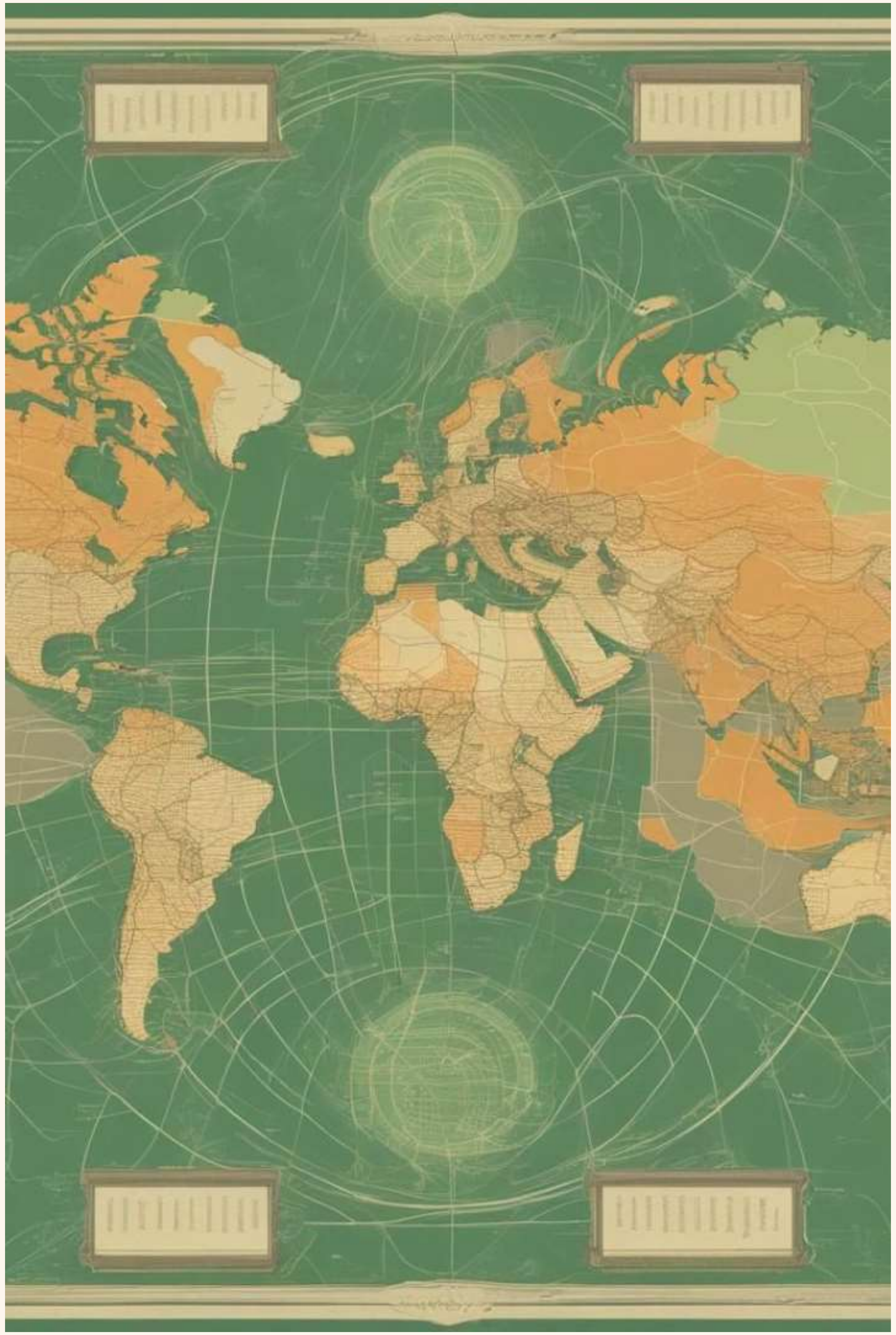
Automatic Datafication and Metadata Creation by AI in the Digital Humanities Research: A Case Study of Archives of Qing Secret Societies

Shu-Jiun (Sophy) Chen, Hsiang-An Wang, Hsi-yuan Chen
Institute of History and Philology, Academia Sinica, Taiwan
Academia Sinica Center for Digital Cultures, Taiwan

