Data Quality Control Patterns in University Research Project Cycles Exploratory Study Postprint

Authors: Tang Jing, Qu Wenjian

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] Aims to provide an effective control approach and method for data quality control in the university research project lifecycle. [Method / Process] Constructs a data quality control architecture system around the research project lifecycle and data quality control lifecycle, and under this framework, implements data quality control from three perspectives—cognitive, managerial, and process—introducing control methods such as the quality gap model, enterprise architecture model, and process analysis to analyze the mechanism of scientific research data quality control in university research project lifecycles. [Results / Conclusion] Constructs a scientific research data quality control architecture system and its data quality control model applicable to the research project lifecycle, providing theoretical support for university scientific research data quality control.

Full Text

Data Quality Control Mode in University Scientific Research Project Cycles

Tang Jing[1], **Qu Wenjian**[2]

[1]School of Information Management, Wuhan University, Wuhan 430061
[2]Department of Information Management, School of Management, Nanchang University, Nanchang 330031

Abstract

[**Purpose/Significance**] This study aims to provide an effective approach and methodology for data quality control within university scientific research project cycles. [**Method/Process**] The research constructs a data quality control architecture system centered on both the research project cycle and the data quality control cycle. Under this framework, data quality control is implemented from three perspectives—cognition, management, and process—while introducing control methods such as the quality gap model, enterprise architecture model, and process analysis to dissect the mechanisms of scientific research data quality control in university project cycles. [**Result/Conclusion**] The study establishes a data quality control architecture system and corresponding control modes suitable for research project cycles, providing theoretical support for university scientific research data quality control.

Keywords: data quality; research project cycle; research environment; quality gap model; enterprise architecture model

1. Introduction

With the emergence of the fourth research paradigm—the data-intensive scientific research environment—research data volumes have grown exponentially with diverse manifestations. Researchers now face staggering amounts of data to store and process, with complex and varied data sources and structures, creating significant obstacles to advancing research projects. Moreover, research projects are filled with data resources from complex sources and diverse storage structures, with poor technical compatibility for data sharing and insufficient data management awareness among project teams, resulting in low-quality research data and underutilization. Universities possess unique advantages in research data resources and serve as the main force in generating and utilizing research data for scientific inquiry. Therefore, universities urgently need to strengthen data quality control in data-intensive research environments and explore, construct, and improve their research data management frameworks under digital research contexts.

Reviewing domestic and international research on university research data quality control reveals that scholarly contributions have focused on establishing theoretical systems and practical applications. In terms of theoretical frameworks, DAMA's division of data quality control into ten functional domains has become a widely recognized data management knowledge system [1]. Bao Dongmei et al. innovatively proposed the CALib framework for university library data governance, promoting effective utilization of library data resources [2]. Qu Wenjian et al. established a university research data quality control architecture based on Krantz's theory and analyzed its operational mechanisms, providing effective theoretical support for research teams [3]. In practical applications, Ma Bo et al. drew upon European and American university library practices to propose strategies for strengthening data governance awareness, infrastructure development, and talent cultivation to improve domestic university research data governance levels [4]. Xu Huifang et al. conducted surveys on domestic research data management practices and performed SWOT analyses to propose effective pathways for implementing research data services and quality control from multiple perspectives [5]. However, existing research reveals that under the digital research paradigm, data quality control modes still have considerable room for improvement when dealing with massive data scales.

Based on this analysis, this paper constructs a university research data quality control architecture system around the research project cycle and proposes control modes from cognitive, management, and process perspectives, providing theoretical reference for research institutions and libraries to undertake research data quality engineering.

2. Necessity of Implementing Data Quality Control in Research Project Cycles

2.1 Connotation of Research Project Cycles and Related Data Activities

The research project lifecycle represents a management concept describing the developmental stages and processes of "planning—initiation—implementation—closure." For analytical convenience, this study adopts Lu Yuquan's four-stage division method, segmenting the research project cycle into research conception, research implementation, results compilation, and results publication [6] [Figure 1: see original paper], to examine data quality control in university research project cycles.

[**Figure 1: see original paper**] *Data-related activities in university research project cycles*

(1) Research Conception Stage. This stage forms the foundation for project selection and proposal submission, reflecting the research team's capabilities and the project's novelty and cutting-edge nature. Primary data activities involve data resource collection, including extensive data gathering, stakeholder identification, project topic selection, and data quality planning. Data sources mainly rely on sharing alliances, active uploads, and direct acquisition [7], encompassing literature review, tracking of domestic and international research trends, and technical feasibility analysis.

