The Application of Data Governance in Universities

2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 978-1-6654-9125-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIBD57115.2023.10206371

Haitao Jiang
Network Information Center
Qufu Normal University
Qufu, China
qfjht@126.com

Wei Yan
Network Information Center
Qufu Normal University
Qufu, China
yanwei@qfnu.edu.cn

***Abstract*—With the wide application of information systems, the universities have accumulated a large number of data resources. Due to the lack of effective data governance, such problems as the difficulty in data sharing, low data quality, lack of security, poor utilization ability, etc. are more and more prominent. In order to solve these problems, this paper designs a university data governance platform, which realizes the whole life cycle management of data, and thus can improve the data management ability, eliminate data islands, and fully exploit the potential value of data. This paper expounds the architecture, the resource development process, the main functions and the typical applications of the data governance platform.**

Keywords—data governance, data quality, data security, data sharing

I. INTRODUCTION

With the development of smart campus, the vast majority of the universities have built perfect campus networks, and begin to apply various business systems to improve their work efficiency, simplify the work procedures, and reduce the operating costs. With the continuous increase of the number of business systems, the scale of data is gradually expanded, which caused such problems as the increasing number of low-quality data, the existence of redundant data, the low accuracy of data, the difficulty of data sharing.

The only way to solve the above problems is to use big data technology to carry out data governance. Through data governance, the universities can improve their data management ability, enhance the data quality, eliminate data islands, increase the value of data and minimize data-related cost and risk [1].

In recent years, many scholars and engineers have carried out research and practice on data governance in many fields. Milne and Brayne discussed data governance in large-scale data sharing of dementia [2]. Mao et al. constructed a new framework for government data governance [3]. Su et al. put forward a proposal to apply block chain technology to the data governance system of the oil and gas industry [4]. Zhang et al. took a mining company as a research case and developed a framework to guide enterprises to configure data governance activities [5]. However, there are few discussions on how to carry out data governance in universities.

In this article, we firstly discussed the problems in university data governance. Then we proposed a university data governance platform. Next, we elaborated on the platform in three aspects: architecture, data resource development process, and data resource management functions. Finally, we introduced two typical applications of the platform.

II. CURRENT STATUS QUO OF DATA GOVERNANCE IN UNIVERSITIES

Through the comprehensive analysis of the current situation of university informatization, it is found that the lack of reasonable planning and cooperation in the early stage of informatization construction has brought about the following problems in the management and use of data [6].

- The phenomenon of data island is widespread. Data cannot be conveniently exchanged and shared between business systems.
- Low-quality data: Due to the lack of a unified data quality management system, there are a large number of wrong data, invalid data, inconsistent data and duplicate data in the business system, and the data quality is not guaranteed.
- The rights and responsibilities of data are vague. The authoritative source of many data is indeterminate, and the process of data circulation is not standardized, leading to the phenomenon of inconsistent data, and the decline of the credibility of the data.
- Lack of complete data life cycle management: The current data life cycle management process and specifications are not perfect, and there are no automated tools to support the management of the data life cycle, leading to potential security risks in data use.
- Lack of big data analysis ability: The purpose of big data analysis is to extract valuable content from large and fragmented data. At present, the schools have fully realized the importance of structured data, and have strengthened the integration and accumulation of structured data. However, the attention paid to unstructured data, such as learning behavior data, Internet behavior data, and activity track, is obviously insufficient, which cannot fully exploit the value of big data.

The current data governance capabilities in schools can be evaluated by using the data governance maturity model. In a data governance maturity model proposed by DQM Group [7], the capabilities are divided into five stages: Aware Stage, Reactive Stage, Proactive Stage, Managed Stage, and Optimal Stage. At present, the data governance capacity of many

universities is still in the second stage. Through the implementation of comprehensive data governance, the governance capacity can be promoted to the fourth stage, and the continuous and effective data governance can make the governance capacity reach the fifth stage.

III. Construction of a Data Governance Platform

With the help of the data governance platform, the schools' data governance can be comprehensively promoted. The platform can provide a standardized data resource development process, a unified data resource management center, an efficient data asset operation center and various data analysis tools.

