数据要素视域下高校数据质量评价体系研究

杨树春，王义
（对外经济贸易大学网络安全和信息化处，北京 100029）

摘　要：数字化时代，数据要素作为新型生产要素，其质量表现直接影响着数据要素效能的发挥。结合高校信息化建设的特点，深入分析了数据要素视域下的高校数据质量困境，归纳提出了涵盖完整性、唯一性、准确性、规范性、一致性、及时性6个方面的高校数据质量属性。从高校业务场景出发，建立了以业务影响为主导的高校数据质量评价体系，有助于高校对业务数据开展有效的质量评价和问题精准识别，为高校数据治理提供方向和依据。

关键词：高校；数据质量；质量评价
中图分类号：G647
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024236

Research on the evaluation system of university data quality from the perspective of data elements

YANG Shuchun, WANG Yi
Department of Cyberspace Affairs, University of International Business and Economics, Beijing 100029, China

Abstract: In the digital era, data factor is a new production factor, its quality performance directly affects the performance of data factor. Based on the characteristics of informatization construction in colleges and universities, the dilemma of data quality in colleges and universities was deeply analyzed from the perspective of data elements, six attributes of data quality in colleges and universities were summarized and put forward, including completeness, uniqueness, accuracy, normalization, consistency and timeliness, and a data quality evaluation system was established in colleges and universities with business impact as the leading role from the business scenario. It is helpful for colleges and universities to carry out effective quality evaluation and accurate identification of problems on business data, and provide direction and basis for data governance in colleges and universities.

Keywords: university, data quality, quality evaluation

0　引言

2020年，国务院将数据定位为新型核心生产要素、国家基础性资源和战略性资源。对于高校而言，数据是实现数字化转型的关键要素，数据治理成为高校数字化转型的首要任务。高校数据治理的目标是通过建立一套包含制度、流程等在内完整的工作体系，提升并保证数据质量，最大化发挥数据资源价值。数据质量是衡量数据治理成效的一个重要指标，直接影响到数据资源价值的发挥。高校业务部门广、系统数量多，如何对高校数据质量进行有效的评价，成为数据治理的重要工作之一。

Maffei最早意识到数据质量问题及数据质量评

收稿日期：2024–10–25
基金项目：2022年度高等教育科学研究规划课题基金资助项目（No.22XX0405）；对外经济贸易大学工会工作专项研究课题基金资助项目（No.2020ghzd001）
Foundation Items：2022 Annual Funding Project for Scientific Research on Higher Education (No.22XX0405), Funded Project of Special Research on the Work of the Trade Union at the University of International Business and Economics (No.2020ghzd001)

估的困难[1]。数据质量是指在具体业务环境下，数据符合数据消费者的使用目的，满足业务场景需求的程度[2]。本文结合高校信息化建设特点，从数据要素的深度视角出发，剖析了当前高校在数据质量方面所面临的困境，系统性地归纳并提出了高校数据质量属性，进而构建了一个以业务影响为核心驱动力的数据质量评价体系，不仅能够有效指导高校对业务数据进行全面、客观的质量评估，更为高校数据治理工作提供了明确的方向指引，从而显著提升数据资产的价值，促进高校管理决策的科学化、智能化水平。

1　高校数据质量困境

数据质量伴随着高校业务数据的产生而来，业务系统的建设成效往往体现在相应业务数据的质量上。在高校信息化建设的过程中，一直存在着顶层规划不足、部门协同不够、数据意识薄弱等问题，导致高校数据质量困境不仅仅有技术上的因素，也包含着管理上的因素。这2个因素交叠在一起造成了当前高校面临的数据质量困境，主要表现在以下几点[3-5]。

1) 数据多源问题。各业务系统长期分散建设，客观上导致部分数据重复采集，相应数据缺乏规范性、一致性，需要围绕“一数一源”开展大量工作。

2) 数据异构问题。各业务部门对于数据有不同的认知，也有不同的规范，导致数据无法直接互通共用，需要进行大量数据清洗转换工作。

3) 数据缺失问题。各业务系统建设前期规划不足，部分关键数据未得到有效采集，部分重要的数据项有缺失，导致数据的可利用性降低。

4) 数据易变问题。部分源头系统常常在未提前通知的情况下进行变更表结构、直接删除数据等操作，导致下游系统数据不一致，降低了数据的可靠性。

5) 数据录入问题。大部分业务系统缺乏对数据完整性、规范性、唯一性等的检验，导致数据录入时就存在不完整、编码不规范、模型不准确、参数不匹配等问题。

6) 数据反哺问题。数据从源头系统共享后，在使用中会有更新或产生衍生数据等情况，大多数情况下未考虑将这些数据反哺到源头系统中，形成另一层面的数据割裂。

2　高校数据质量属性

数据质量体现在数据多个维度的属性上，2019年颁布实施的《信息技术—数据质量评价指标》中将数据质量属性划分为6类：规范性、完整性、准确性、一致性、时效性、可访问性。结合上述质量属性分类，从高校业务场景的特点出发，本文认为高校数据质量属性可归纳为完整性、唯一性、准确性、规范性、一致性、及时性6个方面[5-8]。

