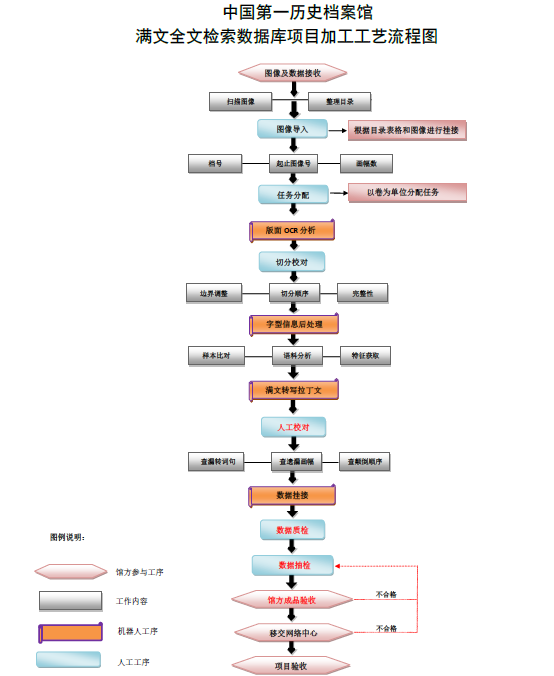
针对此次中国第一历史档案馆满文朱批奏折全文检索数据库加工涉密采购项目，我们仔细阅读了采购文件。满文朱批奏折全文检索数据库加工档案预计58万画幅，转换成拉丁字母约5500万字。项目工作期限200个工作日，在次期间必须完成满文朱批奏折全文检索数据库加工加工以及验收工作。

我们按照本次满文档案全文检索加工项目的招标要求，设计了项目实施流程的总体框架。根据公司既有的档案全文检索项目加工经验，结合项目实际情况，将本项目实施流程分为八个部分，共15道工序，其中质检工序有3道，分别是人工校对、数据质检、数据抽检。我们采用汉王OCR自动识别技术、拉丁字母转满文编码转换技术嵌入程序后台，各工序环环相扣，前后校验的方式进行操作和质检，既提高了工作效率，也提高了工作的质量。

图像数据接收、图像导入、任务分配、版面OCR分析、切分校对、错字标记、错字修改、字型信息后处理、满文转拉丁文、人工校对、数据挂接、数据抽检、馆方成品验收、移交网络中心、项目验收等15个工序。如下图所示：

****

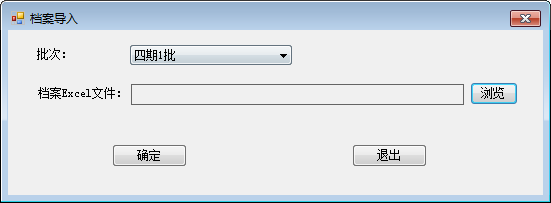
**工艺流程示意图**

该流程的有点在于“工艺流程清晰，责任明确，利于分工协作”，同时由于“专人做专事（即员工固定工序岗位）”，其工作效率和质量都会显著提高。

### 2.1图像及数据接收

图像及数据接收是流程的第一道工序，在下载数据之前，公司管理人员根据项目生产计划，然后跟监管方沟通，经过监管方协调沟通，确定好下载时间，下载完成后结合整理目录对图像数据进行完整性核对。

