综合档案数字化制作系统简介

（V 7.X 版）

随着国家对纸质档案保管、利用的重视力度逐渐增大，每年国家投入到用于纸质档案材料保管的费用也在逐年递增（包括库房、防潮、防火、防虫以及纸质档案的数字化加工等费用），由此衍生了纸质档案数字化加工这个行业。作为档案行业基础上衍生的档案数字化加工行业，其特点是在传统劳动密集型行业基础上，引入档案行业相关知识和技术，专门服务于档案行业大批量纸质档案的电子化转换工作。**其产生与存在是社会生产力发展的必然结果。**随着和平年代社会的快速发展与进步，由此产生了大量纸质材料作为历史依据被归档保存，然而现有编制的档案管理人员数量极其有限，除了日常实体档案的保管之外，根本无力于电子档案的收集、采集与转换。因此，为了满足档案信息化建设的需要，需要大量的外部的专业团队介入，以最短的时间、最低的成本投入，来完成纸质档案的电子化转换工作，从而更好的提高通过电子档案快速查询、利用所带来的效率，“至下而上”更进一步促进社会生产力的高速发展。

纸质档案数字化加工行业特点：

属于劳动密集型行业，故先从组成人员的角度来看：

1、 数字化加工人员年龄普遍较低（入世不深、便于管理）。
2、 数字化加工人员文化程度相对不高(对档案信息的敏感度不高，降低信息安全隐患)。
3、 数字化加工人员计算机水平有限（人员培养需要消耗成本）。
4、 数字化加工人员流动性极大（管理层面、安全层面存在风险）。

从市场角度来看：

1、 社会需求量很大（档案保管编制人员有限）。
2、 社会需求量在逐年递增（社会生产力高速发展的必然结果）。
3、 可持续性较强(光栅图像的电子档案是实体档案的快照，具有相对同等的法律效应)。
4、 市场竞争越来越激烈（进入门坎较低，加工团队日益俱增）。

从管理及成本角度来看：

1、 加工团队越来越专业（经过了多个项目的经验积累）。
2、 加工成本逐渐降低，从而抵消通货膨胀所带来的利润贬值（团队存在的基础）。
3、 团队管理、建设的新思路层出不穷（一人一管法，没有太多可复制的东西）。

从技术角度来看：

1、 档案专业知识日积月累（档案行业是一个“包容”的行业，多多少少都会涉及到各个行业的专业知识。比如搞诉讼档案加工，就要清楚有了“送达回证”意味着可以封卷了）。
2、 软件技术水平逐渐提高（涉足各行业的业务档案越来越多）。
3、 电子档案信息安全问题迫在眉睫（泄密 99%发生在信息保存和利用这个环节，属于电子档案管理软件系统平台，在信息安全方面的设计问题。与有无**保密安全资质**没有太大关系，只是某些同行公司在商业竞标中设置的门坎而已，只要企业注册资本够几千万都可以搞到，所以并不安全。美国五角大楼的机密档案，频频泄密在哪个环节~o~大家都清楚）。

综上分析，档案数字化行业作为一个新兴行业，其发展速度之快已远远超出预料。其成本投入及利润空间，将与一些传统行业平齐甚至赶超（如钢铁、石油、冶金等行业的利润比例，不一定比档案数字化行业高，因此有大批量的房地产开发商转型）。

通过上面对档案数字化加工行业现状的分析，档案数字化加工行业需要一套简单、高效、安全的管理与技术作支撑，在保质、保量的完成加工任务的同时，来降低成本损耗，提高利润空间。我们围绕着这个目的，同时结合十几年的项目经验，逐渐完善起来了一套专门用于纸质档案数字化加工制作的工具软件产品《综合档案加工制作系统》。其软件特点如下：

1、 设计轻巧，便于安装携带

由于档案数字化项目的不定时、不定点、不定人数的不确定性，因此要求软件可以快速复制、安装到位，即便只有一个人也可以立即展开工作。（Win7 下需要以管理员身份安装）。

2、 操作简单，便于快速上手

由于数字化加工人员信息化程度不高，因此，在软件运行与使用方面，将复杂的操作与配置隐藏在“后面”。初次接触的操作人员可以快速上手、进入状态，而把复杂的操作及配置，留到后续的加工过程中逐步学习（循序渐进、由浅入深）。如此一来，所有加工人员经过一个项目后，都有可能成为项目管理人员，来带队做下一个项目。这样将大大减少人员培养的成本，同时也降低了由于人员流动所带来的损失。

