一种面向闭环的数据治理平台与方法设计

张国宝,卞艺杰
(河海大学 网信中心,江苏 南京 210098)

摘　要:数据孤岛现象在高校普遍存在,数据治理的目的是解决高校存在的数据源分散异构、数据质量不高等问题。文中对数据治理的方法论和数据平台的功能设计进行研究探讨。研究了数据质量、数据质量维度和数据质量治理相关概念,结合 PDCA 系统循环理论,提出了闭环数据治理体系的原则和治理的四个维度,即技术维度、业务维度、演化维度、管理维度。设计了符合高校遗留系统环境下的闭环数据治理体系架构和数据归集、对标、提质、共享、服务的治理活动过程,并且给出了数据治理平台的功能设计,在主数据管理、数据质量检测等方面实现可视化管理。通过设计的数据治理平台的实际案例的应用,以及相关关键技术的分析,结果说明平台在数据治理方面是有效的。

关键词:数据质量维度;数据治理;PDCA;闭环
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2019)08-0156-05
doi:10.3969/j.issn.1673-629X.2019.08.030

A Closed-loop Data Management Platform and Method Design

ZHANG Guo-bao, BIAN Yi-jie
(Network and Information Center, Hohai University, Nanjing 210098, China)

Abstract: Isolated data island is common in colleges and universities. The purpose of data governance is to solve the issue of scattered and heterogeneous data sources and low data quality in colleges and universities. The methodology of data governance and the functional design of data platform are studied. The concepts of data quality, data quality dimension and thoughts on data quality governance are studied. Combined with PDCA system cycle theory, the principles of closed-loop data governance system and four dimensions of governance are proposed, namely technical dimension, business dimension, evolution dimension and management dimension. The closed-loop data governance architecture and data governance activities including data collection and management, benchmarking, quality improvement, sharing and service in the legacy system environment of universities are designed, and the functional design of the data governance platform is given. Through the application of the actual case and the introduction of data quality detection methods, the demonstration shows the designed platform is effective in data governance.

Key words: data quality dimension; data governance; PDCA; closed loop

1　研究背景与问题提出

校园信息化建设发展到今天,已经从 IT(information technology)步入 DT(data technology)时代。即利用大数据、物联网、移动计算、云计算等新的技术让校园更具“智能”,实现“智慧校园”。而数据“孤岛”,是利用大数据技术实现“数据”说话、用户少跑腿的智慧化应用服务建设需要面对的现状。遗留系统,泛指校园中已经建成的多个业务应用,一般都具有独立的数据库。数字校园建设的数据共享平台,通过数据的简单共享与交换实现了遗留系统之间的数据共享,但是不能很好地满足流程化服务应用的建设以及未来的智慧校园的需求。

主要表现在:数据标准不能与业务数据进行关联,数据标准不能及时反映业务数据的变化;不能提供全局准确的数据来支撑决策分析及大数据应用;不能提供数据质量溯源与数据关联影响分析。以上问题制约着校园信息化中数据的有效共享和基于共享数据的大数据分析应用。因此,文中从数据治理体系的方法论的角度进行数据治理体系的设计与构建研究,以期找到解决上述问题的方法路径,以及为智慧校园的数据

收稿日期:2018-09-15　　**修回日期**:2019-01-16　　**网络出版时间**:2019-03-27
基金项目:2017 年江苏省教育信息化研究课题(20172067)
作者简介:张国宝(1980-),男,博士在读,研究方向为数据集成、信息融合;卞艺杰,教授,研究方向为信息管理与电子商务、金融工程与投资管理。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1624.040.html

平台的数据治理和数据质量提升研究提供参考与借鉴。

2　数据质量与数据质量治理

对于数据质量与数据质量治理,许多学者进行了相应研究,包括数据质量的概念与定义,数据质量维度和质量属性,数据质量评估方法与数据质量提升,数据质量治理等,研究所针对的问题背景不尽相同。

