面向共享的政府大数据质量标准化问题研究

洪学海[1]，王志强[2]，杨青海[2]
1. 中国科学院计算技术研究所，北京 100190；2. 中国标准化研究院，北京 100191

摘要
回顾了国内外数据质量研究与实践的进展，重点对ISO 8000数据质量国际标准提出的数据质量框架、主数据质量、事务数据质量和产品数据质量进行了探讨，对面向共享的政府大数据质量标准化的方法和测度理论进行了研究，最后对我国政府进行大数据质量控制及其标准化建设提出了建议。

关键词
政府大数据；主数据；产品数据；数据质量；ISO 8000

中图分类号：F253.3，L70　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2017029

Research on the quality control of sharing big data for government

HONG Xuehai[1], WANG Zhiqiang[2], YANG Qinghai[2]
1. Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China
2. China National Institute of Standardization, Beijing 100191, China

Abstract
The progress of research and practice in data quality standardization was reviewed, and the framework of data quality was introduced, which was put forward by the international standard of ISO 8000 data quality. The master data quality, transaction data quality and product data quality were discussed. The method and measurement theory of the large data quality standardization for sharing were discussed. At last, suggestions for China's government in the big data quality control and standardization were put forward.

Key words
government big data, master data, product data, data quality, ISO 8000

1 引言

大数据时代的到来，对我国政府的决策模式、治理模式和工作方式等都提出了新的挑战。推进政府大数据战略对实现政府治理有重要的意义，是政府治理实现的强力助推器。

当前，不论在整个社会的哪个行业、哪个部门、哪个单位、哪个个人，“数据”成为其核心属性，“数据”成为其核心业务纽带或重要的标识工具，即“数据”贯穿着集体或个人业务信息的“采集、存储、传输、处理、应用”的全过程，“拿数据说话”成为共识。

对于政府管理来说，“拿数据说话”就是借助大数据分析、挖掘等技术，对政府获得的方方面面的大数据进行深度分析，建立关系、找出问题、发现规律等，从而辅助政府管理部门和主要领导对政府管理的方方面面的工作进行决策，提高决策的有效性和科学性。而这个前提就是政府数据能够共享，并且共享的数据是准确的，一定程度上是标准化的，只有保障政府大数据能够共享，并且数据准确、完整，那么在此基础上进行辅助政府决策的大数据分析，才能够表现出发现问题准确、建立问题之间联系的关系脉络清晰、发现的规律有迹可循等特点。因此，研究政府大数据，首先要解决的是政府大数据开放与共享问题，其次最重要的是政府大数据的质量标准化问题。

2 政府大数据共享及其数据质量面临的挑战

政府大数据一方面来自政府部门本身的业务积累，如医疗管理部门、交通管理部门、城市经济管理部门等，另一方面来自专门单位的采集，如地理信息、生态环境信息等。来源可谓广泛，种类可谓繁多。政府大数据是国家和全社会的公共财富，价值密度高。然而，在笔者的研究过程中发现，真正要实现政府大数据的潜在价值，不仅技术方面面临着大数据复杂性带来的问题（如数据本身的复杂性、计算的复杂性和信息系统的复杂性），而且政府大数据融合方面还面临着政府大数据资源的管理、质量和标准化等一系列的问题和挑战，主要有以下几个方面。

（1）数据本身的变化

数据的价值，从单一转向多元；政府数据资源的形态，以结构化为主转向以非结构化为主，从离线静态数据转变为在线动态、实时数据；数据资源的战略地位，从机构组织层转向跨机构组织、区域和国家层；数据权由简变繁，并具有不确定性，涉及信息主体的所有权、删除或留存处置权、利用权、授权他人利用的许可和审批权、隐私保护权等，甚至涉及国家数据主权议题等。

（2）数据管理主体的变化

数据管理主体从数据的控制者转变为数据的提供者、保护者和获取权利的协调者；从追求部门局部利益最大化转向追求政府整体效益及社会利益的最大化；从信息孤岛转向跨界、跨领域、跨部门、跨系统、跨层级的信息融合；需要多主体联盟与跨学科复合型数据人才支持。