3、 结构设计灵活，适于多种类型档案

随着社会生产力的高速发展，新兴行业、产业层出不穷，档案类别千变万化（已经不仅仅局限于国档局定义的七大门类）。因此，要完成不同类型档案的数字化加工任务，需要面对不同类型档案的目录及层次结构（或同时面对多种档案门类结构）。故需要软件可以随时修改或更新数据库字段组成结构，以此来满足不同项目、不同门类档案的数字化加工需求。

4、 批处理能力强大，减少人工耗时

面对海量的电子图像数据及著录的条目数据，如果全部加工过程完全由人员逐个、逐条、逐页的手工完成，何谈效率。因此，系统提供了大部分完全可以由计算机自动完成的批处理操作（如图像的转换合成、条目数据批量导入导出的方式来“合库”等等），而加工人员只需把精力放在批处理前后数据的准确性对比上。

5、 按国家标准执行，降低跑偏几率

没有规矩不成方圆，档案数字化加工行业亦是如此。国家档案局（包括中央档案馆）通过多年的实践和研究，总结并颁布了纸质档案数字化加工的技术规范，可以说是“非常科学，很有远见”。因此，软件系统完全按照标准中的规范来设计，在实际实施过程中，在技术层面上防止用户因为“跑偏”而造成的巨大损失。（很多地方档案馆还在扫描多页 TIFF，并没有原始扫描数据的概念，更何谈原始数据备份，如果灾难出现，后果很难预料。同一份材料，扫描生成的多页 TIFF 文件与 PDF 文件，哪个占用的磁盘空间大，哪个可以加密码？同一页纸，扫描生成的真彩色单页 TIFF 文件与 JPG 文件，哪个占用的磁盘空间大？采购磁盘存储硬件需要多花出多少钱？）

6、 自动查错纠错，提高加工质量

“只要是人参与的工作，就一定会存在出错可能”这个命题是真命题。在数字化加工过程中，有谁会说他们一定不会出错。既然是这样，何不在软件设计之初，就把它考虑进去。让软件为其自动的判别错误，甚至纠正错误，把人员解放出来。当然，计算机不是万能的，需要人工干预的地方也很多，但整体上可以减少人员的大部分工作量（传统方式的人工查错，会有遗漏，而计算机一旦设置好规则，则不存在这个问题）。同时，通过软件自动批处理操作，加大了质检力度，从而保障最终挂接应用的数据更加准确（直到目前为止，有哪家单位敢确信，自己的档案管理系统中挂接的加工数据一定没问题）。

7、 备份数据加密，减少安全隐患

国家档案局颁布的标准中，只是规定了加工后的原始数据要进行备份（包括光盘、

硬盘备份），但没有要求备份在磁盘中的图像数据是否进行加密。设想一下，如果单位因为搬家，光盘遗失了怎么办？在档案进馆或数据报送过程中，备份磁盘丢失了怎么办？被网络黑客非法复制怎么办？因此，在某些特殊情况下，我们需要把扫描的图像数据进行加密，增加一层安全措施。当然，在档案图像数据加密过程中，要考虑到主要的两个因素。一是加密前后文件的磁盘存储空间不变（即空间没有变大）。二是加密单个图像文件的时间，几乎等于保存该图像数据到文件的时间（即时间没有多少延迟，本系统初步统计 1 小时加解密 10 万页，CPU 运算速度也要考虑进去）。

8、 免费复制升级，便于传播推广

通过广泛的收集并汇总全国各地数字化团队遇到的技术问题，将解决方案设计到加工软件之中，从而实现软件功能的更新与升级，并将升级后的加工软件发布于网络，大家又可以免费下载使用。如此循环，可以实现档案数字化技术的提高、普及与推广，从而保障档案数字化行业的健康、高效、快速的发展，这也是我们每一个档案人的愿望。