March 21, 2024

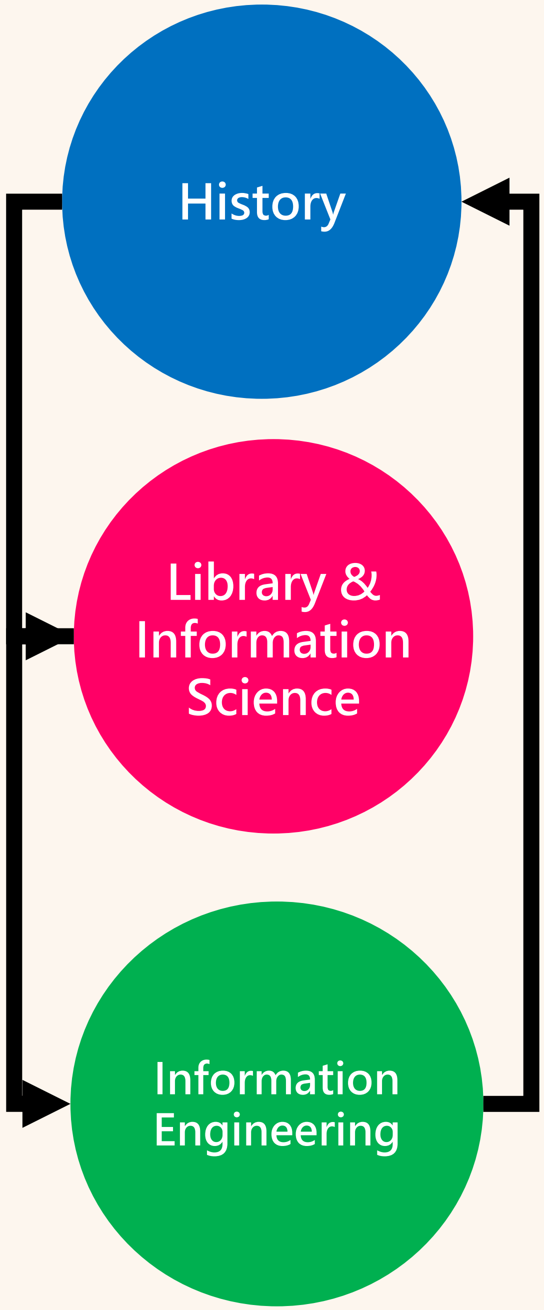
Networked Knowledge Organization Systems (NKOS) Workshop 2024,
at the 18th International Society for Knowledge Organization (ISKO) International Conference 2024
Mar. 20 - 22, 2024. Wuhan, CHINA



An Interdisciplinary DH Project of Archival Research

The Secret Societies in the Chinese Qing Dynasty

- **A humanities project** is initiated by the National Science Council of Taiwan.
- **Topic & Period**: **Official documents on the secret religious groups** of the Chinese Qing dynasty
- **Materials**: Archives of official documents collected in **National Palace Museum (Taipei), Institute of History and Philology at the Academia Sinica (Taipei), and the 1st Historical Archives of China (Beijing)**
- **Focus of Research**: ***Jiaqing* period** **Historical archives on the secret religious groups of the** ***Jiaqing* period** (1796-1820).



History

Library & Information Science

Information Engineering

It seeks to analyze historical data on case frequency, organizational size, geographic distribution, and mobilization methods using DH tools.

The Development of Knowledge Graph, Knowledge Organization System

o-Di

Aiming to transform archive content into structured data by tools of AI (ChatGPT 4.0), integrating data to closely mirror the historical reality.

