档案数字资源备份策略及可行性验证研究

刘菁[1]　姚丹超[2]

1.镇江市人力资源社会保障档案管理中心；2.镇江市人事考试考工中心

摘　要：大数据时代，档案电子化形成档案数字资源已成为主流发展趋势，随着档案管理信息化、数字化的推进和计算机技术的快速发展，加强档案数字资源的安全管理与备份存储成为档案管理工作的关注重点。本文将基于现有档案数字资源备份保存研究，提出了将数字化信息准确地记录于不同存储介质，采用高性能、高稳定、高可靠的分布式存储架构和新一代节能高效蓝光介质，结合磁、光、电不同存储介质特性的磁光电混合存储档案数字资源备份策略。磁光电混合存储是将磁、光和电子存储的优点集于一体的存储方式，数字档案资源长期保存应用具有较强的理论研究意义和实践指导价值。

关键词：资源备份；混合存储；长期保存

【DOI】10.12252/j.issn.2096-6288.2022.12.180

一、我国现有档案数字资源存储备份系统建设

现有传统集中式的以磁盘阵列为主体，磁带做备份的电子档案存储架构，具有磁存储介质损坏率高、系统可靠性低、总体能耗高、安全隐患高，随着电子档案增长存储架构日益复杂等问题，电子档案集中式的存储架构存在单点故障、扩容困难、安全性低、存储架构复杂的问题，设计一个横向扩展、自动均衡、自愈合的分布式存储系统，可以合理利用存储资源并且对数据进行有效的组织和管理，以满足日益增长的电子档案对存储系统越来越高的性能、容量、安全需求。档案数字资源存储备份系统主要需要实现性能扩展、大容量、自动均衡等功能：

性能扩展功能：档案数据粒度小但总体数据量大，且数据增量大，系统需支持性能扩展，以应对越来越耗性能的档案数据读写需求；大容量功能：档案数据存量大、增量也大，系统需要支持高达上千PB的存储容量；自动均衡功能：档案资源根据不同机构需要，存在不同的需求，系统应支持缩容扩容、自动均衡功能，保证不同机构需求。

1.存储介质分析选择。目前存储介质可分为含磁盘、磁带等的磁介质、光盘介质以及SSD、NVME等的电介质，存储介质的耐久性、可靠性、通用性、容量、速度和成本等指标不可能同时达到最优化，我们根据场景应用去关注存储介质的物理稳定性：档案数据重要性高，不宜使用对外界环境敏感的存储介质；存储介质的技术寿命：档案数据长期保存需采用存储寿命长的存储介质；离线存储能力：电子档案如存储配置在线存储系统，存储介质应可与存取设备分离，离线封存保管；数据防篡改能力：档案必须保证其内容的真实、可靠、完整、可用，不可篡改的一次写入的存储介质有更大的优势；存储介质使用成本：档案数字资源保存数据量大，保存时间长，我们在考虑存储介质使用成本时应充分考虑到初次采购的成本以及后期长期使用过程中的维护和升级成本。

2.存储备份策略设置。传统存储架构只采用单一存储介质无法满足“3-2-1”数据存储策略的问题，设计磁光电混合存储管理系统，通过统一存储管理平台，同时纳管固态硬盘、普通硬盘、磁带、蓝光盘等多种介质，一套系统实现档案数据资源多种形式、多种介质的存储备份策略设置。系统存储备份策略设置主要需要实现如下功能：档案数字资源长久保存过程中的自动存储备份功能、档案数字资源数据和各介质检测功能、存储系统数据智能恢复功能、存储系统多协议互通功能、预警故障追踪及处理功能等。

如何将档案资源长期保存，一直以来都是档案部门和档案工作关注的重点。随着信息技术的飞速发展，海量的电子文件和数字化的图像数据等成为当下档案工作的重点管理对象。数字档案资源作为国家、组织、个人具有重要价值的各类数字记录，研究面向档案数字资源备份策略及可行性验证，保障档案资源长期安全保存是十分重要的。

