

地圖中文地名 AI萃取

報告人: 呂祥熙

- 1. MapKurator 簡介**
- 2. 如何應用 MapKurator**
- 3. 如何校正與存入資料庫**

1. MapKurator 簡介

AI歷史地圖文字偵測系統

<https://knowledge-computing.github.io/mapkurator-doc/#/>

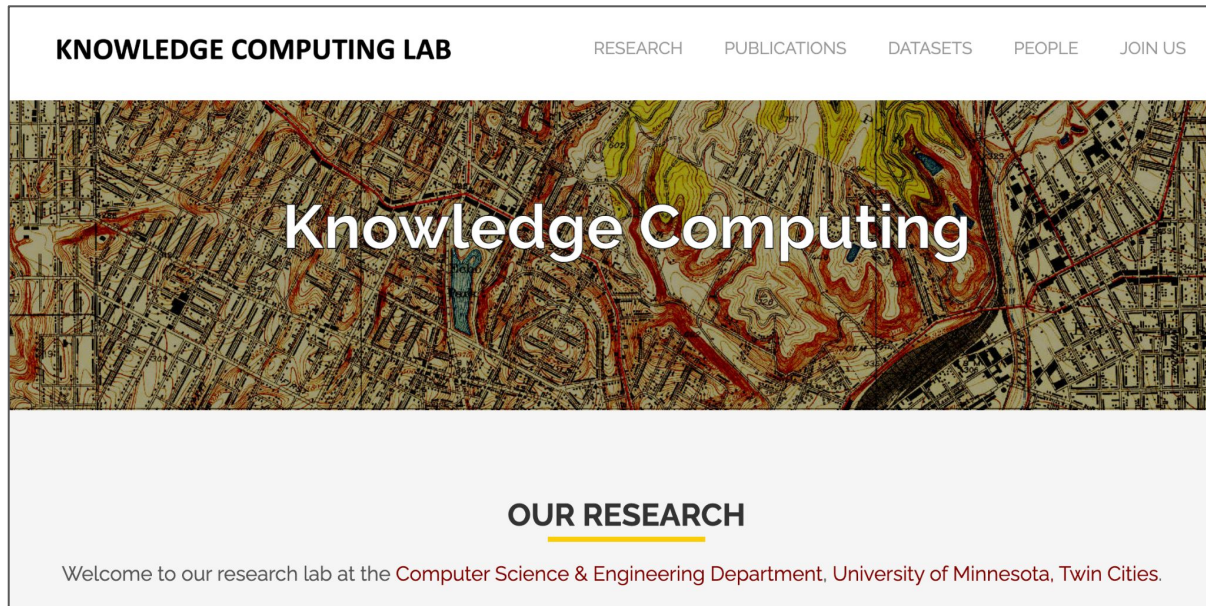
MapKurator 是一套自動化流程，
用深度學習是大範圍萃取歷史地圖上的文字。

