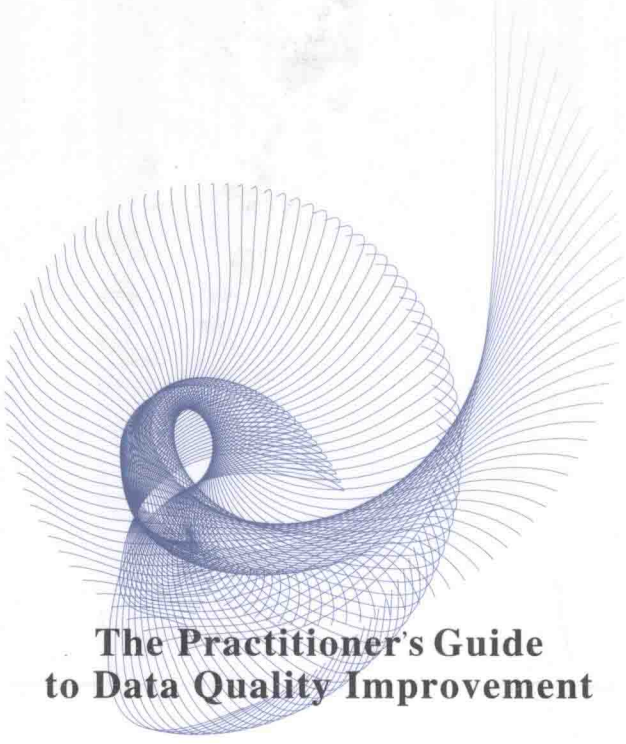
ELSEVIER
爱思唯尔

装备科技译著出版基金

大数据治理与应用丛书



The Practitioner's Guide
to Data Quality Improvement

数据质量改进实践指南

【美】David Loshin 著

曹建军 江 春 等译　刁兴春 审校

國防工業出版社
National Defense Industry Press

数据质量改进实践指南

The Practitioner's Guide to Data Quality Improvement

[美] David Loshin 著
曹建军 江 春 等译
刁兴春 审校

国防工业出版社
·北京·

著作权合同登记　图字：军－2013－193 号

图书在版编目（CIP）数据

数据质量改进实践指南/（美）洛申（Loshin，D.）著；曹建军等译. —北京：国防工业出版社，2016.8

书名原文：The Practitioner's Guide to Data Quality Improvement

ISBN 978-7-118-10813-2

Ⅰ.①数… Ⅱ.①洛… ②曹… Ⅲ.①数据管理－质量管理－研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2016）第 047204 号

数据质量改进实践指南

The Practitioner's Guide to Data Quality Improvement

※

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** 388 千字

2016 年 8 月第 1 版第 1 次印刷　**印数** 1—2000 册　**定价** 99.00 元

（本书如有印装错误，我社负责调换）

国防书店：（010）88540777　　发行邮购：（010）88540776

发行传真：（010）88540755　　发行业务：（010）88540717

《数据质量改进实践指南》翻译组

组织与统稿　曹建军　江　春

审　　校　刁兴春

翻　　译　（按姓氏拼音排序）

陈　爽　丁晨路　高　科　何　焱

蒋国权　刘　艺　刘志强　马金钢

彭　琮　尚玉玲　孙　晓　王丽君

王艳霞　魏祥麟　翁年凤　邢继娟

许永平　姚俊楠　俞　赟　袁　震

张　宾　张　静　张潇毅　郑奇斌

周金陵

译 者 序

大数据战略如火如荼，数据质量问题日益凸显。好产品的典型特征是具有较好的稳定性，而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新挑战。

