Solution Ideas and Practices for Data Governance Engineering in Colleges and universities

Qiao Xie[1,a], HuanMing Zhang[2,b], Yirong Tang[3,c],Min Lin[4,d]

[1]Center for Internet and Educational Technology Jinan University Guangzhou, China
[2]Center for Internet and Educational Technology Jinan University Guangzhou, China
[3]Center for Internet and Educational Technology Jinan University Guangzhou, China
[4]Center for Internet and Educational Technology Jinan University Guangzhou, China

***Abstract*—**Effectively improving and enhancing the data quality of colleges and universities is the most fundamental goal of university data governance. In this paper, first of all, the causes of data quality problems in colleges and universities are analyzed, and then the general idea of improving data quality is put forward, in which introduces how to identify data quality problems, how to solve data quality problems through technical and business means and how to make the data quality management of colleges and universities form a long-term mechanism through the construction of the system.

1 CAUSES OF DATA QUALITY PROBLEMS

In recent decades, the construction of digital campus in colleges and universities has shown an obvious characteristic, that is, it attaches importance to process, but ignores data and lacks standards. Such a background causes a lot of problems in the data of colleges and universities in China at present, which can be roughly summarized in four aspects. The problems and their causes are as follows:

(I) Problems concerning the overall information architecture. (1) In terms of the causes of this problem, the first and foremost is the lack of uniform and standardized information standards. As each department's business system was previously built independently, there is a lack of norms and consensus that can be used by all departments for reference and implementation, so the codes and coding used by each business department in the construction of the system are not the same. For example, we often see professional codes, department codes and teacher numbers, even the most important information such as work number and student number may be inconsistent. (2) In addition, the data interconnection and exchange and sharing in many schools are still inadequate, and there are many data islands, which are manifested by the inconsistent contents of the basic data used by different departments. For example, the fixed asset lists of the national capital department and the finance department may be different, and the personnel lists of the academic affairs department and the personnel department may be inconsistent, which may lead to inconsistency in the caliber of the data at the time of statistics and the inability to find an accurate data source. (3) Another reason is that the authoritative source of data is not clear enough. The Education Informatization 2.0 Action Plan launched by the Ministry of Information Technology and Education requires that a data should correspond to a source, which means that the authoritative source of each data item should be clear. However, in the actual situation of universities, for example, the data for the management of contract staff and the management of external teachers are scattered in various departments and colleges, so it is difficult to find an accurate summary table. In such cases, if something goes wrong with the data, it is often unclear which department should be held responsible. Many other data are in a similar situation.

All of the above involves issues at the level of the overall data architecture that need to be thought about and solved primarily at the information center, which is the main driver for many schools to initiate data governance at the moment.

(II) In the beginning, various data quality problems were caused by the inadequate design or low quality of the functional modules of various business system software.

The software quality problem itself has many manifestations, and this paper only discusses the manifestations related to data quality, which are mainly manifested in the data entry process of the software, i.e., the lack of necessary constraints and checks.

(1) For example, when students and teachers fill in an information form, they will be asked to fill in their unit, political affiliation and major name, but in fact, these contents correspond to a code list or a very definite set, so a reasonable way to fill in the form should be to provide a drop-down box, so that users can selectively input the information. However, due to the insufficient design of the software, such an interface in many cases does not allow the user to enter data by drop-down boxes, but

[a]68207930@qq.com
[b]68207930@qq.com
[c]87274044@qq.com
[d]44342716@qq.com

allows the user to fill in the form freely, resulting in a variety of different ways to fill in the form.

Also, for example, information like home address should have been filled in at the national, provincial, city and county level with a drop-down box to select it, and then the detailed address behind it should be filled in by hand by the user, but many software designers left a long text box for convenience and let the user fill in a long string of text at will. But the address information thus filled out lacks a structured definition that only a human can understand, and a program would have difficulty understanding it. It is for this reason that in the subsequent use and analysis of the data, the program has little good way of structurally identifying this geographic information and can only display it as a text string. If the university wanted to conduct, for example, an analysis of address data such as birthplace, this information would have little way to be used directly.

