人工智能技术在档案划控上的应用

（2023 年度国家档案局优秀科技成果二等奖）

一、成果简介：

国家档案局科技项目“人工智能技术在档案划控上的应用”(以下简称“本项目”)由安徽省档案馆与讯飞智元信息科技有限公司于 2018 年共同申报获批，项目批准编号 2018-X-028。项目于 2018 年 4 月启动，在项目科研过程中，安徽省档案馆负责规则制定与数据保障，讯飞智元信息科技有限公司负责系统设计与技术实现，在双方共同努力下，2020 年 4 月“人工智能技术在档案划控上的应用”项目成果“档案智能划控系统”正式发布。

利用人工智能，基于机器学习技术，以已“划控”开放审核数据为基础，智能学习“划控”规则，构建“划控”规则库和知识库，对档案数据进行智能“划控”分析，是解决档案开放审核工作量大，提高“划控”效率最有效的办法。

主要研究内容：

本项目采用人工智能技术，利用文本分类、语义分析的相关研究经验，对人工智能技术在档案划控中的应用进行研究。主要研究内容：

1.对档案本身内容语义的分析；

2.在对档案 OCR 识别和语义分析的基础上，建立档案数据分词算法，实现语句到短语的切分，识别档案中的短语、名词、命名实体等；

3.机器跟着档案划控领域专家进行学习如何划控，通过训练，提高智能划控的效率和精准度，保证划控的质量；

4.根据不同历史时期档案形成的历史背景，将涉密、涉政治事件、涉案、涉军、涉外、涉宗教、涉民族、涉边界、涉人事、涉诉讼、涉处分等问题的内容进行主题分析，提炼敏感词，积累生成敏感词库；

5.对档案多维分析建模，机器解读，档案原文语义分析及阅读理解，建立不同时期各个档案门类的词库，最终形成一个完整的划控知识库；

6.根据档案馆对开放审核相关规定和要求建立规则库；

7.根据划控词库建立专业文本分类识别引擎；

8.应用相关技术，结合档案鉴定划控流程规范，研究开发智能划控软件系统。

项目研究成果：

项目主要研究成果为集成文本分类识别引擎的专业档案智能划控软件系统及相关文档，对已有的档案数据（支持双层 PDF、双层 OFD、WORD、TXT 等）执行分析划控任务后，获取高可信的划控档案列表和机器给出的建议划控结果，支持人工抽检及划控结果修改。

机器对大量档案进行划控，对于划控成功的档案及划控错误后纠正的结果，机器进行自主学习，逐步提升划控准确率及效率，实现“机器划控+知识推荐+规则采集+流程审核”一体化的档案鉴定开放审核新方法。具体成果如下：

1.研发了一套机器智能划控引擎

2.构建了档案划控敏感词库，包括 2000 多个敏感词

3.构建了档案划控知识库，包括 5000 多条知识点

4.构建了档案划控规则库，包括 2000 多条规则

5.研发了一套智能开放审核平台

6.人工智能技术在档案划控上的应用相关软件著作权（2 份已获证，1 份申请中）

随着互联网技术的发展，人工智能、机器学习、深度学习等技术的不断发展，特别是基于卷积神经网络的文本分类技术的出现和发展，为人工智能在档案领域的应用，提供了广阔的应用前景。机器划控作为人类档案划控专家的助手，能够提供划控初步意见，有效辅助档案专家，作出划控判断，提高划控效率和准确率。人工智能技术在档案智能划控上的广泛应用，将会有效助推海量档案开放利用。

二、成果详细内容：

档案划控是对不同门类档案的开放权限和范围进行界定，本项目从计算机技术角度出发，将自然语言处理中的文本分类技术应用于档案开放权限范围分类工作。对经过 OCR 识别的数字化成果和原生电子档案进行自然语言处理、文本分类研究，并结合划控敏感词、划控知识点库、规则库，对档案是否开放和开放范围作出判断。本项目以提升档案划控的精确性、准确度为研究的重要技术指标，同时鉴于馆藏档案数量庞大、种类繁多，档案划控效率也是项目研究的性能指标。

