建构与解构：生命周期视角下的档案数据质量控制探析

林　凯[1]　周林兴[2]
（1　苏州城市学院城市治理与公共事务学院　苏州　215104；
2　上海大学文化遗产与信息管理学院　上海　200444）

摘　要　大数据时代，档案数据迅速增长、难以管控，数据质量问题严重限制了档案数据价值的发挥，开展档案数据质量控制研究具有重要意义。通过系统梳理有关文献，阐述档案数据质量控制的出场逻辑，借鉴生命周期理论，将档案数据生命周期过程划分为数据生成、数据收集、数据组织与分析、数据归档与保存、数据发布与利用五个环节，并立足于档案数据生命周期，将档案数据质量控制过程建构为前端、中端、后端三个阶段，再对每一阶段的质量控制过程进行解构，以期为提升档案数据质量提供有价值的研究参考。

关键词　生命周期；档案数据；数据质量；质量控制；质量优化
DOI：　10.16065/j.cnki.issn1002-1620.2025.02.014

Construction and Deconstruction: Analysis of Archival Data Quality Control from the Perspective of Life Cycle

LIN Kai[1], ZHOU Linxing[2]
(1　School of Urban Governance and Public Affairs of Suzhou City University, Suzhou　215104;
2　School of Cultural Heritage and Information Management, Shanghai University, Shanghai　200444)

Abstract: In the era of big data, archival data has grown rapidly and is difficult to control. Data quality problems have seriously limited the value of archival data, so it is of great significance to carry out research on archival data quality control. Systematically reviewing relevant literature, this paper elaborates on the logic of quality control for archival data and, drawing on the life cycle theory, divides the process of archival data life cycle into five stages: data generation, data collection, data organization and analysis, data archiving and preservation, and data release and utilization. Based on the lifecycle process of archival data, the quality control of archival data is constructed into three stages: front-end, mid-end, and back-end. The quality control process of each stage is deconstructed to provide valuable research references for improving the quality of archival data.

Key words: life cycle; archival data; data quality; quality control; quality optimization

1　研究问题的提出

大数据时代，档案数据大量生成、海量汇聚，逐渐成为新的信息资源形态，传统的档案管理工作逐渐向档案数据管理工作转型、迈进。作为新型生产要素，档案数据价值红利日益凸显，相较于其他数据而言更加具备真实、可靠与权威价值，能够为资政服务、科技发展、民生建设、城市规划等经济社会发展

的各个方面提供支撑数据。与此同时，数据重复冗余、数据运行异常、数据缺失、数据泄露等质量问题频发，常造成档案数据无法使用、内容失密等损失，严重限制了档案数据价值的发挥，难以有效满足社会发展需要。《“十四五”全国档案事业发展规划》强调“加强档案资源质量管控”[1]。国家档案局局长陆国强在2022年全国档案局长馆长会议上提出“针对影响制约档案事业转型发展高质量发展的瓶颈问题……扎实有效解决”[2]。可见，如何处理档案数据质量问题，推动档案事业高质量发展，是目前学界和业界关注的焦点。

“大数据环境下，数据、信息、文件、档案等概念之间没有共识的边界。”[3]以数据形式存在的档案，其内涵和外延逐渐延展，可以概括为：档案数据是具备档案属性的数据，既包括档案部门已经掌握的各类数字化档案资源、电子档案等，还包括具有长久保存价值但还没有纳入档案部门保管范围的数据，以及档案管理业务过程中产生的各类数据。[4]生命周期理论是事物发展的基本理论，其含义可以通俗理解为从摇篮到坟墓（cradle to grave）的整个过程。世间万物，任何现象或问题研究的演化和发展都要经历一个生命周期过程。[5]档案数据从生成开始就处在系统和网络中，经常会跨库、跨系统运行，导致其在生命周期过程的众多环节中，所遇到的管理人员、方式和标准并非一致，容易产生种类各异的质量问题。因此，有必要从生命周期视角出发，详细解构档案数据的各阶段特征，并有针对性地分析其质量问题，提出科学的应对措施，全方位推动档案数据顺利满足社会各项事业的需求。

研究者从不同视角对档案数据质量问题进行了探讨。一是对档案数据质量的认识。档案数据质量是对其进行管理和开发的基础[6]，质量内容包括数据内容完整规范、数据数量齐全、数据分类科学等方面[7]。二是对档案数据质量问题的分析。目前，档案数据化实践尚处于探索阶段[8]，较难解决数据错误、格式不兼容等问题[9]，提升数据质量刻不容缓。三是档案数据质量提升对策。根据档案数据质量问题的复杂性，可开展档案数据监管[10]，研究档案数据质量管理框架[11]。也要改善档案数据质量的管理理念，将档案数据质量要求融入整体数据质量要求中[12]，并完善相关政策法规[13]，使档案数据质量得到政策上的保障。同时，加强技术的运用，可融入区块链[14]、数据融合[15]、副本冗余[16]等技术，提升档案数据质量检测的准确性。可见，学界对档案数据质量研究已有一定成果，对其的认识和提出的问题解决对策对于后续研究的开展具有重要参考价值。但是，现有研究缺乏对档案数据生命周期的分析，导致提出的策略针对性不足。本文基于生命周期视角，建构生命周期视角下的档案数据质量控制过程，力图提出更具针对性的质量控制举措，使档案数据质量在生命周期的每一阶段均能得到有力维护。