A. Platform Architecture

Referring to the data governance framework proposed by IBM's Data Governance Committee and the big data governance framework proposed by Zhang et al. [8, 9], we designed the architecture of the school data governance platform, as shown in Fig. 1.

The school data governance platform is deployed on the Hyper-Converged Infrastructure (HCI), adopts distributed storage and parallel computing, makes the best of HCI's flexible resource allocation mechanism, and provides the computing resources, storage resources and network resources required by the platform as needed.

The platform can be divided into four layers from bottom to top: data source layer, data storage layer, data management layer and data application layer.

The data sources cover all kinds of original data of the school, such as the basic information of teachers and students, students' grades, scientific research achievements, asset list, financial data, etc. According to the different structures and types of data, the data can be divided into structured data, semi-structured data, log data, document data, IoT (Internet of Things) data, etc. Some of these data are distributed in the application systems, and some are stored in the form of electronic documents in the computer.

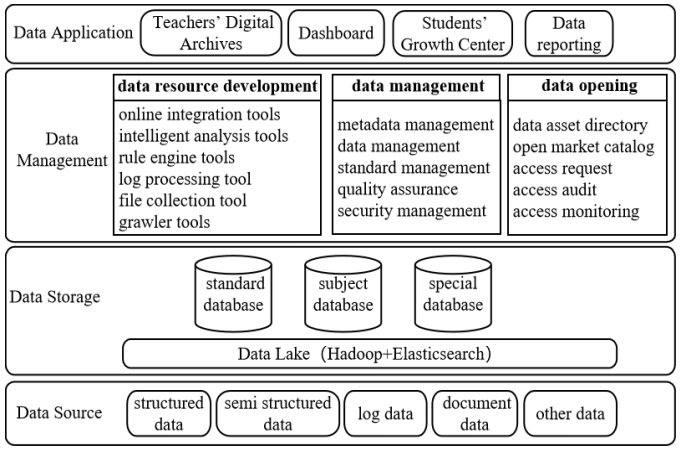

Fig. 1. Architecture of the school data governance platform.

The data storage layer is mainly used for the data integration and storage. This layer contains data lake, standard database, subject database, special database and other types of storage structure.

The platform adopts the combination of batch and stream processing to solve the data collection problem, using data integration, log integration, document integration and other tools to collect data, and then stores a "backup data" consistent with the data source structure and type in the data lake.

Data from the data lake is cleaned and converted and stored in the standard database. The construction of the standard database refers to the national standards and industry standards, and is combined with the individualized needs of the school. In general, the standard database can be built around eight theme domains: teacher domain, student domain, finance domain, asset domain, teaching domain, scientific research domain, management domain and public service domain.

Subject database and special database can provide support for the data analysis of the application layer. Their data models are built according to the application requirements, and thus characterized by being flexible, scalable, and wide table, with the fact table and dimension table as the main construction mode. The subject database contains data analysis topics such as teaching, research, students, consumption, assets and books. The special database stores statistics about schools, departments, individuals and assets. These data can support multidimensional data analysis, decision-making assistance, data reporting and other applications.

Data management layer provides unified data management capabilities, data opening capabilities, and resource development capabilities, and realizes data full life cycle management.

Data application layer is to reflect the value of data governance. Relying on the whole-domain data of the school, the value of the data can be fully mined to provide services for the teachers and students. Typical applications include dashboard, teachers' digital archives, students' growth center, intelligent report, etc.

B. Data Resource Development Process

The platform provides a set of visual standard processes for the development of data resources, which is shown in Fig. 2.

1) Data source registration

The first step of the data governance is to survey the source data in various departments and application systems, confirm the form of data storage, the data update mechanism, and the requirement of data sharing. Next, the data source information is registered on the data governance platform, so the platform can automatically connect the data source and complete the data source registration.