Research Methods and Features of Archives

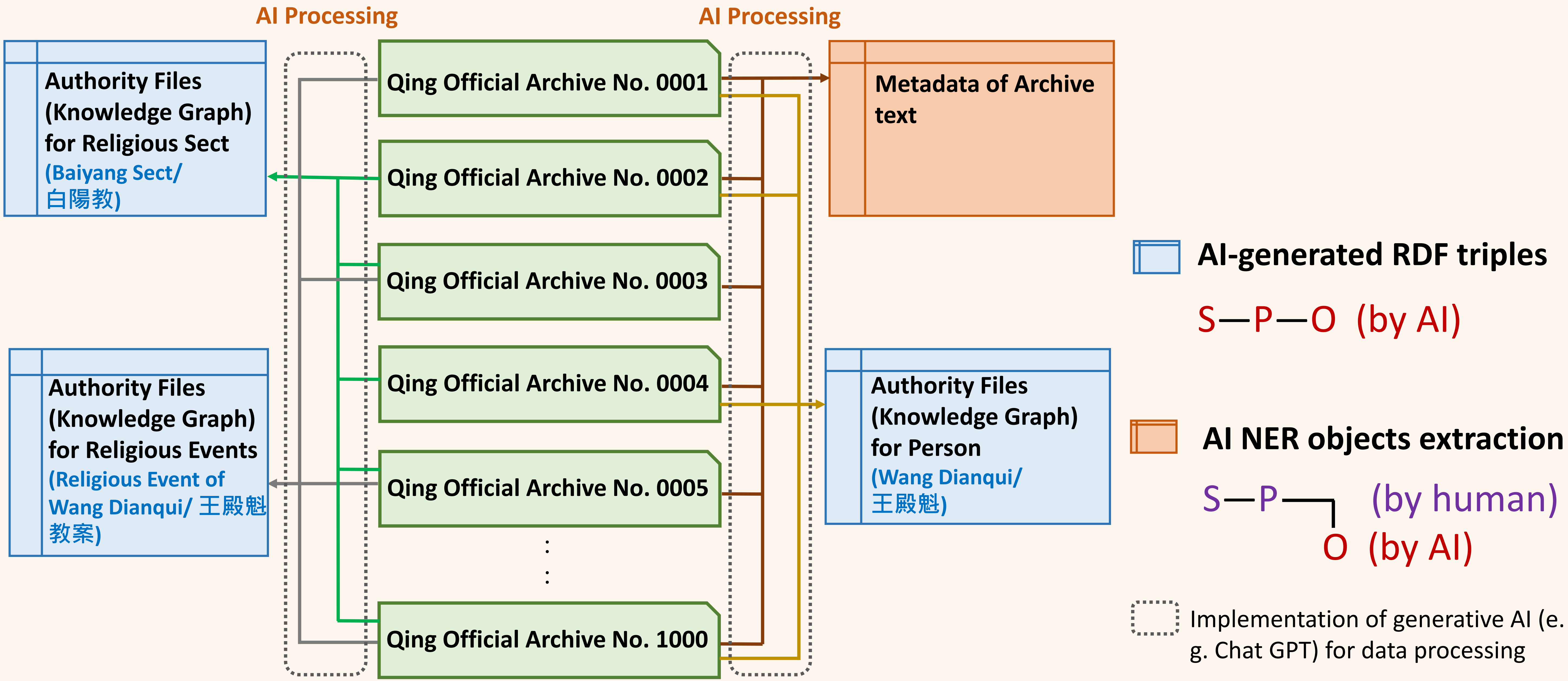
- Utilizing digital humanities approaches to process vast amounts of unstructured historical content.
- Extracting SPO triples using NER technology.
- Transforming text into machine-understandable, structured data in RDF framework.
- Facilitating automated data management and analysis.

Unstructured handwritten content

Names of common people

Lacking punctuation

Process of Automatic Data Creation: A Conceptual Workflow



AI Processing

AI Processing

Authority Files (Knowledge Graph) for Religious Sect (Baiyang Sect/ 白陽教)

Authority Files (Knowledge Graph) for Religious Events (Religious Event of Wang Dianqui/ 王殿魁教案)

Qing Official Archive No. 0001

Qing Official Archive No. 0002

Qing Official Archive No. 0003

Qing Official Archive No. 0004

Qing Official Archive No. 0005

:

:

Qing Official Archive No. 1000

Metadata of Archive text

Authority Files (Knowledge Graph) for Person (Wang Dianqui/ 王殿魁)

AI-generated RDF triples

S—P—O (by AI)

AI NER objects extraction

S—P— (by human)

O (by AI)

Implementation of generative AI (e. g. Chat GPT) for data processing

Process of Automatic Data Creation: AI-generated RDF Triples

After reading the text, ChatGPT automatically generates S-P-O data triples and incorporates the RDF structure.

