非结构化电子档案数据管理探析

尹 悦
（新乡职业技术学院 河南 新乡 453000）

【摘 要】非结构化电子档案数据数量庞大，增长速度较快，对于现有的数据档案存储检索管理而言产生了巨大技术挑战。非结构化电子档案管理需重新调整技术手段，采用 OS 系统文件管理技术、分布式文件管理的对象存储技术以及针对 erms 环境进行管理方案调整等方式，提高存储能力和检索效率，提升管理质量。

【关键词】非结构化档案数据；信息；数据库；存储；检索

非结构化档案数据与结构化档案、半结构化档案相比的主要差别体现在档案数据模型特征之上。非结构化档案的是档案特征通常以不规则、不完整为就基本表现特征，因此相比于结构化、半结构化档案的数据管理来说，非结构化档案数据管理主要面临存储和检索两个方面的难点。随着信息技术的快速发展，非结构化档案数据信息体量快速增多，需要进行相应的配套管理技术机制来带动管理模式升级，提高对于非结构化数据档案的管理能力水平。

一、非结构化数据档案的主要生成来源

（一）来源于自动化办公软件生成

现代办公软件和自动化系统的广泛应用，带给了人们工作极大的便利性，有效提高了信息化工作效率。但是在办公软件自动化的使用过程中，所产生出的诸多电子文件信息存储，均表现为非结构化特征。信息化发展过程中，办公自动化软件普及程度更高，非结构化的电子文件形成方式也更为多样，办公电子文件成为目前非结构化数据的主要来源。

（二）来源于档案数字化发展

《全国档案事业发展规划纲要》明确要求要全面推进传统档案文件资源的存量数字化建设，利用数字化来完成原本纸质档案信息的全面转化。相应的，档案管理方面也积极推进数字化存储机制的建设和升级，数字档案逐渐成为主要的档案管理形式。不过，与一般数据不同，数字档案的存储方式多以图片形式进行存储，同时部分文件需要通过文字识别等技术手段进行转化，这种档案形式的数据信息本身便是非结构化的，因此数字化档案也是目前较为主要和常见的非结构化档案数据来源。

（三）来源于其他渠道

除了上述两个生成渠道之外，非结构化档案数据还来源于其他不同领域。常见的档案信息来源如口述档案等，是较为典型的非结构化档案数据信息。此外在管理工作过程中所必须使用的工作日志、信息浏览记录或者电子邮件等，也属于非结构化档案数据，存储于网络数据系统之中。

二、非结构化档案数据的类型化特征

（一）非结构化档案数据的主要类型

目前非结构化档案数据体量庞大类型众多，由于不同来源所产生的文件编码形式也不同，最终出现在数据系统当中的非结构化档案数据也有着多样性的文件格式特点。目前数据系统当中非结构化档案数据主要可以分为下述几种类型：

首先是文档文件，这种档案数据来源于日常办公过程中所产生的文件信息或者由办公人员主动保存的重要文档内容。这类文档文件以文字内容为主，主要的文件格式由 PDF 文件、doc 文件、xls 文件、txt 文件等。

其次是图片类型文件，这类文件来源于数字化档案工作过程中，常见文件类型如摄影记录、影像资源、设计图纸等内容。主要以 dwg 文件、dxf 文件等图形文件和 JPG 文件、bmp 文件、png 文件等图像文件格式为代表。

其三是音视频类型文件，在现代信息技术发展中，档案存储开始采用音视频方式进行信息记录。包含的主要档案数据文件类型由 wav 文件、MP3 文件、AVI 文件、flv 文件等文件格式。

最后是其他类型文件，主要是指在档案工作当中可能涉及到的工作日志、电子邮件等文件，这类文件也属于非结构化档案数据。

（二）非结构化档案数据的主要特征

首先，非结构化档案数据特征的体量十分庞大。由于非结构化档案数据本身的类型格式相对灵活，应用范围也最为广泛。伴随计算机信息技术的发展和升级，所产生的非结构化档案数据以指数级进行增长。同时近年来档案管理工作持续推进数字化建设，导致档案数据总量增长迅猛，相应的需要更为庞大的存储空间以实现对于数据信息的存储支持。

其次，非结构化档案数据种类类型十分多样。从上述统计可见，非结构化档案数据有着多种不同的类型，所采用的文件数据编码格式也有着较大差别，同时近乎覆盖了日常计算机信息工作当中所有的文件类型，类型化特征鲜明。