本项目结合国家档案局有关划控规定和指导意见以及安徽省档案馆对划控的具体要求，对不同门类的档案和档案的不同时期，按照档案的划控标准分别划控。在业务上，一方面，选取民国时期相关全宗中已完成人工划控的档案进行机器划控学习，通过人机划控结果比对，提炼结果，修正模型。另一方面，选取不涉及敏感政治专题的全宗，对其中未划控的档案直接进行机器划控，并进行人工校核，同时修正模型。通

过不同门类不同年代的档案机器划控，提炼相关规则，达到在不同的档案门类中的应用。同时，考虑到主观性对于模型的影响，项目组通过提供尽可能多的档案数据进行模型的训练及演练，进行逐步完善。在数据训练与验证中，选取了安徽省总工会共 50369 件档案进行数据验证，其中 BERT 模型准确率达到 80％以上，符合项目效果预期。

1为控制，0为公开

题名	预测为公开	预测为控制	模型判断	人工标记
[关于准予许开文等人任巡教二团团员的指令、委任令]	7.97E-06	1.00E+00	1	1
安徽国民军事训练委员会公函[关于函发修正高中以上学校军事教育方案的公函]	0.99998367	1.64E-05	0	0
[关于呈报程德昌对国学有研究之名家调查表的代电]	0.99999464	5.42E-06	0	0
[安徽省政府关于蚌埠市筹备处呈送中心及国民学校专任校长一览表的代电]	0.99999356	6.48E-06	0	0
[关于储一贯请求豁免缴纳学膳杂费问题的批]	0.9999945	5.49E-06	0	0
[关于颁发私立学校规程问题的指令]	0.9999932	6.78E-06	0	0
[关于派员安徽省立霍邱初级农业职业学校监收麦子问题的电]	1.00E+00	5.78E-06	0	0
[关于安徽省立第二临时中学二十八年度第二学期员生昏表及证件的指令]	6.31E-06	0.9999937	1	1
[关于蚌埠教方设小学教材为伪教育部编订其课本内容荒谬给教育厅的电]	5.86E-06	0.99999416	1	1
[关于王同志辞职等情给余旭东的电]	5.79E-06	0.99999416	1	0
[关于特务教育业案付诸实施给教育厅的电]	6.55E-06	0.99999344	1	0
[关于职校修业期满教职员一览表及空白成绩证明书等问题的呈]	7.72E-01	0.2284068	0	0
[关于敌伪盘踞地点及敌化教育情形并深入破坏敌化教育机构等问题的函]	6.49E-06	1.00E+00	1	1
[关于派第四区专署科长黄振秋前往出席教育会议的电]	0.99999404	5.96E-06	0	0
[关于铜城县第四区青草镇中山民众学校后聘教员委令停聘其薪金不予补发的指令]	0.9999945	5.45E-06	0	0
[关于填送安徽省三十一学年度第二学期所属师资训练机关现况调查表的代电]	1.00E+00	6.17E-06	0	0
[关于上报伪小学概况及破坏敌化教育机构有效办法与困难点的呈]	6.14E-06	1.00E+00	1	1
[安徽省政府关于怀宁县政府转呈私立车彤小学艺友一览表准予备查的指令]	0.9999939	6.12E-06	0	0
[关于检送舒城县属乡小学朱子聪及龙王庙乡小学胡继禹履历表问题给安徽省政府的呈]	1.00E+00	5.57E-06	0	0
[关于宿松县政府呈送员生登记证件准予备查的指令]	1.57E-05	0.99998426	1	0
[关于安徽省立立煌师范学校呈报杜教主任干事资历表准予备查的指令]	1.54E-05	0.9999846	1	1
[关于五月份汤家沟敌情及大民会更新宣言的呈]	6.99E-06	0.99999297	1	1
[关于证明刘国超年龄及学籍问题的批]	2.26E-05	1.00E+00	1	0
[关于报告霍邱县教育科长余志明启程日期问题给省教育厅的电]	1.00E+00	5.95E-06	0	0

图　测试集验证样例

相关研究成果及技术研发内容展示如下：

1.档案划控业务系统流程设计

为了更好地满足档案专业领域的需求，机器需要跟随档案专业领域的专家学习划控知识。通过不断地训练和学习，机器可以提高划控的效率和精准度，从而保证划控的质量。

为了实现这一目标，我们需要建立专业文本分类识别引擎，引擎可以帮助机器快速准确地识别出不同类型的档案文件，从而提高划控的准确性。同时，我们还可以利用档案多维分析建模技术，对档案进行深入的分析和研究，以便更好地理解档案的内容和意义。此外，我们还需要研究机器解读档案原文的语义分析及阅读理解技术。通过这些技术，机器可以理解档案中的文字和信息，从而实现对档案的智能解读。这将有

