中国科学: 信息科学 2014 年 第 44 卷 第 7 期: 836–856
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
双清论坛大数据专题 · 论文

中国公共数据库数据质量控制模型体系及实证 ‡

刘文奇

昆明理工大学理学院, 昆明 650500
E-mail: liuwenq2215@sina.com
‡ 特约编辑: 柴天佑, 张军, 王成红

收稿日期: 2013–08–09; 接受日期: 2013–09–23
科技部科技型中小企业技术创新基金 (批准号: 11C26215305906) 资助项目

摘要 在讨论公共数据库的公共产品属性及其制造过程的特点基础上, 确定了公共数据库中数据质量的维度, 提出了公共数据库作为特殊公共产品的生产过程中非技术数据清洗和数据稽查的概念, 并建立了数据质量的变权综合评估模型. 对公共数据库的数据制造过程质量控制进行了技术设计和博弈论模型分析, 给出了各个环节的数据质量控制措施, 并以此分析了当前中国公共数据库建设和运行中存在的问题和政策建议. 最后, 作为实例, 提供了质量技术监督公共数据库的设计、运行、数据制造过程质量控制方法.

关键词 公共数据库 数据质量 数据清洗 数据制造过程控制 数据稽查 博弈分析 Mapreduce 模型

1 引言

社会经济系统是复杂的自适应系统, 系统与环境之间、子系统与子系统之间、要素与要素之间广泛存在着物质流、能量流和信息流 [1]. 随着信息时代的来临和信息技术的发展, 社会经济系统内部的信息流越来越引起人们的重视. 信息和信息流的研究已经对经济学和管理科学研究产生深刻影响, 甚至彻底改变了相关领域研究的面貌. 博弈论与信息经济学的兴起正反映了这种深刻的改变 [2]. 数据作为信息和知识的载体, 对信息时代的社会经济系统的运行和控制起着基础性作用, 从根本上影响决策的信息基础和科学性, 进而影响社会经济生活的方方面面. 伴随着数据和信息的流动, 在市场和政策的作用下, 资源得以有效配置. 甚至信息及其流动机制通过公共政策执行在资源配置中起到决定性作用. 更有甚者, 在信息网络如此发达的环境下, 以公共数据库为基础的信息流动已经引起政治变革.

公共数据库是信息公开和信息自由的基础, 是信息时代的必然产物和社会民主的信息基础, 也是公共政策与决策的支撑要件. 当今世界各国把公共数据库的建设当成公共基础设施建设的主要内容之一, 其重要性不亚于国家高速公路网、通讯网、电网、高铁网的建设. 公共数据库是指具有社会管理功能且基于计算机系统与网络的数据库, 除非特别申明，本文中是指 Web 公共数据库.

有的学者甚至认为, 数据的生命力比软件更持久. 理由是程序可以不停地升级换代乃至退出, 但保存数据的数据库却会继续存在, 其价值很可能与日俱增、日久弥新. 万维网之父 Tim Berners-Lee 就认为, 数据是宝贵的, 它的生命力比收集和处理它的软件系统还要持久 [3]. 美国软件开源运动领袖 Eric

引用格式: 刘文奇. 中国公共数据库数据质量控制模型体系及实证. 中国科学: 信息科学, 2014, 44: 836–856, doi: 10.1360/N112013-00136

Raymond 也说, “一个好的数据结构和一个糟糕的代码, 比一个糟糕的数据结构和好的代码要强多了”. 可见公共数据库比其上的应用软件系统重要得多.

美国是拥有公共数据库最多的国家. 截止 2011 年, 美国联邦政府已经拥有 1 万多个独立的公共数据库, 分属于 2094 所数据中心, 管理着庞大的联邦数据资产, 每年直接的运行费用仅联邦预算就高达 784 亿美元 [3]. 以美国商务部下属的美国普查局 (USCB) 数据库为例, 它目前拥有 2560 TB, 即 2560×2^{40} 字节, 它的大小已经大大超出人类的直接感知能力. 美国国家安全局 (NSA) 和美国中央情报局 (FBI) 都拥有超级巨大的数据库. 据时任 Rand 公司情报政策研究中心主任 John Parachini 接受《巴尔的摩太阳报》记者采访时所说, 美国国家安全局是从数据库保留的电话监控记录中发现了本 · 拉登的蛛丝马迹而将其一举击毙的. 该局对全美电话进行监控, 所收集的数据量是惊人的, 每 6 小时产生的数据量就相当于美国国会图书馆印刷体藏书的信息总量. 而美国国会图书馆是世界上馆藏量最大的图书馆. 再说美国中央情报局, 其本职工作就是收集情报信息, 业内专家普遍认为它可能拥有全世界最大的数据库.

我国公共数据库建设相对滞后, 运行效率比较低下, 而且缺乏有效的数据质量控制机制, 导致数据质量较差, 从而影响了可用性. 随着我国社会经济制度市场化改革的深化和计算机网络的普及, 已经陆续建成、在建和即将建设一些超大规模的公共数据库, 如个人身份信息数据库、组织机构代码数据库、个人医疗保险数据库、社会保险数据库、房屋产权登记数据库、财政税务数据库、中国人民银行个人征信数据库等. 尽管有些数据库尚未实现全国联网, 但由于人口基数大, 经济规模迅速扩大, 使得我国公共数据库建设之初就遇到数据规模庞大、网络结构复杂、法制不健全和社会环境变迁等因素带来的数据管理难题. 可以说, 从全世界范围看, 我国公共数据库建设和运行管理的困难和挑战都是前所未有的, 其中包括技术困难和制度设计困难. 例如, 由于人口基数的原因, 以个人身份信息数据库为基础的国家福利和个人经济活动记录 (医保、社保、信贷、通讯、交通等) 的公共数据库规模就是美国的 5 倍, 而数据库的技术和管理难度随数据量的增加呈现非线性增长.

公共数据库与一般的商用数据库不同, 它往往由政府主导建设和运行的, 其本身具有公共产品的属性. 公共产品的供给和使用是一个囚徒困境问题 [2], 公共数据库也不例外. 政府、企业和公众都提供真实有效的数据来建设公共数据库并保证公共数据库的良好运行, 政府的公共政策制定就会更透明更科学, 企业和公众的权利和利益会得到更有效的保护. 问题是, 如果政府积极组织公共数据库的建设和运行, 并及时、有效、真实地把政府业务流程数据提供给公众, 但企业和公众不按法规和政策提供真实有效数据, 政府又拿不出有效的行政措施进行处置, 那么政府会处于政绩失败的境地并承担财政损失. 同样, 企业和公众按要求提供了完整、有效、真实、及时的数据, 但政府不按政务公开原则让企业和公众通过数据库查询政府业务流程并有效履行隐私保护, 那么企业和公众因为承担过多成本和隐私泄露风险而没有配合政府维护公共数据库的运行的积极性, 并最终导致公共数据库名存实亡. 更糟糕的是, 政府和既得利益者沆瀣一气危害公众利益. 所以, 政府、企业和公众都不愿意投入公共数据库建设和运行, 或消极地给公共数据库提供虚假信息、不完整信息或审核不严. 这样就导致了公共数据库建不起来、运行失效和数据质量低下, 最终形同虚设. 这也是我国各类公共数据库大多未能有效运行的根本原因所在. 解决这样的问题需要进行一系列关于公共数据库建设与运行的机制设计.

经济学认为, 由于公共产品或服务的特性而导致的偏好显示及搭便车问题, 政府提供公共产品比市场提供具有更高的效率. 因此, 我国政府长期以来普遍采用纵向一体化的方式, 垄断了公共数据库的供给和需求. 最近也有一些学者认为: 政府垄断地提供公共产品, 不仅事实上存在着成本高、质量低、不公平等问题, 而且造成某种 “强制消费” 现象, 即国民在获得公共产品方面缺乏必要的选择权.

而且基于公共选择理论、新公共管理理论的公共产品供给的市场化主张逐渐为许多国家所接受 [4]. 例如, 美国等发达市场经济国家已经尝试行业协会、民间智库和跨国公司围绕某些公共领域建立和运行相对独立的市场化的数据库. 这些数据库具备一定的公共数据库功能, 必要时政府可以征用其中的数据. 例如, 企业或行业协会自主建立的产品溯源系统、行业标准管理系统等. 这些数据库已经具备公共产品的某些属性, 属于准公共产品 [5]. 在我国, 学术界和技术管理机构已经就一般公共产品供给的市场化进行了研究并取得了一些理论和实践成果 [4]. 随着我国改革开放的进一步深入, 有必要在兼顾效率与公平、控制与监督原则下, 探讨推进我国公共数据库建设与运行的市场化机制. 但是, 在短期内公共数据库作为特殊公共产品, 其供给的政府主导方式很难改变.

由是观之, 我国公共数据库的建设、运行管理和数据质量控制已经成为管理科学界急待研究的重点领域之一.

2 公共数据库中数据质量的定义及维度分析

2.1 公共数据库的公共产品属性和数据质量的内涵

数据是信息的载体, 好的数据质量是各种数据分析 (如 OLAP 分析、数据挖掘等) 能够得到有意义结果的基本条件. 反之, 如果数据质量得不到保障, 再先进的数据库系统、搜索引擎和决策支持系统也无济于事. 人们常常抱怨所谓的 “数据丰富, 信息贫乏”, 其原因有二: 一是缺乏有效的数据分析技术; 二是数据质量不高如数据残缺不全、数据不一致、数据重复等会直接导致数据不能有效地被利用. 公共数据库中的数据质量管理如同其他公共产品质量管理一样贯穿于数据生命周期的各个阶段, 但目前尚缺乏适合我国公共数据库数据质量管理的系统思路.

