明清历史档案图像
数字化加工质量管理体系
研究报告

中国第一历史档案馆

目　录

1. 研究背景.....6

1.1 档案数字化工作现状.....6

1.2 明清档案数字化现状.....10

1.3 档案数字化质量管理研究现状.....13

1.3.1 国际文献理论研究现状.....13

1.3.2 国内数字化质量研究现状.....16

1.4 明清档案数字化存在的困难.....18

1.5 开展课题研究的重要意义.....20

1.5.1 社会效益.....21

1.5.2 可持续影响.....21

1.6 课题相关概念理解.....22

1.6.1 工作流技术及应用.....22

1.6.2 基于深度学习的 OCR 识别技术及应用.....23

1.6.3 全面质量管理理论.....23

1.6.4 明清档案数字化质量管理.....24

2. 研究目标和内容.....25

2.1 项目研究目标及任务.....25

2.2 课题研究方法及实施.....26

2.2.1 课题研究方法.....26

2.2.2 课题实施基础.....28

2.2.3 课题实施步骤 30
2.2.4 推广应用及完善 31
3.明清档案数字化质量管理体系概述 33
3.1 设计思路 33
3.2 主要成果 35
3.3 成果特点 36
3.3.1 体系特点 36
3.3.2 技术创新点 38
4.明清档案数字化外包项目工作模式介绍 42
4.1 明清档案数字化外包基本工作原则 42
4.2 明清档案数字化外包项目工作流程 43
4.2.1 前期准备 43
4.2.2 加工监管 51
4.2.3 检验接收 61
5.明清档案数字化质量管理体系建设 66
5.1 明清档案数字化质量管理之项目管理 66
5.1.1 组织管理 66
5.1.2 人员管理 67
5.1.3 制度管理 69
5.1.4 设备及软件管理 70
5.2 数字化加工各环节质量管理及技术应用 71

5.2.1 数字化加工环节工作流技术应用 71
5.2.2 数字化前处理环节质量管理 79
5.2.3 数字图像扫描环节质量管理 88
5.2.4 图像质量检查环节质量管理 95
5.3 质量管理体系风险控制 100
5.3.1 档案安全控制 100
5.3.2 数据安全控制 105
6. 课题成果推广应用范围 108
6.1 课题特点概述 108
6.2 课题可推广应用范围 110
7. 总结 112

作为近代古文献四大发现之一，明清档案内容极其丰富，价值弥足珍贵。其在编史修志、学术研究、古建修缮、文化传播、影视创作、对外文化交流等方面发挥着巨大作用，是研究明清历史的第一手资料，也是中华民族历史文化遗产的重要组成部分，有些档案已被列入联合国教科文组织《世界记忆名录》。

不同于现代文书档案，明清历史档案文种复杂，形制多样，同时，由于清朝本身、帝国主义入侵，北洋军阀和国民党时期政权更迭、战乱破坏等一系列原因，档案损毁情况严重，半数以上档案存在不同程度的残破、褶皱、虫蛀、霉变、粘连等情况。为此，开展明清档案数字化，最大限度地减少档案原件使用，保护档案实体，解决档案保护与利用矛盾，成为明清档案事业发展的重中之重。

早在上个世纪80、90年代，国外发达国家档案及古籍文献数字化工作已系统开展，相比较而言，我国档案数字化工作起步较晚，明清档案数字化工作发展尤其缓慢。如今，除台湾地区做过明清档案数字化相关工作外，其他各处相对较少，且已经开展档案数字化工作的明清历史档案数量较少，仅包括清代部分文献、地方志、族谱和图书杂志报纸等，目前尚未形成一个权威的、有代表性的明清档案数字化加工管理体系及标准。

2011年5月6日，中国第一历史档案馆（简称“一史馆”）启动了大规模档案整理及数字化工作。“一史馆”是专门保管明清

两代中央国家机关和皇室档案的中央级国家档案馆。馆藏档案74个全宗，1000余万件（册），其中明代档案3600余件（册）。其浩瀚的明清历史档案资源及复杂的档案状况，具有广泛的代表性，课题组以此为契机，尝试将现代数字化相关技术引入到明清历史档案数字化外包工作中，基于全面质量管理，建立一套数字化加工质量管理体系，以促进我国明清档案数字化科学发展，为明清档案数字资源建设提供助力。

1．研究背景

1.1 档案数字化工作现状

1、档案数字化工作发展历程

档案数字化工作是随着计算机技术、网络信息技术、数据库技术及多媒体技术发展而不断发展的。上世纪80年代，计算机在办公领域中的应用，推动了档案数字化工作的开端。1985年，我国发布了第一个档案著录国家标准《档案著录规则》（GB3792.5-85），计算机应用技术与档案管理正式建立密切关系，但受网络信息技术、多媒体技术和数据库技术所限，计算机在档案行业中的应用仅限于简单的目录著录，进行单机检索、统计、简易编目自动标引等。在这一时期，基本上是手工录入档案原文，与目录进行挂接，制作著录卡等，工作效率较低，局限性较大，但计算机应用技术的引入，使得传统的档案管理开始向现代化管理转型，是档案管理行业迈出的重要一步。

