第十一届上海图书馆开放数据竞赛 · 应用开发及智能体赛道一等奖

一字一砖砌谱局

——从“听说”到“有据”的千年迁徙

刘丽娜

南京图书馆团队

第十一届上海图书馆开放数据竞赛 · 应用开发及智能体赛道一等奖

一字一砖砌谱局

——从“听说”到“有据”的千年迁徙



展示视频

先睹为快

片长 2′55″

第十一届上海图书馆开放数据竞赛 · 应用开发及智能体赛道

谱局



基于上海图书馆家谱知识库的迁徙证词辑录与对勘

众志修谱，初心布局，九年同行，续力传递

南京图书馆团队 · 九次参加开放数据竞赛 · 荣获一等奖

壹 · 先看一段宗谱摘要

一部《程氏宗谱》的馆藏提要

上图家谱知识库 · 编目员所写
原文照录，一个字没改。

始迁祖永寿，字忠，元末兵燹，明洪武初由江西饶州府乐平县瓦屑坝携九子迁居湖北黄州府黄冈县慕义乡东把村

朱笔所圈：谁 · 何时 · 从哪 · 到哪 · 为何

一句话 说清了五件事

壹

谁走的

永寿

贰

什么时候

洪武初

叁

从哪儿走

瓦屑坝

肆

到了哪儿

湖北黄冈

伍

为什么走

元末兵燹

壹 · 一处疑点

同一个瓦屑坝 不同归属地

乐平县

《程氏宗谱》

湖北省 / 黄冈市

「始迁祖永寿（字忠），元末兵燹明洪武初由江西饶州府乐平县瓦屑坝携九子迁居湖北黄州府黄冈县……」

平乐县

《张氏宗谱》

湖北省 / 黄冈市

「始迁祖珏（法名道静），明代自江西饶州府平乐县瓦屑坝迁居湖北黄冈定居。」

↑与「乐平」两字正好倒过来

丰城

《丁氏族谱》

湖北省 / 黄冈市 / 麻城市

「始迁祖宏，原籍江西丰城瓦屑坝，元末迁湖北麻城西隅南湖畈。」

存疑

现考证：瓦屑坝遗址在鄱阳县，不在乐平县。

读出五件事，标出一处可疑；不替前人修改，只是原录存疑。

谱局做的就是这件事——只不过对象不是一部家谱

是整个上图家谱知识库

103,252

条带馆员文言提要的家谱书目

采自上图家谱知识库 SPARQL 接口

谱局 | 数据采集、LLM处理与场景展示

馆员提要 → 迁徙证词 → 姓氏查证

01 数据采集

馆方开放数据
家谱题名 · 文言提要 · URI

→ **只读分页采集**
SPARQL · 限速 · 断点续抓

→ **保存原文快照**
103,252 部唯一著录

→ **筛选与同文去重**
含「迁／遷」· 扣除首轮语料
净池 86,562 部

02 LLM处理与核验

LLM 主抽取
DeepSeek V4 Flash
谁 · 何时 · 从哪 · 到哪 · 为何
各字段附逐字引文

原文没写 → 留空；不推测缘由

→ **引文自检通过？**

是 → **程序逐字核验**
原文.find(引文)

