高新科技译丛

Information Quality

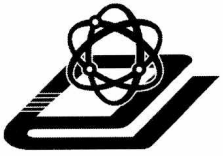
信息质量

【美】 Richard Y. Wang Elizabeth M. Pierce
Stuart E. Madnick Craig W. Fisher 著

曹建军 刁兴春 许永平 译

M.E.Sharpe
Armonk, New York
London, England

国防工业出版社
National Defense Industry Press



装备科技译著出版基金

高新科技译

信息质量

Information Quality

[美] Richard Y. Wang　Elizabeth M. Pierce
Stuart E. Madnick　Craig W. Fisher　著
曹建军　刁兴春　许永平　译

M.E.Sharpe
Armonk, New York
London, England

国防工业出版社

著作权合同登记　图字:军-2012-013号

图书在版编目(CIP)数据

信息质量/(美)理查德(Richard,Y. W.)等著;曹建军,习兴春,许永平译. —北京:国防工业出版社,2013.3
(高新科技译丛)
书名原文:Information Quality
ISBN 978-7-118-08274-6

Ⅰ.①信... Ⅱ.①理...②曹...③习...④许... Ⅲ.①信息资源—信息管理 Ⅳ.①G203

中国版本图书馆 CIP 数据核字(2012)第 214956 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)
北京嘉恒彩色印刷责任有限公司
新华书店经售
*
开本 710×960　1/16　印张 18½　字数 326 千字
2013 年 3 月第 1 版第 1 次印刷　印数 1—3000 册　定价 69.00 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

致中国读者

过去若干年，中国的信息技术发展速度惊人。作为这一发展的组成部分，信息质量领域也在中国快速成长。刁兴春研究员和曹建军博士于2008年成立了信息质量研究组（Information Quality Research Group，IQRG），近年来已发表相关论文30余篇；2010年和2011年，连续在信息质量国际会议（International Conference on Information Quality，ICIQ）上发表论文，并参加了第16届信息质量国际会议（2011 ICIQ），出版了 *Executing Data Quality Projects*（原书由 Morgan Kaufmann 出版社出版），该译著是第一本中文信息质量书籍。另外，2011年，西安交通大学也正式成立了信息质量研究小组，并参加了麻省理工学院信息质量行业研讨会（Information Quality Industry Symposium，IQIS）；西安交通大学黄伟博士正在组织成立中国信息质量协会（Society of Chinese Information Quality，CNIQ）；黄伟博士的研究团队已将 *Journey to Data Quality* 一书（原书由麻省理工学院出社出版）译成中文，并计划用作信息质量培训课程。

过去的30余年，信息质量领域取得了有目共睹的进展。1988年，麻省理工学院的 Stuart Madnick 教授和 Richard Wang 教授启动了全面数据质量管理（Total Data Quality Management，TDQM）计划，开始了研究出版的历程。信息质量是一个多学科交叉应用领域，需要实践人员与研究人员之间的交互与协作。为了满足以上要求，1996年，麻省理工学院 TDQM 计划组织了首届信息质量国际会议，以此鼓励研究人员和实践人员交换知识。麻省理工学院信息质量计划旨在对信息质量进行全方位研究，例如，将信息作为产品管理、开发信息产品图、信息质量实践在组织中的应用。2007年，为了进一步促进实践人员、供应商、学者之间的交互协作，麻省理工学院信息质量计划发起了首届信息质量行业研讨会。会上除报告和专题研讨外，还包括供应商报告、产品公告和咨询方法等，与信息质量国际会议互为补充。2012年，麻省理工学院将主办第2届首席数据官论坛和第6届麻省理工学院信息质量行业研讨会，并将二者合并称为首席数据官与信息质量研讨会（Data Chief Officer and