助于提高档案管理工作的效率，同时也可以为档案开放审核流程规范提供有力的支持。

基于以上技术和应用需求，我们研究开发了智能开放审核平台。这个系统主要分为三个部分：档案划控辅助、知识树管理和档案划控规则管理。

档案划控辅助部分是为了帮助用户更高效地进行划控操作。通过提供一系列的辅助工具和方法，用户可以更方便地完成划控任务，提高工作效率。知识树管理部分则是为了方便用户对档案知识进行管理和查询。通过构建知识树结构，用户可以快速查找到所需的档案资料，同时也可以将新的知识和信息添加到知识树中，形成一个完善的档案知识体系。

最后，档案划控规则管理部分则是为了确保划控过程的规范性和一致性。通过对划控规则进行统一管理和控制，我们可以确保所有的划控操作都符合规定的标准和要求，从而保证划控质量。

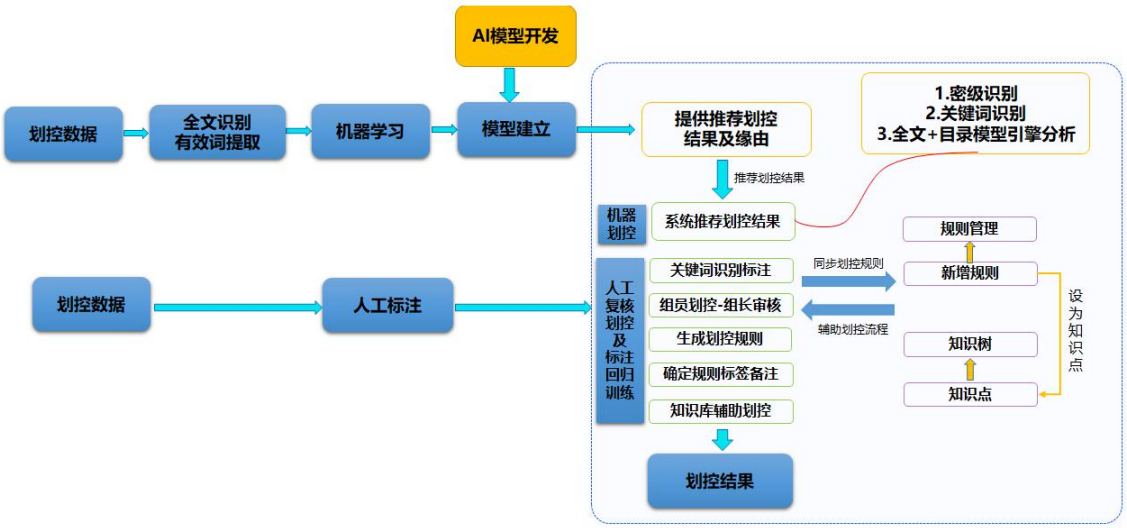

图　档案划控业务流程

2.对档案识别和语义分析，建立档案数据分词方法

系统导入待划控的档案数据，在对档案识别和语义分析的基础上，建立档案数据分词方法，实现语句到短句的切分，识别档案中的短语，名字、命名实体等，对识别出来的信息进行统一管理。

图　档案划控系统-待划控档案页面

3.机器解读，建立智能划控算法引擎

根据单件档案的划控结果，系统采集档案目录数据、提炼原文内容，完成档案分类、文本内容分词；按照标密文件划控、关键词划控。

通过分词分类算法及关联学习训练，形成划控算法引擎，对数据进行划控，结果输出为机器划控结论。可持续对关键数据进行标注，逐步提高划控的精准度，保证划控的质量，对后续划控操作提供划控辅助。