前面提到的是档案数字化加工软件的设计初衷、理念及愿望，接下来了解一下软件具有哪些功能及技术特点。

1、档案扫描模块

特点：

- 通过快捷键操作，快速选择扫描图像的颜色、分辨率和文件存储格式。
- 通过文件列表可以实时的监控扫描的图像是否完整和连续。
- 通过图像预览窗口查看扫描图像的质量。
- 用户可以随时自定义操作快捷键，以便于提高扫描速度。
- 所有操作完全可以由键盘操作控制来完成，无需太多鼠标干预。
- 通过文件加前缀或后缀名区分拼图文件，按字母顺序排序。
- 通过自定义扫描框架，使扫描仪设备只扫描选中框架区域。
- 通过扫描快捷键或菜单，指定扫描后文件的所在位置（F5、F6、F7 键）。
- 自动获取扫描仪硬件支持的功能信息（CAP），便于检查扫描仪对 Twain 的支持力度。
- 通过在文件列表中选中文件，来执行文件删除或批量重命名等操作。
- 支持三种文件存储模式（本地磁盘、FTP 服务器端、档案数字化服务器端），来存放扫描文件，完全内存中压缩，没有本地临时文件存储。

【操作过程】只需两步。根据纸张上是否存在红头、红章，先按下 F2 或 F4 键，来选择指定的颜色类型、文件格式。然后，按下 F5（或 F6、F7）键启动扫描过程。

2、图像处理模块

特点：

- 通过文件夹列表，快速选中文件存放目录，同时打开图像文件。
- 只要档案图像信息被用户处理过，系统会自动保存图像信息。
- 图像优化处理过程，可由键盘和鼠标的组合操作来完成。
- 自动化处理操作，可大大减少了每一页电子档案的处理时间。
- 用户可以随时自定义操作快捷键，以便于提高图像处理速度。
- 真彩色图像纠偏后，图像内容及边界不再存在“锯齿”现象。
- 支持通过掩码图像处理及合成技术，来实现当前图像的高清化处理。
- 通过在文件列表中选中文件，来执行文件删除或批量重命名等操作。
- 支持三种文件存储模式（本地磁盘、FTP 服务器端、档案数字化服务器端），来获取及保存文件数据。完全内存中压缩/解压缩，没有本地临时文件存储。

3、干部档案图像高清处理模块

特点：

- 双路径下文件存储。目标路径下存放高清处理后的图像文件。
- 同步移动、同步缩放，便于高清处理前后图像的比对。
- 支持图像底色替换。用户可以指定任意颜色为图像的背景色。
- 支持局部高清处理。用户可选中图像中某个区域，然后对该区域图像进行清晰化处理。
- 支持掩码图像合成。对某些小斑点较多的彩色图像，可以通过对其掩码图自动去斑点，然后再合成方式，来实现图像的清晰化处理。

【备注】如果来源图像为黑白二值图像（非真彩色图），软件会自动将其转换为真彩色图像，然后再进行清晰化处理。

4、工程图纸拼接模块

特点：

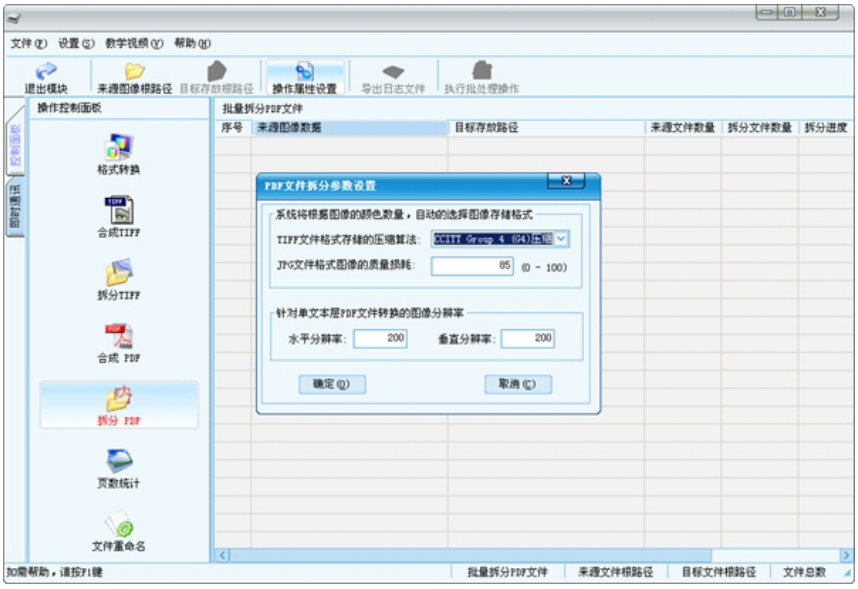
- 用户可以添加多个拼接图像碎片，没有数量限制。
- 支持底层画布尺寸调整。用户可随时根据碎片数量来调整画布的大小。
- 仿 PhotoShop 软件的图像旋转任意角度操作方式。
- 拼接后保存图像，旋转的碎片图像边界无“锯齿”现象。
- 支持拼接碎片定位。通过选中缩略图中的碎片图像，来定位画布上的实际拼接碎片。
- 支持碎片图像透明合成，多图像层间叠加图像透明合成。

