ICS 01.140.20
A 14

DA

中华人民共和国档案行业标准

DA/T 77—2019

纸质档案数字复制件光学字符识别(OCR)工作规范

Specification for optical character recognition (OCR) of digital copies of paper-based records

2019-12-16 发布　　2020-05-01 实施

国家档案局　发布

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由国家档案局提出并归口。

本标准起草单位：国家档案局馆室司、青岛市档案馆。

本标准主要起草人：刘芸、丁德胜、杨来青、邹杰。

纸质档案数字复制件光学字符识别(OCR)工作规范

1 范围

本标准规定了纸质档案数字复制件光学字符识别(OCR)工作的组织、实施和管理。

本标准适用于字迹清晰、文本规范的纸质档案数字复制件的光学字符识别(OCR)工作。

2 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,其最新版本(包括所有的修改单)适用于本文件。

DA/T 13 档号编制规则

DA/T 22 归档文件整理规则

DA/T 31 纸质档案数字化规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

字符 character

供组织、控制或表示数据用的元素集合中的一个元素。

[GB 18030—2005,定义 4.1]

3.2

字符集 character set

多个字符的集合。

注:常见字符集有 ASCII 字符集、GB 2312 字符集、BIG5 字符集、GB 18030 字符集、Unicode 字符集等。

3.3

光学字符识别 optical character recognition;OCR

通过信息技术对图像文件中的字符形状进行识别、文字转换和文本输出、呈现的过程。

3.4

纸质档案数字复制件 digital copy of paper-based record

纸质档案经过数字化加工过程后形成的,存储在磁带、磁盘、光盘等载体上并能被计算机等电子设备识别的数字图像。

3.5

档案 OCR 成果 OCR outcome of record

记录通过 OCR 技术获取的纸质档案数字复制件文字内容的文件。

3.6

识别准确率　recognition accuracy

通过 OCR 技术识别正确字符的比率。

注：识别准确率＝(识别正确字符数/应识别字符总数)×100%

3.7

识别速度　recognition speed

单位时间内通过 OCR 技术识别字符的数量。

4　总则

4.1　档案 OCR 应纳入数字档案馆(室)资源建设范畴，统筹规划、有序实施，逐步实现常态化。

4.2　档案 OCR 应科学开展，有利于实现档案信息检索和计算机辅助编目、编研开发、数据挖掘。

4.3　档案 OCR 应基于档案数字化工作，档案 OCR 成果与纸质档案数字复制件之间应建立准确、可靠的关联关系。

4.4　应当采取有效的管理和技术手段，加强档案 OCR 的过程管理和质量控制，确保档案 OCR 过程规范、成果可靠、数据安全。

4.5　涉密纸质档案数字复制件的 OCR 工作，应符合涉密档案相关的管理和技术要求。

5　工作组织

5.1　机构及人员

5.1.1　应建立档案 OCR 工作机构，配备相应素质和技术水平的工作人员，组织开展档案 OCR 工作的统筹规划、组织实施、协调管理、技术保障、安全保障、监督检查、成果验收和长期保存等。档案 OCR 可与纸质档案数字化工作统筹配置工作机构和人员。

5.1.2　档案 OCR 工作实行服务外包的，应从企业性质、股东组成、安全保密、企业规模、注册资金情况等方面严格审查档案 OCR 服务供方的相关资质；从规章制度的建立健全程度等方面考查服务供方的管理能力，建立权责明确、覆盖工作全过程的监督机制和安全防范机制，确保档案信息安全。对外聘的工作人员，应进行安全审查，按规定进行保密教育。

