Research on Identification, Archiving, and Preservation of Enterprise Structured Data Based on Master Data

Authors: Li Zefeng, Ma Wen, Wang Qian, Wei Nan, Wei Nan

Date: 2023-08-14T00:00:00+00:00

Abstract

Enterprise digital transformation necessitates full lifecycle management of data, and data appraisal and archiving constitutes a crucial approach to overcoming the difficulty that big data technologies face in directly processing traditional unstructured documents. On the foundation of enterprise data governance, master data management is introduced into data archiving, dividing enterprise data into three categories: master data, transactional data, and analytical data. The macro appraisal method is applied to appraise these three types of data and determine the scope of data archiving, with ER diagrams, data dictionaries, data lineage graphs, and similar metadata incorporated into the metadata archiving scope. The optimal pathway for data archiving and preservation lies in integrating archived data sub-lakes into the construction of enterprise data lakes. Archives departments can accelerate their integration into the national big data strategy by implementing a "dual-system" approach for electronic records and data archiving, initiating pilot programs for data archiving in large state-owned enterprises, and enhancing the data literacy of archival work teams to enable active participation in data governance.

Full Text

Research on Identification, Archiving and Preservation of Enterprise Structured Data Based on Master Data Management

Li Zefeng[1],[2],[3], **Ma Wen**[1], **Wang Qian**[1], **Wei Nan**[1] [1]Zhengzhou University of Aeronautics, Zhengzhou 450046 [2]Henan Collaborative Innovation Center for Aviation Economy Development, Zhengzhou 450046 [3]Electronic Records Management Research Center, Renmin University of China, Beijing 100872

Abstract

Enterprise digital transformation requires whole-lifecycle data management, and data identification and archiving represent a crucial means to address the challenge that traditional unstructured documents cannot be directly processed by big data technologies. Building upon enterprise data governance, this study introduces master data management into data archiving, categorizing enterprise data into three types: master data, transaction data, and analytical data. Using the macro-appraisal method to evaluate these three data categories, we determine the scope of data archiving. ER diagrams, data dictionaries, and data lineage diagrams are incorporated as metadata into the archiving scope, with the optimal preservation path being the integration of archived data sub-lakes into enterprise data lake construction. Archives departments can accelerate integration into the national big data strategy by implementing a "dual system" for electronic records and data archiving, piloting data archiving in large state-owned enterprises, and enhancing the data literacy of archival work teams to actively participate in data governance.

Keywords: Master data; Data archiving; Data appraisal; Data lake

---

Introduction

With the rapid development of information technology, various computer systems have been widely applied in office, production, and management work, generating massive amounts of data. Initially, this data accumulated within respective generation systems, isolated from one another. To address these so-called information silos, practices such as common databases and data warehouses emerged to establish connections between isolated data. In recent years, data lakes and data lakehouses have emerged as centralized storage methods that accommodate all forms and massive volumes of data within organizations.

These massive datasets include both structured and unstructured data. Initially, archives organized data according to their relationships to form visualized, easily understandable unstructured data such as various forms, which were then appraised and preserved in paper or electronic formats. In the big data era, to fully utilize big data technologies for mining archival information content, it is necessary to extract key business data from previously unstructured archival information—what we call "archival datafication." This process involves data extraction, syntactic and semantic analysis, and is by no means trivial. Scholars have further proposed the concept of "data archivalization," which involves directly managing and preserving business data 沉淀 in operational systems as archives in structured data form. Both concepts essentially transform archival content into computer-processable data to solve the problem that current big data technologies cannot directly process unstructured data.

The academic community has made valuable theoretical explorations in this

area. Liu Yuenan et al. [**?**] proposed relevant control activities to ensure data resources meet archival management requirements (authenticity, integrity, availability, and security). Zhao Shenghui et al. [**?**] argued that archival management institutions should establish a dual-wheel drive mechanism for data and archives. Zhao Yue [**?**] contended that traditional archiving models have limitations in meeting data era resource preservation needs and proposed a data resource archivalization model that shifts from "control" to "intervention." These studies demonstrate academic consensus that archival work in the big data era should incorporate data into its management scope. However, practical issues such as data appraisal, archiving scope, and archiving methods have received limited attention. Therefore, this paper introduces the concept of master data and applies fundamental archival theories to construct a path and operational methods for data archival appraisal, archiving, and preservation based on master data. For convenience of discussion, this paper defines data archives as archives organized and preserved in structured form, while traditionally understood archives are referred to as unstructured archives.