（3）数据管理活动过程的变化

政府数据资源的采集，从单一来源转向多源异构，从基于目标的局部采集转向基于场景的全面采集；政府数据的存储，从分布式、冷备份存储转向云端、热备份存储；政府数据的利用，从个别部门的数据公开转向政府数据集的整体开放，从处置边界明确转向互联互通，边界模糊；

政府数据的维护，从信息化管理转向数据化、网络化、智能化、“互联网+”的现代化治理。

上述这些挑战在笔者研究“宁波市政府大数据项目的数据开放与社会化利用”等课题的过程中已经充分暴露出来。突出表现首先是政府各个部门的数据标准不一、质量千差万别，没有基准（benchmark），甚至同一个市民的个人属性数据在公安、社保等部门的数据项、数据集等都不统一，同一个人的属性数据甚至还“打架”。上述存在的这些问题和挑战可归结为：如何在技术和政策上保障政府大数据共享目标能够实现；在技术保障上，除了共享的信息网络系统体系外，作为政府大数据本身，如何保障共享的数据可用、可融合，就是政府大数据开放共享最基础性的工作。若数据不准确或数据缺失，即使共享也没有价值；若数据标准没有统一，即使共享也难以发挥大数据融合带来的令人期盼的效果。政府大数据质量问题在现阶段比较突出，这给依赖于政府大数据进行政府重大事项的决策带来很大的风险。

3 大数据环境下数据质量标准化与传统的数据质量标准化的差异

大数据质量问题是数据质量问题在这个新阶段（大数据环境）表现的一个新形式，是数据质量历史的一个阶段。可以预见，伴随着信息技术的发展和不断演化，数据质量会呈现出不同的变化形式。

20世纪80年代以来，国际上对数据质量的概念也从狭义向广义转变，准确性不再是衡量数据质量的唯一标准。20世纪90年代，美国麻省理工学院（Massachusetts Institute of Technology，MIT）开展的全面数据质量管理（total data quality management，TDQM）活动，提出基于信息生产系统生产的数据产品的质量管理体系，在数据生产过程中形成的数据质量（如精度、一致性、完整性等）成为基本要求。数据用户要求的满意程度也成为衡量数据质量的重要指标，认为数据质量就是要“反映出数据对特定应用的满足程度”[1]。例如，在智能制造系统中，数据是应用程序的初始原料和最终产品，并经过应用程序的组织，提供给用户[2]。同样的一组数据，面对不同的应用要求，可能表现出不同的数据质量。

传统的数据质量的研究和实践总体上可归纳为“自上而下”和“自下而上”两种方式[3]。“自上而下”方法通常是先提出数据质量框架（data quality framework）和数据质量维度（data quality dimension），数据质量维度也称为数据质量属性、数据质量元素、数据质量衡量指标、数据质量特征等，然后在应用中通过与具体的需求相结合，构建可执行的细化的数据质量维度；而“自下而上”则是从具体需求出发，提炼出一系列的数据质量维度，通过实际应用的验证，最后归纳形成数据质量框架。在具体的应用实践中，既存在理论上构建数据质量框架但不细化到可操作的维度的现象，也存在仅在具体操作层面定义数据质量维度、改善数据质量状况但不上升到数据质量框架的具体应用，而且在实际实践中后者更多。

当前，在大数据环境下，研究数据质量标准化问题，一个显著的不同于传统的数据质量标准化的问题是强调保障多目标数据融合的实现，这也是发挥大数据价值的重要方式。由于数据来源不同、数据种类异构以及数据类型繁杂，使得用传统的数据质量标准框架和质量维度定义大数据质