上世纪90年代到本世纪初，随着计算机的普及和网络信息技术的快速发展，档案数字化逐步走上了快车道。2000年，王刚同志提出了全国档案信息化建设任务的要求，随后国家档案局组织有关部门对档案信息化建设开展研究，提出了全面推进档案信息化建设战略部署，制定了《全国档案事业发展“十五”计划》。计划中提出：“加快现有档案的数字化进程，在北京、天津等地开展档案工作应用数字化和网络技术的试点。”在这一阶段，档案数字化理论研究和实践探索开展较多，出现了大量针对档案应用开发

的专业档案管理系统，档案数字化加工、流转、利用及控制过程基本形成，并开始提出全文数字化的概念，但全国层面来看，档案数字化的效率和进程还是比较缓慢的。

进入新世纪，计算机技术、数据库技术、多媒体技术以及存储技术的迅猛发展，为档案数字化建设高速发展提供了有利条件。2002 年，国家档案局发布《全国档案信息化建设实施纲要》，全国档案信息化建设进入全面推进阶段。2005 年，国家档案局发布《档案事业发展“十一五”规划》，将档案信息化作为新时期档案事业发展的一项重要工作，提出了“以为党和国家中心工作和各项建设事业有效服务为目标，以国家档案资源建设为核心，以档案信息化建设为重点”的指导思想，并提出“根据‘统一领导、标准先行、利用优先、分布实施’的原则，有序推进传统载体档案数字化进程”。在档案局一系列政策标准的支持和引导下，各地档案部门纷纷开展不同程度的数字化工作探索。档案数字化工作迅速发展，规模不断扩大，档案数字化规范日渐成熟，档案数字化加工市场逐渐形成，档案数字化对象不断丰富，从纸质档案扩展到照片、图像、录音、缩微胶片等各种载体档案数字化。2006 年，国务院信息化工作办公室和国家档案局联合进行“档案信息资源开发利用试点工作”，档案数字化作为重要内容之一被列入试点范围；2008 年，国家档案局在中国科学院召开了“中央国家机关档案数字化和整理现场会”，杨冬权局长强调“档案数字化工作促进了档案工作由传统管理模式项现代管理模式转变，由传统利用模

式向现代利用模式转变，不但可以便捷全面的提供档案信息，而且可以有效的保护档案原件，确保档案实体安全，应该积极开展，大力推进，以此为抓手，全面提升档案工作的水平”；2010 年，国家档案局正式提出建立档案安全保密体系，档案数字化成为安全保密体系建设的重要内容，档案数字化作为一项档案基础业务的地位得到了确立。

2．当前档案数字化工作现状

随着数字化工作的不断深入，档案部门在数字化建设中逐渐形成了一套完整、成熟的管理方法，国家也相继出台了一系列数字化加工标准规范，如《纸质档案数字化技术规范》、等。虽然各地数字化建设的思路和方法各有特点，但基本思路和做法大同小异，标准和手段基本一致。具体呈现以下特点：

1．数字化加工环节不断细化。随着档案数字化实践的开展，档案数字化加工的环节越来越精细，逐渐形成了以档案整理、档案扫描、图像处理、图像存储、数据挂接、数据验收、数据备份、数字化成果管理等环节组成的一整套数字化加工流程，并且每个环节都有相应的标准和要求，从流程控制到质量监督都能够通过计算机进行过程管理，数字化的效率和质量得到很大提高。

2．数字化标准不断提高。在早期的数字化实践中，数字化成品的标准主要以满足网络利用为出发点，以纸质档案数字化为例，图像基本上采用黑白 200dpi 扫描，个别地方采用灰度扫描。随着存储成本的下降和现代信息技术的发展，越来越多的档案部门在

开展数字化工作时，除了满足基本利用需求外，数字化的精度不断提高，彩色扫描大量使用，数字化成品的质量显著提高。

3. 数字化生产日趋社会化。当前，档案数字化规模越来越大，由于受人员、设备、规模、经验和管理等条件限制，数字化质量和效率难以得到保证。为此，数字化生产由档案部门自主加工转为了社会化服务，除涉密部门外，如今开展数字化生产基本首选社会化专业机构服务。

近年来，各级档案部门认真贯彻落实国家档案数字化相关工作部署，中央统战部、审计署等一些中央机关单位率先开展大规模档案数字化工作。据 2012 年底数据统计，全国副省级以上档案馆全部开展传统载体档案数字化，已数字化档案占馆藏总量比率大幅上升。北京市、上海市、山东省、江苏省、浙江省等各地依据各自不同的特点和经济条件采取不同的选择。比如，北京市的“全面数字化”战略，提出将个别（实物、会计凭证）除外的全部馆藏实现数字化；长春市档案馆则根据现实需要和客观条件，坚持“有所为，有所不为”，有重点、分层次的进行数字化，按照利用目的、频率、年代、所占比重等分类排队，精选馆藏进行数字化，提出“现用现扫”“以用定扫”和“常用先扫”等行之有效的工作原则。

整体来看，我国档案数字化工作多数已纳入各省市档案工作规划，工作定位已从“方便利用”向“替代原件”转化，对数字化范围有较清晰的鉴定原则，具备较好的规范基础及灵活的工作

