ICS

T/GXDSL

团　　体　　标　　准

T/GXDSL —2026

档案智能分类著录管理导则

Guidelines for Intelligent Classification and Cataloging Management of Archives

（工作组讨论稿）

（本草案完成时间：2026 - 6 - 12 ）

2026 - - 发布　　2026 - - 实施

广西电子商务企业联合会　发布

目　　次

前　言

本文件依据GB/T 1.1-2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西产学研科学研究院提出。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

档案智能分类著录管理导则

1 引言

为深入贯彻落实国家数字经济发展战略、智慧档案建设总体部署，严格对标《“十四五”全国档案事业发展规划》中“加快档案工作数字化转型、智能化升级”“推动人工智能技术深度赋能档案开放鉴定、智能检索、精准利用等核心业务”的工作要求，破解全国档案管理领域普遍存在的人工分类效率偏低、著录标引标准不统一、元数据采集碎片化、资源挖掘深度不足等行业痛点。依托自然语言处理、机器学习、知识图谱等新一代人工智能技术，构建标准化、规范化、智能化的档案分类著录技术体系与管理机制。是推动我国档案工作从数字化向知识化、智慧化转型升级的关键支撑，能够有效提升全国档案资源集约化管理、精细化治理、精准化服务能力，助力档案资源价值深度挖掘、活化利用，夯实新时代国家档案资源治理体系和治理能力现代化建设基础，为数字中国、智慧政务建设提供坚实的档案数据支撑。

2 范围

档案智能分类著录工作的总体原则、核心要求、智能著录模型、智能分类体系构建、著录数据项规范、质量管控体系、系统功能配置及数据安全防护等核心内容。适用于全国各级综合档案馆、专业档案馆、党政机关档案室、企事业单位档案部门，基于人工智能技术开展的档案智能分类标引、自动化著录、元数据治理、档案信息智能化管理等业务工作；同时适用于全国档案智能管理系统、档案 AI 应用平台的研发、部署、运维与评测，可作为全国档案行业智能化建设的统一技术依据和管理标准。

3 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件；凡是不注日期的引用文件，其最新有效版本（包含所有修改单、增补公告）适用于本文件。

GB/T 3792.5-1985 档案著录规则（废止，仅作为历史参考依据）

GB/T 4880.1-2005 语种名称代码第 1 部分：2 字母代码

GB/T 7156-2003 文献保密等级代码与标识

GB/T 9704-2012 党政机关公文格式

GB/T 15418-2009 档案分类标引规则

GB/T 18894-2016 电子文件归档与电子档案管理规范

GB/T 26163.1-2010 信息与文献 文件管理过程文件元数据 第 1 部分：原则

GB/T 29194-2012 电子文件管理系统通用功能要求

GB/T 33190-2016 电子文件存储与交换格式 版式文档

GB/T 39362-2020 党政机关电子公文归档规范

GB/T 42727-2023 政务服务事项电子文件归档规范

DA/T 13-2022 档号编制规则

DA/T 15-2020 磁性载体档案管理与保护规范

DA/T 18-2022 档案著录规则

DA/T 31-2017 纸质档案数字化规范

DA/T 38-2021 档案级可录类光盘 CD-R、DVD-R、DVD+R 技术要求和应用规范

DA/T 46-2023 文书类电子档案元数据方案

DA/T 58-2019 电子档案管理基本术语

DA/T 68-2017 档案服务外包工作规范

DA/T 78-2019 录音录像档案管理规范

DA/T 89-2022 实物档案数字化规范

4 术语和定义

DA/T 18-2022、DA/T 58-2019 界定的术语和定义适用于本文件，下列新增术语和定义同样适用。

4.1 智能分类

依托机器学习、深度学习等人工智能核心技术，对档案全文内容、元数据信息、格式特征等多维度数据进行智能特征提取、语义分析与关联研判，自动或辅助人工完成档案类目匹配、归类划分的智能化档案管理核心业务过程，是档案智慧治理的基础核心环节。

