基于DMBOK的电动汽车多源信息决策支持系统数据质量管理过程研究

齐 娜 赵 辉 张英杰
（中国科学技术信息研究所，北京 100038）

摘 要：依据国际数据管理协会数据管理知识体系（DMBOK）中的数据质量管理过程，分析了电动汽车多源信息决策支持系统中的数据特点以及建设中存在的问题，提出改进建议，并初步构建适用于电动汽车多源信息决策支持系统的数据质量管理流程。

关键词：DMBOK；DAMA；决策支持系统；数据质量；电动汽车
中图分类号：TP311 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2014.04.004

Study on Quality Management Process of Decision Support System for Electric Vehicle Based on the DMBOK

Qi Na, Zhao Hui, Zhang Yingjie
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: This article introduces the DMBOK brought by International Data Management Association at the beginning, then analyses the problems in the building process of database of electric vehicle multi-source decision support system. At last it gives the suggestions, and builds the Management process of database quality on electric vehicle multi-source decision support system.

Keywords: Data Management Body of Knowledge(DMBOK), Data Management Association International(DAMA), decision support system, data quality, electric vehicle

1 引言

多源信息具有显著的动态性、分布性、多元性和无序性等特征，它造成了信息环境下信息开发利用的局部有序和全局无序的矛盾，或组织内有序和组织外无序的矛盾，使信息查找与管理变得越来越困难[1-7]。只有依据高质量的事实型数据，得出的分析结果才会对决策行为起到有效支撑。在实际的业务系统建设项目中，数据质量问题不仅仅包括校正数据，同时还包括管理从数据生成、存储、管理、流转、使用、销毁、失效等整个生命周期的各阶段，从而确保生成的信息满足组织中全部数据用户的需求。

本文拟借鉴国际数据管理协会的数据管理框

作者简介：齐娜*（1979- ），女，中国科学技术信息研究所助理研究员，研究方向：信息资源管理；**赵辉**（1971- ），女，中国科学技术信息研究所副研究馆员，研究方向：科技资源管理；**张英杰**（1979- ），男，中国科学技术信息研究所助理研究员，研究方向：信息资源管理。

基金项目：国家科技支撑计划项目“电动汽车专题数据库建设”（2013BAG06B02）；“电动汽车多源信息决策支持系统研发”（2013BAG06B03）；国家社科基金“网络环境下科技信息资源建设中的质量元数据及评估应用研究”(12BTQ016)。
收稿日期：2014年5月19日。

架，梳理支撑电动汽车多源信息决策支持系统的数据质量管理流程。

2 数据质量管理框架

数据质量管理（DQM）是数据管理框架的重要职能之一，它是与数据治理职能交互并受其影响的数据管理职能[8]。数据质量管理是组织变革管理中一项关键的支撑流程，其定义为：通过计划、实施和控制活动，运用质量管理技术度量、评估、改进和保证数据的恰当使用。其目标是：适度改进数据质量，满足既定的业务预期；定义需求和规格说明，将数据质量管理整合至系统开发生命周期；为度量、监控和报告数据质量水平的一致性提供既定的操作程序。

质量管理的一种通用方法是戴明质量环，即“计划－实施－检查－行动”。该模型对数据质量管理同样有效。将此模型应用于数据质量管理时，包括制定数据质量现状评估计划和识别数据质量度量关键指标；实施度量和提升数据质量流程；监控和度量根据业务预期定义的数据质量水平；执行解决数据质量问题的行动方案，以提升数据质量，从而更好地满足业务预期。经过DAMA组织的重新定义，数据质量管理的主要活动为计划、控制、开发、操作几个步骤；实施步骤又分为12项，如图1所示。

3 电动汽车多源信息决策支持系统

电动汽车多源信息决策支持系统是纯电动、混合动力和燃料电池三大领域相匹配的技术、管理、市场和动态类数据，通过决策数据挖掘、信息分析和可视化等关键技术，构建支撑技术管理决策的信息服务系统和规范报告生成、管理系统，形成电动汽车技术决策和计划管理的一站式支撑服务系统。

该系统的数据支撑由20个电动汽车领域专题数据库构成。这些专题数据来源不同，主要表现为有的是从现有成熟数据库经过检索筛选后导入，例如文献、专利库，有的则是根据需求重新采集加工；数据类型各异，包括了结构化数据、