你是一位精於古典文獻研究的圖書資訊學專家，下方文本是中國清代官方偵辦民間宗教案件的相關檔案。 (1)

請閱讀完下方文本，使用「思路鏈（Chain-of-Thought）」與「第一性原則」找出所有可能的主題。 (2)

同時針對每個主題進行RDF結構辨識，其結構固定為「實體-關係-實體」與「實體-屬性-屬性值」一組成對資訊。 (3)

一個主題內部可以包含不止一組資訊。

盡量使用能完整表達文意的最小詞彙來進行實體填寫，而topic、summary、關係、屬性、屬性值不在此限。

語言使用zh-tw，請注意結果只要輸出RDF格式，不要給我多餘非格式資訊。

```
[{
   "topic":"主題",
   "summary":"語意摘要",
   "實體-關係-實體":"(多值、隔開)",
   "實體-屬性-屬性值":"(多值、隔開)"
}]
```

prompt design in AI (ChatGPT)

Process of Automatic Data Creation: AI-generated RDF Triples

After reading the text, ChatGPT automatically generates S-P-O data triples and incorporates the RDF structure.

You are an expert in the field of library & information science and specialized in the classical literature research. The text below is related to the official investigation on the secret societies of the Qing Dynasty. **(1)**

Please read the text below and use “Chain-of-Thought” and “First Principles” to find all possible topics in the text. **(2)**

Meanwhile, please identify the RDF structure of each topic, and fix it as a set of information “entity-property-entity” and “entity-property-property value”. **(3)**

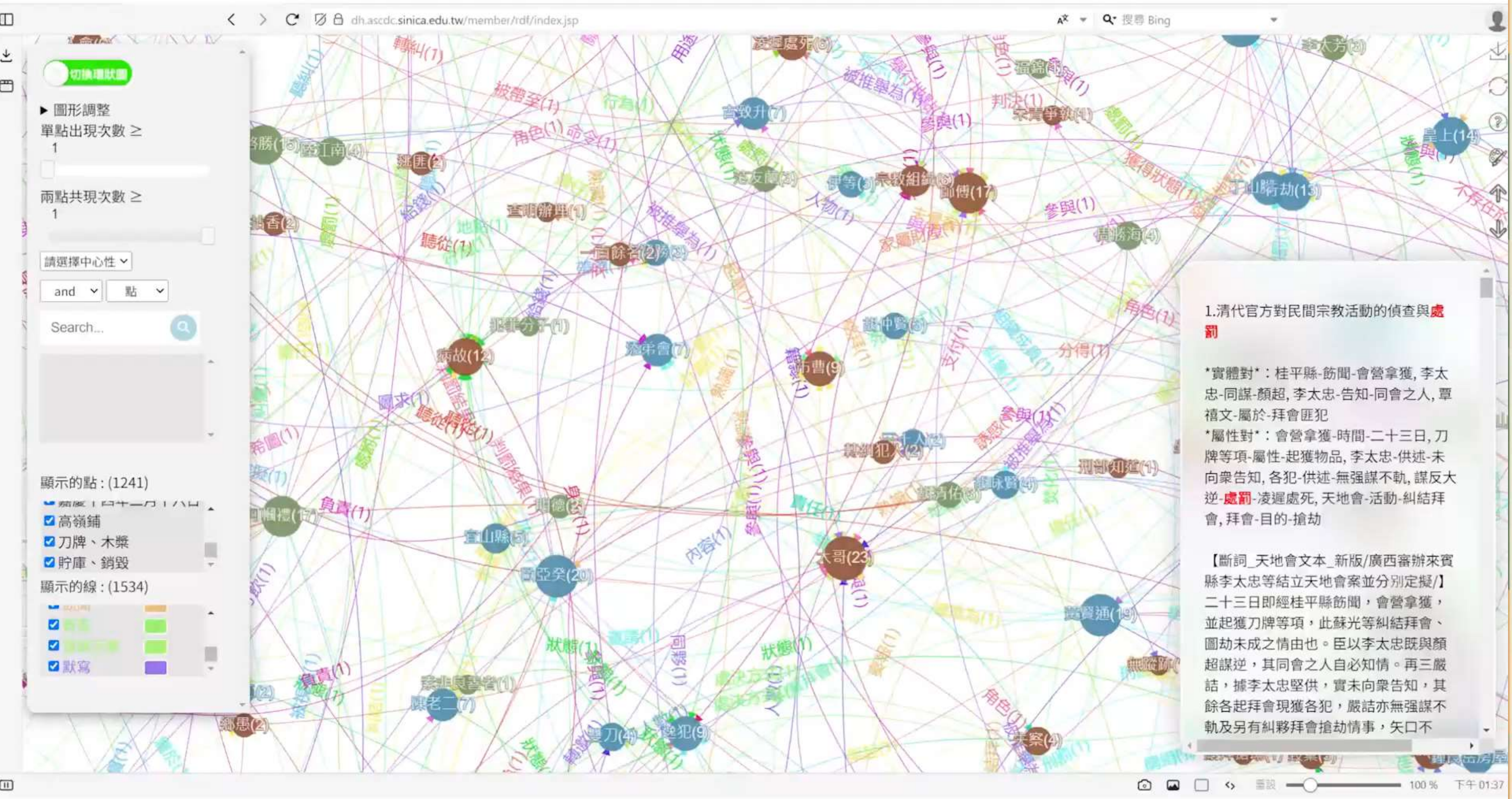
A topic can contain more than one set of information.