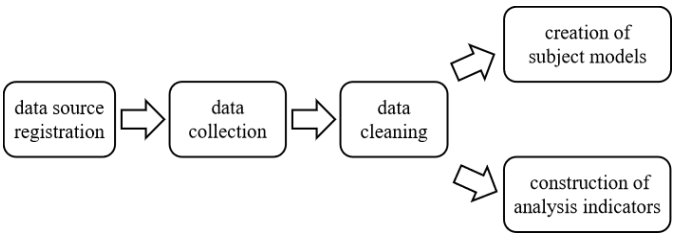

Fig. 2. Data resource development process.

2) Data collection

According to different data source access methods, the platform provides corresponding tools to complete the data extraction, mainly supporting three methods: full extraction, incremental extraction and real-time extraction. The platform can automatically identify the metadata in the data source through such tools as "automatic back identification", "dictionary document identification", and "knowledge base identification", etc. so as to realize the automatic collection of data resources. The platform can monitor the progress of data collection, and generate quantitative statistical indicators accordingly.

3) Data cleaning

Data cleaning is the process of identifying and trying to fix "dirty data" [10]. Data Cleaning can solve the problem of attribute errors, missing data, and approximately duplicate records in the source data. Normalization of data is to solve the problem of data irregularities such as inconsistent coding and inconsistent format, through "rule processing engine", "standard code mapping" and other methods. These processes will gradually improve the versatility and portability of data, and provide a reliable guarantee for data analysis.

4) Creation of subject models and construction of analysis indicators

After cleaning, the standardized data can be reorganized according to different subject, and the relevant data can be integrated to form a new data model through attribute fusion and relationship fusion. For example, when taking teachers as the subject, the data resources described by teachers as the main body can be extracted, including basic information, scientific research achievements, and salaries, etc. Meanwhile, corresponding indicators can be calculated for different data analysis tasks.

C. Data Resource Management

The data governance platform supports the full life cycle management of data, whose main functions include metadata management, standard management, master data management, quality management, security management, etc.

1) metadata management

The metadata is the data describing the data (data about data), is a structured description of the information resources. Metadata quality is the key to the efficient data utilization and to providing high-quality information services [11]. The platform can automatically extract the metadata in the data source and establish a correspondence with the metadata in the standard database. The platform supports administrators to set different labels on metadata according to the data source, type and content, so as to better understand the data. The platform can automatically form a metadata map according to the data source and destination to support data lineage analysis and impact analysis. Fig. 3 shows an example of lineage analysis of teacher number field. The ZGH field of table T_KHXX in the teachers' digital archives is derived from the ZGH field of table T_JZGKH in the standard database, and the ZGH field of table T_JZGKH is derived from the EmployeeNo field of view V_assessment in the HR management system.

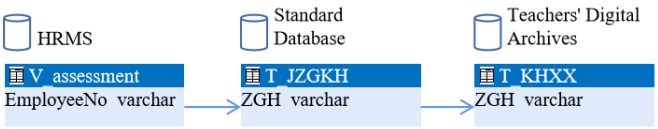

Fig. 3. Lineage analysis of teacher number field.

2) Standards management

Data standardization is an important guarantee to eliminate data islands and realize data exchange and sharing. The information standards should be formulated in combination with the actual situation of the school, the national standards and industry norms. All the data entering the standard database should be cleaned and standardized according to the information standards. Data standards management covers five stages: carding of data standards, draft of data standards, review of data standards, release of data standards and implementation of data standards. Standards management can unify the code standards, data interface standards, data integration standards and data model standards, and lay a good foundation for school informatization. The platform provides standards maintenance, reference analysis, code correction, management standards release, standards statistics and other functions.

3) Master data management

The master data management module provides functions of such as global data query, advanced data query, offline data import, data statistics and so on. The global query supports the online query of the master data in the data lake and the standard database. Advanced query provides the function of creating subject views and full-text retrieval. A subject view is a view that allows administrators to access the data more flexibly through the "UNION ALL" operations of multiple associated data tables. Full-text retrieval can be conducted according to the keyword to search in all data. With the help of full-text retrieval service provided by Elasticsearch, the platform supports the rapid search in a large amount of data.

