

Beyond Boundaries: Integrating Humanities and Generative AI in Archival Resource Development

Dr Yaming Fu

School of Cultural Heritage and Information Management, Shanghai University



Prof. Simon Mahony

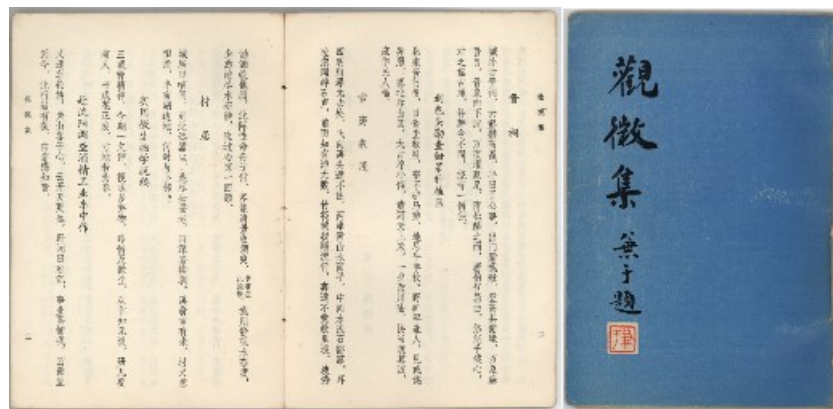
Department of Information Studies, University College London



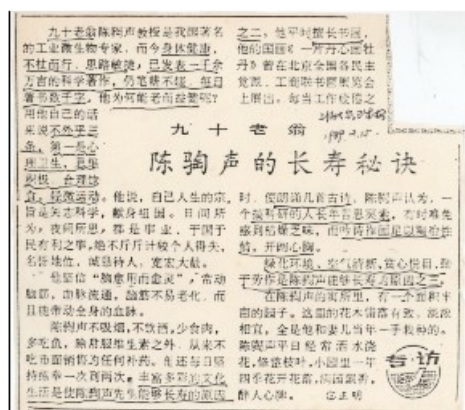
Archival Data in this Project

Fonts diverse and lack proofreading

Layout complex and manuscript type varied



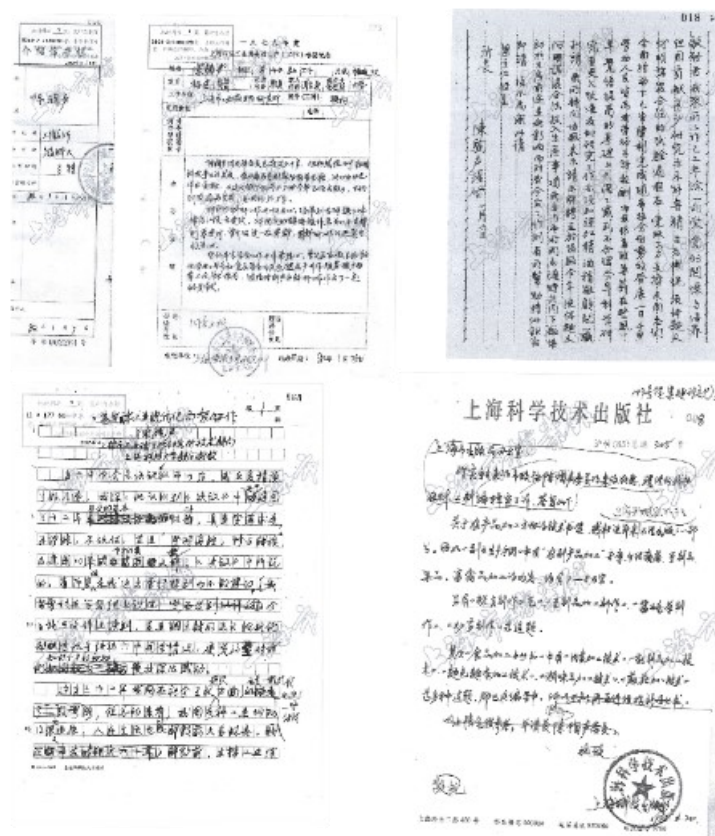
Unpublished poems



Newspapers



Photos



administrative archives, manuscripts



Paintings

Limitations of Traditional Archival Workflows



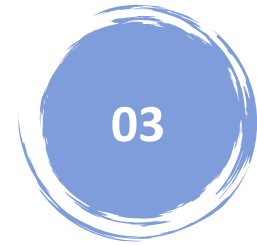
Partially focus on the surface content

The traditional development model only focuses on the surface layer of archives and pays little attention to the development and dissemination of archives as data.