Informatin Quality Symposium, CDOIQ)。

前期的投入已初见成效。其中最令人振奋的进展之一就是 2006 年“ACM Journal of Data and Information Quality”的创办,将信息质量确立为信息技术研究的重要领域。另一个显著成效就是信息质量教育的发展,尤其是研究生教育:由 John Talburt 教授和 Elizabeth Pierce 教授领导的阿肯色大学小石城分校的第一个此类项目——信息质量理学硕士(Master of Science in Information Quality)不断发展壮大;继 2006 年硕士计划后,2007 年阿肯色大学小石城分校又确立了信息质量博士计划。

多年来,麻省理工学院举办了大量的会议和专题研讨并组织社团,如 SIGMOD 的信息系统中的信息质量专题研讨(SIGMOD Workshop on Information Quality in Information Systems)、CAiSE 的数据和信息质量专题研讨(CAiSE Workshop on Data and Information Quality),以及德国信息和数据质量协会定期组织的会议与专题研讨(German Society for Information and Data Quality)。目前,研究人员和实践人员已将信息质量从作为解决包括定义、测量、分析和改进信息质量问题的领域,拓展到研发提高信息质量的工具、方法和流程方面。因此,目前有很多可用的信息质量资源供读者使用。行业中,也有许多供应商和实践者在积极推动信息质量发展,如 Acxiom、A. I. D.(法)、Deloitte Consulting、EDS、FAST、Firstlogic、FUZZY!、Informatik AG(德)、IBM、Informatica、SAS 和 Serasa S. A.(巴西)等。作为一个团体,我们为已取得的成绩感到自豪!

感谢曹建军博士、刁兴春研究员、许永平博士等将 *Information Quality* 一书译成中文,此举有助于推动信息质量领域的发展。通过该项工作,中国的学者和实践者将得到信息质量方面的前沿知识。最后,希望中国能够成为信息质量领域的领导者!

Richard Y. Wang
美国马萨诸塞州剑桥市
麻省理工学院
rwang@ mit. edu
http://mitiq. mit. edu

译者序

为了系统开展信息质量(数据质量)研究与实践,我们于2008年成立了信息质量研究组(Information Quality Research Group,IQRG)。“数据”是信息时代的标志性产品,会像机械产品、电气产品、电子产品、软件产品一样给军事、经济、生产、生活带来巨大变化和深远影响。信息质量伴生于数据,但其影响远远超出数据范畴,并且更多时候是隐蔽的、潜在的和不确定的。

我们深切体会到国内信息环境与美国不同且更加复杂,有些问题还涉及到政策机制层面,现有的信息质量方法与技术并不完全适用于国内环境。然而,在国内企业级数据集成业务需求尚不明确,对信息质量认识还未达成完全一致,相关研究与实践工作刚刚起步的现阶段,将国外优秀的信息质量著作译成中文,成体系地引入信息质量理论方法与实践经验,无疑是一项意义深远的基础性工作。

《信息质量》是信息质量领域的第一本基础理论著作,汇集了Richard Y. Wang、Elizabeth M. Pierce、Stuart E. Madnick、Thomas C. Redman、Craig W. Fisher等一批最早开展信息质量研究与实践并至今活跃在领域内的学者们的智慧结晶,他们的观点代表了当前信息质量领域的发展方向。《信息质量》将与另一本面向实践的译著《数据质量工程实践》(2010年11月由电子工业出版社出版)互为补充,共同为国内信息质量研究与实践服务。

原著第一作者Richard Y. Wang教授首次使用了“信息质量”这一术语,是麻省理工学院全面数据质量管理(Total Data Quality Management,TDQM)计划和信息质量国际会议(International Conference on Information Quality,ICIQ)的发起人之一,于2009年从麻省理工学院调往美国白宫负责美军数据质量工作。Richard Y. Wang教授对该书的中译版非常关心,并专门为中译版拨冗作序,我们在此向Richard Y. Wang教授表示诚挚谢意。

本书由曹建军、刁兴春、许永平、张健美、朱俊、江春、彭琮译校,参加翻译

工作的还有谭明超、温俊、张潇毅、邓波、严浩、王俊、邹攀红、蒋国权、袁震、翁年凤。

在此，我们还要感谢在信息质量工作中付出努力以及提供过帮助的其他人员和同事：崔之祜、杜鹢、吴建明、汪挺、瞿雷、朱爱平、丁鲲、王芳潇、黄宇、张慧、李凯齐、李戈、陈爽、刘艺。