Offline data import can import the local data (such as EXCEL files, etc.) into data lake, through the data template, and then realize the data collection and sharing.

The statistics function can comprehensively count data assets and describe the current data governance situation through indicators of data asset and distribution of asset.

4) Data quality management

Using wrong data for analysis and calculation usually leads to wrong conclusions, and high quality data is the key to realizing the value of data. The platform provides a data quality management system including detection rule defining, problematic data discovering and quality monitoring. The basic process of quality management is as follows. First, the quality rule configuration function is used to define various types of data detection rules. Then, using the quality detection configuration function, you can define the detection rules for specific data items. Next, a periodic detection task is defined and initiated, and the platform automatically completes the data quality detection. Finally, the detection results will be pushed to the relevant personnel according to the data rights and responsibilities. Relevant personnel can query the details of the

problem data and locate the problematic data to correct them. The platform can also analyze the data quality and establish data quality indicators, so as to display the data quality in real time and intuitively.

5) Data security management

In the era of data sharing, big data not only brings convenience, but also causes the problem of personal privacy leakage [12]. While meeting the needs of the data sharing, the data governance platform should also fully consider the security and protection of the data, and effectively avoid the privacy leakage caused by the data mismanagement. In order to ensure the security of data in the collection, storage, transmission, sharing and use, the data governance platform has established a complete set of security system. The main measures are as follows.

- Identify the authoritative source of each data according to the principle of "one data, one source", the departments of data source (the producers of the data) are responsible for determining the scope and mode of data sharing. When other departments (the data users) need to use the shared data, they can submit an application, which will be approved by the departments of data source. For the sharing of sensitive data, both parties can also agree on the responsibility of data security protection.
- Confirm the data's level based on its importance and the impact caused by its leakage, and select the security protection strategy accordingly. The platform can automatically identify sensitive data such as ID number, mobile phone number and salary and mark the data level.
- In order to ensure the security of data during storage and transmission, the platform supports a variety of encryption algorithms and data masking algorithms, which can automatically identify important data and encrypt or mask them.
- The platform provides the data watermark processing function, which can effectively prevent data from tampering and facilitate the tracking of reasons after data leakage.
- The platform provides a perfect authentication, access control and authorization mechanism, which can authorize users based on their roles, departments and positions, and can grant different operation privilege for data with different levels.
- The platform can monitor key data operations, warn or prevent noncompliant operations, so as to prevent data leakage and data damage caused by malicious operations.
- The platform is able to audit the operational behaviors and record all data access behaviors so as to detect potential risks in time.

D. Construction of Data Opening Capability

Data opening capability is an important guarantee for data circulation and extensive sharing. Through standardized data exchange, data sharing can be realized between cross-platform application systems, and more application systems can be connected to the data governance platform in the future. The platform publishes the school data asset directory and data open directory, and provides five types of data accessing: API interface, online query, file download, open view, and data subscription. According to different data security requirements, the platform can provide both "real data" and "encrypted data". The application system can apply for data access according to the needs, and the platform can approve it based on different approval processes. In addition, the platform can count the indicators related to data sharing, so as to comprehensively reflect the status of data sharing.

IV. Typical Applications

As a result of data governance, the school will have rich data assets, and based on which, many practical applications can be derived. Here are two typical applications.

A. Dashboards

Through the dashboards, the school's data assets can be visually analyzed according to different dimensions such as human resources, financial and material resources. The platform can provide dashboards with various independent subjects for different needs of school leaders and department heads, such as basic statistics on school, faculty structure, information of students, financial situation, scientific research, teaching situation, discipline construction, etc.. It can also combine dashboards of different subjects into a set of IOC(intelligence operation center), which can present the school's current work progress and operation status in a complete and real-time manner. Fig. 4 shows the partial content of the dashboard for scientific research.