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，对国内数据质量现状及特点的认识也逐渐清晰。

当前，国内数据质量研究与实践面临以下困境：一是成熟的理论技术没有很好地本地化，理论研究以跟踪国外前沿动态为主，与国内实际信息环境脱节，形不成从研究到应用再到研究的良性可持续发展的闭环；二是成功的应用案例大多集中在具体行业中（如电信行业），这些案例以实现数据增值为基本驱动，尽管解决数据质量问题是实现项目目标的前提，而且占据了较大比重的工作量，但相关工作往往隐藏在业务流程之下并与业务流程紧密捆绑，得不到良好体现且通用性差；三是许多数据资源建设与利用项目，重心仍局限于数据获取与共享，沿用的是信息系统的建设思路，加之数据的应用需求不明确，对数据质量的重视程度和资源投入还远远不够。以上背景下，将国外优秀的数据质量著作译成中文，成体系地介绍数据质量理论方法与实践经验，为国内数据质量工作提供借鉴，无疑是一项意义深远的基础性工作。

《数据质量改进实践指南》以“从始至终控制信息生产过程的质量，确保在产生实质影响前尽早识别、排序和修复数据缺陷”的核心理念贯穿始终，指导管理者和实践者以相同的方式交流、获得支持、规划和建立数据质量计划，给业务影响分析、数据质量指标定义、检查和监测、修复以及数据质量工具的使用提供了模板和流程。本书系统阐明了如何制定和执行数据质量计划，如何有效使用各种数据质量工具，对业务层和IT层数据质量涉及的人员、过程、技术等进行了全面分析，特色鲜明。

《数据质量改进实践指南》是继《数据质量工程实践》（电子工业出版社，2010）、《信息质量》（国防工业出版社，2013）之后信息质量研究组完成的第三本译著。《数据质量工程实践》是第一本面向工程实践的数据质量专著，《信息质量》是数据质量领域的第一本基础理论著作，与前二者相比，《数据质量改进实践指

南》中数据质量改进的方法论更为完善系统，操作性更强。

近年来，信息质量研究组得到了国家自然科学基金项目（No. 61371196）、中国博士后科学基金特别资助项目（No. 201003797）、中国博士后科学基金面上项目（No. 20090461425）、江苏省博士后科研资助计划项目（No. 0901014B）、解放军理工大学预研基金项目（No. 20110604，41150301）的支持。

在本书翻译过程中，译者力求忠实原著，并保留原著风格；但受水平所限，书中若有错误和不妥之处，恳请广大读者批评指正，并欢迎与译者直接交流 E - mail：xinxizhiliang@163. com。

信息质量研究组（IQRG）

2016 年 1 月

序　言

我的工作是和许许多多的组织一起努力最小化乃至完全消除数据质量问题。我们提出了很多大力处理这些问题的手段：从建立业务案例的方法，一直到数据质量问题的解决策略。组织环境的多样性是惊人的，不同规模、行业和结构的组织都有由于各种不同方式、功能和流程所产生的各种棘手的数据质量问题。

但有一件事对所有组织中的研究人员而言是共同的，即期望实用的、易于实施的和立竿见影的方法来解决问题。事实上，问“如何做?”，如何开始，如何划定范围，如何聘用合适人员，最重要的是，如何从他们的工作中产生商业价值。他们需要实用的和有效的成熟方法。

我也读过许多以数据质量改进为主题的书。许多人抽象地叙述该学科。实际上许多讨论的都是高深理论乃至学术问题。如果是了解理想情况下如何能使数据质量最优化的高层原则和基本原理，这是适合的。但是，我们并非是生活在完美世界中，因此，我们需要能在组织中直接实行有效行动的实用方法，这一行动将以好的方式开始数据质量改进方案，并迅速产生商业价值。

本书中，David Loshin 做了大力传播这种实用方法的工作。书中讨论了一个完整数据质量改进计划的基本模块，重点关注组织方法、基本流程和技术角色。我认为，这正是组织需要采用的正确的优先级次序。首先，选用合适的人员并确保他们具有合适的技能（组织）；其次，让这些人员执行规定的可重复的活动（流程）；最后，在合理的时间和地点给予他们工具支持（技术）。在诸如数据治理（Data Governance）、主数据管理（Master Data Management）以及更广泛的企业信息管理的关键相关行动中，数据质量改进的作用日益重要。