三、非结构化档案数据管理面临的主要问题

（一）非结构化档案数据存储压力十分巨大

《全国档案行政管理部门和档案馆基本情况摘要》统计数据显示，节制 2017 年年底，国内共有示范数字档案共计十六个，国家级数字档案馆数量为二十七个。数字化档案建设成果中，馆藏电子档案总数量达到 162 万 GB，其中包含 84 万 GB 的文书档案，26 万 GB 图片信息数据档案，52 万 GB 的音像资源数据。而数字化存储的档案副本总量更是高大 1700 万 GB。如此庞大的非结构化档案数据体量，导致数据信息的存储压力十分巨大。大部分档案机构在档案数据信息存储上仅能够达到 TB 级别，现有材料介质对于如此庞大规模的数据存储在信息量存储服务中捉襟见肘。而随着后续数据量的持续增大，现有数据存储服务将更为窘迫。

（二）非结构化档案数据信息检索困难

在庞大的数据信息体量之下，非结构化档案数据还面临检索难题。非结构化档案数据检索问题表现在检索效率和检索安全性两个方面。

首先是检索效率方面。现代存储信息内容的检索主要依托信息数据的特征量来进行对比验证，从而完成数据库内部相关特征量数据信息的调取。但是非结构化档案数据本身的结构性特征不明确，同时很难进行有效的数据信息挖掘，导致检索过程相对复杂。在实际的检索中，数据库需要对已有数据存储信息进行相关特征量的遍历，导致检索过程缓慢，速度低下，效率不高，影响非结构化档案数据的正常使用。

其次是数据检索的安全性问题。非结构化档案数据具有极高的保存价值，同时大量的非结构化档案数据为绝密、机密的敏感内容，涉及到关键性信息或者隐私信息，需要对外进行保密。而在检索过程中，特征量信息检索事实上对已有的数据特征信息进行识别，因此很容易受到不同渠道的信息介入影响，导致

特征信息遭到窃取、复制甚至是篡改，最终导致检索环节安全性大打折扣，数据库的非结构化档案数据信息也将面临较大的安全威胁。

四、现代非结构化电子档案数据管理升级需要提升的能力

（一）需要提升数据档案的存储能力

保存存储是现代档案数据管理当中的基本能力，同时也是当前非结构化档案数据管理中面临的主要问题。在新兴技术的影响之下，国家正在积极推进数字档案的管理工作，但同时非结构化档案数据的增长趋势十分明显，现有的档案存储管理技术在不断升级过程中，仍然面临存储管理的困境。我国省级档案馆在多年档案存储方面，平均每年非结构化档案数据增长幅度约为 20%，部分地区数据档案信息增长量可以达到 30% 以上。非结构化档案基数巨大，同时增长速度极快，导致相应的档案存储压力也不断增大。针对非结构化电子档案数据所开展的管理优化，必须首先进行存储能力的提升，通过存储技术创新等手段来实现更高水平的档案管理能力。

（二）需要具备数据档案管理的环境适应能力

我国目前数据档案的管理模式主要采用分布式机构管理，档案数据存储于各行政级别的各类档案馆、档案室或者档案机构当中。其中档案馆是最为普遍且职能最为全面的档案管理机构，根据非结构化数据档案的基本类型，档案馆在档案管理中主要可以分为国家综合档案馆、国家专门档案馆等类型。档案馆下属设置档案室，常见档案室类型有机关档案室、科技档案室、声像档案室等，分别进行不同类别档案信息存储。在实际的档案管理运营当中，档案管理所处的环境，相关工作人员本身的信息素质，以及现代化信息管理当中所必须具备的信息技术手段，甚至在档案管理的财政预算等，都会对档案管理的实际效果产生影响。可以认为现代化非结构化档案除了是技术层面的管理创新之外，还需要依托技术优势，催动管理体制方面的升级创新，来实现高水准、高质量的管理服务。

（三）需要兼顾安全与管理成本

数据安全应当作为数据管理的最核心要义，非结构化档案数据存储与管理并不仅仅是完成数据信息的妥善保存，还需要具备档案服务眼光，认识到档案系统的本身的服务作用，才能够实现高质量的档案管理。其中档案信息的检索和信息的服务提供等，也是档案管理当中的重要环节，因此非结构化档案数据管理还会涉及到一定程度的数据安全问题。从非结构化数据特征来看，结构多样性、类型多边形导致各类信息编码方式和存储机制各有不同，导致在实际的安全管理中，需要投入更多的精力和管理成本，通过不断进行安全防御，来保证其安全性。对于管理来说，在安全管理之上，便会面临到成本问题。非结构化电子档案数据管理必须在保证安全基础之上，通过管理技术和管理机制的不断创新，来实现成本投入的优化，避免造成不必要的成本浪费。