注意事项：下载数据过程中，如出现数据下载问题，请及时将问题反馈到监管方安全管理工作人员，如有需要，积极配合监管方处理问题，保证生产计划的顺利进行。



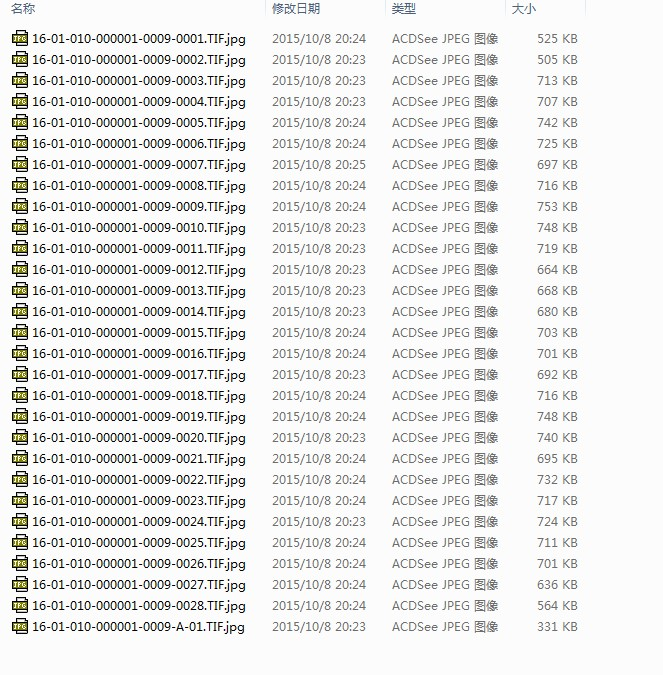
**下载管理功能示意图**

### 2.2图像导入

下载完成的档案图像，在监管方和公司安全员的共同监督下，组织人员分批次通过馆方网络服务器下载档案图像和目录表格。

注意事项：根据整理目录表格以及下载的档案图像进行挂接，挂接内容有案卷级、文件级、档号、卷内文件起止号、件数等。

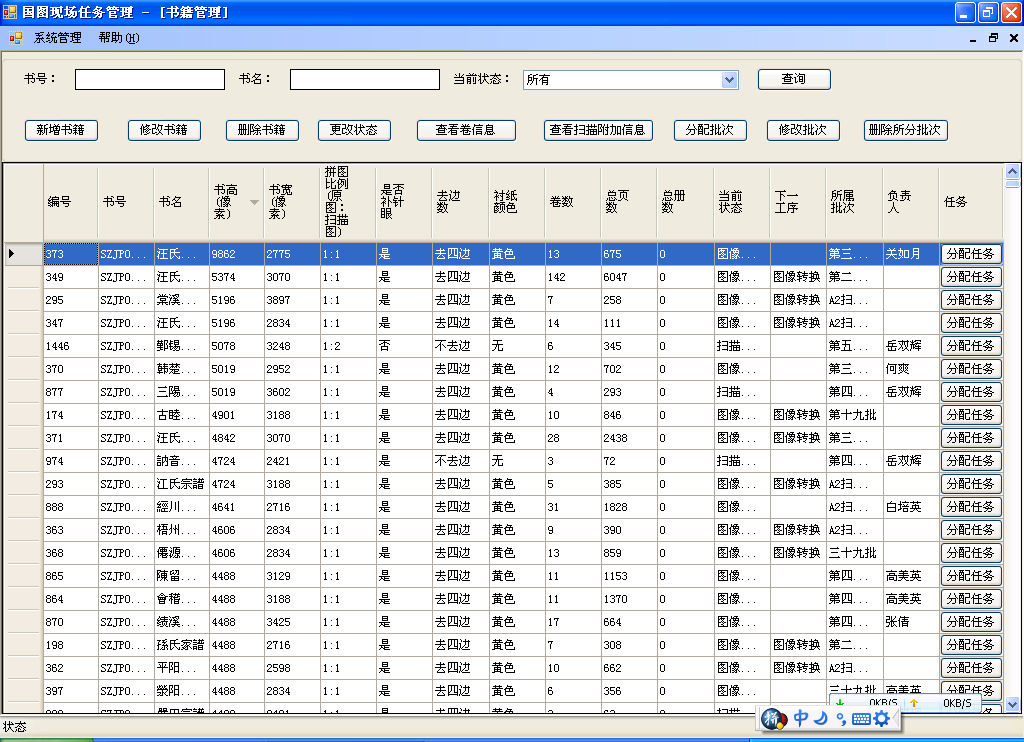
程序可根据相关目录数据和对应的满文扫描图像进行目录结构自动挂接，为档案自动识别做好准备工作，如下图：



**目录结构自动挂接**

### 2.3 任务分配

通过服务器将档案图像和目录数据下载到本地，然后将馆方提供的档案数据导入自行设计的软件系统中，按照数据生产需要进行流程分配。我们采用实名制登陆数据管理端，通过管理端进行任务分配，由专门管理人员将以案卷为单位的任务分配给工作人员。