5.2　流程控制

5.2.1　档案 OCR 流程包括图像导入、图像预处理、比对识别、修改校正、成果整理输出五个业务环节。应依据相关技术标准，对档案 OCR 全过程进行有效控制。

5.2.2　应加强对档案 OCR 工作全流程的质量管理和安全管理，建立完善的质量、安全问题发现、修正机制，确保 OCR 成果质量和档案信息安全。

5.3　工作文件与元数据

5.3.1　应建立档案 OCR 工作方案、技术方案、工作审批材料、流程控制材料、数据验收材料、项目验收报告、成果移交材料等的管理工作文件，采取服务外包的还应包括项目招标文件、投标文件、中标通知书、项目合同、保密协议、操作规程、监管记录等，以加强对档案 OCR 工作的管理。

5.3.2　应参照相关标准，提出档案 OCR 工作流程中相关元数据设计、捕获、著录和管理的基本要求，与对应的纸质档案数字复制件管理过程元数据实施融合管理，并纳入数字档案馆(室)应用系统数据库。

6 方案制定

6.1 确定工作策略

6.1.1 OCR 工作开展前，应当依据纸质档案数字复制件 OCR 项目的计划、合同、招投标书等有关项目文件，对 OCR 工作的识别处理系统、网络系统、基础设施、保障能力等方面进行业务评价。

6.1.2 评价通过后，应根据以下因素，制定档案 OCR 的工作策略：

——图像资源：符合导入标准的可识别的彩色(24 bits)、灰度(256 阶)和黑白二值图像。一般应为 TIFF、BMP、JPG、PDF(图像)、OFD(图像)格式文件。

——OCR 引擎：对图像包含文字进行高速度和高准确率识别的 OCR 软件开发包。

——OCR 软件：装备 OCR 引擎的软件，可高速、准确输出识别成果，支持人工比对和校正。应根据需要识别的目标，按照项目资源的成本风险平衡原则确定 OCR 的范围、质量、效率、技术等要求。

——基础设施：支持系统运行的场所、设施和设备，包括 OCR 设备及工作间、介质的场外存放场所、备用的机房及辅助设施等。

——专业技术支持能力：对系统的运转提供支撑和综合保障的能力，以实现系统的预期目标。包括硬件、系统软件和应用软件的问题分析和处理能力，网络系统安全运行管理能力，沟通协调能力等。

——运行维护管理能力：保障系统相关的设备和软件正常运行，提供长期、及时、全面的技术支持的能力。包括运行环境管理、系统管理、安全管理和变更管理等。

——灾难恢复预案：对系统灾难实行快速、有效的响应和恢复。包括灾难紧急响应，灾后系统重建及重续运行，通信、后勤、技术等相关保障机制建设。

6.2 制定技术方案

6.2.1 应当根据确定的档案 OCR 工作策略制定 OCR 各工作系统技术方案，包含 OCR 的数据管理系统、OCR 识别处理系统和网络系统。技术方案中所涉及的系统应满足如下条件：

——与档案管理系统相当的安全保护级别；

——具有可扩展性；

——对档案管理系统无明显可用性和性能影响。

6.2.2 为确保技术方案满足档案 OCR 工作策略的要求，应对技术方案进行确认和验证，并记录和保存验证及确认的成果。按照确认的 OCR 软件技术方案进行开发，实现所要求的数据管理系统、OCR 识别处理系统和网络系统。

6.2.3 应按照经过确认的技术方案，制定 OCR 软件各阶段的系统安装及测试计划，以及支持不同关键业务功能的系统安装及测试计划，并组织最终用户共同进行测试。确认以下各项功能可正确实现：

——对识别图像进行预处理；

——数据识别及校验；

——输出档案 OCR 成果；

——数据安全管理。

7 档案 OCR 的实施

7.1 图像导入

7.1.1 档案 OCR 实施前，应先评估纸质档案数字复制件质量是否符合 OCR 的基本要求。评估内容

一般应包括图像分辨率、偏斜度、清晰度、失真度、亮度、对比度、灰度等。

7.1.2 纸质档案数字复制件的图像分辨率应不低于 200 dpi。特殊情况下，如文字偏小、密集、清晰度较差等，可以适当提高分辨率。文件命名应符合 DA/T 13、DA/T 22、DA/T 31 的规定。