Yao-Yi Chiang 蔣耀毅



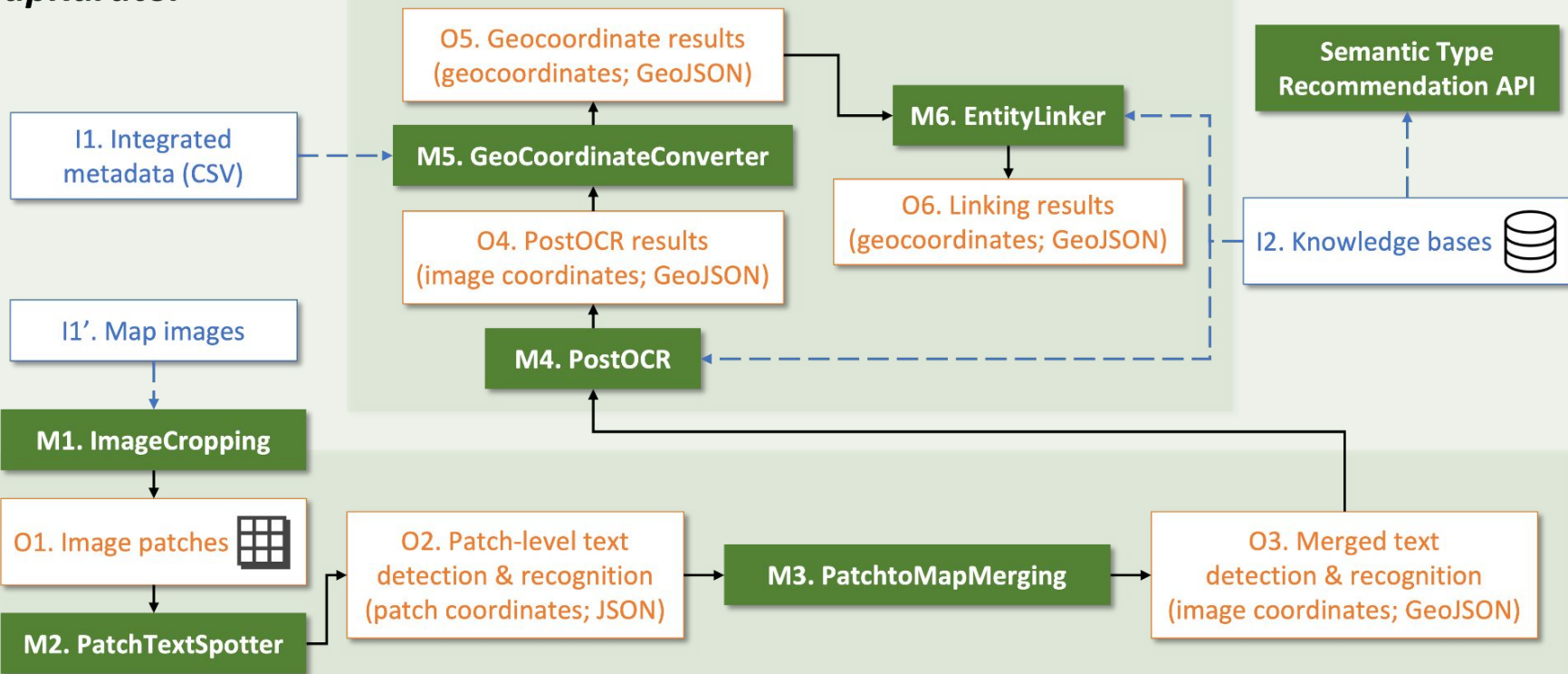
Contact

✉ yaoyi@umn.edu

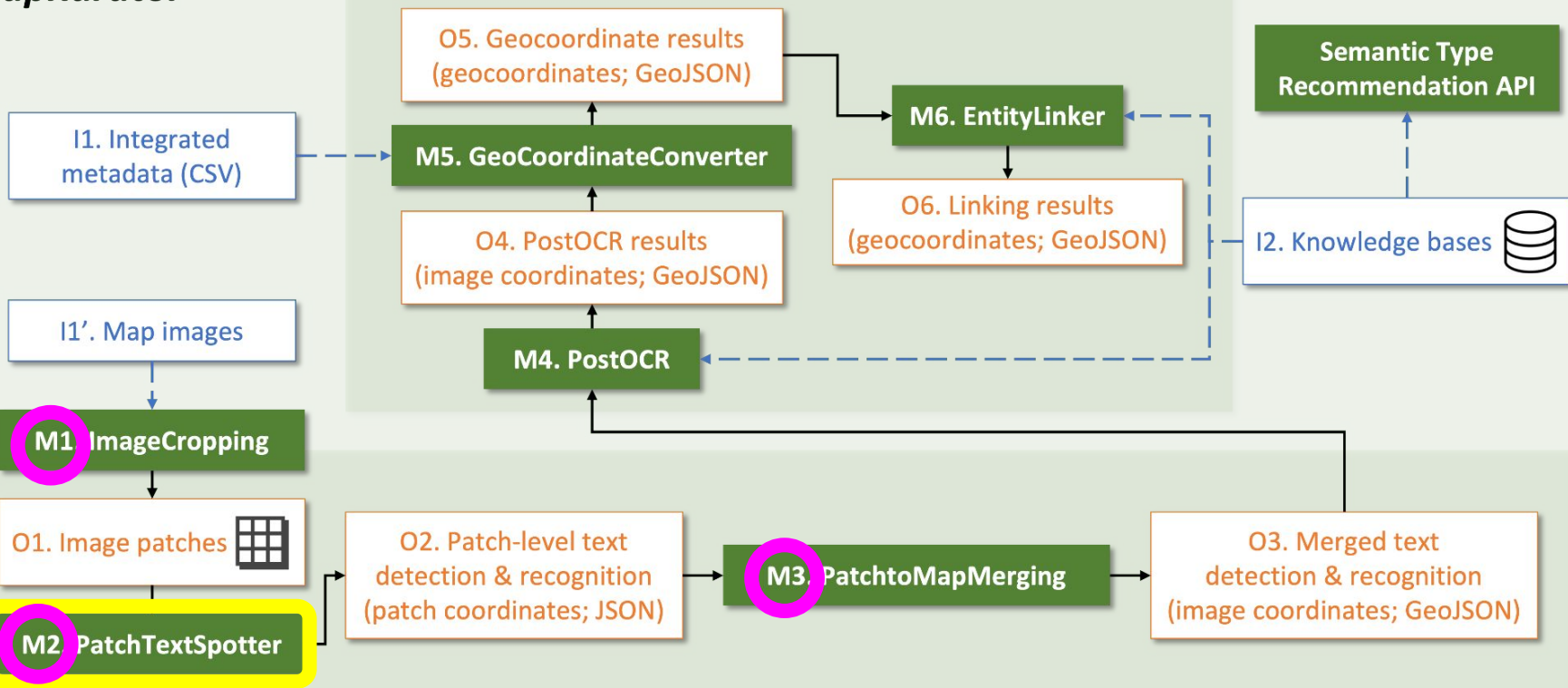


<https://knowledge-computing.github.io/>

mapKurator



mapKurator

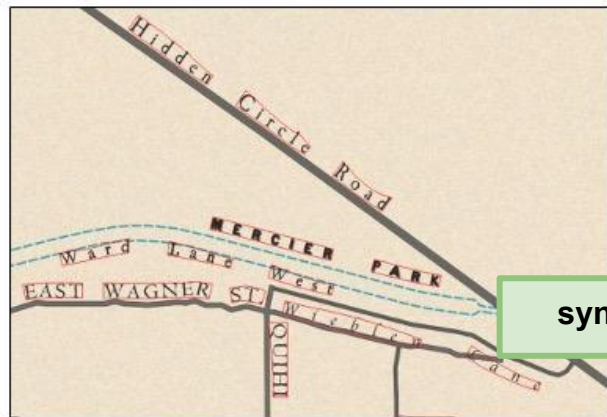


Patch Text Spotter