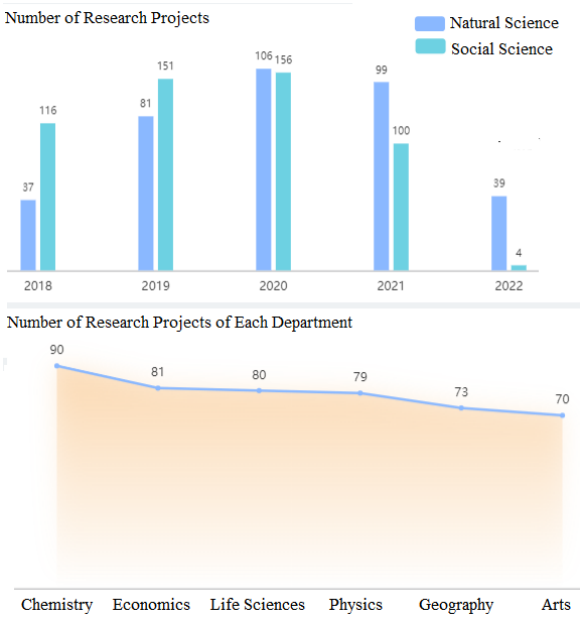

Fig. 4. Partial content of the dashboard for scientific research.

B. Teachers' Digital Archives

Because of the school data governance, we can collect, store, process and calculate the teachers' data, which scattered

in various application systems, so as to establish digital archives with the personal data as the center. Eventually each teacher's teaching, scientific research, projects, Internet records, books borrowing information can be visually display, so that all teachers can better understand themselves, and thus summarize and improve their teaching, scientific research and learning ability. In addition, the big data analysis technology and intelligent algorithm can provide teachers with more valuable data analysis services. For example, the growth and change process of a teacher since his/her induction into the profession can be traced. Fig. 5 shows partial screenshot of a teacher in the teachers' digital archives.

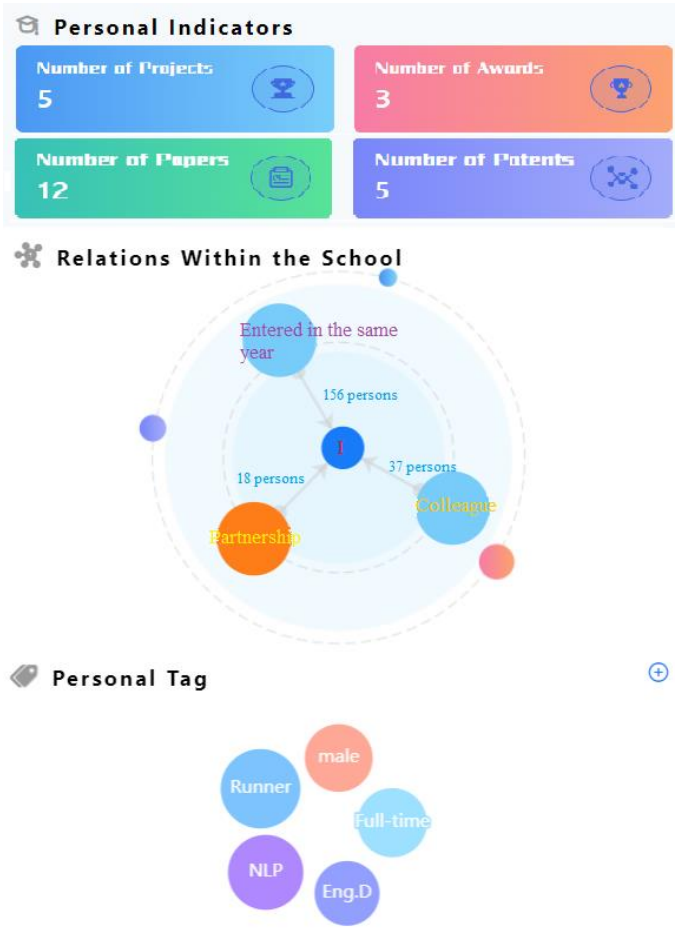

Fig. 5. Partial screenshot of a teacher in the teachers' digital archives.