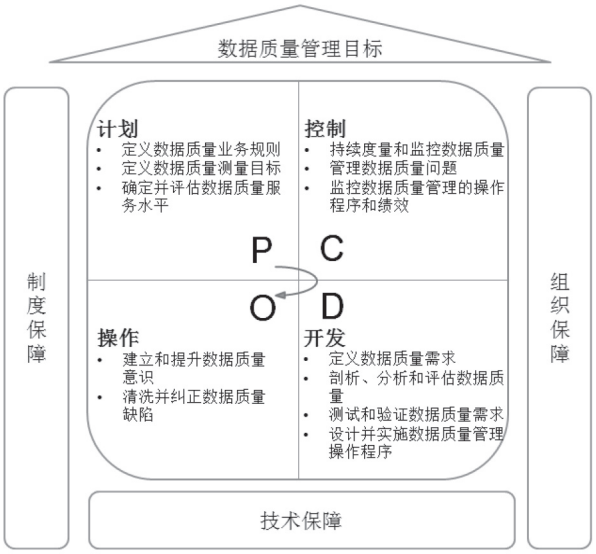

图1 DMBOK数据质量管理框架

非结构化数据；数据类型全面，涵盖了纯电动汽车、混合动力汽车、燃料电池汽车，从技术范围上覆盖了电池、电机和电控。这些多维多源信息如何进行精细化组织，实现资源与决策需求的对接，这需要对多资源类型、多形态资源进行精细化整合，对多源数据进行高效的数据质量管控。

电动汽车多源信息决策支持系统从业务或业务关系上讲，该系统涉及多种业务；从数据产生环节来说，该系统数据涉及从计划到运行的各个阶段；从数据本身来说，该系统数据既包括原始数据，又有分析后的成果数据等。对照DMBOK的数据质量管理流程，在数据清洗加工过程中，分析了系统内的数据特点和提出了主要的数据质量要求，如表1所示。

由于该系统设计建设周期有限，采用的是系统开发与数据库建设同步进行的方式实施，基于全生命周期的数据质量管理相关规范也是随着数据库的建设在不断完善。在此过程中，表1所列出的数据质量要求有些可以在数据采集计划阶段得到满足，有些可以在采集阶段满足，有些可以在数据清洗入库阶段满足。

4 电动汽车多源信息决策支持系统数据质量管理流程

数据生命周期是对数据生产加工过程的系统

表1　电动汽车多源信息决策支持系统数据库情况

序号	类别	数据库名称	数据特点及主要质量要求
1	基础类	电动汽车人才数据库	人才类别划分
2		电动汽车机构数据库	机构归一性
3		产业化及产业组织形态数据库	机构归一性；机构演变关系
4		电动汽车项目数据库	项目主题类别及项目金额度量衡一致性
5		电动汽车计划数据库	各类计划的演变及与项目的归属
6	输出产出类	电动汽车政策法规数据库	对法律法规的准确性要求高
7		电动汽车学术文献数据库	成熟数据库导入，对查全性与查准性要求高
8		电动汽车技术标准数据库	技术标准的时效性
9		网络动态类数据库	时效性强
10		电动汽车规范报告数据库	规范报告的来源、分类和权威问题
11		电动汽车专题研究报告库	成熟数据库导入，对查全性与查准性要求高
12		世界电动汽车专利数据库	成熟数据库导入，对查全性与查准性要求高
13		电动汽车产品样品数据库	计量标准统一
14		车型数据库	车型多维度分类
15		电池、电机关键零部件数据库	计量标准统一
16	基础设施类	配套基础设施及使用状况数据库	配套设施统计数据的全面性、及时性。
17		示范推广运行数据库	多为非结构化数据，需转为结构化数据后为统计分析服务
18	指标与度量类	电动汽车商业化数据库	时效性强，计量标准需统一
19		电动汽车资源与环境评价数据库建设	数值型数据；对准确性、规范性、权威性要求高。
20	语义类	电动汽车词系统数据库	准确性、全面性要求高

表述。数据生命周期管理应贯穿到整个数据的生命周期中，从数据采集计划的制定到最终对外提供数据服务，应根据数据不同阶段采取不同的应对措施而加以管理，达到数据质量管理的预期目标。其目的在于帮助项目管理者在数据生命的各个阶段以最低的整体控制成本获得最大的效益。它是一个针对数据生产过程进行主动管理的策略。

