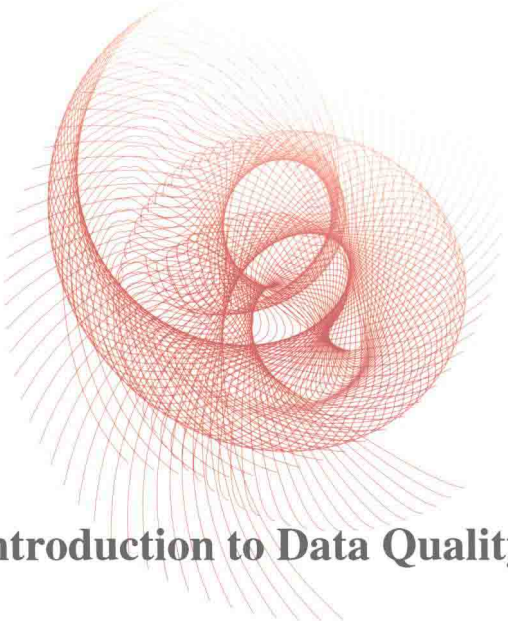
大数据治理与应用丛书



Introduction to Data Quality

数据质量导论

曹建军 刁兴春 著

国防工业出版社
National Defense Industry Press

大数据治理与应用丛书

数据质量导论

Introduction to Data Quality

曹建军　刁兴春　著

国防工业出版社

·北京·

内容简介

本书结合国内信息环境特点，系统分析了数据质量以及数据全生命周期质量管理的内涵，构建了数据质量研究和数据清洗系统框架，并引入了数据质量管理的并行发展模式；深入研究了实体分辨、不完整数据、不一致数据三类实例层数据质量问题的数据清洗技术，提出了若干数据清洗技术方法；归纳了数据质量工具的发展概况，提出了两种数据质量工具设计方法；总结提出了大数据质量面临的十大挑战，构建了适用于国内信息环境特点的数据治理系统框架。

本书内容由浅入深，系统性强，易读性和可操作性强，既可作为数据质量领域的入门和进阶用书，又可作为数据资源建设与利用、信息技术等相关学科的教学参考用书。

图书在版编目（CIP）数据

数据质量导论/曹建军，刁兴春著. —北京：国防工业出版社，2017.10

ISBN 978-7-118-11405-8

Ⅰ.①数… Ⅱ.①曹… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2017)第 243903 号

※

国防工业出版社出版发行

（北京市海淀区紫竹院南路 23 号 邮政编码 100048）

三河市众誉天成印务有限公司印刷

新华书店经售

*

开本 710×1000 1/16 **印张** 20½ **字数** 402 千字

2017 年 10 月第 1 版第 1 次印刷 **印数** 1—2000 册 **定价** 79.00 元

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

序　一

我初次涉足“数据质量”和主数据管理(Master Data Management, MDM)是1995年,当时正致力研发算法与应用软件,以实现来自不同组织用户记录的近似匹配。那时,数据质量行业只有一批出售数据标准和清洗工具的公司,以及一些主张过程控制而忽视应用技术的咨询师。

当时面临的主要挑战是如何在两者之间寻求平衡——尽管能够使用工具“清洗”数据,但如何才能明确界定“高质量数据”的真正含义呢? 因此,我意识到需要通过探究如何使用数据剖析与清洗工具,以及相关技术促进数据质量规约,进而基于明确的数据质量期望指标来增强合规性以实现两种方法的有机结合。这也是我2001年专著《企业管理知识——数据质量方法》的主题。

如果向前看20年,就可以理解数据质量社区是如何出现的。有定期会议供从业者讨论众多不同业务场景下的各类数据质量技术;建立了监督数据方针规范的工作流程,企业使用这些方针创建、使用和保护数据资产;随着越来越多的组织任命负责数据资产管理的首席数据官(Chief Data Officer),数据治理(Data Governance)和数据统管(Data Stewardship)得到快速普及;有许多致力推动数据质量最佳实践的国际组织,我及其他优秀数据质量作者的书正被译成不同语言,促进信息利用服务的全面改善。