2　生命周期视角下档案数据质量控制的出场逻辑

充分利用档案数据，最大限度地释放其价值，前提是要使之保持高质量状态。因此，档案数据质量控制应运而生，在实践需求变化和跨学科理论的倒逼和指导下，不断充实其研究内容。

2.1　生成逻辑：档案数据的大量形成

大数据时代，信息技术的飞速发展推动了人类信息存储能力和计算机数据处理技术的进步，大量数据不断涌现。作为社会信息资源的重要组成部分，档案不可避免地加入了这场数据化浪潮。在技术驱动和数据驱动下，信息的存在方式越来越数据化，档案数据逐渐成为大数据时代档案信息资源的重要形态[17]，其呈现出从“模拟态”“数字态”到“数据态”的形态跨越，并在数量上以较快的趋势逐年增加。如2023年底，全国各级综合档案馆馆藏电子档案2289.6TB，其中，数码照片211.4TB，数字录音、数字录像1207.6TB，馆藏档案数字化成果28849.2TB。[18]同时，档案数据的生成范围和空间逐渐延展，其整个生命周期过程的生存环境呈现出网络化和数据化特征，任何能接入互联网的单位和个人都可以生成并运行档案数据，这也导致档案数据在来源上变得广泛，数据内容纷繁复杂，利益相关者众多，管控难度增大。若不能采取合理措施有效管控档案数据，将直接影响其质量，难以满足各方需要，开展档案数据质量控制已成为档案工作的迫切需求。

2.2　实践逻辑：业务工作需求的推动

一方面，档案存在形态的改变给实践部门带来了巨大挑战。大数据环境下，档案数据在其生命周期的全过程中，普遍存在于以数据为尺度的业务系统空间中，业务的原始记录脱离了传统介质的载体，可直接以离散的数据形式存在。[19]而以离散的数据化形式存在的档案信息结构复杂，包括原生数据、衍生数据，以及半结构化、结构化数据。需要将这类

档案信息收集齐全，提取出其外在特征、来源信息、背景、人物、时间、地点等，让离散的数据保持逻辑上的关联，使反映客观事实的档案内容记录完整。这将增加数据挖掘与分析，多元异构数据融合的困难，易造成档案数据流失、格式冲突、数据异构等，最终导致质量下降。如浙江省绍兴市档案馆在建设档案数据中心过程中，经常出现数据异构问题[20]，制约了该城市跨地域的档案资源数据化开发质量。

另一方面，档案数据管理主体之间的业务协同不足。档案数据在生命周期过程中生存环境网络化和数据化的特征，使其分散保存在不同单位、不同系统中，其所有者、处理者、控制者不一定是数据形成单位。[21]这也使得档案数据的管理权限分散到了多个不同主体中，这与传统环境下档案最终移交到档案馆，并由其集中统一保管的方式存在差异性。如近些年来，各省市大数据局、大数据管理中心等数据管理机构纷纷成立，与档案部门在数据管理标准、数据质量目标需求上未能做到统一。其在收集整理社会数据资源时，与档案部门缺乏有效协同，未能严格按照档案管理规定采集、接收、整合数据记录，使得归档数据难以符合来源可靠、程序规范、要素合规等相关要求[22]，较大限度地影响了数据质量。档案业务工作面临的质量问题对档案管理实践提出了新诉求，迫使档案部门思考如何开展档案数据质量控制，探索采取科学合理的控制方法。

2.3 理论逻辑：数据科学理论的助力

一方面，数据思维模式赋能质量控制理念的产生。数据科学的概念由图灵奖获得者彼得·诺尔（Peter Naur）于1974年提出，其指出数据科学是一门基于数据处理的科学，侧重于解决数据管理中的问题。数据科学理论将“数据现象”和“数据问题”从信息科学中独立出来，强调从数据出发，让数据说话，用数据导控。[23]其使得数据和信息、知识的边界日渐模糊，能在尚未从数据中提炼出知识的情况下，直接用数据化解面临的难题。[24]即强调利用数据满足需求，使数据可以不进行知识转化而直接使用的范式。这对档案数据管理思维模式产生了显著影响，档案管理思维开始考虑从内容驱动向数据驱动，直接利用开放数据解决困难的方式转变。如《“十四五”全国档案事业发展规划》提出“推动档案馆定期通过网站或其他方式公布开放档案目录，稳步推进开放档案全文在线查阅”，“鼓励有条件的综合档案馆全年向社会公众开放或延长开放时间”[25]。但是，大数据环境下，数据在其生命周期的任一环节都可能产生质量问题，让整个数据资源池中充斥着重复、冗余、篡改、错误的数据。“数据行业依托自身的技术优势，过度追求数据的抓取与拥有。”[26]只注重数量的积累，忽视质量的提升，最终只能使数据无法满足需求，不能达到数据驱动的目的。如在“长三角一体化建设”背景下，上海市电子健康档案数据质量堪忧，主要存在关键信息缺失，重要健康信息逻辑错误的问题[27]，对该项目的推进产生了不利影响。因此，掌握高质量的数据才是掌握了“数据原油”。档案数据化实践尚未成熟，如果任由轻视数据质量的思想蔓延，档案数据管理必将陷入泥沼。开展档案数据质量控制，以高质量档案数据驱动社会发展才是档案数据管理的根本遵循。

