结构化数据环境下政务类原生数据采集的研究

文◆杭州易康信科技有限公司　张　雷

引言

结构化数据即数据库，一般是指可以使用关系型数据库表示和存储，用二维逻辑表达实现的数据。以行为单位、一行数据表示一个实体信息，且每一行数据的属性相同，存储在数据库中。此外，结构化数据能用数据或统一的结构加以表示，如数字、符号等。因此，结构化数据的存储和排列具有规律性，有助于实现查询和修改等操作。典型运用场景包括企业 ERP、财务系统、医疗 HIS 数据库、政府行政审批、政务办理等。为了构建结构化数据环境，2022 年 10 月，国务院办公厅印发《全国一体化政务大数据体系建设指南》，要求整合构建标准统一、布局合理、管理协同、安全可靠的全国一体化政务大数据体系，加强数据汇聚融合、共享开放和开发利用，促进数据依法有序流动，充分发挥政务数据在提升政府履职能力、支撑数字政府建设以及推进国家治理体系和治理能力现代化中的重要作用。

1 结构化数据环境下政务类原生数据采集的特点

政务类数据原生数据是指政府机构在正常运行和履行职责过程中产生的最初的、未经加工、保持原始状态的数据。通常由政府部门自身采集和生成，用于支持政务管理、决策制定和公共服务，包括统计数据、行政数据、教育数据、卫生医疗数据、环境数据和公共安全数据等。在结构化数据环境下，采集特点包括数据标准化、数据精确性、数据完整性、数据一致性 [1] 和数据安全性。

数据标准化要求原生数据按照一定标准进行组织和存储，包括使用统一的数据格式、命名规则和编码体系，在共性采集标准下，确保数据的一致性和可读性。数据精确性要求数据的准确性和精确度高，来源可靠。通过明确的数据字段和数据类型，减少数据录入和转换过程中的错误和偏差。数据完整性应确保数据的全面和完整，即原生数据不仅要保证现役数据的完整性，还要参与退役原生数据的拯救和长久保存。同时，数据的所有必要字段都应正确填写，避免数据遗漏和不完整导致信息丢失和误解。数据一致性是指原生数据采集要求数据在不同系统和平台间保持一致的定义和格式，以实现数据的互操作性和集成性，方便数据归档和分析。数据安全性在采集过程中注重数据的安全性保护，包括敏感数据的加密和权限控制，防止未经授权地访问和数据泄密。

结构化数据环境下政务类原生数据采集面临着一系列挑战。例如，共性和差异化的行业特性采集标准虽然存在，但大部分尚未形成国家级或者省级标准，使后续数据整合和共享变得困难；在数据质量管理方面，目前数据的完整性 [2]、准确性和一致性仍缺乏有效的数据采集和验证机制，且数据来源模糊，程序不合理，溯源困难。此外，原生数据采集需要相关的技术能力和资源投入，部分政府部门在该方面存在限制，影响了数据采集的效率和质量。

2 政务类原生数据采集原则

2.1 因地制宜

在政务类原生数据采集过程中，应秉持因地制宜原则，在遵循共性采集标准的基础上，根据不同行业要求，制定具有行业特性的采集标准。共性采集标准包括规范性文件引用、数据编码规则、保存格式、保管期限等。例如，某市政务类原生数据编码规则以全宗号、门类、年度和保管期限为命名基础，归档数据包、结构化数据档号和年度归档数据

【作者简介】张雷（1972—），男，北京人，研究方向：结构化信息数据处理、政务信息处理。

补缺包命名方式，并不断优化延伸。其中，“全宗号”是各级档案馆为已进馆或计划进馆的单位分配的编号，用于区分各立档单位，便于日常管理和查询[3]；“门类”是指政务业务系统结构化数据档案的归档门类按照“一级门类·二级门类”的方式进行编制；“年度”共4位，以数字字符表示，表示数据的生成年份；“保管期限”分为定期和永久，定期以字母“D”和保管时限表示，如“D30”；永久以字母“Y”表示。共性采集标准制定保证了后期溯源查证，也利于政府各个部门和机构进行数据交互和共享。针对不同政务领域的特殊要求和业务特点，采集标准应考虑到特定领域的数据要求、数据格式和数据分类，确保数据满足不同的管理和决策要求。例如，住房公积金业务要求专业的个人信息采集标准、缴存信息采集标准、贷款信息采集标准及业务办理信息采集标准，包括数据匹配表、字段长度、字段类型和字段描述等，故对公积金中心的数据采集归档能力提出了更高要求。