【备注】可扩展功能很多，有待用户提出。

5、图像文件批处理模块

特点：

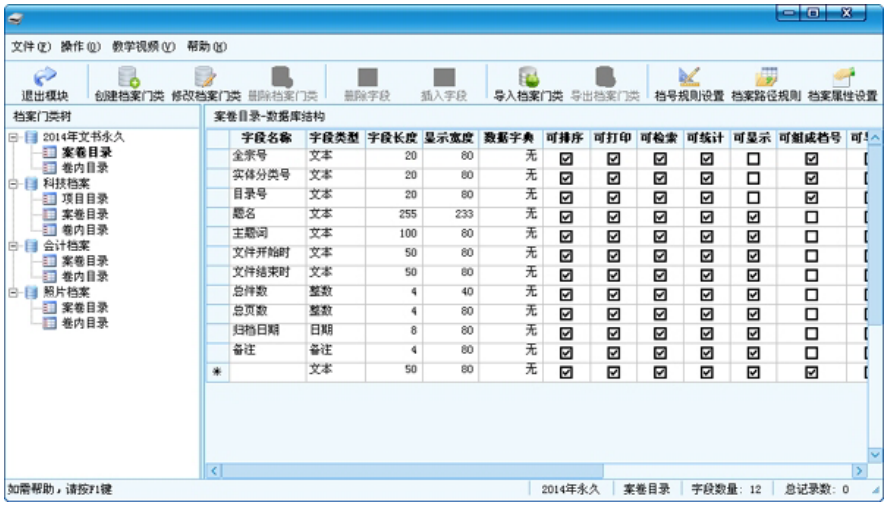
- 用户指定文件存放的来源及目标根路径，系统会自动完成海量文件数据的批处理操作。
- 动态显示处理结果，便于用户发现问题。
- 支持批处理多线程任务的停止（终止）操作。
- 处理结果列表可导出到文本文件，以便于用户及时的改错。



6、档案门类维护模块

特点：

- 系统内置了全国各档案馆通用的档案门类数据库结构模板。用户通过选择档案门类结构模板库的方式来创建档案门类。然后，在各门类结构基础上，简单修改数据库字段结构。
- 支持多种档案门类的创建、修改和维护功能。
- 采用分层目录结构，建立档案门类与其对应的档案目录表间的层次关系。
- 采用表格方式列举出档案目录库的字段组成结构。用户可直接进行修改，改后立即生效。
- 支持档案门类结构的导入和导出功能，方便用户移植档案门类结构，统一所有条目著录终端的数据库结构。