(2) In addition, many information systems lack post-entry validation of data. For example, these systems do not adequately test the required or the fill in the data is not filled in, or fill in the data is correct, whether it contains illegal characters, whether beyond the reasonable value of the boundaries and range. This results in improper Spaces, commas, semicolons, and even line breaks after data entry. Also, sometimes when users enter their mobile phone number, they may accidentally fill in one digit less, or fill in one digit more, or the number of digits in the ID number is incorrect, such factors will lead to a large number of unreasonable data.

(III) Insufficient information literacy of the operators. Many staffs of business departments are lack of information literacy, and the information center has spent a lot of energy and funds to help these departments to build the information system, but the result is still delayed, or after the built system is handed over to the department, the department complains by various phone calls every day that the system is not easy to use or can't be used and other problems, and finally they are not willing to adopt it. Another situation is that the department will often use the system for a period of time on the side, then they ignore it, and return to the original off-line manual work state. In this case, it will obviously lead to a lot of data not being recorded correctly, which is also a very important aspect of the data quality problem.

(IV) The adaptability of software and business. The business processing system used by many departments has been built and used for many years, and the version is very old. However, at this time the vendor neither upgraded the system nor performed maintenance. Since it is actually very difficult to approve funding for school maintenance, and the school's business is always changing, it leads to the software not being able to keep up with the needs of the business, so they have to stop using the business system. There are also some non-core business departments that do not use information technology to build the system at all, but use very primitive electronic forms, and even use paper documents for business management. Such factors will lead to the situation that the data that should be recorded is not recorded, and the data that should be put online is not online. Departments often use offline data recording, which is difficult to share and use, and paper-based data is even more difficult to process informally.

In this paper, the four major reasons for the data quality of colleges and universities are listed one by one. In fact, the problem of data quality in universities is far more complicated than this, and the various situations are strange and bizarre. The above list is just some typical situations, although not necessarily comprehensive, but they are more representative and common problems, which the author believes that many teachers responsible for information technology in colleges and universities should be empathetic to this.

If the quality of data is to be addressed, the cause of the quality problem must first be traced back to the source, just as with a doctor's visit, the prescription can only be effective if the diagnosis is correct.

2 IDEAS FOR IMPROVING DATA QUALITY

At the level of how to improve data quality, the process of solving data quality problems can be divided into 4 stages, which are identifying the problem, analyzing the problem, giving feedback and solving the problem.

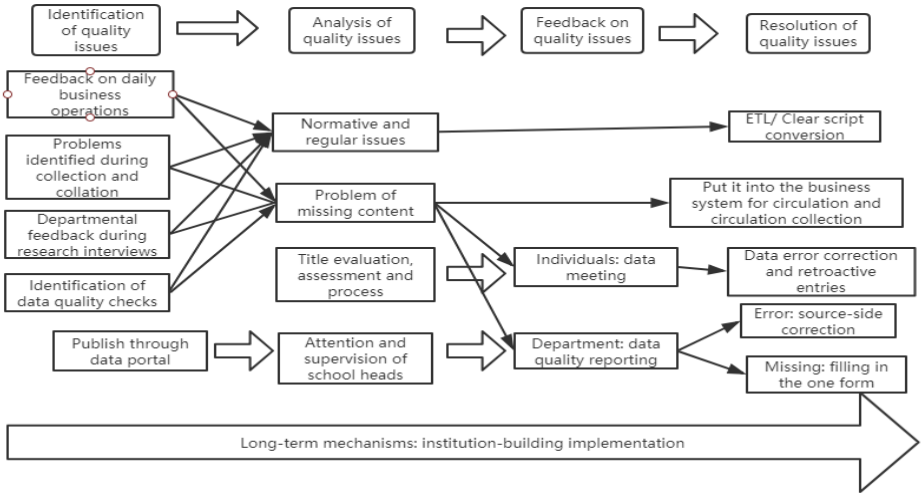

Figure 1 General approach to improving data quality

(I) Identify data quality issues

Let's first look at how to spot and identify quality issues. In fact, the most straightforward scenario is that when various business departments and teachers and students are running various operations, they encounter data problems that often lead to process errors, so they make a phone call to the information center to ask what's going on. I believe many teachers of the information center of the university have encountered this situation before, of course, this is a very effective way to find out the quality problem, but it is too passive.