模式，数字化工作已普遍采取文件分级管理方式、划分工作环节，通过制定工作制度严格控制数字化项目管理及成果质量。但也存在质量管理急需统一标准，缺少系统而具体的研究。

1.2 明清档案数字化现状

据调查统计，现存于我国大陆、台湾及世界各地的明清档案约有 2000 多万件，主要保存于中国第一历史档案馆，其次台北故宫文献馆、中央研究院保存了一部分，辽宁、四川、山东、西藏、黑龙江等各地档案馆和博物馆、图书馆中也保存了一些明清地方政权、家族和民间档案。其他单位如国家图书馆、故宫博物院图书馆、中国科学院图书馆、中国国家博物馆、中央民族大学图书馆等收藏有一定量的满文档案。此外，日本、美国、俄罗斯、德国、英国、澳大利亚等国家都保存了一定数量的满文档案文献，虽然这些档案文献与我国保存的明清档案相比较，数量十分有限的，但作为明清档案文献的重要组成部分，其价值和作用却是不可忽视的。

由于国外数字化发展起步较早，上世纪 90 年代前后，流存于国外的明清档案基本完成数字化，并开始提供网络化服务，其中日本京都大学图书馆建制的中国清代民国公私文书数据库最具成就。该数据库收录了京都大学法学部旧日本法史研究室所藏康熙至民国年间的 295 件中国公私文书的图像数据。内容包括田地、房屋、鱼池等典卖关系文书；租佃关系文书；所有权确认官给文

书等。

国内方面，台湾地区自1998年推出“数位博物馆专案计划”后，陆续进行了“数位典藏国家型科技计划”“数位典藏与数位学习国家型计划”等，涉及文化、学术、经济、教育、外交、社会等方面。其中，数位典藏国家型科技计划开始于2002年，是一个人文与科技并重的数位典藏计划，目的是将台湾地区重要的文献（含古籍）、文物典藏数位化，着重对精选出的最具代表性的文化资产进行数位化工作，并建立相应标准与规范，开发出包括“档案”“金石拓片”“善本古籍”在内的16个主题的数位典藏项目。

以台湾“中央研究院-历史语言研究所”（以下简称“史语所”）为例，其珍藏的内阁大库明清两朝档案约31万多件，其中明代档案约有4000多件，多数为清代档案，其1996年开展了明清档案的数字化工作，将所藏内阁大库档案进行了数字化。史语所所藏明清档案数量较少，加上开展数字化较早，其主要是由本单位人员针对档案形制分别采取扫描和拍照两种方式来完成的。扫描图像分辨率为典藏级300dpi，利用级降阶转存为150dpi或72dpi。其基本流程为整理、数字扫描（拍照）、图像校对（修正）、数字图像制作（包括接图、嵌入水印等）。后期，其以扩展更多珍贵典藏的数字化、创造更多元的数字资源为目标，并结合内容与技术专家创新发展，将数字资源再造为数字知识，推展科普应用、教育应用与学术加值应用。2013年，其制定了《史语所学术创新数

字深耕计划》，并以此推动数字化工作长期发展。一方面仍持续进行珍贵典藏数字化、数据库内容与功能增建等工作，以丰富数字人文学的基础建设。另一方面，利用现有成果结合信息科技，进行一源多用，跨库整合，让数据库由单一数据库单方面供用户提取信息，进一步演化为具有标注、分析、探索功能，并成为能与各数据库横向链接的查询系统，以发掘材料间的内部连结与脉络，进而找出值得深入研究的主题。

台湾“国立故宫博物院”收录清代宫中奏折及军机处档案折件等约15万件，军机处折件约19万件，内容涵盖清代国政大事、国家政策、军事外交、文化习俗等。目前其已将数字化档案建成数据库，并能够提供相关内容的检索，检索结果以标题索引和原版影像呈现。

而大陆地区则起步较晚，已经开展档案数字化工作的明清历史档案数量较少，已知的有四川省档案局实施的“国家重点档案抢救工程”、“国家清史纂修工程”等项目，已抢救重点档案超45万卷，其中“清代南部县衙门档案”、“咸丰朝巴县档案整理3万件”和“清代档案图片3000张”，上述三个项目已入选国家清史纂修工程。

辽宁省档案馆所藏明代档案，主要是辽东都司及其所属卫、所的档案。年代从洪武到崇祯， 延续200多年，其中以嘉靖、万历两朝居多。还有少量系明兵部和山东备倭都司的，总共1081卷。

所藏清代档案包括满文老档、实录、圣训、玉牒等。但更多部分则是属于清代盛京地方档案，计28个全宗，约16万卷，主要为旗务档和行政官署文书档。其自20世纪初就开始了明清档案的缩微拍照工作，目前正在开展数字化扫描过程中。

另，黑龙江省档案馆在重点档案抢救保护计划实施中，现已完成4万余卷清代档案的数字化前处理工作，完成3万余卷档案的缩微复制、修裱加固计划。而其余各地数字化较少，仅包括清代部分文献、地方志、族谱和图书杂志报纸等。