(2) Research Implementation Stage. In this phase, teams conduct academic writing and scientific experiments according to the plan developed during conception. This represents the golden period for data production and discovery, requiring strict adherence to data standards established earlier and appropriate preprocessing tools to obtain high-quality data from the outset.

(3) Results Compilation Stage. After obtaining raw research data, teams must employ data mining, database management, and quality detection technologies to extract, integrate, and mine data for more accurate experimental results. The focus is on integrating, analyzing, and visualizing data to intuitively reflect genuine relationships between phenomena.

(4) Results Publication Stage. Following refinement and adjustment, research outcomes take shape and enter review, examination, and publication phases. Project evaluation committees assess result quality, and upon approval, data is uploaded to sharing platforms for archiving and dissemination while the team prepares for subsequent projects.

2.2 Data Quality Issues and Causes in University Research Project Cycles

Despite growing scholarly emphasis on its importance, research data—critical components of university project cycles—still face numerous quality problems, including errors, incompleteness, inconsistency, and untimeliness . Causes include: (1) **Complex and diverse data sources.** Team members from different disciplinary backgrounds generate data without unified standards, exacerbating data silos and challenging collection and organization. (2) **Small data scale.** While individual project teams produce limited data, the sheer number of teams makes high-quality data difficult to locate. (3) **Insufficient quality control awareness.** Team members have yet to develop data management habits, showing deficiencies in data collection capabilities, sharing willingness, and security consciousness.

**** *Primary data quality issues in research project cycles*

Comprehensive analysis of these issues reveals that to transform disordered, chaotic research data into reusable resources, universities must establish reliable data quality evaluation standards based on accuracy, completeness, consistency, and timeliness, construct a practical architecture for data quality engineering, and achieve innovative value-added data utilization, improved project management efficiency, and enhanced team quality control awareness.

2.3 Significance of Implementing University Research Data Quality Control

Considering the characteristics of university research data, data requirements, and research environments, implementing data quality control in research project cycles offers several key benefits: (1) **Promoting cross-disciplinary data organization and integration.** Research data inherently features interdisciplinary characteristics due to project needs and heterogeneous team backgrounds, making universal standards difficult to formulate and reducing utilization rates. Quality control enables multi-dimensional organization and classification of data across project stages, enhancing reusability. (2) **Meeting demands for high-quality data.** Defective, incomplete, or misleading data can cause experimental failure, skyrocketing costs, or even project termination. Accurate, complete data can accelerate project progress, alleviate research pressure, and increase data value. (3) **Essential requirement for integrating into the fourth research paradigm.** In this new paradigm, research data has become a fundamental resource [8], with existing data serving as valuable assets for new studies [9]. High-density research big data has profoundly impacted traditional management, making quality control architecture construction imperative.

3. University Research Data Quality Control Architecture System in Project Cycles

University research data quality control is a long-term, complex process with varying data sources, quality requirements, and control methods across project stages. This study constructs a control architecture from two dimensions—the research project cycle and the quality control cycle [Figure 2: see original paper]. Data activities involve five quality roles (project planners, data collectors, technical supporters, data analysts, and records managers) who conduct quality control through assessment, attribution, and action phases. Based on assessment results, they identify root causes and develop improvement solutions from cognitive, management, and process perspectives, forming a continuous quality control cycle.

[**Figure 2: see original paper**] *Data quality control architecture system in research project cycles*

3.1 Key Roles in University Research Data Quality Control

Five primary roles influence data quality and value throughout project and control cycles:

(1) Project Planners. Including project leaders and data architects, they collect data requirements during conception, develop data management plans, and integrate quality improvement into project design to reduce risks and allocate resources rationally.

(2) Data Collectors. Team members who conduct experiments or gather data, they control initial data sources and are responsible for capturing, creating, or maintaining data.

(3) Technical Supporters. Database administrators and IT support staff who develop storage platforms, implement metadata management, establish validity rules, and ensure logical integration and spatial scalability.

(4) Data Analysts. Knowledge workers who apply data mining and modeling technologies during results compilation to analyze raw data, detect and cleanse low-quality data, and uncover underlying relationships.

(5) Records Managers. Responsible for collecting and archiving data throughout both cycles and documenting status at each lifecycle stage. Since data is often dispersed among team members, maintaining complete documentation of status, operations, and environmental information is crucial for reducing data loss and strengthening management [10].