2.2 确保归档数据“一数一源”

在采集过程中，应遵循的另一原则是归档数据“一数一源”，即每个数据项只有一个来源且所有相关的数据副本都指向一个源头，以提高数据可信度和管理效率，避免数据冗余、不一致及混乱。运用数据源标识，保证每个数据项能够追溯到原始数据；建立合适的数据传输和同步机制，确保数据在不同系统之间的传输和同步可控、有序；所有数据更新和变更应经过正确流程和授权，并将其同步到所涉及的地方，保持一致性；实施严格的数据访问权限控制，避免非授权人员更改数据。

2.3 确保数据来源可靠、程序规范、要素合规

确保数据来源可靠，且采集程序规范、要素合规，提高数据质量和可信度，保护数据安全和隐私。为了实现这一要求，相关部门对归档数据包、补缺数据包和归档 / 补缺数据包过程管理文件可通过区块链、可信时间戳和数字摘要等技术手段进行归档数据存证，并按数字档案、数据档案最小利用单元为原则进行存证保护，确保其真实性、完整性和可用性。具体归档数据存证方式有“数据级 + 文件级”存证、数据级存证、文件级存证3种方式。在条件许可情况下，优先选用“数据级 + 文件级”存证方式执行，即对拟归档的结构化数据逐条进行数据级存证，确保各类原始数据可独立进行验证验真。同时，在档案整理、组包、交接等管理过程时，对各管理环节所有发生变量管理信息进行存证保护（按过程逐级叠加），确保过程管理信息真实可靠、可追溯、可校验。

此外，除了进行数据存证外，当归档数据包、归档数据补缺包和过程管理文件产生变量或位移时，有关部门应进行数据验证。在验证验真通过后，方可继续实施后道归档管理工序。且数据档案利用时，各类出证数据和数据归档全过程信息均可正、逆向全过程溯源。

3 政务类原生数据采集意义

在数字中国战略带动下，社会原有生产方式、运行机制和发展动能均发生重大变革，政务类业务档案工作也产生了颠覆性变化。大量泛在、多源异构的原生数据受困于对源系统的依赖，游离于国家档案资源范畴之外，数据档案化治理陷入了“首席缺位、中段停滞、末端无奈”的尴尬状态。原生数据“各自为政”，脱离原生环境应用，存续难，数据档案真实性权威性难以确认，且缺乏标准创建体系，数据尚未达到“可控易用”阶段。因此，结构化数据环境下，原生数据采集的最终目标是实现原生业务数据归档后的治理，做好大数据时代档案资源建设和共享利用的基础性工作，同时强化多跨协同、制度再塑、流程再造，努力构建融合式、闭环式、区域化原生数据归档治理新模式，实现原生数据档案化的单元智治和整体智治。2021年，《“十四五”全国档案事业发展规划》颁布[4]，提出了“加快推进档案信息化建设，引领档案管理现代化”的发展任务。具体措施如下，完善档案信息化发展保障机制[5]，加强电子文件归档和电子档案移交接收，加速数字档案馆（室）建设，推进档案信息资源共享平台建设等。该规划体现了国家完善档案治理体系建设的决心，给政务类业务数据归档治理带来了政策导向和变革动力。

4 相关案例分析

杭州易康信科技有限公司成立于2007年，深耕大数据管理，特别是在政务业务类办事应用系统的研究和探索，创新性针对政务业务搭建信息系统开发平台（OEA），开创了全系列数据档案管理先河，突破了结构化数据归档的瓶颈，在细分领域卓有成效。合作案例包括嘉兴住房公积金、嘉兴医疗保障公共事务信息系统、桐乡人力资源和社会保障金信息体系等政府关键部门数据管理部门业务。