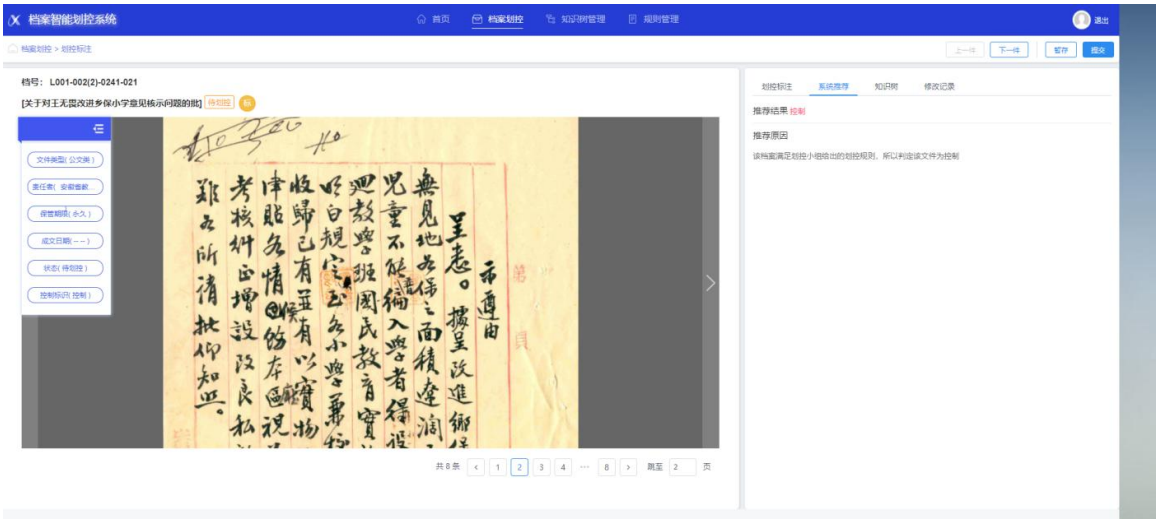

图　档案划控系统-标注详情页面

4.根据不同类型档案标注，积累生成划控知识点

根据不同历史时期档案形成的历史背景，将涉密、涉政治事件、涉案、涉军、涉外、涉宗教、涉民族、涉边界、涉人事、涉诉讼、涉处分等问题的内容进行主题分析，提炼敏感词、关联句，积累生成知识点。建立不同时期、不同档案门类、不同控制缘由的关键词、关键规则库，最终形成一个完整的划控知识点库。

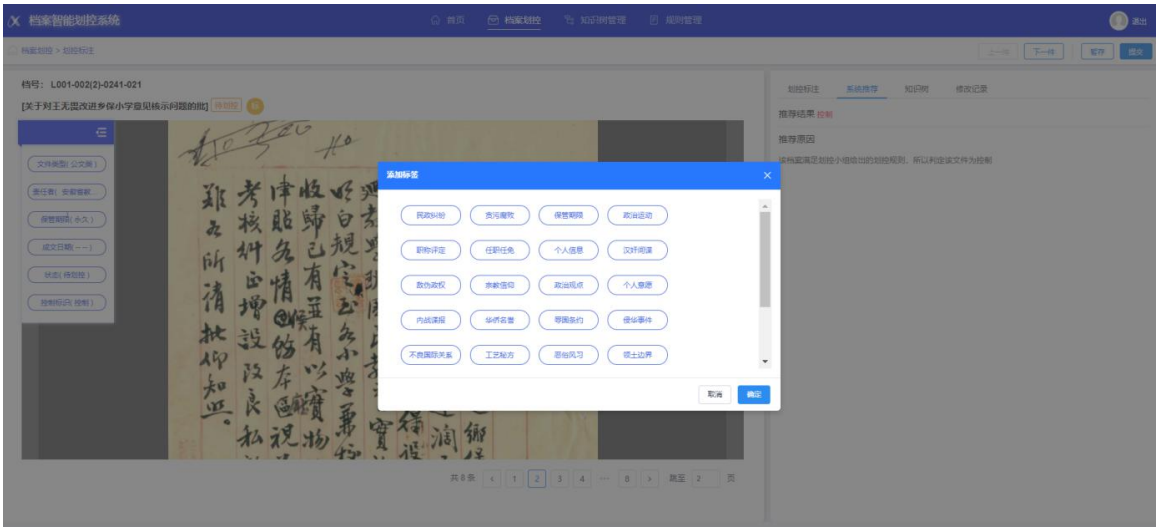

图　档案划控系统-标签管理

5.建立规则库，形成知识体系，辅助人工划控

通过标注数据，收集规则后，可根据文种类型、文件标签对规则进行自动分类及标签化处理，并赋予划控结果属性。通过规则可辅助提高划控准确度；同时规则可形成知识点并关联关键词、标记划控结果属性形成知识树，通过专家录入/接入第三方知识数据库，形成档案关键知识信息总库。