**任务分配功能示意图**

### 2.4 版面OCR分析

数据下载后，以案卷级为单位，然后将每一个案卷分解到文件级。

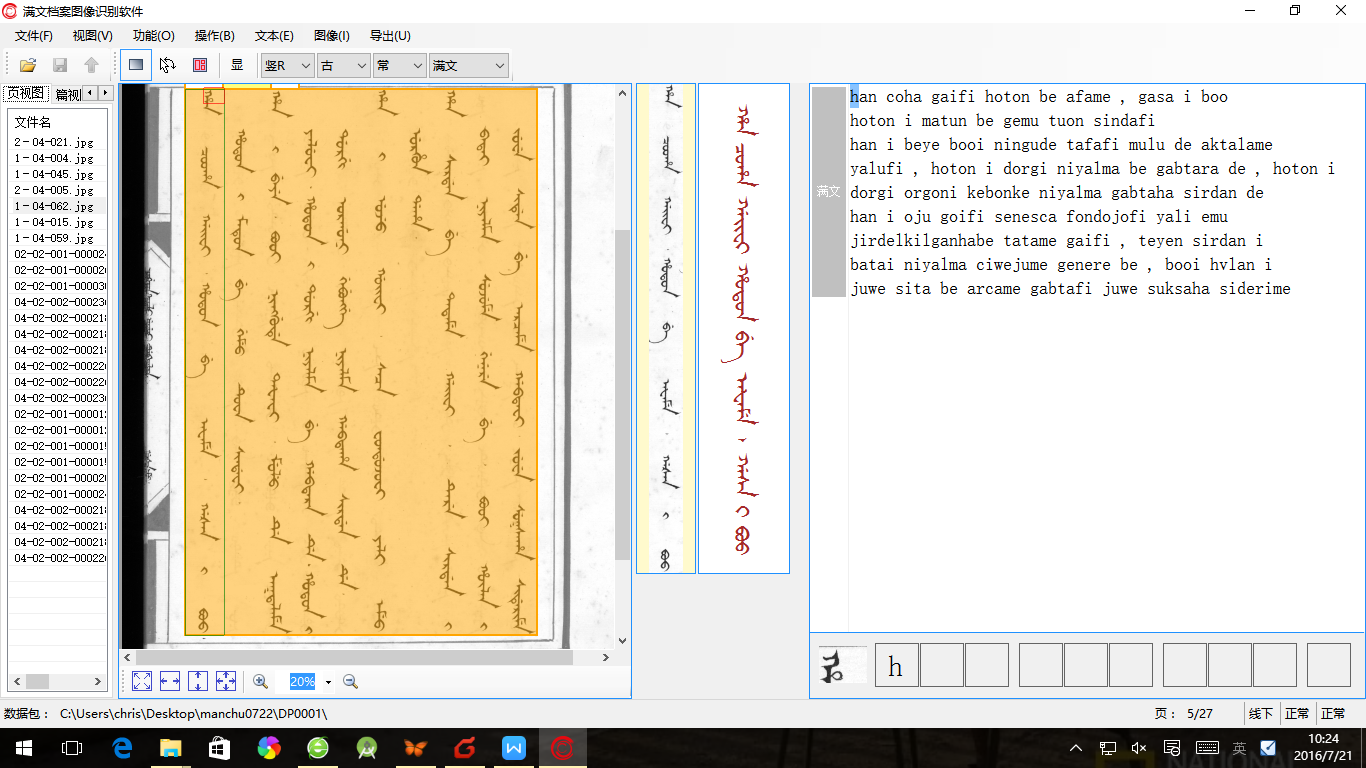
依据项目要求，程序设置完成后，满文档案经过扫描，图像处理得到黑白二值图像。经过了前期的去噪处理，得到的二值图像只包含黑色的文字部分和白色的背景部分。利用汉王成熟的版面分析技术，定位到待识别的文字区域，每个区域只包含单栏文字信息。对单栏文字区域进行行切分，再根据单词之间的空白区域提取出单词图像。

应用OCR技术，在图像二值化图像基础上，对满文朱批奏折扫描图像进行自动识别，生成拉丁字母转写形式。识别结果保存两份，一份作为原始结果用于保存，另一份用作人工校对。本方案以音节识别为基础。采用如下技术路线：



**满文识别原理**

满文朱批奏折扫描图像识别软件，主要包括目录栏、原图像可视区、满文识别校对区、以及满文转拉丁识别校对区。可以加工通过满文档案图像识别软件自动生成的拉丁字母数据，将识别后的拉丁字母数据提供给数据加工者使用。软件应易学习、易操作，界面美观，菜单定义清楚，工序设定明确，可灵活的调整软件以适应生产流程。满文识别软件界面如下：