各级政府作为公共数据库的主体责任者, 对公共数据库中的数据质量控制起到至关重要的作用. 对此国外政府颁布了一些政府层面的数据质量控制措施, 来保证数据收集、使用、发布过程中数据的客观性、实用性、完整性、时效性、数据管理流程的科学性和数据救助机制的可行性. 例如, 1980 年美国国会要求联邦政府行政预算管理局 (OMB) 制定公共数据库中数据质量控制的具体措施. 为此，OMB 制定了数据质量控制的指导原则 [3]:

(1) 数据质量标准: 政府各部门必须保证数据的真实性、实用型、完整性和时效性.

(2) 科学的数据质量管理流程: 政府各部门必须针对数据质量, 完善信息管理的流程, 防止低质量的数据出现.

(3) 完善的数据质量救助机制: 政府各部门必须建立一个行政机制来应对公众对数据质量的质疑和挑战. 如果政府发布的数据质量确实存在问题, 必须有相应的纠错机制来补救.

我国政府也在《中华人民共和国统计法》和《食品安全法》等法律条文中对企业和政府部门提供数据的必要性和真实性作了定性的要求, 但没有制定过特别针对数据质量的专门法规.

在学术界的许多文献中, 数据质量 DQ(data quality) 与信息质量 IQ(information quality)2 个术语通用, 定义多种多样. 其中, 比较流行的是 Wang 等 [6] 提出的定义, 即将数据质量定义为数据使用的适合性. 此定义的基础是当时全面质量管理中广泛接受的质量概念, 因此关于数据质量的这个定义迅速而广泛地被学术界接受. 按此定义, 数据质量判断依赖于数据使用者, 即数据用户, 不同环境下不同数据用户使用的适合性不同. 故数据质量概念是相对的, 不能独立于数据用户来评价数据质量. 据此, 识别数据质量维度成为一项有价值的研究工作. 文献 [6] 采取二阶段调查方法识别出 4 类共 15 个数据质量维度.

固有质量包括: 正确性, 客观性, 可信性, 声誉;

可访问性质量包括: 可访问性, 访问安全;

语境质量包括: 相关性, 增值性, 及时性, 全面性, 数据量;

表达质量包括: 可解释性, 易理解性, 简明性, 一致性.

文献 [6] 所采用的统计方法中选择的调查对象是具有工业背景的人士、MBA 学生和普通数据消费者, 第一阶段设计的 179 个初始数据质量属性指标中的多数指标明显倾向数据消费者对商业数据质量的关注. 因此, 可以判断文献 [6] 关注的是商用数据库或企业数据库中的数据质量问题. 另一方面, 由于当时网络还不发达, 导致可访问性受到数据用户的高度关注, 显然在高速网络环境里的 Web 数据库中, 可访问性不是数据质量的维度而是权限设置范畴, 甚至是可忽略的. 公共数据库与商用数据库不同，其使用者往往不是以纯粹的消费为目的, 与文献 [6] 中的数据消费者内涵不同, 而且数据用户使用公共数据库时没有多余的选择, 因此声誉和增值性也不是公共数据库中数据质量维度.

之后, 针对 Web 数据库，多位学者增加了网络环境下的数据质量维度, 包括: 理解, 正确, 清晰, 适用, 简明, 一致, 恰当, 流通, 方便, 适时, 可追溯, 交互, 可访问, 安全, 可维护, 快捷 [7~13]. 这些维度较好地体现了数据库中数据质量网络应用层面的内涵, 但还是不能很好地涵盖公共数据库中数据质量的公共产品的质量内涵.

若不考虑数据用户的不同需求, 仅仅针对数据的固有质量, 即狭义的数据质量, 则可以从数据的制造过程来看数据质量. 利用数据库生产过程和普通产品制造过程的相似性, 建立起数据产品与物质产品的联系. 原始数据对应原材料, 数据加工对应材料加工, 数据产品对应物质产品. 这样, 全面质量管理 (TQM) 的原则、方法、指南和技术就可以用于数据质量管理. 因此, 在数据产品制造过程中有 4 种角色: 数据提供者, 数据生产者, 数据消费者, 数据管理者. 文献 [14] 给出的数据制造系统模型, 通过建立表达数据单元和系统构件关联关系的数据制造系统分析矩阵, 系统地追踪数据产品相关属性, 这些属性的测量值可以用于数据制造系统的改进. 但是此文献中的研究对象也是商用数据制造, 反应了商用数据制造过程质量控制原理. 也就是说, 把商用数据库中数据的质量问题视为普通商品生产过程质量控制问题.

公共数据库中数据制造的过程与一般公共产品制造过程有相似性, 有别于普通商品的制造过程. 在运行环境上又与商用 Web 数据制造系统较为一致, 与当今物流环境下普通商品全球供应链制造模式具有相似性. 因此, 把公共数据库中数据质量视为公共产品制造过程质量控制的结果是一个科学的思路.

对于狭义的数据质量, 也可以从另一角度来解释. 在一个或多个数据库中, 同一个现实对象可能具有多种描述方法. 因此, 数据质量可以用数据和其对应实体的 “完美表达” 间的差距来衡量. 实体识别在数据质量管理中起着重要作用, 是数据质量管理的主流研究方向之一. 实体识别的目的是在一个或多个数据库中辨识描述同一个实体的不同表示方法, 正确地识别出数据库中的所有不同实体. 实体识别的结果是数据库中所有不同实体的集合以及每个实体的不同描述方法. 实体识别的结果可以在数据质量管理的其他阶段得到广泛应用, 如冗余数据去重、错误数据发现、不一致数据发现与冲突消解等. 在不同的文献中, 实体识别有着不同的名称, 包括对象识别、冗余发现、实体消解等 [13].

在公共数据库中数据制造过程的各个关键环节产生的中间产品 (中间数据) 采用实体识别技术进行数据清洗, 来达到关键过程质量控制目的是一种值得推荐的方法.

公共数据库是具有公共产品性质的特殊的数据库, 其数据质量也有特殊的维度, 专门针对公共数据库中数据质量的研究成果尚未见文献报道.

2.2 公共数据库的特性

我们可以从公共数据库这一公共产品的提供者、使用者和管理者的角色分析与数据生产过程来阐述公共数据库的数据质量与数据质量控制. 公共数据库的用户角色和数据制造关键过程控制的特性表现在以下几个方面:

(1) 公共数据库参与者具有角色多重性

公共数据库的数据源提供者、数据的生产者、数据的管理者和消费者往往有多重身份, 特别现在 Web 公共数据库越来越占主流的情况下, 更增加了角色的多重性. 例如, 大多数公共数据库的责任主体是公共部门 (以各级政府部门为主), 他们在监管和决策过程中是数据的使用者, 在业务审批流程中又是数据的提供者, 在数据管理法规的制定中又是数据的管理者; 企业在被监管的过程中是数据的提供者, 在数据知情权下又是数据的使用者; 公众作为传统的公共数据库的数据消费者已经逐步进入参与角色, 通过举报、投诉、发帖等形式为公共数据库提供数据, 由此触发公共部门处理举报、投诉过程中的业务流程数据再造. 与商业数据库不同, 公共数据库的数据用户不等同于商业数据消费者和其他公共产品使用者.

(2) 公共数据库必须体现公平和效率的原则

公共数据库作为公共产品, 无论何种用户都必须秉持在公正与合法的前提下保护用户的数据使用权益和履行数据提供者的法律义务. 这一点和普通商用数据库有着本质性的差别. 一方面要打破数据垄断, 另一方面要通过法制和市场手段保护数据提供者和生产者的权益. 如何实现二者之间的平衡是一个值得研究的课题.

(3) 公共数据库最小数据集确认和数据成本听证制度

出于公共事务的处理和政府信息公开需求, 公共数据库的部分数据来源具有强制收集法规支撑, 在原始数据收集过程中必然产生社会成本. 同时, 原始数据收集过程中也必须遵从隐私保护原则, 藉此保护公民个人隐私和企业的商业机密、独有配方和工艺等. 因此面向数据源提供者强制收集的数据范围和种类必须尽可能小. 对不同的数据源强制收集的数据指标之间尽量不重合. 也就是说, 公共数据库中强制性数据收集必须遵从数据最小原则. 最小数据集 (MDS) 的概念起源于美国的医疗领域 [3]. 1973 年, 在美国国家生命健康统计委员会 (NCVHS) 的主导下, 为了规范出院病人的信息收集工作, 美国第一次制定了统一的出院病人最小数据集. 既然是出院, 核心环节就是付钱, 所以这些数据又被用于创建统一的医疗账单 (UB), 成立了国家统一账单委员会, 并于 1982 年统一制定了 UB-82 的数据格式, 1992 年又升级到 UB-92, 且扩大应用到医疗保险和索赔领域. 由于其实用性, 最小数据集的概念在美国已经演变成一个一般概念, 它指代国家管理层面针对某个业务管理领域强制收集的数据指标. 不少领域的最小数据集甚至被上升到立法的高度, 对公共数据库建立和运行起到至关重要的作用.

我国各类公共数据库的建设基本属于起步阶段, 公共数据库的建设处于混乱状态, 其中重要的原因包括: (a) 信息收集和公开的立法滞后, 主要以条例或部门规定存在, 确定最小数据集缺乏法律依据; (b) 数据收集和报送制度的缺失; (c) 最小数据集研究薄弱, 最小数据集设定不合理, 缺乏科学性; (d) 部门分割造成数据孤岛, 很多数据重复收集和生产. 据我们在 CNKI 网络查询结果看, 研究我国各领域公共数据库最小数据集的工作还很少, 更没有研究不同公共数据库的数据共享机制的文献报道.