作为馆藏明清档案最多的档案馆，一史馆自1973年开始，先后采用缩微复制、数码拍照、数字化扫描等方式进行明清历史档案数字化工作。2011年，一史馆启动了五年档案整理数字化项目，开始大规模明清档案数字化外包工作，明清档案数字化工作进入快车道。但由于明清档案种类繁多，形制多样，目前尚未形成一个权威的、有代表性的海量明清档案图像数字化加工体系及标准。

1.3 档案数字化质量管理研究现状

1.3.1 国际文献理论研究现状

在国外，档案数字化质量管理多以专题项目或研究发布的报告、指南为主要形式。

在美国，其2004年发布的《档案材料数字化的技术指导方针》，就列举了数字化影像捕获、最小元数据、文件格式、文件命名、文件存储和质量控制等方面的一些技术上的要求。内容非常详细，

可以用来对数字化过程中产生的各种数据指标进行评测，规范数字化产品质量，以期达到不同使用等级的不同质量要求。《美国国家档案与文件署1571号文件档案存储标准》，这个标准主要是针对美国国家档案与文件署的各个档案处置场所和永久保存场所设定在建筑结构、建筑物防水性、制热、通风、空调系统、防火防水、光源、虫害控制和档案库安全等方面的要求。在数字化过程中要时刻考虑到档案原件的安全。进行数字化工作的场所和暂存档案原件的场所必须符合标准涉及的质量要求。此外，2007年由美国发起的美国联邦机构数字化指南动议（FADGI，Federal Agencies Digitization Guidelines Initiative）发布的《数字成像框架指南》（Digital Imaging Framework）等文件都较为系统而具体的对数字化质量管理提出相关建议。

澳大利亚在2010年制定了《信息与文献—文件数字化实施指南》（ISO/TR13028-2010），该文件为机构在业务流程中的文件与存量文件的数字化项目提供了实践指南，从数字化过程规划、管理方面提出了多项要求。包括：第一，所有的数字化项目都应该制定规范的政策和程序，以确保数字副本的真实性和完整性；第二，为维护数字副本的真实性和完整性，尤其是保证是原载体档案的真实拷贝，要记录对数字图像的各种操作，包括裁剪、模糊、去斑等；第三应尽可能的捕获和保存各类元数据；第四是应详细记录复制设备与复制过程、复制的政策和程序、质量管理记录、测

试结果等。

澳大利亚国家档案馆在2011年4月颁布的《数字化存量实体档案》指导规范。该规范主要用于指导数字化项目的计划和开启。该规范比较详细地提出了数字化过程所应遵循的技术标准和规范。主要内容包括：（1）数字化的原因，包括节省空间、与业务系统的融合、更好地利用以及保护文件；（2）数字化项目进行的前期计划，包括了解文件的特性、数字化的分类、大型数字化醒目介绍、业务案例介绍、项目管理、质量管理和品质保证、数字化设备、机构内数字化与众包、知识产权等方面的规定；（3）文件的相关决策，包括文件的等级、数字化缩微胶片、封装数字化文件、数字化文件的管理数字化文件的存储、数字化保管、数字化文件的源文件的处置、源文件的销毁或迁移、元数据要求等：（4）数字化过程，包括源文件的准备工作、其他格式的处理、数字化过程中文件的处理、部分被数字化的源文件、管理文件衍生品、安全等级划分的相关资料的处理；（5）相关的技术标准等。

日本国家档案馆—国立公文书馆，收藏有以《日本国宪法》原件为代表的日本政府各行政机关移交的行政档案以及古代、现代其他文书档案1003 万卷(册)，古籍、古代文献资料、地图、绘画等54 万5 千余册(幅)，如诏书、条约、敕令、阁令等原件。它既是日本的国家级档案管理中心，也是面向社会各界广泛提供档案利用的中心。2004 年，日本国立公文书馆发表《独立行政法人

国立公文书馆档案数字化建设推进纲要》，明确提出了其数字化建设的工作目标。对于列入重要文化遗产的文书档案和绘画等超大型的档案资料，将现存的彩色负片逐步转换为高清晰的数字化图像档案信息；完成古籍、古文书档案中所附插图的数字化工作。

此外，一些档案数字化政策，如美国《NARA 指令-促进利用的数字化活动》，加拿大 LAC《内部（档案数字化标准)》，澳大利亚《通用处置授权-适用于复制、转化或迁移的源文件》等，也部分涉及了数字化技术质量控制，尤其是数字化格式选择、压缩方法和存储的国际标准和国家标准。还有一些国际标准化组织（ISO）制定的数字化实施过程的原则、要求等，较为典型的是 ISO/TR 13028:2010 信息与文献—档案数字化实施指南（Information and documentation -Implementation guidelines for digitization of records)。

1.3.2 国内数字化质量研究现状

在国内，台湾地区早就颁布了《资料数字化与命名原则草案》《数字数据委外制作需求规范》等一系列数字化加工标准。近年来，随着全国各地档案数字化外包业务的蓬勃开展，国家档案局和各地也纷纷在档案数字化管理方面出台了相关的政策、规范和相关标准。如国家和行业标准《纸质档案数字化技术规范》、《文献档案资料数字化工作导则》以及《数字档案馆建设指南》等，地方标准如《北京市档案数字化工作规范》、《四川省资料数字化