7.1.3 对质量不能达到档案 OCR 工作基本要求的纸质档案数字复制件，应按照 DA/T 31 的要求重新数字化后导入。

7.2 图像预处理

7.2.1 二值化

7.2.1.1 在识别处理前，应对彩色图像进行灰度化和二值化处理，对灰度图像进行二值化处理。应采取局部自适应二值化等算法，并支持自动或手动调节。

7.2.1.2 应具备亮度和对比度值自动、手动调节功能。亮度和对比度值的设定以调整后的图像中文字的笔画连贯清晰为准。

7.2.2 图像降噪

7.2.2.1 对图像中印刷体字符进行识别处理前，需要根据噪声的特征对待识别图像进行降噪处理，提升识别处理的精确度。

7.2.2.2 降噪处理应去除在扫描过程中产生的污点、污线、黑边等影响图像质量的杂质，去除档案页面原有的纸张褪变斑点、水渍、污点、装订孔等影响识别的地方。

7.2.3 倾斜校正

7.2.3.1 对图像进行识别前，应进行图像方向检测并进行自动水平或垂直倾斜校正。

7.2.3.2 应支持由用户指定图像倾斜的角度，采用相应的图像旋转算法进行手工倾斜校正。

7.2.4 图像监测

图像质量控制程序应自动检测图像处理质量。对无法达到质量要求的图像进行标注。

7.3 比对识别

7.3.1 版式分析

7.3.1.1 比对识别前应对图像中的字符块结构进行版式分析，把图像中相似的版块信息划分到一起。如横排文本、竖排文本、表格、图形等。

7.3.1.2 版式分析可采取多种分析方法，自动检测各版块类型，对图像内部区域进行逻辑归类，记录各版块的位置，存储版面信息。

7.3.2 档案特征分析

7.3.2.1 归档章分析。建立归档章式样库，自动识别图像中的归档章，并根据归档章样式，识别出字段位置，如全宗号、年度、机构、保管期限、件号、页数等。

7.3.2.2 公文要素分析。建立公文格式库，可准确识别公文的版头、主体、版记三部分，识别公章、签章等区域，比照公文样式，识别密级和保密期限、紧急程度、发文字号、签发人、标题、主送机关、正文、附件说明、发文机关署名、成文日期、附注、附件、抄送机关等公文要素。公文要素 OCR 识别要求见附录 A。

7.3.2.3 表格分析。建立单独表格处理模块，建立专用表格模板定义工具，自定义文件处理单、发文稿纸、备考表等表格模板，识别表格中的字段位置。

7.3.2.4 印章分析。识别印章图像位置，存储印章图像，建立印章名称与印章图像的关系库，用于版式

恢复。

7.3.3 识别和匹配

7.3.3.1 识别时应抽取字体、字号、粗体、斜体、首行缩进等字符特征，通过相似度计算方法，与特征数据库比对，识别为计算机文字内码。

7.3.3.2 特征数据库应存储多种印刷体字符、常用签名和批注手写体字符，具备可更新和可扩充性。对使用频率高的汉字、英文、数字以及常用的符号、常用签名和批注手写体字符应建立高频库。应将无法识别的手写体筛选出来，通过人工识别，并将识别成果存入字符库。

7.3.3.3 应通过将比对后的识别文字根据上下文在可能的相似候选字群中找出最合乎逻辑的字词对识别文字进行除错或更正，以提高 OCR 识别准确率。

7.4 修改校正

7.4.1 应对识别的文本进行自动语义识别和校正，通过词汇库和语义库对识别后文本中的字符、词汇、语句自动进行逐层分析更正。词汇库和语义库应具备更新和自动学习功能。

7.4.2 应对候选字、拒认字和可能有问题的字词、语句进行标记。

7.4.3 应支持以人工方式对 OCR 成果进行图像与识别文字对照、修正等校正的功能，以满足更高识别准确率的特殊要求。

7.5 成果整理输出

7.5.1 成果整理

7.5.1.1 支持按照纸质档案数字复制件的版式对 OCR 成果的段落和表格进行版面理解与重建。重建后 OCR 成果的段落编排、表格样式应与纸质档案数字复制件图像一致。