另一方面，数据科学指导下的档案数据研究纵深发展需要高质量数据支持。数据管理是数据科学理论中的重要内容，其将数据视作重要的信息资源，运用云计算、物联网、大数据、智慧工程等现代技术对数据资源进行有效收集、处理、存储、挖掘、利用，保障数据长期可用的效果，实现数据价值，提高组织运行效率和核心竞争力。[28]在该理论指导下，数据研究将朝着纵深方向发展，对数据的加工和处理不再仅限于数据的简单过滤，而是更加注重数据价值的开发和创造。其倡导通过算法和工具迅速获取数据，揭示数据之间的关联特征，洞见数据隐藏价值，并在此基础上开展数据深加工，开发数据产品，最大限度地释放数据价值。档案数据作为重要的社会信息资源，是数据管理的对象和场域，在其指导下，档案数据管理理论、方式将会产生重要变革。数据管理将会深化档案数据的研究范畴，真正盘活档案数据资源，通过指导档案数据的分析和加工，探讨构建数据之间人的关联、物的关联、人与人的关联、人与物的关联、时空的关联，建设互相联系、纵横交错、网络状的数据连接格局[29]，从而开发档案数据，使其价值得以增值。然而，就目前档案数据管理实践而言，管理理念滞后，且档案工作人员的技术水平未能和数据管理发展水平相适应，出现“技术隔阂”[30]，导致档案数据管理停留在简单的数据捕获、数量积累和价值探索阶段，不能使档案数据在其全生命周期过程中维持高质量，使得更深层次的数据价值开发和增值研究难以进行。如囿于“数字增量”意识，“目前不少档案机构虽已掌握着大量PDF、图片格式的数字档案资源，但这些资源目

前仅停留在简单的信息检索、组织、利用上”[31]。因此，档案数据的深入研究需要数据质量控制作为辅助支撑，提供高质量数据源以满足需求，才能保证研究的顺利开展。

3　生命周期视角下档案数据质量控制的阶段性过程建构

立足于生命周期理论指导，探寻档案数据生命周期过程，能为档案数据质量控制在方法选取、理念重构、长期保存等方面提供全面、立体的分析视角。数据研究中，生命周期理论得到灵活运用，数据生命周期理论被提出。数据生命周期被认为是支持数据保存和管理实践的重要因素[32]，其是指“数据产生，经数据加工和发布，最终实现数据再利用的一个循环过程”[33]。档案研究中，文件生命周期理论是生命周期理论在文件档案管理领域的具体应用，该理论揭示了文件从最初生成到完成使命后被最终销毁或是具有保存价值而被保存的运动过程。按照文件生命周期的划分阶段，应从文件档案生成的前端出发，实行前端控制和全程管理。[34]档案数据虽已脱离了文件格式的“封装”[35]，但依然具备档案的原始记录性等固有性质，亦具有数据的流动性、结构复杂性、可复制性等特征。根据档案数据所具有的档案和数据的双重属性，笔者认为应综合借鉴数据生命周期和文件生命周期理论，划分档案数据生命周期，并根据其生命周期阶段性特征开展质量控制。基于此，本文将档案数据生命周期过程划分为数据生成、数据收集、数据组织与分析、数据归档与保存、数据发布与利用五个环节。其中，数据生成属于档案数据生命周期的前端；数据收集以及数据组织与分析属于其生命周期的中端；数据归档与保存以及数据发布与利用属于其生命周期的后端。档案数据质量控制应从档案数据产生的前端出发，立足于其整个生命周期过程进行全程管理，并依据其生命周期每一阶段的特征和所处环境的变化，采取针对性的策略管控（图1）。

3.1　档案数据生命周期的前端

此时的档案数据正在日常事务工作中大量生成，数据量虽不断积累，但尚未开始发挥作用。同时，该阶段较多数据分布零散，并不容易被发现和重视，这也为档案数据质量控制增添了工作难度。档案部门对此应该具备敏锐的眼光，认识到档案数据的重要性及其广泛的社会效益，在档案数据生成的前端明确如何管控档案数据，即应当确定档案数据质量需要达到何种程度，怎样使其质量达到目标需求才能满足后续的工作需要；并对如何提供数据获取、利用等有系统、合理的规划，从而使档案数据质量控制工作在档案数据生成的前端准备就绪，以便随时开展工作。

3.2　档案数据生命周期的中端

该阶段的档案数据已经开始显现出数据价值，并在业务系统中快速流转。各类单位、公众等用户群体都迫切需要获得真实、完整、准确、可靠、可用的档案数据，以满足工作需求，解决遇到的难题。在此过程中，档案数据正处于现行的流转过程，用户的信息需求表现为信息粒度细，需求量大且质量要求高，实时性强，对档案数据质量控制的要求也更为苛刻。因此，需要根据需求，收集相应数据信息，并对其进行组织与分析，使之可获取、利用。一方面是档案数据收集。收集的档案数据必须完全满足需求，并且能准确收集足够量的数据来解决工作难题。对此，在该环节中，应当明确档案数据的收集范围，掌握收集流程、方法，并确保收集数据的质量，使其准确可用。另一方面是档案数据组织与分析。应当依据一定的标准、技术手段对收集的数据开展数据类目、元数据元素、数据格式、数据语义等规范性的组织与分析，使档案数据满足可读、可控、可