标准》等。但是，由于资源匹配的不平衡，各个地区、行业数字化工作开展不平衡，存在要求不协调、标准不统一的现象。

2008 年 11 月，国家档案局局长杨冬权在中央和国家机关档案整理和数字化现场会的讲话中指出:“要加强基础工作和全程控制，保证档案数字化质量。”“数字化工作流程复杂，任何一个环节出现差错，都将影响整个数字化工作的质量。因此要做好档案数字化的全程控制，注意全面的质量检查，加强数据的全方位质量控制，对全过程进行管理，保证档案数字化的质量。”国家档案局从档案工作科学发展的角度，提出了档案数字化工作的整体质量要求。

2014 年，中国档案学会档案信息化技术委员会开展了“关于档案数字化质量控制体系的研究”专项课题，从规范档案数字化工作流程、强化档案数字化产品质量管控、提高档案信息资源质量、保证档案信息安全的视角，提出了档案数字化工作“6+1”质量控制框架体系；并将 ISO9000 质量管理体系的理念融入档案数字化工作，并贯穿于档案数字化工作全过程。该研究报告诠释了档案数字化概念与质量管理理论的关系，介绍了档案质量控制的方法、建设原则，从组织体系、流程体系、项目文档体系、规范体系、安全体系及监理体系等几个方面做了详细的介绍，提出了质量控制体系建立的相关要素，对全国档案数字化质量管理工作的规范化、标准化管理提供了参考。

1.4明清档案数字化存在的困难

与现行档案相比，明清档案具有这样几个特点，一是内容丰富，价值珍贵；二是文种复杂，形制多样；三是年代久远，状况较差。其前后跨越550多年，内容涉及明清两代政治、经济、军事等社会发展各个方面，不但文种较多，且档案形制不尽相同，包括折件、簿册、舆图以及实物等；另有部分档案或体量较大（如幅面超过45cm×60cm，部分簿册厚达15-20cm），或粘连较为严重（多件档案粘连在一起），或档案中有夹条、贴条等，状况较为复杂。加上政权更迭、战乱破坏等一系列原因，档案损毁情况严重，较大部分档案存在不同程度的残破、褶皱、虫蛀、霉变、粘连等情况，数字化加工难度较大，在数字化过程中，主要存在以下困难。

1．缺乏完善的历史档案数字化标准。

截至现阶段，我国共形成了 12 项历史档案数字化标准，其中包括 3 项明清档案、4 项民国档案、5 项革命历史档案，上述均是建立在著录细则、主题标引细则、分类标引细则及其机读数据交换格式的基础之上。我国政府自 2005 年出台《明清档案目录中心数据采集标准》、《明清档案机读目录数据交换格式》后，其相应的历史档案数字化标准便停滞不前。除此之外，我国相关部门尚未针对于不同时期制定出历史档案编号标准，从而给予历史档案数字化标准统一工作造成了极大的负面影响。

2．各地档案馆缺乏相关技术和人力资源

历史档案数字化对现代信息数字化处理技术以及技术人员需求既巨大又迫切，仅凭档案馆自身力量难以满足这一需求。专业公司有技术实力雄厚的专业化人力资源队伍，在长期实践中积累了处理具体技术性问题的丰富经验，能够为档案数字化处理提供高质量和高效率、高效益的服务。鉴于此，一些档案馆开始尝试将自身力量同社会力量结合，实行业务外包，选择有一定经验和能力的专业公司来协助完成历史档案数字化处理工作。

3.数字化过程中档案原件容易受到损伤。

众所周知，安全是档案数字化外包工作一条不可逾越的红线，尤其是明清档案数字化加工，由于价值珍贵，文物特征明显，档案原件安全更是重中之重。一方面，档案从出库、数字化加工到最后的入库，环节较多，接触人员范围较广，易出现安全防护隐患；另一方面，由于档案年代久远，残损状况较重，数字化过程中，一旦档案原件的操作手法、加工方式稍有不慎，都会造成档案损伤，加速档案老化。如何做好档案原件不丢、不坏是明清档案数字化外包管理中的一个难点。保护档案原件安全，抢救性保留档案信息，成为数字化首要原则与目标。

4．明清档案数字化成品质量不易保证。

（1）档案信息容易漏扫。为保存明清档案原貌，减少对档案原件的损坏，档案业务部门在数字化过程中一般不编号、不拆装，

不破坏档案原始状况。此种情况下，一些簿册类、幅面较多的折件类档案，极易出现档案信息漏扫；同时，明清档案书写文字除汉字外还包括满、蒙、藏、察哈台、拉丁、英、俄、日、法、德等多种文字，尤以满文居多，满文人才的匮乏也导致加工过程中对图像检查难度进一步加大，更容易出现重复扫描、漏扫等情况。此外，由于档案粘连状况复杂、夹页、夹条情况较多，给数字化过程质量管理带来很大困难。

（2）成品质量不易稳定。从明清档案特点来看，一方面，不同形制和文种档案数字化加工要求不同，很难采用同一个数字化加工标准来规范操作；另一方面，明清档案色彩丰富，对色彩还原要求较高，从而对设备使用和质量检查提出了更高的要求。从人员管理角度来看，因为加工人员流动性比较大，整体业务素质及水平容易起伏，在标准规范的理解上不容易做到统一，从而导致明清档案的数字化加工质量不易保持稳定。同时，明清档案较高的残损比例一定程度上也增加了加工难度。如何确保明清档案质量统一、稳定成为数字化加工过程中的一大难题。