作为一本实践手册，书中使用了行业专业人士和 IT 专业人士都易于理解的术语，并且使用了读者能够对应到自己情况的具体实例和建议。关键是，数据质量改进不仅涉及 IT 领域，还需要强有力的行业领导和约定。制定新规定的组织需要这些基本原则指导其前期工作。在处理数据质量问题方面具有广泛经验的组织，当它们调整和拓宽工作范围时，也会从这些原则中受益。

特别是在数据日益被视为最关键资产的当今，引入的数据质量改进是一个极好的方法。成功仍然很难，它需要热情、专注和一如既往的坚持，还需要对基本原则

的全面掌握以及对有效技术的充分利用。应用书中的观点，您将踏上掌握最重要武器的正确之路。祝您在持续数据质量改进工作中取得成功！

Ted Friedman
副总裁兼著名分析师
Gartner Inc.

前　言

在我从事数据质量工作的15年里，注意到数据质量管理的方式发生了巨大变化。数据质量正从以简单清理名称和地址为主，迅速转变成更加鲜明反映企业生产管理环境的行业。人们越来越认识到，高质量数据能有效促进企业目标的实现，这就意味着需要制定一项企业数据质量计划。

但为了建立这一计划，不仅需要名称和地址清理工具，而且更需要维护可度量高质量数据的管理与治理框架。本书旨在为制定企业数据质量计划提供基本原则，并为在组织范围内建立可操作数据质量控制措施的管理者和实践者提供指导，其要点如下：

（1）为制定数据质量计划建立业务案例的能力。

（2）数据质量成熟度水平的评价。

（3）评价数据质量和确定与实现业务目标相关的度量指标的准则和技术。

（4）基于这些度量指标测量、报告，以及采取行动的技术。

（5）为数据质量改进开发数据质量工具和技术所用到的方针和流程。

数据质量知识传递

为了传递我们公司过去多年积累的知识，汇编此书以帮助那些在诸如数据质量、数据治理、主数据管理、客户数据集成（Customer Data Integration）领域承担任务角色的个人，以及在这些类型活动中取得成功的许多其他数据管理角色，这些活动包括：

（1）为制定数据质量计划建立业务案例。

（2）制定企业数据管理、数据治理和数据统管（Data Stewardship）策略。

（3）制定实施方案。

（4）分配角色和职责。

（5）制定数据质量评价、数据质量度量指标，以及持续监测与报告的方针与规程。

（6）使用数据质量工具和技术。

（7）创建数据标准管理计划。

（8）监测数据质量和绩效趋势。

物尽其用

更重要的是：在《企业知识管理——数据质量方法（*Enterprise Knowledge Man-*

agement — The Data Quality Approach)》出版后的10年里，我一直致力于帮助组织策划改进信息质量。失败的数据管理活动的经历使我结束了原来的职业道路，开始了证明存在使数据获得高质量更佳途径的咨询工作。我成立了旨在帮助组织形成成功的信息质量、数据治理和主数据管理计划的 Knowledge Integrity 公司，以使我的工作与其他咨询公司相区别。针对经营方式还制定了一些重要业务规则：

（1）我们的任务是担负并普及企业数据质量改进方法。与相关的技术、方法和流程专利保护相反，我们公开出版我们的思想，以便使花费时间和精力关注我们所倡导思想的人们受益。

（2）我们鼓励客户在他们的成功模式中采纳我们的方法。走进组织，告诉在工作上已取得成功的人们需要放弃现有的并且改变他们工作方式的每一方面是一项挑战（或许在某种程度上是无礼的）。我们相信每个组织都有自己的成功方法，而我们的工作是建立一种将基于绩效的信息质量管理整合进现在组织成功结构的途径。

