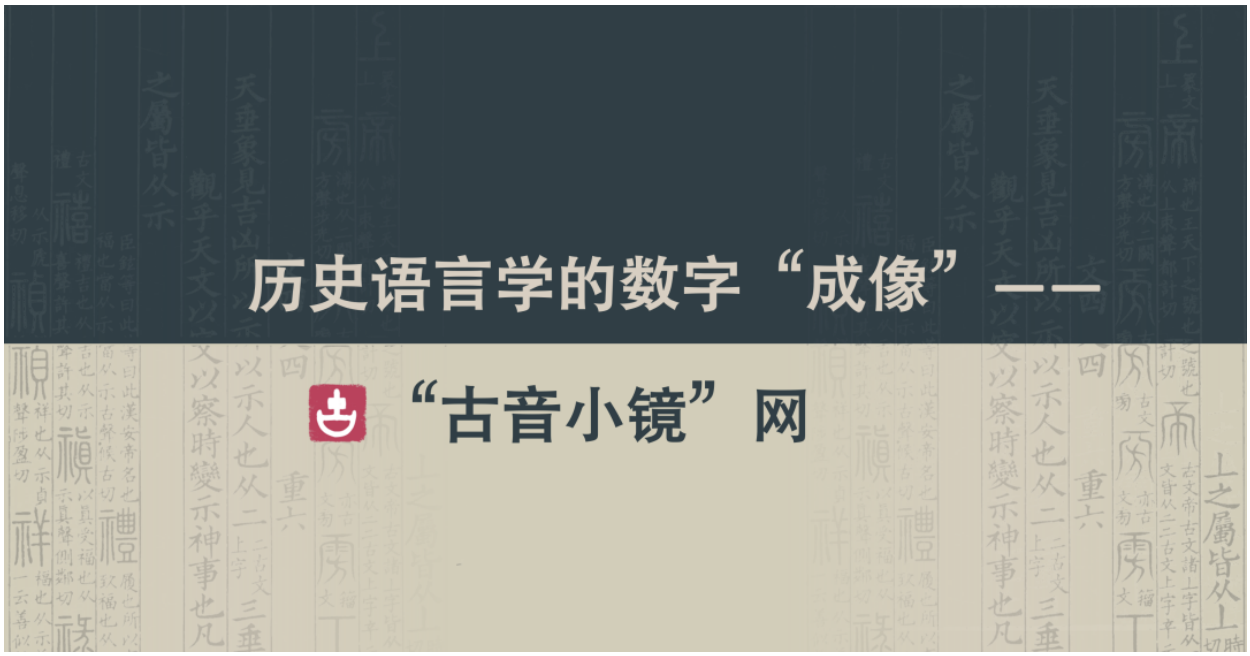
由中国人民大学信息资源管理学院冯惠玲教授、加小双副教授主讲的《数字人文导论》是中国人民大学本科生“数字人文荣誉辅修学位”系列课程之一，选课学生涵盖本、硕、博三类。课程内容包括数字人文导论（理论讲解、案例分享）和数字人文项目设计（实践操作）。
本系列推文共12期，是该课程学生课堂展示的成果报道。第二期由国学院2019级本科生盛一涵和信息资源管理学院2019级本科生田雨娇共同完成，以顾国林等民间小学爱好者合作建设的“古音小镜 • 历史语言学共享网站”项目为研究对象，对该项目的背景、主要内容、建设过程和项目特色进行系统梳理，并总结对今后数字人文项目的启示。



“古音小镜”是一个历史语言学领域的数字人文网站，学术性和专业性很强。小镜建设起始于2017年，最早放置个人积累的历史语言学材料。域名kaom.net，“kaom”取自「古ka-音om」两字的古音。小站用于探索汉语的早期历史，主要领域是上古音，兼及古文字、民族语、音韵训诂、汉语地理等。