另一方面, 由于公共数据库属于公共产品, 它的收集、加工成本和有偿使用价格都必须经过听证. 从博弈论的角度理解, 最小数据集设置不合理会导致数据收集成本过高或侵犯隐私权, 数据提供者必然会采用隐瞒、错报、假报的策略来规避成本和保护隐私, 进而影响整个公共数据库的质量, 也加大数据清洗的成本和数据稽查的难度.

(4) 公共数据库的数据清洗具有强制性

数据清洗 (datacleaning, data cleansing 或者 data scrubbing) 是指检测数据中存在的错误和不一致, 剔除或者改正它们, 以此提高数据的质量 [15,16]. 在单个数据源中可能存在质量问题. 例如, 某个字段是一个自由格式的字符串类型, 比如地址信息、电话号码等; 错误的字段值, 由于录入错误或者其他原因, 数据库中一个人的年龄为 485 等; 数量字段单位不统一造成的数据错误; 主线名称不统一造成的项下数据重复, 如 "× × × 有限责任公司" 和 "× × × 有限公司" 本来是同一家公司, 但名称不规范导致项下的产品信息和设备信息重复. 考虑多个数据源的情形, 比如数据仓库系统、联邦数据库系统, 或者是基于 Web 的公共数据库系统, 问题更加复杂. 来自不同数据源的数据, 对同一个概念有不同的表示方法, 同一数据实体在不同的数据源中表述不同. 比如某一家生产企业项下的数据包含了组织机构代码库下的注册信息、工商注册库下的注册信息、生产许可证库下的行政许可信息、特种设备库下的设备使用登记信息等. 在集成多个数据源时, 需要消解模式冲突, 主要就是为了解决这个问题. 还有相似重复记录的问题, 需要检测出并且合并这些记录. 在网络环境中的公共数据库大多数是 Web 数据仓库模式, 加上公共数据库的公共产品属性, 使得公共数据库中的数据清洗更为复杂. 比如信息博弈带来的数据造假问题, 以及假数据的识别和数据造假者的行政处罚措施等. 大多数商用数据库中的数据绝大多数来自自动采集装置, 如沃尔玛的销售记录主要来自条码、二维码和 RFID 识别系统. 而公共数据库的数据源提供者往往是具备社会属性的组织或个人, 本身具有主观性和自身利益, 采集的数据中大部分靠人工采集或自主申报. 这种方式产生的数据可能数据格式和逻辑上并不存在质量问题, 但本质上真假难辨. 此类数据的数据清洗不能靠单纯的计算机技术手段而更多地靠数据稽查手段来处理. 数据稽查是指官方或其代理人依据法规和国家力量, 通过数据源提供的数据与原始凭证和真实情况作对比来鉴定数据真实性的过程. 数据稽查的主要特性是官方强制性和实务性.

因此, 对公共数据库而言, 数据清洗应该分为技术性数据清洗和非技术数据清洗 2 类. 迄今为止, 国内外绝大多数数据清洗文献论及的是技术性数据清洗 [10~19]. 技术性数据清洗的目标是最大限度地满足数据的一致性 (consistency)、正确性 (correctness)、完整性 (completeness) 和最小性 (minimality). 非技术性数据清洗的目的是最大限度内满足数据的真实性 (truth)、及时性 (timeliness)、权威性 (authority) 和隐私保护 (privacy).

公共数据库的设计、建设和维护往往是以法律为基础的. 因此, 公共数据库中数据制造关键过程控制中有强制性手段, 比如数据造假者除了可能付出经济代价外还要负刑事、民事或行政处罚责任. 基于这一特点, 公共数据库数据质量控制比商用数据库数据质量控制和普通产品 (非公共产品) 制造的质量控制多一些司法和行政等非市场手段. 在我国, 政府部门对公共数据制造过程的强势介入更为普遍, 行政处罚措施种类繁多. 一方面会加大数据制造成本, 另一方面, 如果使用得当, 也可以提高非技术数据清洗的效率.

(5) 公共数据库的数据制造过程的多阶段性

公共数据库的建设和运行的主体责任者是政府, 而各级政府在数据制造过程中的管理职能和对数据使用的需求也不同. 各级政府以不同的规则和质量维度对下一级数据源进行质量考核、评估和数据清洗. 例如, 基层政府数据管理部门负责对企业的日常监管, 他们评估的数据主要来自辖区内被监管或服务的企业上报的数据和本级业务人员填报的业务台账数据. 数据的真实性和录入的及时性是基层数据管理者首要关注的维度. 在数据清洗过程中伴随着大量的数据稽查和行政执法行为. 而地市一级政府数据管理部门很少开展对企业的具体数据稽查工作, 而是根据举报投诉等信息进行数据抽样并对下级政府发出稽查指令或进行数据管理考核. 省一级政府数据管理部门关注的是横向和纵向的数据实体

识别和决策支持系统应用, 基本不接触企业的具体数据真实性. 中央级数据管理部门则更多地关注数据法规政策制定和督察 (顶层设计). 不同级别的政府数据管理部门处于数据制造过程的不同环节, 数据质量的维度不同, 数据质量控制的方法也有区别, 非技术数据清洗的手段也不同.

2.3 公共数据库中数据质量的维度和数据制造过程质量控制

综上所述, 我们把公共数据库视为一个特殊的公共产品, 技术上体现为大型 Web 数据仓库系统模式. 我们将公共数据库的建设和运行视为公共产品的制造过程, 公共数据库中数据的质量是数据制造过程质量控制的结果. 数据制造过程质量控制也是一个完整的数据质量溯源系统, 与普通产品质量溯源系统 [16] 极为相似. 不同的只是数据的迁移对数据质量的影响可以忽略不计. 数据制造过程质量控制的手段包括 Web 数据仓库设计、最小数据集确定、数据成本控制、数据听证制度设计和数据清洗, 数据清洗包括技术性数据清洗和非技术数据清洗. 每个数据制造的关键环节的数据质量维度有所不同, 处于数据源阶段的数据质量维度是一致性、正确性、完整性、最小性、真实性、及时性、权威性和隐私保护.

3 公共数据库的网络模型与数据清洗机制设计

公共数据库的用户多重角色和数据制造关键过程控制的上述 5 个特性决定了其设计有别于一般商用数据库的设计要求. 由于数据用户角色的转换和数据清洗的强制性, 使得各种数据用户都在信息公开和隐私保护之间寻求某种平衡, 形成复杂的博弈 [2]. 这种数据用户之间的互相制衡给公共数据库设计和非技术性数据清洗机制设计带来极大困难. 迄今为止, 中国公共数据库的一般设计原理和非技术性清洗机制设计尚未见诸国内外文献.

3.1 公共数据库的网络数据仓库模型

由于公用性, 公共数据库用户多、数据类型多、数据来源广、用户权限复杂等特点, 公共数据库宜采用当今流行的多层 Web 数据仓库模式进行设计. 这样设计的优点在于可以适当按数据需求和数据来源分解各类用户和数据源到不同的 Agent, 这种分解可以是多层次的. 本质上讲就是将数据和应用进行类别和级别划分, 将数据仓库及其中数据视为数据制造过程的最终产品, 不同的数据和用户层次对应于各类数据原料和不同的数据制造关键环节, 从而减轻核心数据仓库的压力和控制数据仓库中的数据质量. 同时, 这样的的设计可以在数据收集和处理中充分保护数据提供者的隐私权.

近几年数据仓库又成为数据管理研究的热点. 按主体架构解决方案的不同, 流行的数据仓库分为并行数据库、Mapreduce 和混合架构 3 种类型 [20]. 这 3 种类型的数据仓库架构各有特点. 在实际的公共数据库的架构选择中, 应该根据数据分析结果和未来数据挖掘的需要进一步选择符合其自身特点的数据仓库架构. 文献 [20] 已经对上述 3 种类型及其亚类的数据仓库的优劣做了全面的比较分析.

Mapreduce 是 2004 年由 Google 公司提出的面向大数据集处理的编程模型, 起初主要用于互联网数据的处理, 例如文档抓取、倒排索引的建立等. 但由于其简单而强大的数据处理接口和对大规模并行执行、容错及负载均衡等实现细节的隐藏, 该技术一经推出便迅速在机器学习、数据挖掘、数据分析等领域得到广泛应用 [20~22]. 该模型是面向有数千台中低端、异构服务器组成的大规模机群而设计的, 建构成本远低于并行数据库. 基于 Mapreduce 模型的分析无需复杂的数据预处理和写入数据库的过程, 而是直接基于平面文件进行分析, 设计的初衷是面向非结构化的数据处理. 该模型采用的计算

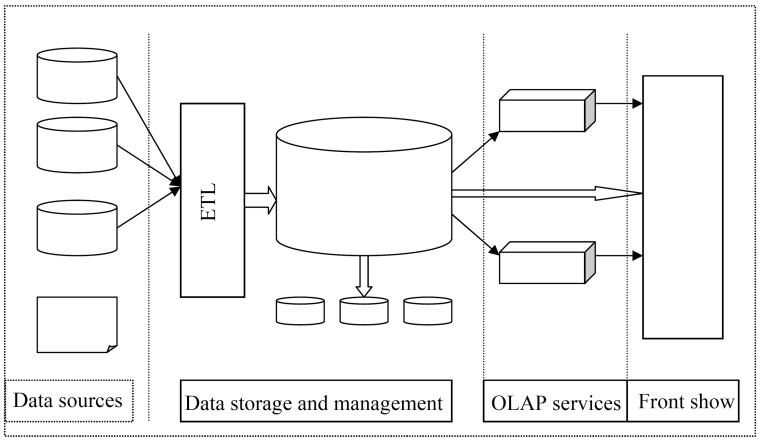

图 1 公共数据库数据仓库模型

Figure 1 Data warehouse model of public database

模式是移动计算而非移动数据, 可以使分析延迟最小化. Mapreduce 模型的开源实现以 Hadoop 平台最为成熟. 在此平台下, Mapreduce 模型卓越的扩展能力已经被 Google, Facebook, Baidu, Taobao 等商用业务充分验证. 更重要的是, 作为开源系统, Mapreduce 具有完全的开放性, 其 ⟨key,value⟩ 存储模式具有较强的表现力, 可以存储任意格式的数据. 同时, Mapreduce 模型的 Map 和 Reduce 两个函数接口给用户提供了足够的发挥空间, 从而实现各种复杂的数据处理功能. 为了弥补开放性带来的数据库系统处理 BI 报表分析等能力的不足, HadoopDB 平台下的 Mapreduce 模型能较好地融合调度层、沟通层和关系数据库, 尽可能将数据查询推入数据库层处理.