量标准体系存在不适应问题，因为传统的数据质量体系是针对单一来源数据和单一类型数据的。同时，现在大数据环境下的数据质量体系是将各种单一来源甚至单一数据类型的数据进行“混合”，形成非单一来源、非单一数据类型的“数据集”，应围绕数据融合的目标而定义新的大数据质量体系，并且数据融合的粒度大小决定了大数据质量框架和质量维度是细粒度还是粗粒度。因此，研究大数据环境下的数据质量体系需要在传统数据质量体系的基础上，再研究新的大数据质量体系框架和质量维度。

国际上到目前为止，对于大数据质量标准化的研究和制定工作都还在起步阶段，主要是依赖数据技术体系，从基础、技术、产品和应用的不同角度进行分析，形成大数据质量标准化体系框架。主要由ISO/IEC JTC1 SC32的“数据管理与交换”分技术委员会、ISO/IEC JTC1 WG9大数据工作组、国际电信联盟（International Telecommunication Union，ITU）以及美国国家标准技术研究院（National Institute of Standards and Technology，NIST）等相关组织和机构开展此项研究和标准编制工作。我国主要是全国信息技术标准化技术委员会在进行大数据标准化工作，期望与国际标准接轨。但是可以预见的是，考虑大数据质量问题的标准化工作难度较大。

4 国际标准ISO 8000与面向共享的政府大数据质量标准体系框架

4.1 数据质量国际标准——ISO 8000

ISO 8000是一套国际通用的数据质量管理标准，立足于工业数据质量，旨在为政府、公共机构和各类公司、制造企业以及应用提供更可靠、可信数据的国际标准。ISO 8000涵盖从概念设计到废弃处置整个数据生命周期中的质量特征。ISO 8000列出的特种数据包括但不限于：主数据、事务数据和产品数据。ISO 8000给出了一个用于改善某种特定数据的数据质量框架。该框架可独立使用，也可与质量管理系统协同使用。ISO 8000定义了一组特征，数据供应链中的任何组织都可用其测试数据是否与ISO 8000保持一致。

ISO 8000是ISO 9000质量管理体系的扩充，以满足质量管理体系内数据产品质量的需求。实践证明，如果不能保证数据质量，ISO 9000是不能真正实现其质量目标的。ISO 9000标准家族是国际标准化组织于1987年制定并经过后续不断修改完善而成的系列标准，可帮助组织实施、有效运行质量管理体系，是质量管理体系通用的要求或指南[4]。它不受具体的行业或经济部门限制，可广泛适用于各种类型和规模的组织。

图1显示了ISO 8000、ISO 9000和其他数据产品标准之间的关系。数据描述标准规定交换数据的模型和格式，ISO 8000以这些标准为基础，增加了关于这些标准

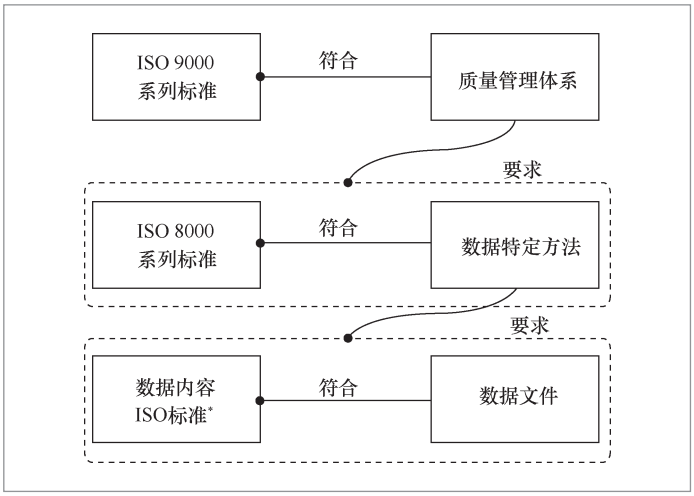

图1 ISO 8000、ISO 9000 和其他数据产品标准之间的关系

的使用要求，以确保交换数据的高质量。ISO 8000弥合了ISO 9000和数据产品标准之间的差距[5]。

ISO 8000数据质量国际标准由系列部分组成[6]，各部分的侧重不同，ISO 8000由以下部分组成。

- 第1~99部分：数据质量综述。
- 第100~199部分：主数据质量。
- 第200~299部分：事务数据质量。
- 第300~399部分：产品数据质量。

其中，主数据标识和描述了个人、组织、地点、物品、服务、过程、规则和标准。该系列标准描述定义主数据质量的特性，规定了一些主数据信息，这些信息应在总体上确保信息发送方和接收方数据通信的可靠性。