项目背景

01

研究背景

语言学领域是人文研究中与数字技术最适配的领域，中国历史语言学研究材料众多，有古代的各种韵书、字书，还有陆续出土的文献。韵书和字书都有固定的体例，材料量大，而且基本都处于半结构化的状态，古代就已经出现了韵图之类“可视化”的方法帮助研究，现代的历史语言学研究更是基于统计的方法有了重大突破。

“古音小镜”网站汇集了大量的文字、音韵、方言材料，将它们全部整理成格式化的形式，并对大部分数据提供了非常直观的可视化表现。

图1 “古音小镜”网站首页

02

制作团队

古音小镜起始于2017年，由民间小学爱好者支持和搭建。站长顾国林先生是计算机行业的资深从业者，用业余时间学习小学并搭建了这个网站。资源提供者们按照古代韵书的体例，将文字材料整理录入表格，再由站长建设数据库，呈现在网页上。

图2 “古音小镜”致谢资源提供者

项目内容

网站内容分为上古音、古文字、汉语地理、地名、小功能、工具书六个板块，主要的材料、功能如下。

图3 “古音小镜”网站菜单栏

01

工具书

工具书栏目提供了所有材料原书的检索，基本囊括了中国历史语言学研究的重要材料。

网站共整合了：

- 韵书、字书35种
- 训诂工具书22部
- 沙门（佛家释经工具书）10部
- 16位学者对上古中古音的构拟
- 出土简帛11种
- 4种汉字声符
- 民族语词汇集15部
- 通假字书12部
- 汉语大词典
- 方言大词典
- 明清吴语词典

图4 工具书检索页面

02

上古音

- 材料：出土简帛、《诗经》用韵、《广韵》《集韵》异读、两汉经注的声训材料、《后汉三国梵汉对音谱》等。

- 主要功能：系联节点图、关系统计表

图5 《诗经》韵系联节点图

03

古文字

- 材料：

「甲骨字形库」，约44万单字切片；

「金文字形库」，约13万单字切片；

「楚简字形库」，约10万单字切片（未上线）

- 资料来源：爱好者整理分享的网络资源、PDF资源
- 功能：字形检索、字形溯源

图6 甲骨文字形检索结果

04

汉语地理

- 材料：方言大词典、方言研究著作
- 功能：以地图、桑基图的形式呈现方言的时空对比（已完成2023种方言的统计）

具体可概括为以下四个方面——

1. 单字的声母/韵母/声调等在不同方言中的对比

图7 “江”字的声母在不同方言中的对比

2. 方言字音统计

图8 单一方言的语音统计

3. 方言音系整体对比

图9 两个方言语音的对比

4. 方言古今演变对比

图10 《广韵》音系与某个方言今音的对比

05

地名

- 材料：中国392万自然村名
- 功能：输入地名用词（如厝、浜、峪、畈、舍、寮、垟、埭、嵛、涌等），程序生成点状分布地图。

图11 “各“字在中国自然村名中的出现情况

项目建设

01

栏目建设过程

在众多繁复的功能之下，是一套相同的严谨、踏实的流程。

1. 资源数字化
   1）获取语言学词典资源（实体书或扫描版）
   2）人工数字化
2. 整理并导入数据库（一本词典一个库）
   1) 数据库：MySQL
   2) 建表方法：
   1. 数据量小（几十万以下）：普通的建表方法，一部字典一张表； 查询时，先后查30张表，集中起来显示在网页上（对韵书字典来说，不加优化，总时间都可控制在2-3秒）
   2. 数据量大（百万条以上，比如方言字音）：使用数据库的索引功能提速
3. 网站建设
   1) 搭载数据库
   2) 网页设计
   3) 构建基础检索与可视化功能