（3）我们并不将自身视为固定不变的。我们相信信息质量管理是组织应该掌握的核心能力，每项约定的努力目标是建立计划的基本方面，把技术转化成内部资源，然后继续下去。我常说，我们的工作如果做得好，将必然超出合同约定的工作内容。

（4）我们不“出售产品”，而是忙于解决客户问题。我们更关心的是确保客户的核心问题得到解决，如果我们的方法与组织的方法相适应，这说明我们找到了工作的最佳途径。我常说，当客户对我们的观点满意时就成功了。

（5）有效沟通是管理变更的关键。阐明良好的信息管理技术如何提高组织的效能和绩效是吸引客户并确保得到他们支持和资助的第一步。我们会将每个合约的部分收入用于建立一个附有企业内和跨企业接受的相关信息的业务案例。

以这些规则为基础，我们的第一份工作是将面向语义和规则的数据质量管理组织整理成书，于是在2001年由 Morgan Kaufmann 出版了《企业知识管理——数据质量方法》。许多读者告诉我，此书在他们的数据质量管理计划改进行动中至关重要，这几年来，针对基于规则的数据质量监测所提出的新技术思想已整合进所有主要数据质量供应商产品中。

随后的几年里我们为纽约大学（New York University）制定了数据质量研究生水平课程，并为数据仓库研究所（Data Warehousing Institute，www. tdwi. org）制定了多门日间课程；出席了许多数据管理协会（the Data Management Association，DAMA）的会议讨论和理事会议；在 Robert Seiner 的数据管理通信（Data Administration Newsletter，www. tdan. com）上开设了专栏，在数据管理评论（DM Review，www. information-management. com）上开设了每月专栏，在 BetterManagement（www. bettermanagement. com）上提供了可下载数据质量资源；在 BeyeNETWORK

（www. b-eye-network. com）管理一个专家通告和每月通信。

我们经常要为跨领域的供应商提供分析报告，并考虑许多领域数据管理的领导，我们曾为联邦州以及其他全球政府机构的公共部门提供咨询。我们深入了解了一些行业的数据质量管理和数据治理方法，包括财务服务业（Financial Services）、卫生保健业（Health Care）、制造业（Manufacturing）、石油和矿业服务业（Oil and Mining Services）、保险业（Insurance）和社会服务业（Social Services）。

自从公司成立以来，对信息质量管理价值的认识，已成为高管需要考虑的最重要的主题之一。实际上，这种情况已经出现在企业数据开发利用中，例如，企业资源规划（Enterprise Resource Planning，ERP）、供应链管理（Supply Chain Management，SCM）、客户关系管理（Customer Relationship Management，CRM），其中存在对不同的经营理念的关键实例进行高质量描述的统一视图需求，增强的管理监督以及增加的信息交换需求。企业绩效管理和面向服务体系结构的价值正在驱使人们对与公共信息资产的可访问性（Accessibility）、一致性（Consistency）、流通时间（Currency）、新鲜性（Freshness）和可用性（Usability）相关的面向绩效企业数据产生更多的关注。

本书简介

本书包括三个部分。

第1部分（第1章～第5章）关注数据质量的组织特征：了解低劣数据质量的影响、数据质量计划的特征、组织准备状态和成熟度、企业数据质量与其他企业行动的关系，建立业务案例、蓝图和路线图，以及将数据质量改进作为形成竞争优势的重要因素。包括以下各章：

第1章低劣数据质量的业务影响；

第2章组织的数据质量计划；

第3章数据质量成熟度；

第4章企业行动整合；

第5章建立业务案例和数据质量路线图。

第2部分（第6章～第13章）讨论数据质量计划核心流程的实现：度量指标和数据质量绩效改进、数据治理、数据质量维度定义、数据需求分析、数据标准、元数据管理、数据质量评价、修复以及数据质量服务水平协议。包括以下各章：