4.2 智能著录

融合光学字符识别、自然语言处理、知识图谱、命名实体识别等智能技术，自动完成档案著录信息的捕获、提取、校验、补全与规范化整改，生成符合国家行业标准的结构化档案目录数据与元数据体系的自动化作业过程。

4.3 置信度

档案智能分类、自动著录模型对输出结果的准确性、合规性进行量化评估的概率数值，以 0%-100%百分数形式呈现，是判定智能结果是否合规、是否需要人工干预的核心量化指标。

4.4 人机协同

适配档案智能化管理全流程的新型作业模式，依托人工智能系统完成批量、标准化基础作业，依靠档案管理人员开展精准复核、异常处置、规则优化、模型迭代，通过人机互补、双向校验、持续优化，实现档案分类著录高效化、精准化、规范化的协同工作机制。

5 缩略语

本文件适用缩略语释义如下：

AI：人工智能（Artificial Intelligence）

NLP：自然语言处理（Natural Language Processing）

OCR：光学字符识别（Optical Character Recognition）

XML：可扩展置标语言（eXtensible Markup Language）

ASR：自动语音识别（Automatic Speech Recognition）

6 总体原则与要求

档案智能分类著录工作立足国家档案事业高质量发展大局，坚守标准化、智能化、安全化、长效化发展底线，遵循以下核心原则，全面适配全国档案智慧化建设统一要求。

6.1 标准化兼容原则

智能分类著录全体系、全流程、全数据需严格对标国家及行业现行档案标准，全面兼容 DA/T 18-2022等核心规范，确保全国跨地区、跨部门、跨层级档案数据互通、资源共享、系统互联，保障全国档案智慧化建设的统一性、规范性和通用性。

6.2 多维全域著录原则

严格遵循国家档案多级著录管理体系，全覆盖支撑全宗级、类别级、案卷级、文件级四级智能著录，实现不同层级档案数据的关联匹配、统一治理，满足国家档案精细化、层级化、全域化管理需求。

6.3 人机协同精准原则

构建“机器批量预处理、人工精准复核、异常专项干预”的标准化作业体系，明确智能风控阈值，系统输出分类、著录结果置信度低于 85%的，必须强制触发人工审核流程，杜绝智能化作业系统性偏差，保障档案数据精准度符合国家级归档标准。

6.4 动态迭代优化原则

建立全国统一的模型长效优化机制，智能算法模型具备增量学习、自主迭代能力，依托人工复核、整改优化的真实业务数据，按季度、半年度开展模型迭代升级，持续提升全国档案智能分类著录的适配性、精准性与通用性。

6.5 安全合规底线原则

严格落实国家网络安全、数据安全、保密管理相关法律法规及标准要求，对涉及国家秘密、工作秘密、商业秘密、个人隐私的档案数据，必须先脱敏、后处理；智能运算、数据存储、传输交互全过程置于安全隔离环境，坚守档案数据安全底线，筑牢国家档案资源安全屏障。

7 智能分类体系构建

立足全国档案分类标准化统一要求，构建“标准规则+海量样本+智能模型+流程管控”的一体化智能分类体系，适配全国各级各类档案机构智能化分类作业需求。

7.1 分类规则体系构建

7.1.1 基础标注数据集：以《中国档案分类法》（第 4 版）及国家、行业专项分类标准为核心依据，搭建国家级标准化档案智能分类标注数据集。训练样本总量不少于 10000 条，涵盖精准标注的档案目录数据及全文文本，覆盖各机构核心职能活动类别占比不低于 95%，保障模型适配全国党政机关、企事业单位各类档案分类场景。

7.1.2 行业特征词库：结合全国各行业、各层级机构职能特点、机构沿革、公文规范、业务特性，构建通用性与专项性相结合的档案智能分类特征词库，词向量维度统一设置为 300 维。词库涵盖机构名称沿革、公文字号规范、行业专有术语、业务核心关键词等专项特征，全面提升跨领域、跨行业档案分类的精准适配能力。

7.2 模型选型与性能标准

7.2.1 模型选型要求：优先选用 BERT、RoBERTa 等成熟通用预训练语言模型及其轻量化、国产化优化变体作为基础分类器，兼顾运算精度与运行效率，适配全国不同层级、不同算力条件的档案机构部署需求。

