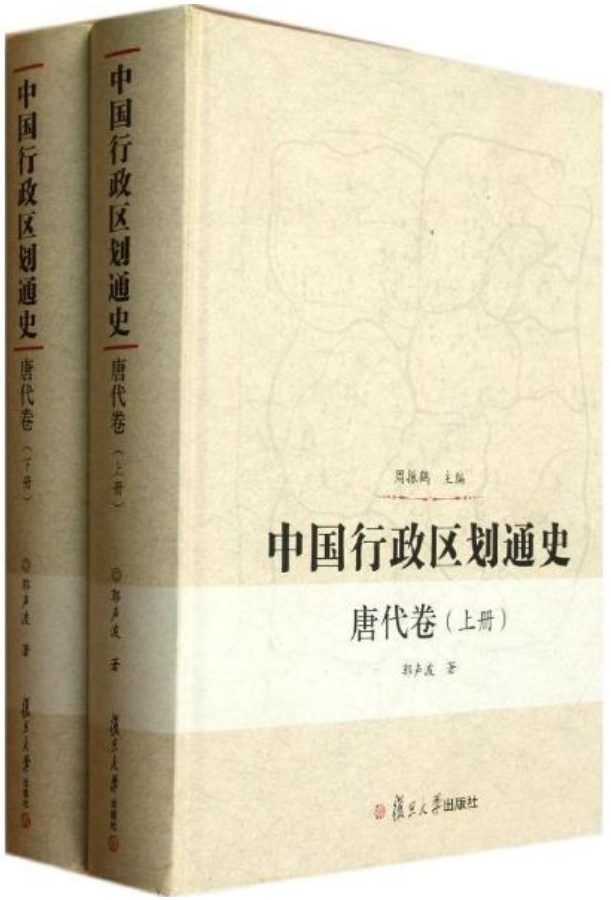
唐代地名信息更新

邢云 2018.11.8

CBDB唐代地名的增补

- CBDB原有2500多条唐代地名数据，主要来自于CHGIS

- 将《中国行政区划通史·唐代卷》中的地名增补到CBDB中



CBDB唐代地名的增补

- 预处理
- 筛选工作
- 导入工作*

*导入工作同时涉及上层政区的匹配及地理坐标的确认

预处理

- 将书中地名信息抽取为数据表格形式
- 利用python程序，转换数据表格使其具有与CBDB地名表相似的结构形式

name	time_ft_1	time_lt_1	…	level1	level2	…
酒泉	618	766	…	河西道	凉州都督府	…

抽取整理的数据表结构

name	time_ft_1	time_lt_1	time_ft_2	time_lt_2
开封	618	627	712	907

name	time_ft	time_lt
开封	618	627
开封	712	907

转换后的数据表结构

筛选工作

- 一、程序初步比对
  - 编写python程序，将从书中获取的6083条记录与CBDB唐代2564条地名数据进行逐一比对
  - 比较项为**地名、起止年代、*隶属关系**
  - 根据比较结果分类打标签（tag）

*郭书以天宝十三年（754年）为基准年代编写，从书中提取的地名数据，其上级隶属政区也都以754年为准，故而并不是准确的上级隶属关系，无法作为地名是否匹配的严格依据。由于唐代正州县中名称和起止年月存在一定的匹配关系，而隶属关系不同的情况，出现概率并不高。因此，最终自动比对结果只选取地名和起止年月作为比较项。

筛选工作

- 二、人工初次检查6083条对应关系
  - 去除647条无效比对
    - 郭数据无意义：187+69
      - 187条研究用名（如前**县，安氏大燕国，已修正）
      - 69条自定区域名称（**直辖地区、**直属地区，可忽略）
    - 比对无意义——391条

      层级不同的重名地名的错误比较
  - 将5436条有效比对进行筛选
    - 按分类不同采用抽查、重点检查等方式人工检查
    - 将隶属关系作为检查中的重要参考信息

筛选工作

- 三、利用程序对5436条有效比对进行筛选
  - 2411个郭书中有而CBDB中没有的地名
  - 3025个郭书与CBDB数据存在某对应关系的地名，按时间要素区分：
    - 657条郭真包含于CBDB
    - 1021条CBDB真包含于郭
    - 426条有交集
    - 542条空集
    - 379条完全匹配