从项目管理的角度看，电动汽车多源信息决策支持系统的建设其实包括了3个层面的建设工作：首先是系统建设过程，从计划、分析、设计、编程、调试、测定直至运行、跟踪、评价的过程；其次是用于支撑决策支持系统的数据生产过程，从数据的计划、采集、清洗、入库、维护、服务的数据流过程；最后是从质量控制的角度来看，它还是一个质量信息的收集、质量管理策划、质量控制、质量评估的质量管理过程。

经过梳理，我们认为支撑电动汽车多源决策支持系统数据的全流程生命周期以及数据质量管理流程如图2所示。

在基于DMBOK的电动汽车多源信息系统数据质量管理框架内，具体的质量控制过程可以细分为以下几个步骤。

（1）培养数据质量意识。在本项目组中配备专职数据管理人员，要求这些专职人员能够发现数据质量问题的存在，能将数据质量问题与其对决策支持系统的实质影响联系起来，对组织内数据质量有全面的洞察，并向项目团队内成员传达“数据质量问题不能仅靠技术手段解决”的理念。与20类专题数据提供单位密切合作，协同支持数据质量管理项目。

（2）定义数据质量需求。目前基于用户需求的数据管理成为主流，所以数据的质量必须具备适用性。可以根据定义好的数据质量维度来度量数据是否符合适用性需求，并生成数据质量指标的报告。电动汽车决策支持系统的用户为电动汽车领域的决策者以及专家顾问组成员。需要根据他们的具体使用需求来定义数据质量需求。数据质量维度体现了高层次指标度量的特点，可以据此对决策支持系统业务规则进行分类。根据实施

的需要，对度量的粒度进行细化，如数据值、数据元素、数据记录和数据表。本项目所采用的主要数据质量维度详见表2。

（3）剖析、分析和评估数据质量。可选择采用自下而上或自上而下的方法进行数据质量评估。自下而上法：进行数据分析可较为直观地检测出数据异常，问题数据需要专业人员进行有效性验证和分析。这种方法强调潜在问题，包括出现率分析、重复性分析、跨数据集的关联关系、孤立单一数据记录和冗余分析。自上而下法：需要理解决策支持系统的业务流程如何使用数据，各业务模块需要调用哪些数据库的哪些数据，哪些数据字段对于业务应用的成功至关重要。依据此对数据质量进行分析评估。结合电动汽车决策支持系统的实际开发情况和项目实施周期，我们选择了“自上而下法”为主、“自下而上法”为辅的评估方法。

（4）定义数据质量指标。数据质量问题必须在源头得到修正，这是数据质量管理的一项基本原则[9-10]。定义数据质量指标需要在电动汽车决策支持系统的专题数据库设计阶段进行，随着项目的实施，可以对指标进行修正。数据质量的指标应该合理地反应数据质量维度所定义的数据质量特性，具体可包括可度量性、业务相关性、可接受程度、数据认责制度、数据管理制度、可控性、可跟踪性。相对于修正，预防的意义更大，