本书是对数据质量知识体系的加强。特别值得一提的是,本书既回顾了数据质量工具与技术相关的传统知识,又在诸如不完整数据检测,基于马尔可夫模型的数据填补,以及改进数据清洗的机器学习算法方面提出了新思想。另外,作者还讨论了大数据质量的时代特征,这些特征很快会成为我们面临的重大挑战。

很荣幸受邀为本书作序,希望作者的工作对世界范围的数据质量从业者产生深远影响。

David Loshin
Website: www. knowledge – integrity. com
Email: loshin@ knowledge – integrity. com
LinkedIn: David Loshin
Twitter: @ davidloshin
2017年2月13日

David Loshin:被誉为信息管理行业的思想引领者,著有"*The Practitioner's Guide to Data Quality Improvement*"(2011)、"*Master Data Management*"(2008)、"*Business Intelligence—The Savvy Manager's Guide*"(2003)等,开设了多门数据管理最佳实践课程,并多次做相关专题报告。

序　二

智能设备、云计算和物联网，这些技术是如此让人钟爱！它们就像闪闪发光的金币一样吸引眼球。然而，事实上这些技术的价值源自对所创建、存储及分发数据与信息的利用。立足数据与信息的新应用，各类组织正在谋求更大的成功。为此，它们必须知道所拥有的数据是正确的并且是可信的。那么，是谁在关注这些炫目科技中的数据与信息呢？

众所周知，所有组织都有诸如财务（Finance）、人力资源（Human Resources）管理系统，那么是否有数据与信息管理系统呢？在过去，这样的系统称为信息技术（Information Technology）或直接称作 IT。IT 作为一项重要技术，只是故事的一部分。我们需要一个和数据与信息自身相关的管理系统，不仅包括技术，而且涉及人员、组织、流程、标准方方面面。以数据质量为中心是数据管理系统的基本特征。不管是以营利为目的的企业，还是政府、教育、医疗、慈善，及其他非营利机构，任何组织的成功都离不开信息。用户、客户、雇员和商业伙伴使用数据进行决策、处理事务、相互协作，并提供产品与服务。他们规划、创建、更新、转换、存档和删除数据时，也会影响数据质量。自动化系统基于高质量数据顺利完成交易，无需人为干预。自然，以数据质量为中心是为了确保上述所有组织和人员满怀信心地使用数据。再次重申，数据质量对信息利用和期望结果至关重要。

高质量的数据与信息是适用的（Fit for Purpose）。这意味着能够找到（能够获取并访问）所需要的数据或信息，当需要时是可用的（及时而不滞后），包含所需要的全部信息（无缺失），是安全的（未经授权不得访问与操作），是可理解的（能够对其解释），并且是正确的（是对真实世界的准确反映）。这样才能确保所获取的信息是可信的，并且满怀信心地利用它们。为了拥有高质量数据与信息，需要综合考虑多种因素。只有在数据全生命周期过程中优化协调大量流程、技术、人员、组织以及各种各样的数据，才能提供组织所需的高质量可信信息。尽管这一过程错综复杂，但从数据质量的视角，能够综合考虑多种因素，进而找出全新解决方案。

能够从事数据质量职业是令人兴奋的。作为第一本译成中文的数据质量专著——《数据质量工程实践——获取高质量数据和可信信息的十大步骤》的作者，很荣幸受邀为中国的第一本数据质量专著作序。我最近一次去中国是 2016 年 11 月，当时作为特邀国际演讲专家参加了“首届中国数据标准化及治理大会”。看到中国对相关领域表现出的热情以及所做的工作非常振奋。期间，我还与大数据专

业的 MBA、管理数据的技术用户、致力于增强数据质量的公司，以及乐于分享新闻的媒体进行了交互。

作为一个职业和专业领域，数据质量还不如计算机科学、工程学、法学、会计学这些职业领域成熟。任何新兴职业在成为主流并为众人所知之前，都要经历一个漫长的发展过程。对于数据质量研究者来说，这是一个激动人心的时代，因为我们已经从前人的工作中获得了许多经验。尽管在很多地方取得了重大突破，但仍有广阔的创新和发展空间。于是该书应运而生。如前所述，这是中国第一本数据质量专著；它建立在对数据质量的认知之上，并会为数据质量领域的发展作出贡献；它还讨论了中国信息环境的独特性。本书为技术研发等数据质量管理工作提供专业知识。