公共数据库的多层 Web 数据仓库模型图 (图 1).

3.2 公共数据库的数据清洗机制设计

3.2.1 公共数据库的技术性数据清洗技术设计

从技术层面来讲, 技术性数据清洗是通过利用应用程序按照一定规则从数据源数据库自动提取数据实例 (instance) 来提高数据质量过程. 在公共数据库的数据仓库系统前端设计前置交换系统, 用于数据的自动检测和数据的反馈. 自动检测的事项包括: 重复对象检测、缺失数据处理、异常数据检测、逻辑错误检测、不一致数据处理和数据转换.

同源数据之间和其他数据源数据之间具有一定的业务逻辑. 数据清洗应满足这些业务逻辑的要求, 对违反业务逻辑的数据必须予以处置. 被过滤的数据称为不合格数据, 经过过滤后的剩余数据称为合格数据.

数据清洗应满足:

1) 保证部门报送的信息类有关键字属性 (primary key);

2) 尽量放松清洗规则, 保证数据的原样性;

3) 在清洗过程中, 只能作数据映射, 不能修改用户数据, 不对错误数据进行纠正;

4) 清洗的数据要保存.

数据转换是将数据的格式、形式统一化的过程. 首先要对数据的格式、表示形式等方面有统一的规则, 再按照转换规则将各数据源符合规则的数据进行转换, 转换的原则是不能改变数据本身所表示的意义.

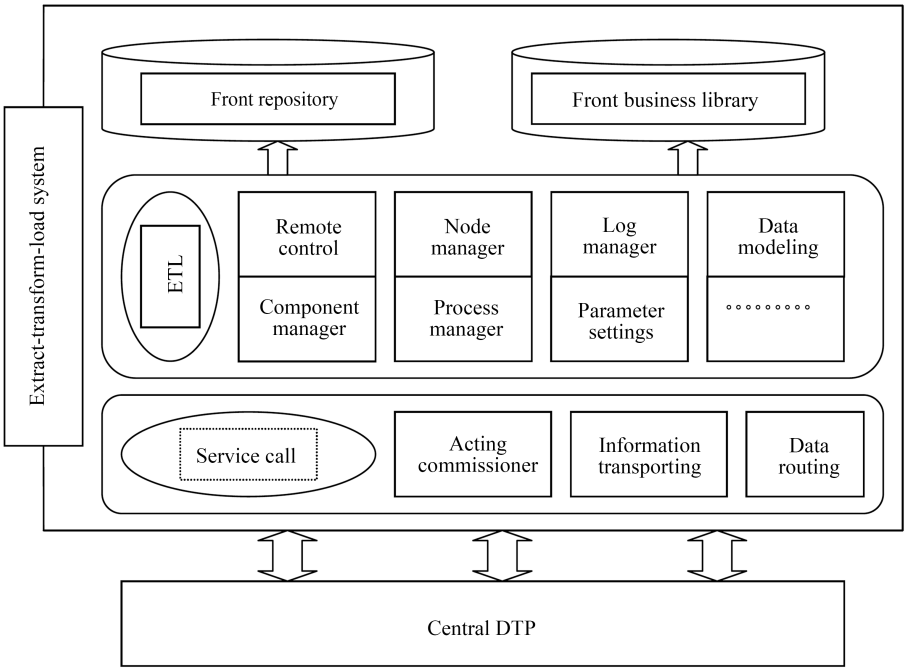

图 2　前置交换系统结构

Figure 2　Structure of ETL (extraction-transformation-loading) system

为了自动实现数据清洗过程所设计的应用软件统称为前置交换系统, 包括服务调用、ETL 工具和数据源数据副本库 3 部分. 其结构如图 2.

3.2.2　公共数据库非技术性数据清洗的博弈论模型

公共数据库非技术数据清洗的核心是人工数据核实和数据打假, 统称为数据稽查. 很显然, 不可能由计算机系统的网络设计和算法设计来实现. 被监管的企业、个人和下级政府部门在提供数据的过程中因为各种利益驱动会存在数据造假、瞒报等动机. 尽管上级数据管理和审核部门有法律和行政措施等手段对数据源提供者的假报、瞒报行为进行处罚的权力, 但也受行政成本的限制与面临行政复议、行政诉讼和侵犯隐私权诉讼的风险, 因而存在核实与否和处罚与否的两难选择.

在公共数据库中的相当大的部分是靠手工采集而不是利用传感器等自动采集装置记录的, 这导致数据造假和瞒报的几率大为增加, 数据的核实更加困难.

迄今为止, 数据稽查 (data checking) 没有统一的定义. 结合辞海的解释, 我们可以将数据稽查定义为: 通过与原始凭证及旁证进行比对确认上报数据是否与账物一致的过程. 在实践中, 数据稽查一直存在, 例如税务稽查、财务审计、特种设备安全监察、产品质量的国抽和省抽等, 都属于数据稽查范围. 在信息时代数据稽查更加重要, 数据稽查的范围更加广泛, 需要进一步加以诠释. 这是一个值得司法界、公共管理学界关注的课题.

数据的报送与稽查是一个信息不完全的动态博弈过程. 文献 [23] 对统计数据质量可靠性进行了博弈分析. 尽管我国有《统计法》, 但是统计数据造假只占公共数据造假的很少的部分, 而且企业的违法动机不是很强烈. 在其他公共数据库建设和运行中危及公共安全的数据造假更为普遍. 可见, 公共数据库中数据造假和瞒报的博弈分析研究还刚刚起步.

为了便于分析博弈均衡解的存在性, 可以将公共数据库的数据制造质量控制过程中广泛存在的数

表 1 政府 – 企业 (个人) 博弈

Table 1 Government-to-enterprise (personal) game

		Enterprice (Personal)	
		Cheating	No cheating
Government	Checking	a, b	$d, 0$
	No checking	$0, c$	0,0,

据稽查多方博弈分解为 3 个简单静态博弈, 即下级政府部门 – 上级政府部门、政府部门 – 企业 (个人)、委托人 – 中介部门 (技术机构). 这里的中介机构是指数据稽查中涉及的专业知识提供方, 比如说认证公司、评估机构、审计事务所、特种设备检定单位、产品质量检验机构等, 它们在专业性数据稽查中提供技术标准. 同时, 做如下假设:

(1) 局中人都是理性的;

(2) 报送的数据是通过技术性数据清洗合格的, 即非技术性数据清洗不含数据格式核查;

(3) 数据稽查者进行数据核查的条件为 (a) 接到举报或投诉 (b) 随机抽查 (c) 例行巡查.

(4) 局中人都具备本职工作所需的专业知识, 自动收集的数据都是准确无误的.

也就是说数据的造假和瞒报都是数据源提供者故意造成, 而非过失或意外事件造成的.

博弈分析一: 政府 – 企业 (个人) 博弈

为了获得模型 Nash 均衡解存在的条件, 我们将企业 (个人) 的瞒报归入数据做假, 同时假设第三方提供的技术报告和举报投诉信息是真实可靠的且企业 (个人) 在稽查过程中没有足够时间修改数据. 局中人, 行动及支付水平由表 1 给出. 模型参数分析如下:

1) $b < 0, d < 0, a > d$. 如果政府进行稽查, 而且企业 (个人) 提供假数据, 那么企业 (个人) 将会被罚款或被吊销许可证等; 如果政府进行稽查, 但企业提供的数据真实, 则政府将付出稽查支出 $-d$.

2) $c > 0$. 企业 (个人) 之所以做假, 是因为如果政府不稽查, 那么企业 (个人) 可以通过提供假数据获得收益.

3) 若 $a < 0$ 且 $c > 0$, 则博弈的 Nash 均衡解为 (不稽查, 做假); 此时, 政府通过稽查带来的好处不足以抵消执法成本或政绩不显著, 因此放弃执法, 不作为, 企业 (个人) 则因为有好处而提供假数据, 从而公共数据库里该项数据就是假数据.

4) 其他情况下该博弈没有 Nash 均衡解.