编号	分类	数量
1	书有而CBDB没有的地名	2411
2	郭＜CBDB	657
3	CBDB＜郭	1021
4	CBDB∩郭	426
5	CBDB∅郭	542
6	CBDB=郭	379

筛选工作

- 四、人工再次对3025条时间要素对比进行查验

编号	分类	原数量	去掉	新数量
1	书有而CBDB没有的地名	2411	37	2374
2	郭＜CBDB	657	89	568
3	CBDB＜郭	1021	22	999
4	CBDB∩郭	426	19	407
5	CBDBØ郭	542	107	435
6	CBDB=郭	379	0	379

导入工作

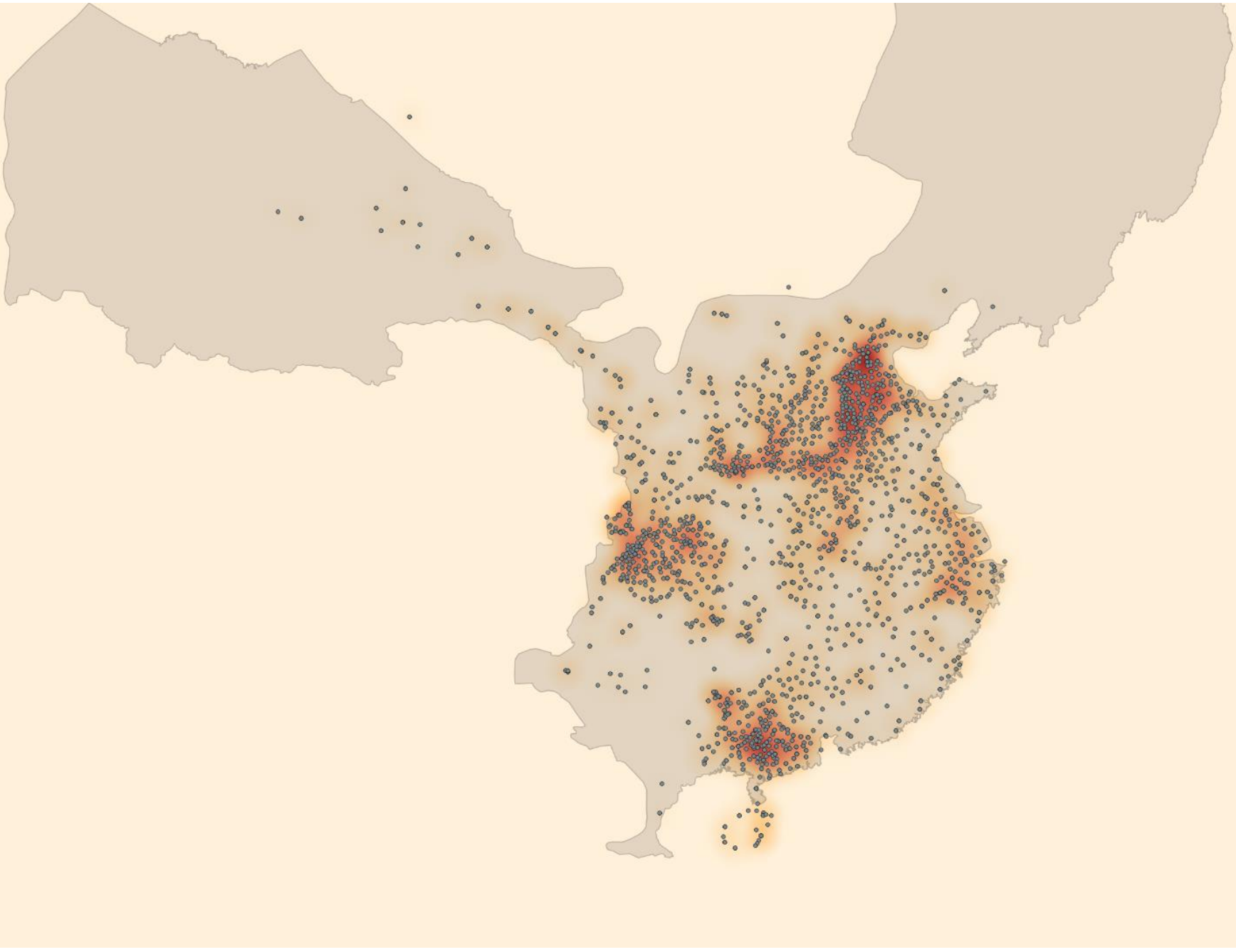
- 一、数据增补入库策略
  - A. 更新起止年（Modify_fy_ly）
    - 包括第3类999条、第6类379条，共1378条
  - B. 更新起止年和隶属关系（Modify_fy_ly_belongs）
    - 在4和5类中所发现的时间和隶属关系都不对的，均用书中数据替换掉CBDB，共842条
  - C. 导入数据（Insert）
    - CBDB中完全没有的数据，包括：第1类中2411条，第2类中发现的89条CBDB所没有的，第3类中发现的22条CBDB所没有的，共2522条