In this situation, the information center is under all kinds of pressure, and they all want to solve the problem as quickly as possible, so the process of solving the problem becomes a kind of coping state, so, basically, they don't collect and organize the problems systematically.

There are three main ways to identify data quality issues more systematically and effectively, which are:

(1) Find problems with the data when cleaning and integrating the data after it was collected.

(2) Find out what problems exist with the data and the business through departmental feedback during departmental interview research.

(3) Use a special data quality check tool to perform quality checks and output detailed quality check reports.

(II) Analysis of data quality issues

After understanding the manifestation and location of the data quality problem, the next step is to analyze the identified data quality problem. First of all, we need to classify the problem, and the classification is based on the characteristic that whether the problem can be dealt with by technical means. If it can be dealt with by technical means, then we call it a normative and regular problem, while problems that cannot be dealt with by technical means, such as missing data content and semantic errors, are called irregular problems.

(III) Data quality feedback and resolution

For normative and regular data quality problems, we can take ETL class software or the use of database scripts and other methods to clean the conversion and processing; for irregular data quality problems, such as some content errors, which basically cannot be solved by technical means, we need to feedback these problems to the data responsibility unit or the relevant subjects.

For some data attributed to individuals, which are generally only better known to the person to whom the data is linked, we can correct and re-mediate missing and erroneous data by allowing individual students and teachers to see what data exists for them at the school, and then target the missing and erroneous data. Accordingly, schools should provide a page that displays individual student and teacher data and supports students and teachers in initiating the process of correcting errors.

For data belonging to the functional department, this data can be used to provide feedback on issues by way of data quality reports. Correspondingly, the school should accurately analyze the quality of the school's data through data quality analysis related tools or platforms and output data quality reports that form the basis for business departments to revise errors and missing data. The revision referred to here is the revision of errors and missing data by each department by logging into their respective business management systems, such as the original business systems of Academic Affairs, Personnel and OA.

At this point, some people will surely raise questions: (1) with so many teachers and students across the school, who would normally take the time to look at that data, and would they follow it when you ask them to correct errors; (2) also, even though those departments receive data quality reports, it's actually very difficult to get those departments to really take the quality reports seriously and consciously correct erroneous data. Indeed, this is the most difficult part of the whole data quality improvement process.

It seems to us that some external push is needed to make these tools work. For example, there is a good time to carry out error correction and filling in of personal data, i.e., at some critical points in time such as title reviews and annual appraisals, and if the school is going to be on a business system such as a comprehensive business system such as a one-stop shop, or a micro-service platform, the school often needs to verify personal data in bulk, and therefore The information center can seize this opportunity to promote the verification and reporting of all personal data in the university. The processing of departmental data requires us to adopt another mechanism, for example, we can make public the data quality issues and information management level of each department. In this way, we can rely on the supervision of the school leadership or the information office or through information technology assessment and other such means to push the pace of each department to address their respective data quality issues.

In fact, in addition to the data feedback to the various relevant subjects to solve the problem, there is a particularly important way, that is, to put the data into the business system to make it work and flow, and among these systems, the typical system is the comprehensive business platform just mentioned, such as one-stop service platform and net-com.

These previously proposed measures to solve the quality problem are actually temporary and technical in nature, and can even be said to be treating the symptoms rather than the root cause. In order to achieve long-term stability, we also need to carry out the supporting construction of information management system specifications and effective implementation. The figure below shows an overall idea about improving the quality of university data, which strives to achieve effective data management in a comprehensive manner.

3 DATA GOVERNANCE IN PRACTICE

Guided by the idea of improving data quality, Jinan University has carried out data governance practice. Jinan University organizes the information technology construction department to conduct information technology data and business research for all units on campus, and then collects source business data through ETL workers for aggregate integration to form a data warehouse. The data tables and table fields in the data warehouse are identified and annotated by combining the

unit survey results and routine business analysis. Through the data warehouse management system, the metadata, master data and other objects of the data warehouse are under standardized management and data quality detection, and the data quality report is generated. The unified open data platform encapsulates the data in the data warehouse for business application invocation, so that the data can be tested in real application scenarios and the closed source business pays attention to data quality in the production data link, which ultimately promotes the gradual improvement of data quality. The data correction and re-recording system allows teachers and students to see the data relevant to them, and can initiate data correction and re-recording to continuously improve data quality from the source data production side. At the same time, the data portal presents data quality reports to the university leaders and departments, so that the leaders can grasp the university's data quality situation intuitively, thus achieving the purpose of urging all units to improve their awareness of data quality work, and all units can also clearly grasp the data quality situation of their own units. The above are the practical steps of data governance in Jinan University, as shown in the figure 2 below:

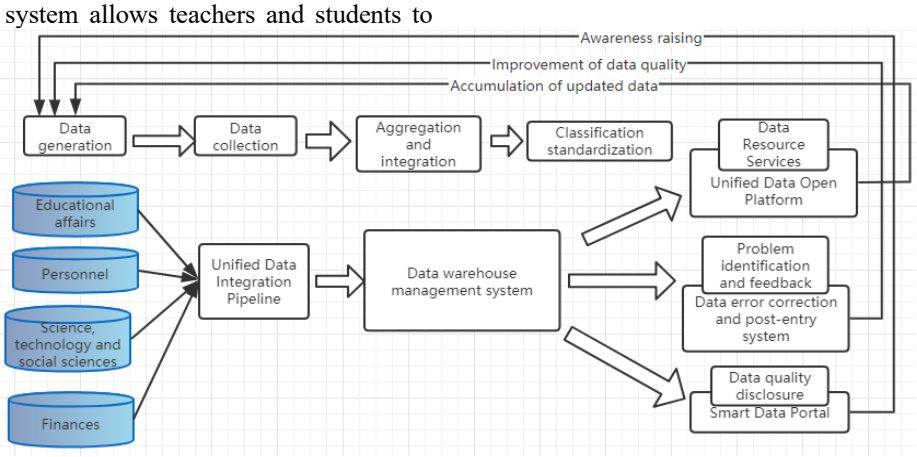

Figure 2 Diagram of Data Governance Practice Steps at Jinan University

4 Conclusion

After From the above discussion, we can find that although it is very difficult to solve the problem of data quality in schools, it is not impossible. The difficulty lies in the fact that there are many factors involved in it, which involve a lot of work in the information center, the attention of school leaders and relevant management measures, as well as various departments, manufacturers and various business systems and software. Therefore, a lack of work in any aspect will exert a great impact on the final result. However, it doesn't mean that this matter can't be handled well, because after all, information technology is a major trend in China and has received more and more attention from the country. At the same time, the information literacy of teachers and students is getting increasingly higher, and the work assessment of each department is getting stricter and stricter. And we also cannot deny the convenience and help that informatization brings to teachers and students as well as to all departments and leaders at all levels. As long as such an atmosphere is formed and guided by a scientific method, we can manage it through a sound system, so that all faculty and all departments of the university can be held accountable for their respective responsibilities, and then we take a step-by-step approach to improve such an idea as we move forward, believing that it is only a matter of time before we finally solve the data quality problem of the university from the ground up.

Acknowledgement

This paper is one of the achievements of Guangzhou Science and Technology Innovation and Development Special Fund Project (2019) "artificial intelligence-based human-computer interaction intelligent learning platform development and application" (project number: 201902010041)

References

1. Zhao Hongwei, Feng Tao and Yu Haitao, Application of comprehensive data quality management framework in power grid industry, Information Technology&Standardization, Tianjin, no.7, pp. 62-65, 2018. *(references)*
2. Lu Letian, Yang Menghua and Deng Yingwen, Research on government data governance system, Telecom Engineering Technics and Standardization, Hu'nan, vol.32, no.1, pp.29-33, 2019. *(references)*
3. Li Qing and Han Junhong, Data governance: means and methods to improve the quality of educational data, Distance Education in China, Beijing, no.8, pp.45-53, 2018. *(references)*
4. Wu Lili and Zhang Bo, Research on the method of improving data quality in university data governance, Journal of Chongqing University of Technology: Natural Science, Chongqing, vol.33, no.8, pp. 149-156, 2019. *(references)*
5. Dang Fangfang, Mei Lin, Gao Feng, Wang Ning and Jiang Wei, Research on data governance technology of electric power enterprise based on life cycle

management, Power Systems and Big Data, Henan, vol.22, no.3, pp. 66-70, 2019