Labour-consuming way of data identification and description

Diverse types of archival materials under the traditional development model require additional effort to identify content, and the archival metadata is manually recorded, which is labour intensive and time consuming.



Limited way of presentation

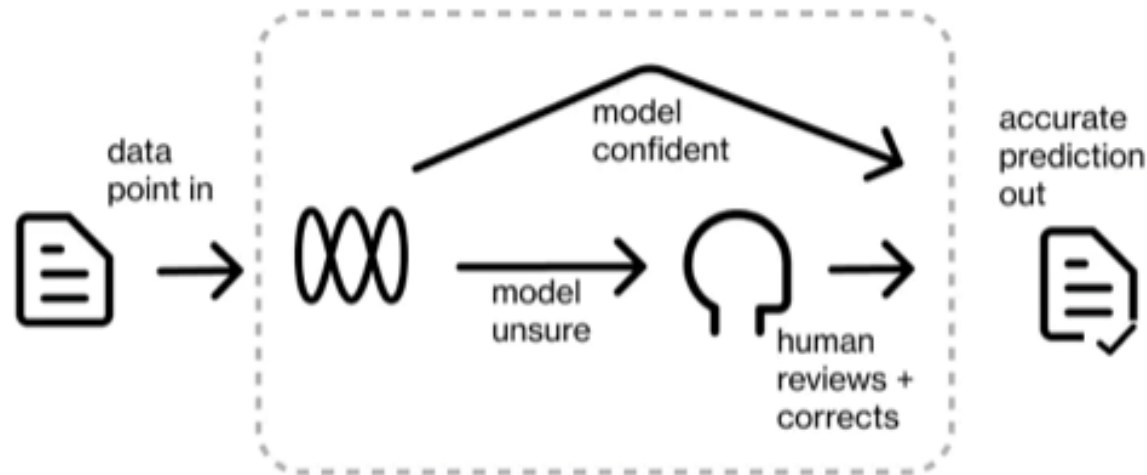
The traditional development mode presents archival resources mainly by display, and the interaction mode is relatively simple, and lacks visualisation and interaction.

Research Goals

- Design and implement an IIIF-based OCR environment to support archival transcription, recognition, comparison, and description. IIIF (International Image Interoperability Framework), as 'a set of open standards for delivering high-quality, attributed digital objects online at scale', could provide standardised image processing protocols, enhancing image-formatted archival data management, improving the efficiency and ensuring accuracy of archival data.
- Build Classic RAG and GraphRAG to deeply mine and analyse archival data, improve the interpretation ability of archival data with scanned archival images, and design the 'chat' function with AIGC, forming a new paradigm of archival usage.
- Design a human-machine collaborative framework for archival knowledge services based on the core concept of "human-in-the-loop" (HITL) to ensure credible control and trustworthy outcomes.