事务数据规定和描述了时间事件，包括个人、组织、地点、物品、服务、过程、规则和标准。该系列标准描述定义事务数据质量的特性，规定了一些业务事务数据信息，这些信息应在总体上确保信息发送方和接收方数据通信的可靠性。

产品数据质量是产品数据正确性和适用性的度量，产品数据可保证数据能及时地提供给需要这些数据的用户，产品数据是产品从概念到制造需要的数据。

在政府大数据相关开发与利用的应用实践中，数据质量标准化具有极其重要的战略地位。可以借鉴国内外业已成功应用ISO 8000数据质量国际标准的行业经验，研究ISO 8000数据质量国际标准在政府大数据领域的应用，建立和完善数据质量管理体系，提高政府大数据质量，深化质量标准体系，为发掘政府大数据价值提供保障。

4.2 面向共享的政府大数据质量标准体系框架

到目前为止，对政府大数据的范围或边界还没有形成共识，因此，在研究政府大数据质量体系的过程中，要遵循“循序渐进”的策略，从政府各个相关管理部门的管理职责范畴考虑政府大数据的最小元数据集，由此逐步向外延展。

政府大数据数据质量框架是面向政府管理的数据质量问题的基本概念及其解决方案、实施指导的抽象化结构表达。它表现为一组构件及构件实施指导、实例交互方法，能够在具体应用中灵活定制质量工作架构，较适合政府管理部门范围内数据质量问题复杂多样且统一解决方案的需求。

从一般意义上来看，国家大数据标准体系由6个类别的标准组成，分别为：基础标准、数据处理标准、数据安全标准、数据质量标准、产品和平台标准及应用和服务标准。而从政府大数据角度看，面向共享的政府大数据质量标准体系是政府大数据质量标准体系的有机组成部分。

由此建立的政府大数据质量指标体系主要有：数据源质量、数据规模质量、数据结构质量、数据时效质量、数据价值密度质量。这5个指标体系是政府大数据质量标准的5个一级指标，数据源质量指标是数据一般性质量，另外4个质量描述的是大数据的四大特征质量。一直以来，数据质量框架是粗粒度研究数据质量问题和解决方案的重要内容和方向。笔者提出的政府大数据质量体系框架是一个参考模型，在评价各个政府大数据质量的过程中，需因地制宜。

此外，还需要考虑政府大数据质量维度问题。有些参考文献将数据质量问题直接定义为一组属性（特征），如正确性、适时性、完全性、一致性和相关性等。数据质量判断依赖于使用数据的个体，不同环境下不同人员使用的适合性不同，数据质量是相对的，不能独立于使用数据的消费

者来评价数据质量。由此可见，政府大数据的质量问题从数据质量维度来看，可以为建立面向共享的政府大数据质量评价体系的二级乃至三级指标体系提供多维度的指标，从而可以构建不同目标、不同方式的面向共享的政府大数据的质量评价体系框架。

在以后的研究中，需要分析面向共享的政府大数据标准化需求，研究大数据质量的特殊性，研究大数据标准化的特殊性。针对典型应用，理解大数据共享的主要价值，研究政府大数据共享现状，研究政府大数据质量现状，分析政府大数据质量标准化需求。根据当前信息技术及其应用的发展趋势，研究政府大数据资源共享的未来前景，研究典型应用中政府大数据质量问题，研究政府大数据质量标准化当前以及未来的总体需求。同时，需要提出标准体系框架与明细表，梳理政府大数据质量技术标准，研究政府大数据质量标准与技术发展、业务领域的关联性，对政府大数据质量标准进行全景式分类研究，给出适用的政府大数据标准分类描述体系。在此基础上，提出政府大数据标准体系框架，建立政府大数据标准明细表。