近年来，信息质量研究组的工作得到了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项目（No. 20090461425）、江苏省博士后科研资助计划项目（No. 0901014B）的支持。

在本书翻译过程中，我们力求忠实原著，并保留原著风格；但受译者水平所限，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。

信息质量研究组（IQRG）
E - mail：xinxizhiliang@163.com
2012 年 6 月

致　谢

本书编辑在此向为本书问世而付出了大量时间和精力的人们表示感谢。首先感谢评阅人对本书的仔细推敲，正是他们在百忙之中抽时间对本书进行了认真阅读和审查。他们是 Donald Ballou、Paul Bowen、InduShobha、Chengular – Smith、WooYoung Chung、Ronald Coleman、Bruce Davidson、Adenekan Dedeke、Frank Dravis、Martin Eppler、James Funk、Tor Guimaraes、Markus Helfert、Yatin Karpe、Barbara Klein、Rita Kovac、Eitel Lauria、Yang W. Lee、Liping Liu、Jennifer Long、Anne Matheus、Felix Naumann、Pamela Neely、Jack Olson、Jeff Parsons、Leo Pipino、Frank Ponzio、Tom Redman、Marc Rittberger、G. Shankaranarayanan、Keng Siau、John Erickson、Diane Strong、Giri Kumar Tayi、Stephen Tu 和 Eitel Von Maur。还要对 M. E. Sharpe 公司的联系人 Elizabeth Granda 表示感谢，她为这本论文集的最终汇编成册提供了帮助和指导。最后，还要感谢 Vladimir Zwass，在他的领导下才使丛书的出版成为可能。

目　录

第四部分　信息质量中的组织问题

第五部分　信息质量的教育及能力建设

第1章 引 言

Elizabeth M. Pierce

摘要:数据和信息质量是管理信息系统领域中重要但仍不成熟的一个研究方向。本引言探讨了组织追求更高数据和信息质量的动机。这一追求因数据和信息质量的定义、测量、分析和提高面临重重困难而充满挑战。为了寻求帮助应对这些挑战,相关组织转向一门正在发展中的包括本书在内的数据和信息质量研究体系,本书包括出自该领域内重要研究者和实践者的14篇新近撰写的富有创新思想的论文。

"水呵水,到处都是水,
船上的甲板却在干涸;
水呵水,到处都是水,
却没有一滴能解我焦渴。"
——摘自《古舟子咏》(The Rime of the Ancient Mariner),
Samuel Taylor Coleridge

1 提高数据和信息质量的动机

就像漂泊于大海却几乎要渴死的柯勒律治老水手的悲哀一样,许多组织发现自己虽然被数据包围着,但却没有多少数据能够真正满足他们的信息需求。今天我们疲于应对所拥有的巨量信息,这些信息表现为多种形式:记录、指南、设计图、蓝图、地图、图像、声音、元数据、详细数据、概要数据等,所列出的也只是少数几种。信息可存储于从档案柜到数据库、从图书馆书架到互联网的任何地方。今天的组织已获取大量的数据和信息,但二者在质量上都没有达到要求,这意味着从满足使用的角度来看,数据或信息还缺少一个或多个必要特质。当许多组织为了知识管理与组织记忆而试图改进他们的系统时,数据和信息质量带来的问题会对它们造成更复杂的影响。

数据、信息和知识之间联系紧密。数据通常被看作是简单的事实;当数据有

一定的语境并兼具某种结构时，信息就出现了；当通过释义赋予信息一定含义时，信息就变成了知识。尽管表面看来，似乎是知识建立在信息的基础上，信息建立在数据的基础上；但 Tuomi（1999—2000）却极力主张先对所需知识进行界定，然后才能描述表达这一知识所需的信息，也只有界定了信息，才能描述原始数据以及将这些数据转换成信息所需的流程。例如，某个组织为了完成销售可能需要了解关于客户意愿的相关知识，该需求驱使这一组织确定某一销售订单，这一信息产品由关于客户、产品、服务的具体的原始数据组成，这些数据必需予以处理并填入协议表格中。