7.5.1.2 应自动分析、提取党政机关公文的各公文要素，包括密级和保密期限、紧急程度、发文字号、签发人、标题、主送机关、正文、附件说明、发文机关署名、成文日期、附注、附件、抄送机关等。档案 OCR 成果中各公文要素位置应与纸质档案数字复制件图像一致。

7.5.1.3 应支持调用、编辑、备份、导出 OCR 成果，支持对文字、符号的搜索等功能。

7.5.2 成果输出

7.5.2.1 档案 OCR 成果应同时保存为纯文本形式和双层 PDF/OFD 文件形式。

7.5.2.2 应以纸质档案的件或页为单位输出、保存纯文本形式档案 OCR 成果。纯文本形式 OCR 成果保存规则参见表 1：

表 1 OCR 成果保存规则

纸质档案数字复制件保存形式	档案 OCR 成果保存形式	用途
一件档案保存为一个文件	一个 txt 文件	便于纸质档案数字复制件和 OCR 成果管理
一件档案分组件(收发文处理单、正文、定稿等)保存为多个文件	一个纸质档案数字复制件文件保存一个 txt 文件	
一件档案按页保存为多个文件	一页保存一个 txt 文件	便于全文检索后原件页面的准确定位和呈现

7.5.2.3 应以档号为基础对纯文本形式档案 OCR 成果命名，命名方式的选择应确保档案 OCR 成果

命名唯一性。一件档案保存为多个档案 OCR 成果文件时，应按档号结合 OCR 成果顺序流水号为档案 OCR 成果命名。

示例 1：档号为 A001-001-0001-0001 的纸质档案数字复制件，对应的 OCR 成果文件名为 A00100100010001.txt。

示例 2：档号为 A001-001-0001-0002 的纸质档案数字复制件包含收文处理单、文件正本两个文件，对应的 OCR 成果文件名分别为 A00100100010002_01.txt 和 A00100100010002_02.txt。

7.5.2.4 应根据纸质档案数字复制件版式文件格式，自动形成支持全文检索的双层 PDF 或 OFD 文件，方便全文检索后对文件的阅读。

7.5.2.5 应支持按照档案著录规则和电子档案元数据规范，自动保存档案 OCR 成果中的党政机关公文要素。相关公文要素应保存到数字档案馆(室)应用系统数据库。

7.5.2.6 应支持档案 OCR 成果中文简繁体的自动转换功能。

7.5.3 成果验收

7.5.3.1 应采用计算机自动检验与人工检验相结合的方式对纸质档案 OCR 成果进行验收检验。

7.5.3.2 验收检验内容包括 OCR 成果、提取的党政机关公文要素、数据挂接情况、OCR 工作文件和存储载体等。

7.5.3.3 能够采用计算机自动检验的项目应采用计算机自动检验的方式进行 100%检验，对于无法用计算机自动检验的项目，可根据情况以件或卷为单位采用抽检的方式进行人工检验。抽检比率不得低于 5%。