5 面向共享的政府大数据质量标准化方法

面向政府大数据共享，开展大数据质量标准化方法研究意义重大，包括标准化循环改进过程研究和标准化演化机理研究。大数据质量标准化循环改进过程如**图2**所示。以政府的行政管理为主要应用领域，基于过程控制方法，建立大数据质量保证方法，通过构建大数据质量评估模型，实现大数据质量的改进和完善。通过大数据质量计划、大数据质量实施、大数据质量评价、大数据质量改进来实现大数据的质量目标。对大数据的质量评价应建立在与大数据质量标准化、大数据标准体系密切关联的大数据质量测度模型的基础上。大数据质量标准化与质量改进，需要满足大数据质量需求，并实现大数据质量效益的目标。

从时间维、空间维和业务维3个维度探索大数据质量标准化发展变化的客观规律，研究大数据质量标准化的动态演化机理。研究大数据质量标准化过程的主要特点和规律，包括其复杂性、网络化、自组织等特性。

- 复杂性包括涉及大数据生命周期各阶段的时间复杂性、涉及不同层级相关组织的空间复杂性、涉及各领域应用对象的业务复杂性。
- 网络化是指在大数据质量标准化演化中，不同层级的相关组织形成的多种形式的关联关系。
- 自组织是指大数据质量标准化的过程是一个自行改进、优胜劣汰的系统过程。

6 面向共享的政府大数据质量测度理论方法

政府大数据质量具有其特殊性，一是数据来源的多样性，带来丰富的数据类型，增加了数据质量评测的难度；二是数据规模的海量性，使得难以在合理的时间内判断数据质量的好坏；三是数据变化的快速性，使得难以形成相对稳定的数据质量评测体系和方法。这也就决定了在大数据环境下，数据质量的测度理论和评价方法与传统数据质量测度和评价相比会有显著不同。大数据质量是