1. Enterprise Data Analysis Based on Master Data

1.1 Enterprise Data Relationship Analysis

Enterprise data refers to all data information related to enterprise production, management, and operations. Examining modern enterprise operations, enterprise data can be categorized by source into internal data and external data. Internal data refers to historical data directly generated during business activities, such as employee data, production data, financial data, customer data, and product data. These data are primarily generated and stored in structured form within various enterprise systems (such as HR, SCM, ERP, CRM, etc.) and can be transformed into unstructured data when utilized. External data is data obtained from outside the enterprise to ensure normal operations, such as market data, competitor data, relevant national regulations and standards, and industry data. External data is collected and preserved in both structured and unstructured forms.

As the national big data strategy advances, big data technologies are increasingly applied in enterprises, and data has become an important component of enterprise assets. Scientific management of data assets based on master data management is currently a relatively mature and widely applied method. Master data management categorizes data from the perspective of enterprise data assets and governance into master data, transaction data, and analytical data.

1.2 Master Data

The national standard "Data Management Capability Maturity Assessment Model" (GB/T 36073-2018) [**?**] defines master data as "core business entity data within an organization that needs to be shared across systems and departments." The Data Management Association International (DAMA) defines

master data as "authoritative, most accurate data about key business entities that provide context for business transactions" [**?**]. Although definitions vary, there is academic consensus that master data reflects the basic information of an organization's core business entity status attributes, such as enterprise personnel, products, customers, materials, and project data.

Master data possesses characteristics of globality and shareability. Globality means that master data is required by all functional departments and business processes within an organization, while shareability means master data is commonly used and shared by multiple business systems. Precisely because of these characteristics, master data items cannot be modified once defined—only expanded after a stable period. Simultaneously, master data must be compatible with various heterogeneous business systems. This is crucial for the archivalization of heterogeneous, multi-source, and multi-modal enterprise data, establishing a foundation for the master data path in enterprise data archivalization.

1.3 Transaction Data

Transaction data is data generated when various departments within an enterprise use business applications to process departmental business transactions according to business processes. It represents data produced during the fulfillment of departmental responsibilities, with complete business data thoroughly documenting all business transactions during enterprise operations.

Transaction data is primarily generated and stored in structured form within various business systems, such as personnel, sales, procurement, and financial data. Through clustering and correlation relationships between transaction data, various documents, forms, vouchers, and other unstructured data are formed. In traditional enterprise archival work, these unstructured documents, forms, and vouchers are appraised, organized, and become integral components of enterprise archives.

It is important to note that transaction data includes conditional data—transaction data generated under special circumstances or scenarios, such as pricing policies, environmental requirements, and credit ratings. These conditions originate from external data; under external data constraints, enterprises generate transaction data, which changes accordingly when external data changes.

Analytical data refers to structured or unstructured data formed through in-depth analysis of the previous two data types using data analysis techniques as needed, such as BI analysis, various reports, and audit data. This type of data frequently appears in current enterprise archiving scopes.

Considering current enterprise archival management objects, Figure 1 [Figure 1: see original paper] illustrates the relationships among the three data types and their relationships with enterprise archives. The arrow direction indicates the process of enterprise data supporting archive generation, the opposite direction

represents archival datafication, and the dashed line direction represents data archivalization.

2. Data Appraisal Methods

2.1 Applicability of Macro-Appraisal Method

The academic community has reached consensus on incorporating data into archival management as an object of archival work. Enterprise data is dispersed across various business systems, massive in volume, and diverse in type. The discrete characteristics of enterprise data make it difficult for archivists to appraise each data item individually. How to apply fundamental archival value appraisal theories to improve data appraisal and archiving is an urgent problem requiring resolution.

Various archival value appraisal methods exist. This paper selects the macro-appraisal method as the fundamental appraisal theory. The macro-appraisal method refers to archivists no longer appraising individual documents but instead appraising the various creation contexts of these documents and their current utilization situations—that is, appraising the functional origins of documents. The classic textbook *Archival Management (5th Edition)* considers the macro-appraisal method an inevitable choice for archival appraisal work in the information age and an important theoretical method and practical tool for batch processing and identification of records generated, stored, and archived in various business information systems [**?**].

