陈骝声档案知识平台

"骝"声依旧

ARCHIVES

作品概述

陈駒声（1899—1992） **工业微生物学家**

中国近代工业微生物学的奠基人和开拓者

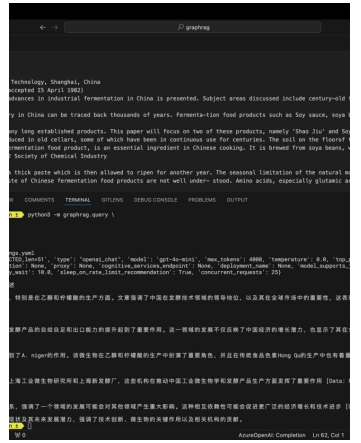
推动**科学家精神**进校园、进课堂、进头脑。

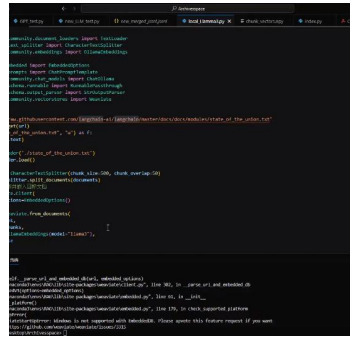
系统采集、妥善保存**科学家学术成长资料**，深入**挖掘所蕴含的学术思想、人生积累和精神财富**。

——《关于进一步弘扬科学家精神
加强作风和学风建设的意见》

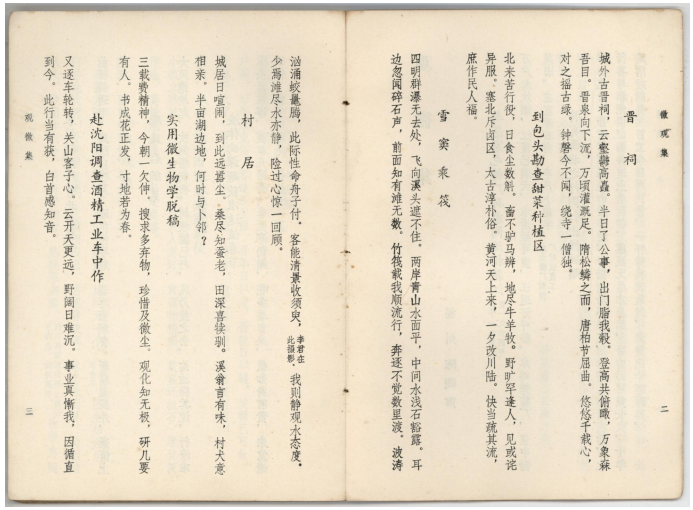
作品概述

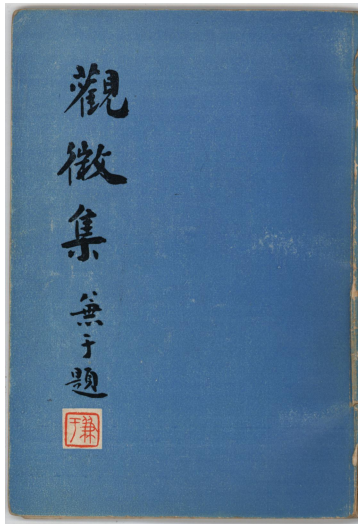
本作品以陈駒声科学家档案为例，利用开源档案管理软件ArchivesSpace、IIIF（国际图像互操作框架）和OCR技术，对档案数据进行标准化管理且支持档案著录、自动化识别、对比和分析。并搭建了基于档案数据的Classic RAG及GraphRAG，以语言重构与挖掘能力释放数据的佐证及解释能力。项目结合档案学理论、数字人文方法、大语言模型等对科学家的档案材料进行处理，并深入挖掘档案材料中蕴含的数据。**本作品不仅关注科学家的学术成果，也重视他们的人生积累与精神财富，从档案材料中探索老一辈科学家生活的多个侧面，弘扬老一辈科学家的精神。**