COCO dataset

An Example **SynthText** Image



synthetic datasets

An Example **SynMap** Image

Deformable Transformers

模型权重

2. 如何應用 MapKurator

大範圍偵測 Web Map Tile Service (WMTS) 圖層文字

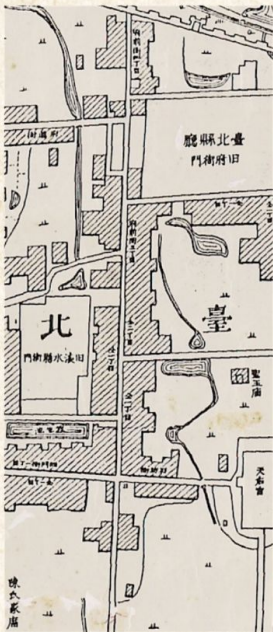


臺灣百年歷史地圖



地理資訊科學研究專題中心
Center for GIS, RCHSS, Academia Sinica

臺北



臺中



臺灣



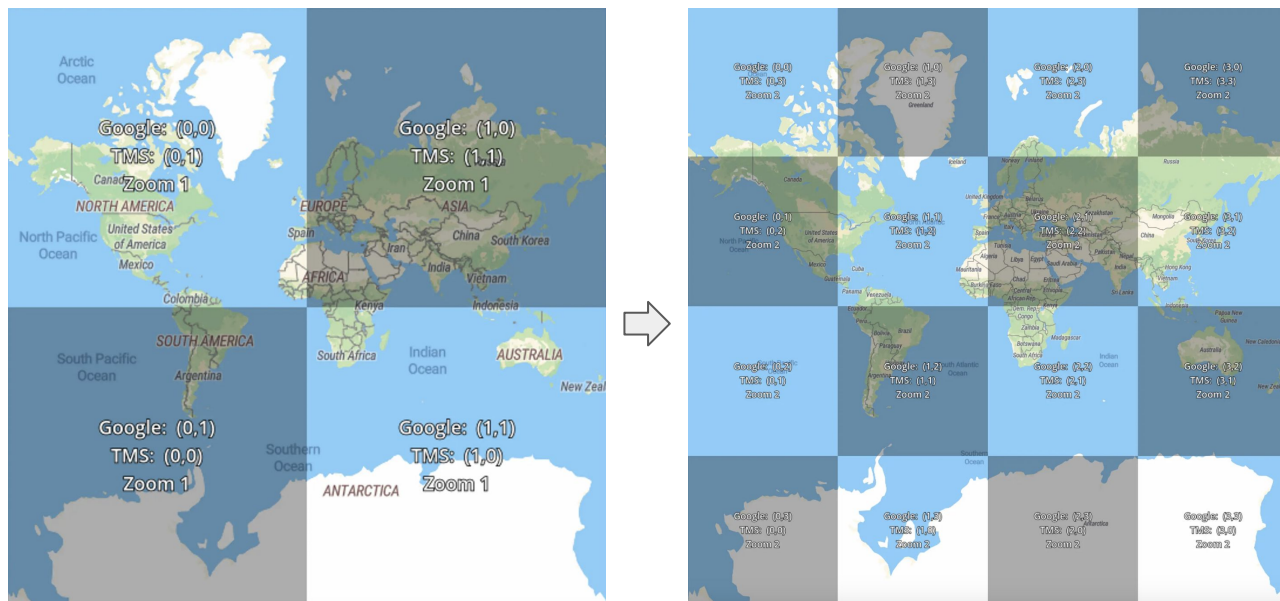
臺南



高雄

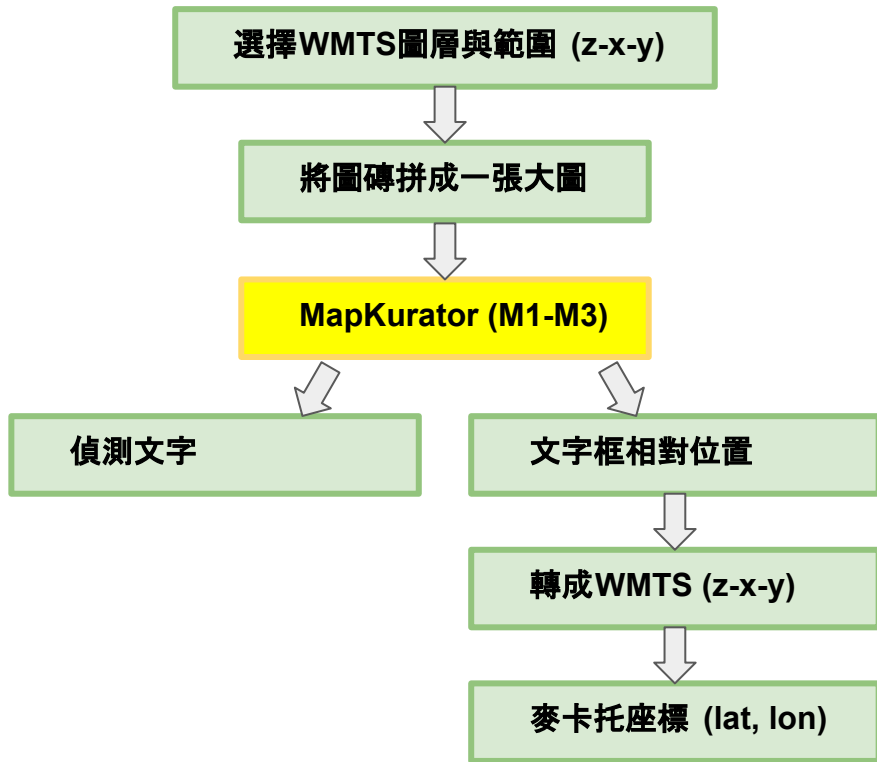


WMTS 影像金字塔原理



Z: 縮放層級 (zoom level); X, Y: 圖磚座標 (X 水平編號、Y 垂直編號)

使用流程



MapTextAI

選擇偵測語言 (Select a Language)

Chinese