否 → **Pro 复核一次**
选引文未命中更少的版本

辑出 138,470 条

→ **每个非空字段均有原文引文？**

是 → **证据入库**
138,410 条
原文 + URI + 字符位置

否 → **整条弃用**
60 条

03 场景处理与展示

标准化与分片
纪年换算 · A/B分级
64 检索片 + 64 详情片

→ **用户输入姓氏**
简繁互认

→ **检索与聚合**
逐批加载 · 按记录去重

→ **展示迁徙证据**
省域印阵 · 五元组列表

→ **点开证词回查**
原文高亮 → 馆方著录

同库衍生：星台 · 行旅簿 · 传说 · 档房 · 数据集导出

LLM离线抽取；网页读取已生成数据

99.99% 为引文匹配通过率，不代表史实正确率

依据本地采集脚本、核验脚本与 guixing.html | 2026-09-23

103,252 条书目 —— 138,410 条迁徙证词

① 采集

103,252 条家谱书目

从 SPARQL 接口采集
带有馆员文言提要

② 筛选与去重

86,562 段“迁徙”相关提要

迁徙关键词筛选
+ 同文去重

③ 大语言模型处理

大模型

逐字原文比对
+ 语义偏移度复合评估

④ 纯代码最终复核

46万5千多 个字段引文

为了去除 LLM 大模型
语义丢失

⑤ 生成并入库

138,410 条迁徙证词

作品网站上的信息，
都源于这批证词

⑥ 数字人文知识增值

12 个功能页面

检索用的数据分片、星图、
移民集散地书证等

最后以可溯源的迁徙数据集回流馆方：迁徙证词 + 两万余条迁徙缘由统计 + 240 组古今地名对照表等衍生数据

叁 · 先说不编

让 AI 读文言提要：会不会编？会不会读错？

名　南宋 · 郑樵《通志·氏族略》

「历代并有图谱局」

《谱局》之名由此而来——可见考据的传承。

法　刘向《别录》校雠

「一人持本，一人读书，若怨家相对，故曰雠。」

设计思路借鉴两千年前的校雠学：两人逐字对质。

五个 AI 角色接力

模型

辑录官

五要素抽取

谁、何时、自何处、至何处、为何；每个非空字段附原文逐字引文

→

程序

校勘官

逐字回查

拿每段引文回原文逐字比对；对不上，整条弃

→

程序

考异官

跨谱候选

同姓、估年相差五十年内才配对；评分达标进候选

→

模型

驳议官

对抗式挑错

默认怀疑；三条全中才判同案，缺一驳回

→

落判

总纂官

三方印定案

整印勘合
半印存疑
无印驳回

模型管规模 规模化阅读与对抗性质疑

程序管精度 确定性校验守住可信边界

叁 · 先说不编

读的规矩 有三条

壹

原文没写的字段，
留空

贰

非空字段，
必须附上提要原文

叁

只许输出Json格式
结构化数据

符号规矩 方可验收为一条迁徙证词；一段提要，往往能读出不止一条证词。