五、强化非结构化档案数据管理质量的实现策略

（一）应用 OS 文件管理系统进行非结构化档案数据管理

OS 文件系统主要是指利用计算机操作系统的方式进行计算机内部文件的管理控制。伴随着系统运行的不断优化，非结构化档案数据管理可以借助用户接口和文件名重命名设置等方式，来进行整合性的文件数据信息管理。

目前 OS 文件系统管理主要面向个人用户提供小体量 的非结构化档案数据管理服务。其中主要的管理方法有资源管理器管理、虚拟文件夹管理等管理手段。其中资源管理器是 windows 系统自带的计算机辅助管理工具，在档案管理当中，工作人员可以直接通过关键性信息检索的方式，来获取相关非结构化档案数据信息结果；虚拟文件夹管理主要是通过针对非结构化档案数据进行虚拟文件夹命名的方式来进行管理。windows 系统中有关存储系统项的文件内容被称为虚拟文件夹，这类文件夹能够建立起不同类型的非结构化档案数据及索引关系，工作人员可以借助动态检索操作进行文件夹模拟合并，呈现相关非结构化数据信息。

（二）基于大数据的分布式文件系统应用

大数据技术背景下，建立对象存储模式，完成对于 SAN 存储和 NAS 存储优势的整合，来实现分布式存储和数据共享，成为解决非结构化档案数据存储问题，提升档案数据管理质量 的关键。目前主流技术领域较为知名的当属 open stack 所提供的对象存储服务模式，该服务模式将海量的非结构化数据信息以“对象”为基本单位，利用分布式系统进行数据信息存储。系统内部搭建多服务器文件分享的方式，实现数据库网络连接。在进行对象访问中，客户端不再对底层数据存储区块进行访问，而是直接与对象目标服务器建立联系，进而提高数据存储和数据检索调用能力。目前 open stack 在对象存储服务方面应用广泛，其中腾讯公司、中国移动、中国邮政储蓄银行等，都采用 open stack 对象存储服务，利用数百个服务节点搭建分布式的数据存储平台，来参与非结构化档案数据的存储、运营、调用、分析。近年来，智能档案技术发展，越来越多档案馆开始采用分布式档案数据对象存储服务进行智能化升级。

（三）erms 环境下的数据库管理

erms 是目前档案数据库管理中较常运用的数据存储环境。在实际管理当中，针对 erms 环境可以采用数据库挂接的方式，实现与非结构化档案数据的连接，提高处理速度的同时，能够在一定程度上压缩数据存储空间。部分管理机制中可以尝试引入 blob 嵌入式存储技术，来进行字段信息的数据内容集成，提高管理效率。

（四）非 erms 环境的管理

对于未采用 erms 标注环境的非结构化档案数据管理，可以重新进行管理方式的调整。例如在实际的档案管理当中，可以改变原有的数据文件存储格式，采用 DAT 文件格式进行存储。DAT 数据文件能够以较低的存储成本来完成对于多种类型任意内容的数据文件存储，并借助随机生成文件名称的方式，来提高存储和信息调用的效率，提高管理质量。

（五）NoSQL 数据库存储

针对海量非结构化数据和不断处于增量状态的信息数据，需要进行存储方式的创新。档案管理除了需要硬件数据库建设，来保证存储空间，还需要采用存储技术创新手段来推进管理创新。其中 NoSQL 数据库是一种关系型数据库，该数据可以飞恩威键值数据库、文档数据库、列族数据库和图数据库等不同的数据库类型，可以实现面向非结构化数据的针对性存储，提高存储和特征量检索效率。在当前数据增量提高中，具有较高的应用价值。

参考文献：

[1] 朴承哲 . 大规模非结构化数据的分布式存储方法 [J]. 太原师范学院学报 (自然科学版),2021,20(03):57-61.

[2] 李小飞 , 杨其 , 李媛 , 聂琪鹤 . 大数据与 AI 技术在智能辅助评标中的应用 [J]. 集成电路应用 ,2021,38(09):54-55.

[3] 赵顺存 . 三大融合引领下一代数据湖架构演进 [J]. 软件和集成电路 ,2021(08):48-49.

作者简介：

尹悦，1986.02，女，汉族，河南新乡人，新乡职业技术学院，本科，馆员，档案管理。