2.1　数据质量与数据质量维度

数据质量是满足使用的需要,是数据“使用适合性”(fitness for use)的一种刻画[1]。文献[2]认为有 12 个数据质量维度属性。文献[3]列举了 19 个数据质量属性,其中所列的数据质量不仅仅有质量方面的属性,如完整性,还有“非质量”方面的属性,如可访问性、表达质量等。文献[4]给出了一个大致的属性划分。但是一般认为最主要的数据质量属性是正确性、完整性、一致性、最小性[5-6]。文献[7-8]认为最主要的数据质量属性是正确性、完整性、时效性、一致性。基于数据质量属性的研究,衍生出数据质量量化和评估的研究,例如文献[9]给出了基于数据质量属性、准确性和完整性的度量量化方法。文献[8]给出了针对数据质量属性(正确性、完整性、时效性、一致性)的约束规则的质量量化算法。正如文献[2]所给出的,数据质量评估流程一般是确定维度、定义约束、确定算法、实施评估、质量问题检查与质量问题解决。文中主要围绕质量问题检查与质量问题解决。

2.2　数据质量治理

正如文献[10]提到的,数据质量的问题来源是多方面的。既有业务方面的原因,例如实际的业务还未发生,相应数据没有产生。也有技术方面的原因,或者格式校验缺失,以及人为录入错误的原因。文中主要针对的是异构多数据源集成后带来的数据质量的治理问题,进行数据质量治理框架体系的设计构建。文献[11]提出了基于 PDCA 循环和六西格玛改进方法 DMAIC 的闭环的数据质量管理模型。文献[12]基于数据质量评估维度,提出了数据质量管理的整体框架。文献[13]提出了 ERP 数据质量评估指标体系,以及相应的企业级数据治理方案。建立起包含组织、标准、流程、质量、安全、技术多个目标的方案框架。文献[14]指出数据治理是一个具有决策和审计的系统体系,依据一致的治理模型对数据进行处理和操作,模型规定了在什么条件下,谁能对什么样的数据进行何种操作,采用什么样的方法。文献[15]通过应用一个元数据质量框架实现数据治理,框架主要从元数据的语法、语义、程序三个层次进行分析和处理。

综上,数据治理是一个系统工程,需要相应的方法论作为依据,需要依赖设计数据治理的平台完成治理的功能。数据质量的治理不是一次性和短期性的问题,需要迭代构建整体的治理体系完成数据治理。

3　PDCA 循环与数据治理原则

PDCA 循环是美国质量管理专家休哈特博士首先提出的,由戴明采纳、宣传,获得普及,所以又称戴明环。PDCA 循环的含义是将质量管理分为四个阶段,即计划(Plan)、执行(Do)、检查(Check)、调整(Adjust)。文献[16]提出基于 PDCA 的云数据治理模型。文献[14]提出了云数据治理的 9 个关键维度,其中 5 个维度适用于非云环境。文中借用 PDCA 理念构建面向高校遗留系统的数据质量治理体系,从四个维度来构建数据治理框架体系,即技术维度、业务维度、演化维度、管理维度,如图 1 所示。

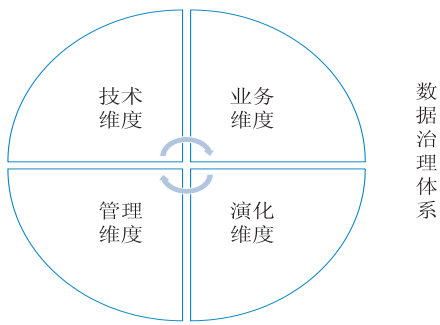

图 1　数据治理体系维度

数据治理的目的,即提高数据质量。治理体系是一个运行的动态过程,根据实践认为治理过程中需要坚持以下原则:

3.1　数据源唯一性原则

同一个数据来源只能有一个出处。其他都应该是数据出处的引用部门。这里同一个数据是指数据对象,一般应当以数据表为单位。

3.2　数据按需迭代原则

业务数据伴随着具体业务的办理进行沉淀和积累,数据的使用与交换也需要依据数据的需求开展,坚持最小化原则,够用就行,提高效率。

3.3　标准伴随原则

数据标准在数据的使用和交换过程中,起到很好的“桥梁”作用。因此数据的集成与使用,数据标准应当伴随着业务数据。

3.4　业务场景原则

数据治理是为了满足数据的使用需求,数据的使用需求需要符合实际需要的业务场景,所以数据的治理需要结合业务场景,满足业务场景需要。

3.5　先管理后治理原则

数据治理不是目的,而是手段。要做到数据治理,

首先应该进行数据的管理,把已有的各种遗留系统的数据能够登记注册,通过主数据平台能够对需要治理的数据进行集成汇聚,而后逐步构建数据治理体系。

4 数据质量治理体系设计

数据治理体系借鉴 PDCA 系统管理理念,它是一个系统工程,需要从系统的视角进行考虑。可以从提出问题(目的)到解决问题(方法)然后到验证问题(效果)再回到提出新问题(需求)构成一个闭合的工作链条。具体到治理体系的四个维度:

4.1 技术维度

通过分散异构数据源的数据聚集,实现主数据的条目化管理,通过数据对标、质量规则约束检查,实现数据治理体系中数据源端到数据使用端的有质量的“流通”。在技术(工具)层面上实现数据从产生到使用,从集成管理到质量提升治理的实现。技术维度的 PDCA 过程要素是:集成与主数据管理,可视化的质量监测,场景化数据服务,数据修正与迭代。

技术维度是数据治理体系的基础,是实现数据采集、汇聚的工具和平台。通过技术手段进行数据的登记、交换、集成、管理与数据服务提供。通过数据服务向应用提供经过治理的数据。技术维度的主要功能包括数据源的管理、数据交换与集成、数据标准管理、数据标准发布、数据归档、数据服务接口、数据质量管理、数据脱敏等。

4.2 业务维度

通过数据的场景化业务需求,进行数据的按需共享和数据服务实施,通过数据需求驱动数据的有效共享,同时反过来根据业务场景的需求,从业务侧进行共享数据的满足业务使用的验证,同时反馈推动共享数据的使用质量的提升。业务维度的 PDCA 过程要素是:场景化数据需求申请,数据共享服务实施,业务数据验证,数据修正校核。

业务维度是数据治理体系的目的,数据的采集、汇聚与数据质量检测是为了满足业务应用的场景需求。业务应用的主要功能包括全局的校情分析、数据统计分析、跨业务的数据类应用等。

4.3 演化维度

伴随着业务系统的技术演化和业务对数据需求的更迭,面向遗留系统的数据集成的过程是一个动态的增量迭代过程,业务系统的升级和业务需求的更新共同驱动着数据集成和数据治理内容的范围与边界。演化维度的 PDCA 过程要素是:业务系统升级或新业务数据接入,数据增量迭代与集成方案,方案评估与实施,方案调整与修订。

演化维度是数据治理体系的必然选择。从单一的信息系统发展到系统集成再到现在的云计算,数据伴随着业务和组织的更迭而不断累积和继续保持发展,需要有一种技术体系来管理和使用分散和异构的业务数据,最大化挖掘数据价值,支撑全局的数据分析应用。

4.4 管理维度

数据治理体系需要具有一套规范可行的管理制度(包括组织、人员、职责)。主要是从制度上明确数据的产生来源以及数据流通过程各个环节的各部门的职责。管理维度的要素是:形成数据共享使用规范,规范的实施与执行,规范修订,规范运行成效。

管理维度是数据治理体系的实施保障。数据源具有唯一性,唯一性判别的依据是数据产生和数据责任部门。管理维度针对数据的来源与责任部门可明确数据的职责划分,从管理的角度给出了数据的唯一提供部门、流通部门以及数据准确性的职责划分。

这四个维度是数据治理体系构成的有机统一。从四个维度综合考虑构建数据治理的体系,四个维度缺一不可。四个维度构成了整个的 PDCA 过程,对应到数据治理活动中,体现为数据归集、数据对标、质量提升、数据共享、数据服务 5 个步骤的一个数据治理的循环迭代过程。技术维度,提供技术工具和平台以实现异构数据源管理,数据质量检测与治理,数据发布与数据共享。业务维度,提炼基于数据的业务需求,包括数据共享与集成、主题数据分析、全局数据分析以及面向用户的基于业务场景的数据查询。演化维度,提出数据集成与数据治理的发展方向,基于异构的遗留系统实现数据集成与数据治理的统一,既满足于业务应用的发展、数据共享又服务于全局数据应用和统计分析。管理维度,提供组织内关于数据的使用的规范和原则。通过规范约束数据来源、权属与职责,明确数据职责划分。