7.2.2 核心性能指标：智能分类模型测试集宏平均 F1 值不低于 92%，单条档案智能分类响应时长控制在 2 秒以内，确保全国规模化、批量式档案分类作业的高效、精准开展。

7.3 标准化分类流程

7.3.1 文本预处理：对电子档案原文、纸质档案 OCR 识别文本进行标准化预处理，完成文本去噪、精准分词、停用词剔除、格式统一等操作，为智能分类研判提供标准化数据基础。

7.3.2 智能自动归类：系统依托档案题名、发文字号、核心关键词、全文首段核心语义等多维度信息，对照 GB/T 15418-2009 标准分类号体系，自动测算档案归属各类目的匹配概率，完成初步智能归类。

7.3.3 规则优先修正：坚持“规则优先、AI 辅助”的核心逻辑，对于存在历史合规档号、既定分类号、固定归档规则（年度-机构-问题）的档案，优先执行标准化规则匹配，人工智能预测结果作为补充校验，保障分类结果符合国家档案归档规范。

8 智能著录管理规范

严格对标国家档案著录、元数据管理核心标准，构建全覆盖、标准化、结构化、深度化的档案智能著录体系，实现档案基础信息、深层信息的全自动、规范化采集与治理。

8.1 核心必备著录项

依据 DA/T 18-2022、DA/T 46-2023 国家标准，智能著录需全覆盖档案核心必备字段，实现精准自动提取、规范整改，核心项目如下：

8.1.1 题名：依托 OCR 识别与 NLP 语义分析技术，自动提取档案标准题名；针对无题名、题名不规范的档案，自动提炼核心关键词，生成符合党政机关公文规范的标准化题名。

8.1.2 责任者：通过公文署名、印章识别、电子公文元数据抓取等多渠道自动捕获，统一规范为机构标准全称或国家认可的通用规范简称，实现全国责任者名称标准化统一。

8.1.3 日期：智能识别档案成文日期、签批日期、生效日期等核心时间信息，统一规范化为 YYYYMMDD 标准格式，保障全国档案时间数据口径统一。

8.1.4 文号：通过正则匹配、语义识别自动抓取公文标准发文字号，精准匹配国家公文字号规范格式，实现文号标准化提取与校验。

8.1.5 保管期限：结合档案智能分类结果、全文核心关键词，自动匹配国家档案保管期限标准，完成保管期限智能判定与著录。

8.1.6 关键词/主题词：采用 TF-IDF 算法结合语义权重分析，自动抽取 3-8 个核心关键词，形成标准化主题词体系，适配全国档案检索、统计、利用需求。

8.2 全文深度智能著录

8.2.1 全文关联著录：所有数字化档案副本需建立标准化全文数据库，实现档案元数据与全文内容双向关联，支持上下文精准定位、全文溯源式智能著录，提升档案信息挖掘深度。

8.2.2 命名实体智能识别：系统具备高精度命名实体识别能力，自动抓取档案内人名、地名、组织机构名、时间、数量、事件等核心实体信息，其中地名识别准确率不低于 95%，人名识别准确率不低于 90%，组织机构名识别准确率不低于 85%，所有实体信息统一纳入档案元数据体系闭环管理。

8.3 结构化数据管理

8.3.1 标准化数据输出：智能著录成果自动生成符合国家通用标准的 XML、JSON 结构化数据，可无缝对接全国各级各类档案管理系统，实现数据互通、批量导入、统一治理。

8.3.2 声像档案专项著录：依据 DA/T 78-2019 标准，针对录音、录像等声像档案，配备高精度 ASR 语音转文字功能，自动完成音频、视频内容文字转化、整理与结构化著录，补齐声像档案智能化治理短板。

9 质量控制与评价体系

立足全国档案数据质量统一管控要求，构建量化、可考核、可迭代的智能分类著录质量管控体系，建立全流程审核、全周期反馈、常态化优化机制，保障全国档案智能著录数据精准、规范、统一。