5) 除了 3) 的情况, 企业 (个人) 提供假数据的动机取决于数据违法期望收益 $U = pb + (1-p)c$(其中 p 为企业 (个人) 对政府稽查的主观概率). 若 $U > 0$, 则企业 (个人) 选择提供假数据; 否则企业 (个人) 提供真实数据. 解 $U = 0$, 得临界概率为

$$p^* = \frac{c}{c-b},$$

可见, 对企业 (个人) 做假惩罚力度越大 ($|b|$ 越大), 企业的做假动机越弱, 稽查的频率可以越小, 即形成数据稽查的高压态势. $b \to -\infty$ 时, 数据稽查对数据造假者威胁最大. 另一方面, 在对查获的数据做假者的处罚一定的情况下, 要使数据提供者提供假数据的期望收益为零来杜绝数据做假, 政府的数据稽查频率不得小于 p^*.

政府进行稽查的动机取决于稽查期望收益 $V = qa + (1-q)d$(其中 q 为政府对企业 (个人) 数据做假的主观概率), 临界概率为

$$q^* = \frac{-d}{a-d}.$$

表 2　上级政府 – 下级政府博弈

Table 2　High-to-lower governments game

		Lower government		
		High in checking rate	Middle in checking rate	Lower in checking rate
High government	Assessing	m, a	$0, c_2$	$0, b$
	No assessing	$0, c_1$	$0, c_2$	$0, c_3$

结论: 一般而言, 政府稽查中遇到数据造假, 如果处罚太轻执法成本太高将导致 $a < 0$, 从而使公共数据库失效.

例如, 迄今为止, 在所有质量技术监督数据库中强制检定计量器具 (如衡器、血压计、验光器、测速仪等) 信息多数是假数据或瞒报、漏报, 原因是由于此类计量器具种类数量太多, 执法成本高, 违法罚款少 ($\leqslant 2000$), 违法结果不涉及生命事故, 政府稽查的积极性不高, 从而基本处于未监管的状态 ($a < 0$). 在稽查成本不变的情况下, 提高稽查率的办法是加大违法的处罚力度 (如提高罚款额度).

博弈分析二: 上级政府 – 下级政府博弈

我们假定, 下级政府 (基层政府部门) 对企业 (个人) 提供的数据只有形式审核和稽查的权利，而且形式审核只涉及流程记录和格式审核. 因为公共数据库中的流程数据和数据格式审核都可以由技术性数据清洗来实现, 因此下级政府提交的的数据质量只是取决于稽查记录的真实性和数据稽查的覆盖率. 也就是说, 下级政府提交真实数据的意愿由稽查率来体现.

在上级政府和下级政府的博弈中, 上级政府的行动集为 {稽查, 不稽查}, 下级政府理论上的的行动集为 {稽查记录做假, 稽查记录不做假}$\times\{p|0 \leqslant p \leqslant 1\}$. 其中 p 为下级政府的数据稽查率. 但是, 如果下级政府的稽查记录做假的话, 有理由认为他可以将稽查率改为 100%, 故上级政府对下级政府的稽查内容退化为下级政府对企业 (个人) 稽查记录的真实性. 因此, 下级政府的行动集为 {稽查记录做似}$\cup[0,1]$, 更进一步, 由于稽查记录做假, 等同于对企业 (个人) 提供的数据未作任何核实, 即 $p = 0$, 也就是做假的稽查记录等同于全部数据未核实. 所以, 下级政府的行动集实际上就是 {高, 中, 低}. 此时, 上级政府的稽查变为考核, 即上级政府的行动集为 {考核, 不考核}. 我们得到上级政府和下级政府的博弈表 2:

1) 若上级政府选择不考核, 下级政府承担的稽查成本随着稽查率升高而升高, 所以 $c_1 < c_2 < c_3 < 0$; 若上级政府选择考核, 则一定会对稽查率低的下级政府进行惩罚, 而给稽查率高的下级政府奖励. 故 $b < c_3 < 0$ 且 $a > 0$.

2) 若 $m < 0$, 那么上级政府不会选择考核, 则 Nash 均衡解为 (不考核, 稽查率低); 若 $m \geqslant 0$, 则 Nash 均衡解为 (考核, 稽查率高).

因此公共数据库建设的顶层设计必须满足 $m \geqslant 0$. 否则公共数据库建不起来, 或建起来也无法运行, 或运行了数据也都是假的.

博弈分析三: 委托人 – 中介机构博弈

在中国，长期以来中介机构并不是独立于政府的第三方机构, 传统上属于事业单位, 实际上是受政府委托的. 到现在为止, 仍然有一些事业单位行使着政府职能甚至享有执法权. 但是, 随着我国的市场化改革的推进, 大量的中介机构在事业单位改制中被推向市场, 逐步过渡为提供专业知识的第三方. 这些中介机构中常见的如会计事务所、审计事务所、招标公司、环境评估事务所、产品质量检验所、特种设备检定研究院、计量鉴定所、车辆检验站等. 作为技术支持单位已经成为代表技术公正的第三方, 而不再单纯依赖于政府和代表政府说话. 例如, 在产品质量检验中, 质检机构出具的产品质量检验

表 3 上委托人 – 中介机构

Table 3 Principal-to-professional organization game

	Providing false reports	Providing ture reports
Believing	a, b	c, d
No believing	0,0	$0, e$

报告代表的是第三方结论, 在产品质量仲裁中应该具有公正性, 从而使政府质量稽查执法部门和受稽查的生产企业在技术上处于平等地位.

尽管对中介机构的业务有法律约束和行政许可, 但是, 实际中由于它们长期依附于政府部门或受到商业利益的驱使, 并不能真正严守技术中立, 出示虚假技术报告 (如财务审计报告、环境检测报告、产品检验报告等) 时有发生. 这些虚假的专业技术报告会给公共数据库提供虚假数据, 使得公共数据库中数据质量受到影响. 它们会接受委托方指使提供虚假技术数据. 例如, 在 SARS 流行初期某疾病预防控制机构就提供了虚假的病例报告, 严重妨碍了疫情数据的真实性.

由于专业性要求很高, 中介机构提供的数据真实性很难受到稽查, 无论政府还是企业 (公众), 比起中介机构处于信息不对称中弱势的一方. 有时这样的中介机构还处于垄断地位, 更是加重了他们数据做假的可能性. 委托人 – 中介结构的博弈如表 3.

1) $a < 0$ 且 $b > 0$ 且 $e < 0$. 因为委托人相信了一个虚假报告, 在商业上或行政诉讼中一定会带来更大损失, 报告提供者一定从中得到好处 (如商业贿赂). 反之, 中介机构花钱提供了真报告, 但委托人不相信, 那么一定不付钱, 所以会给中介机构带来损失.

2) 若 $c \leqslant 0$, 则 Nash 均衡解为 (不相信，提供假报告). 若 $c > 0$ 且 $d \leqslant b$, 则 Nash 均衡解也为 (不相信, 提供假报告). 所以, 技术机构可能提供真实报告的条件是委托人在真报告中得到足够多的好处而且技术机构提供真实报告的好处大于提供假报告的好处. 但是, 即使 $c > 0$ 且 $d > b$, (相信, 提供真报告) 也不是 Nash 均衡解.

因此, 要保证中介机构给公共数据库的数据是真实的, 即提供真报告, 必须进行机制设计, 即给予技术机构足够的补偿 $r \geqslant -e$. 在实际中, 技术报告的委托方为多个时, 情况就复杂得多, 报告真实的充分条件之一是没有真正的委托人, 或匿名委托人.

3.3 公共数据库中数据制造过程质量控制的机制设计

从前面博弈论分析看, 如果不进行机制设计, Nash 均衡的结果就是公共数据库失败. 经过机制设计可以达到新的 Nash 均衡并实现对各个环节的数据质量控制. 机制设计的目的是经过制度创新使公共数据库的数据源提供者及时提供真实数据, 同时使数据管理者能够及时信息公开且有效保护隐私权.

根据公共数据库中数据制造过程、数据质量维度分析和数据质量的博弈论分析, 提出公共数据库数据制造非技术性质量控制图. 如图 3 所示.

4 公共数据库数据质量评估的变权综合模型

按照前述的数据质量控制机制设计, 从非技术角度讲, 要确保公共数据库的数据真实的条件是: (1) 上级政府对数据真实性进行考核; (2) 基层政府的数据稽查率为 100%且企业 (个人) 提供假数据违法成本高; (3) 中介机构可以得到足够多的补偿.