**满文识别版面分析软件示意图**

满文识别软件按照逐行校对，程序识别过程中，对识别可能错的音节、单词自动爆红提醒校对员，对爆红的内容着重检查，起到只能辅助人工的作用，不仅保证质量，还提高了效率。

可识别用相对规范的满文楷书和行书抄写档案的扫描图像，通过识别软件对图像的文本区定位，经过行分割，如对满文档案的音节切分和音节识别，经过语料获取。

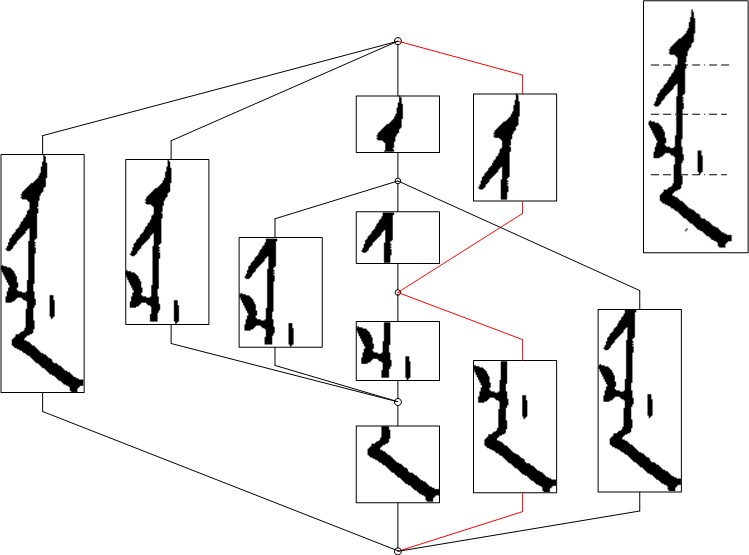
### 2.5 切分校对

文字图像中文字行所在的区域黑色信息较多，行间的空隙上黑色信息较少（主要文字的圈点）甚至没有，基于此特点进行分割的方法是行切分常用的有效方法，具有速度快，效果好的特点。但是跟中文等不同，满文字母高度，宽度极为不统一，文本行稍有倾斜，文字行间隙的特点就会不明显；而且，部分字符有主体部分和圈点部分组成，文字行间的空白间隙有可能是行边界，也可能是字符主体部分和圈点部分的空白。用自适应的行间阈值指导行选取得到文字行的框架，再用字符轮廓跟踪方法得到精确的行边界可以有效的解决这些问题。

满文单词具有字符连接的特点，这给音节切分带来很大的困难。本项目拟采用轮廓线极值点切分技术，将音节切分成更小的部件(称之为基元)。先提取单词主体部分的连通体，通过对连通体的左右轮廓线的进行分析，把基线位置的极值点作为候选切分点，得到候选切分位置。所得候选切分位置将通过后续的识别距离比较策略进行过滤，进而得到最终的音节切分位置。

|  |  |
| --- | --- |
| 原始图  (a)满文单词图像 | (b)满文单词轮廓图及切分线 |

**音节候选切分点示意图**



**分层构筑的动态规划算法示意图**

### 2.6字型信息后处理

#### 2.6.1语料分析

语料获取的主要目的是从大量图像中获取不同音节的多幅图像，并对这些图像进行自动标注，主要步骤为：

1）扫描样本，获取大量满文样本图像。

2）对满文图像进行行切分，得到单行的满文图像。

3）对单行的满文图像进一步切分，得到满文单词。

4）利用印刷体满文识别技术，将满文单词切分为音节序列，通过聚类算法归类相同音节，同时过滤掉切分错误的音节，并通过人工方式对归类后的音节进行标注，从而获得大量的已正确标注的满文音节。



**满文样本语料获取示意图**

#### 2.6.2音节识别特征提取

特征提取是字符识别很重要的组成部分，是为了选择区分度高的音节特征，便于识别软件区分不同的音节。同时，由于手写字符变化较大，需要采用鲁棒性比较强的特征。本项目采用音节的折线拟合信息作为特征进行识别。

在分类器学习阶段，提取样本音节的轮廓并进行折线逼近，将折线信息作为模板特征。样本音节为利用语料获取模块获得的大量的已正确标注的满文音节，模板即为用数据的方式描述样本音节的特征。

在特征提取阶段，计算待识别音节的轮廓，在轮廓线上进行等间距采样，拟合出等长的小线段，并记录其位置和方向作为特征，即待识别音节的数据表示方式。

在识别阶段，将提取特征后的待识音节，与事先训练好的模板特征进行匹配，使用多对一的匹配方法，即多个小折线特征与模板的长折线特征进行匹配。



**音节识别特征提取示意图**

目前，满文识别软件对行书音素识别准确率为92% ，单词识别准确率80%；楷书音素识别准确率为94%，单词识别准确率85%，满文经过ODR自动识别后，再经过人工校对、质检后，转写成拉丁字母的满文朱批奏折数据差错率不超过1%；满文朱批奏折原扫描图像及转写成的满文拉丁字母数据挂接准确率为100%；全文检索应用软件在规定条件下检索准确率为100%。