4.1 嘉兴住房公积金——现役的原生业务系统连续性归档

2023 年，针对“数据孤岛”问题，“易康信科技”携手嘉兴市档案部门深入推进档案数字化改革，创新创建档案管理新模式，先后出台《嘉兴市市级机关政务业务系统数据归档暂行标准（通用部分）》（以下简称《标准》）和针对各单位数据的专用标准，确定了原生数据档案在线获取、规范管理、有效利用的方法和途径。嘉兴住房公积金便是试用业务之一。《标准》明确了适用范围和规范性引用文件，将住房公积金综合业务数据、依据性附件、办理流程等数据整体划入归档范围，制定了共性和专业编码规则，规定了数据采集规则，即每月 1 日 0 点采集上月全量应归档数据，包括结构化数据和非结构化数据，保证数据来源的可持续性。对于数据存证，《标准》采用“数据级 + 文件级”存证方式，以区块链为存证技术手段，且归档数据中包含存证的数据摘要，提供离线状态下验证，确保原生数据档案来源真实可靠、全流程篡改留痕、数据固化保真。此外，根据公积金业务数据采集需求和特征，《标准》制定了一系列样表，如《公共基础数据（单位信息）采集与归档元数据匹配表》《公共基础数据（个人信息）采集与归档元数据匹配表》《归集和提取基础数据（单位账户信息）采集与归档元数据匹配表》等，规定了具体业务数据采集项，包括字段类型、字段名称和字段长度，给数据采集提供方向和标准规范，确保现役的原生业务系统连续性归档，以实现原生数据档案全流程、全管控、全生命周期管理运行。

4.2 嘉兴人社——退役的原生业务系统的数据拯救及长久保存

退役原生业务系统数据是指已经停止使用或者废弃的数据，包括过去时间段内的历史数据和需要归档的数据，这些数据不再频繁使用但仍具有重要价值。通常需要进行合理保存和处理，数据采集和归档是重要的保存手段。但由于技术及历史原因，部分退役原生数据并没有进入“国统”或者“省统”范围，面临着遗失和残缺的风险，形成了数据“孤本”。为了妥善处理退役数据，嘉兴人力资源和社会保障局与“易康信科技”共同搭建信息开发平台，拯救退役数据，实现了多元汇聚、长期保存、有效利用，推动了退役原生数据早日纳入国家档案资源管理范畴，最终实现数据贯通，形成业务闭环。

4.3 区块链存证技术在政务类原生数据采集中的应用

区块链存证作为电子数据存证的一种，通过区块链技术实现电子数据证据的固定与取信。区块链本质上是一个分布式的共享账本和数据库，具有去中心化、集体维护、多方透明、安全防护等特点，存储于其中的数据或信息具有不可篡改的特征。基于这些特征，区块链技术在数据采集中的合理运用可以确保原生数据的完整性、真实性和隐私性，提高数据的可信度和可验证性。“易康信科技”所采用的区块链存证技术贯穿采集、预归档处理、校验整理、入仓、移交、接收、馆藏的全流程，实现了归档 / 补缺数据存证和档案管理过程文件存证，探索了区块链技术赋能对业务系统原生数据档案的真实性保障和溯源验证的有效方式，完成区块链技术跨部门组网和对原生数据档案存证验真的落地应用。

结语

结构化数据环境下，政务类原生数据采集具有重要的战略意义，是实现原生业务数据归档后治理的必要步骤，也是政府科学决策和促进服务创新的关键途径。然而，原生数据采集仍然面临着严峻挑战，包括采集标准规范缺乏、存证技术不成熟、“数据孤岛”和“数据孤本”现象普遍等。因此，政府部门和企业应该共同努力，不断优化采集标准，推动国家级和省级标准规范的创建，科学探索原生数据全生命周期的管理模式，助力数字化转型，实现智慧政府目标，提升公共服务水平，增强社会效益。

引用

[1] 龚剑超,徐国华,戴建军.关于数据采集分析的结构化思考[J].水电站机电技术,2019,42(11):6-9.

[2] 王俊松,朱辰,边荟淞,等.基于数据中台的高职院校数据画像系统建设[J].无线互联科技,2022,19(20):143-145.

[3] 肖佳祥,侯荣泽,邓元旭,等.基于SSM的供电所电子档案管理系统的设计与实现[J].科技创新与应用,2020(29):34-36.

[4] 陈泊舟.电子档案单套制管理的困境及破解策略[D].湘潭:湘潭大学,2022.

[5] 宫晓东,丁海悦.论乡镇档案信息化建设的四个维度[J].档案与建设,2022(4):40-43.