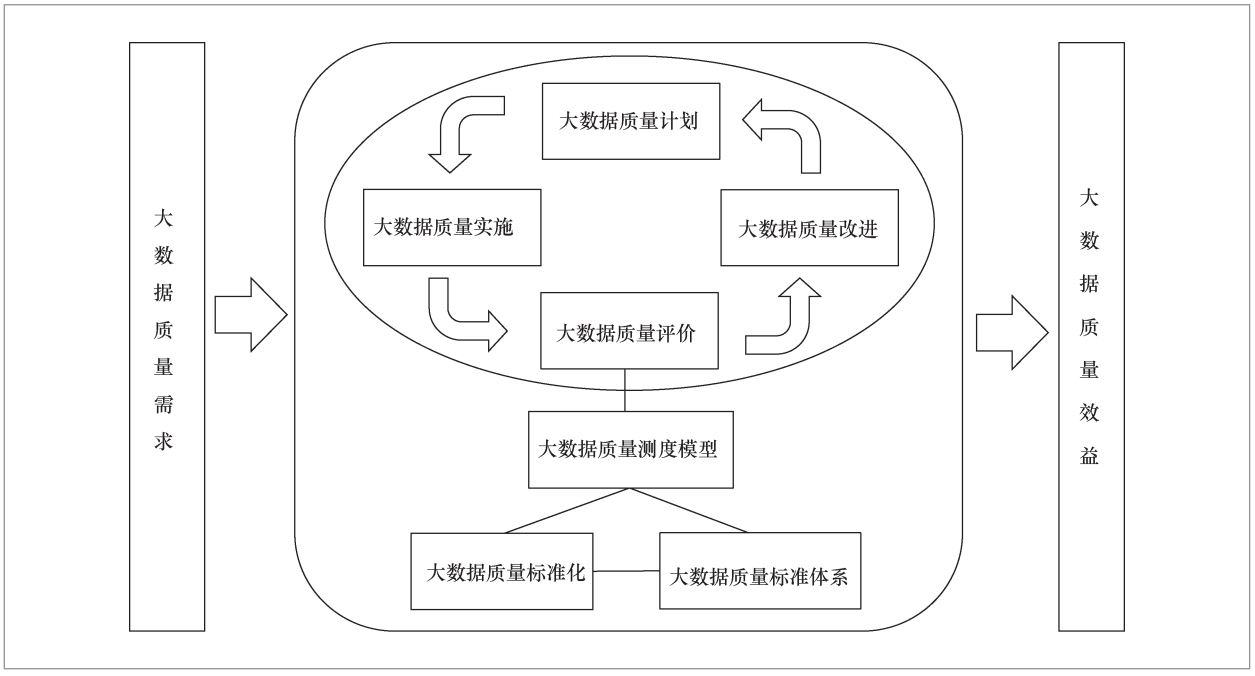

图 2 大数据质量标准化循环改进过程

全景式的数据质量，包括面向数据生命周期的时间维、面向不同层级逻辑组织的空间维、面向不同领域应用对象的业务维。

面向政府大数据共享，开展大数据质量测度理论方法研究，包括测度模型的研究和评价方法的研究。图3为大数据全景式数据质量测度模型，分为时间维、空间维、业务维3个维度。时间维面向大数据生命周期，包括大数据产生、采集、传输、存储、分析、使用等环节。空间维面向大数据的逻辑组织空间，包括企业/子部门内、行业/部门内、跨行业/部门等多个层级。业务维面向大数据的主要业务对象，包括商贸、监管、工业、交通、安全、医疗等政府相关业务领域。时间维、空间维、业务维3个维度以及各个维度上的要素，反映了大数据质量的多个视角、关键影响要素，共同构成大数据共享质量测度的模型，为大数据共享质量评价奠定理论基础。

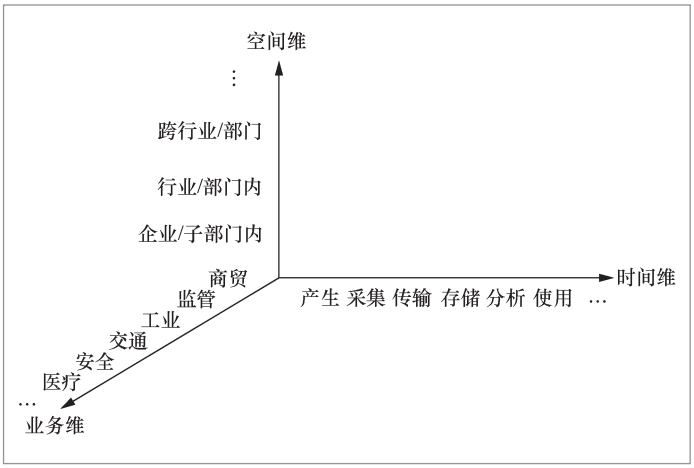

图 3 全景式数据质量测度模型

大数据的来源和应用都是多学科的，对大数据的质量评价不是单一阶段、单一组织、单一领域的技术问题，而是一个跨周期、跨部门、跨业务的综合性问题，是一项系统工程。需要研究测度模型及其各个视图投影，研究多维度的综合评价方法以及某个视角的特定评价方法。

针对以上特点，大数据质量测度需要

建立全景式测度模型，从而综合考虑各个环节、各个层级、各个领域的特殊性和普遍性，以提升大数据质量测度模型的科学性和适用性。

7 结束语

大数据时代的到来对我国政府的决策模式、治理模式和工作方式等都提出了新的挑战。推进政府大数据战略对实现政府治理有着重要的意义，是政府治理实现的强力助推器。当前，这一工作的推进面临着数据公开缺乏社会规范条件、数据格式缺乏统一、数据共享缺乏有效体制保障和大数据数据质量参差不齐等问题。更为重要的是，政府数据质量问题对于我国政府大数据共享至关重要。政府大数据质量的提高涉及技术、设计、流程、人员和基础设施等多个方面。对政府大数据质量开展研究，进而提出改善数据质量的方法和对策，保障政府大数据质量，具有非常重大的意义。