【备注】在导出的门类结构文件中包含用户设置的档号规则信息。

7、档案目录著录模块

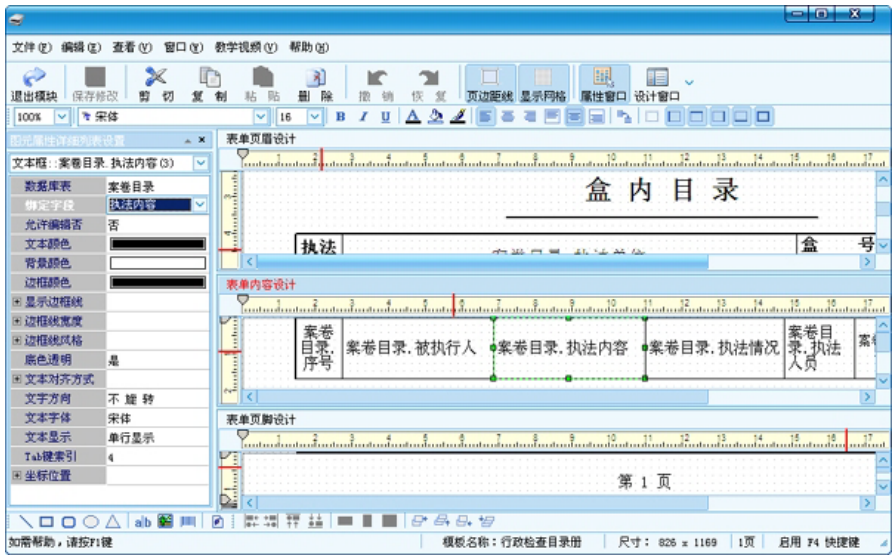
特点：

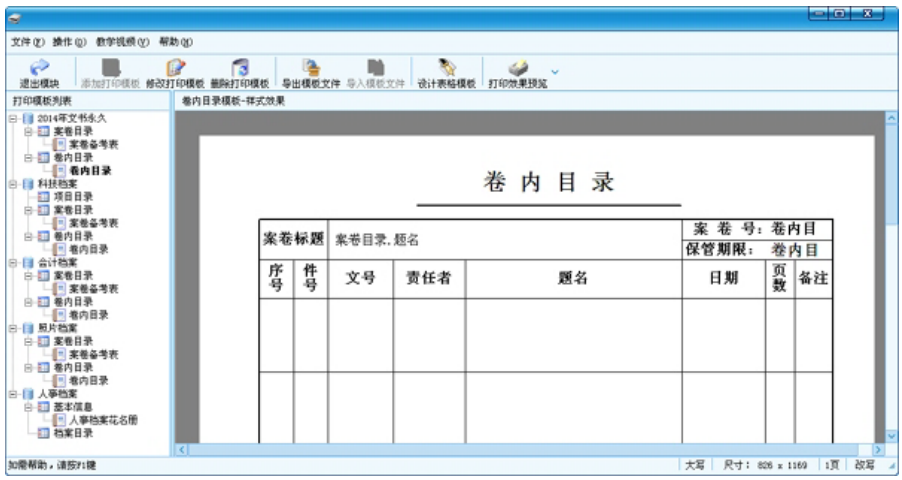
- 各层目录之间的层次结构清晰，出错几率大大降低。
- 支持多种档案门类同时著录。用户随时可以选择准备著录的档案门类。
- 著录方式为表格方式，数据著录操作比较直观。
- 著录数据时可以完全使用键盘进行操作，著录速度很快。
- 数据字典的关联，用户经常著录的内容可以存于字典之中，再次著录时只需选择即可。
- 隐藏无需著录字段（或内容完全相同的字段），使著录界面变得十分简洁、重点突出，降低出错几率。
- 自动填充字段内容。用户可以指定某一列的填充内容，由系统来完成填充操作（其填充的内容可以是连续递增的数字，或内容相同的文本）。
- 自动生成或批量更新“归档号”。著录完毕的条目数据，符合档案管理软件挂接要求。
- 级联字段自动填充。用户可以定义卷内目录中的某个字段，与案卷目录中某个字段的数据内容相同，以此来建立对应关系，称这两个字段是级联字段。在著录过程中，系统自动提取案卷目录中级联字段的数据内容，然后填充到卷内目录中与之关联的字段中。
- 支持多字段列排序，方便著录人员观察条目之间的顺序关系。
- 支持 Excel、Access 文件数据的导入与导出，方便数据迁移、汇总及阶段性备份工作。
- 支持从系统粘贴板中复制或粘贴条目数据。可以在 Excel 文件中选中一块区域复制，然后到著录条目的末尾行，粘贴从 Excel 中复制的数据。或将选中的条目数据行复制，粘贴到 Excel 文件的单元格中。
- 支持组合条件查询方式的打印操作。同时也可以选中多条记录行，直接进行打印。
- 允许用户自定义打印表单。用户可以自行设计打印表单样式。

8、打印模板设计模块

特点：

- 操作直观，通过鼠标拖放图元来改变生成表单的样式效果。
- 通过可编辑图元绑定数据库字段的方式，加载档案目录数据。
- 支持表单和表格两种类型模板的创建与设计。
- 支持打印模板的导入和导出，以便于打印表单样式的统一。