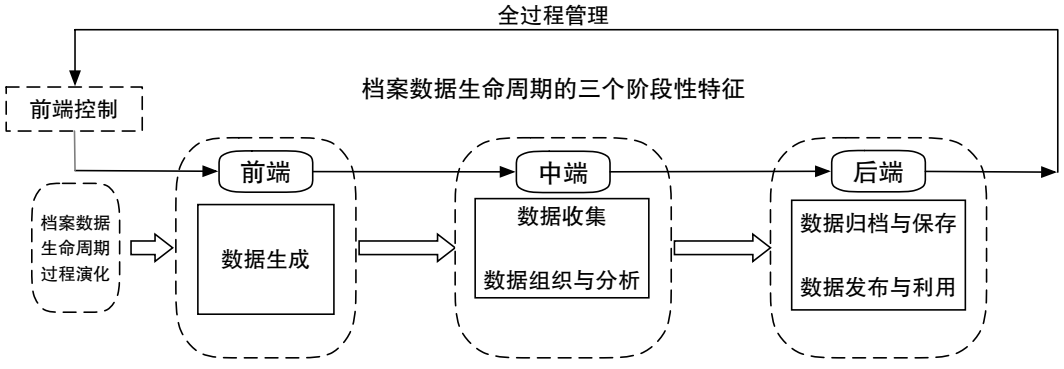

图1　档案数据生命周期过程

利用等通用技术指标，从而确保档案数据满足工作需要。

3.3 档案数据生命周期的后端

该阶段的档案数据已完成现行阶段的任务，需要归档保存，以备查考利用，并在经过漫长的保存期后，向社会提供利用，供其开展研究，创造出更高的数据价值。在此过程中，档案数据的质量问题依然存在，不可放松警惕。一方面，档案数据归档与保存。应当掌握数据归档方式，依据科学的归档程序，确保所有相关数据记录都能合理归档。同时，重视对处于保存期的档案数据质量的管控，防止发生数据变质、外部网络攻击造成的数据损失事件。另一方面，档案数据的发布与利用。该环节已完全向社会开放，数据质量隐患将更为复杂，既要保障向社会提供开放获取的档案数据都是高质量的，也要确保数据在满足用户需求后依然能保持质量效果，防止发生损坏、丢失、泄密、被窃取等质量问题。因此，需要创造安全合理的空间环境以保障档案数据质量，使其能以准确的数据内容、高效的获取速度来满足用户需要。同时，要重视对用户利用行为的规范，严密监视用户的利用动向，从而保证档案数据的复用质量。

4 生命周期视角下档案数据质量控制的阶段性过程解构

不同生命周期阶段的档案数据质量控制要求在目标、管理标准、技术方法、控制行为等方面各有侧重。既要从档案数据产生的源头出发保障其价值和质量，也要在档案数据生成后的收集、分析等中端阶段进行数据监督和数据行为的引导，从而加强数据运行过程的控制，还要确保处于长期保存过程中的档案数据质量稳定，使其以高质量面向社会提供开放利用。基于此，应对档案数据前端、中端、后端三个生命周期阶段进行过程解构（图2），分析各阶段的特征和需求，为档案数据质量控制提供更为微观和清晰的视角。

4.1 前端：明确控制目标需求，科学规划行动方案

首先，明确档案数据质量控制的目标需求。档案数据质量控制应明确为何开展，要达到何种目的，从而为后续控制方案、控制标准、控制行动等规划提供方向指引和实践依据。一是明确档案数据管理的质量目标。即在管理档案数据时制定对应的目标，并朝着该目标开展质量控制。如应明确档案数据管理符合档案管理规范，保证归档数据记录来源可靠、要素合规，并且“应收尽收”“应归尽归”；确保档案数据在整个生命周期过程中保持真实性、准确性、可靠性、可信性、完整性、关联性、一致性、可控性的效果；档案数据要有完善的防灾减灾机制，并做好数据备份，预防并有效应对安全事故的发生。二是重视档案数据利用者的质量需求。即档案数据在提供利用时，保证其质量效果能满足用户的需要，并要确保档案数据在利用行为完成后，依然保有质量，能不断重复利用。如应保证档案数据满足易获取、可读、可运行、格式兼容性强等通用技术指标；档案数据开放获取的时效性强，能第一时间提供给用户并满足其需求；档案用户的利用行为能得到有效指导和监控，保证档案数据在满足需求后依然保持高质量。

其次，整体规划档案数据质量控制方案。制订档案数据质量控制方案旨在对整个生命周期过程中如何开展控制行动、怎样保障数据质量进行宏观层面的规划。一是精确划分档案数据管理主体的权责义务。档案数据质量控制方案应以制度条文的形