1.5 开展课题研究的重要意义

总体来说，本课题研究将带来良好的社会效益和极为深远的可持续影响力，不但对档案馆未来的发展奠定了坚实基础，而且为其他档案行业的数字化质量管理提供了参考。

1.5.1 社会效益

1．课题研究对明清档案原件的延年保护作用显著。

开展明清档案数字化质量管理课题，实现纸质档案向数字档案转化的科学化管理，加工出高质量的档案数字化图像，将更加有效地支持图像著录、电子利用和查阅，最大限度地减少档案原件的使用，保护了档案实体。在数字化质量管理过程中，对粘连、破损、残缺、褶皱等各类档案开展的展平、除霉、修复等处理工作，使部分破损档案实现“起死回生”，修补缺损、去除污渍、消除霉变，使档案实体得到更好保护，有效地延长了档案实体寿命。

2．课题研究对提高明清档案资源管理水平作用显著。

课题研究制定的数字化相关技术标准规范和管理制度等标准性文件，建立起明清档案数字图像标准，形成了统一的电子秩序目录，全面记录档案实际状况，为改进明清档案资源著录方式，提高著录工作效率奠定了基础，为提高明清档案资源数字图像查询利用信息化水平，推动明清档案信息资源建设提供了有利条件。

1.5.2 可持续影响

1．对明清档案数字化项目目标实现的促进作用显著。

目前，仅一史馆数字化项目尚有数百万件档案尚未完成数字化加工，该课题研究成果将进一步完善数字化加工管理机制、制度体系、质量体系，积累丰富的数字化加工经验，有效降低数字化加工过程风险，培养一批数字化加工项目管理人才，将进一步

加快明清档案数字化加工进程，促进明清档案数字化项目目标的快速实现。

2．课题研究填补了明清档案数字化理论研究空白，具有较强的推广利用价值。

课题结合明清历史档案数字化的特殊要求，对明清历史档案数字化加工项目管理的理论、方法、手段等进行有效总结，具备一定的学术研究价值。同时能够很好的起到填补领域空白、引领行业趋势、规范行业标准的效果，对于明清档案数字化事业具有重要意义。课题顺应信息化时代要求，将进一步探索大规模图像数字化加工中计算机信息处理技术的应用及融合，具备较强的现实指导意义，能够对国内外其它历史档案数字化工作起到了借鉴作用。

1.6 课题相关概念理解

1.6.1 工作流技术及应用

根据国际工作流管理联盟(Workflow Management Coalition，WFMC) 的定义，工作流是指一类能够完全自动执行的经营过程，根据一系列过程规则，将文档、信息或任务在不同的执行者之间进行传递与执行。工作流所要解决的主要问题是：为了实现某个业务目标，利用计算机在多个参与者之间按某种预定规则自动传递文档、信息或者任务。工作流技术已广泛应用到各个行业。在档案行业，为了提高档案的利用率，降低档案管理的成本，实现

无纸化自动办公，推进档案信息化建设，推进档案管理现代化进行，在档案行业信息化过程中引入了工作流技术。

1.6.2 基于深度学习的 OCR 识别技术及应用

OCR （Optical Character Recognition，光学字符识别）是指使用电子设备（例如扫描仪或数码相机）检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程；即，针对印刷体字符，采用光学的方式将纸质文档中的文字转换成为黑白点阵的图像文件，并通过识别软件将图像中的文字转换成文本格式，供文字处理软件进一步编辑加工的技术。

近年来，深度学习技术发展迅速，在各个领域得到广泛使用。深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。将 OCR 识别技术与深度学习技术相结合，就可根据客户具体需求，为各行各业在终端智能文字录入应用领域提供全方面的开发与定制，从而方便地将 OCR 技术集成到他们的应用流程和设备中。

1.6.3 全面质量管理理论

质量管理是指确定质量方针、目标和职责，并通过质量体系中的质量策划、控制、保证和改进来使其实现的全部活动。质量管理的发展大致经历了 3 个阶段。质量检验阶段、统计质量控制

阶段以及全面质量管理阶段。质量管理发展到全面质量管理，是质量管理工作的一大进步，相对于前两个阶段，其更加适应现代化大生产对质量管理整体性、综合性的客观要求，从过去限于局部性的管理进一步走向全面性、系统性的管理。

欲有效开展质量管理，必须设计、建立、实施和保持质量管理体系。质量管理体系是组织内部建立的、为实现质量目标所必需的、系统的质量管理模式，是组织的一项战略决策。它将资源与过程结合，以过程管理方法进行的系统管理，根据企业特点选用若干体系要素加以组合，一般包括与管理活动、资源提供、产品实现以及测量、分析与改进活动相关的过程组成。其具有符合性、唯一性、系统性、全面有效性、预防性、动态性、持续受控7个特性。

