大数据的治理。全书分为五个部分：

1. 开篇

第 1 章～第 5 章，对大数据治理框架、成熟度评估、业务案例和路线图进行了概述。

2. 大数据治理准则

第 6 章～第 12 章，讨论了如何应用组织、元数据、隐私、数据质量、业务流程整合、主数据整合和大数据生命周期管理的信息治理准则。

3. 各类大数据的治理

第 13 章～第 17 章，介绍了与网络(Web)和社交媒体数据、机器对机器的数据、大体量交易数据、生物计量学数据和人工生成数据的治理有关的最佳实践。

4. 行业视角

第 18 章～第 20 章，介绍了医疗保健机构、公用事业部门和通信服务提供商中，大数据治理的最佳实践与案例研究。

5. 大数据技术

第21章和第22章，提供了大数据的参考架构，并对IBM、甲骨文、SAP、SAS、Informatica、微软和其他公司的大数据平台进行了介绍。

本书不讨论大数据分析。与此相反，本书聚焦于治理大数据的重要性。大数据治理可能不像大数据分析那样引人入胜，但是，大数据治理是从大数据平台中获得最大价值的关键推动力量。Gartner的最近研究表明，组织希望利用的大部分数据处于其控制力之外，更少结构化，并且比以往处理的交易数据更难以理解，因此，大数据将给信息治理带来巨大挑战。

原始形态的大数据一般不能被验证和确认，除非出于证券欺诈等特定目的而对其进行处理。到那时，为时已晚。由于数据集处于不断增长和变化中，大数据分析并不总能被复制。在数据中，副本、遗漏和大体上的不完整性，是可以预见到的。由于大数据并不总是像传统数据那样可信，其治理更加重要。

最后要说明的一点是：我采用“big data”（小写字母）而不是“Big Data”（首字母大写）指代“大数据”，目的是强调将新数据类型集成到既有治理计划和技术设施的重要性。

本书的读者对象

对负责大数据分析和治理的任何从业者来说，本书都包含了有价值的内容。本书面向业务层面的读者，偏重于非技术性知识。以下是可能对本书感兴趣的有代表性的部分读者：

首席信息官
信息治理总监/管理者
数据治理官
信息技术总监/管理者
信息管理总监/管理者
商业智能和数据仓库总监/管理者
商业智能能力中心总监/管理者
社交媒体分析总监/管理者

数据主管
数据发现官
数据科学家
数据分析总监/副总裁
主数据管理总监
首席数据官
首席信息安全官
首席隐私官
首席市场营销官
首席风险官
首席财务官
首席保险清算师
首席供应链官
首席采购官
首席医疗信息官
首席(社交)聆听官

大数据开启了发现商业洞察的诸多机会。但是,大数据治理极其重要,特别是在涉及数据质量、元数据、隐私和与主数据的整合时。我希望,阅读此书后,您和我一样,对大数据治理感到振奋。

桑尼尔·索雷斯