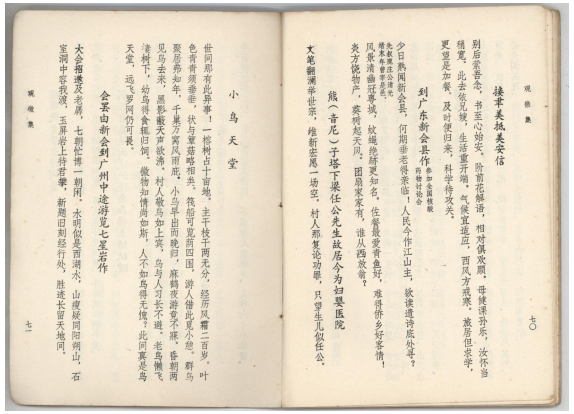


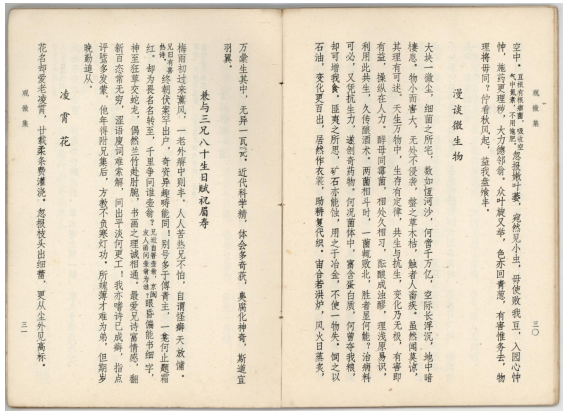


特色档案与挑战







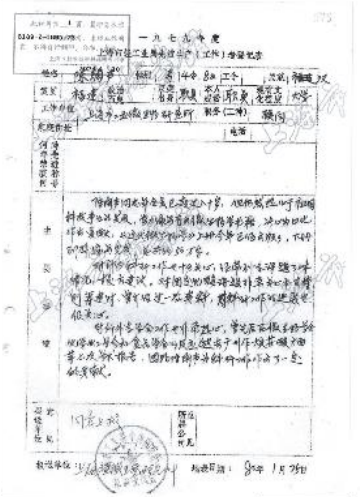


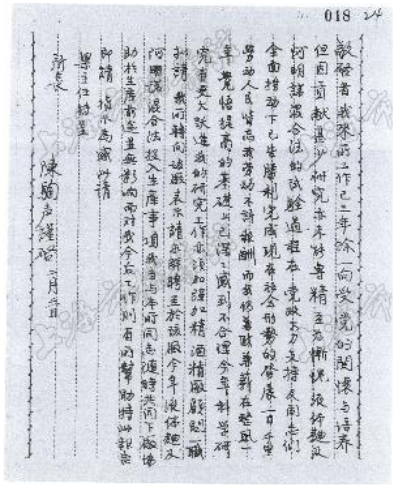
字体多样、缺乏校勘

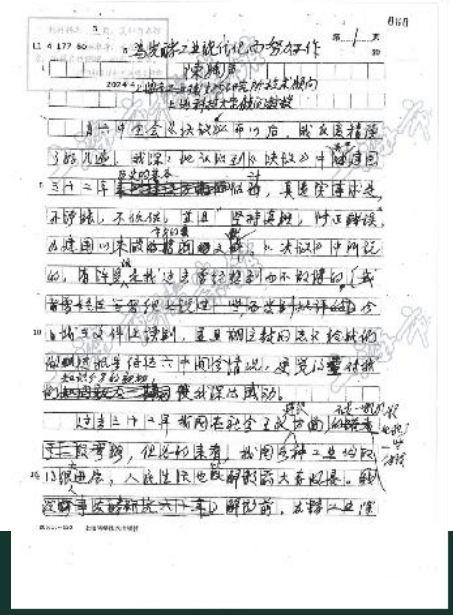
陈駒声 × 《观微集》

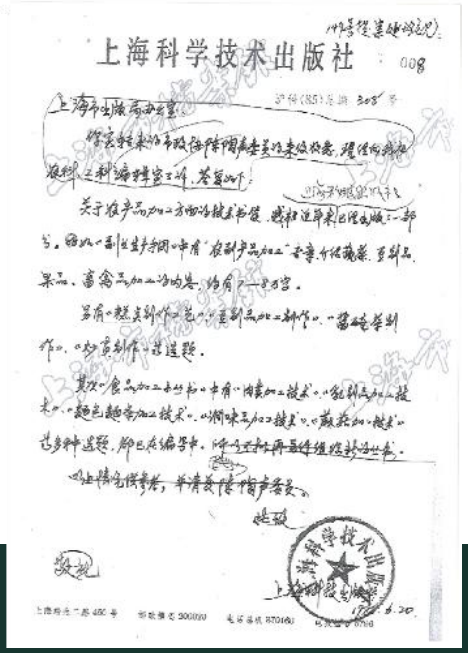
陈駒声 × 手稿

布局复杂、稿本多变









现有方案不足

面对科学家档案的传统开发模式

- **只关注表层内容，未关注档案作为数据的价值**

传统开发模式只档案的内容，对于档案作为数据的开发和传播关注甚少。

- **使用传统的内容识别和著录方式，耗费大量的人力与时间**

传统开发模式下较为特别的档案材料需要人工对识别出的内容进行额外的调整，且档案元数据全由人工著录，消耗大量的人工和时间成本。

- **档案资源呈现方式单一，整合与可视化效果不佳**

传统开发模式呈现档案资源的方式以展示为主，互动方式较为单一，可视化效果不佳。

作品概述

作品以微站的形式呈现，可以自适应各种设备（使用flutter框

原始档案处理过程


1（1）.jpg


1（2）.jpg


1（3）.jpg


1（4）.jpg

干部自传 Resource

Basic Information

Title 干部自传

Identifier 陈驹声 上海大学档案馆 1

六十四年前的一场较量 Resource

Basic Information

Title 六十四年前的一场较量

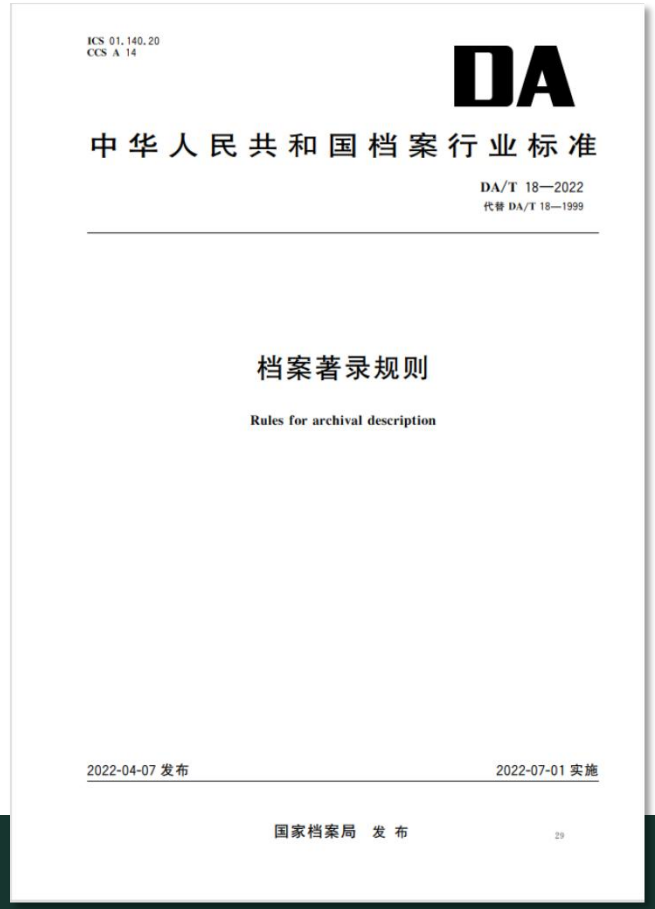
Identifier 陈驹声 九三学社 1

上海市轻工业局关于同意解除陈驹声兼上海酒精厂的顾问职务的批复 Resource

Basic Information

Title 上海市轻工业局关于同意解除陈驹声兼上海酒精厂的顾问职务的批复

Identifier 陈驹声 上海市档案馆 1


ICS 01.140.20
CCS A 14

DA

中华人民共和国档案行业标准

DA/T 18—2022

代替 DA/T 18—1999

档案著录规则

Rules for archival description

2022-04-07 发布　　2022-07-01 实施

国家档案局　发布

29

组件

著录

分类

著录工具：ArchivesSpace

著录标准：档案著录规则 2022（行业标准）


1（1）.jpg


1（2）.jpg


1（3）.jpg


1（4）.jpg

档案“尊重三原则”

