ICS 01.140.20
A 14
备案号:15821—2005

DA

中华人民共和国档案行业标准

DA/T 31—2005

纸质档案数字化技术规范

Specification for digitization of paper-based records

2005-04-30 发布　　2005-09-01 实施

国　家　档　案　局　发布

前　言

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。

本标准主要起草人：王良城、马淑桂、蔡伟、宋涌、王大众、韩冬、田军、孙森林。

纸质档案数字化技术规范

1 范围

本标准规定了纸质档案数字化的主要技术要求。

本标准适用于采用各种设备对纸质档案的数字化加工处理及数字化成果的管理。

2 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准的条款。凡是注日期的引用文件，其随后所有的修改单(不包括勘误的内容)或修订版均不适用于本标准，然而，鼓励根据本标准达成协议的各方研究是否可使用这些文件的最新版本。凡是不注日期的引用文件，其最新版本适用于本标准。

GB/T 17235.1—1998 信息技术 连续色调静态图像的数字压缩及编码 第1部分：要求和指南

GB/T 17235.2—1998 信息技术 连续色调静态图像的数字压缩及编码 第2部分：一致性测试

GB/T 18894—2002 电子文件归档与管理规范

3 术语和定义

下列术语和定义适用于本标准。

3.1

数字化 digitization

用计算机技术将模拟信号转换为数字信号的处理过程。

3.2

纸质档案数字化 digitization of paper-based records

采用扫描仪或数码相机等数码设备对纸质档案进行数字化加工，将其转化为存储在磁带、磁盘、光盘等载体上并能被计算机识别的数字图像或数字文本的处理过程。

3.3

数字图像 digital image

表示实物图像的整数阵列。一个二维或更高维的采样并量化的函数，由相同维数的连续图像产生。在矩阵(或其他)网络上采样——连续函数，并在采样点上将值最小化后的阵列。

3.4

黑白二值图像 binary image

只有黑白两级灰度的数字图像。它对应于黑白两种状态的文字稿、线条图等。

3.5

连续色调静态图像 continuous-tone still image

以多于两级灰度的不同浓淡层次或以不同颜色通道组合成的静态数字图像。在纸质档案数字化过程中，通常表现为灰度扫描和彩色扫描两种模式。

3.6

分辨率 resolution

单位长度内图像包含的点数或像素数，一般用每英寸点数(dpi)表示。

3.7

失真度 distortion measure

对档案进行数字化转换后，数字图像与档案原件在色彩、几何等方面的偏离程度。

3.8

可懂度 intelligibility

数字图像向人或机器提供信息的能力。

3.9

图像压缩 image compression

清除图像冗余或对图像近似的任一种过程，其目的是对图像以更紧凑的形式表示。纸质档案数字化过程中，较常见的有 TIFF(G4)、JPEG 等压缩格式。

4 纸质档案数字化基本要求

4.1 基本原则

纸质档案数字化的基本原则是使档案信息资源准确、方便、快捷地提供利用，使可以公开的档案信息资源得到共享，以满足社会对档案利用的需求。

4.2 数字化对象的确定原则

应当对所要进行数字化的对象按照一定的原则和方法进行确认，只有符合一定要求的纸质档案文献才能进行数字化。

4.2.1 符合国家法律法规的原则

纸质档案的数字化，必须符合国家档案开放规定以及有关规定。

4.2.2 价值性原则

属于归档范围且应永久或长期保存的、社会利用价值高的档案可列入数字化加工的范围。

4.3 基本环节

纸质档案数字化的基本环节主要包括：档案整理、档案扫描、图像处理、图像存储、目录建库、数据挂接、数据验收、数据备份、成果管理等。

4.4 过程管理

4.4.1 应加强纸质档案数字化各环节的安全保密管理机制，确保档案原件和数字化档案信息的安全。

4.4.2 纸质档案数字化的各个环节均应进行详细的登记，并及时整理、汇总，装订成册，在数字化工作完成的同时建立起完整、规范的记录。

5 档案整理

在扫描之前，根据档案管理情况，按下述步骤对档案进行适当整理，并视需要作出标识，确保档案数字化质量。

5.1 目录数据准备

按照《档案著录规则》(DA/T 18)等的要求，规范档案中的目录内容。包括确定档案目录的著录项、字段长度和内容要求。如有错误或不规范的案卷题名、文件名、责任者、起止页号和页数等，应进行修改。