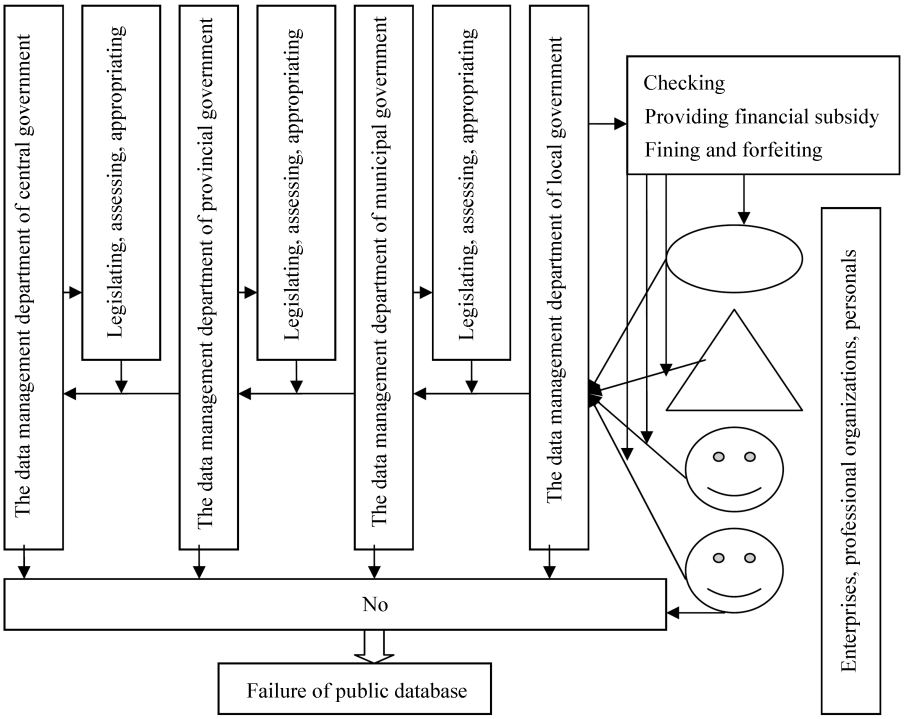

图 3　公共数据库数据质量控制

Figure 3　Data quality control system for Chinese public database

但是, 实际中基层政府部门的稽查率不可能达到 100%, 即便企业 (个人) 违法成本很高仍然会有企业 (个人) 提供假数据或不提供数据. 所以, 无论如何进行机制设计, 公共数据库中一定存在假数据或数据不完整. 这就使公共数据库中最终的数据质量评估仍然有必要. 选择的维度是一致性 (f_1)、正确性 (f_2)、完整性 (f_3)、和最小性 (f_4)、真实性 (f_5)、及时性 (f_6)、权威性 (f_7) 和隐私保护 (f_8). 在不同的公共数据库中可能这些维度的重要程度不同, 但是不管是哪一个维度评价太低都会导致数据库由于数据质量差而不可用, 即数据质量综合评价为零. 这与变权综合的思想相吻合 [24,25].

下面是公共数据中数据质量的变权综合模型.

设 x_j 是 f_j 的属性值. 代表对应维度下数据质量的评价值, $x_j \in (0,1]$; $w_j^{(0)}$ 为 f_j 的重要程度, 即初始权重. 建立如下变权:

$$w_j(x_1,x_2,\ldots,x_8)=\frac{w_j^{(0)}x_j^{\alpha-1}}{\sum_{k=1}^{8}w_k^{(0)}x_k^{\alpha-1}},\quad (j=1,2,\ldots,8,0<\alpha<1)$$

及变权综合评价模式:

$$V_\alpha(x)=\sum_{j=1}^{8}w_j(x_1,x_2,\ldots,x_8)x_j,$$

其中 α 为惩罚系数.

5 质量技术监督公共数据库的数据质量控制与数据质量评估

5.1 质量技术监督公共数据库建设和运行的法律支撑体系

目前, 中国有关质量技术监督公共数据库的法律和条例有 7 部: 《中华人民共和国政府信息公开条例》(2008 年)、《中华人民共和国保密法》(2010 年)、《中华人民共和国知识产权法》、《中华人民共和国食品安全法》(2009 年)、《中华人民共和国产品质量法》(2000 年)、《特种设备安全监察条例》(2009 年修订)、《中华人民共和国认证认可条例》(2003 年). 另外一些还有各省、市、自治区立法机构和政府通过的地方性法规. 关于企业或个人的隐私权保护还没有法律和条例支持, 所以在数据质量控制过程中的隐私保护很难界定, 一些内容可以参考《保密法》和《知识产权法》裁定. 例如, 在要求食品生产企业填报生产原料来源、数量、食品添加剂进货索证、关键过程控制数据时, 有的企业以泄露产品配方和制作工艺为由拒绝提供真实数据时有发生. 因此, 我国数据隐私保护立法滞后已经影响了公共数据库中的数据质量 [26,27], 质量技术监督公共数据库中数据质量也不例外.

5.2 质量技术监督公共数据库的最小数据集的形成和数据制造成本控制

根据上述法律和条例, 我国质量技术监督部门已经着手建立质量技术监督公共数据库. 其中包括组织机构代码数据库、产品质量档案数据库、质量信用数据库、特种设备安全监察数据库、标准备案登记数据库等. 但是, 迄今为止除了国家组织机构代码数据库运行良好外, 其他国家级质量技术监督公共数据库建设混乱, 多数基本处于瘫痪状态. 存在的主要问题是重建设轻设计、轻维护、轻考核、政出多门、条块分割. 经过调查发现, 仅 2004 年要求云南省基层质监部门提供数据的国家级和省级数据库就有 9 个, 而且互不兼容. 2008 年由国家级别出面整合全国的质量技术数据库系统, 形成的数据库也有 8 大系统, 数据集交叉重合且互相兼容性差, 反映了国家层面公共数据库数据管理条块分割, 数据集远未最小化, 造成企业 (个人) 和基层质监部门数据收集和录入难度大、成本大, 不能反映 “千条线连着一根针” 的基层和企业的实际. 最终这些数据库基本名存实亡. 2004 年, 根据国家法规和质量技术监督工作日常需要, 结合原有各个数据库提出的数据需求和质监部门的实际情况, 我们用文献 [28,29] 提出的粗糙集方法进行质量技术信用的属性数据挖掘, 并指导昆明市质量技术监督局提出了《巡查记录表》的概念, 即提供了基础表格给一线的巡查人员在巡查过程中使用. 此《巡查记录表》收集的数据加上日常审批记录即可覆盖国家各类数据库且适应日常监管工作需要. 之后, 经过在使用过程中反复提炼, 消除一些数据的重合 (如设备型号中含有设备参数, 产品标准中含有质量参数等) 并规避了可能泄露企业商业机密的属性, 进一步形成了更加简明的《三查三责表》, 以此收集质量技术监督监察工作的基本数据, 形成了质量技术监督公共数据库的最小数据集. 2008 年, 国家质检总局在昆明召开了讨论会肯定了数据集的科学性, 并将数据收集范围扩大到举报投诉和产品召回, 并作为国家 “金质工程” 的数据规范之一. 2009 年, 经过对企业的回访, 扩充后的数据集得以最终确定, 形成质量技术监督移动执法系统的信息资源库最小数据集.

为了进一步控制数据收集成本, 减轻企业、基层质监分局一线巡查人员和后台录入人员的工作量, 经过对企业内部质量管理和设备管理系统的现状, 提出企业报送、质监局审核的数据生成方式. 填报的内容仍然以《三查三责表》为基准, 并辅以关键过程质量台账、设备运行台账和从业人员资格管理台账等附件, 抽样进行数据稽查, 形成对现场数据稽查的有力补充. 而且这些基本信息和附件直接生成企业许可证年审、换证和申请新证的支撑材料进入审批流程, 大大降低了数据收集和处理的成本并且提高了数据质量.

在此基础上, 2011 年昆明市质量技术监督局又融合了特种设备检定报告管理系统、组织机构代码管理系统, 从而为数据稽查提供了第三方技术数据支撑, 大大提高了数据的准确性和及时性.

5.3 质量技术监督公共数据库的架构和开发工具

质量技术监督公共数据库系统的主数据库是 Linux 操作下的 ORACLE 11g 数据库. ORACLE 数据库系统是美国 ORACLE 公司 (甲骨文) 提供的以分布式数据库为核心的一组软件产品, 是目前最流行的客户/服务器 (Client/Server) 或 B/S 体系结构的数据库之一. 比如 SilverStream 就是基于数据库的一种中间件. ORACLE 数据库是目前世界上使用最为广泛的数据库管理系统, 作为一个通用的数据库系统, 它具有完整的数据管理功能; 作为一个关系数据库, 它是一个完备关系的产品; 作为分布式数据库它实现了分布式处理功能.

主数据库的数据源是各个应用系统和企业上报数据, 这些系统的数据库一般为 Microsoft SQL Server, 我们对有条件提供数据接口的应用系统数据库进行设计.

系统的构架设计图 4.

系统界面设计分为资源平台界面和资源平台导入界面, 都用 JAVA 实现.

主数据库界面: 实现基层局的各个应用系统的查看、导入、导出等功能. 导入分为手工导入和自动导入 2 种方式, 手工导入可以实现 EXCEL, SVC, XML 格式的文件导入, 例如, 特种设备监察网、质量档案系统等. 提供数据结构的应用系统可以实现自动导入, 例如, 基层质量技术监督管理信息系统、昆明市获证企业管理信息系统等.

主数据库的前置界面: 主数据库系统的前置导入界面布局在数据源服务器显示任务栏里, 可以实现导入过程的查看、成功导入的数据的查看、未导入成功的数据的查看. 成功导入的数据和未成功导入的数据都可以生成 TXT 或者 SVC 格式的文件, 这样未成功导入的数据就可以发回给对应数据源提供者进行修改, 修改反馈后再次进行导入. 未成功导入的数据是软件自动进行比对出的格式不正确或者重复的数据.

5.4 对基层政府部门的数据质量考核

如前所述, 从机制设计 [30] 角度讲, 要提高公共数据路数据质量, 除了良好的技术设计外，还必须加强数据稽查的制度建设. 为了提高基层分局数据稽查的力度, 昆明市质量技术监督局结合政府部门导入 ISO9001 质量体系认证, 从 2010 年起将数据填报和稽查制度化并列入季度和年终考核范围, 且占分逐年提高, 2012 年占年终考核总分的 15%, 并且其他各项考核的依据以公共数据库里的数据为准, 从而大大提高了数据质量和规模. 表 4 是该局年终数据考核指标和评估方法.

表 5 是 2012 年基层分局各基层数据质量工作部分考核成绩表.

2012 年基层分局数据质量变权综合评估结果如表 6.