从件内顺序上

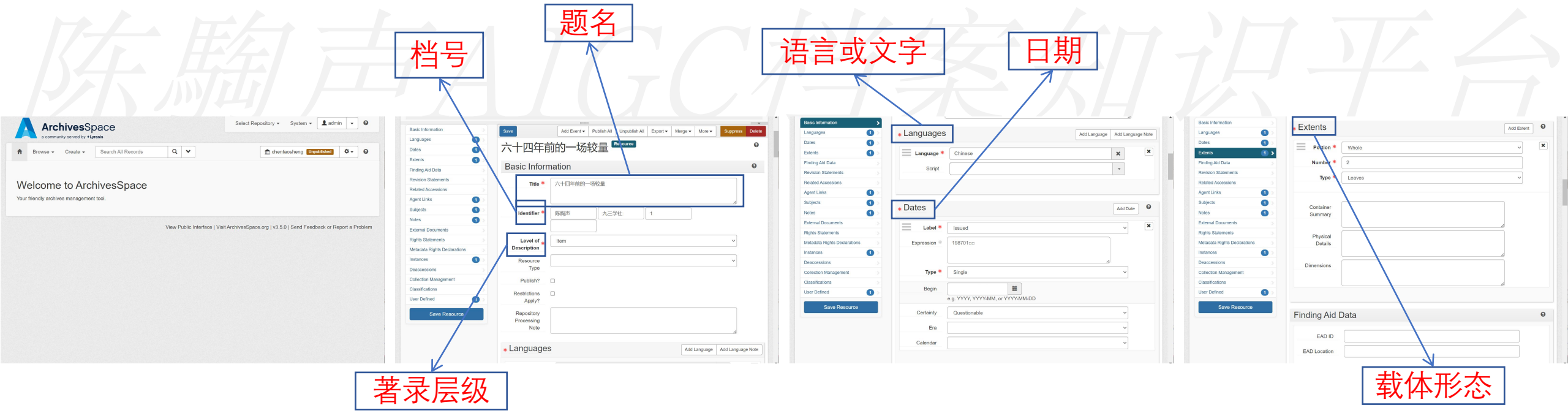
Title 上海市轻工业局关于同意解除陈驹声兼上海酒精厂的顾问职务的批复

Identifier 陈驹声 上海市档案馆 1

从全宗内顺序上

尊重全宗　尊重来源　尊重原始顺序

ArchivesSpace的DACS标准与档案著录规则（2022）的融合



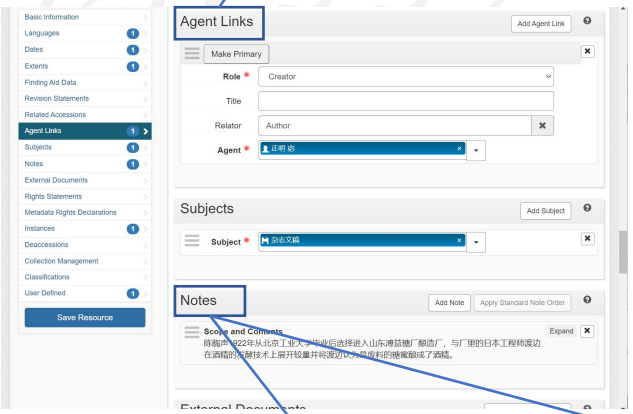
和DACS相比，档案著录规则的著录条目对于日期、载体形态、责任者的描述没有那么细致。