Tuomi（1999—2000）的知识、信息、数据这一倒序层次思想富有启发性，因为这意味着数据和信息质量标准也必须根据这一倒序层次来定义。一个组织不但应确定知识类型，还应确定日常工作和决策所需知识的质量层次，然后才能充分说明信息产品及其质量标准，继而才有可能保留和表达知识。一旦很好地理解了信息质量标准，组织才能在如何建模、表达和处理原始数据方面做出明智决策，以满足必要的标准要求。因此，在一个销售订单实例中，组织在没有从整体上完全理解销售订单所需的质量层次之前，无法充分说明构成销售订单不同数据成分的数据质量标准，而这又依靠对所需知识的期望质量来完成对客户的销售。

对组织而言，了解和提高数据和信息质量的动机比过去更为紧迫。越来越多的组织不再与客户、供货商、官方管理人员甚至雇员保持面对面的接触，而是像交换货物和服务一样，主要通过携带在当事人间所传递信息的数据进行联系。就像孩子们的电话游戏一样，一个孩子将一条消息通过耳语传递给另一个孩子，而另一个孩子又将其传递给下一个孩子，最后收到的信息可能变得无法理解或已失去消息的原始语境。没对来自信息供应链上所有参与人员间的个人联系数据进行验证和核实，低劣的数据和信息质量很可能逃过检测，最终导致组织在理解其存储的知识上出现问题。

以下几个例子说明了低劣数据和信息质量造成的麻烦。2002 年 7 月 24 日，美国宾夕法尼亚州西部 Quecreek 煤矿的 9 名矿工在井下 24 层作业时，意外打通了附近的一个废弃矿井，7700 万加仑冰冷的水涌入矿井，导致被困井下。幸运的是，4 天后矿工获救了，这多亏了几百名志愿者从一片农田里打通了一个通道，使营救工作进展顺利。据估计，该事故的营救工作及后果处理总花费超过 200 万美元，动用了 4 个州和联邦部门调查，而且多名被困矿工提出诉讼，起诉煤炭公司没有告知他们开采区紧邻废弃煤矿（Erdley 2003a）。同年晚些时候，宾夕法尼亚州环保部门发布了一个报告草案，指出关于该废弃煤矿的不完全的地图信息是导致 Quecreek 煤矿事故（Erdley，Prine 2003）的主要原因之一。现在，

宾夕法尼亚州需要有关老矿的地图(一类信息)与煤炭开采记录(另一类信息)之间进行前后对照,以提醒开矿规划人员地图上可能存在不准确的信息,即数据质量问题,这种不准确的信息使得开采公司无法知道在什么地方可以安全采煤,这便是知识所能起的作用(Erdley 2003b)。

1998 年 12 月 11 日,NASA 发射了火星气象卫星,这颗卫星是其 23 590 万美元的 Mars'98 项目的一部分(NASA 1998)。卫星经过 10 个月的旅程到达火星,对火星表面进行为期 1 火星年(687 天)的观测,观察火星的季节变化(Beoing 1998)。不幸的是,卫星于 1999 年 9 月 23 日到达火星后就消失了。工程人员推断该气象卫星可能是进入了火星很低的大气层后烧毁了(Jet Propulsion Laboratory 2003)。NASA 喷气推进实验室的一次内部审查得出结论:科罗拉多航天器研究组与加利福尼亚任务导航组之间传递的信息有一个错误,没能识别并改正这个错误导致了卫星在机动时进入了错误的轨道。很显然,在关键的飞行器操作中,一个小组用的是英制单位(如英寸、英尺和英磅),而另一个小组使用的是公制单位(喷气推进实验室 1999)。设计文档中这种数据单位的混用致使工程人员缺乏正确计算气象卫星进入火星大气层的轨道位置所必需的信息。