Try to use the smallest vocabulary that can fully express the meaning of the text to fill in the entities. Topic, summary, relationships, properties, and property values are not limited to this.

The language mark is specified as zh-tw. Please note that the results only need to be output in RDF format and do not response other unnecessary non-format information.

```
[{
    "topic":"topic",
    "summary":"Semantic summary",
    "Entity-Property-Entity":"(Multiple values, separated)",
    "Entity-Property-Property Value": "(Multiple values, Separated)"
}]
```

prompt design in AI (ChatGPT)



切換環狀圖

▶ 圖形調整

單點出現次數 ≥
1

兩點共現次數 ≥
1

請選擇中心性

and 點

Search...

顯示的點：(1241)

☑ 高嶺鋪

☑ 刀牌、木槳

☑ 貯庫、銷毀

顯示的線：(1534)

☑ 默寫

1.清代官方對民間宗教活動的偵查與處罰

實體對：桂平縣-飭聞-會營拿獲, 李太忠-同謀-顏超, 李太忠-告知-同會之人, 覃禧文-屬於-拜會匪犯
屬性對：會營拿獲-時間-二十三日, 刀牌等項-屬性-起獲物品, 李太忠-供述-未向衆告知, 各犯-供述-無强謀不軌, 謀反大逆-處罰-凌遲處死, 天地會-活動-糾結拜會, 拜會-目的-搶劫

【斷詞_天地會文本_新版/廣西審辦來賓縣李太忠等結立天地會案並分別定擬/】二十三日即經桂平縣飭聞，會營拿獲，並起獲刀牌等項，此蘇光等糾結拜會、圖劫未成之情由也。臣以李太忠既與顏超謀逆，其同會之人自必知情。再三嚴詰，據李太忠堅供，實未向衆告知，其餘各起拜會現獲各犯，嚴詰亦無强謀不軌及另有糾夥拜會搶劫情事，矢口不

Accuracy and Reliability Questions of AI-Generated RDF Triples

- **Challenges of "hallucinations" and "randomness" by using GPT:**
  How can we ensure the data quality and credibility that extracted by using ChatGPT? AI hallucination in this study belongs to the type of “factuality hallucination”. Its Reason can be attributed to the “**Knowledge Boundary**” in LLM and “**Domain Knowledge Deficiency**” (Huang et al. 2023).

- **Relying on GPT to determine important information:**
  How to address the information that missed by AI-Generated RDF triples?