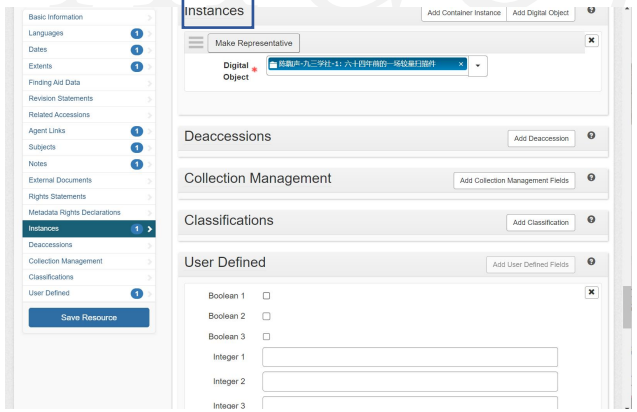
ArchivesSpace的DACS标准与档案著录规则（2022）的融合

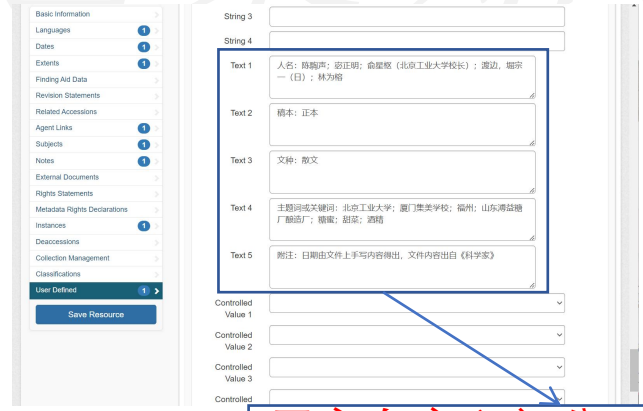
责任者

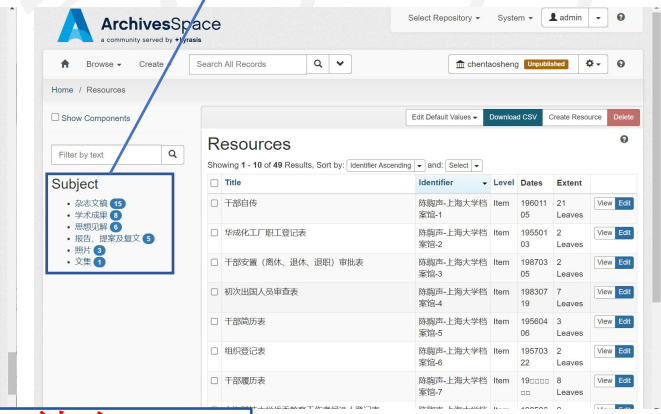
实例：可以添加档案图像、实体的相关链接

与条目中的subject对应，标引档案的类型，便于后续大语言模型对于档案材料的利用。









Custodial History（档案保管沿革）

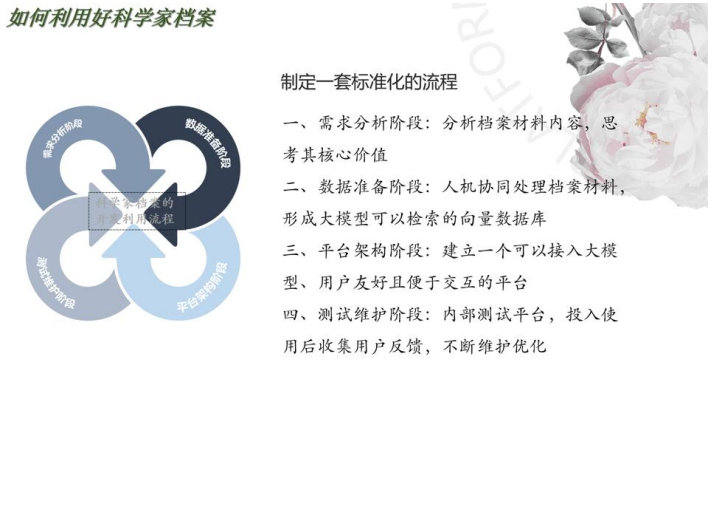
Scope and Contents（范围与提要）

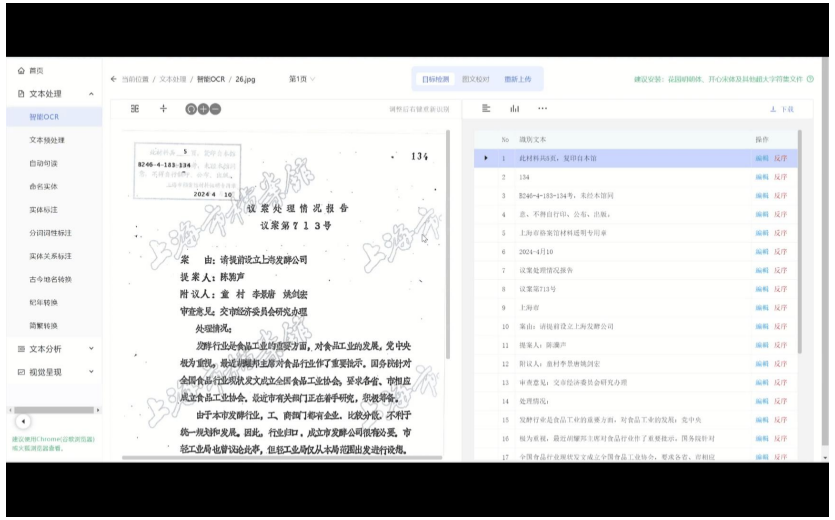
用户自定义部分：补充DACS中所未提及的人名、稿本、文种、主题词或关键词、附注等

解决方案

陈駒声AIGC档案知识平台如何解决困境:

1.以IIIF（国际图像互操作框架）[网址：https://viewers-dev.etu.wiki/]和OCR（光学字符识别）[上海图书馆开放数字人文工具]技术为基础框架，将纸质档案材料转化为图像原生的档案材料，维护档案材料的**原始性**，尊重档案材料**实存形态**，赋予其作为**“数据”**和**“证据”**的可能性，保证档案材料可以长期地保存。



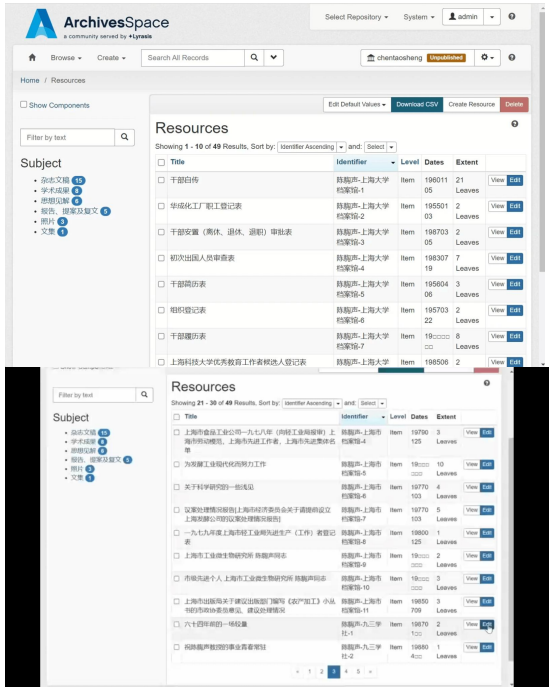


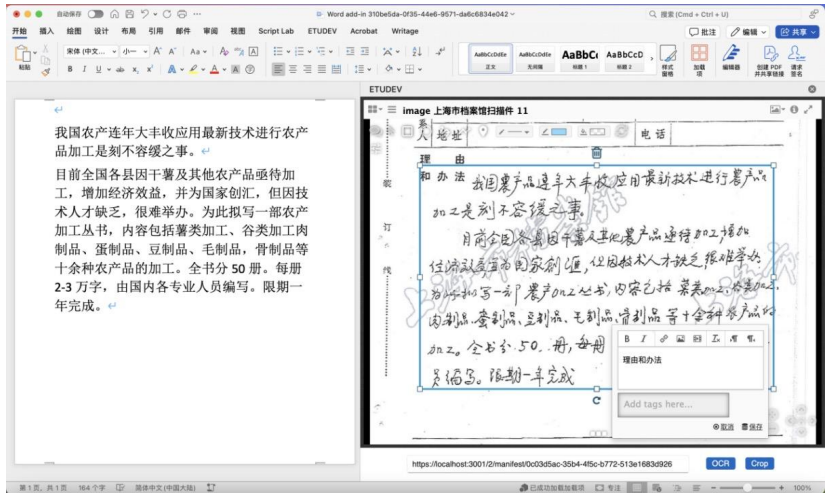
制定一套工作流程　　IIIF　　OCR

解决方案

陈駉声AIGC档案知识平台如何解决困境：

2.整理科学家的照片、文学作品、专著、专利等档案，在ArchivesSpace中对其**主题、稿本、文种**进行标引，配置了office+IIIF+OCR的插件，同时，借助人工智能对档案材料的元数据进行高效地识别，节省了大量的人力成本。