5.2 拆除装订

在不去除装订物情况下，影响扫描工作进行的档案，应拆除装订物。拆除装订物时应注意保护档案不受损害。

5.3 区分扫描件和非扫描件

按要求把同一案卷中的扫描件和非扫描件区分开。普发性文件区分的原则是：无关和重份的文件要剔除，有正式件的文件可以不扫描原稿。

5.4 页面修整

破损严重、无法直接进行扫描的档案，应先进行技术修复，折皱不平影响扫描质量的原件应先进行相应处理(压平或熨平等)后再进行扫描。

5.5 档案整理登记

制作并填写纸质档案数字化加工过程交接登记表单，详细记录档案整理后每份文件的起始页号和页数。

5.6 装订

扫描工作完成后，拆除过装订物的档案应按档案保管的要求重新装订。恢复装订时，应注意保持档案的排列顺序不变，做到安全、准确、无遗漏。

6 档案扫描

6.1 扫描方式

6.1.1 根据档案幅面的大小(A4、A3、A0 等)选择相应规格的扫描仪或专业扫描仪(如工程图纸可采用 0 号图纸扫描仪)进行扫描。大幅面档案可采用大幅面数码平台，或者缩微拍摄后的胶片数字化转换设备等进行扫描，也可以采用小幅面扫描后的图像拼接方式处理。

6.1.2 纸张状况较差，以及过薄、过软或超厚的档案，应采用平板扫描方式；纸张状况好的档案可采用高速扫描方式以提高工作效率。

6.2 扫描色彩模式

6.2.1 扫描色彩模式一般有黑白二值、灰度、彩色等。通常采用黑白二值。

6.2.2 页面为黑白两色，并且字迹清晰、不带插图的档案，可采用黑白二值模式进行扫描。

6.2.3 页面为黑白两色，但字迹清晰度差或带有插图的档案，以及页面为多色文字的档案，可采用灰度模式扫描。

6.2.4 页面中有红头、印章或插有黑白照片、彩色照片、彩色插图的档案，可视需要采用彩色模式进行扫描。

6.3 扫描分辨率

6.3.1 扫描分辨率参数大小的选择，原则上以扫描后的图像清晰、完整、不影响图像的利用效果为准。

6.3.2 采用黑白二值、灰度、彩色几种模式对档案进行扫描时，其分辨率一般均建议选择大于或等于 100 dpi。特殊情况下，如文字偏小、密集、清晰度较差等，可适当提高分辨率。

6.3.3 需要进行 OCR 汉字识别的档案，扫描分辨率建议选择大于或等于 200 dpi。

6.4 扫描登记

认真填写纸质档案数字化转换过程交接登记表单，登记扫描的页数，核对每份文件的实际扫描页数与档案整理时填写的文件页数是否一致，不一致时应注明具体原因和处理方法。

7 图像处理

7.1 图像数据质量检查

7.1.1 对图像偏斜度、清晰度、失真度等进行检查。发现不符合图像质量要求时，应重新进行图像的处理。

7.1.2 由于操作不当，造成扫描的图像文件不完整或无法清晰识别时，应重新扫描。

7.1.3 发现文件漏扫时，应及时补扫并正确插入图像。

7.1.4 发现扫描图像的排列顺序与档案原件不一致时，应及时进行调整。

7.1.5 认真填写相关表单，记录质检结果和处理意见。

7.2 纠偏

对出现偏斜的图像应进行纠偏处理，以达到视觉上基本不感觉偏斜为准。对方向不正确的图像应进行旋转还原，以符合阅读习惯。

7.3 去污

对图像页面中出现的影响图像质量的杂质，如黑点、黑线、黑框、黑边等应进行去污处理。处理过程中应遵循在不影响可懂度的前提下展现档案原貌的原则。

7.4 图像拼接

对大幅面档案进行分区扫描形成的多幅图像，应进行拼接处理，合并为一个完整的图像，以保证档案数字化图像的整体性。

7.5 裁边处理

采用彩色模式扫描的图像应进行裁边处理，去除多余的白边，以有效缩小图像文件的容量，节省存储空间。

8 图像存储

8.1 存储格式

8.1.1 采用黑白二值模式扫描的图像文件，一般采用 TIFF(G4)格式存储。采用灰度模式和彩色模式扫描的文件，一般采用 JPEG 格式存储。存储时的压缩率的选择，应以保证扫描的图像清晰可读的前提下，尽量减小存储容量为准则。

