 DOI: https://doi.org/10.12345/daglyqyfzyj.v3i7.30312

Practice and thinking of science and technology information resource management platform

Jianqin Yang

CNOOC Research Institute Co., Ltd., Beijing, 100028, China

Abstract

To address the issues of poor correspondence between paper-based and electronic data in offline scientific archives, weak inter-data correlation, and severe information silos, a unified digital scientific archive management system has been developed. This platform enables one-stop access to comprehensive information covering archive collection, storage, management, and utilization. Through practical project implementation, the effectiveness and significance of this archival management platform have been validated. The unified management of archival data enhances data interoperability, while the establishment of standardized management systems strengthens operational efficiency and reduces redundant infrastructure development.

Keywords

digital science and technology archives, management platform, graphic recognition, full-text retrieval

科技信息资源管理平台实践与思考

杨建钦

中海油研究总院有限责任公司，中国·北京 100028

摘　要

针对线下科技档案部分纸质档案资料与电子数据缺乏对应关系，资料间关联程度差，信息孤岛问题严重等问题。建设了一个统一的数字化科技档案管理系统，能够一站式获取多方面的档案采集、存储、管理、利用等信息。最后通过项目实践验证了档案管理平台应用效果及意义。通过档案数据的统一管理，提升了档案数据的共享性；通过档案管理规范体系建设，加强了档案业务管理和使用，减少重复建设。

关键词

数字化；科技档案；管理平台；图文识别；全文检索

1 引言

随着新时代下档案及资料管理信息系统的迅速开展，档案线上各种信息资源管理系统分散，线下部分纸质档案资料又与电子数据缺乏对应关系，导致资料间关联程度差，信息孤岛问题严重。科研人员迫切需要一个统一的档案信息资源平台，能够一站式获取多方面的档案资源信息 [1]。

（1）在科研档案工作中，对于文件在线浏览，线上借阅流程的需求十分强烈，花费在资料的查找和搜集的时间占比高，当前已归档的综合类科技档案资料缺乏对归档项目名称及编号的管理，致使科研人员无法通过项目名称去查找资料，需要花费大量的时间通过资料内容去检索筛选；线上电子资料均只进行条目管理，缺乏全文检索及在线 浏览功能，容易导致借阅资料非所需资料，事倍功半，档案资料整编也需要一套档案全面整理利用的标准工作平台。

【作者简介】杨建钦（1982–），男，中国山东郓城人，硕士，高级工程师，从事GIS，信息系统集成，档案管理研究。

（2）针对上述问题，科技信息资源管理平台实现了档案资料从接收开始的全生命周期管理。其中整编工作平台的建立，降低了整编工作人员的工作量及劳动强度，在线归档和在线借阅功能则实现了档案资料从源头的获取，既降低了整编工作人员的工作量，也通过与其它多个应用系统的对接，充分利用了已建设资源，打破了数据壁垒、信息孤岛，成为了智能化油田建设的重要组成部分 [2]。

2 科技信息资源管理平台建设研究思路

科技信息资源管理平台建设总体思路是设计一套能够管理总院各类信息资源的系统，实现各类档案资料（包括科技、专项、汇交、图书、期刊和标准）的管理与利用、库房的在线监控，以及与其他系统之间的数据交换接口设计，为科研人员提供一站式的获取多方面信息资源的渠道。具体如下：

（1）经过对总院原有档案与资料管理系统的功能分析及现有需求调研，分析新增的业务数据内容，设计完善系统的数据结构，完成相关功能优化，使其功能更加完善，同时

完成新功能设计并完成新功能的开发；

（2）深入分析数据库现有的数据管理模块，完善其功能，让用户更加快捷的加载、查找与对比资料；

（3）本次科技信息资源管理平台建设，统一线上管理专项档案、汇交资料数据。并将原档案管理系统、标准管理系统、图书管理系统及原科技信息门户系统中所管理的科技档案资料、标准、图书等数据按新制定的档案资料管理规范迁移至新建的信息数据资源库统一管理。该系统升级完成后会与勘探成果数据管理平台以及电子数据在线管理系统形成数据共享。

3 科技信息资源管理平台解决方案与关键技术难点

采用三层架构体系，包括数据访问层、业务逻辑层和业务表现层，其中整编工作平台运行模式和开发架构采用 C/S 模式，在线应用平台运行模式和开发架构采用 B/S 模式。基于 ES 的全文检索服务平台和图文识别及关键词扫描工具作为业务系统的科研创新成果，运行模式和开发架构采用 B/S 模式 [3][4]。