The macro-appraisal method provides an excellent approach for data appraisal. First, enterprise data classification is precisely based on data creation context and utilization situations, demonstrating good compatibility with the macro-appraisal method. Second, master data governance requires consideration of organizational background, culture, internal functions, policies, and standards. In particular, master data management demands the division of functional domains based on business work content (not institutional departments) and business analysis of each functional domain. This aligns with the macro-appraisal method's conceptual foundation of constructing based on functional origins. Third, examining the sources of transaction data and analytical data makes the application of macro-appraisal method appropriate. Various departments within an institution formulate and implement continuous or one-time work projects and activities to fulfill their responsibilities, which trigger specific actions and transaction processing; to effectively handle these transactions, information systems are established [**?**], thereby generating corresponding transaction data and analytical data.

2.2 Data Appraisal Method Based on Macro-Appraisal

The macro-appraisal method posits that the focus of appraisal work lies in analyzing and identifying the importance of the limited number of functions,

projects, activities, and interactions, rather than on the massive volume of documents. From a data appraisal perspective, we can combine functions, activities, and interactions with the corresponding information systems that generate data for analysis and identification.

(1) Confirming the Functional Source of Data

This involves sorting out and confirming the functions of various departments within the enterprise and the various information systems established by departments to effectively handle departmental affairs. By establishing relationships between enterprise departmental functions and business systems, we can further confirm the functional sources of data. Together with business personnel and system developers, we analyze software business processes, data flows, and data processing to confirm core business and core basic information. This process should include both vertical and horizontal dimensions: vertical includes group and subsidiary unit systems, while horizontal includes different business systems within the enterprise.

Analyzing the functional source of data for a single business application system is relatively straightforward. Analysis across the entire enterprise or even the entire group company is much more difficult. In the analysis, three issues should be addressed: First, which enterprise and departmental functions and activities should be documented, reflected in which business information system and which functional module of the system; second, which departments and positions formulate and execute critical functions, and what data and documents are formed during departmental business processing; third, which functions are most important. Since software development follows software engineering methods, business system development documents such as business flowcharts and data flow diagrams will be important analytical tools and bases.

Unlike the singularity of document generation, data generation may have multiple sources. For example, student data exists in admission systems, academic management systems, and student work systems, with likely inconsistencies across the three systems. Analysis according to the three questions above becomes particularly important. If we understand that the academic management system manages student status and reports data to education administrative departments, it can be regarded as formulating and executing critical functions.

(2) Confirming Data Lineage

Data lineage is a common concept in data governance, referring to the various relationships formed between data and data throughout the entire data lifecycle—from generation, processing, and storage to utilization and disposal—similar to human blood relationships [**?**]. It records the link relationships of data origins and paths. Through lineage relationships, we can relatively easily determine data origins, intermediate source databases, files, applications, and the departments and positions that create and maintain this data, thereby establishing deeper relationships with the functional sources of data.

Figure 2 [Figure 2: see original paper] is a schematic diagram of data lineage relationships (not a standard data lineage visualization diagram). Table X data is the final business data, Tables A, B, C, and D are original data, Tables E, F, and G are intermediate table data calculated during processing, and Table H is data possibly used from other business processes. The diagram clearly shows the data relationship links. Sankey diagrams are commonly used in big data management to visualize data relationship links.

(3) Confirming Data Standardization

Confirming whether data is standardized and whether there is a large amount of standardized, interconnected business data is a prerequisite for data archiving. Data that is not standardized and remains in information silos is difficult to appraise. Saving all data without appraisal neither conforms to archival management principles nor causes management costs to rise sharply. More importantly, as time extends and data volume increases, storage of non-standardized data will inevitably become a data swamp, making retrieval, querying, and utilization difficult. The master data introduced earlier represents one of the achievements of data governance and standardization.

(4) Applying Macro-Appraisal Based on Functional Source Concepts

Based on the above analysis, archivists can clearly understand the importance of various enterprise departmental functions and business activities, identify the simplest, most accurate, and most important data formed by critical functional departments within the enterprise, and preserve them as archives.