8 档案 OCR 质量要求

8.1 识别准确率

8.1.1 档案 OCR 对档案中文、数字、英文印刷体的识别准确率在 95%以上。

8.1.2 档案 OCR 对常用签名识别准确率达到 90%以上，手写体识别准确率达到 80%以上。

8.2 强抗噪能力

8.2.1 档案 OCR 应当具备对噪点的强抵抗能力，识别过程中能够有效屏蔽较大程度的噪点干扰。

8.2.2 档案 OCR 应能准确判别纸质档案数字复制件上的污点、污线、黑边、纸张褪变斑点、水渍、污点、装订孔等，提高识别准确率。

8.3 识别速度

8.3.1 识别速度指标与识别准确率指标应同时适用。

8.3.2 在主流计算机软硬件平台下，A4 纸幅面中文识别速度不低于 1 000 字/s，英文识别速度不低于 2 000 字/s。

8.4 版面还原度

8.4.1 应实现复杂版面的精确还原，采用分栏技术，智能分析中文(简体、繁体)、英文字体，文、表、图混排文本，识别后无需人工干预，自动还原排版。

8.4.2 识别后的文档与原导入图像版面还原度应达到 90%以上。

9 档案 OCR 成果的管理与应用

9.1 成果管理

9.1.1 应保持档案 OCR 成果各组成要素对应的纸质档案数字复制件、档案目录、元数据之间的逻辑层次和关联关系。

9.1.2 以纯文本形式保存的档案 OCR 成果应使用档号作为文件名，可在存储载体中以档号为基础逐级建立层次文件夹单独保存，也可与纸质档案数字复制件统一保存。

9.1.3 支持全文检索的双层 PDF 或 OFD 文件可与对应的纸质档案数字复制件统一存储。数字档案馆(室)应用系统应记录并维护不同文件版本之间的联系。

9.1.4 档案 OCR 成果文件管理权限应与纸质档案数字复制件相同。

9.1.5 OCR 成果应与纸质档案数字复制件同步开展数据备份工作。

9.2 成果应用

9.2.1 档案 OCR 成果应通过数字档案馆(室)应用系统实现全文检索，提高档案信息检索效率。

9.2.2 可发挥档案 OCR 成果提取的归档信息和党政机关公文要素的作用，辅助开展档案自动著录、目录质量核查，以及纸质档案数字复制件挂接准确性核查等业务工作。

9.2.3 可利用档案 OCR 成果，结合数据挖掘技术开展数据分析、知识管理、词库建设等工作。

附 录 A
（规范性附录）
公文要素 OCR 识别要求

公文要素 OCR 识别要求见表 A.1。

表 A.1 公文要素 OCR 识别要求

公文要素	是否识别	识别要求
1 正本		
1.1 份号	否	—
1.2 密级和保密期限	是	正常识别，识别成果辅助档案著录
1.3 紧急程度	是	正常识别，识别成果辅助档案著录
1.4 发文机关标志	否	—
1.5 发文字号	是	正常识别，识别成果辅助档案著录
1.6 签发人	是	正常识别
1.7 标题	是	正常识别，去除软回车，识别成果辅助档案著录
1.8 主送机关	是	正常识别，去除软回车
1.9 正文	是	正常识别，去除软回车，表格基本符合原貌，单元格内文字内容完整
1.10 附件说明	是	正常识别，去除软回车
1.11 发文机关署名	是	正常识别，自动拆分联合发文机关，识别成果辅助档案著录
1.12 成文日期	是	正常识别，识别成果辅助档案著录
1.13 印章	否	—
1.14 签发人签名章	否	—
1.15 附注	是	正常识别，去除软回车
1.16 附件	是	正常识别，去除软回车，表格基本符合原貌，单元格内文字内容完整
1.17 抄送机关	是	正常识别，去除软回车
1.18 印发机关和印发日期	否	—
1.19 页码	否	—
2 文件处理单/发文稿纸		
2.1 起草人	是	正常识别，手写体自动学习提高识别效率，识别成果辅助档案元数据记录
2.2 签发人	是	正常识别，手写体自动学习提高识别效率，识别成果辅助档案元数据记录
2.3 起草时间	是	正常识别，识别成果辅助档案元数据记录
2.4 签发时间	是	正常识别，识别成果辅助档案元数据记录
2.5 阅办意见	是	正常识别，手写体自动学习提高识别效率，识别成果辅助档案元数据记录
2.6 批办意见	是	正常识别，手写体自动学习提高识别效率，识别成果辅助档案元数据记录
2.7 办理结果	是	正常识别，手写体自动学习提高识别效率，识别成果辅助档案元数据记录