数据存储代表系统运行所需要的数据库服务器，数据访问层表示具体与数据库进行交互的部分，业务逻辑层定义了档案与资料管理系统的各类业务规则、处理流程、管理对象以及抽象出的各个业务相关的类和接口，业务表现层调用数据访问层、业务逻辑层的相关接口。业务表现层是用户和系统之间交流的桥梁，它一方面为用户提供了交互的工具，另一方面也为显示和提交数据实现了一定的逻辑，以便协调用户和系统的操作。档案与资料管理系统建设课题结合现场应用的实际情况，经过项目组的研究讨论，确定了双数据库的设计思路：工作库和归档库。

工作库：归档、整编工作数据库，主要由整编工作平台使用。归档、整编的著录项数据优先进入工作库，在工作库内完成增加、删除和修改操作；著录项数据以及对应的电子文档数据在入库前完成关键词的过滤筛选工作，并将扫描结果反馈至整编工作平台。

归档库：系统主数据库，主要由在线应用平台使用。受归档、整编过程的影响，归档、整编的档案资料在完成著录项信息的入库后，还需要标签打印、档案装盒、库房存放等一系列后续过程。因此，项目组确定了 14 天的数据同步时间间隔期。

工作库的数据结构与归档库的相关数据结构保持一致，方便数据的同步更新工作。归档库作为系统的主数据库，其关联业务更多，数据库结构更复杂。双库的同步策略：定时自动同步策略。工作库入库 14 天之后，由工作库同步数据至归档库。其中，自动定时机制由数据库的定时任务实现，数据的同步策略则由存储过程实现。档案与资料的修改和删除，由整编工作平台实现。由于工作库与归档库存在 14 天的缓冲周期，14 天之内的工作库的数据修改、删除不受影响。14 天之后的数据的删除和更新，则增加了数据的更新、删除记录，并由系统的同步更新服务程序在第一时间完成归档库的同步更新和删除。

系统提供了一套完整的权限管理机制，结合研究中心的 Share 库，实现了组织机构、用户、资源、角色、密级等的关联关系。系统的组织结构、用户和岗位信息采用总院的 Share 库，用户 AD 登录时在写入系统自身用户表 [U_USER]，通过角色表 [U_ROLE]、资源表 [U_RESOURCE]、密级表 [U_SecretClass] 之间的关联关系，实现用户挂角色、角色挂资源、用户挂密级、角色挂密级，最终实现用户的权限控制。

用户表的属性信息包括用户 ID、用户登录名、用户密码、用户真实姓名、电话、邮箱、AD 域账户、有效标识、描述、IP 地址、用户编号、DN、组织结构 ID 和岗位 ID。

角色表的属性信息包括角色 ID、角色名称、角色描述和角色类型。资源表的属性信息包括资源 ID、资源名称、资源代码、父资源 ID、资源层级、排列序号、资源路径、资源类型和资源描述。

4 科技信息资源管理平台关键技术实现与核心难点解析

密级表的属性信息包括 ID、密级名称、密级级别和排序。用户角色表的属性信息包括 ID、用户 ID 和角色 ID。角色资源表的属性信息包括 ID、角色 ID 和资源 ID。密级角色表的属性信息包括 ID、名称、密级 ID、档案类别和描述。主要关键技术难点如下：

4.1 ChineseOCR 图文识别组件

档案资料整编过程中，上传的扫描图件采用 ChineseOCR 进行图文识别，结合关键词 DFA 扫描算法，实现上传文档关键词汇的筛选功能 [5][6]。目前支持的文档类型包括文本、Office 常用文档，图片、PDF 等。关键词汇的过滤筛查时，使用 ChineseOCR 组件作为后台图像识别引擎提取图片中的文字信息，通过 DFA 扫描算法与关键词词库进行比对，从而发现某些特定的关键词。

ChineseOCR 是基于 yolo3/darknet 和 crnn（卷积循环神经网络）实现的中文自然场景文字检测与识别工具，支持多种深度学习框架，拥有良好的识别模型，也支持使用自己训练的模型文件。经测试图文识别组件的图片文字识别率可以达到 95% 以上，清晰排版工整的图片识别率可以达到 99% 及以上。

4.2 Docker 技术

Docker 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到任何流行的 Linux 机器或 Windows 机器上。

基于 ES 的全文检索服务平台使用 Docker 技术，预装

Elasticsearch、全文检索服务接口、Java 环境、依赖组件等，预装后打包为 Docker 镜像，以便于快速化部署。一台宿主服务器可部署多个 Docker 镜像，达到部署多套 ES 或者 ES 集群的目的 [7]。