通过调研分析现有档案数字资源备份策略现状，研究结合磁、光、电不同存储介质特性的磁光电混合存储档案数字资源备份策略可在档案数字资源长久保存备份

工作中的有效应用，设计并实现磁光电混合存储系统，从而实现档案数字资源的自动异质备份。

二、国内外对档案数字资源备份的研究现状

国外关于数字资源如何长期保存方面的理论研究和实践探索已较成熟，包括国际标准组织（ISO）、美国等国际组织或国家相关研究机构、联盟，针对数字资源长期保存都制定了相应的标准[1]。中国国家科技图书文献中心的国家数字科技文献长期保存示范系统（National Digital Preservation Program，ND-PP）[2]项目，美国的“国家数字信息基础设施和保存计划”项目都已形成大量成果。

在国内，对电子文件及元数据等组件要素形成的归档信息包做到长期保存的研究也纳入数字资源长期保存研究范畴，当前学术研究集中于长期保存目标的研究。如张美芳提出真实性、永久性、有效性、安全性和通用性是电子文件长期保存的目标[3]；钱毅、刘力超提出了通过 ODBC 实现集中归档、采用 SIARD 等格式进行 XML 封装达到长期保存技术策略的研究；关于长期保存格式的研究[4]。钱毅提出电子档案长期保存格式的选择需考虑：符合既有需求或指南要求，环境适应性强，标准化基础高，且具备自描述、自包含与自校验等核心功能[5]。

总的来看，数字档案资源的长期保存与其他数字资源一样，也涉及政策法规、存储介质、存储技术、存储格式以及软硬件基础设施等方面。政策法规和存储格式是相对独立的，而存储介质则在一定程度上决定了需要相应配套的安全管理制度、基础设施建设需求，以及建设和运维的成本。

在实践中，档案部门更多以异地异质备份来保证数据可用。当前较为成熟的存储介质包括光介质（如DVD、CD、BD盘等）、磁介质（如机械硬盘、磁带等）、电介质（如 SSD、NVME）和缩微胶片等。不同的存储介质在总体使用成本、防篡改、容量、寿命和读取速度等性能指标上各有特点，但整体看来蓝光光盘是最适合作为不经常访问的数字档案资源长期存储的介质。

三、档案数字资源备份采用的研究方法

针对档案数字资源备份在调研收集了有关档案数字资源的备份策略及相关保存技术的资料的基础上，研究了磁存储、光存储和电存储等存储技术的优缺点和适用范围，梳理了数字档案资源长期保存介质的相关标准、实践应用、性能需求，分析了各存储厂商研发能力及实践成果。调研发现磁光电混合存储在各行业已开始得到应用。相关企业打造新一代节能高效蓝光及光磁电一体化智能存储应用系统已在全国33个城市数据湖进行了部署与使用。存储协议方面，国内外主要存储厂商基本兼容AWS提供的标准对象存储S3协议，磁光电混合存储系统可设计支持S3、NFS、CIFS等存储协议，通过标准协议对接主要存储设备。在档案数字资源备份策略方面，根据已有相应的分级存储技术设计系统，实现数据热度智能分类和更新，数据自动检测与迁移。

在此基础上，下一步，针对档案数字资源长久保存、备份策略的需经全面的分析，将从三个方面进行研究：

1.存储介质和存储设备的分析和选择。目前常见的主要存储介质分为磁（磁盘、磁 带）、光（光盘）、电（SSD固态硬盘）等类型，通过介质的稳定、耐久、可靠、通用、容量、速度和成本等基本指标进行对比选择。通常一种介质的指标不可能同时达到最优化，需要根据应用场景找到最合适的平衡点。

档案数据非常重要，要求保存周期长、数据量大、真实和完整性要求高、不可丢失篡改数据。档案数字资源导入、检测、检索等数据的频繁读写，对存储设备和存储介质的响应及读写速度有较高的要求。研究市场上的存储介质，发现蓝光介质具备长寿命、稳定性好、一次写入不可篡改等优点，但蓝光介质访问速度慢、介质与驱动器分离的问题给数据恢复、数据管理带来较多问题。具有较高读写速度的高速磁盘或固态硬盘作为存储介质更适合频繁读写、响应读写速度高的场景。

2.存储系统架构设计。档案数字资源备份策略需要一套存储系统实现，经过需求分析，系统硬件部分需要集高性能固态硬盘、大容量磁盘阵列、高密度光盘库、磁带库于一身，充分发挥磁光电存储介质的各自优势。系统软件需将固态硬盘、机械硬盘、光盘库、磁带库等硬件资源进行整合成一个瘦供给的资源池，并以文件存