Human-in-the-Loop, HITL

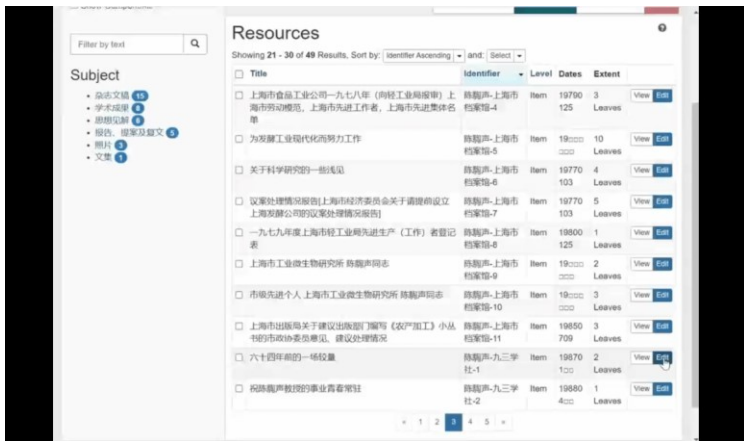
- The concept of Human-in-the-Loop (HITL) is to use human domain knowledge to facilitate the automation of machine learning and improve the accuracy of models.
- Archivists, as agents who possess professional knowledge in the field of archiving, play a significant role in the process of machine involvement in the management and development of scientists' archives.



Archival Data Management System

‘ArchivesSpace is an open-source, browser-based archival information management software application’. The underlying descriptive standard it uses is DACS (Describing Archives: A Content Standard).

ArchivesSpace allowed the users to add, arrange, describe, evaluate materials, manage locations, control permissions, and link subject tags when storing archival metadata. DACS includes descriptions of archival material and authoritative records representing the people and organisations that created them (SAA, 2022). DACS contains 25 archival elements, many encoded with multiple EAD tags.



著录项目				
Basic Information (基础信息)	英文条目	对应中文条目	约束性	补充说明
	Title	题名	必著	
	Identifier	档号	必著	
Languages (语言)	Level of description	著录层级	必著	
	Languages	语言	必著	
Dates (日期)	Label	标签	必著	描述档案的发表形式
	Expression	发布日期	必著	
	Type	日期类型	必著	
Extents (载体形态)	Certainty	确定性	必著	描述档案日期的确定性
	Portion	完整性	必著	描述当前著录的档案材料的数量
	Number	载体数量	必著	描述档案材料数量的单位
	Type	载体类型	必著	
Agent Links (关联人物或机构)	Role	人物或机构的作用	有则必著	
	Relator	人物或机与档案的关联	有则必著	描述实体对档案的贡献
	Agents	人物或机构	有则必著	与Agent关联，填写时直接选择或新建
Notes (注释)	Note Type	注释类型	选著	可以选择Custodial History (档案保管沿革) 和 Bibliography (参考文献)
User Defined (用户自定义)		人名	选著	
		稿本	选著	
		文种	选著	
		主题词和关键词	选著	
		附注	选著	