1) 完整性指数据主体的完整程度，包括数据字段完整、数据值完整、数据属性完整3个方面。数据字段缺失和部分数据项缺失会降低数据的可利用性，导致数据价值打折扣。完整性对所有数据主体都具有普适性。

2) 唯一性指数据主体不存在重复记录，数据内容唯一的程度。重复的数据主体不仅会导致数据无法正常共享，更会导致业务之间无法协同、数据追溯困难等诸多问题。唯一性是数据主体能够被共享共用的基础。

3) 准确性指数据主体与其对应的客观实体的特征相一致的程度，数据不准确表现在数据值的长度、内容出现异常或者错误，不符合客观实际。数据不准确不仅会导致数据可用性降低，还可能会造成错误的数据导向。

4) 规范性指数据主体遵循数据标准、数据模型、业务规则等预定语法规则的程度，具体表现在数据的命名、类型、格式、值域等方面遵从既定约束，不规范的数据需要进行大量清洗转换，增加数据共享共用的成本。

5) 一致性指数据主体与其他上下文数据主体的一致程度，即相同的数据主体在不同的应用场景/数据集中是否遵循了统一的规范、统一的格式等。数据的一致性不单单意味着数值上的相同，也要考虑收集、处理的方法和标准的一致。

6) 及时性指从业务发生变更开始计算，到结论性数据得到更新的时间间隔。及时性是业务处理和管理效率的关键指标，对于业务间的协同非常关键。业务操作人员的职责规范、处理效率能够直接影响到数据的及时性。

3　高校数据质量评价体系

高校对于数据质量进行评价，应该围绕着具体的业务场景来构建科学的、符合实际的评价体系。

在明确高校数据质量属性基础上，结合造成高校数据质量困境的因素，接下来选择对数据主体所属业务有较大影响的质量属性来设计评价规则和权重，构建数据质量评价体系，如图1所示。

1) 确定数据主体的质量属性。不同数据主体所需要评价的数据质量属性是存在着差异，需要数据管理者根据数据所对应的业务场景来进行明确。例如，针对师生基础数据而言，身份证号字段需要考虑的质量属性包括完整性、准确性、规范性、唯一性，而姓名字段涉及的质量属性只包括完整性、准确性。

2) 根据质量属性编制针对数据主体的质量评价规则 $R=\{R_1,R_2,\cdots,R_n\}$，其中 n 表示规则数量。不同数据主体的评价规则存在差异，规则数量也存在差异，具体评价规则由业务数据的意义和作用而决定。例如，对于唯一性而言，涉及单个字段、多个字段组合的重复性检验，如可根据身份证号这一个字段是否重复来检验人员数据的唯一性，干部任免数据的唯一性则需结合职工号、任职日期、任职职务3个字段进行组合检验。表1展示了不同质量属性下的评价规则。

3）为数据质量评价规则逐一设定权重值，得到 $w=\{w_1,w_2,\cdots,w_n\}$，其中 n 表示规则数量。需要注意的是，同一个数据主体下各评价规则的权重是不相同的，不同数据主体下同一个评价规则的权重也是不同的。例如，学工号字段的“数据标识唯一性”权重较高，而姓名字段的“数据值完整性”权重较高。实际操作中，所有评价规则的权重应根据