《诗经》韵部分的录入与整理由站长一人完成，其他部分的资源数字化工作大多由小学爱好者交流群的群友完成，站长进行全部的数据库和网站建设、维护。

图12 项目建设导图

02

网站建设过程

起始：

时间：2017年

作用：放置历史语言学数字化资料

扩展：

时间：2017-2021年

内容：上古音为主，兼及古文字、民族语、音韵训诂、汉语地理

功能：语言学词典数字化查询、汉语字音时空关联可视化、古汉语学习小工具等

优化：改善代码提升速度、流畅性，减少经济负担，保证可长期运行【用户体验】

项目特色

1. 数据量庞大

结构化电子词典与注书共133部，全部实现字头查询，部分可进行全文查询。

2. 功能丰富

古音小镜充分利用了网页可交互的优势，对知识节点、时间和空间的可视化呈现技术的运用具有极强的数字人文特色。

3. 自由查询

1）单本词典电子查询

2）多部文献联动查询

4. 实体书扫描本全文阅读

图13 《说文解字诂林》

5. 穷尽、准确、可溯源

“穷尽”指穷尽可用资料。由于是民间爱好者合力积累而成，能最大范围地覆盖有关资料，并完整著录。

“准确”指知识准确、识读准确和著录准确。知识准确依赖于所凭借的资料；识读和著录准确取决于专业素养和责任感。由于是非营利性网站，这些工作也并非有偿，全凭爱好，准确性有所保障。

“可溯源”包含两方面，一是每条数据都出自权威编纂成果，有据可查，能通过层层点击进入文献扫面页；二是单本词典著录者也基本只有一人，一旦发现错误，找到当事人就可以全面修改。

6. 速度快

与多数商业化语言学网站相比，“古音小镜”的加载和运行速度都具有显著优势，这也是本职工作是程序员的站长引以为傲的一点。

7. 为爱发电，全免费

“小站是非盈利网站，所有功能都是免费的，不设广告。”

——摘自古音小镜网站简介

8. 建设周期长，持续更新

网站首页呈现了最近的更新内容。

图14 “古音小镜”网站更新栏目

最近一次更新在2020年12月6日。站长和他的群友们至今也一直在爱好的驱动下，利用业余时间扩充网站的数据库和功能。热爱不灭，更新不停。

9. 首创性

“古音小镜”是众多汉语语言学研究方向的集合，有些功能并非首创，但也有创新，比如假借字系联，其他网站尚未尝试，小镜属首例。

项目评析

01

项目启示

1. **用数字方法实现学科领域集成。**相信每一个第一次打开“古音小镜”的人都会被它的数据量震撼到，为它丰富的功能而惊叹。古音小镜实现了对汉语语言学的资料汇集和串联，为后人的学习和研究带来极大便利。
2. **以众包实现纸质资料数字化和数据结构化。**在数字人文项目建设中，我们倾向于让机器取代人工，尤其是基础性的文字识别、数据录入等工作，但又不得不承认人工识别的正确率更有保障。“古音小镜”让我们看到人工识别不一定是效率低下的，也不是不能完成的。
3. **计算机技能和数据科学素养对数字人文项目的构建非常重要。**站长顾国林先生的专业与职业都是计算机方向，这也是“古音小镜”能令人惊艳的最主要原因。他在毕业后对小学产生兴趣，才把知识和技术结合于一身，完成了这项造福后学的工程。

图15 顾国林先生线上访谈截图

02

研究批判性分析

在启示之外，基于以上的研究认识，我们对网站也有一些浅薄的建议以供商榷：

1. 利用庞大的结构化数据实现知识发现。现在网站的功能以查询为主，知识挖掘潜力还可以进一步。
2. 留存网站建设过程，吸引更多技术人员加入。
3. 数据录入过程中加入互查和相互商讨环节，进一步保障准确性。

“古音小镜”是承袭了古汉语语料库语言学的传统、典型的数字人文项目，涉及数字人文所需的各个环节，并且都做出了成功且成熟的尝试。无论是它的优点还是缺点，都能给数字人文项目的建设带来启发。

编辑 盛一涵/田雨娇
排版 盛一涵/田雨娇

公众号账号：rucdh2019
网址：http：//dh.ruc.edu.cn
邮箱：rucdh@ruc.edu.cn

中心简介

中国人民大学数字人文研究中心集人民大学多学科优势，秉持融合文理、协同创新之理念，开展数字人文理论研究、实践探索、人才培养和学术交流。