The multi-source, heterogeneous, and discrete characteristics of data create certain difficulties in practical application of the above appraisal methods. Even with the help of business personnel, data management personnel, and IT personnel, tracking data lineage and confirming the importance of data-generating functions is not easy. Fortunately, with the rapid development of data management concepts and technologies, the emergence and application of data governance, master data management, data lineage diagrams, data lakes, and data lakehouses have brought new opportunities to data appraisal work. Combining the above appraisal methods with enterprise data governance processes or utilizing the results of data governance will greatly reduce the workload of archivists and particularly alleviate their technical apprehension about data appraisal and archiving.

3. Enterprise Data Archiving

Enterprise data governance addresses issues of data quality, availability, and security throughout the entire lifecycle. Following the concepts and practices of master data management, we can discuss archiving scope around the three types of enterprise data.

3.1 Data Archiving Scope

(1) Complete Archiving of Master Data

As basic information reflecting the status attributes of an enterprise's core business entities, master data is consistently shared across departments and business systems, meeting the needs of both individual departments and inter-departmental business collaboration. Therefore, master data should be completely archived. For example, enterprise-wide business standard type master data (organizational structures, customers, suppliers, etc.) must be fully archived. If an enterprise has built project master data, product master data, material master data, or equipment master data according to its needs, these should also be archived.

(2) Selective Archiving of Transaction Data

According to data lineage relationships, transaction data can be further divided into original real-time data, result data, intermediate data, and conditional data. Real-time data records enterprise real-time business, describing business behaviors occurring at specific time points; result data represents transaction data generated after departments complete their responsibilities; intermediate data is data generated during statistical, correlation, and other calculations on real-time data to produce result data. A simple example is shown in Figure 2.

There are two logical approaches to transaction data archiving:

The first approach archives original real-time data, while other data may not be archived, such as data in Tables A, B, C, and D in Figure 2. From a data lineage perspective, all other data originates from original data. This archiving scope is simple to operate—that is, according to the macro-appraisal method, determine the responsibilities and importance of various departments, then archive the original data they generate. The advantage of this approach is comprehensive data, large volume, and fine granularity, facilitating big data technology analysis. Its disadvantages are also obvious: the archived data volume is enormous, information granularity is excessively fine, and if not properly managed after archiving, it may create a data swamp in the archives department.

The second approach archives result data, while original data and intermediate data are preserved in the forming departments but not archived. In Figure 2, Table X data would be archived. Conditional data mostly belongs to result data in most cases; if result data is archived, conditions should be archived together with the data to ensure completeness of data archiving. This method can relatively easily determine core critical data as archiving content based on the most important functions of departments. The sum of key data from all departments still constitutes the enterprise's full data volume, which is relatively large but easier to manage after archiving.

(3) Analytical Data May Not Be Archived

For current enterprise archiving scopes, unstructured analytical data is often

the archiving object. However, from a data management perspective, unstructured analytical data is generally derived from result data through induction, statistics, correlation, clustering, etc. When enterprise full-volume result data is preserved, analytical data may not need to be archived.

For example, an enterprise treats personnel (employees, customers, partners, etc.), finance, and product-related data as master data. Taking the human resources department as an example, one of its core businesses is personnel assessment. The department may have established a complex assessment system and indicators, requiring collection of large amounts of basic data from other departments, then conducting weight analysis and other calculations according to the indicator system, finally producing quantitative evaluation results. Following the second archiving approach, master data needs complete archiving, so the personnel assessment business archives the final quantitative evaluation data, while other large amounts of basic data and intermediate calculation data do not need to be archived.

(4) Utilizing Data Lineage Analysis Tools to Assist in Appraising Transaction Data Value

Since relevant laws and standards for data archiving have not yet been established, the above methods depend on archivists' familiarity with business and their experience in original archival work, placing high demands on archival staff. Data lineage analysis can be used as an auxiliary tool, together with business personnel and data management personnel, to appraise transaction data value based on actual data application value. This appraisal is based on practical data application value and can serve as a good auxiliary tool for data appraisal. Data lineage diagrams visually display data inflow nodes, outflow nodes, usage volume, etc. This paper does not elaborate on data lineage in detail but briefly explains how to appraise data value.