用IIIF处理海量手稿并在ArchivesSpace中链接

在office工作环境中可以对手稿做有针对性的OCR

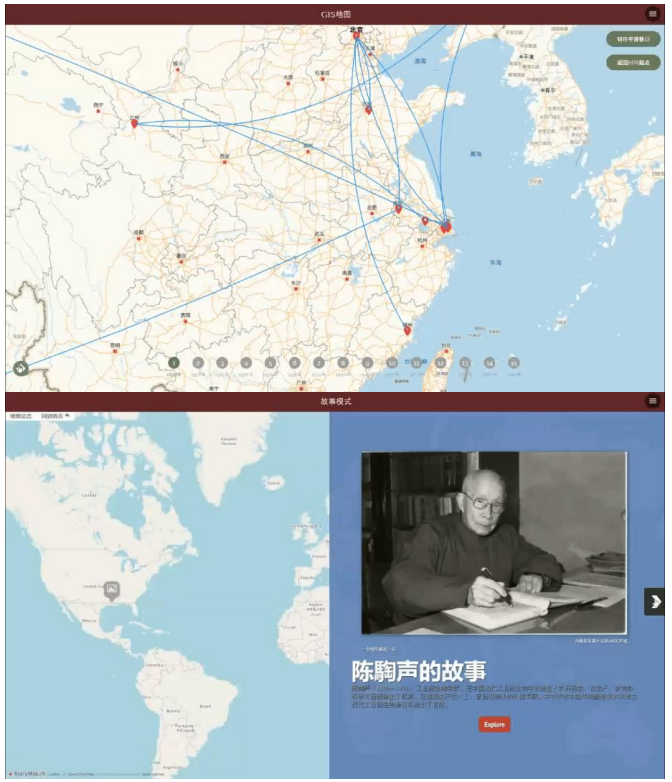
该流程对于处理手稿或排版复杂的文档有特别的价值

解决方案

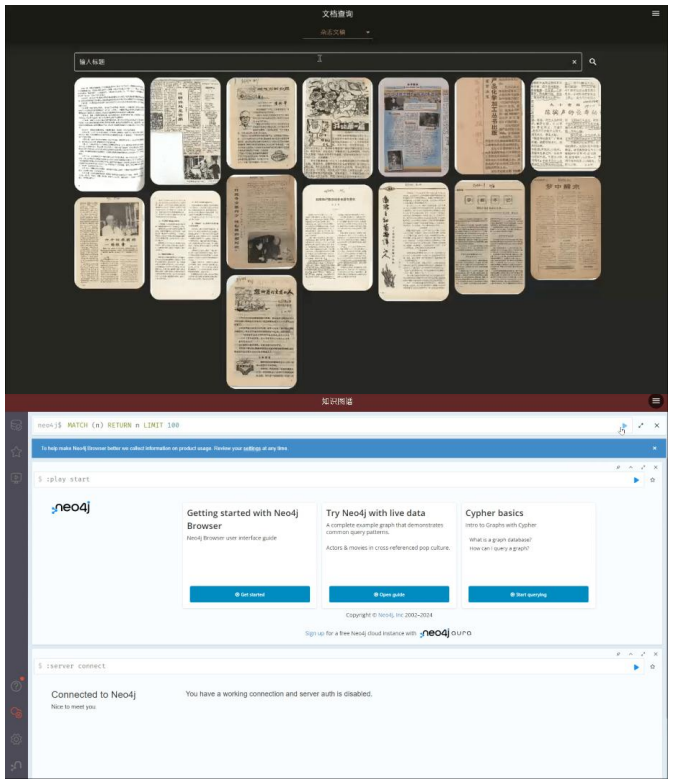
陈骝声AIGC档案知识平台如何解决困境：

3.集合、整理不同来源的科学家档案资料，为有价值的档案**提供保存的平台和多种可视化方案**，为后续研究者科学地挖掘和分析档案提供便利。

大事年表、照片墙



gis地图（故事模式中所有图像均为AIGC生成）

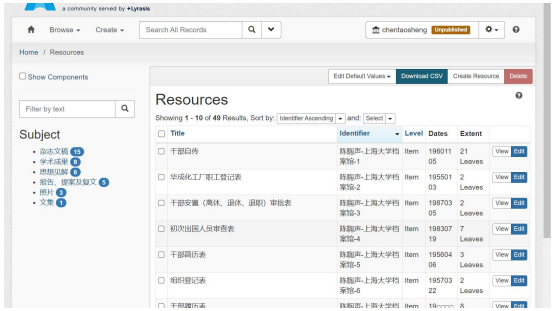


检索、知识图谱

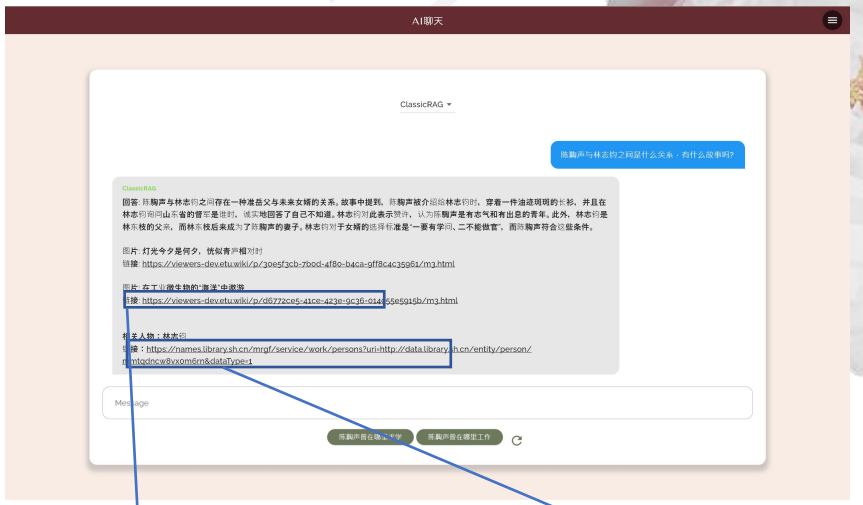
如何利用好科学家档案

陈鬷声档案开发模式

- 分析用户利用需求，关注档案多方面价值
- 关注档案资源特点，突出档案资源的核心价值，深入挖掘档案材料中蕴含的数据和关联
- 引入AIGC技术，人工智能协同著录档案材料
- 兼顾档案的原始形态与数字图像，使用“图像原生”档案数据，提供系统化、面向未来数字化的档案资源