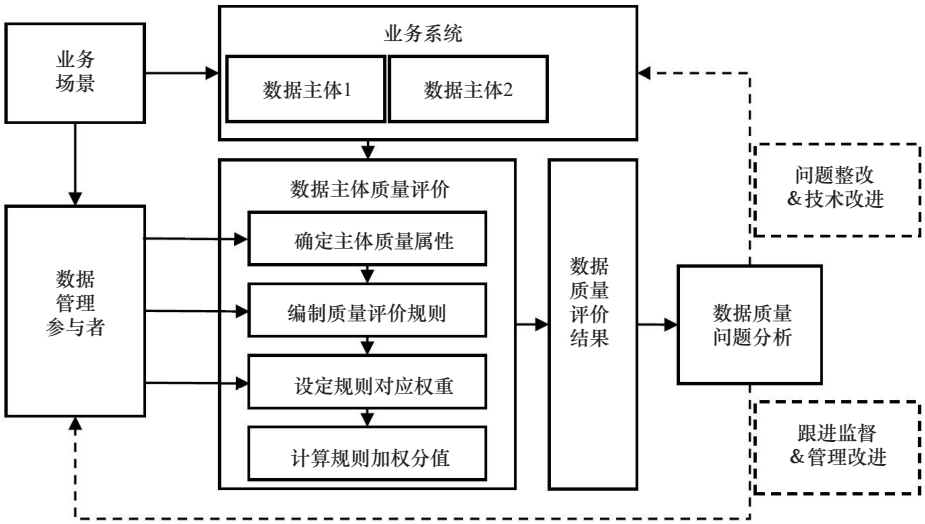

图1　高校数据质量评价体系

表1　　**不同质量属性下的评价规则**

数据质量属性	评价规则	规则含义
完整性	数据字段完整性	数据表满足业务需求的字段覆盖程度
	数据值完整性	数据字段中值的完整程度
	数据属性完整性	数据表/字段属性描述的完整程度
唯一性	数据标识唯一性	数据表是否有唯一主键的程度
	数据条目唯一性	数据表中数据内容唯一的程度
准确性	数据值准确性	数据值在有效值域范围的程度
	数据属性准确性	数据表/字段属性描述的准确程度
规范性	数据编码规范性	数据编码与学校标准一致的程度
	数据类型规范性	数据字段类型与相应业务一致的程度
一致性	数据属性一致性	同一数据字段在不同数据表的格式、注释一致的程度
	数据内容一致性	同一数据字段在不同数据表的取值一致的程度
	数据逻辑一致性	同一数据字段在不同位置逻辑计算一致的程度
及时性	数据更新及时性	实际业务变更后，数据能及时更新的程度
	数据更新差异性	不同数据字段更新频率差异的程度

数据主体的应用场景，由有高校数据参与经验的专家（包括但不限于从事管理、技术、业务工作）通过打分而综合评判得出。这种多指标权重赋值常采用层次分析法（AHP, analytic hierarchy process）来实现，但其判断矩阵的一致性检验较为严格且存在不通过的情况，因此本文采用标度扩展法来构建判断矩阵[9]，进而计算每条评价规则的权重值具体步骤如下。

①参与专家对评价规则 $R=\{R_1,R_2,\cdots,R_n\}$ 分别评价，按照重要程度不减的方式进行排序，不失一般性，假设某一专家的排序为 $R_1 \geqslant R_2 \geqslant \cdots \geqslant R_n$。

②分别对 $i=1,2,\cdots,n-1$，将规则 R_i 与 R_{i+1} 进行比较，并将其对应的比较标度值记为 t_i，其取值范围可以选取表2内容。

表2　　比较标度值

标度值	解释
1	R_i 与 R_{i+1} 同等重要
1.2	R_i 与 R_{i+1} 略微重要
1.4	R_i 与 R_{i+1} 相当重要
1.6	R_i 与 R_{i+1} 明显重要
1.8	R_i 与 R_{i+1} 绝对重要

③设计判断矩阵 $\boldsymbol{P}=(p_{ij})_{n\times n}$，遍历 i,j，如果 $i=j$，则 $p_{ij}=1$，否则 $p_{ij}=t_it_{i+1}\cdots t_{j-1}$，$p_{ji}=\dfrac{1}{p_{ij}}$，得到如下的判断矩阵

$$\boldsymbol{P}=\begin{bmatrix} 1 & t_1 & t_1t_2 & \cdots & t_1t_2\cdots t_{n-1} \\ \dfrac{1}{t_1} & 1 & t_2 & \cdots & t_2t_3\cdots t_{n-1} \\ \dfrac{1}{t_1t_2} & \dfrac{1}{t_2} & 1 & \cdots & t_3t_4\cdots t_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \dfrac{1}{t_1t_2\cdots t_{n-1}} & \dfrac{1}{t_2t_3\cdots t_{n-1}} & \dfrac{1}{t_3t_4\cdots t_{n-1}} & \cdots & 1 \end{bmatrix}$$

④计算上述判断矩阵的特征向量，归一化后计算得出参与专家认定的各规则的权重，记为 $w_s=\{w_{s1},w_{s2},\cdots,w_{sn}\}$，其中 s 表示第 s 位参与专家。