Data lineage theory holds that the more data users, the larger the usage volume, and the more frequent the updates, the more valuable the data is. Examining data lineage diagrams, outflow nodes on the right represent data users; more data users indicate greater data value. Thicker data flow lines indicate larger data update volumes, reflecting data value to some extent. Shorter data flow line segments indicate more frequent updates, fresher data, and higher value.

This is a value assessment method based on actual data application, appraising current data value, and should be combined with comprehensive analysis of the importance of functions involved in the data chain.

3.2 Metadata Archiving Scope

Due to the discrete characteristics of structured data, in addition to archiving the data itself, the most important aspect is how to maintain inter-data relationships and preserve these relationships as metadata for data archiving. Limited by space, this paper focuses on discussing metadata that characterizes data rela-

tionships, without delving into metadata that may be consistent with electronic records archiving.

China's National Standardization Administration released a series of metadata registration system standards in 2009, such as the GB/T 18391 series, GB/T 23824 series, and GB/T 30881, which provide important guidance for enterprises to understand data metadata and define metadata in a standardized and unified manner during data governance. For archives departments, this means incorporating the results of enterprise metadata management according to standards into the data archiving scope.

During the archival governance process, a series of relevant data and documents are generated, including data governance standards, master data management, data quality management, and data security management. These documents should be metadata elements of the data. As shown in Figure 3 [Figure 3: see original paper].

(1) ER Diagrams

ER diagrams, or Entity-Relationship diagrams, describe real-world entities, attributes, and relationships [**?**]. Figure 4 [Figure 4: see original paper] shows a partial entity-relationship diagram for an enterprise. In practice, ER diagrams are structural diagrams used for database design. In relational database design, one relationship typically corresponds to one database table, with attribute values being data in the table. Through ER diagrams, primary and foreign key constraints between data tables (such as employee ID attributes in employee and purchase list entities) establish logical connections between data. The diagram clearly shows the logical relationships between employees, purchasers, and purchased goods.

ER diagrams can be collected during the development of various business systems. Best practice is to collect them during data governance, as ER diagrams established during data governance identify all enterprise entities and their relationships, while ER diagrams from various business systems are only partial entity-relationship diagrams.

(2) Data Dictionaries

Data dictionaries are collections of information describing data, defining and describing all data elements used in enterprise business systems through definitions of data items, data structures, data flows, data storage, and processing logic [**?**].

Data dictionaries typically include five aspects: data items, data structures, data flows, data storage, and data processing procedures. The following examples of data items, data storage, and processing procedures illustrate the necessity of data dictionaries as data metadata.

Data items include data item names, meaning descriptions, aliases, data types, lengths, value ranges, value meanings, and logical relationships with other data

items, providing accurate descriptions of entity attributes in ER diagrams. Data processing describes the process by which original data generates result data through operations and other processing. Data storage describes and explains the storage of transaction data such as original data, intermediate data, and result data.

Data dictionaries are established after detailed analysis of data used by various departmental functions and businesses, reflecting data itself and inter-data relationships. Using Figure 3 as an example to establish a data dictionary, we can know which employee purchased what quantity of which product from which supplier at what time.

Like ER diagrams, data dictionaries are best collected when data governance is completed.

(3) Data Lineage Diagrams

Data lineage relationships are important methods for metadata analysis. Data lineage diagrams use visualization methods to show how data comes into being, what processes, stages, and computational logic it goes through, essentially visualizing data flows and processing procedures in data dictionaries. Data lineage diagram applications allow data dictionaries to focus on describing and explaining data items. If data lineage relationship diagrams are archived as metadata, data dictionaries can archive only the data items section.

Data lineage diagrams should be generated using big data governance tools after data governance is completed and then collected.

(4) Pointer Links

Analyzing up to this point reveals an interesting phenomenon: electronic records (unstructured data) have structured data as metadata, while structured data may have unstructured metadata. Therefore, pointer link methods can be applied to establish relationships between structured data and unstructured metadata, with these pointer links also preserved as part of metadata.

3.3 Determining Retention Periods

Enterprise data can be assigned different retention periods according to the importance of functions and business chains. A currently workable method is mapping enterprise data to the retention schedule of the National Archives Administration's Order No. 10, using existing enterprise archival retention schedules to guide retention periods for archived data. For example, Order No. 10's enterprise management retention schedule 13 .10 divides customer information into important customer information (permanent preservation) and general customer information (30-year preservation). Mapping customer information to data, customer information belongs to the master data category, and corresponding master data can be divided into permanent and 30-year retention periods.