数据质量领域历经近 24 年仍在不断发展，而我自诩该领域第二代先锋。很庆幸自己曾经与几位首先提出数据质量重要性的先驱们并肩工作。我不断向不同领导者、管理者、从业者，以及咨询师请教学习。你是组织中少数数据专家之一吗？你理解数据的重要性吗？你想让你的组织认识到数据的重要性并想得到数据质量管理的资源吗？你是公司的致力于保证高质量数据的第一位首席数据官吗？你想引领数据质量潮流吗？如果是，那么你也是一名数据质量先锋。

成为数据质量先锋并非易事，要不畏挑战，勇往直前。要让更多人认识到数据质量的重要性，并提升相关技能。纵然要不断开拓进取，但能成为世界范围的弄潮儿又倍感欣慰。数据质量的理念超越了语言、文化和国家的界限。很高兴所到之处皆能与志同道合的人士并肩战斗，很高兴数据质量原理正在产生积极影响。数据质量在我心中是一股团结的力量。我们因数据质量走到一起，团结协作，和谐发展。我在世界范围内结交了众多同道中人，并得到了国际数据质量社区的高度赞赏。这个世界真小，不论走到哪里，都能学习他人的工作动机及协作精神，对他们表现出的友善我心存感激。这也正是数据质量社区的内在精神！

那些真正理解数据质量工作的人还认识到了“人的因素”对成功的重要性，包括工作过程中的沟通（Communication）、协调（Coordination）、互助（Cooperation）、合作（Collaboration），还包括另外一个 C——勇气（Courage）。尽管在数据质量领域很少提及“勇气”，但是创新、改革、进取的确需要勇气，放弃安逸、提升自我、不断突破也需要勇气。要鼓励公司相关成员像管理人员与财务一样对数据进行管理。

尽管投身数据质量领域多年，但我依然热情不减，坚信这项重要工作切实影响组织发展和人们的日常生活。人生苦短，希望热情和理想激励我们一路前行。希望读者利用从本书中汲取的营养实现组织受益并为社会作出贡献。

感谢信息质量研究组(Information Quality Research Group,IQRG)以及本书的所有贡献者,感谢你们的分享。鉴于本人不谙中文,感谢曹建军博士将本序言译成中文,感谢御数坊(Data Governance Workshop)的刘晨先生对本书的评论。

Danette McGilvray
Website:www. gfalls. com
Email:danette@ gfalls. com
LinkedIn:Danette McGilvray
Twitter:@ Danette_McG
2017 年 1 月 31 日

Danette McGilvray:国际数据标准化、治理与质量领域知名专家,自诩为数据质量领域的第二代先锋,著有"*Executing Data Quality Projects—Ten Steps to Quality Data and Trusted Information*"(2008),御数坊"数据质量十步法"在线课程主讲专家,2012 年企业数据世界(Enterprise Data World,EDW)大会最受欢迎讲师。

前　言

大数据战略进展如火如荼，数据质量问题日益突显。好产品的典型特征是具有较好的自身守恒能力，能够稳定保持用户期望的产品使用价值，较之其他有形产品或软件产品，数据产品的这种能力恰恰较差。同时，数据的价值主要体现在“流通”，而非“存储”，所以，数据质量问题较传统产品质量面临更多挑战。

信息质量研究组（Information Quality Research Group，IQRG）成立于 2008 年，以结合我国信息环境特点系统开展数据质量研究与实践为己任，随着相关工作的深入推进，对国内数据现状及特点的认识也逐渐清晰。