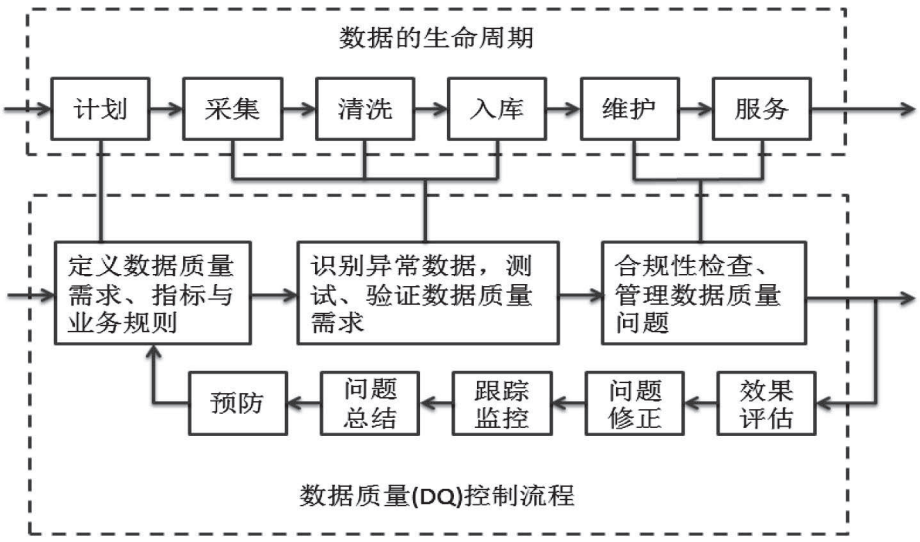

图2 数据质量管理流程图

表2 数据质量主要维度

序号	层级	维度	主要定义
1	一级维度	准确性	指数据准确反映其所建模的“真实世界”实体店程度。通常，度量数据值与一个已确定的正确信息参照源的一致性可以度量准确性
2		一致性	指确保一个数据集的数值与另一个数据集的数值一致
3		唯一性	主要体现在一个数据集中，没有实体多于一次出现，并且每个唯一实体有一个键值且该键值指向该实体
4	二级维度	完整性	完整性的要求之一是一个数据集的特定属性都被赋予了数值，或一个数据集的全部行记录都在
5		时效性	指信息反映其所建模的当前真实世界的程度
6		精确度	指数据元素的详细程度
7		隐私	指需要对数据进行访问控制和使用监控
8		合理性	使用数据合理性考察与一些特定的运营场景相关的数据一致性
9		及时性	指对信息可访问性和可用性的时间预期
10		有效性	指数据实例的存储、交换或展现的格式是否与数据值域一致，是否与其他相似的属性值一致

可以防止产生新的数据质量问题。

针对电动汽车多源信息决策支持系统定义数据质量指标的过程可概括为：选择决策支持系统中一项重要的业务模块；评估与业务模块相关的数据元素以及数据创建、更新流程；对于每一个数据元素，列出与之相关的数据需求；对于每一项数据需求，定义相关的数据质量维度以及一个或多个业务规则，以便确定数据是否满足需求；对每一个选中的业务规则，描述度量需求满足度的流程；定义可接受程度的阈值；根据上述定义出一系列度量流程，可提供原始数据质量的评分，并可将评分汇总并量化为数据质量需求的满足程度，未达到可接受程度阈值的度量说明不符合数据质量需求，需要采取必要的纠正措施。

（5）定义数据质量业务规则。需清晰定义检查数据质量是否满足业务流程并监控这些业务规则的符合度。需要将不满足业务需求的数据值与有效的数据值分别标记，及时向本项目数据管理员警示潜在的数据质量问题，并建立自动或时间驱动的缺陷数据纠正机制，以满足业务期望。制定并提供各方认可的规则模板，这有助于建立业务团队与技术团队之间的沟通，这部分规则模板也可转化成代码嵌入数据质量检查工具中。

（6）测试和验证数据质量。数据质量检测工具可分析数据状态并发现潜在异常数据，同时也可以使用项目组开发的数据质量检测工具对质量规则进行验证。在数据质量评估阶段识别或定义的规则，将作为业务流程的一部分用以验证数据的合规性。

（7）确定与评估数据质量服务水平机制。在数据质量服务水平协议中定义的日常数据质量管理内容包括：协议涉及的数据项范围；数据缺陷对相关业务的影响；与各数据项有关联关系的数据质量维度；系统对数据项的质量需求；针对数据质量进行的各种度量方法；各项测量的可接受阀值；当达不到质量要求时及时通知相关人员；问题上报机制。

（8）持续测量和监控数据质量。数据质量管理的操作流程取决于可用的数据质量测量和监控服务。对于数据质量是否符合业务规则，有两条控制和测量的方式，即随时测控与批量测控。同时，测量可应用于3种颗粒度，即数据值、数据实例、数据集。这样组成了6种可能的测量方案。针对本项目，对于网络动态信息库可采用随时测控方式，对于批量导入的永久存储的数据集中的数据可进行批量测控。

（9）管理数据质量问题。数据质量服务水平的有效实施需要建立数据质量事件解决报告与跟踪机制。具体的步骤涉及以下领域：将数据质量问题和活动标准化，制定数据问题的处理过程，管理问题上报程序，管理数据质量解决流程。

（10）清洗和校正数据质量缺陷。对于有质量问题的数据，需要进行两种活动：确定和消除错误发生的根本原因；清洗与校正不正确的数据项。清洗校正的方式主要有3种，即自动校正、人工指导校正、人工校正。