Enterprise production and operations generate numerous reports, most of which are formed by extracting data from business systems/cross-system databases to establish inter-data relationships (such as various operations). If these are within the enterprise archiving scope and retention periods have been divided, mapping relationships with corresponding data can also be established to determine data involved in reports. Following the data lineage relationship diagram to trace its data chain and applying the transaction data archiving approach in Section 4.1, corresponding data retention periods can be determined.

4. Data Archiving Preservation Methods Based on Data Lakes

Current archival management systems are primarily designed and developed for unstructured archives and basically lack the capability to manage data archives. Structured data archives require reconstructing the preservation environment. Data lakes represent current best practices for data storage platforms, capable of centrally storing massive, multi-source, heterogeneous, and multi-type data within enterprises while supporting rapid data retrieval, processing, and analysis.

4.1 Data Preservation Based on Data Lakes

Data lakes lack a unified concept, but examining various definitions reveals several common characteristics: a data lake is a centralized data repository where enterprises can store their full-volume data [**?**]. Therefore, data lakes have sufficient data storage capacity and complete data management capabilities to preserve an enterprise's full-volume data, including structured, semi-structured, and unstructured data, while managing key information elements such as data sources, formats, and permissions.

Generally, data in data lakes is a complete original copy of enterprise business data, maintaining consistency with data in business systems. Therefore, data lakes also possess data lifecycle management capabilities, meaning they not only store original data but also preserve various intermediate and result data from analysis processing, while completely recording data analysis and processing procedures to enable users to trace in detail the generation process of any data.

As a centralized distribution center for enterprise full-volume data, data lakes basically meet the conditions for archival data sources, thereby avoiding the trouble and difficulty of archives departments collecting data from individual business systems. Simultaneously, data lakes have complete data acquisition and publishing capabilities, enabling regular and irregular acquisition of full-volume or incremental data from relevant business systems for standardized storage. This also provides conditions for incremental data archiving.

Existing archival management systems cannot effectively collect, store, and manage data. The powerful data collection, storage, and management capabilities

of data lakes compensate for archival management system deficiencies. Consideration can be given to reserving specific storage areas in data lakes to establish data archive sub-lakes for preserving archived data, as shown in Figure 5 [Figure 5: see original paper].

In Figure 5, the data archive sub-lake can be part of the enterprise data lake or a separate centralized storage platform for data archives. The enterprise data lake is primarily used for fine-grained data management and business, while the archive lake is a storage platform with larger granularity that supports both current enterprise utilization and preservation of enterprise historical data.

4.2 Data Lake Construction and Data Archiving in a Large State-Owned Enterprise

Currently, many large state-owned enterprises such as Sinopec, CNPC, CNOOC, Southern Power Grid, and China Mobile have all undertaken data lake construction in their industrial digital transformation. Among them, CNOOC is conducting phase two of its group-wide data lake project. The author participated as a consulting expert in the data sub-lake construction of a Tianjin subsidiary of CNOOC.

(1) Data Lake Construction Planning of a State-Owned Enterprise

The group headquarters proposed that the data lake is the support for group digital transformation, data governance, and core business. The planned unified data lake is designed to achieve efficient data collection, comprehensive aggregation, high-quality data asset control, unified sharing, and multi-dimensional services, as shown in Figure 6 [Figure 6: see original paper].

In the diagram, each professional technical service company builds its own data lake platform according to unified technical architecture to support internal data applications within its unit. Data standards are consistent with headquarters, and data is aggregated to the group unified data lake. Data management functions achieve centralized management of upstream production data catalogs, architecture, standards, security, quality, and governance achievements such as metadata and master data, while supporting upstream production data operations. The group headquarters data lake aggregates and integrates standard data from various subsidiary and professional companies, uniformly constructing an upstream data lake. Based on the data lake, data analysis is conducted to provide unified data services for various applications.

(2) Data Archiving Management

Both the group headquarters data lake and various company data sub-lakes, as intensive infrastructure, explicitly propose data archiving to achieve full-lifecycle data management, as shown in Figure 6. Archiving here includes two meanings: first, from an IT perspective, storage strategies are formulated based on data hot/cold analysis to achieve three-tier storage of online, nearline, and offline;

second, from an archival perspective, data is appraised and preserved in specific archival storage areas.