综上所述,数据治理体系的闭环框架体系如图 2 所示。

5 关键技术分析

5.1 对标技术

对标技术是数据治理前进行数据集成时对来自数据源的数据进行必要的代码标准比对,即业务数据的参照代码是否采用了标准的代码表。分两种情况,如果采用了标准的代码表,那么完成对标。如果未采用或者未全部采用标准代码表,那么可能有三种情形:标准代码表包含业务数据;业务数据包含标准代码表;业务数据与标准代码表不一致。三种情形中,对前两种进行标准代码表的完善或修订,对于第三种,则需要在业务数据校对后进行标准代码表的扩充完善。完善或

修订后的标准代码表是所谓的执行标准代码表。

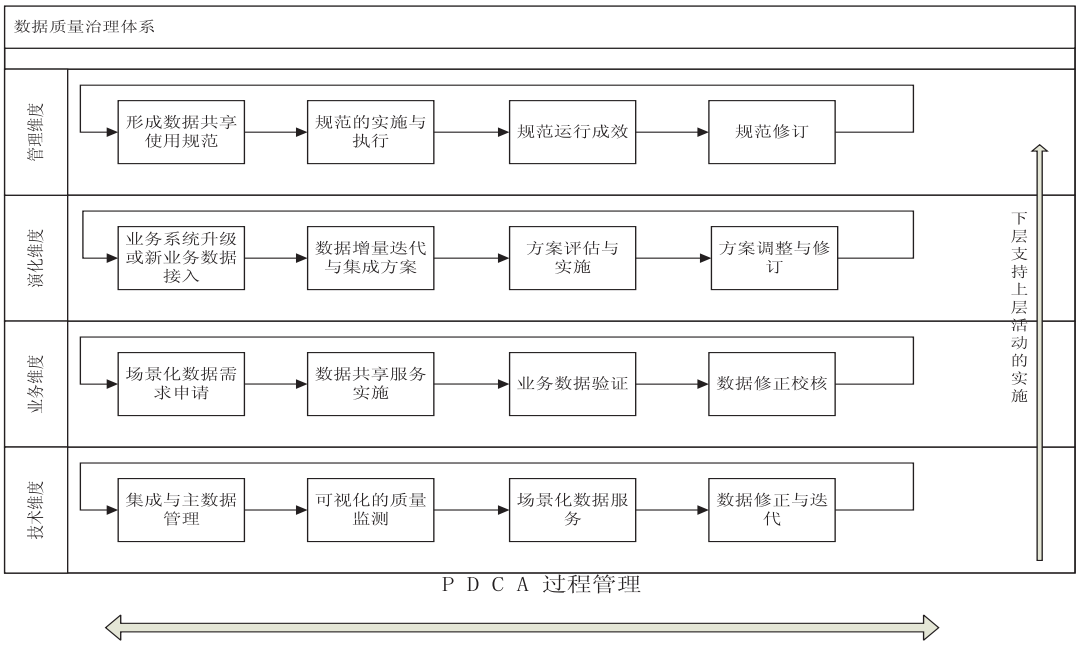

图 2　数据质量治理体系结构

5.2　基于规则的质量检测技术

基于规则的质量检测技术通过预制的质量约束规则,如唯一性、不为空、证件号码校验等,设置为需要检测的数据表的相应数据属性上。检测可以通过定时任务作为守护进程自动进行。这样对于数据质量检测点能够有效地进行质量检测,定时获取数据问题的明细报告,作为业务数据验证和修订的依据。基于规则的数据检测技术以身份证号码校验为例,

Sql clause:SELECT * FROM@ {TABLE_NAME}　　(1)