快、慢读取的双重策略

DeepSeek-V4-Flash · DeepSeek-V4-Pro

选型依据

文言处理能力已在本团队关于《史记》的研究中得到证实——该研究入选 2026 年中国数字人文年会（**CDH2026**）

叁 · 不编：纯代码逐字回查

46 万多处字段引文

60 处未过关

并非语义错读
原因如下：

简繁体　半全角　空格换行

所呈现的信息出处 能够精确标到字符

再说读错

不编还不等于读对 进行人工抽查

96.4%

严格口径精确率

抽查 140 条

1	4	2
读错	漏读	拿不准

错的那条：错判了提要里的迁徙年份

叁 · 标准化处理

统一纪年 生成跨谱配对

第一 统一纪年

年号元年对照表
＋ 粗粒度时间词

→ 年份转换

第二 跨谱配对

姓氏相同 ＋ 估年相差不超过 50 年

↓

共生成5万6千余对

↓

AI 驳议官：人物、时间、迁徙地三条全中，才判同案

→ 后面的「公堂」判案环节中有具体呈现

补齐家谱主题赛道两块空白

壹

迁徙证词
多维度 强关联
可视化呈现

贰

数字人文叙事
让数据开口
道明「从哪到哪」
「为何迁」

宋明清三朝的迁徙发现

13 余万迁徙证词，按年号排开



再看「为什么离家」

19,957

条证词写明迁徙缘由 · 原数据结构中不存在，是我们赋予的数据增值



仕 宋代

乱 元末明初

商 明中叶

荒 明末

垦 清初

六百年里，从「仕」变成了「乱」，又变成了「生计」

谱局的发现
也只是家谱的记忆

参照赵世瑜先生

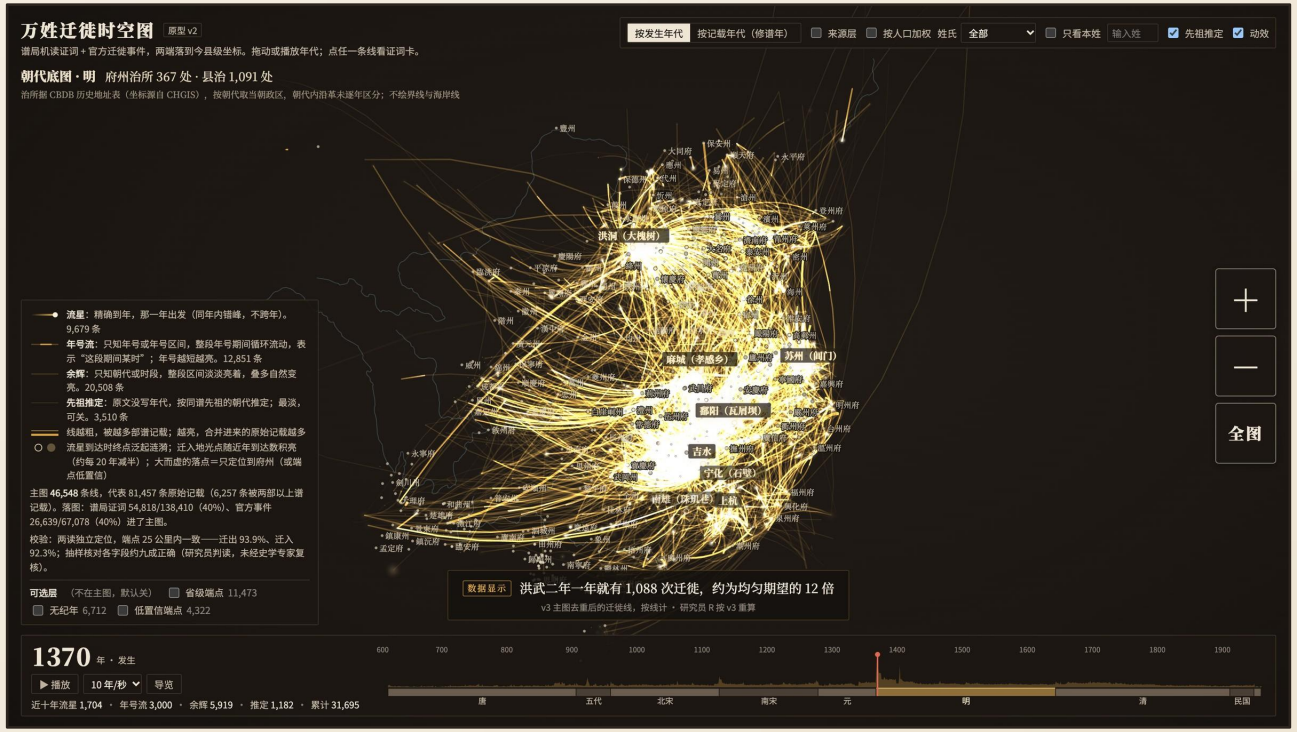
山西洪洞大槐树集体记忆推论

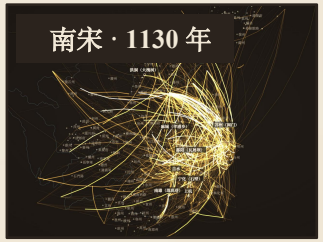
谱局 提出审视角度
不替历史学作定论

万姓迁徙时空图

基于整库采集的上图家谱知识库全量数据的动态迁徙图

接近完成









动态播放 · 只看本姓

4.6 万条迁徙线随年代流动；输入姓氏只看本姓；每条线都能回到提要原文。

全量数据（团队整库采集）

1.23 亿条三元组重新定位、定年；主图 46,548 条线，代表 81,457 条原始记载。

不确定性表达

纪年到年号的是「亮而短」的流星，只知朝代的是「淡而长」的余辉，不插值。

线上体验

扫码，或点击下方网址

review2pass.com/zaoban/shikong/

《谱局的一天》

智慧谱游：寓教于乐 身临其境

线上体验 · 扫码或点击
review2pass.com/zaoban/yitian/

接近完成













研究价值：公众参与式校勘——背对背双判、暗题质控、分歧仲裁

谱局 IP 与产品

专属 IP“橘猫”形象 多维度创意产品



IP 核心

世界观 清代江南谱局 · 一天十二时辰

角色 门房橘猫“大橘”