信息质量研究组成立以来，我们陆续出版了译著《数据质量工程实践》、《信息质量》和《数据质量改进实践指南》，后两者受到了装备科技译著出版基金的资助。“御数坊”在介绍第 20 届企业数据世界（Enterprise Data World）大会（加利福尼亚州圣迭戈，2016 年 4 月 17—22 日）时，向关注数据质量的同学推荐了《数据质量工程实践》。三本译著在国内普及数据质量理论与实践体系、提升数据质量认识层次上发挥了积极作用。为了有计划地推出研究成果，立足我国信息环境特点逐步构建数据治理与应用理论技术体系，2016 年上半年，受国防工业出版社之邀，信息质量研究组启动了“大数据治理与应用丛书”出版工作，译著《数据质量改进实践指南》是丛书开卷，本书是此丛书的第二个成员。

本书共分 12 章。第 1 章至第 3 章是本书的总述部分。第 1 章为绪论，引出数据质量问题，介绍了数据质量以及数据全生命周期质量管理的含义，分析了数据质量问题的来源并归纳其研究发展简况；第 2 章分析构建了数据质量研究和数据清洗系统框架，引入了数据质量管理的并行发展模式，构建了数据质量控制层次框架，分析了其实现所涉及的关键问题，在进一步辨析数据清洗概念的基础上，构建了数据清洗的一般性系统框架；第 3 章综述了典型数据清洗技术的发展动态，系统归纳了实体分辨、不完整数据、不一致数据三类实例层数据质量问题的数据清洗技术发展动态。第 4 章至第 10 章是以上三类数据清洗技术的研究成果。第 4 章研究了实体分辨中的数据分块问题，第 5 章研究了实体分辨中的相似度算法，第 6 章研究了基于关系的实体分辨；第 7 章研究了不完整数据的分类与检测，第 8 章研究了不完整数据的估计与填充；第 9 章研究了条件函数依赖挖掘及其优化方法，第 10 章研究了基于规则的不一致数据检测与修复方法。第 11 章研究了数据质量工具的发展概况及设计方法，分别研究了基于表达式树的数据质量工具设计和基于

流程的数据质量工具设计方法。第12章研究了大数据与大数据质量问题，归纳了大数据时代的特征，总结提出了大数据质量面临的十大挑战，构建了适用于我国信息环境特点的数据治理系统框架。

本书由曹建军、刁兴春全面筹划，并负责了第1章至第3章、第12章的研究撰写工作，指导参与了其他各章的研究撰写；谭明超、周星负责了第4章至第6章的研究撰写；郑奇斌、谭明超负责了第7章的研究撰写；郑奇斌、谭明超、陈爽负责了第8章的研究撰写；周金陵负责了第9章的研究撰写；高科负责了第10章的研究撰写；江春、翁年凤、高科负责了第11章的研究撰写。许永平参与了第9章、第10章的编辑整理，刘艺、冯钦参与了部分章节的编辑整理。江春、彭琮负责了全书的文字编辑润色；尚玉玲、刘艺、李红梅、张磊、冯钦负责了全书的规范性审核与修改工作。

感谢两位国际著名数据质量领域专家 David Loshin、Danette McGilvray 为本书拨冗作序，感谢二位对信息质量研究组相关工作的支持与肯定。

本书是作者在数据质量领域研究成果的梳理小结，试图传递三个信息：一是国内数据质量领域的发展模式要紧贴国内信息环境特点与数据应用实际；二是数据质量控制技术研究要紧贴国际前沿；三是数据质量管理实践既要重视具体的数据质量工具又要重视体系化的数据治理平台。通过阅读本书，甚望读者能够在概念层面对数据质量有全面客观的认识，在技术层面能够管中窥豹，在实践层面获得可用参考。

本书可作为数据资源建设与利用、信息技术等领域科研和工程技术人员进行数据质量研究与实践的入门指导及工程参考用书。

在本书内容的研究整理过程中，广泛参考了国内外相关成果，并与多家兄弟科研团队及专家同仁进行有益的经常化交流研讨，在此一并致以诚挚的谢意。

受水平所限，书中若有错误和不妥之处，恳请广大读者批评指正，并欢迎与作者直接交流。

作者

2016年10月

目　录

Contents