9、数据校对挂接模块

特点：

- 支持多档案门类数据校对。系统为每个档案门类，单独提供校对配置信息的数据库存储。
- 原文与条目的对应关系比较直观。用户可以选择原文目录，来查看匹配的数据条目，或通过选择条目来查看匹配的原文。
- 原文图像内容级校对。用户可以点击文件列表中的文件，来打开浏览扫描原文。通过原文中的内容，来判断匹配的条录记录是否著录正确。
- 系统辅助自动校对。系统隐藏了复杂的检测校对过程，用户只需根据实际的操作逻辑，选择相应的操作按钮，即可自动完成校对任务。
- 以错误列表方式显示校对结果。用户可以选中列表中的错误记录，定位到原文或档案目录上，分析鉴别错误原因，及时纠正错误。
- 采用“原文表”方式存储挂接后条目。系统完成自动挂接操作后，会在数据库的原文表中产生挂接原文记录。在后续的数据备份（光盘制作）制作完毕后，导出的光盘 Access 数据库中包含有中文字段的原文表，其主键字段为卷内级档号（或文件级档号）。
- 系统单独为每个档案门类的原文路径设置及档案路径规则配置进行保存。并且不同的数据类型（原始数据或成品数据）分别进行保存。以此来保障各档案门类、各种数据类型，在校对时不会产生**数据来源混乱**现象。

10、档案数据备份模块

特点：

- 支持多档案门类数据备份操作。用户可为不同档案门类创建数据备份信息集。
- 支持三种类型光盘描述符：VCD、DVD、自定义类型。其中自定义类型的光盘描述符，可用来创建硬盘数据备份包（可以是几十个 GB）。
- 每个档案门类下可以创建多个不同类型的光盘描述符。
- 支持两种类型数据的备份，原始数据和成品数据。
- 无需指定备份数据的硬盘存放信息。不论是原始数据的备份工作，还是成品数据的备份工作。其文件存放于磁盘路径信息及路径规则设置情况，是在数据校对模块中指定的，并保存到数据库字段中。因此在实际的数据备份操作中，用户无需关心数据的来源情况。
- 通过向已经定义的光盘包（光盘描述符）中，移入移出档案条目数据的方式，来完成备份数据包的逻辑划分过程。
- 以标尺方式显示数据容量。在实际划分备份数据包时，用户可以通过观察标尺上显示的容量刻度，来判断是否需要再添加条目数据，还是移出一部份条目数据。

- 支持加密备份数据。备份数据加密是原文图像级加密，而非数据库条目级加密。也就是说，系统会对每一页扫描图像文件的内容进行加密，而在实际浏览时，输入正确密钥才能解密浏览图像文件的内容。此加密操作只针对原始扫描图像（单页 TIFF 和 JPG），而成品数据（多页 TIFF 和 PDF）不支持加密。
- 即时备份，延迟制作。就是说用户可以随时创建数据备份包（创建光盘描述符，添加条目数据），软件系统是在逻辑上对用户划分的数据备份包进行保存，而非实际原文在磁盘上的存储。当需要制作成备份数据包时，系统会根据用户指定的磁盘路径，创建并复制档案文件及数据库条目到该磁盘路径下。

【备注】只有在数据校对模块中进行了“自动挂接”操作后，才能在光盘数据备份中获取到实际数据磁盘容量。才能实际制作过程中，复制原文到目标文件夹下。

11、光盘数据浏览模块

特点：

- 可脱离软件环境，在光盘上单独运行。无需安装档案数字化系统或其它软件。
- 可以直接查询、定位档案条目数据。光盘浏览器程序支持组合条件查询档案条目数据。
- 支持 PDF 文件的直接浏览。无需安装 Adobe PDF Reader，光盘浏览器程序可直接加载 PDF 文件格式的档案数据。
- 支持浏览加密后的图像数据。用户只要输入一次正确的数据加密密钥，便可以浏览全部的加密的扫描原文。
- 支持自解密批处理过程。光盘浏览器程序，可以自动批量解密光盘中的图像文件，到本地的磁盘文件夹下，用户只需输入正确的解密密钥。
- 与档案管理系统软件无缝连接。光盘数据库中包含中文字段的档案目录及原文表，并且以档号作为主键字段，关联各个层次的数据库表集。在实际挂接到档案管理软件的过程中，对于专业的技术人员而言，只需要一两条 SQL 语句，便可以完成全部的挂接操作。最后，将档案图像文件复制到本地磁盘上，档案软件指定的文件夹下即可。