使用大语言模型解决用户获取档案信息难度大的问题

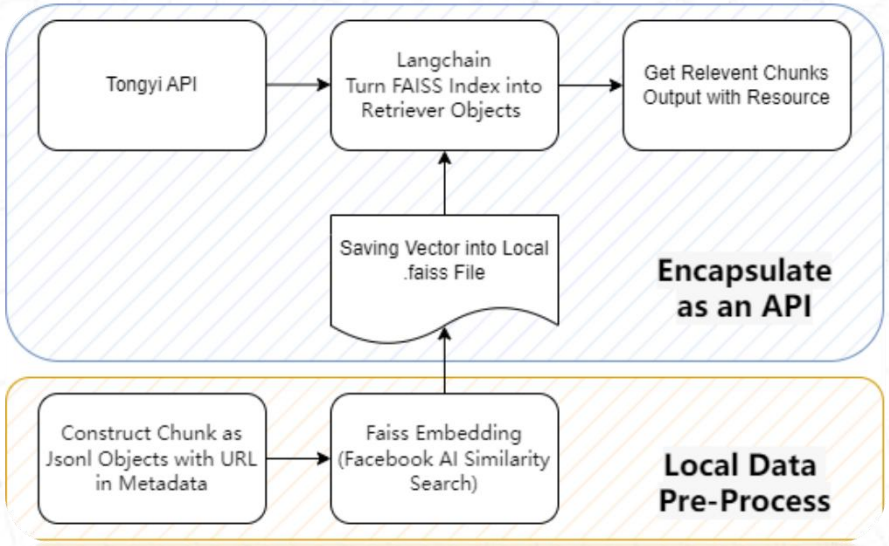


图像原生实现档案循证
跨模态信息的查询、确保数据真实性

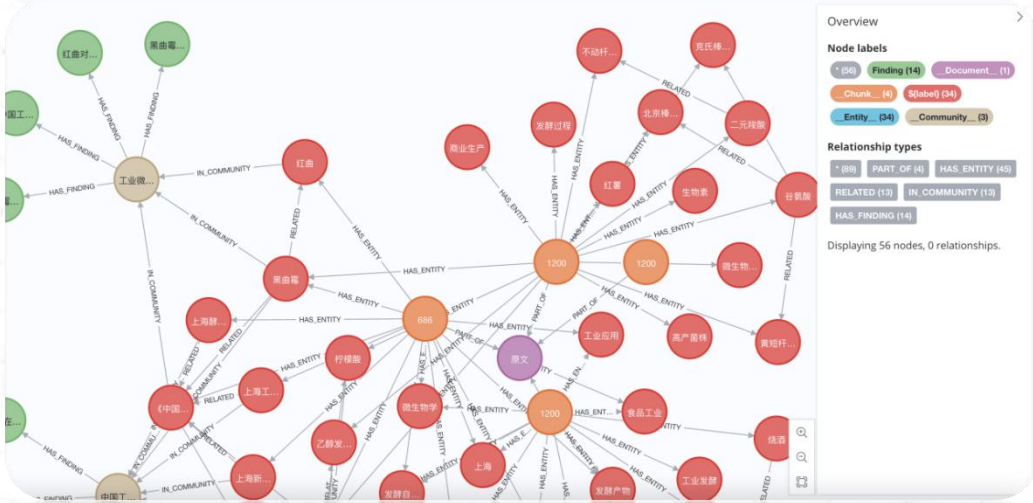
技术实现方案

检索增强生成（RAG）在解决大型语言模型（LLM）的挑战方面取得了显著的成功，而无需重新训练。RAG通过引用外部知识库，优化了 LLM 输出，有效缓解了信息“幻觉”、缺乏特定领域知识和提供过时信息等问题。对于图书馆数据库来说是一个高性价比的选择。

- 经典RAG系统首先通过OCR技术将文档转换为Markdown格式。利用此格式的层次结构，系统使用重叠滑动窗口方法分割文本，保持语义完整性。分割后的文本块被转换为Jsonl对象，包含文本内容和URL等元数据。这些对象随后被转化为嵌入向量，存储在向量数据库中。在使用时，系统通过最近邻搜索在数据库中找到相似文本，以此扩充上下文窗口。



- GraphRAG 沿用了这一流程，但增加了额外的推理层。首先大型语言模型（LLM）会对数据集中的每个句子进行推理操作，进行实体与关系抽取，命名实体识别（NER）识别文本中的实体，并确定实体间的关系及其关系的强度。最后将抽取的关系形成一个知识图谱，包含节点（实体）和边（关系）。



技术实现方案——Graph RAG 技术综述

Graph-Based Indexing： 建立结构化的图索引

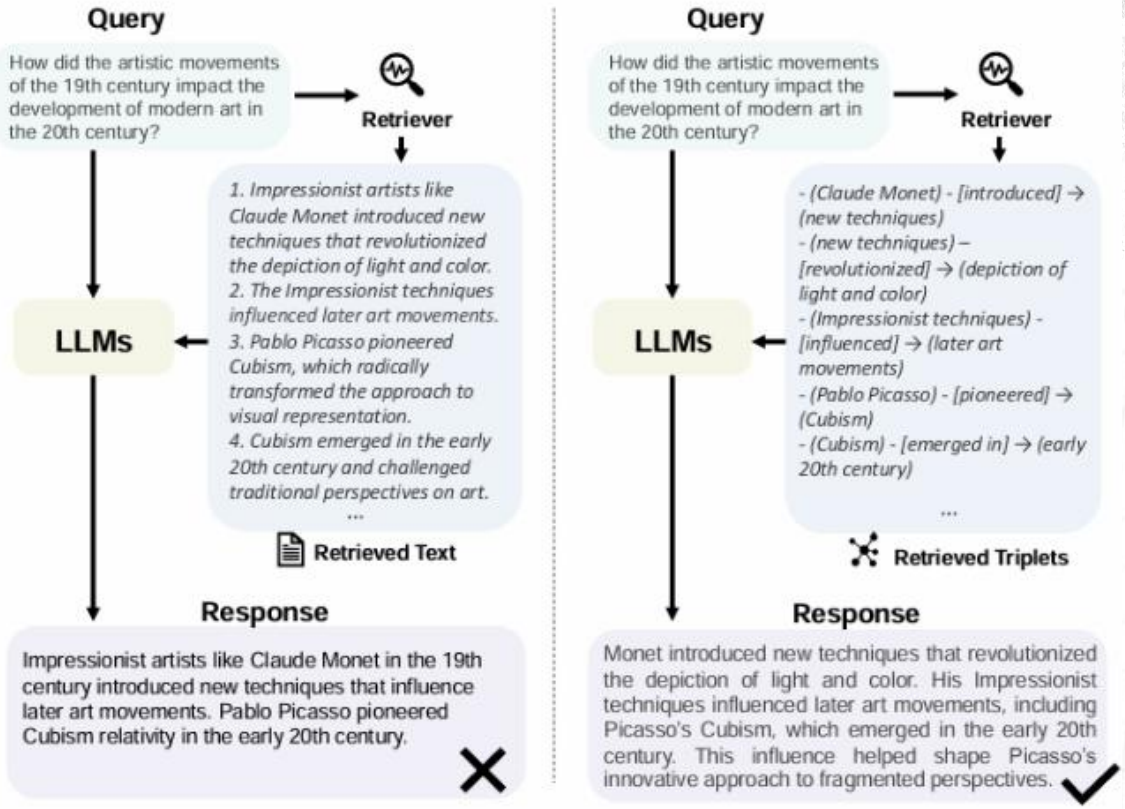
GraphRAG使用图数据库作为其核心。**图索引类似于图书馆的目录，帮助快速定位信息。**索引方法包括图索引（保留完整图结构）、文本索引（将图数据转换为文本描述）和向量索引（将图数据转换为向量）。实际应用中常混合使用这些方法以优化检索效率。