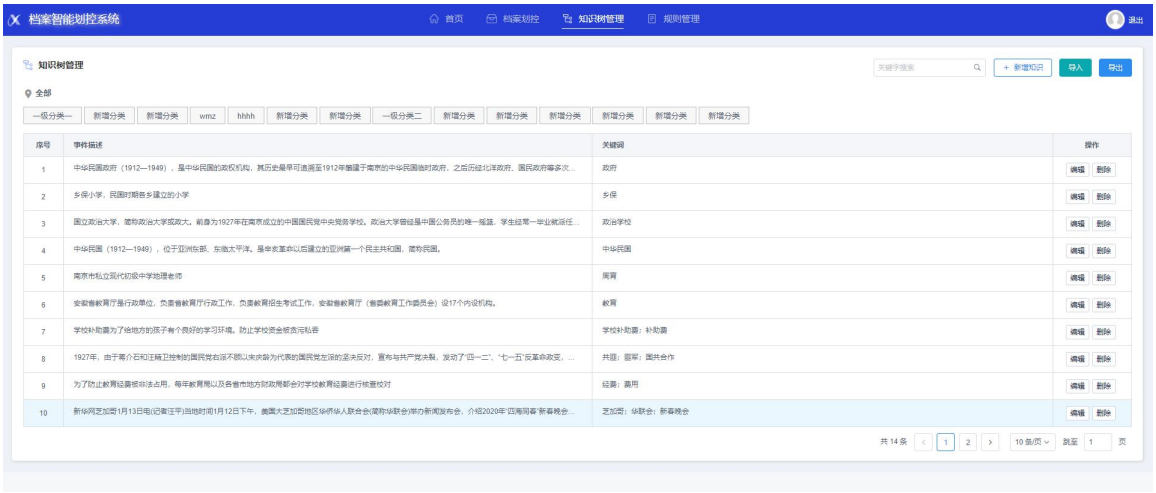

图　档案划控系统-知识树管理

6.知识树自动匹配，推荐划控知识

根据全文图文识别结果，识别关键知识点、规则、关键词。同时推荐知识树中的知识点进行结果匹配。并给出相关知识点词条、注释、信息描述，对人工划控提供数据知识支撑。支持关键信息跨页定点定位阅览、主动提示。