During the data governance process, the archives department proposed the internal and external dual-cycle concept: an internal cycle where archival management rules are embedded in the data governance process to ensure compliant control and efficient operation of data archives, and an external cycle where archival value is embedded in core business scenario data (such as equipment operation and maintenance). Based on the dual-cycle concept, the archives department actively participates in the data governance process. Grounded in the core values of archival business—authentic data preservation, authoritative data distribution, business evidence locking, and data asset control—the department reasonably embeds archival management rules and values into the main lines of industrial and data chains. It actively participates in data governance to complete key business system data archiving and intelligent data supervision services, with practical operations following the appraisal and archiving storage approaches proposed in this paper.

5. Recommendations for Data Archiving

(1) Implementing a "Dual System" for Electronic Records and Data Archiving

Data archivalization and data archiving currently lack mature concepts, models, and best practices. To meet the challenges of the big data era, accelerate integration into the national big data strategy, and stimulate data element vitality, we can implement a "dual system" for electronic records and data archiving, similar to the dual-system archiving of paper records and electronic files.

(2) Pilot Programs First

Data archiving can be piloted in large state-owned enterprises, especially those that have implemented data governance, achieved master data management, and initially established data lakes. These enterprises basically have the infrastructure conditions for data archiving, have established data standardization systems and complete data full-lifecycle management systems, and possess relatively powerful data management tools. More importantly, large state-owned enterprises, in accordance with national digital transformation and data element vitality enhancement policies, have urgent needs for data archiving.

(3) Enhancing Archivists' Data Literacy and Capabilities, Actively Participating in Data Governance

The big data era places increasingly high demands on archival work. Archives are not only the information destination after business completion but also important tools for data quality supervision and management optimization throughout the business data lifecycle. Archival work teams should recruit data management talents. Archivists should possess not only knowledge of archival informatization and electronic records management but also gradually develop

data literacy and certain data management capabilities, mastering the ability to explore, understand, and communicate using data. They should actively participate in data governance as archival personnel, proposing archival management concepts and requirements at key nodes of the enterprise data full-lifecycle and business data chain, to achieve effective management of enterprise data identification, archiving, and preservation.

References

[1] Liu Yuenan, He Siyuan, Wang Qiang, et al. Collaborative Management of Enterprise Archives and Data Assets: Problems and Countermeasures[J]. *Archival Science Research*, 2022(6):94-102.

[2] Zhao Shenghui, Hu Ying, Huang Yihan. Microscopic Mechanism Analysis of Data, Archives and Their Symbiotic Evolution[J]. *Archival Science Communication*, 2022(2):4-12.

[3] Zhao Yue, Sun Jingqiong, Duan Xiane. Archivalization: Rational Thinking on the Intervention of Archival Science in Data Resource Management[J]. *Archival Science Research*, 2020(5):83-91.

[4] National Standardization Administration. *Data Management Capability Maturity Assessment Model* GB/T 36073-2018[EB/OL]. Beijing: China Standards Press, 2018.3.

[5] DAMA International. *The DAMA Guide to the Data Management Body of Knowledge*[M]. New York: Technics Publications, 2009.

[6] Wang Yingwei, Chen Zhiwei, Liu Yuenan. *Archival Management*[M]. Beijing: China Renmin University Press, 2021:107.

[7] Wang Yingwei, Chen Zhiwei, Liu Yuenan. *Archival Management*[M]. Beijing: China Renmin University Press, 2021:108.

[8] Li Chunmei, Zhang Xing, Geng Huizheng, et al. Construction of Data Analysis Methods Based on Data Lineage[J]. *China New Telecommunications*, 2020,22(20):50-51.

[9] Wang Shan, Sa Shixuan. *Introduction to Database Systems*[M]. Beijing: Higher Education Press, 2014:215.

[10] Hu Ming. Digitalization of Grassroots Social Grid Governance and Its Regulation[J/OL]. *Social Science Journal*:1-13[2023-07-19].

[11] Chen Qing, Zhang Zhi. Research on Data Lake Architecture Integrating Multi-Source Heterogeneous Data Governance[J]. *Journal of Intelligence*, 2022,41(5):139-145.

Note: Figure translations are in progress. See original paper for figures.