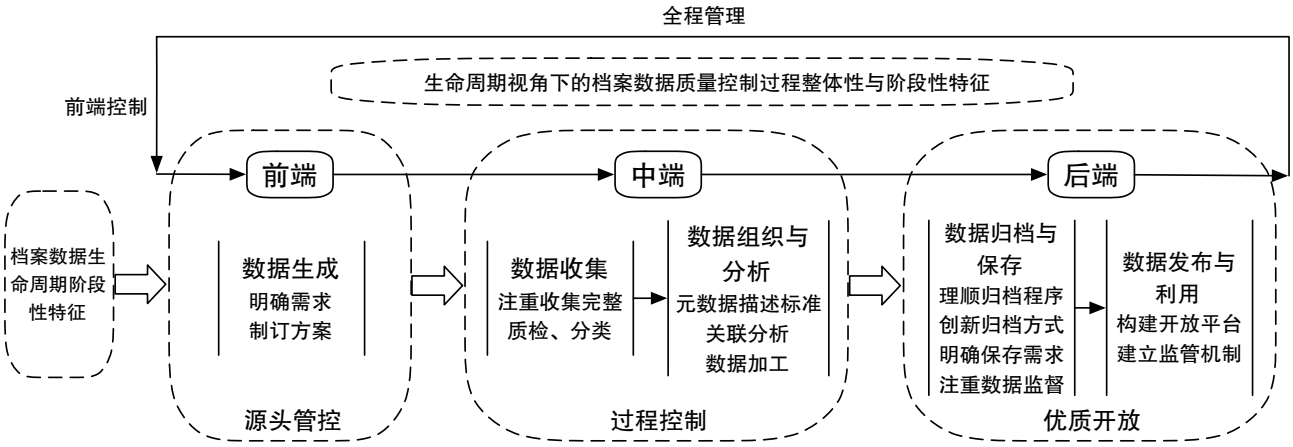

图2 生命周期视角下的档案数据质量控制过程

式将管理主体的权责义务进行清晰划分，使相关主体控制行为得到规范，从而保障档案数据管控的合理性，降低由于管控原因导致的质量问题的发生概率。二是从宏观上设计档案数据质量控制计划。科学的档案数据质量控制计划能够有效指导档案数据生成、收集、组织与分析、归档与保存、发布与利用等各阶段的控制行为，使其有章可循。具体包括在生成阶段制定元数据、数据格式、数据结构等规范体系；在收集阶段明确收集方法、收集流程、整合集成规则等；在组织与分析阶段制定合适指标选取数据处理工具、掌握分析方法、数据支撑软硬件等；在数据归档与保存阶段明确归档方法、归档流程，确定保存方式，掌握保存工具，制定长期保存的质量维护策略等；在利用发布阶段应划分用户利用权限、制定用户利用行为规范、明确开放利用格式。如2016年浙江省人民政府发布《浙江省促进大数据发展实施计划》（浙政发〔2016〕6号），其中规划了通过制定大数据归档范围、标准，统一归档平台等举措，促进了大数据证据保全、长期保存和再利用，对合理管控档案数据、保障数据质量具有积极意义。

4.2 中端：全面收集数据资源，分析组织档案数据

4.2.1 重视数据收集的完整性，提高质量检测的准确度

首先，收集齐全完整的档案数据。即强调数据收集的全面性，足以支撑后续各类分析需求。一是构建档案部门与各层级数据保管机构的协同联动机制。由于大数据环境下，档案数据分散保存在不同机构和单位，档案数据资源收集要突破单一主体界限，在跨层级、跨系统、跨部门、跨区域间实现档案数据资源集成。[36]对此，档案部门应主动出击，疏通并协调其与各类数据保管机构之间的沟通协作关系。具体应明确各机构数据权责，签订数据共享协议，建设数据整合渠道，推动数据之间的有序流通与共享，为各单位收集齐全完整的数据资源以满足其自身需求提供便捷通道。二是建立档案部门与社会主体之间的沟通协作机制。产生档案数据的不仅有数据保管机构，还有各种类型的社会主体，包括公众、社会组织、企事业单位等都可生成并保存有高价值的档案数据，像网络舆情数据、社交媒体数据、医疗数据、企业业务数据等，都具有保存价值。由于该类数据分布零散，且收集管辖长期以来都处于被忽视的状态，所以收集该类档案数据资源时，应加强档案部门与社会主体之间的协同联动。可建设档案部门面向社会沟通交流的通道，加强档案宣传教育，鼓励社会各界向档案部门主动提交有价值的数据记录。促进社会档案数据资源融入档案部门的管控范畴，使档案数据收集范围延伸到社会生活的各个角落，保证数据资源的完整性。如美国国家档案和记录管理局（NARA）构建了跨政府部门主体合作、与社会领域资本合作、与用户沟通的机制[37]，有效加强了各主体之间的沟通，利于数据的互相流通共享。