V. The Value of Data Governance

Data governance can improve data quality, enhance data security and improve data availability. Besides, it can also promote the healthy development of schools through reducing operating costs, improving processing efficiency and supporting decision-making. For example:

- In the scenario of information filling such as professional title declaration and employment assessment, a large amount of authoritative data can be directly obtained from the teachers' digital archives. This can avoid the repeated filling of data, and simplify the review process, so as to improve the work efficiency and cut the cost.
- In the students' growth center module, by analyzing students' elective courses, grades, credits, GPA (Grade Point Average), rewards and punishments information, the platform can evaluate academic standing and give corresponding suggestions in combination with the training programs.
- Through visual analysis of scientific research data, the academic committee can intuitively understand various statistical indicators, find out the scientific research weaknesses of the school, and formulate corresponding measures to improve the academic level.

VI. Summary

In view of the difficulties and problems of university data governance, this paper designs a data governance platform, and after its implementation we have achieved the expected goal. This platform has also been applied in several other universities and has got consistent praise, which fully shows that the data governance architecture designed in this paper meets the actual needs of universities.

In the process of school informatization, data governance is the only way through; at the same time, it is also a long-term, complex, systematic task. The construction of data governance platform not only provides the standard process and technical method of data governance, but also includes the definition of various data specifications, the formulation of various data management regulations and the establishment of data governance organizations at all levels.

There are also some problems and challenges in the process of data governance.

- With the wide application of the Internet of Things, 5G communication and other technologies, the amount of data has increased rapidly, and the data types are more diversified, which brings more challenges to data integration.
- The contradiction between data utilization and privacy protection is more prominent. How to give full play to the value of data while protecting personal privacy needs better solutions.
- How to effectively transform the results of data governance into the benefits to schools?

Data governance is a task that needs all departments of the school, all teachers and students to participate together. We are the promoters, participants and beneficiaries of the school's data governance task.

Acknowledgment

This project is funded by the Smart Campus Project of Qufu Normal University. We thank our colleagues from all departments for their pleasant cooperation in the project.

References

[1] R. Abraham, J. Schneider, and J. Brocke, "Data governance: a conceptual framework, structured review, and research agenda," International Journal of Information Management, vol. 49, pp. 424-438, December 2019.

[2] R. Milne, and C. Brayne, “We need to think about data governance for dementia research in a digital era,” Alzheimer's Research & Therapy, vol. 12, January 2020.

[3] Z. Mao, J. Wu, Y. Qiao, and H. Yao, “Government data governance framework based on a data middle platform,” Aslib Journal of Information Management, vol. 74, pp. 289-310, February 2022.

[4] J. Su, S. Yao, and H. Liu, “Data governance facilitate digital transformation of oil and gas industry,” Frontiers in Earth Science, vol. 10, May 2022.

[5] Q. Zhang, X. Sun, and M. Zhang, “Data matters: a strategic action framework for data governance,” Information & Management, vol. 59, March 2022.

[6] X. Wu, B. Dong, X. Du, and W. Yang, “Data governance technology,” Journal of Software, vol. 30, pp. 2830-2856, September 2019.

[7] A. Gregory, “Data governance--protecting and unleashing the value of your customer data assets,” Journal of Direct, Data and Digital Marketing Practice, vol. 12, pp. 230-248, February 2011.

[8] A. Wróbel, K. Komnata, and K. Rudek, “IBM data governance solutions,” 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), 2017, pp. 1-3.

[9] S. Zhang, R. Pan, and Y. Zong, Big data governance and services, Shanghai: Shanghai Scientific & Technical Publishers, 2016, pp. 1-224.

[10] N. Tang, “Big data cleaning,” in Web Technologies and Applications, L. Chen, Y. Jia, T. Sellis, and G. Liu, Eds. Cham: Springer, 2014, pp. 13-24.

[11] Z. Liu, J. Wang, and Q. Li, “Review of metadata quality evaluation studies,” Information Theory and Practice, vol. 45, pp. 42-48, July 2022.

[12] L. Huang, M. Tian, and H. Huang, “Preserving privacy in big data: a survey from the cryptographic perspective,” Journal of Software, vol. 26, pp. 945-959, April 2015.