### 2.7满文转写拉丁字母

**1. 满文转拉丁文要求**

1）严格按照《满文档案转写拉丁字母标准》将满文朱批奏折、满文实录原图转写为拉丁字母。

2）转写成拉丁字母的满文朱批奏折数据与满文朱批奏折原图保持一致，转写成拉丁字母的满文实录数据与满文实录原图保持一致，不得出现漏转词句、遗漏画幅、颠倒顺序等情况。

3）转写成拉丁字母的满文朱批奏折数据及满文实录等数据不抬格。

4）遇到满文中夹杂少量非满文单词的，如：汉文、蒙文、藏文等。每个单词用一个“#”表示。

5）遇到残缺不全或模糊不清的单词，用“\*”代替，不可推断，不可只转写单词的一部分。

6）遇朱批 (含墨批和蓝批)部分不转写,每一处朱批在适当位置用一个“+”表示。如朱批文字、朱圈、朱点、朱杠等。

7）遇有含有夹签的档案图像，选择满文未被夹签遮挡且顺序靠前的档案图像转写，与之内容重复的满文档案图像不转写。

8）遇到拼写特殊的单词，或者拼写错误的单词要及时向一史馆项目组汇报，不可随意转写。

**2. 满文转写拉丁文**

我们按照《满文档案转写拉丁字母标准》，利用专用工具将满文朱批奏折原图转写为拉丁字母。该工具实现拉丁字母到满文以及满文到拉丁字母的转换功能，并具备以下特点：

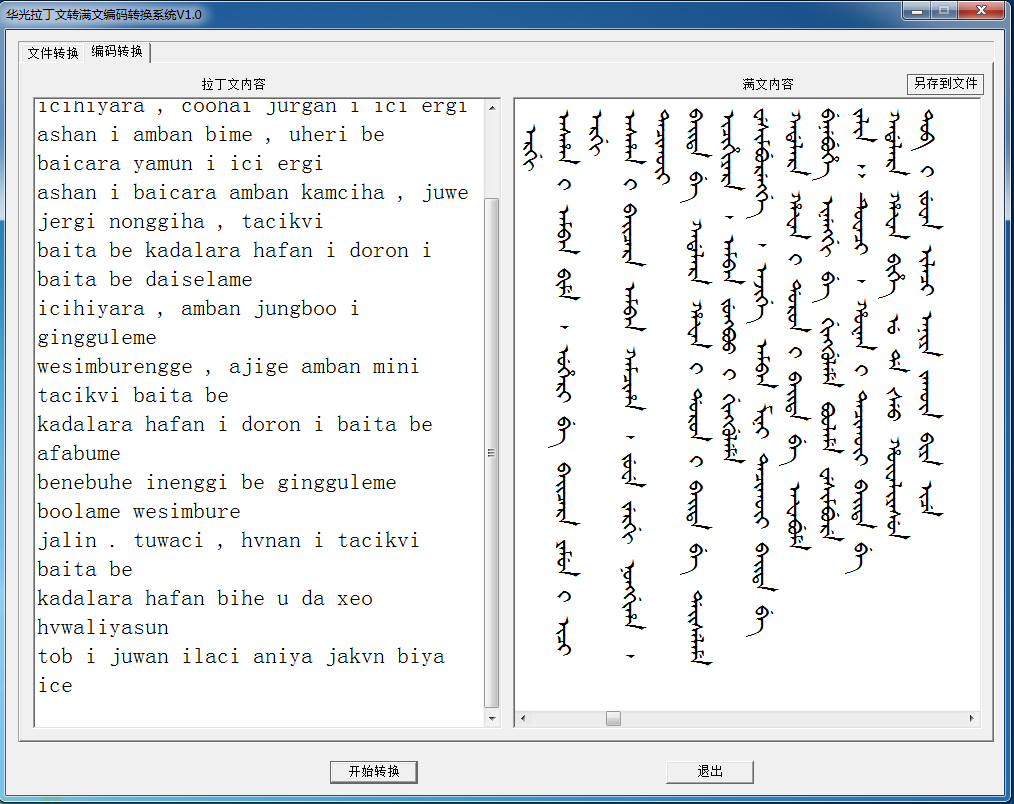
1）转换工具功能强大，用户可手动转换，也可以监控目录自动转换，拉丁字母转满文编码转换工具可实现文件转换，转换准确率为100%；

2）使用方便，系统提供文件转换和内容转换两个方式，拉丁字母转满文编码转换工具可实现内容转换，转换准确率为100%；

3）查看方便，系统提供满文竖排显示控件，编码转换工具可实现满文竖排显示转换结果一目了然；

4）安全可靠，自动转换原始文件存放到中转目录，保证文件安全；

转换工具可自由选择是拉丁字母转满文还是满文转拉丁字母。



**拉丁转写与满文编码转换软件**

### 2.8满文校对

根据项目实施的实际情况，满文识别校对软件同时具有原扫描图像、识别后的满文、以及转写的拉丁文，具有满族语言文学专业知识的工作人员根据每件档案的内容进行质检，按照原扫描图像，在满文可编辑区进行逐行校对。