9.1 核心质量量化指标

9.1.1 分类准确率：定义为智能自动分类后、经抽检核验无需人工修改的合规档案占比，全国统一目标值≥95%，抽检覆盖率不低于 5%，保障档案分类整体精准度。

9.1.2 著录项填充率：题名、责任者、日期、文号四大核心必备著录项非空占比，全国统一目标值为 100%，实现核心档案信息完整归集。

9.1.3 著录一致性：同一全宗、同一批次档案的责任者名称、文号格式、时间格式、关键词规范等内容的标准化统一程度，全国统一目标值≥99%，杜绝同类数据格式混乱、标准不一问题。

9.2 全流程审核与迭代机制

9.2.1 三级审核机制：建立全国统一的“AI 一审、人工二审、管理三审”三级审核体系。一审由人工智能系统完成自动校验、初步审核；二审由档案业务人员对低置信度、高风险、特殊类别档案开展全面人工复核；三审由档案管理部门负责人开展常态化抽检，抽检比例不低于 1%，实现全流程质量闭环管控。

9.2.2 模型迭代反馈机制：搭建全国档案智能著录反馈数据集市，将所有人工复核、整改优化的业务数据纳入样本库，按月度、季度开展模型增量训练与迭代升级，持续优化全国智能分类著录模型适配性与精准度。

10 系统功能与性能要求

面向全国档案智慧化建设通用需求，明确智能分类著录系统架构、性能、接口标准，保障系统通用性、扩展性、稳定性，适配国家档案数字化规模化建设要求。

10.1 系统架构要求

系统采用轻量化、高适配、可扩展的微服务架构，实现 OCR 识别、NLP 语义处理、智能分类模型、档案业务管理等模块解耦，支持 GPU 集群并行加速运算，可适配国家、省、市、县四级档案机构分级部署、互联互通。

10.2 核心性能指标

10.2.1 单条处理性能：在标准服务器配置（不少于 2 颗 Intel Xeon Gold 6248R CPU、4 张 NVIDIA Tesla T4 及以上 AI 加速卡、256GB 内存）环境下，A4 幅面 300DPI 纸质档案图像 OCR 识别、智能分类、全项著录全流程平均处理时长不超过 3 秒。

10.2.2 批量处理能力：系统支持高并发批量处理，8 小时标准工作时长内，档案处理能力不低于 20000 件或 50000 页，满足全国档案数字化批量攻坚、常态化治理的作业需求。

10.3 数据接口规范

10.3.1 通用数据接口：系统配置标准化 RESTful API 接口，全面兼容 OFD、PDF、JPEG、TIFF 等国家通用档案格式，支持多格式档案数据的接入、处理、导出与共享。

10.3.2 业务系统对接：严格遵循 GB/T 29194-2012 标准，可与全国各级电子文件长期保存系统、数字档案室、智慧档案管理平台无缝对接，实现全国档案业务一体化、智能化、闭环化管理。

11 数据安全与隐私保护

严格落实《中华人民共和国数据安全法》《中华人民共和国档案法》《中华人民共和国个人信息保护法》等法律法规要求，建立全维度、全流程、可溯源的档案智能处理安全防护体系，守护国家档案数据资源安全。

11.1 加密安全管控

档案数据智能处理、传输、存储全流程采用国家商用密码 SM2、SM4 算法进行加密防护，杜绝数据泄露、篡改、丢失风险，保障档案数据全生命周期安全。

11.2 外部服务安全规范

如需调用公有云 AI 服务、第三方智能接口开展档案处理，必须签订专项数据保密协议，明确第三方不得将档案数据用于模型训练、商业应用、对外共享，严格界定数据使用边界，压实安全责任。

11.3 模型风险应急处置

建立国家级档案 AI 模型风险应急管理机制，实时监测模型运行精度与稳定性。当系统检测到模型分类著录准确率连续 3 天低于 85%，或出现全域系统性分类偏差、批量数据错误等问题时，立即关停自动入库功能，自动回退至人工著录模式，同步启动模型紧急核查、修复与复测工作，防范规模化档案数据质量风险与安全风险。