⑤针对每一项规则权重，分别计算所有专家评价权重的平均值（或按照参与专家的权威性分配权重），进而得出各项规则的最终权重 $w=\{w_1,w_2,\cdots,w_n\}$。

4) 依照规则进行评价，并根据规则权重计算加权分值。在逐条依照评价规则对数据主体展开评价时，可采用定性评价与定量评价相结合的方式进行评分。定性评价侧重于主观描述，即通过用户反馈法、专家评议法对数据主体进行评价，从而得出百分制表示的评价结果。对于及时性属性而言，因其与业务操作的相关性较强，适宜采用定性评价的方法。定量评价完全依据数据客观表现，因此采用“简单比率法”进行评估，即通过计算没有问题的数据单元数在总数据单元数中的比率进行评价，适宜于完整性、准确性等其他属性评价，具体公式为

$$q_i=\frac{\text{第}i\text{个数据质量评价规则下没有问题的数据单元数}}{\text{数据主体总数据单元数}}\times 100$$

其中，i 表示第 i 个数据质量评价规则。在计算出数据主体的每个规则下的评分后，就可以通过对评分进行加权求和 $Q=\sum_{i=1}^{n}w_iq_i$，得到数据主体的最终评价得分，进而可以得出对相应业务系统整体数据的质量评价。

4　结束语

数据质量是数据资源发挥价值的关键。本文首先分析高校数据质量困境表现，然后归纳高校数据质量属性，最后建立以业务影响为主导的数据质量评价体系，实现了对高校数据质量的有效评价。实际工作中，高校还需要根据评价结果来进一步分析定位数据产生质量问题的原因，为数据质量提升找准解决方案。作为业务系统运维单位，要及时进行问题整改，完善技术和管理手段，从源头上提升数据质量。作为数据管理管理，要对整改过程进行跟进监督，验证整改成效，形成

整改闭环。这样双方协作，才能促进高校数据质量稳步提升，使数据为高校日常的管理和决策提供更好的支撑。

参考文献：

[1] 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, 12.
HAN J Y, XU L Z, DONG Y S. An overview of data quality research[J]. Computer Science, 2008, 35(2): 1-5, 12.

[2] 巫莉莉, 张波. 高校数据治理中提升数据质量的方法研究[J]. 重庆理工大学学报(自然科学), 2019, 33(8): 149-156.
WU L L, ZHANG B. Research on the method of improving data quality in university data governance[J]. Journal of Chongqing University of Technology (Natural Science), 2019, 33(8): 149-156.

[3] 缪亚琴, 陈丽蓉. 数据质量提升之道[J]. 中国教育网络, 2016(4): 25-27.
MIAO Y Q, CHEN L R. The way to improve data quality[J]. China Education Network, 2016(4): 25-27.

[4] 康军广, 周静. 浅谈高校数据治理过程中存在的共性问题及其对策[J]. 信息系统工程, 2021(5): 39-40.
KANG J G, ZHOU J. On the common problems and countermeasures in the process of data governance in colleges and universities[J]. China CIO News, 2021(5): 39-40.

[5] 盛小平, 田婧, 向桂林. 科学数据开放共享中的数据质量治理研究[J]. 图书情报工作, 2020, 64(22): 11-24.
SHENG X P, TIAN J, XIANG G L. Research on data quality governance in open sharing of scientific data[J]. Library and Information Service, 2020, 64(22): 11-24.

[6] 陈远, 罗琳, 沈祥兴. 信息系统中的数据质量问题研究[J]. 中国图书馆学报, 2004, 30(1): 48-50.
CHEN Y, LUO L, SHEN X X. A study of data quality in information system[J]. The Journal of the Library Science in China, 2004, 30(1): 48-50.

[7] 丁海龙, 徐宏炳. 数据质量分析及应用[J]. 计算机技术与发展, 2007, 17(3): 236-238.
DING H L, XU H B. Data quality analysis and application[J]. Computer Technology and Development, 2007, 17(3): 236-238.

[8] 数据治理. 什么是数据标准？如何制定数据标准？[R]. 2020.

[9] 黄德才, 郑河荣. AHP方法中判断矩阵的标度扩展构造法[J]. 系统工程, 2003, 21(1): 105-109.
HUANG D C, ZHENG H R. Scale-extending method for constructing judgment matrix in the analytic hierarchy process[J]. Systems Engineering, 2003, 21(1): 105-109.

[作者简介]

杨树春（1989– ），男，山西孝义人，对外经济贸易大学工程师，主要研究方向为教育信息化、数据治理。

王义（1975– ），女，山东淄博人，对外经济贸易大学工程师，主要研究方向为教育信息化、IT治理。