导入工作

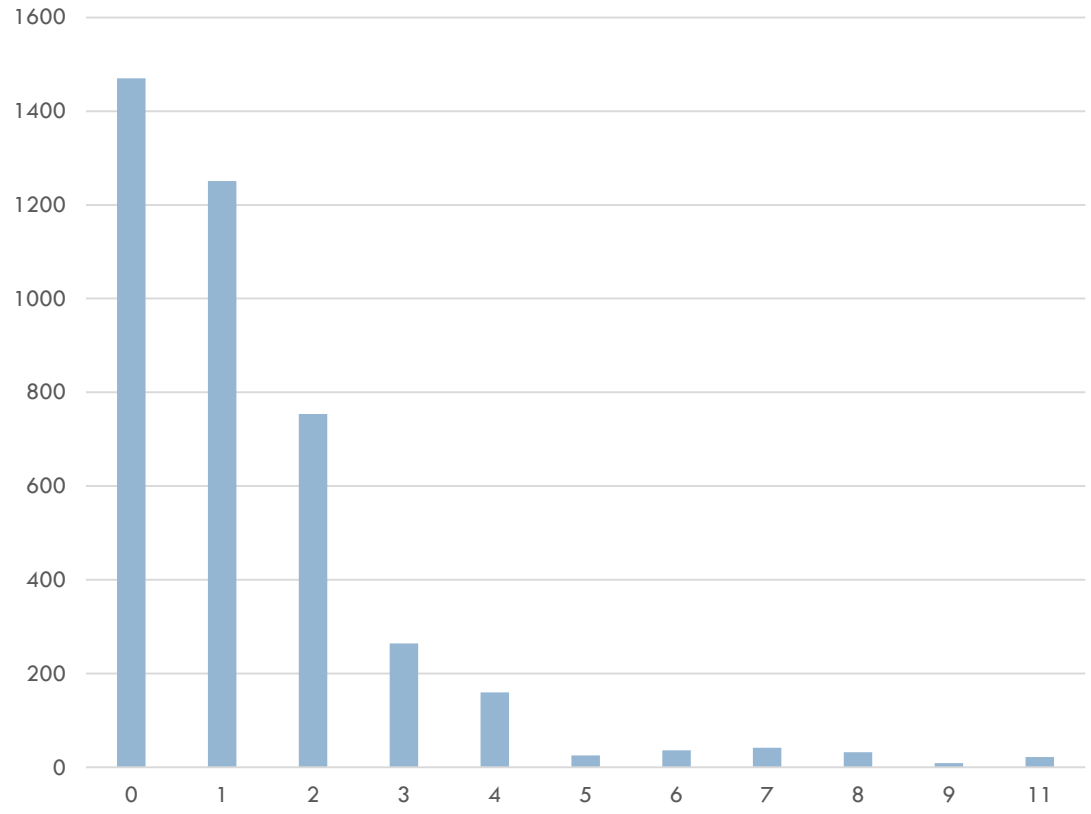
- 二、上层政区的匹配
  - 目前已有754年政区框架下的所有唐代州、县政区
  - 运用程序提取并补入隶属关系
    - 提取原则
  - 各郭ID明确隶属关系后，将其按照之前的规则导入CBDB

导入工作

- 三、地理坐标的确认
  - 3137个郭ID有地理坐标
    - 郭书与CBDB数据存在对应关系。且CBDB ID有坐标
  - 共有3250个郭ID没有地理坐标
    - 郭书中有而CBDB中没有的地名
    - 郭书与CBDB数据存在对应关系的地名，但CBDB没有相应的地理坐标
  - 运用Python从TGAZ上抓取地理坐标（by Yuying）
  - 运用Regex从文本中抓取地理信息



地理坐标匹配结果



匹配结果的比例