Graph-Guided Retrieval： 图结构指导检索过程

检索器分为非参数、基于语言模型和基于图神经网络三种类型。**检索策略**包括一次性检索、迭代检索和多阶段检索。**检索粒度**可以是单个节点、节点对、路径或子图。为提高搜索质量，还可以通过扩展问题或合并、剪枝搜索结果来优化。

Graph-Enhanced Generation： 生成相应社群与发现构建图谱

生成器使用了大语言模型。**LLM需要将图数据转换为自然语言格式。**转换方法包括使用图语言（将图数据转换为自然语言描述、代码形式、节点序列或树状图）和图嵌入（将图数据压缩成向量）。主要挑战在于如何有效结合图嵌入和文本数据。



技术实现方案-技术对比

从Classic RAG到Graph RAG，展现了信息检索和知识表示的重大进步。Classic RAG主要依赖向量数据库处理非结构化文本，而Graph RAG通过引入知识图谱，将实体和关系等价于单词，构建了一个超大规模的词汇表。通过这种方式，Graph RAG 在检索时能够将实体和关系作为单元进行联合建模，从而更准确地理解查询意图，并提供更精准的检索结果。

Classic RAG:

- 向量数据库存储
- 非结构化文本处理
- 向量相似度检索
- 有限上下文理解（基于chunk）

Graph RAG:

- 图数据库/知识图谱
- 结构化信息表示
- 图遍历和关系分析
- 丰富上下文理解（基于图）

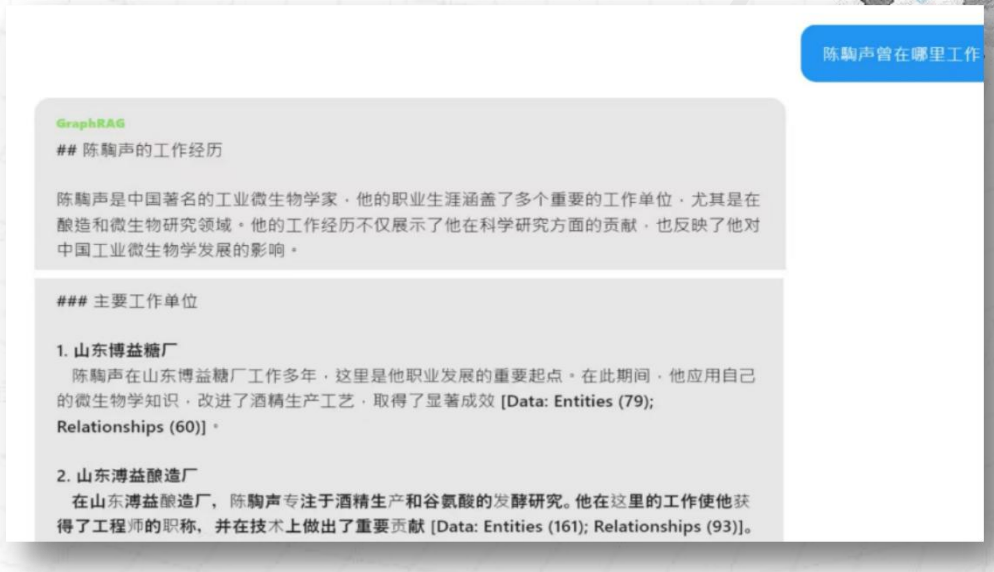
而Classic RAG则在处理大规模非结构化文本和实现简单性方面更有优势。在我们的实施例中，能够给出相应的图片、文字Resource，增强了大语言模型的**多模态输出能力**。Graph RAG 在处理复杂关系、提供丰富上下文和支持高级推理方面显示出明显优势化文本和实现简单性方面更有

技术实现方案-问题与解决方案

问题：LLM无法识别到关键词之间的关系，导致无法对问题做出可信和可溯源的回答，只能给出推测。**因为相关信息原因和结果跨度大，并且没有总结性语句。**

解决方案：通常通过增大Chunk的尺寸，增大重叠语段长度来使得语义关系更完整，使用不同大小的窗口进行多个层级的提取，但是对于大范围的概述并不能很好的解决（可以换用Graph RAG）

问题：出现了**命名实体的歧义**，例如文献中出现的"陈老"和"陈教授"都指的是陈鲸声老先生，然而，在构建知识图谱时，这些指称可能被错误地识别为不同的实体；**再例如图中所示对话，因OCR误差会加剧这一现象**。



解决方案：需要进行基于聚类的实体消歧（Entity Disambiguation）即确定一个实体指称项所指向的真实世界实体。原理是同一指称项具有近似的上下文，利用聚类算法进行消歧。

创新点与潜在价值

构建了一个集成科学家档案材料，用户友好的平台

平台集成科学家聊天、科学家知识图谱、科学家足迹地图、科学家大事年表、科学家档案材料展示与检索功能，提供处理过的档案材料，方便用户获取需要的信息。

形成科学家生平及其所在领域大事记

采取时间轴的形式记录陈駒声本人、酒精发酵工艺、酱油发酵工艺以及上海科学技术大学的重要历史事件，探究陈駒声的一生与其所在领域（酒精发酵、酱油发酵、教育）的关联性。

多模态大模型处理科学家档案

搭建rag及graphrag，实现科学家聊天功能，借助人工智能形成科学家知识图谱、科学家聊天功能等，更好地保存和传承科学家的研究成果，揭示科技发展与民族工业崛起的关联性，展现科学家与近代民族工业的“同构共生”。