Whereclause:WHERE not regexp_like (@ {COLUMNS},″(^[1-9]\d{7}((0\d)|(1[0-2]))(([0|1|2]\d)|3[0-1])\d{3} $)|(^[1-9]\d{5}[1-9]\d{3}((0\d)|(1[0-2]))(([0|1|2]\d)|3[0-1])((\d{4})|\d{3}[Xx]) $)″)　　(2)

(1)加上(2)组成完整的质量约束规则,约束规则通过关联数据检测点属性,实例化具体的可执行质量规则实现对数据质量对应维度的检测。

5.3　场景化集成技术

数据集成技术是实现遗留系统环境下数据交换共享的最主要方法,在实施数据集成的过程中一般有数据表或视图、数据服务 api 等集成方式,(称之为数据集成切面)对外提供需要的数据服务。数据集成基于需要满足的业务场景需求,需要对数据集成切面进行准确的场景化描述。一般可以采用基于语义表达的描述方法、场景化数据切面标注方法。实现对数据集成切面的准确表达,可以在此基础上进行数据集成切面的重用或集成审计。

5.4　数据跟踪技术

基于质量规则进行数据质量监测和评估,能够有效地提供数据质量检测结果,作为集成数据质量提升的参考依据。通过主数据平台(master data system, MDS),对分散异构的数据源和数据进行注册登记,通过基于规则的数据质量监控实现主数据的质量监测,对于质量监测的结果进行问题数据的可视化输出,包括所属数据条目、数据源、问题数据记录、问题数据字段和具体数据等信息。通过问题数据结果的可视化输出,同时关联到所属的数据条目和数据源,通过这些信息就能够对上述的问题数据进行反馈、修正和再评估。

6　案例应用与分析

该校在数字校园一期建设了数据共享交换平台,打破不同业务应用之间的数据孤岛,能够通过集成技术对需要的数据进行共享。但随着业务应用的迭代优化以及数据分析决策需求的出现,原有数据交换与共享的平台不能很好地满足新的需求。主要表现在:

(1)遗留系统的共享数据未能有效管理;

(2)缺少可视化的数据管理和检索平台;

(3)不能进行数据质量的检测和数据标准的输出。

为了解决以上问题,建设了基于共享数据的治理平台。通过技术手段实现遗留系统的主数据条目化管理,对分散异构的主数据进行数据源的注册登记、数据的集成和对标、数据集成监控、数据质量检测以及数据主题化分析应用。到目前已经实现登记注册数据源 39 个,完成数据集成接口 276 个,具备数据共享交换

的业务数据表 64 个,配置数据质量检测规则 10 个,进行数据质量检测的表 3 个,检测属性 18 个。进行业务数据主题分析 2 个。数据源注册登记截图如图 3 所示,数据质量检测结果如图 4 所示。

数据源	驱动	连接信息	用户名
人事管理系统	oracle.jdbc.(	jdbc:oracle:thin:@(ES RI TC N =	SR_hr
科研管理系统	com.micros(	jdbc:sqlserver://20 .1 .1 .1 0:14	jc
财务（常州）	com.micros(	jdbc:sqlserver://21 .2 97. 1 3;D	kuser
财务工资	com.micros(	jdbc:sqlserver://20 .1 .1 .1 1:14	r_gzcx
研究生管理系统(挂	oracle.jdbc.(	jdbc:oracle:thin:@ 2. 19. 12 4:1	ydb_gx
一卡通管理系统	oracle.jdbc.(	jdbc:oracle:thin:@ 2. 19. 12 68:	r_ykt
研究生管理系统(T	oracle.jdbc.(	jdbc:oracle:thin:@ 2. 19. 12 4:1	ydb_gx
教务管理系统（常	oracle.jdbc.(	jdbc:oracle:thin:@(ES RI TC N =	wjw_share
教务管理系统	oracle.jdbc.(	jdbc:oracle:thin:@(ES RI TC N =	wjw_share