储、对象存储、Restful API等多种接口方式提供存储服务，所以系统的总体架构应该是横向扩展、自动均衡、自愈合的分布式存储系统，系统能横向扩展有效提升系统IOPS性能和存储容量。

3.档案数字资源备份策略设计。磁光电混合存储管理系统可通过统一存储管理平台，同时纳管固态硬盘、普通硬盘、磁带、蓝光盘等多种介质，一套系统实现档案数据资源多种形式、多种介质的存储备份策略设置。数字档案资源可分为热数据、温数据和冷数据，且热、温、冷数据的划分会随着利用需求的变化而变化，系统策略可实现热、温、冷数据三者之间的智能化迁移。目前项目采用将数据存放在不同级别的存储设备的分级存储技术：光盘库存放冷数据，磁盘或磁盘阵列存储温数据，固态硬盘和内存保存热数据。实现分级存储涉及数据热度分类和更新数据检测策略、数据迁移策略等策略的设计。

随着市场的发展，磁存储、光存储和电存储等存储介质都有了较大的发展，分布式存储系统随着发展越来越成熟。磁光电混合存储以分布式存储架构为基础，集磁、光、电存储介质优点于一体，在数字档案资源长期保存及备份的实践应用中具有较大的实践指导价值。磁光电混合存储安全、海量、高效、绿色的优点，使得其逐渐开始得到应用。根据需求，国内部分企业开始开发存储容量大、扩展灵活、安全性高、保存时间长、能耗低的城市数据湖，能够达到绿色节能环保、总体拥有成本低、存储寿命达100年以上、数据安全可靠并能支持多种备份模式等多方面的要求。

四、研究中遇到的问题及解决办法

1.不同存储设备统一管理问题。在磁存储、光存储和电存储等存储介质及设备厂商，采用各自的存储标准和系统，存在不同存储设备统一管理问题。国际上，各国政府或档案部门主要通过制定相应指南来规范数字档案存储介质的选择。可根据国家相关要求，以数字档案资源建设为基础，采用标准化和封装技术策略实现存储设备统一管理。

2.存储协议、接口对接匹配问题。磁光电混合存储采用分布式架构，包含NVME、SSD、HDD、磁带、蓝光等不同介质，不同介质设备及档案管理系统采用协议、接口不同，系统需尽可能支持较多的存储协议及接口。

国内主要存储厂商有华为、阿里、浪潮、华三、xsky等，国外主要存储厂商有EMC、AWS等，AWS提供的标准对象存储S3协议为业内标准，各大厂商基本都兼容这套标准协议，可通过软件开发，系统开发对接市场主流协议及接口。

3.档案数字资源备份策略设计问题。数字档案资源可分为热数据、温数据和冷数据，且热、温、冷数据的划分会随着利用需求的变化而变化，系统策略需实现热、温、冷数据三者之间的智能化迁移。根据档案数字资源需要，将数据存放在不同级别的存储设备进行策略设计，通过系统应自动实现分级存储及档案备份策略。

4.存储设备故障检测问题。存储设备多，架构复杂，需要借助一个自动化、智能化的统一运维平台，将数据进行数据化、可视化，帮助档案管理员完成对存储设备的统一运维管理与监控。磁光电混合存储系统以保障数字档案资源真实、完整、可用、安全为出发点，设置档案数字资源保存备份机制、数据检测机制和载体检测机制。

参考文献

[1]毛义春，美国数字资源长期保存的研究进展及经验借鉴，北京档案，2009.7

[2]张晓林、吴振新等，国家数字科技文献资源长期保存体系的战略与实践，图书馆杂志，2017.12

[3]蒋术、吴明霞，电子文件长期保存回顾与展望，档案天地，2018.9

[4]钱毅、刘力超，数据库电子文件归档与长期保存技术路径研究，档案学研究，2017.4

[5]钱毅，数据态环境中数字档案对象保存问题与策略，研究员成果，2021.6

作者简介：刘菁，年龄：47，出生年月：1975年4月12日，性别：女，学历：本科，职称：副研究馆员，专业：档案，籍贯：镇江。

姚丹超，年龄：46，出生年月：1978年7月22日，性别：女，学历：本科，职称：馆员，专业：档案，籍贯：丹阳。