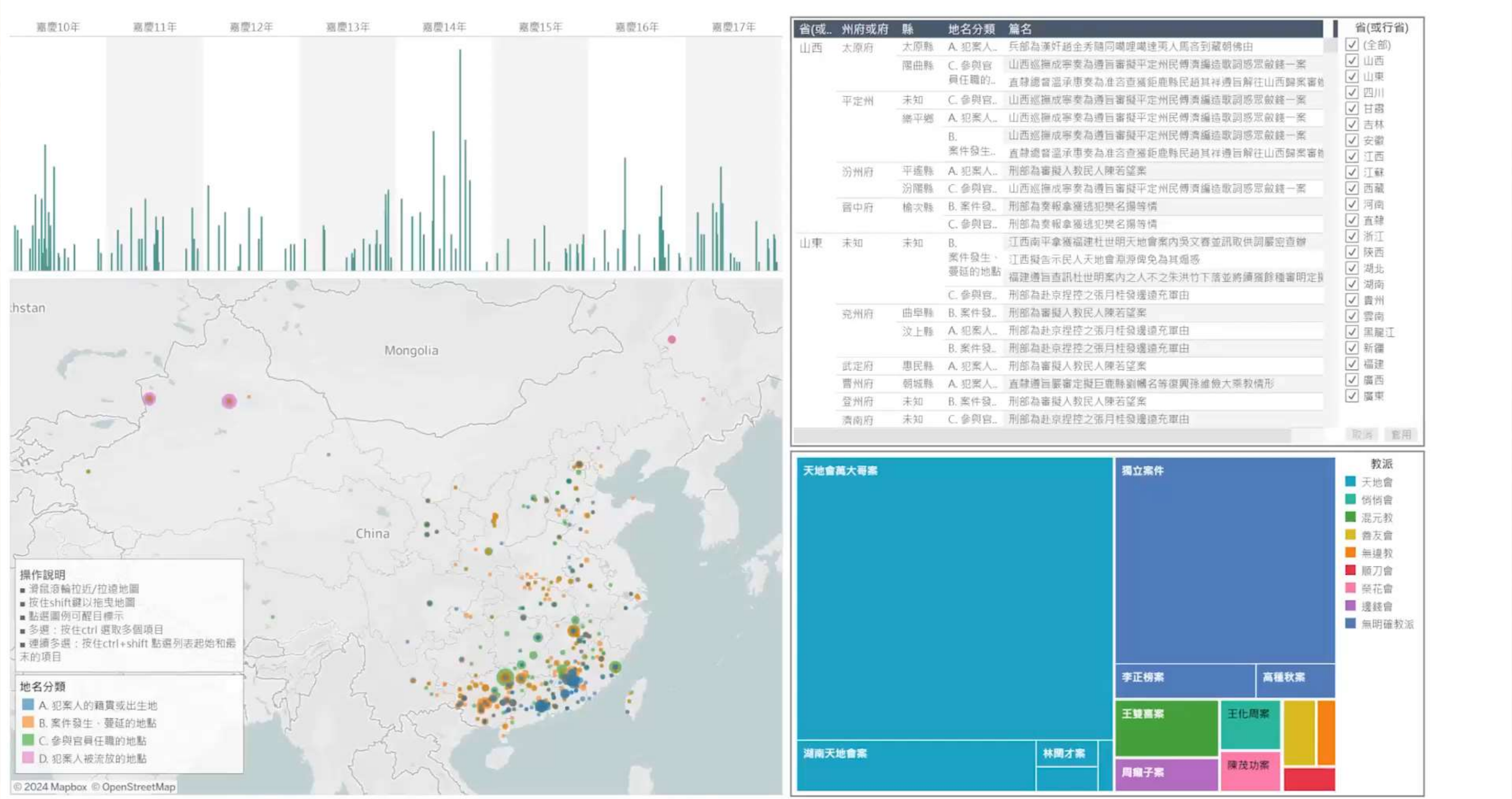
Strategies to Improve Accuracy and Reliability of Using AI

- AI hallucinations might occur when there are no corresponding and relevant texts in the existed LLM training data. **Providing original text** can reduce the incorrect AI responses.
- **Adjusting AI temperature in GPT response** by calling the API. AS the temperature is set to zero, randomness of the answer will be minimized.
- **Specifying AI responses in RDF format** to reduce the randomness and make machine easier to filter out meaningless answers.
- **Adding terminologies** as "Chain-of-Thought" and "First Principles" into prompt that makes ChatGPT incline to generate professional answers.
- **Using method of AI NER Objects extraction** to supplement detailed information that is missed by ChatGPT.

Optimizing Methods for AI-Generated Data

- Employing method of "**AI NER Objects Extraction**" for accurate information extraction based on detailed prompts.
    - Compared to AI-generated RDF triples, the AI NER Objects extraction approach requires deeper textual understanding, producing cleaner and more organized data.
- Combining "**AI-generated RDF Triples**" and "**AI NER Objects Extraction**" for a comprehensive approach, compensating for each other's limitations.
    - Flexibly choosing methods based on research needs and current circumstances for mutual enhancement.