其次，强化对收集数据的质检和分类。收集阶段的档案数据质量控制相当于一个关卡，需要提高对收集数据质量检测的准确性和效率，详细识别出不同种类的档案数据，为后续数据的利用和价值开发提供高质量的归口数据。一是重视对档案数据信息的审查核验。由于档案数据数量巨大、结构复杂的特征，人工操作已难以满足质量控制需求，可采用大数据、人工智能等技术实现档案数据的自动审核。重点核验收集数据的元数据描述是否规范，数据是否可用、可重复使用、可读、可运行，数据格式能否与通用的操作系统兼容等档案数据自身的质量。还要检测收集的数据与需求的相关性，并检测有无重复、冗余的数据，将其进行清洗。二是将档案数据进行科学分类。通过数据审核程序后，应对档案数据进行分类，使杂乱无章的数据从无序变成有序，表现出数据之间的逻辑关系，使反映客观事实的档案信息变得清晰可见。可根据档案数据的来源，按照功能、组织结构或是主题进行分类[38]，并辅以机器学习技术和人工帮助的方式，使档案数据种类得到智能识别，实现自动分类，提高分类的准确性。如西北民族大学利用朴素贝叶斯分类器对甘肃省档案馆提供的档案数据资源进行档案文本内容主题分类和内容识别[39]，提高了档案数据信息分类的准确性和质量检测效率。

4.2.2 提升数据组织的科学性，分析数据资源的关联度

首先，科学组织档案数据。即通过特定的标准规范，采用一定的工具，对档案数据进行类型、形态、属性的分析，从而实现档案数据的科学组织，使档案数据具备易获取性、可查找性、可访问性、增值性、互操作性。[40]档案数据组织关键在于合理进行元数据描述，确保元数据描述的高质量与准确性，使档案数据各元素之间在逻辑上保持相关性和一致性，内容上能完整反映客观事实。在实际操作中，应以足够量、完整全面的档案数据集为基础，对数据集内部包含的创建者、时间、地点、人物、机构、

上下文等信息进行分析描述，在数量庞大、数据分散并且结构各异的档案数据集内部梳理各数据之间的逻辑关系，实现对档案数据的语义检索和价值挖掘，以便用户对数据的获取和利用。此外，由于元数据标准纷繁复杂，在元数据描述标准选取时，应尽可能统一标准规范，并采用现存的、已形成广泛共识的标准体系，以扩展和引用的方式描述档案数据，防止发生标准不一致造成管理漏洞并导致质量下降的事件。如澳大利亚联邦科学与工业研究组织（CSIRO）强调尽可能利用已经形成广泛共识的元数据标准[41]，以此加强数据的互操作性，提高使用质量。

其次，分析处理档案数据。即在庞大的数据资源池中将异构、离散、跨界的档案数据建立数据关联，开发数据价值，促进档案数据之间以及档案数据与其他数据资源之间的兼容性与互操作性。同时为满足多样化用户需求，需对数据进行加工分析，提高利用效果。一是强化档案数据的关联分析。可使用关联数据技术建立档案数据多维语义关联框架[42]，建立异构、跨界、来源不同的档案数据资源之间的语义关联，实现档案数据的智能控制和准确检索，增强数据获取的简易性和互操作性。二是确保档案数据的加工质量。档案数据的加工意在挖掘数据价值，开发数据产品，满足多样化的科研或是工作需求。加工过程中需要高质量的数据支持，随时检测数据质量问题，确保加工产物的可用性和可靠性。如可综合采用信息抽取技术、数据关联技术增强档案数据的同一性、相关性、隶属性等语义关联，可在增强数据关联的同时有效提升数据异常情况的检测[43]，保证加工过程的万无一失。

4.3 后端：保障保存质量稳定，实现数据优质利用

4.3.1 强化数据归档的合理性，确保长期保存的高质量

首先，优化档案数据归档管理体系。应改善归档理念，探讨文件归档向数据归档的范式转型，提升数据归档效率，保障归档数据记录的高质量。一是理顺档案数据的归档程序。目前，档案数据化实践正处于初步发展的态势，以数据驱动为特征的相关理论和方法尚未在数据归档保存领域取得实质进展。[44]对此，应由国家档案部门主导，建立统一的数据移交、呈缴规范制度，将分散、碎片化的数据记录统一归档保存，保证档案数据的完整性。具体应在制度中明确档案数据的归档格式，完善归档数据记录的接收流程，注重数据在接收、传输过程中的安全防护，制定系统的安全保障条文，从而达到档案数据高效、安全、优质归档的目的。二是创新归档方式。档案数据在归档程序中应根据数据特征选取合适的方法，在技术创新、管理创新上探讨适应性的实践策略，合理构建技术设施布局、数据资源管控方式等内容。如空客德国公司根据其海量业务数据的归档需求，建设了可以对数据进行提取、访问、维护、保存，以及检查反馈的归档管理系统——ZAMIZ系统[45]，提升了数据归档效率，也保障了归档数据质量。

其次，制定科学的长期保存策略。档案数据能否以可识别、可运行、可复用的方式向整个社会提供开放获取服务，其在长期保存过程中的质量控制是关键。一是明确档案数据的保存需求。大数据环境下，由于档案数据摆脱了“文件”格式的束缚，在对其的长期保存策略选取上，应从各自业务系统内在关联出发，基于各自场景解决档案数据语义表达与管控的难点[46]，超越原有的文件形式保存体系，构建维护数据语义完整的保存体系。二是注重对保存阶段档案数据的监督。档案数据的长期保存持续时间及管控周期长，应加强对保存期档案数据的监督力度，定期开展质量核验，在发现质量问题时及时预警，做到有效防范安全风险或是降低质量问题发生时遭受的数据损失。如可设立档案数据监督委员会，明确监督管控的职责分配，合理开展对保存期档案数据的监督和业务指导。亦可构建档案数据的长期保存质量评价机制，从所保存的档案数据自身特征出发，制定档案数据质量评价指标，赋予各评价指标对应的权重，建立各评价指标的分值区间，形成完善的评价指标体系[47]，从而使保存期的档案数据在数据质量核验过程中有据可依。