从数据质量评估结果看, 质量技术监督数据库的数据质量仍不容乐观, 主要原因是特种设备和强检计量器具数据陈旧, 更新不及时. 造成数据陈旧的原因包括: (1) 特种设备和强检计量器具数量庞大, 分布广, 基层分局安全监察力量薄弱, 导致数据稽查不力; (2) 对特种设备和强检计量器具过期未检行政处罚过轻, 以致下发处罚决定书后无明显改善; (3) 技术机构 (特检院、计量院) 检定人员严重不足, 设备检定到期而且企业已经报检但技术机构派不出技术人员进行检定; (4) 大量的特种设备和强检计量器具使用单位在服务行业, 分布较广, 而质监部门的工作重点集中在生产环节, 忽视了服务业的特

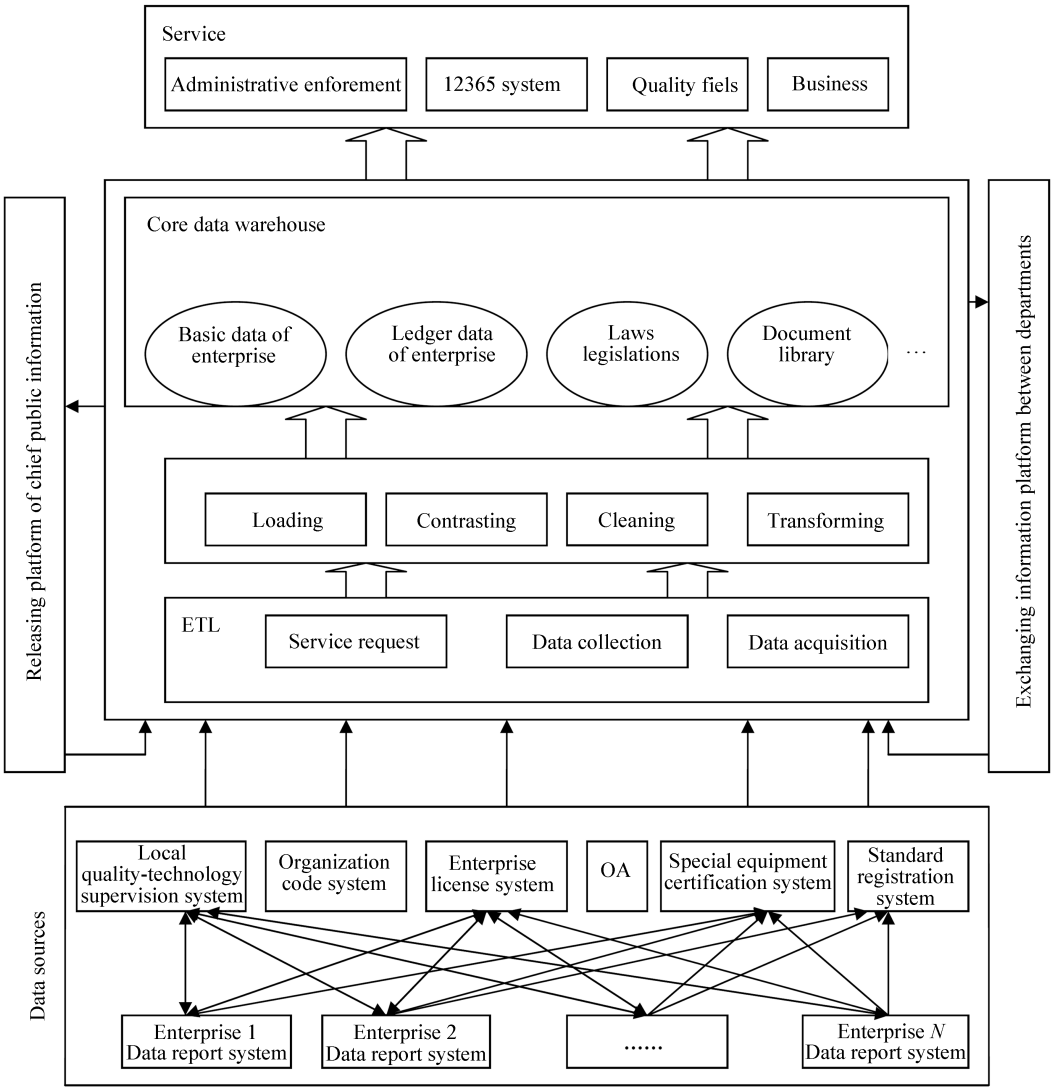

图 4 质量技术监督公共数据库系统架构设计

Figure 4 The frame of controlling data quality in the quality-technology supervision database

种设备、强检计量器具监察和标准备案登记; (5) 目前企业数据网上上报只限于获证的生产企业, 其他企业数据上报困难, 有些企业已经鉴定了设备但相关数据无法更新.

6 结论

通过以上的模型体系分析, 我国目前公共数据库建设和运行中存在很多问题, 造成了公共数据库的建设效果不理想和数据库中的数据质量较差, 从而影响了公共数据库在公共管理决策和国民生活中的应用效果. 存在的主要问题及政策建议如下:

(1) 公共数据库立法严重滞后, 对公共数据库建设和运行财政投入太少, 而且重建设、轻运行维护, 也存在部门利益凌驾于总体利益的不合理利益格局, 导致中央级政府部门缺乏对地方政府 (下级政府)

表 4　质量技术监督数据库数据质量考核指标和评分办法

Table 4　Assessment items and their assessment scores

No.	Assessment items	Assessment scores
1	The entry of the daily patrol and safety supervision from.	The monthly supervision and inspection work is not less than 10% of the total number of the certification enterprises of the county (district), and 0.2 points will be dropped with the lack of one percent; the monthly safety supervision work is not less than 10% of the total number of enterprises with equipment of the county (district) and 0.2 points will be dropped with the lack of one percent.
2	All the production licenses of the enterprises belonging to the jurisdiction in pairs can register, and the number of production licenses must coincide with the business department archives information.	The number of production licenses must coincide with the business department archives information, and 0.2 points will be dropped if there is a missing or overstating mistake.
3	The work of license year careful and new registration application must be first done in the system, and then paper-based materials submitted by companies can be accepted.	The work of enterprise license year careful must be acted as planned，the request of enterprise year careful cannot be delayed, and 0.2 points will be dropped if there is a un-audit information.
4	Establish the special equipment and measuring equipment electronic parameter, and updated the old data in real time.	0.2 points will be dropped if there is a outdated information; the system data must coincide with the equipment stand-books of, 0.5 points will be dropped if there is a discrepancy or overstating, underreported.
5	The entry of the enterprise product sampling information.	The monthly product sampling information is not less than 10% of the total number of the companies leading products of the county (district), and 0.5 points will be dropped with the lack of one percent.
6	The key process of replenish onr's stock, and sales of electronic parameter of the certification enterprises within the jurisdiction.	The monthly enterprise parameter input rate is not less than 80% of the total number of the certification enterprises of the county (district), and 0.1 points will be dropped with the lack of one percent.
7	Entry the completion status of Yunnan province famous brand products and Kunming famous brand products every month, and keep the records (including the trace data).	The monthly famous brand product entry rate is not less than 80% of the total number of famous brand product of the county (district), and 0.1 points will be dropped with the lack of one percent.
8	The audit information of enterprises must be dealt with within 24 hours	0.5 points will be dropped if there is a un-audit information.
9	Send all the notice or document to the certification enterprises through the system.	0.1 points will be dropped if there is a lack of record of year careful on the certification enterprises.
	Remark:	Remark:

表 5　2012 年基层分局各基层数据质量工作考核成绩表

Table 5　Annual assessment results of branch in 2012

No.	Branch names	Industrial production licenses: total/expired	Food production licenses: total/expired	Number of inspection	Number of cases chosed	Number of product sampling	Records of measuring instrument total/expired	Records of special equipment total/expired	Rectification notice
1	Jingkai	57/6	250/29	1441	115	339	1060/10	2567/44	369
2	Gaoxin	49/1	60/3	217	22	96	2003/859	1357/90	109
3	Dujia	1/0	9/0	153	8	10	367/102	992/343	60
4	Guandu	245/140	845/453	1546	7	78	1097/551	11582/9533	454
5	Panlong	58/9	154/48	794	96	186	643/502	4854/2487	114
6	Xishan	102/3	113/10	416	67	149	2688/1155	6281/582	44
7	Wuhua	93/28	70/1	337	1	68	2771/2166	4219/5	42
8	Dongchuan	13/4	33/25	43	0	5	275/272	807/476	9
9	Chenggong	22/11	16/14	19	0	19	381/367	1198/716	1
10	Anning	111/7	32/0	212	33	33	2049/1559	4859/3672	16
11	Yangzong	7/0	36/9	103	47	6	324/189	1192/1031	79
12	Jinning	57/10	49/8	256	42	266	423/404	1203/986	40
13	Yiliang	62/10	73/17	364	24	30	750/531	803/653	50
14	Shilin	12/5	54/10	203	36	122	477/251	364/236	143
15	Fumin	22/0	34/2	341	45	107	336/159	523/17	167
16	Luquan	15/0	17/13	93	13	25	307/18	216/32	10
17	Xundian	24/3	27/3	253	8	170	292/243	186/85	30
18	Songming	46/0	67/5	350	29	56	341/259	752/179	51
	Total	996/232	1579/636	7141	593	1765	16584/9597	43955/21167	1788

表 6　数据质量评估表

Table 6　Data quality values of branch in 2012

No.	1	2	3	4	5	6	7	8	9
Data quality value	0.87	0.82	0.90	0.71	0.60	0.82	0.60	0.57	0.55
No.	10	11	12	13	14	15	16	17	18
Data quality value	0.54	0.50	0.51	0.55	0.59	0.81	0.79	0.67	0.65

部门进行数据库建设和数据质量考核积极性, 同时也削弱了基层政府部门数据稽查的积极性. 建议各级政府加大公共数据库建设和维护的常年财政预算, 从而提高上级政府部门对下级政府部门的公共数据库数据质量考核力度.