Preliminary Research Findings : Historical Thematic Map of Religious Events



省(或..	州府或府	縣	地名分類	篇名
山西	太原府	太原縣	A. 犯案人..	兵部為漢奸趙全秀隨同噶哩噶達喇人馬吝到藏朝佛由
		陽曲縣	C. 參與官員任職的..	山西巡撫成寧奏為遵旨審擬平定州民傅濟編造歌詞惑眾斂錢一案
				直隸總督溫承惠奏為准咨查獲鉅鹿縣民趙其祥遵旨解往山西歸案審
	平定州	未知	C. 參與官..	山西巡撫成寧奏為遵旨審擬平定州民傅濟編造歌詞惑眾斂錢一案
		樂平鄉	A. 犯案人..	山西巡撫成寧奏為遵旨審擬平定州民傅濟編造歌詞惑眾斂錢一案
			B. 案件發生..	山西巡撫成寧奏為遵旨審擬平定州民傅濟編造歌詞惑眾斂錢一案
				直隸總督溫承惠奏為准咨查獲鉅鹿縣民趙其祥遵旨解往山西歸案審
	汾州府	平遙縣	A. 犯案人..	刑部為審擬人教民人陳若望案
		汾陽縣	C. 參與官..	山西巡撫成寧奏為遵旨審擬平定州民傅濟編造歌詞惑眾斂錢一案
	晉中府	榆次縣	B. 案件發..	刑部為奏報拿獲逃犯樊名揚等情
			C. 參與官..	刑部為奏報拿獲逃犯樊名揚等情
山東	未知	未知	B. 案件發生、蔓延的地點	江西南平拿獲福建杜世明天地會案內吳文春並訊取供詞嚴密查辦
				江西擬告示民人天地會淵源俾免為其煽惑
				福建遵旨查訊杜世明案內之人不之朱洪竹下落並將續獲餘種審明定
			C. 參與官..	刑部為赴京控控之張月桂發邊遠充軍由
	兗州府	曲阜縣	B. 案件發..	刑部為審擬人教民人陳若望案
		汶上縣	A. 犯案人..	刑部為赴京控控之張月桂發邊遠充軍由
			B. 案件發..	刑部為赴京控控之張月桂發邊遠充軍由
	武定府	惠民縣	A. 犯案人..	刑部為審擬人教民人陳若望案
	曹州府	朝城縣	A. 犯案人..	直隸遵旨嚴審定擬巨鹿縣劉幗名等復興孫維儉大乘教情形
	登州府	未知	B. 案件發..	刑部為審擬人教民人陳若望案
	濟南府	未知	C. 參與官..	刑部為赴京控控之張月桂發邊遠充軍由

省(或行省)

- ☑ (全部)
- ☑ 山西
- ☑ 山東
- ☑ 四川
- ☑ 甘肅
- ☑ 吉林
- ☑ 安徽
- ☑ 江西
- ☑ 江蘇
- ☑ 西藏
- ☑ 河南
- ☑ 直隸
- ☑ 浙江
- ☑ 陝西
- ☑ 湖北
- ☑ 湖南
- ☑ 貴州
- ☑ 雲南
- ☑ 黑龍江
- ☑ 新疆
- ☑ 福建
- ☑ 廣西
- ☑ 廣東

取消 套用

操作說明

- 滑鼠滾輪拉近/拉遠地圖
- 按住shift鍵以拖曳地圖
- 點選圖例可醒目標示
- 多選：按住ctrl 選取多個項目
- 連續多選：按住ctrl+shift 點選列表起始和最末的項目

地名分類

- ■ A. 犯案人的籍貫或出生地
- ■ B. 案件發生、蔓延的地點
- ■ C. 參與官員任職的地點
- ■ D. 犯案人被流放的地點

© 2024 Mapbox © OpenStreetMap

教派

- ■ 天地會
- ■ 俏俏會
- ■ 混元教
- ■ 善友會
- ■ 無邊教
- ■ 順刀會
- ■ 榮花會
- ■ 邊錢會
- ■ 無明確教派

Conclusion and Near Future

- **Using generative AI to extract named entities** from the historical text, to make **automatic metadata creation**, and compile **knowledge graphs**. It can improve the collection and generation of research data, allowing scholars to focus on research analysis and interpretation.
- **Extracting RDF triples** from different archives to be integrated and **reused to construct knowledge graphs in the way of distant- and close-reading**.
- Applying method of **prompt engineering to regulate AI responses in RDF format** to avoid problems as AI hallucination and randomness.
- Continuing to **create RDF knowledge graphs for the religious groups and religious events by using LLM-based AI** as demonstrative method of "human-machine collaboration" to conduct historical DH research and to construct of the “Research Platform on the Qing Secret Societies”.