1.6.4 明清档案数字化质量管理

明清档案数字化质量管理，是将质量管理及质量管理体系的思路融入明清档案数字化加工外包管理中，通过设定质量目标，建立组织，采取工作流控制的方式，规范档案数字化产品的生产全过程，实现抢救保护明清历史档案，以档案图像替代原件提供利用的目的。是针对明清档案的特点及其数字化过程中遇到的困难所采取的一种质量管理方法。其体现了质量管理理论中工序控制的重要思想，也同样具备质量管理体系的7个特性，报告中我们将一一加以体现。

2. 研究目标和内容

2.1 项目研究目标及任务

本课题主要研究在海量数据加工外包条件下，如何建立一套完备的图像数字化加工质量管理体系。主要是立足明清档案实体状况，依据一史馆对明清档案数字化项目具体要求，借鉴国内外同行业先进经验和发展趋势，从制度、标准、管理、技术层面上研究并建立一整套适应明清档案图像数字化加工的质量管理体系。具体如下：

1．建立基于质量管理的明清档案数字化外包工作模式。

依托中国第一历史档案馆数字化加工外包项目，借助全面质量管理的理论思维，梳理明清档案数字化加工外包工作业务流程，完善数字化外包组织管理、制度管理、人员管理、设备管理等外包项目管理方式，理顺明清档案数字化加工前处理、扫描、质检等各环节工作，最终建立一种基于质量管理的明清档案数字化外包工作模式。

2．建立相关制度标准规范体系。

针对明清档案特点，针对性建立数字化外包项目管理、技术标准、操作要求等一系列明清档案数字化加工制度规范体系，促进明清档案数字化工作规范化、科学化，为明清档案数字化外包工作提供制度依据，为其它行业、类型档案数字化提供参考。

3．探索计算机技术应用及融合。

基于明清档案数字化加工业务流程，引入工作流技术、表格登录识别技术、OCR 识别技术、图像自动处理技术等，设计、完善数字化外包加工软件，探索计算机技术在明清档案数字化加工过程质量管理中的应用。

4．设计整体质量管理目标，满足现实需要。

针对一史馆明清档案数字化工作要求，制定具体的数字化成品质量管理目标，满足客户需要，促进项目目标实现。课题设计数字化主要质量管理目标如下：图像参数正确；扫描顺序正确；图像清晰、完整、无变形；无倾斜、异物、折角、压字、透字、彩线、彩晕等；验收质量标准为：目录录入差错率、图像技术质量差错率不超过千分之一，目录与图像挂接差错率、图像信息漏扫率不超过万分之一。

2.2 课题研究方法及实施

2.2.1 课题研究方法

课题采用调查研究、实验研究、分析归纳等理论与实践相结合的研究方法。以现代档案数字化加工技术、软件的特点及应用为基础，针对明清历史档案的特点及加工过程中遇到的问题，开展重点难点技术研究和科学实验，通过不断的优化及实践检验，最终建立了一套完备的明清档案数字化加工质量管理体系。

1．调查研究。在研究过程中，通过行业咨询、实地考察等方式，先后到中国科学院档案处、中国航天科技集团档案馆、中央

档案馆、盛赞公司等单位，就数字化档案的寿命、真实性（还原性）、成本等问题进行广泛调研，形成了档案数字化思路；同时，向相关数字化服务公司了解数字化加工软件功能、应用环境及应用范围，结合一史馆明清历史档案数字化加工需求以及在数字化加工过程中存在问题，寻求解决方案。

2．实验研究。

通过与汉王科技股份有限公司合作，充分发挥其技术研发优势，将相关技术优化并引进明清历史档案数字化加工。通过一史馆数字化加工项目应用，实际检验应用效果，并根据实际结果不断调整、改进，以实现最终目标。期间，多次对表格登录识别技术、图像匹配技术及OCR识别技术在目录录入、图像查重、图像自动处理等方面进行数据测试，并不断加以改进。

3．分析归纳。

课题研究过程中，项目组对相关数据进行归纳、整理，对研究与实验结果进行分析、总结， 最终形成图像数字化加工质量管理研究报告。同时，课题组成员基于课题研究制定了一史馆数字化加工项目手册，还在馆内、馆外撰写、发表了《浅谈明清档案数字化图像加工的若干思考》《档案数字化加工外包过程质量控制研究》《谈明清档案数字化外包管理难点与对策》《明清档案数字化外包安全管理初探》等多篇学术文章，成果显著。

2.2.2 课题实施基础

1．组织及人员基础。

课题组自立项开始，先后吸纳明清行业专家、项目管理人才以及公司技术专家 10 余人专项开展工作，全部具有本科及以上学历，在明清档案专业和档案数字化领域各有专长，长期参与明清档案数字化项目，具有丰富的管理与实践经验。中国第一历史档案馆馆长及班子成员高度重视，处室各级领导、汉王科技股份有限公司及技术服务团队，公司项目组大力支持，为课题组研究实验提供了人员、场地和技术上的充分支持，为课题的最终完成奠定了坚实的基础。

图 1-数字化加工场地

2．规模最大的明清档案数字化加工工区。

课题组所在中国第一历史档案馆数字化工区，建筑面积 300 多平米的数字化加工工区，场地通风条件良好，全部铺设防静电地胶，拥有高速平板扫描仪、A2 扫描仪、A1 大型扫描仪等专业数字化设备 50 余台，配套计算机 100 余台，拥有完善的配电、照明、空调、监控系统，