4.3.2 优化数据服务质量效果，建立用户利用监管体系

首先，提升档案数据服务质量。大数据时代，用户对档案服务的需求越发多样化，应秉持社会信息资源共建共享理念，推动档案数据服务的创新升级，构建精准、优质、高效的档案数据服务体系。探讨构建档案数据开放共享平台，向社会公众提供“一站式”数据开放共享服务。通过强化平台功能模块，统一档案数据开放利用的格式、结构、标准，并借助本体建模、可视化分析、人工智能等技术对档案数据进行同构化与细粒度处理，促使档案数据提质增效，增强其易获取性、可操作性、可用性，从

而使用户容易检索、编辑与利用数据，提升数据驱动解决难题的效率。如中国联通电子档案馆集成档案信息资源库和档案知识库，探讨构建档案数据挖掘服务平台和档案多维展示平台，并考虑运用数据挖掘与机器学习的方法实现档案数据的智能推理与检索[48]，有助于提升平台的数据服务质量，为企业运转提供智能化数据支持。

其次，加强档案数据的利用监管。一是设置档案数据访问权限。应加强档案数据资源库的访问控制权限设置，对用户进行身份认证，设定不同的等级，并赋予相对应的数据利用权限。同时，制定档案数据许可协议，允许用户在确保档案数据质量安全的前提下复制、分发、传输、重复利用数据，规范用户的利用行为，从而降低因操作不当导致的数据质量下降概率。二是注重档案数据利用过程的监管。科学监管有助于实时发现档案数据在提供利用过程中的不安全、不规范的利用行为，加强对档案数据在传输、运行过程中的风险预警，遇到用户操作不当或是恶意篡改、窃取、越级访问数据库的行为时应及时制止，并采用溯源技术追究其责任。如山东省国土测绘院建设的“地理信息档案数据安全系统”，能实现对地理信息档案数据的加密管理、访问控制、全程跟踪，妥善保障了档案数据的质量安全。[49]

5 结 语

大数据时代，数据海量涌现，广泛渗透，驱动着众多行业的档案管理工作向档案数据管理工作转型。现有的管理理论、方式缺乏足够的力度来保障档案数据在生命周期各阶段的质量。通过借鉴生命周期理论，划分档案数据生命周期过程，并根据其生命周期各阶段的特征，采取科学的数据质量控制方式，有助于优化档案数据生态环境，保障档案数据在生命周期过程中都能以高质量满足用户需求。考虑到大数据仍在持续向前发展，档案数据还可能会遇到更多的新问题，在之后的研究中，应不断适应档案数据管理的现实需求，全面优化质量控制举措，从而使档案数据能为经济社会发展提供源源不断的高质量数据支撑。

本文系国家档案局科技项目“面向‘数据要素X’行动的档案数据安全治理体系及关键技术研发与应用研究”（2024-X-016）的阶段性研究成果。

（通讯作者：周林兴）

注 释 及 参 考 文 献

[1][25] 国家档案局. 中办国办印发《“十四五”全国档案事业发展规划》[EB/OL].[2023-01-12]. https://www.saac.gov.cn/daj/yaow/202106/899650c1b1ec4c0e9ad3c2ca7310eca4.shtml.

[2] 国家档案局. 在全国档案局长馆长会议上的报告[EB/OL].[2023-01-12]. https://www.saac.gov.cn/daj/yaow/202203/5b5257a20b964995b22afc1d585382b1.shtml.

[3] 杨太阳. 把握科技脉动 探索管理创新——第八届中国电子文件管理论坛在京召开[N]. 中国档案报，2017-12-18(1).

[4][23] 金波，添志鹏. 档案数据内涵与特征探析[J]. 档案学通讯，2020(3)：4-11.

[5] 杜彦峰，相丽玲，李文龙. 大数据背景下信息生命周期理论的再思考[J]. 情报理论与实践，2015(5)：25-29.

[6][17] 金波，杨鹏. 大数据时代档案数据治理研究[J]. 档案学研究，2020(4)：29-37.

[7] 陈慧，罗慧玉，陈晖. 档案数据质量要素识别及智能化保障探究——以昆柳龙直流工程项目档案为例[J]. 档案学通讯，2021(5)：49-57.

[8] 周林兴，崔云萍. 大数据视域下档案数据质量控制实现路径探析[J]. 档案学通讯，2022(3)：39-47.

[9] 周林兴，林凯. 大数据时代档案数据质量控制：现状、机制与优化路径[J]. 档案与建设，2022(2)：4-8.

[10] 周林兴，黄星. 大数据时代档案数据开放共享监管：价值、机制与推进理路[J]. 档案与建设，2023(8)：6-10.