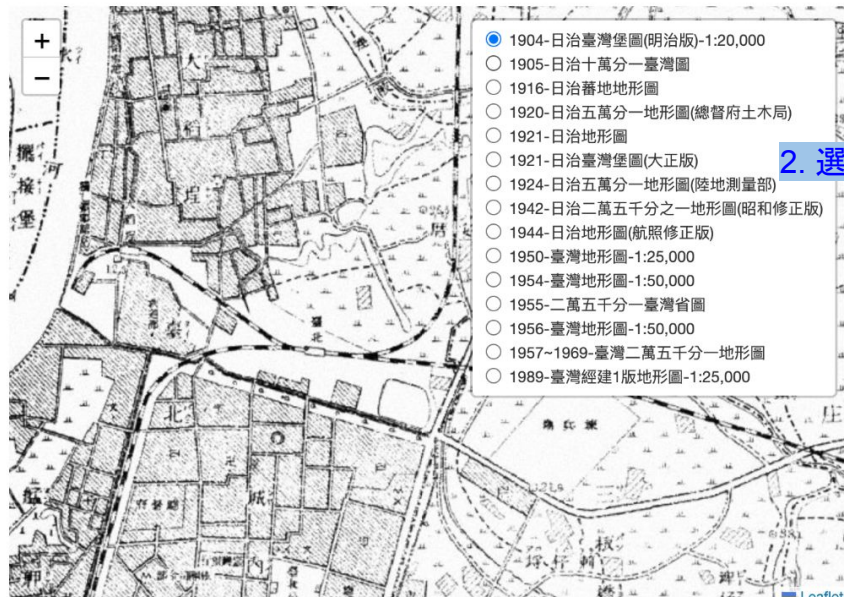
選擇偵測模組 (Select a Module)

百年歷史地圖 (WMTS)

You selected: 百年歷史地圖 (WMTS)

1. 選擇偵測語言與模組

Powered by Palette and the Pretrained
Palette Model licensed by Kartta
Foundation



2. 選擇WMTS地圖圖層

3. 選擇偵測範圍: 視窗內、台灣各區

No tile information available.