第6章度量指标和绩效改进；

第7章数据治理；

第8章数据质量维度；

第9章数据需求分析；

第10章元数据与数据标准；

第11章数据质量评价；

第 12 章修复和改进计划；

第 13 章数据质量服务水平协议。

第 3 部分（第 14 章 ~ 第 19 章）讨论了用于支持第 2 部分描述的数据质量流程的各类工具、技能、算法以及其他技术。包括数据剖析、分解和标准化、身份分辨、审核和监测、数据增强，以及主数据管理。包括以下各章：

第 14 章数据剖析；

第 15 章解析和标准化；

第 16 章实体身份分辨；

第 17 章检查、监测、审核和跟踪；

第 18 章数据增强；

第 19 章主数据管理和数据质量。

最后一章第 20 章回顾了全书所讨论的概念，对数据质量实践者而言可用作速查指南。

致　谢

本书是我多年来从事主数据管理项目的经验总结，内容涵盖了主数据管理工具、技术、流程、人员等多个方面。许多人都对本书的编写做出了重要贡献，借此机会对他们的支持致以由衷的感谢！

首先要感谢我的妻子 Jill 在本书写作过程中给予的信任和鼓励，也要感谢我的孩子们 Kra，Jonah，Brianna，Gabriella，Emma，他们也给予我非常多的帮助。

感谢 Richard Ordowich，Knowledge Integrity 公司主要顾问之一，为促进建立数据质量管理计划给予了很多宝贵的建议，并在多年间始终支持我的工作。

我曾汇编了一些厂家如 DataFlux，Informatica，IBM，Initiate Systems，Microsoft，以及 Pitney Bowes Business Insight 等在数据质量和主数据管理领域的资料，本书的很多重要灵感来源于这一时期的工作。除此之外，我通过专家通道在 www. b-eye-network. com 网站和 Wilshire Conferences，DebTech International，The Data Warehousing Institute，MDM – DQ 大学举办的会议以及供应商组织的网络研讨会和在线活动中获得了非常宝贵的资料。

Dataflux 公司的 Tony Fisher，Katie Fabiszak，Daniel Teachey，James Goodfellow 以及 Dan Soceanu 为本书的撰写提供了重要的信息。

Gartner 公司的 Ted Friedman 总是不断反馈意见，并提供了数据质量的行业现状，以及管理流程和最佳实践如何弥补工具的不足。

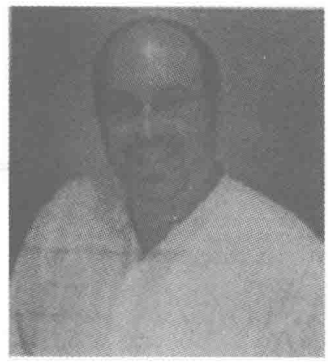
Initiate Systems 公司的首席技术官 Marty Moseley 在数据治理、主数据管理方面提出了独特见解。

Business Intelligence Network（www. b-eye-network. com）公司的 Ron Powell，Shawn Rogers，Jean Schauer 以及 Mary Jo Nott（仅列举少数人）等同事给我提供了编写本书所用资料的平台。

特别感谢 Wilshire Conferences 的 Tony Shaw，每年 DAMA/ Meta-Data and Data Governance 会议的演讲者，The Data Warehousing Institute 以及 DebTech International 的 Davida Berger，使我得以开设一门介绍本书概念的技术课程。同时也非常感谢我的客户们，如 Greg Wibben 向我提供环境，以证明本书所述流程的价值。

最后，感谢 Elsevier 出版社的编辑对本书的支持，他们是 Diane Cerra，Greg Chalson 和 Rick Adms。

作 者 简 介



David Loshin，美国 Knowledge Integrity 公司董事长，该公司专门从事数据管理咨询业务。David 在高性能计算和数据管理方面已出版数本专著，包括“*Master Data Management*”（2008）、“*Business Intelligence—The Savvy Manager's Guide*”（2003）等，并开创了多门数据管理最佳实践诸方面的课程和专题报告，被誉为信息管理行业的思想引领者。

目　录