3.2 Data Quality Control Cycle

The control cycle comprises three high-level steps—assessment, attribution, and action—providing a simple framework for discussing "control" [11].

Assessment Phase. This foundational phase involves analyzing data requirements and environments, defining quality dimensions (correctness, completeness, consistency, and timeliness), and conducting evaluations. Comprehensive planning ensures targeted project implementation, while standardized dimensions provide unified measurement criteria.

Attribution Phase. The core task is identifying root causes of quality issues. Managers collect background information on low-quality data, trace data flow paths, prioritize problems, and determine root causes. If causes fall outside predefined dimensions, the process returns to assessment for reanalysis; if within scope, specific recommendations are developed and prioritized.

Action Phase. This phase aims to correct current errors and prevent future ones. Based on attribution results, improvement solutions primarily involve three modes: quality gap model, enterprise architecture model, and process analysis. These provide theoretical frameworks and practical foundations for quality enhancement.

4. Analysis of University Research Data Quality Control Modes in Project Cycles

To enable managers to control data quality throughout project cycles and identify, prioritize, and repair defects before substantial impact, this study proposes an integrated control mode based on the three-phase architecture [Figure 3: see original paper]. The assessment and attribution phases continuously monitor data quality, evaluate it according to rules and dimensions, identify and classify issues, and develop improvement solutions. The action phase employs three perspectives:

[Figure 3: see original paper] *Overall data quality control mode for university research project cycles*

4.1 Quality Gap Model-Based Control

Marketing scholars A. Parasuraman and L. Berry's service quality gap model analyzes quality issues by comparing perceived versus expected service quality [12], while J. R. Evans advocates a "customer-driven quality" approach [13]. In university research cycles, gap analysis between expected and perceived data quality is critical for identifying root causes. This model precisely repairs quality issues through gap comparison [Figure 4: see original paper].

[Figure 4: see original paper] *Quality control mode based on the quality gap model*

Factors influencing expected data quality include historical data reliability, institutional promotion, and data demand levels. Perceived quality derives from collection reliability, standard scientificity, analytical capabilities, and team data literacy. Control personnel employ surveys to understand these gaps and conduct analyses covering expectation gaps, standard gaps, transmission gaps, communication gaps, and overall perception gaps—the latter being pivotal, as the first four directly affect its magnitude. Based on gap analysis, managers conduct rapid configuration tests, record datasets requiring correction, and perform root cause analysis focusing on identifying error sources and understanding error introduction mechanisms. Prioritizing causes links business value with high-quality data and may reveal previously overlooked issues. If causes are within scope, corrective actions include data cleansing, root cause elimination, and monitoring/prevention to improve quality.

4.2 Enterprise Architecture Model-Based Control

Given the interdisciplinary and cross-departmental nature of university research, enterprise architecture—widely applied in information resource integration [14] and data governance [15]—can be introduced for data quality control. Enterprise architecture typically comprises four layers: business, application, data, and technology [16] [Figure 5: see original paper].

[**Figure 5: see original paper**] *Data quality control mode based on enterprise architecture model*

The **technology layer** forms the foundation, involving network transmission services, IT infrastructure, big data processing technologies, and intellectual support. Network and IT infrastructure provide 保障 for data activities, while big data technologies—including association analysis, intelligent prediction, knowledge graph analysis, and clustering—enable effective raw data processing.

The **data layer** encompasses data collection, quality control, and storage, providing multi-dimensional data sources and preprocessing tools. Sources include sharing alliances, network resources, and project histories. Quality control follows the cycle's steps to prevent and correct errors, with standardized storage formats ensuring archival consistency.

The **application layer** further improves quality through extraction, integration, and mining. Extraction involves selection and cleansing; integration involves filtering and merging to reduce redundancy; mining uncovers underlying knowledge through algorithms to yield ideal research outcomes.

The **business layer** provides an integrated window for monitoring and maintaining data activities throughout the project. It includes requirement/environment analysis, data standard formulation, and data management planning during conception, establishing unified scales for quality assessment and control.

4.3 Process Analysis-Based Control

Data quality comprises both process and result quality, with the latter reflecting each project phase. Improving process quality enhances result quality. Therefore, process analysis enables planned, organized, and continuous quality improvement from a micro perspective [Figure 6: see original paper].