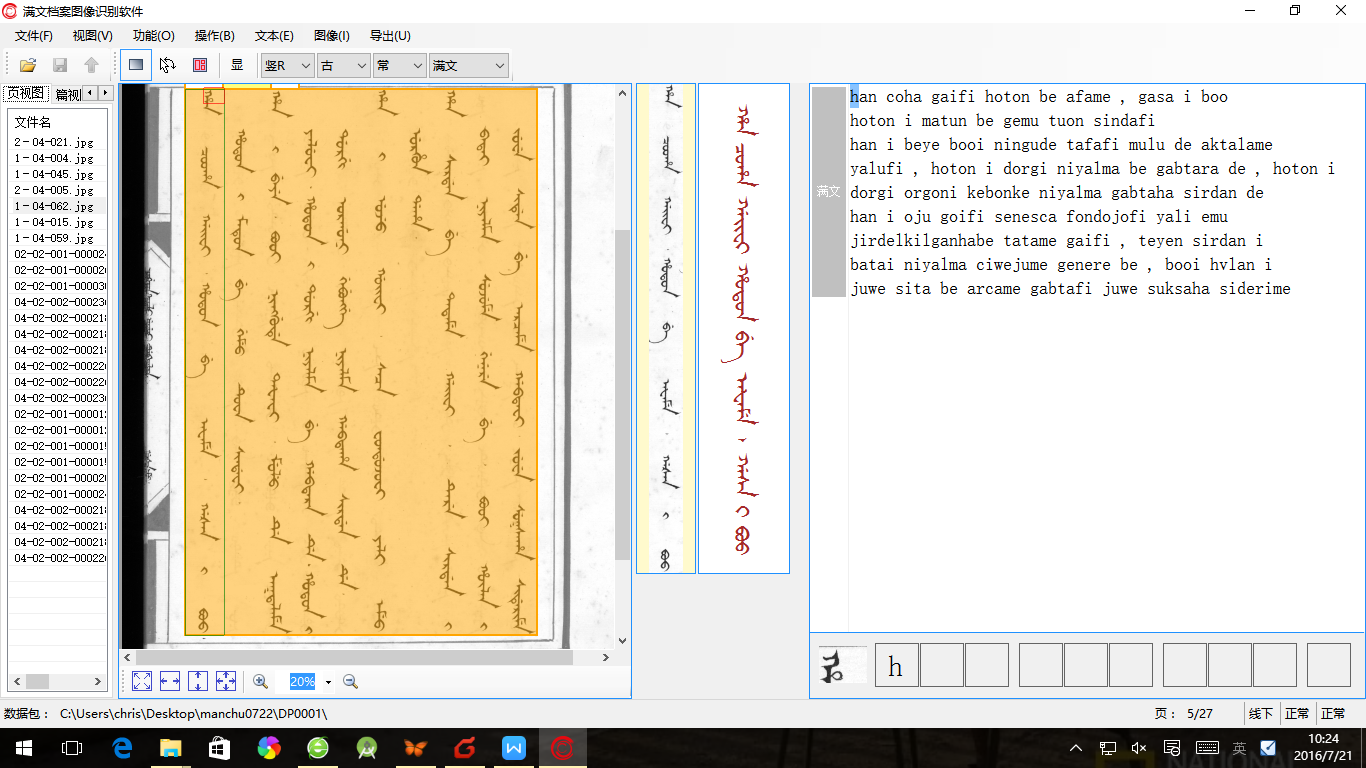
1）逐行校对功能。文字校对是按照档案原图像进行校对修改，它是一种逐行逐字地把识别文本与相应图像做对比的校对方式，特点是直观明了，可以利用上下文信息，软件应提供清晰的拉丁字母识别结果和档案原文图像，校对者发现识别生成的拉丁字母数据错误可以进行修改。

2）任务提示和任务列表。为了方便校对，系统把识别可信度差的文字以醒目颜色标出。

3）软件提供可追溯的生成数据与过程数据。并具有抽检比例、差错率、工作量等统计汇总与报表等功能。

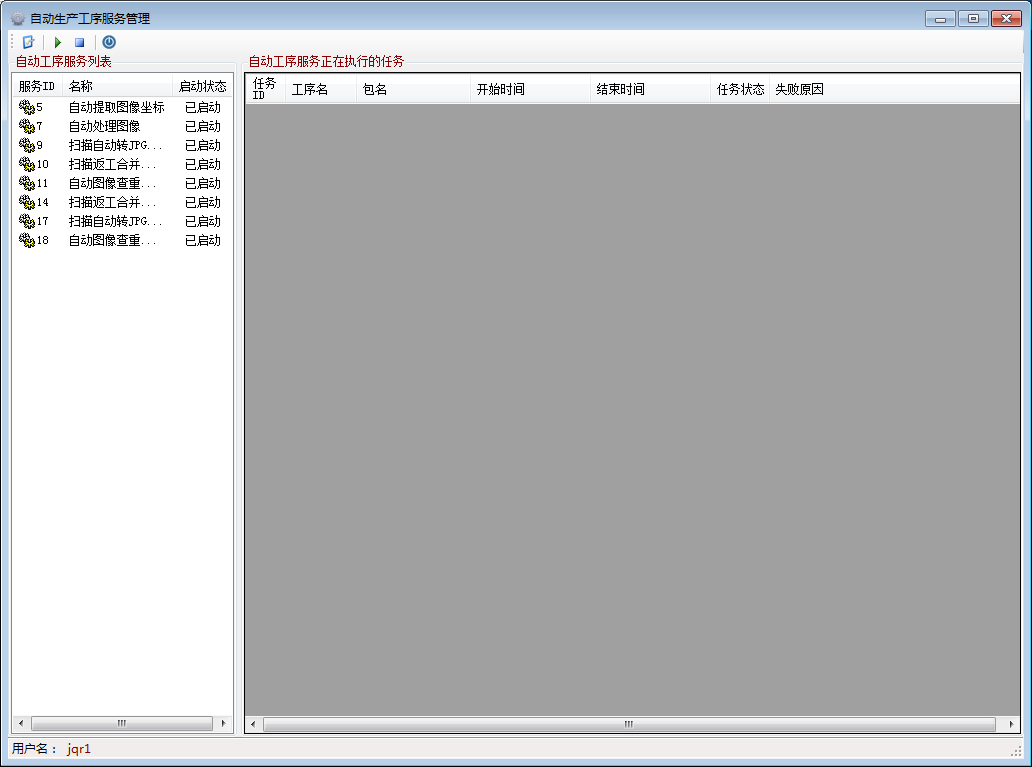
4）转换成拉丁字母的满文朱批奏折数据与满文朱批奏折原图保持一致，不得出现漏转词句、遗漏画幅、颠倒顺序等情况。转换成拉丁字母的满文朱批奏折数据不加标点，不抬格。