4.3 PDF 在线预览组件

PDF.js 主要用于在 HTML5 平台展示 PDF 文档，是一款开源的 PDF 文档读取解析插件，无需任何本地技术支持，支持所有的主流浏览器，提供 PDF 文档的预加载，PDF 安全水印等功能 [8] 。

4.4 RestFul 接口

在线应用平台系统的基础框架采用 VSEAF4.3，利用 VSEAF4.3 快速构建系统的核心层和框架层，基于 Maven 仓库和插件化开发模式，将各个功能模块封装成相对独立的插件，便于开发、调试和部署。在线应用平台的后台功能开发，都是基于 MVC 架构开发模式实现的业务功能；前台采用 LayUI 框架，界面开发高效、快捷。建设过程中提供的所有的 API 接口皆遵循 RestFul 标准规则。RestFul 是一种网络应用程序的设计风格和开发方式，基于 Http，可以使用 Xml 格式定义或 Json 格式定义。

5 档案资料在线归档与整编流程优化方案

针对研究中心档案资料的归档过程，在线应用平台严格遵循研究中心的管理办法规定，结合研究中心档案与资料的归档的业务现状，实现了科技档案、合同档案和汇交资料的在线归档流程。在线归档流程建设过程中充分利用总院已建设资源，通过与总公司合同管理系统项目组、总院科技管理平台的相关负责人进行了数据接口的对接，以实现档案资料的便捷归档。档案资料在线归档后，整编人员可在整编工作平台直接提取归档数据，减少人工录入工作，降低整编难度，提高工作效率。在线应用平台进行档案与资料的在线归档时，依据研究中心的归档管理办法加入了流程审核功能，集成了研究中心统一的协同工作流程引擎。

档案与资料整编工作平台的构建，实现了档案资料从接收开始的全生命周期管理。平台构建过程中，实时对接总院档案资料整编人员的业务需求，以降低工作人员的工作量及劳动强度为主旨，以确保平台落地。通过平台的试运行对比，可有效提高整编效率 30% 以上，受到了整编人员的一致好评。

档案与资料在线应用平台的在线归档，配合档案与资料整编工作平台的整编提取，通过与其它多个应用系统的对接，充分利用总院已建设资源，打破数据壁垒、信息孤岛，实现了跨系统间的数据信息共享。

6 结语

该系统主要针对当前档案与资料管理过程中存在的问题，通过对总院各类信息资源实行“线上、统一、规范”的管理，实现“资源管理规范化、资源利用关联化、资源服务个性化”的目标。在实际工作中取得了很好的应用效果。

构建科技信息资源管理平台，实现档案资料从接收开始的全生命周期的管理；

实现馆藏档案与资料的全文检索与利用，电子资料在线全文阅览，档案资料库房在线监控，总院托管国家地质汇交资料在线管理，馆藏数据资源报表统计，电子类数据资源借阅及审批，完成设计与其它平台数据接口规范 [9]。

下一步工作中，保持研究总院科技信息资源管理平台建设时明确的“线上、统一、规范”管理的总体思路，逐步实现“资源管理规范化、资源利用关联化、资源服务个性化”的建设目标，为将来实现档案从勘探开发一体化协同研究平台开始的一键归档打下坚实基础，对系统平台进行功能优化、拓展和运维，提供 GIS 数据的采集及对外提供数据服务的规范，促进数据的共享与统一管理，最大化发挥平台效用。

参考文献

[1] 安家宏.事业单位档案信息化建设的影响因素及 解决策略探讨 [J] .文化产业，2021（21）：80—81.

[2] 高娟.事业单位档案管理信息化建设中的问题与 解决路径分析 [J] .劳动保障世界，2018（1）：56.

[3] 仝宁宁 . 近年我国智慧档案馆的研究现状 [J]. 档案天地 ,2019(12).

[4] 陈勇，廖琼，崔葳 . 智慧城市背景下我国智慧档案馆建设研究 [J]. 中国档案研究 ,2018(01).

[5] 许燕梅.事业单位档案信息化建设工作探讨 [J] . 档案与建设，2019（12）：56—58.

[6] 谭德新，李秀云，李莉.事业单位档案信息资源 共享问题及措施 [J] .办公室业务，2021（8）：103—104.

[7] 任杰.事业单位档案信息资源共享问题及处理办法 [J] .城建档案，2021（10）：135—136.

[8] 黄林英，傅荣校.智慧图书馆与智慧档案馆对比分析[J].档案与建设，2016(11)：26-29.

[9] 容依媚，肖秋会.我国智慧档案馆研究综述[J].北京档案，2019(8)：15-18.