(2) 受到执法成本的压力和执法力量薄弱的制约, 基层政府部门无力提高对各种数据源数据的稽查率. 建议通过提高拨款和增加编制来加强基层政府部门数据稽查力量, 提高数据稽查率.

(3) 对各种数据源提供假数据的处罚力度太小, 使数据源提供假数据的违法成本过低, 增强了企业(个人) 提供假数据或瞒报的主观意愿. 建议提高对数据造假和瞒报者的行政处罚力度, 甚至在行政许可中一票否决, 加大他们的造假成本, 增强他们向公共数据库提供真实数据的意愿.

(4) 中介机构过于市场化且疏于监管, 机构认证过滥, 使得中介机构在利益驱使下提供大量的虚假数据. 建议适当增加给予中介机构的资金支持并加强对其监管, 从而降低中介机构的造假意愿.

(5) 各种数据源提供数据的成本过高, 渠道不畅, 无法及时提供真实数据. 建议公共数据库在安全的条件下尽量让各地数据源联网上报, 梳理数据收集范围, 实现数据最小化, 从而减少上报成本.

参考文献

1 Xu G Z, Gu J F. Systems Science. Shanghai: Shanghai Kejijiaoyu Press, 2000. 249–260 [许国志, 顾基发. 系统科学. 上海: 上海科技教育出版社, 2000. 249–260]

2 Zhang W Y. Game Theory and Information Economics. Shanghai: Shanghai People's Publishing House, 2004 [张维迎. 博弈论与信息经济学. 上海: 上海人民出版社, 2004]

3 Xu Z P. The Big Data Revolution. Guilin: Guangxi Normal University Press, 2012 [涂子沛. 大数据. 桂林: 广西师范大学出版社, 2012]

4 Li H. Market Mechanisms in the Process of Public Goods Supply. Tianjin: Nankai University, 2010 [李慧. 公共产品供给过程中的市场机制. 天津: 南开大学, 2010]

5 Chen Q L, Han X T. The quality of accurate communal product: definition, classification and category basis. Economist, 2010, 10: 13–21 [陈其林, 韩晓婷. 准公共产品的性质: 定义、分类依据及其类别. 经济学家, 2010, 10: 13–21]

6 Wang R Y, D M Strong. Beyond accuracy: what data quality means to data consumers. J Manage Inform Syst, 1996, 12: 5–34

7 Knight S, Burn J. Developing a framework for assessing information quality on the world wide web. Inform Sci J, 2005, 8: 159–172

8 Song M, Qin Z. Reviews of foreign studies on data quality management. J Inform, 2007, 2: 7–9 [宋敏, 覃正. 国外数据质量管理研究综述. 情报杂志, 2007, 2: 7–9]

9 Han J Y, Xu L Z, Dong Y S. An overview of data quality research. Comput Sci, 2008, 35: 1–5 [韩京宇, 徐立臻, 董逸生. 数据质量研究综述. 计算机科学, 2008, 35: 1–5]

10 Guo Z M, Zhou A Y. Research on data quality and data cleaning: a survey. J Softw, 2002, 13: 2076–2082 [郭志懋, 周傲英. 数据质量和数据清洗研究综述. 软件学报, 2002, 13: 2076–2082]

11 Han J Y, Song A B, Dong Y S. Approach of quantifying data quality dimensions. Comput Eng Appl, 2008, 44: 1–6 [韩京宇, 宋爱波, 董逸生. 数据质量维度量化方法. 计算机工程与应用, 2008, 44: 1–6]

12 Feng Z, Hu W J, Gao Y B. The research of data quality management plan in CRM system. J Elec Power, 2010, 29: 171–173 [冯舟, 胡文江, 高永兵. CRM 系统数据质量管理方案研究. 电力学报, 2010, 29: 171–173]

13 Wang H Z, Fan W F. Object identification on complex data: a survey. Chinese J Comput, 2011, 34: 1843–1852 [王宏志, 樊文飞. 复杂数据上的实体识别技术研究. 计算机学报, 2011, 34: 1843–1852]

14 Ballou D P, R Y Wang, H Pazer, et al. Modeling information manufacturing systems to determine information product quality. Manage Sci, 1998, 44: 462–484

15 Rahm E, Do H H. Data cleaning: problems and current approaches. IEEE Data Eng Bull, 2000, 23: 3–13

16 Zhang G B, Ren X L, Li M, et al. Dynamic quality traceable system based on MES and CAPP. Comput Integ Manuf, 2010, 16: 349–355 [张根保, 任显林, 李明, 等. 基于 MES 和 CAPP 的动态质量可追溯系统. 计算机集成制造系统, 2010, 16: 349–355]

17 Wang Y. Method for data quality to improve the provincial corporate credit underlying database. E-Government, 2012, 9: 85–90 [王云. 提高省级企业信用基础数据库数据质量方法研究. 电子政务, 2012, 9: 85–90]

18 Geng Y B, Yu L, Zhao H. ITS data quality control techniques and applications. China Safety Sci J, 2005, 1: 85–90 [耿彦斌, 于雷, 赵慧. ITS 数据质量控制技术及应用研究. 中国安全科学学报, 2005, 1: 85–90]

19 Wang S, Wang H J, Qin X P, et al. Architecting big data: challenges, studies and forecasts. Chinese J Comput, 2011, 34: 1741–1752 [王珊, 王会举, 覃雄派, 等. 架构大数据: 挑战、现状与展望. 计算机学报, 2011, 34: 1741–1752]

20 Wang H Z, Li J Z, Gao H. Data model for dirty databases. J Softw, 2012, 23: 539–549 [王宏志, 李建中, 高宏. 一种非清洁数据库的数据模型. 软件学报, 2012, 23: 539–549]

21 Peng Q, Guo M, Lin P. An algorithm for the maximum clique based on MapReduce. Syst Eng Theory Pract, 2011, 31: 150–153 [潘全, 郭鸣, 林鹏. 基于 MapReduce 的最大团算法. 系统工程理论与实践, 2011, 31: 150–153]

22 Huang L T, Deng S G, Dai K, et al. Automatic service composition in parallel with MapReduce. Acta Electron Sin, 2012, 40: 1397–1403 [黄龙涛, 邓水光, 戴康, 等. 基于 MapReduce 的并行 Web 服务自动组合. 电子学报, 2012, 40: 1397–1403]

23 Fang D C. Game analysis of the quality of statistical data reliability. Stat Decis, 2009, 25: 30–31 [方大春. 统计数据质量可靠性的博弈分析. 统计与决策, 2009, 25: 30–31]

24 Liu W Q. The ordinary variable weight principle and multiobjective decision-making. Syst Eng Theory Pract, 2000, 20: 1–11 [刘文奇. 一般变权原理与多目标决策. 系统工程理论与实践, 2000, 20: 1–11]

25 Li D Q, Gu Y D, Li H X. Results on axiomatic definition of state variable weight vector. Syst Eng Theory Pract, 2004, 24: 97–102 [李德清, 谷云东, 李洪兴. 关于状态变权向量公理化定义的若干结果. 系统工程理论与实践, 2004, 24: 97–102]

26 Wang H W, Zhou M, He S Y. Empirical research of individuals' intention to provide privacy information online. Syst Eng Theory Pract, 2012, 32: 2186–2197 [王洪伟, 周曼, 何绍义. 影响个人在线提供隐私信息意愿的实证研究. 系统工程理论与实践, 2012, 32: 2186–2197]

27 Meinert D B, Peterson D K, Criswell J R, et al. Privacy policy statements and consumer willingness to provide personal information. J Elec Com Org, 2006, 4: 1–17

28 Zhang S N, Liu W Q. Dynamics approximate rule mining inference approach based on rough set theory. Control Theory Appl, 2003, 20: 93–96 [张仕念, 刘文奇. 一种基于粗集理论的动态近似规则挖掘推理方法. 控制理论与应用, 2003, 20: 93–96]

29 Ding D Q, Liu W Q. Enterprise credit evaluation system of quality technology based on rough set and variable weight synthesis. Nat Sci J Hainan Univ, 2010, 28: 343–347 [丁德琼, 刘文奇. 基于粗糙集理论与变权综合的企业质量技术信用评价. 海南大学学报, 2010, 28: 343–347]

30 Hurwicz L, Reiter S. Designing Economic Mechanisms. London: Cambridge University Press, 2006

Modeling data quality control system for Chinese public database and its empirical analysis

LIU WenQi

Faculty of Science, Kunming University of Science and Technology, Kunming 650500*, China*
E-mail: liuwenq2215@sina.com

Abstract The dimensions of data quality of public database were determined, based on discussing the properties of public database and the characteristics of data manufacturing process to public database. The new concepts of non-technical data cleaning and data check were presented for public database as special communal product, and the variable weight synthesis model was set to evaluate data quality. The procedure of quality control in data manufacturing process was designed by computer network technology and modeling game theory, to suggest a policy for the operation of public database in China. As an example, the programming to the quality and technology supervision database was presented to control the data manufacturing.

Keywords public database, data quality, data cleaning, data manufacturing process control, data check, game analysis, Mapreduce model

LIU WenQi was born in 1965. He received his Master degree in Science from the Beijing Normal University, Beijing, in 1995. Currently, he is a professor at Kunming University of Science and Technology. His research interests include data mining, decision making, information fusion and information system.