5）对于满文语种的专业性，对信息内容项不确定的，则有专门的满文专业人员进行分析、校对，如音节、单词错误，则应及时改正，然后将准确的内容进行拉丁文转写。



**校对功能示意图**

满文识别软件，除了OCR自动识别，我们还嵌入了程序目录自动挂接、满文与拉丁文自动进行转换等功能，如下图：

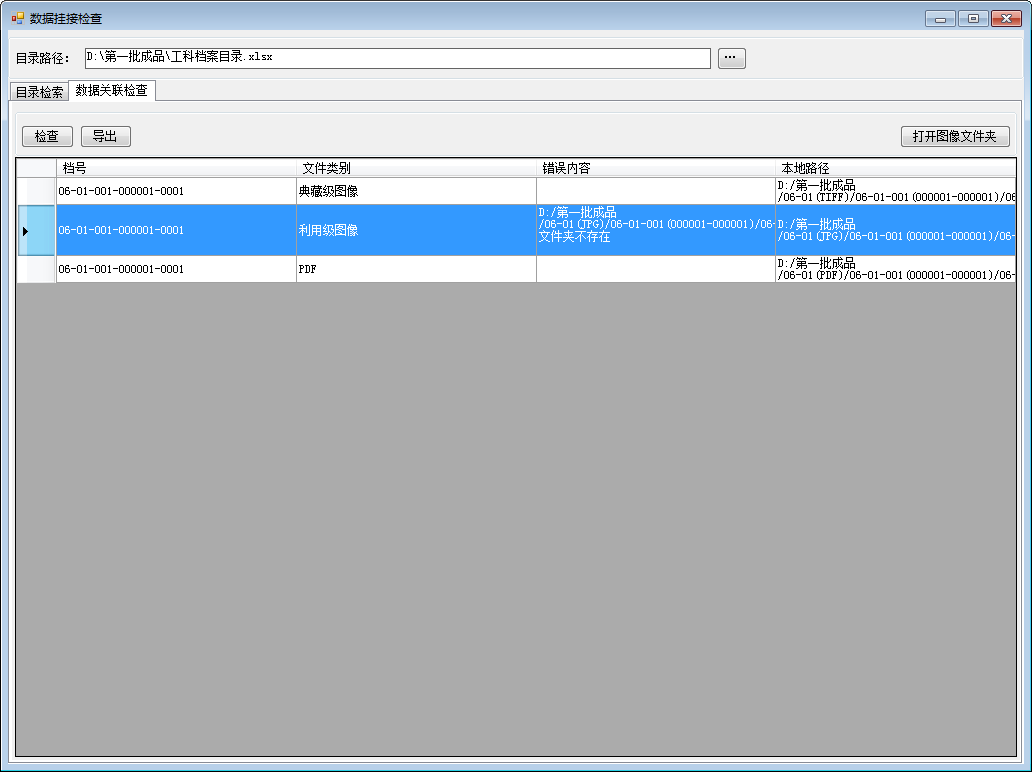


**程序自动转换功能示意图**

OCR自动识别模块，可以准确高效的识别文字。满文与拉丁文自动转换，可以严格按照《满文档案转写拉丁字母标准》转换。这些自动工序既提高了工作效率，又可以减少人为的失误，以确保工作任务按时完成，保证项目进度顺利进行。

### 2.9数据挂接

人工质检完成后，将转写成拉丁字母的满文朱批奏折数据与满文朱批奏折原扫描图及目录进行挂接，同时再次进行错误排查，逐项进行核对，发现错误及时更正，数据合格后准备质检。