8.1.2 提供网络查询的扫描图像，也可存储为 CEB、PDF 或其他格式。

8.2 图像文件的命名

8.2.1 纸质档案目录数据库中的每一份文件，都有一个与之相对应的唯一档号，以该档号为这份文件扫描后的图像文件命名。

8.2.2 多页文件可采用该档号建立相应文件夹，按页码顺序对图像文件命名。

9 目录建库

9.1 数据格式选择

目录建库应选择通用的数据格式。所选定的数据格式应能直接或间接通过 XML 文档进行数据交换。

9.2 档案著录

按照《档案著录规则》(DA/T 18)的要求进行著录，建立档案目录数据库。

9.3 目录数据质量检查

采用人工校对或软件自动校对的方式，对目录数据库的建库质量进行检查。核对著录项目是否完整、著录内容是否规范、准确，发现不合格的数据应要求进行修改或重录。

10 数据挂接

10.1 汇总挂接

档案数字化转换过程中形成的目录数据库与图像数据库，通过质检环节确认为“合格”后，通过网络及时加载到数据服务器端汇总。通过编制程序或借助相应软件，可实现目录数据对相关联的数字图像的自动搜索、加入对应的电子地址信息等，实现批量、快速挂接。

10.2 数据关联

以纸质档案目录数据库为依据，将每一份纸质档案文件扫描所得的一个或多个图像存储为一份图像文件。将图像文件存储到相应文件夹时，要认真核查每一份图像文件的名称与档案目录数据库中该份文件的档号是否相同，图像文件的页数与档案目录数据库中该份文件的页数是否一致，图像文件的总数与目录数据库中文件的总数是否相同等。通过每一份图像文件的文件名与档案目录数据库中该份文件的档号的一致性和唯一性，建立起一一对应的关联关系，为实现档案目录数据库与图像文件的批量挂接提供条件。

10.3 交接登记

认真填写纸质档案数字化转换过程交接登记表单，记录数据关联后的页数，核对每一份文件关联后的页数与档案整理、扫描时填写的页数是否一致，不一致时应注明具体原因和处理办法。

11 数据验收

11.1 数据抽检

11.1.1 以抽检的方式检查已完成数字化转换的所有数据，包括目录数据库、图像文件及数据挂接的总体质量。

11.1.2 一个全宗的档案，数据验收时抽检的比率不得低于5%。

11.2 验收指标

11.2.1 目录数据库与图像文件挂接错误，或目录数据库、图像文件之一出现不完整、不清晰、有错误等质量问题时，抽检标记为“不合格”。

11.2.2 一个全宗的档案，数字化转换质量抽检的合格率达到95%以上（含95%）时，予以验收“通过”。

合格率＝抽检合格的文件数/抽检文件总数×100%

11.3 验收审核

验收“通过”的结论，必须经分管领导审核、签字后方有效。

11.4 验收登记

认真填写纸质档案数字化验收登记表单。

12 数据备份

12.1 备份范围

经验收合格的完整数据应及时进行备份。

12.2 备份方式

为保证数据安全，备份载体的选择应多样化，可采用在线、离线相结合的方式实现多套备份，并注意异地保存。

12.3 数据检验

备份数据也应进行检验。备份数据的检验内容主要包括备份数据能否打开、数据信息是否完整、文件数量是否准确等。

12.4 备份标签

数据备份后应在相应的备份介质上做好标签，以便查找和管理。

12.5 备份登记

填写纸质档案数字化备份管理登记表单。

13 数字化成果管理

13.1 应加强对纸质档案数字化成果的管理，确保其安全、完整和长期可用。

13.2 纸质档案数字化成果提供网上检索利用时，应有制作单位的电子标识，并根据具体情况分别采用可下载或不可下载的数据格式。