1999 年 12 月,医学研究院发布了一篇报告,估计美国每年有 44 000 人 ~ 98 000人死于本可避免的医疗事故,包括处方错误、血样标记错误以及难以辨认的纸质手写患者数据(Dash 1999)。这使得医疗事故成为美国死亡的第八大因素,每年因医疗事故造成的相关经济损失超过 170 亿美元(Hamblen 2000)。医疗机构正在通过执业规范和技术手段来减少因数据质量问题而导致的医疗事故,比如,采用手持系统来关联药物、病人和化验样本,使医务人员能够获取有关病人的更详细、准确的信息(Hamblen 2000)。

2001 年 10 月,美国邮电业邮政服务特别行动组(Postal Service Mailing Industry Task Force)建议提高地址信息质量并提供捕获与报告地址错误的反馈环路,以减少无法投递的邮件数量。邮电业每年因这些无法投递的邮件耗资 190 亿美元,而行动组估计整个行业耗资额是该数字的 2 倍(USPS 2003)。这一问题说明了信息产品中的数据质量问题虽像邮寄地址标签一样简单,但却能阻止组织获取正确的知识(如客户地址)。

Betts(2001a)描述了普华永道会计事务所(PricewaterhouseCoopers)在纽约进行的2001 年的一项研究,该研究发现被调查的599 家公司中有75% 的公司经历过因缺陷数据而引起的财政损失。报告指出,因不良数据管理,全球企业每年有 140 多亿美元浪费在结账、记账和库存混乱上(Betts 2001b)。此外,三分之一的公司表示"脏数据"迫使他们推迟甚至放弃新系统(Betts 2001b),只有 37% 的公司对自己的数据质量"非常有信心",只有 17% 的公司对其贸易伙伴的数据质

量“非常有信心”(Betts 2001b)。从该项研究得出的结论是“低劣数据质量正在威胁甚至破坏其他方面进行的大量投资”,如客户关系管理系统和供应链管理系统。因此,离开数据质量,人们将无法获取用以支持那些可提高机构知识管理系统所必需的高质量信息。

结果,诸如此类的经历促使许多组织正在寻找提高数据和信息质量的途径,因此,当前出现了包括学者、顾问和公司的一种新兴行业,提供解决这一问题的产品和服务。他们所提供的某些解决方案和建议是非常好的,而有些则不尽理想。不管怎样,鉴于多种原因,对很多组织来说要解决信息和数据质量问题都是极为困难的。

2 信息和数据质量面临的挑战

2.1 低劣数据和信息质量造成的损失难以量化

低劣数据和信息质量造成的损失通常都难以量化,因为其中不仅包含有形损失还包含无形损失。没有准确的损失估计,组织很难认识到低劣数据和信息质量对其盈亏底线正在产生的影响,从而,也不会优先考虑提高数据和信息质量。据 Redman(2003)估计,没有适当的主动质量规划,低劣数据和信息质量常常要损失一个组织约 20% 的收益。由于担心宣传会造成不良影响,很多公司对此都选择沉默,但据 Knight(1992)估计,公司数据库中劣质数据每年给美国工商企业和政府部门造成数十亿美元的损失。

劣质数据和信息的存在可通过以下几种途径造成更高损失:一是纠正因劣质数据或信息造成错误的代价与纠正数据或信息问题自身的代价并存,纠正因劣质数据和信息带来的影响这可能涉及到数据清理工作,可能会付出生命的代价,浪费宝贵的设备或生产时间,还可能会导致返工、诉讼或处罚、通过部分退款或道歉信的方式对客户予以安抚。Redman(1996,1 - 16)还引用了其他与质量有关的损失,例如,因为同一个公司内谁都不信任其他部门数据库里的信息,不同部门都保存着大量冗余信息;经理人花了更长时间做出的却是更糟糕、更不自信的决定;还有在采用诸如数据仓库或企业流程再设计(business - reengineering)项目新技术时组织上的困难。