**数据挂接总检功能示意图**

### 2.10数据抽检

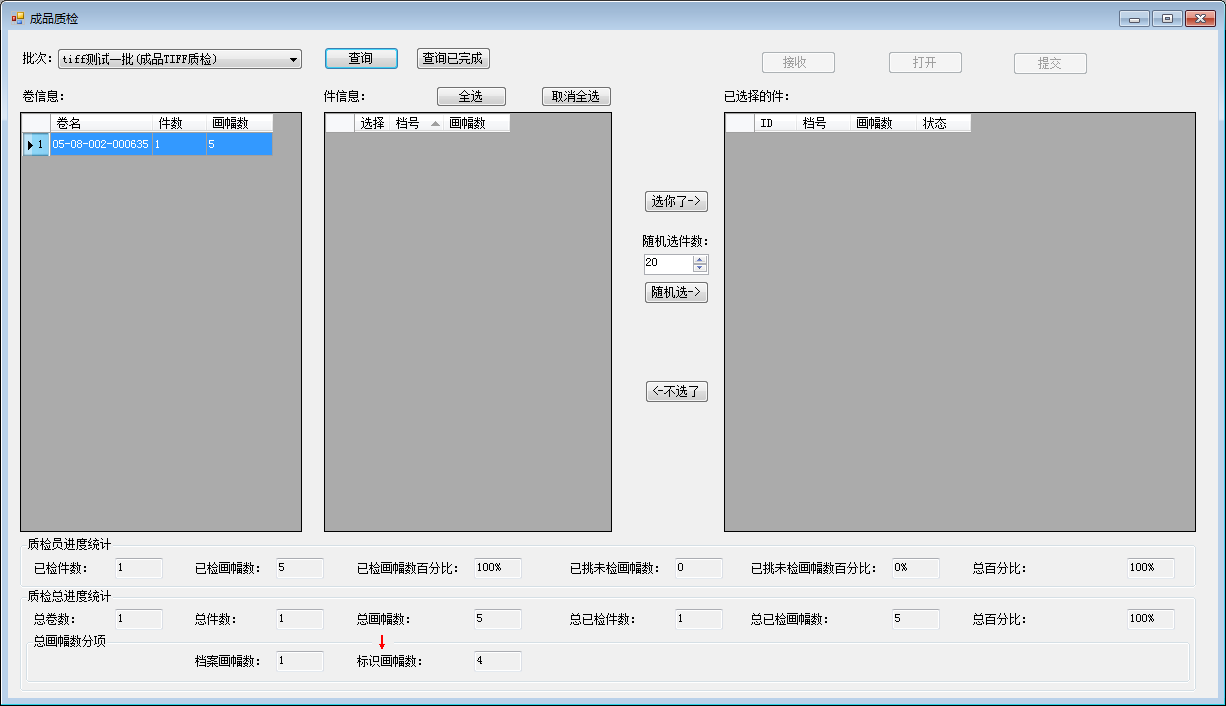
数据挂接总检无误后，程序会将所有数据按照馆方要求的结构化数据形式进行整合，然后对所有数据再次进行检查核对，我们公司项目组在提交给用户前对加工成品数据进行内部100%整体性的质量验收，保证数据完整性和规范性。若数据合格无误，则由工作人员在管理端分批次提交成品数据至中间临时服务器，则可准备提交。

在满文识别校对整个加工生产过程，质检贯穿于整个数据加工流程，除对每一个环节的质量进行检查核对，在成品提交前，项目组对项目的成果进行100%整体性的验收，保证数据完整性和规范性，避免出现漏转词句、遗漏画幅、颠倒顺序等问题。

### 2.11馆方数据验收

按照之前的项目经验，按批次将质量抽检后合格的成品数据提交给客户，客户对提交的成品数据进行质检。质检合格的成品及时上传至馆内服务器。

甲方按照招标文件中《中国第一历史档案馆满文档案全文检索数据库项目质量控制与要求》按提交的批次成果数据进行审核验收，质检抽查比率不应低于20，如果差错率超标，按批次退回返工直至合格。设计客户验收程序如下：



**成品验收功能示意图**

### 2.12移交网络中心

在满文识别、转写数据全部提交完毕的情况下，且数据经客户验收合格后，会将所有成品数据按照馆方要求进行上传，成品数据格式与馆方要求的结构化数据保持一致。提交按批次移交网络中心，提交全文检索成果数据。

### 2.13项目验收

在满文识别、转写数据全部提交完毕的情况下，可准备结项验收工作。

首先开始收集、整理项目加工过程表单，统计项目加工数据，包括总体数据和批次明细数据。然后及时总结档案特点和加工特点，及时撰写项目总结报告，并向客户申请结项并支付尾款。在验收会召开之前，需提前准备好验收材料和项目总结相关文案，并报请公司领导和客户审核通过。根据客户安排时间参加验收会，汇报项目加工情况，回答评委专家问题，完成验收。