图 3 数据源注册登记

检测对象	检测字段	检测规则	扫描量	违规量	操作
T_JZG[教职工基	SFZJH[身份证	不符合身份证	5429	47	违规明细
T_YJS[研究生基	SFZJH[身份证	不符合身份证	48723	237	违规明细
T_BZKS[本专科	SFZJH[身份证	不符合身份证	103597	417	违规明细
T_JZG[教职工基	XM[姓名]	空值	5429	0	
T_JZG[教职工基	SFZJH[身份证	空值	5429	201	违规明细
T_YJS[研究生基	XM[姓名]	空值	48723	0	
T_YJS[研究生基	XBDM[性别代	空值	48723	401	违规明细
T_YJS[研究生基	XH[学号]	空值	48723	0	
T_YJS[研究生基	SFZJH[身份证	空值	48723	2620	违规明细

图 4 数据质量检测

通过数据治理平台的部署,以及相关关键技术(对标技术、数据质量检测技术、场景化集成技术、数据跟踪技术)的运用,为数据治理体系的构建提供了技术手段。同时基于技术维度和业务维度,探索形成数据的共享与管理的相关制度和规范,推动构建责任边界清晰的数据管理和治理体系,让数据能更好地发挥作用,满足用户需求。

7 结束语

数据治理过程是一个持续和不断完善的过程,需要按照系统化思维在每一个层面和维度同步推进。数据质量的提升不是数据治理的目的,满足业务需求和提升应用的服务质量才是数据治理的动力和方向。文中从系统化角度构建高校遗留系统环境的数据治理体系,也在实际的项目中进行治理理念的推行实施,取得了阶段成果。下一步会加强管理维度方面治理内容的建设,并且深入研究场景化集成技术在集成数据服务方面的成效。

参考文献:

[1] BOUFARES F, BEN S A. Heterogeneous data-integration and data quality: overview of conflicts[C]//2012 6th international conference on sciences of electronics, technologies of information and telecommunications. Sousse, Tunisia: IEEE, 2012: 867-874.

[2] 文 峰. 数据组织过程中的数据质量评价研究[J]. 软件导刊, 2013, 17(11): 132-134.

[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.

[4] 张国宝, 卞艺杰. 智慧校园数据质量治理的关键问题[J]. 中国教育网络, 2018(1): 51-52.

[5] 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, 13(11): 2076-2082.

[6] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2): 1-5.

[7] 丁小欧, 王宏志, 张笑影, 等. 数据质量多种性质的关联关系研究[J]. 软件学报, 2016, 27(7): 1626-1644.

[8] 梁吉胜, 李天阳, 王惠霞, 等. 基于约束的数据质量评估算法研究[J]. 科学技术与工程, 2012, 12(3): 551-554.

[9] 韩京宇, 宋爱波, 董逸生. 数据质量维度量化方法[J]. 计算机工程与应用, 2008, 44(36): 1-6.

[10] 曹建军, 刁兴春, 汪 挺, 等. 数据质量控制研究中若干基本问题[J]. 微计算机信息, 2010, 26(9): 12-14.

[11] 吴永欢. 数据质量管理模型及应用研究[J]. 中国电业: 技术版, 2013(8): 46-50.

[12] 贾春燕, 赵亚萍, 程艳旗. 高校数字校园数据质量管理研究[J]. 广西大学学报: 自然科学版, 2011, 36(s1): 272-275.

[13] 苏 博, 陈 溯, 唐成功. ERP 数据质量评估与数据治理方法研究[J]. 信息系统工程, 2012, 31(8): 140-144.

[14] ALRUITHE M, BENKHELIFA E, HAMEED K. Key dimensions for cloud data governance[C]//2016 IEEE 4th international conference on future internet of things and cloud. Vienna, Austria: IEEE, 2016: 379-386.

[15] MYRSETH P, STANG J, DALBERG V. A data quality framework applied to e-government metadata: a prerequsite to establish governance of interoperable e-services[C]//International conference on e-business and e-government. Shanghai, China: IEEE, 2011: 1-4.

[16] ALRUITHE M, BENKHELIFA E. Cloud data governance maturity model[C]//2017 8th IEEE international conference on software engineering and service science. Beijing: IEEE, 2017: 517-520.