除造成额外开支外,劣质数据还因客户的不满而造成合作伙伴流失,从而导致收益减少。某些组织存在太多重复的供应商记录(如 IBM Corp., I. B. M. Corporation, Intl. Bus. Mach.)导致他们缺少一个完整的、准确的业务往来视图,以至于他们无法谈成更好的交易和总额折扣(Betts 2001b)。应付劣质数据和信

息的影响可能会使员工有挫败感，降低工作乐趣，增加对组织的不信任度。在紧缺的劳务市场上，组织会发现劣质数据和信息限制了他们吸引和留住熟练员工的能力。这些发现引起了 Hansen 和 Wang 的共鸣(1991)，他们发现数据和信息质量会影响组织提供客户服务、管理支持和提高生产率的能力，从而影响公司的收益。

2.2 缺少一个完全定义的知识基础

关于数据和信息质量问题的研究早已有之。在过去的 40 年间，许多学科发表论文讨论了针对他们领域内劣质数据的相关问题和可能解决方案。例如，在统计领域，研究人员研究了如何使用计算机进行数据检查和修正来提高统计数据的质量。Joseph Naus(1975)介绍了计算机数据编辑程序如何使用变量间的关联信息来检测缺失值、异常值和不一致数据。一旦数据被标记为可疑数据，Naus 建议采取以下三种选择之一：① 追溯、检查、核对并修正标记数据；② 丢弃坏数据；③ 通过某些合理步骤估算数据。

图书馆管理学领域已经与在线数据库的质量问题纠缠了许多年。Toney (1992)介绍了一个旨在提高主要书目数据库质量的项目，该项目历时两年，修改了 210 万个数据字段，并因重复问题删除了 8.2% 的记录。

在会计领域，如何最好地检测、修正财务数据中的错误，以及评估错误影响的研究由来已久。像信息系统审计与控制协会(Information Systems Audit and Control Association, ISACA, 2003)这样的专业社团，提供了关于如何保护公司数据资产和保证数据完整的丰富信息。

在信息系统管理领域，例如，Morey(1982)、Ballou 和 Pazer(1985)、Laudon (1986)、Oman 和 Ayers(1988)等已意识到评估和提高信息系统中数据质量的必要。除了这些早期的研究，Mathieu 和 Khalil(1997)也发现虽然 1997 年的信息系统模型(Information Systems Model)课程包含了多种影响数据质量的主题，如“EDP 审计”、“数据字典”和“软件开发规程”，但并没有专门的课程指导信息系统的学生们如何在组织内测量、跟踪和提高数据质量。他们调查了 5 本有代表性的数据库系统教科书，发现结果也是相似的：虽然有关于数据质量方面的内容，如数据库设计，但没有明确提到数据质量。

所缺少的是一个可完整而非零散方式讨论数据质量的统一知识体系。为了应对这一挑战，20 世纪 90 年代初 Stuart Madnick 和 Richard Wang 在麻省理工学院启动了全面数据质量管理 (Total Data Quality Management, TDQM) 研究计划。该研究计划制定了几个目标。长期目标是创立基于相关学科的数据和信息质量理论并使之成为数据质量管理的统一知识体系，这些学科包括计算机科

学、统计学、会计学、全面质量管理、组织行为学等。第二个目标是成为数据质量技术实践人员的卓越中心，并将其作为有效方法和项目实践经验的交换场所（Madnick，Wang 1992）。通过 TDQM 研究计划发起的每年一次的信息质量国际会议，学者和实践者可以就如何提高数据和信息质量一起交换意见和建议。过去 10 年的协作已开始形成一个不断成长的专门研究与提高数据和信息质量的知识体系。