【备注】采用专门的“原文表”方式，来存放挂接原文记录的数据库设计，是当今档案管理软件设计中比较通用的方式，此种方式优势很多（具体技术问题，可以去查资料）。如果您的档案管理系统（或平台），不是采用这种方式挂接扫描原文，建议您可以去升级档案管理软件了。

12、档案路径

如果说唯一可以描述一条无重复档案条目的关键字段是“档号”，则唯一可以描述一件电子档案图像文件无重复存放的则是其文件夹名。那么存放扫描原文的文件夹有哪些特点？接着分析一下。

磁盘文件存放的路径组成形式：**盘符+文件夹名+文件名**。例如“C:\AAA\001.jpg”，表示盘符=“C：”、文件夹=“AAA”、文件名=“001.jpg”。由于文件夹可以多层，比如“C:\AAA\BBB\CCC\001.jpg”，则文件夹=“AAA\BBB\CCC\”。

在此基础之上，如果把文件夹逻辑上划分为两种类型，分别为“**绝对路径**”文件夹和“**档案路径**”文件夹。那么上例中的路径可以变化为“C:\扫描原文\2013\Y\005\001.jpg”，其中，**绝对路径**文件夹=“扫描原文\”，**档案路径**文件夹=“2013\Y\005\”。

如此设计之后，我们的扫描原文存放的磁盘路径，可以随意更改盘符和绝对路径文件夹，其下档案路径文件夹的结构不会发生任何变化。只要结构化后档案路径文件夹，确保其不会发生重复即可。为避免文件夹重复（互相覆盖）最简单的方法，就是使用数据库中的档号作为文件夹名。条目不重复，文件夹名也不会重复。上例中可以变换为这样的“**C:\扫描原文\2013-Y-005\001.jpg**”。到此，原文档案存放的磁盘路径，可以在逻辑上划分出档案路径。

档号是如何生成的？是根据档号规则，从数据库字段中提取出数据，经过与“分隔符”的顺序组合而产生的。只要确定数据库中组成档号的字段存在数据内容，则档号就有会产生并存在。

既然档号可以这样产生，那么档案路径是否也可以如此方式产生？答案是肯定的。上例中可以变换为这样的“**C:\扫描原文\2013\Y\005\001.jpg**”。档案路径=“2013\Y\005\”。举一反三，档案路径的产生规则与档号的生成规则可以不同。如此，档案路径便具有了自主的独立性，同时又具有了可规则化配置的特点（数据校对模块中的档案路径规则设置，就是充分利用这一点，来实现软件自动校对条目和原文）。

档案路径：是根据档案条目字段的数据内容，在指定“绝对路径”文件夹下产生的“相对路径”分层文件夹。是描述一件电子档案图像文件的无重复的唯一存在。

档案路径的理解，贯穿于整个档案数字化加工软件系统，是区别于其它文字材料扫描的主要原因。从档案扫描开始，到数据校对挂接，都离不开对档案路径概念的理解。因此，花较长的篇幅来描述档案路径的概念。

如果说归档号是目录级档案管理系统软件应用的必然条件，那么档案路径则是基于扫描

原文级采集、加工、转换合成乃至管理利用的必然条件。当然有人会说，我们用档号直接作为文件夹名（或用 64 位数字编码作为文件夹名绝对不重复），也可以实现数字化加工的全部过程，没错“坐家”说得对（但如果要考虑到采集、转换的效率，及人员操作出错误的几率，及人员改错的成本。那么几百万页的扫描加工下来，看看谁的成本最低就一目了然了）。从档案路径组合成档号，很容易且很准确。但要将档号拆分成档案路径，较费时且存在拆分错误（如 K1・2005.001 和 2005-Y-001—017）。

档案数字化加工软件，在保证加工质量的前提下，最大限度的降低人工成本，提高个体在单位时间内的生产效率，降低出错几率，减少改错成本，从而降低项目整体的加工成本、提高项目的利润率。这是本软件系统开发初衷，而非技术上的“炫酷”。

欢迎广大档案数字化加工团队用户，提出更好的建议、更有创意的新思路，我们可以洽谈与合作。

北京航星万博数据科技有限公司

2014-10-21