Archival data description elements

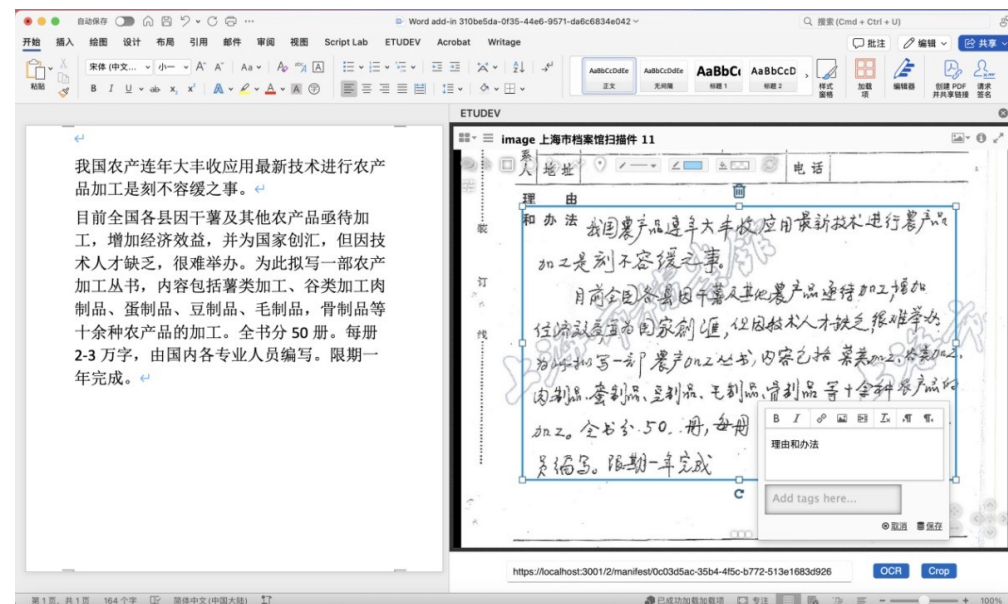


Large Archival Digital Images Solutions and Full Text Recongnition

IIIF was integrated with the OCR text editing plugin in the Office Word environment to realise a commonly used archival work environment and the seamless linking of raw images with their corresponding OCR texts, ensuring comprehensive documentation and accessibility.



IIIF Image API

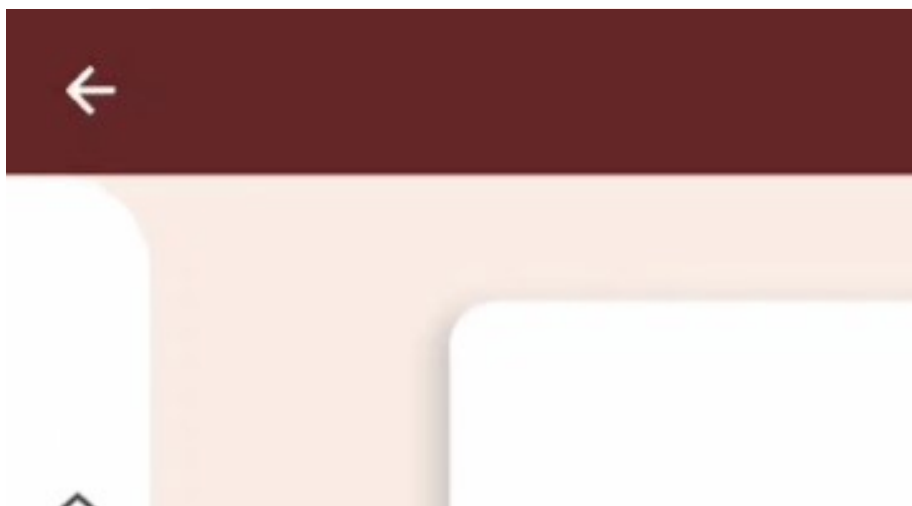


IIIF integrated Word environment
with OCR plugin

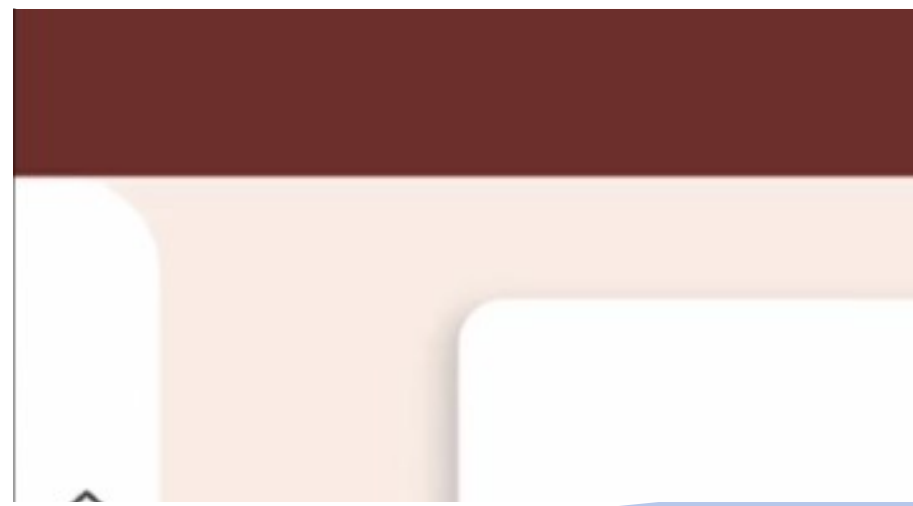
Archival Data AIGC Knowledge Base and LLM Solutions

Archival images (unstructured data) are transformed into structured and semi-structured data through the IIF-based OCR tool.