2.3 数据和信息质量是多维的

在测量数据质量之前必须先对其进行定义，这不是一项简单的任务。依据传统观点，大多数人都会认为数据质量是与准确性（accuracy）或可靠性（reliability）类似的东西。然而，处理数据质量问题的研究人员已经远远超越了给数据质量一个简单定义的地步。数据质量是多维的（Wang，Strong 1996）。Wang 和 Strong（1996）使用一个两阶段调查和两阶段分类研究提出了一个层次框架，将从数据用户那里收集的 118 个数据质量特征合并为 15 个维度，并进一步将这些维度归为 4 类。

- 内在数据质量（intrinsic data quality）：包括可信性（believability）、准确性（accuracy）、客观性（objectivity）、信誉度（reputation）几个维度，这意味着信息本身具有质量。
- 上下文数据质量（contextual data quality）：包括增值性（value - added）、关联性（relevance）、合时性（timeliness）、完整性（completeness）和数据适量性（appropriate amount）等几个维度，强调了信息质量需求应在当前任务上下文中予以考虑。
- 表达数据质量（representational data quality）：包括可解释性（interpretability）、易理解性（ease of understanding）、表达一致性（representational consistency）和表达简洁性（concise representation）等几个维度，关系到计算机系统存储和表示信息的方法。
- 可访问性数据质量（accessibility data quality）：包括可访问性（accessibility）和访问安全性（access security），强调计算机系统必须是可访问的，并且是安全的。

信息质量具有多维性意味着组织必须使用多种测量方法，以全面评估其数据是否适合特定用户在特定时间以特定目的使用，还意味着信息产品的质量必须由数据项的质量，包含这些数据项的处理质量，以及数据产品本身的设计质量决定。

2.4 数据质量难以测量

虽然现在有很多数据质量相关的维度已经得到认同，但还是很难获取每一

个维度的严格定义,以便利用它进行测量并随时比较测量结果。Wand 和 Wang (1996)在其研究中讨论了这一问题,他们应用一个本体框架提出了真实世界系统到信息系统状态映射中的 4 个定义良好的缺陷,即数据是否是完整的(complete)、明确的(unambiguous)、有意义的(meaningful)和正确的(correct)。Pipino、Lee 和 Wang(2002)提出了数据质量的主观(subjective)和客观(objective)评价方法,包括简单比率(simple ratios)、最小或最大算子(min or max operators)、加权平均值(weighted average),以辅助制定实用的数据质量测量标准。最近的研究进一步阐明了如何最好地定义和测量与数据质量相关的不同维度,以及如何将这些测量结果汇集成信息产品质量的全面评估,该信息产品可由几种不同的数据成分构成。

2.5 信息质量管理需要新的管理模式

将信息视作副产品的传统观点导致很多管理人员、分析人员和研究人员将注意力集中于硬件和软件、烟囱式系统的(stovepipe systems)成本控制以及系统是否运行,而不是关心这些系统生产的产品是否真实地满足消费者需求。将信息视为产品的新概念模型,与工业产品系统极其类似,易于将明确的角色、技术、管理原则、统计控制和测量方法应用于信息。Wang 等人(1998)明确了将信息作为产品时组织必须遵循的 4 个原则:

(1) 理解客户的信息需求。

(2) 将信息作为良好定义生产过程的产品来管理。

(3) 将信息作为一个有生命周期的产品来管理。

(4) 指定一个信息产品经理来管理信息过程和最终产品。

这种将信息作为产品管理的思想开创了一种崭新的方法论——全面数据质量管理,全面数据质量管理遵循了以下任务周期(Wang,1998)。

(1) 定义信息产品:是指根据针对数据消费者的设计目的,定义信息产品的特征、基本的原始数据单位、中介子系统及其相互关系方面定义信息产品。还意味着从信息产品供应商、生产者、消费者和管理者的角度定义信息产品的需求,并确定产生信息产品的信息生产系统。

(2) 测量信息产品:通过跟踪基于信息产品定义的信息度量可以随时监测信息产品质量。

(3) 分析信息产品:利用测量结果可以检测调查现有数据质量问题的根本原因。

(4) 改进信息产品:一旦完成分析,为了生产出更好质量的信息产品,可以着手消除引起数据质量问题的根本原因。