图 2-数字化加工场地

是国内规模最大的明清档案数字化加工基地。这也是课题组能够顺利开展课题研究，积累大量明清档案数字化实践经验的基础。

3．课题实施软硬件基础。

课题组依托的一史馆数字化加工外包项目具有完善的网络运维服务和数据管理平台，软硬件基于国家运维服务标准，制定一史馆特有的数据业务模式的数据存储策略和数据库备份策略方案，建立了 PB 级的数据中心，包含底层硬件存储平台构建及上层档案信息化管理平台的开发，选用了 oracle11g 数据库，采用 Oracle Data Guard 技术对数据库进行备份；对于非结构化数据（TIFF，JPEG2000、JPEG 三种格式的图像）采取集中管理模式，以多介质多套存储形式数据安全备份。通过硬件平台、档案信息化管理平台软件、归档备份软件，并配合数据库，利用相应的策略对数据进行管理和保护。

软硬件环境如下：

软硬件名称	规格型号	数量	备　注
HP服务器	Xeon x7542，8核，8GB内存	4	包含windows server2008企业版操作系统
HP计算机	Core i5 650，内存，1TB硬盘	82	包含windows 7专业版操作
A3幅面零边距扫描仪	精益A300	6台	含6台备机
A3幅面快速平板扫描仪	精益A380	44台	
存储系统	EMC,日立	各一套	FC-SAN, SAS盘+SATA盘
网络环境	主干线：2条千兆光纤链路聚合；交换机到计算机：千兆网线。		
数字化加工软件	专业网络版	2套	
杀毒软件	网络版		

2.2.3 课题实施步骤

课题共分三个阶段实施，具体如下：

第一阶段（2013 年 4 月），主要任务为明确质量管理体系整体思路及预期目标。课题组首先对项目团队成员工作职责进行了划分，明确了项目目标和工作计划。随后，课题组梳理了一史馆明清历史档案数字化加工流程、工作需求及遇到的问题，对现有数字化加工软件功能进行了测试，确定了质量管理体系整体思路及预期实现目标。

第二阶段（2013 年 5 月-7 月底），本阶段主要任务为建立质量管理体系框架，搭建质量管理体系运行平台，为项目实施的关键步骤。本阶段，项目组进一步明确并改进了现有质量管理基本制度，图像质量标准及操作规范。在原有工作流程上，项目组一方面根据一史馆档案数字化工作要求，对整个数字化加工流程进行重新梳理，包括改进档案扫描前处理工作，完善并规范扫描前档案基本数据，调整质检及验收工作流程。另一方面，针对目录录入效率低、差错率高以及图像漏扫的问题。引进表格登陆识别技术及图像匹配技术，在目录录入及图像检查上加入自动处理程序。通过上述工作，项目组搭建了一个质量管理体系运行平台，并通过制度规范等加强对人的控制，初步建立起一史馆档案数字化质量管理体系。

第三阶段（2013 年 8 月-12 月），本阶段为实践检验及总结优

化阶段。主要任务为验证及完善质量管理体系有效性，即通过质量目标控制，开展大规模明清档案数字化加工实践。建立定期沟通会晤机制，及时发现并处理管理体系存在的问题，不断完善体系组成部分。在一史馆数字化加工工作标准下，本阶段项目组共完成 348253 条目录的著录工作，3395735 画幅的图像查重及抽检工作，图像数据挂接目录 348253 条，对本项目质量管理体系及应用的关键技术进行了测试，根据测试结果不断优化，最终达到项目预期目标。最终测试结果如下：

测试项		数量	单位	技术参数/要求	准确率
目录录入	表格登陆识别录入	348253	条	录入速度 30 条/分钟 识别区域±偏差≤2mm	>99.9%
图像处理	查重识别	3395735	画幅	图像相似度≤85%	100%
图像抽检	TIFF 格式	256798	画幅	300dpi，24 位真彩，图像清晰、完整、无变形，反映档案原貌。无倾斜、异物、折角、压字、透字等，JPG 图像压缩率 50%。 随机抽检，卷级覆盖率＞90%。	99.9%～99.99%
	JPG 格式	724702	画幅		99.9%～99.99%
图像漏扫		3395735	画幅	漏扫率＜1/10000	无漏扫
数据挂接		348253	条	加 md5 码	>99.9%

在上述工作基础上，项目组对各项工作进行了整理及总结，形成了最终研究报告。

2.2.4 推广应用及完善

基于明清历史档案数字化加工质量管理体系，2014 年至今，

我们完成了1140多万画幅的数字化图像扫描、抽检以及154万余条档案目录的元数据录入及挂接工作。扫描图像质量差错率、数据挂接差错率、数据录入差错率以及图像漏扫率均达到了课题组设计目标，满足了一史馆数字化加工质量标准要求，同时保证了数字化加工过程中档案实体的绝对安全，为一史馆明清历史档案数字化加工项目优质高效的完成提供了坚实的保证。同时，课题组通过汉王科技公司在北京市社保局档案组织机构代码和身份证号识别、银行和医院的单据识别等项目中，就借鉴了该体系中表格登陆识别技术应用，取得了良好的效果。