MapKurator

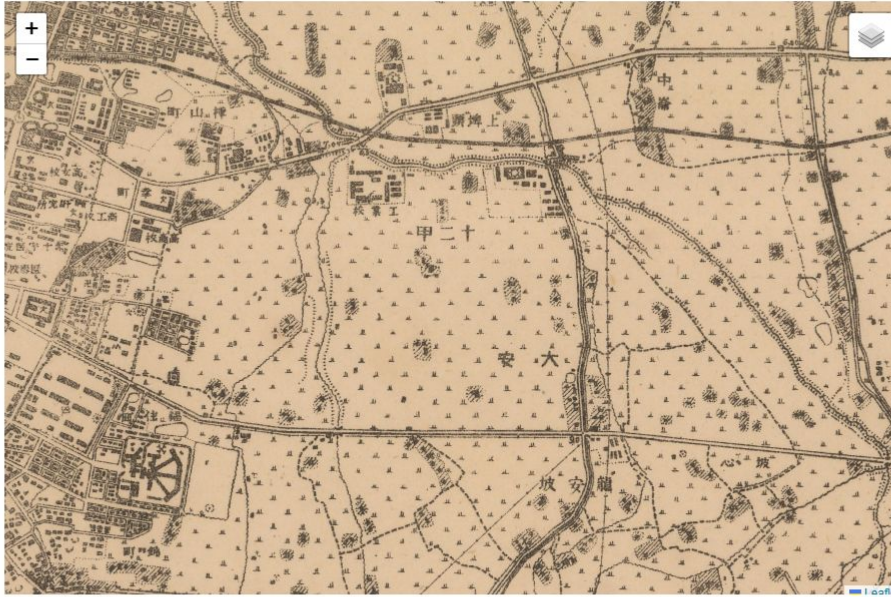
選擇偵測語言 (Select a Language)

Chinese

選擇偵測模組 (Select a Module)

百年歷史地圖 (WMTS)

You selected: 百年歷史地圖 (WMTS)



Get Cropped Map

4. 跑模型中

Get North TW

Get South TW

Zoom Level

15

X Tiles: 27443 to 27448

Y Tiles: 14026 to 14029

Layer: 1921-日治地形圖

Running `upload_and_send_data(...)`.



MapKurator

選擇偵測語言 (Select a Language)

Chinese

選擇偵測模組 (Select a Module)

百年歷史地圖 (WMTS)

You selected: 百年歷史地圖 (WMTS)

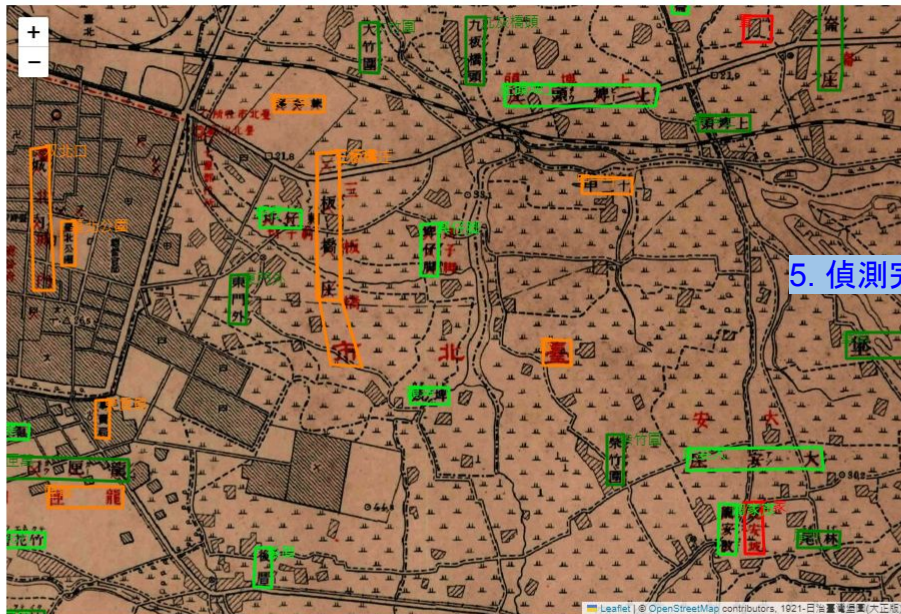
Zoom Level

15

X Tiles: 27443 to 27447

Y Tiles: 14026 to 14030

Layer: 1921-日治臺灣堡圖(大正版)



5. 偵測完可在互動式地圖預覽偵測結果

文字偵測清單 (List of Spotted Text)

[...]