绘制可视化地图

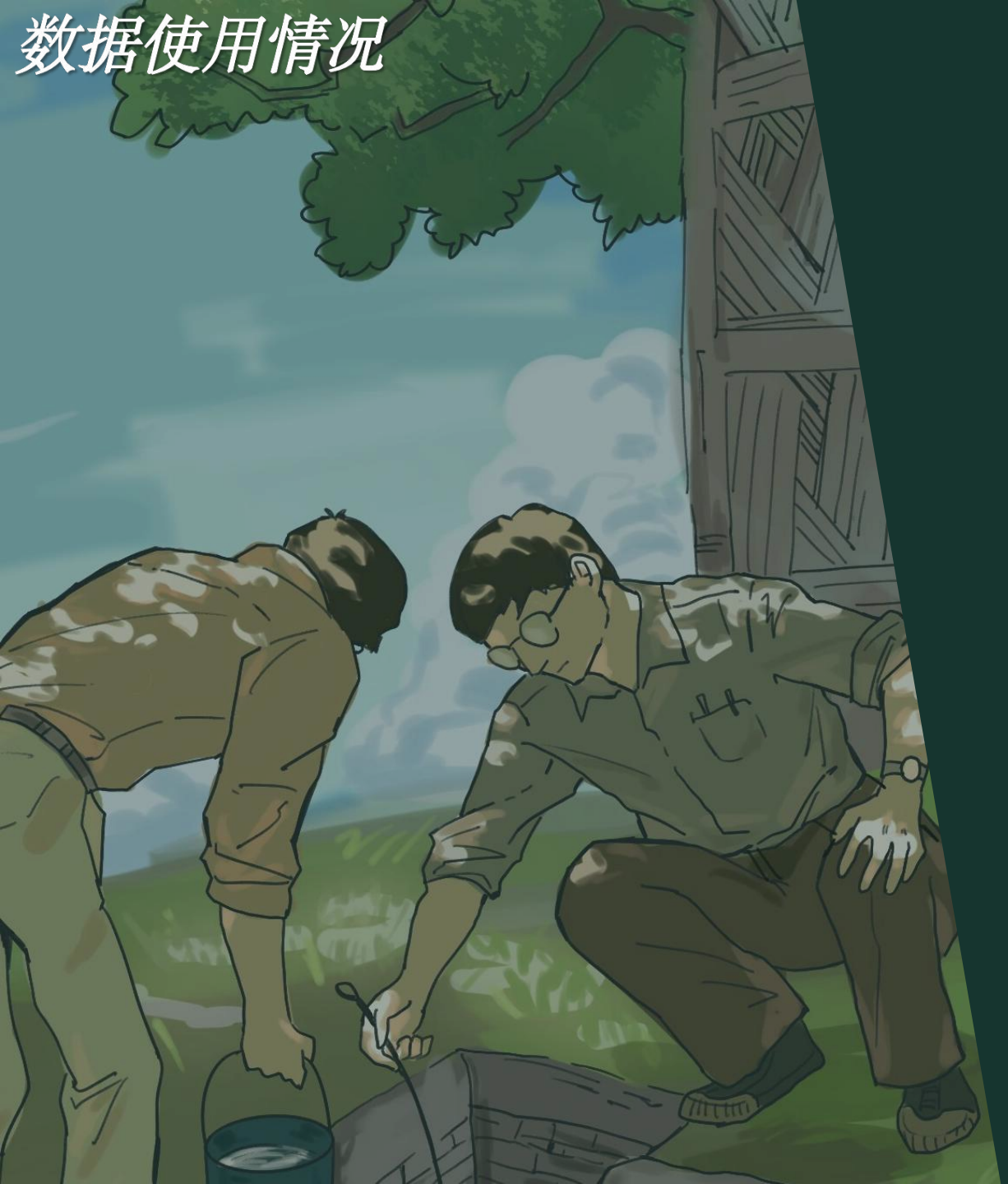
绘制可视化地图展现陈駒声的足迹，地图可切换故事模式，配合图文阐述陈駒声一生足迹。

挖掘科学家档案材料中所展现的文学情怀与精神世界

本作品不仅关注科学家的学术成就，还重视他们的人生经历和精神财富，通过档案材料展现科学家的多面性。这有助于公众更全面地理解科学家作为个体的复杂性，以及他们对社会的贡献和影响，从而弘扬老一辈科学家的科学精神和人文情怀。

本作品利用AGI时代的先进技术，高效处理科学家档案的大量图像数据，实现图文高质量交互，并确保档案内容的著录准确性与安全性。我们提出了面向人工智能环境的科学家档案管理与开发新模式，旨在充分展现科学家的学术思想、人生积累和精神财富。这一模式不仅适应了新时代用户对科技名人档案利用的需求，也为解决科学家档案管理的现存困境提供了新思路。

数据使用情况



1.九三学社：

出版过的陈骝声相关杂志和报纸、陈骝声本人发表或提及陈骝声的论文、陈骝声未出版的诗集《观微集》

2.上海市档案馆：

包含陈骝声的名单、陈骝声对科学研究的见解、陈骝声的提案及其复文、陈骝声的工作介绍信与存根、陈骝声评选先进个人的材料

3.上海大学档案馆：

陈骝声工作过程中填写的表格、陈骝声的思想见解、原上海科学技术大学相关校志、陈骝声相关聘书

4.上海市图书馆开放数据

参赛感想

- 合适的切入点
- 合理的团队结构与时间规划
- 持续改进

团队介绍

付雅明（指导教师）

上海大学文化遗产与信息管理学院档案学专业讲师

宋杰（指导教师）

上海慧游文化传播有限公司CEO

人工智能组

毕景云

物联网与人工智能数据分析专业

前端开发组

夏卉樊

计算机科学与技术专业

俞隽烨

计算机科学与技术专业

美工组

范舒恩

软件工程专业

成卓元

临床医学专业

档案预处理组

张欣然（领队）

档案学专业

莫家佳

信息管理与信息系统专业

王晶

信息管理与信息系统专业


感谢观看
RCHIVES