[11] 葛泽钰. 基于PDCA循环的档案数据质量控制探究[J]. 档案与建设，2023(8)：40-43.

[12] 刘越男. 数据治理：大数据时代档案管理的新视角和新职能[J]. 档案学研究，2020(5)：50-57.

[13] 周林兴，黄星. 大数据时代档案数据开放共享机制探析[J]. 档案与建设，2023(3)：8-12.

[14] LEMIEUX V L. Trusting records: is Blockchain technology the answer?[J]. Records Management Journal, 2016(2)：110-139.

[15] 何玉颜. 档案部门参与政府大数据治理的路径研究[J]. 浙江档案，2018(8)：23-25.

[16] DIMAKIS A G, GODFREY P B, WU Y, et al. Network coding for distributed storage systems[J]. IEEE transactions on information theory, 2010(9)：4539-4551.

[18] 国家档案局政策法规司. 2023年度全国档案主管部门和档案馆基本情况摘要(二)[EB/OL].[2024-12-27]. https://www.saac.gov.cn/daj/zhdt/202409/a277f8b3bfe942ca88d3b7bcf6ddf120.shtml.

[19][35] 钱毅. 数据态环境中数字档案对象保存问题与策略分析[J]. 档案学通讯, 2019(4): 40-47.

[20] 周国刚. 浙江绍兴"树牢档案数字化思维"[EB/OL].[2023-05-04]. http://www.zgdazxw.com.cn/news/2021-09/07/content_325016.html.

[21][22] 杨鹏. 大数据时代档案数据权利及其体系构建[J]. 档案学通讯, 2022(4): 51-57.

[24] 朝乐门, 邢春晓, 张勇. 数据科学研究的现状与趋势[J]. 计算机科学, 2018(1): 1-13.

[26] 刘德寰, 李雪莲. 数据生态的危险趋势与数据科学的可能空间——兼谈中国市场调查业的现状与问题[J]. 现代传播(中国传媒大学学报), 2016(1): 21-27.

[27] 江跃中, 方翔, 潘高峰. 上海居民电子健康档案数据质量堪忧 委员建议制定电子健康档案建设应用规范[EB/OL].[2023-05-04]. https://baijiahao.baidu.com/s?id=1655847991850559992&wfr=spider&for=pc.

[28] 金波, 晏秦. 数据管理与档案信息服务创新[J]. 档案学研究, 2017(6): 99-104.

[29] 大数据战略重点实验室. 块数据2.0: 大数据时代的范式革命[M]. 北京: 中信出版社, 2016: 71.

[30] 王向女, 袁倩. 美梦还是陷阱? ——论数据科学背景下的档案数据管理[J]. 档案与建设, 2019(9): 4-7, 12.

[31] 赵跃. 大数据时代档案数据化的前景展望: 意义与困境[J]. 档案学研究, 2019(5): 52-60.

[32] CORTI L, EYNDEN V V D, WOOLLARD M, et al. Managing and sharing research data: a guide to good practice[J]. Records Management Journal, 2014(3): 252-253.

[33] 武彤. 基于数据生命周期的美国研究图书馆科学数据开放共享服务研究[J]. 图书与情报, 2019(1): 135-144.

[34] 黄霄羽. 文件生命周期理论对机关文档管理的启示[J]. 档案学通讯, 2003(5): 65-69.

[36] 金波, 陈坚, 李佳男, 等. 大数据时代档案数据资源整合探究[J]. 档案与建设, 2022(9): 18-23.

[37] 白文琳, 安小米. 政府电子文件协同管理: 美国经验及其启示[J]. 档案学通讯, 2020(4): 103-112.

[38] 于英香, 刘茜. 论计算档案学的出场逻辑[J]. 档案学通讯, 2021(5): 22-31.

[39] 杨建梁, 刘越男. 机器学习在档案管理中的应用: 进展与挑战[J]. 档案学通讯, 2019(6): 48-56.

[40] 张洋, 肖燕珠. 生命周期视角下《科学数据管理办法》解读及其启示[J]. 图书馆学研究, 2019(15): 37-43, 13.

[41] CSIRO. Introduction to research data management[EB/OL].[2023-01-22]. http://libguides.csrio.au/ResearchDataManagement.

[42] 王志宇, 熊华兰. 语义网环境下数字档案资源关联与共享模式研究[J]. 档案学研究, 2019(5): 114-119.

[43] 牛力, 曾静怡, 刘丁君. 数字记忆视角下档案创新开发利用"PDU"模型探析[J]. 档案学通讯, 2019(1): 65-72.

[44][47] 杨文娜. 大数据环境下归档政务信息长期保存研究[J]. 档案学通讯, 2022(1): 109-112.

[45] 高闯, 柳林集. 合规与妥协: 空客德国产品数据归档的现状及其启示[J]. 档案学研究, 2021(2): 119-124.

[46] 钱毅. 新技术环境下电子文件管理纵深发展关键问题分析[J]. 档案学通讯, 2020(2): 4-9.

[48] 杨茜雅. 中国联通电子档案数据挖掘与智能利用的研究[J]. 档案学研究, 2018(6): 105-109.

[49] 赵君. 共建共享背景下地理信息档案安全系统建设实践[J]. 中国档案, 2022(10): 62-63.