（11）设计并实施数据质量管理操作程序。采用预定义的规则进行数据质量验证，主要包括检查和监控、诊断和评估补救方法、解决问题、报告等程序。检查和监控：通过自动化或人工方式对全部数据进行扫描或抽检，从而测量数据集对数据质量规则的满足程度。诊断和评估补救方法：评审数据质量事件所反映的根源问题，跟踪错误数据的来源与用途，诊断问题的类型与起源。解决问题：数据质量团队应要求业务数据所有者选择多种解决方案中的一种。报告：为保证数据质量管理过程的透明度，应对过程的运行情况进行定期报告。数据质量运营团队负责开发和发布这些报告。

（12）监控数据质量管理操作程序和绩效。责任制是执行监控数据质量规范的关键。所有问题必须指定给专人、专门团队或组织负责，并以文件形式规定数据问题责任人。这对决策支持系统的长期运行维护意义重大。

5　结论

在DMBOK数据质量框架的指导下，规范了

（下转第30页）

6　结语

电动汽车产业数据库的建设在国际上仍然处于起步阶段，随着电动汽车产品应用的普及和产业发展的深化，建设电动汽车产业数据库，实现车辆信息资源和应用环境信息资源的数据化已成为一项重要的基础性工作。本文基于电动汽车产业链的发展特点，从技术和应用环境两大角度系统地构建了“VP+ISBCG”产业数据库框架，填补了我国电动汽车产业信息资源建设领域的空白。

在大数据时代背景下，随着数据分析与数据挖掘技术的进一步发展，产业数据库可以为不同领域的决策者或者用户如电动汽车用户、科研机构、政府等提供权威的信息服务，满足及时的、定制化的数据需求，既必要又可行，对我国目前的发展阶段而言，这既是挑战又是机遇，在参照国外同行数据库建设经验基础上，需要加强研究，以建设符合我国自身发展特色的电动汽车产业数据库。

参考文献

[1] 李跃明.数据库系统的创新发展[J].电脑知识与技术,2011,7(2):269−270,273.
[2] 向海华.数据库技术发展综述[J].现代情报,2003(12):31−33.
[3] 叶瑞克,陈秀妙,朱方思宇,等.“电动汽车－车联网”商业模式研究[J].北京理工大学学报:社会科学版,2012,14(6):39−44.
[4] 许海洋.电动汽车产业链格局探究[J].汽车工业研究,2013(11):22−26.
[5] 隋忠海,王震坡.对我国电动汽车产业链发展模式的思考[J].企业研究,2011(4):58−61.
[6] 郭理桥.建设行业多层次数据库建设的思考[J].中国建设信息,2009(22):6−9.

（上接第24页）

针对电动汽车多源信息决策支持系统完善的数据质量管理流程，对系统内的数据质量实施全过程、全领域管理，将数据质量管理以制度化、规范化的方式落实到数据生成、传递和使用的各个过程之中。相对以往的数据质量管理流程，它细化了管理流程，更加注重对研究人员数据质量意识的培养，要求提高系统内各数据库建设人员的协同配合。按此流程实施后，在一定程度上做到了数据质量从源头开始控制，优化了数据字段设置，提高了数据采集加工入库的速度，系统内数据能够做到可检索、可统计、可展示，为电动汽车多源信息决策支持系统提供有力的数据支撑保障。

参考文献

[1] 商广娟.有效的数据质量管理体系——21世纪管理的基石[J].航空标准化与质量,2005(2):18−22.
[2] 常宁.IMF的数据质量评估框架及启示[J].统计研究,2004(1):27−30.
[3] 侯瑜.基于DQAF框架的我国统计数据质量管理及改进[J].统计研究,2012(12):24−30.
[4] 宋敏,覃正.国外数据质量管理研究综述[J].情报杂志,2007(2):7−9.
[5] 韩京宇,徐立臻,董逸生.数据质量研究综述[J].计算机科学,2008(2):1−5, 12.
[6] 郭志懋,周傲英.数据质量和数据清洗研究综述[J].软件学报,2002(11):2076−2082.
[7] 叶宏伟,金中仁,陈振宇.图书馆信息集群研究[J].中国图书馆学报,2008(1):99−101.
[8] DAMA International. DAMA数据管理知识体系指南[M].北京:清华大学出版社,2012.
[9] 孙中东.企业级数据治理框架下的数据质量管理[J].金融电子化,2011(6):57−60, 6.
[10] 谷斌.信息系统建设中的数据质量管理体系研究[J].情报杂志,2007(5):65−67.