视觉 刻本活字 · 朱印 · 秋日江南

玩法 印书即交卷 · 合符即判案 · 落印即结业

内容 上图家谱数据 · 句句有出处



馆内智能终端

智能大屏、一体机、互动桌、服务机器人



手机小程序

过所、谱牒、有缘谱随身带



主题展陈

迁徙长卷、家谱主题展



文创

过所护照、印章、活字套件、大橘周边



研学课程

活字印谱、家族史、乡土课



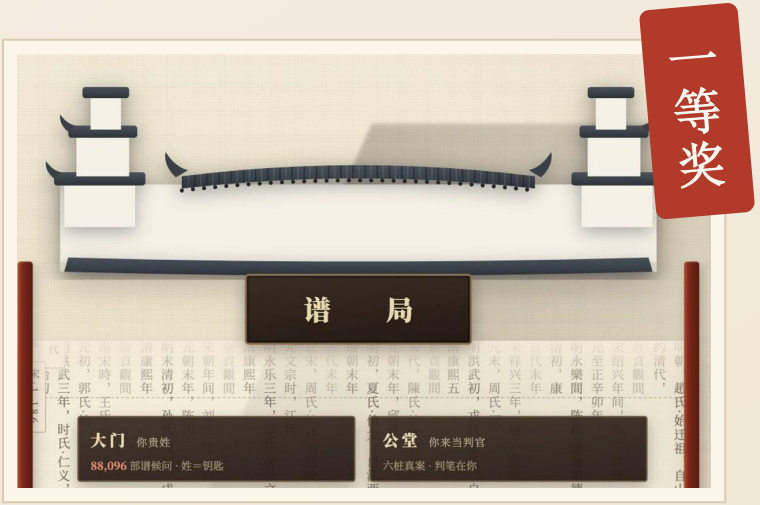
出版与影音

大橘绘本、动画

注：④–⑥ 为产品概念示意图，尚未生产

同一届竞赛 三条赛道 一个源头

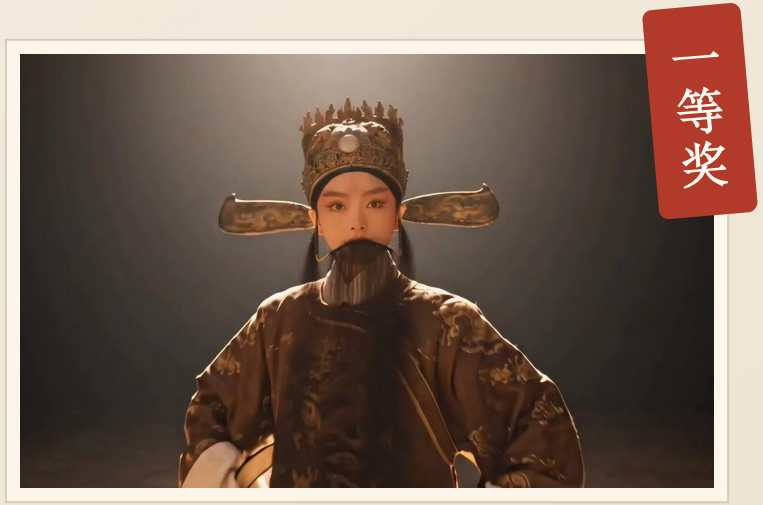
上海图书馆开放数据·承古启今



应用开发及智能体赛道

《谱局》

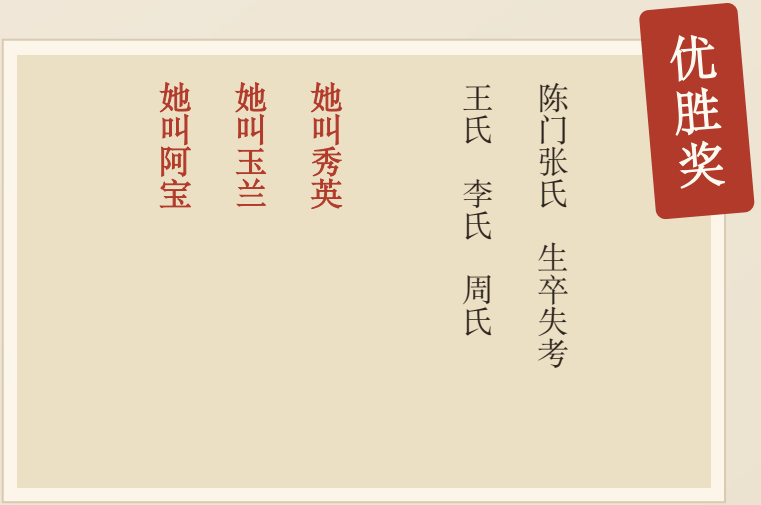
家谱知识库里十万多条馆员提要



AIGC 微电影

《替》

电影库与戏单里的一出戏



AIGC 歌曲

《氏》

家谱里只记作“某氏”的女性

谱已修就
局方新开

《谱局》的门会一直开着，欢迎你来扣门探寻

www.review2pass.com

感谢上海图书馆 · 感谢各位评委与同行