结合目前我国政府大数据的数据标准、数据质量管理等现状以及ISO 8000等数据质量国际标准，建议从以下4个方面着手提高政府数据质量。

（1）建立政府大数据质量标准

在深入研究ISO 8000等数据质量标准体系的基础上，结合我国政府大数据现状，建立面向共享的政府大数据质量标准，为政府大数据质量管理提供全面的遵从依据，从数据权属和治理的角度，提出大数据标准化运行机制。

（2）建立政府大数据数据质量管理流程

基于ISO 8000 等数据质量标准体系，结合我国各地政府部门大数据现状，建立数据质量管理体系流程，规范数据质量管理过程，提升数据质量管理的科学性，保障数据质量标准在政府大数据共享中的落地，也确保政府大数据不仅能“共”，而且还能共“享”。

（3）构建政府大数据数据质量评价模型及考核方式

基于ISO 8000的数据质量标准体系，构建政府大数据数据质量评价模型并固化，结合现有政府大数据数据质量通报等考核方式，为全面管控各级政府数据质量情况提供支撑。

（4）建立政府大数据质量管理信息化支撑工具

继承并扩展现有政府大数据管理信息化系统，为政府大数据质量标准落地、管理流程落地、评价模型落地及考核落地提供信息化支撑。

参考文献:

[1] LEE Y W, STRONG D M. Knowing-why about date processes and data quality[J]. Journal of Management Information System, 2003, 20(3): 13-39.

[2] LEE Y W, PIPINO L, STRONG D M, et al. Process-embedded data intergerity[J]. Journal of Datebase Management, 2004, 15(1): 87-103.

[3] 胡良霖, 黎建辉, 刘宁, 等. 科学数据质量实践与若干思考[J]. 科研信息化技术与应用, 2012, 3(2): 10-18.
HU L L, LI J H, LIU N, et al. Practice and some thoughts on quality of scientific data[J]. e-Science Technology & Application, 2012, 3(2): 10-18.

[4] 王军玲, 李华, 王强. ISO 8000 数据质量系列标准探析[J]. 标准科学, 2010(12): 44-46.
WANG J L, LI H, WANG Q. Research on ISO 8000 series standards for data quality[J]. World Standardization & Quality Management, 2010(12): 44-46.

[5] STRONG D M, LEE Y W, WANG R Y. 10 potholes in the road to information quality[J].

IEEE Computer, 1997, 30(8): 38-46.

[6] 国际标准化组织. 数据质量第1部分：综述：ISO/TS 8000-1:2011[S]. [出版地不详：出版者不详], 2011.

International Organization for Standardization. Data quality-Part 1: overview: ISO/TS 8000-1:2011[S]. [S.l:s.n.], 2011.

作者简介

洪学海(1967-)，男，博士，中国科学院计算技术研究所研究员，信息技术战略研究中心常务副主任，兼任中国科学院计算机网络信息中心信息化战略与评估中心主任，主要从事高性能计算、信息服务计算以及信息技术与信息化发展战略等方面的研究工作。发表文章40余篇，合著中文专著5本。

王志强(1975-)，男，中国标准化研究院高新技术与信息标准化研究所副研究员、副所长，主要研究方向为工业数据标准化、数据质量标准化、信息资源开发利用、电子政务标准化等。

杨青海(1965-)，男，博士，中国标准化研究院高级工程师，主要研究方向为工业数据标准化、产品模块化，出版著作1本、译著1本，发表论文10余篇。

收稿日期：2017-04-01

基金项目：国家自然科学基金资助项目(No.91646127)

Foundation Item: The National Natural Science Foundation of China(No.91646127)