[**Figure 6: see original paper**] *Data quality control mode based on process analysis*

In the **conception stage**, planners and records managers focus on data resource collection, requirement/environment analysis, and quality standard determination. Collection from sharing platforms must ensure accuracy and completeness from the source. Requirement analysis examines team quality needs and cutting-edge trends, while standard formulation ensures reliability and scientificity.

In the **implementation stage**, producers generate raw experimental data, and technical supporters preprocess it using big data technologies—including deduplication, error correction, and format standardization—to obtain structured data meeting standards. Rigorous scientific activity fundamentally impacts subsequent quality.

In the **compilation stage**, the focus shifts from data discovery to knowledge discovery, and from collection to creation. Data extraction, integration, and mining depend on technical proficiency and data literacy, revealing underlying relationships for more precise results.

In the **publication stage**, knowledge workers integrate and mine compiled data into preliminary reports. After removing outliers and making appropriate adjustments, results are archived in research data repositories that structure data, ensure consistency, and update in real-time for high-quality outcome sharing.

5. Conclusion

In the big data era, the fourth research paradigm ultimately revolves around mining and controlling massive datasets [6], making research data quality control critical for in-depth study. This research elucidated the necessity of quality control by examining project cycle connotations and data activities, constructed a university research data quality control architecture from project cycle and control phase perspectives, and developed three complementary control modes from cognitive, management, and process viewpoints. These provide theoretical support and effective pathways for research data quality control in university project cycles.

References

[1] DAMA International. *DAMA Guide to the Data Management Body of Knowledge*[M]. Beijing: Tsinghua University Press, 2012:4-9.
[2] Bao Dongmei, Fan Yingjie, Li Ming. Data Governance and Its Framework in

University Libraries[J]. *Library and Information Service*, 2015, 59(18):134-141.
[3] Qu Wenjian, Tang Jing, Chen Danzhi. Research on University Scientific Research Data Quality Control Architecture and Mechanism[J]. *Information Studies: Theory & Application*, 2018, 41(11):45-50.
[4] Ma Bo, Li Yu. Practice and Enlightenment of Research Data Management in European and American University Libraries[J]. *Library Work and Study*, 2018(8):17-24.
[5] Xu Huifang. Investigation on Current Status of Research Data Management Practice in China and Its Implications for Libraries[J]. *Research on Library Science*, 2019(17):45-53.
[6] Lu Yuquan. Research on Service Model and Construction Path of University Scientific Research Data in Big Data Era[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017.
[7] Xiong Wenlong, Li Ruinan. Research on Library Data Services Based on Scientific Data Management[J]. *Library and Information Service*, 2014, 58(22):48-53.
[8] Sun Jizhou. Research on Pathways for University Libraries to Carry Out Scientific Data Management and Sharing under E-Science Environment[J]. *Library*, 2016(5):66-71.
[9] Marchionini G, Yang Guancan, Lu Kun. Research Data Management: Ensuring Data Quality, Promoting New Scientific Research in iSchools[J]. *Documentation, Information & Knowledge*, 2013(4):4-9.
[10] Ding Ning, Ma Haoqin. Comparative Study and Reference of Foreign University Library Scientific Data Lifecycle Management Models[J]. *Library and Information Service*, 2013, 57(6):18-22.
[11] McGilvray D. *Executing Data Quality Projects*[M]. Translated by Cao Jianjun, Diao Xingchun, Zhang Jianmei. Beijing: Publishing House of Electronics Industry, 2010.
[12] MBA 智库百科. Norbert Wiener's Control Theory[EB/OL]. [2012-04-14]. http://wiki.mbalib.com/wiki/Control_{theory}.
[13] Evans J R, Lindsay W M. *The Management and Control of Quality*[M]. 5th ed. St. Paul: West, 2002.
[14] Qu Wenjian, Zhou Xiaoyu, Meng Shurong. Research on Integrated Services of Digital Library Science and Technology Information Resources Based on Enterprise Architecture[J]. *Research on Library Science*, 2017(13):32-39.
[15] Guo Lusheng, Liu Chunnian. Research on Emergency Data Quality Governance in Big Data Era[J]. *Information Studies: Theory & Application*, 2016, 39(11):101-105.
[16] CIO Council. *Federal Enterprise Architecture Framework*[EB/OL]. [2016-02-03]. http://www.enterprise-architecture.info/Images/Documents/Federal%20EA%20Framework.pdf.

Author Contributions

Tang Jing: Conceived research ideas, designed methodology, wrote manuscript;
Qu Wenjian: Conceived research ideas, supervised manuscript revision.