Sort by Score

6. 下載JSON偵測檔

7. 最後將不同區的JSON合併成一個偵測檔

下載偵測檔 (Download JSON)

實際操作

```
"type": "FeatureCollection",
  "features": [
    {
      "geometry": {
        "coordinates": [...],
        "type": "Polygon"
      },
      "properties": {
        "rec_score": [...], 個別文字 confidence score
        "score": 0.91423, 文字框 confidence score
        "text": "外埔",
      },
    },
  ],
}
```

3. 如何校正與存入資料庫

Post-OCR to Elasticsearch

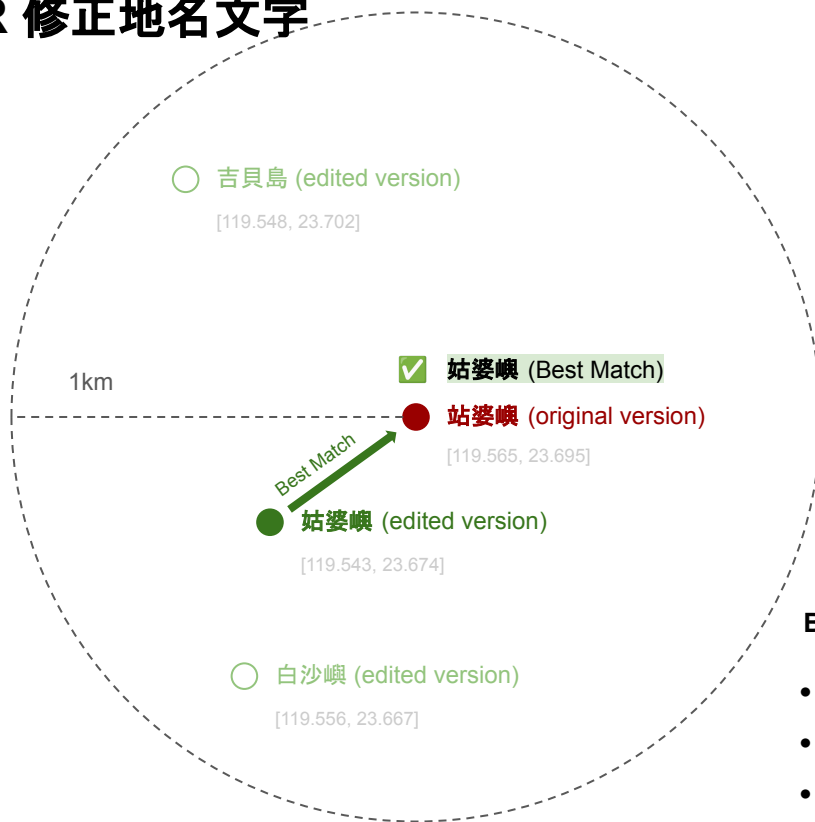
第一步：人工校對



第二步：Post-OCR 修正地名文字



第二步：Post-OCR 修正地名文字



Best Match = arg min(edit distance)

- Filter Search Radius: within a 1 km radius
- Filter Text : Fuzzy text similarity ($\geq 50\%$ match)
- Edit Distance: Levenshtein (0 = exact match)

第三步:存入 Elasticsearch

圖層名

正確地名

```
"_index": "jm25k_1921"  
"_source": {  
  "geometry": {"type": "Point", "coordinates": []}, 座標  
  "text": '火燒厝',  
  "year": 1921 地圖年份  
}
```

相關資料

MapKurator

<https://knowledge-computing.github.io/mapkurator-doc/#/>

Deformable Transformers

<https://arxiv.org/abs/2010.04159>

中央研究院臺灣百年歷史地圖 WMTS 服務

<https://gis.sinica.edu.tw/tileserver/>