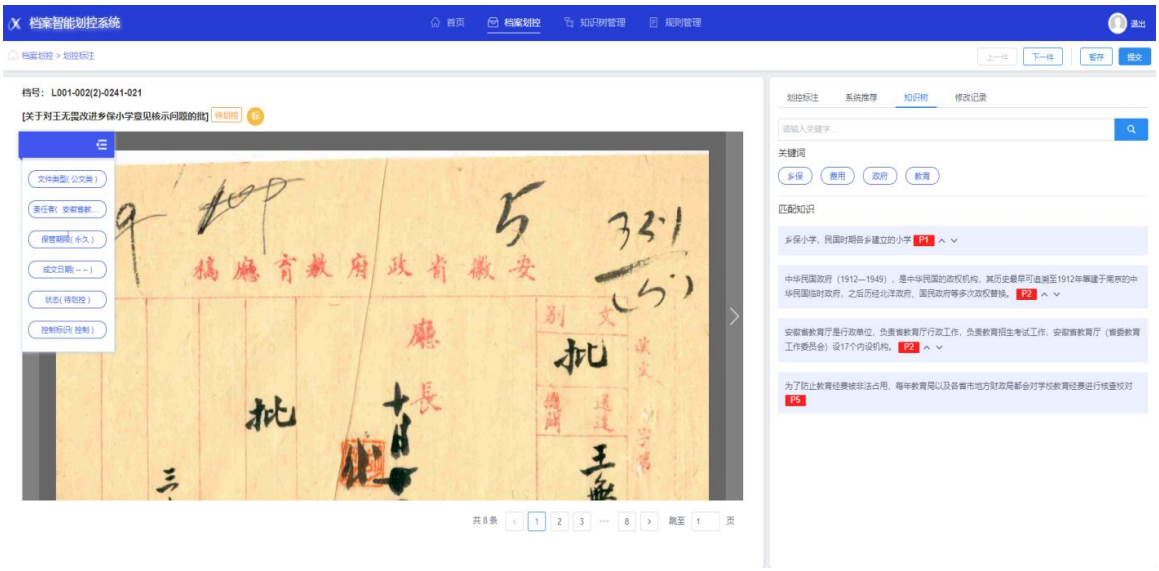

图　档案划控系统-划控详情知识树页面

三、成果的创新点：

项目组通过深入调查研究、建立合作机制、开发应用系统和全面部署实施等步骤，认真做好项目研究。本项目结合人工智能技术，利用基于 CNN（卷积神经网络）、语言技术平台 LTP 分词、文本分类技术的研究成果，构建人工智能计算机档案划控引擎、构建了档案划控知识库、规则库，开发了智能档案开放审核平台，主要创新点如下：

1.构建机器智能划控引擎

档案划控引擎本质上是一种先进的文本分类系统，在该引擎中，我们采用了 BERT 预训练模型，该模型在大规模文本数据上进行了预训练，学习到了丰富的语义信息。BERT 预训练模型的核心架构是多层 transformer，它在自然语言处理领域取得了显著的成果。与传统的循环神经网络（RNN）和卷积神经网络（CNN）相比，Transformer 模型具有并行计算能力强、具有 Self-attention 机制、捕捉长距离依赖关系等优点。在档案划控任务中，我们利用 Transformer 模型的强大表征能力，

并通过微调该模型，有效地提取了文本的特征信息，实现高精度的分类。同时，为了更好地利用文本数据的特征，我们的模型同时获取了题名特征和原文特征，这种多维度的特征融合策略可以有效地提高模型的精准度，使得模型能够更好地理解和挖掘文本数据的内在信息。

为了进一步提高模型的泛化能力，我们采用了文本对抗 fgm（Fast Gradient Method）策略。这是一种强大的正则化技术，训练模型时通过在输入数据中添加微小的扰动，从而提高模型对噪声和异常数据的鲁棒性。通过这种方法，我们的模型在面对具有挑战性的数据集时，仍然能够保持良好的性能。另外，在模型训练过程中，我们对训练数据还进行了清洗和增强。清洗操作主要是去除无关的信息和噪声，确保模型能够在干净的数据上进行训练。增强操作则是通过一些技巧（如同义词替换、句子重组等）来扩充训练数据，从而增加模型的泛化能力。经过多轮迭代训练，我们最终利用开发集挑选出了最优模型，并封装成智能划控引擎。

2.构建档案划控知识库

采用开放审核过程伴随快速人工采集标注的方法，获取划控依据规则点，机器采集公共源词条、词库、文库数据比对词条，对规则点数据知识化关联处理，形成划控知识点。通过不同历史时期档案形成的历史背景，将涉密、涉政治事件、涉案、涉军、涉外、涉宗教、涉民族、涉边界、涉人事、涉诉讼、涉处分等问题的内容进行主题标签分类，最终实现分级分类的综合档案开放审核知识库构建。目前，已经积累知识点约 5000 个，持续积累中。

3.构建档案划控规则库

通过构建划控规则库，标注数据，收集规则后，根据文种类型、文件标签对规则进行自动分类及标签化处理，并赋予划控结果属性。通过规则可辅助提高划控准确度，降低档案鉴定工作的难度，提高档案鉴定的效率。目前，规则持续积累中。

4.智能开放审核平台

通过智能开放审核平台，针对档案划控鉴定审核的结果，机器智能划控引擎首先给予“开放/控制”结论及依据，实现划控初筛；平台利用文本分析、关键字词定位技术，对档案原文中的内容知识点自动识别提取，知识库、标签关联推荐，实现人工审核过程的快速阅读、知识获取；平台具备划控依据采集标注、规则管理、知识管理、多人多级审核功能；创新传统人工划控流程，整体实现“机器划控+知识推荐+规则采集+流程审核”一体化的档案鉴定划控审核新方法。

在项目实施前，安徽省档案馆与讯飞智元信息科技有限公司签署科技项目合作研究保密协议，讯飞智元信息科技有限公司为安徽省档案馆提供人工智能核心技术研究服务，安徽省档案馆提供用于研究的档案数据，双方在项目研究期间保守档案数据秘密等有关事项，遵守相关保密协议，圆满完成了科研项目的任务，达到了预期效果。