Traditional manual description tasks are converted into data mapping tasks that can be automatically completed by LLMs. The high-quality electronic archival database can serve as a private AIGC database for Classic RAG (Retrieval Enhanced Generation) or GraphRAG needed by LLMs. This enables advanced content mining, content analysis, and other tasks that unlock the value of archival data.



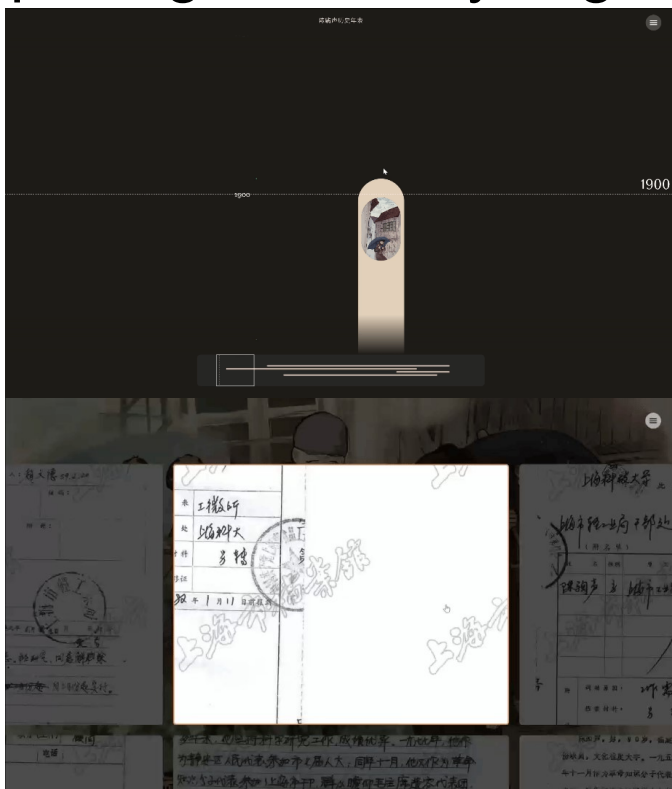
ClassicRAG



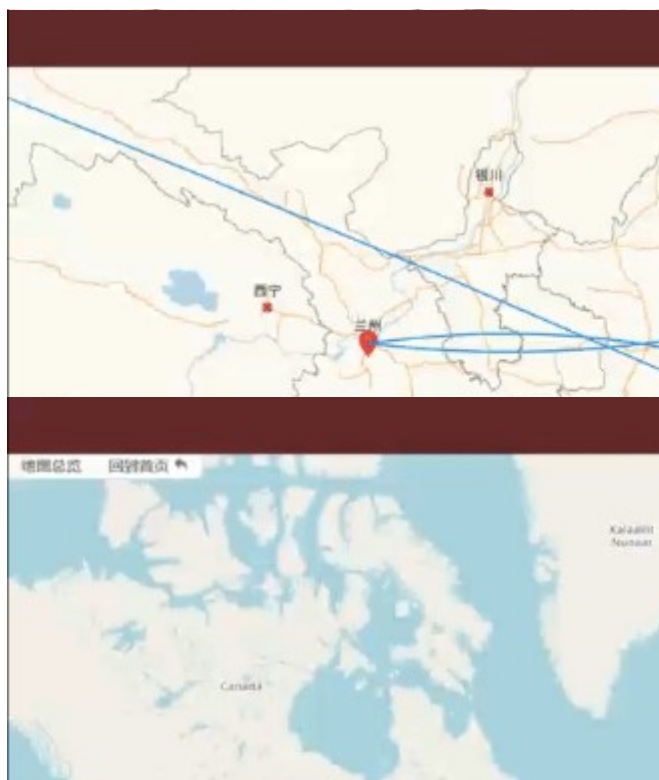
GraphRAG

Visualisation Solutions

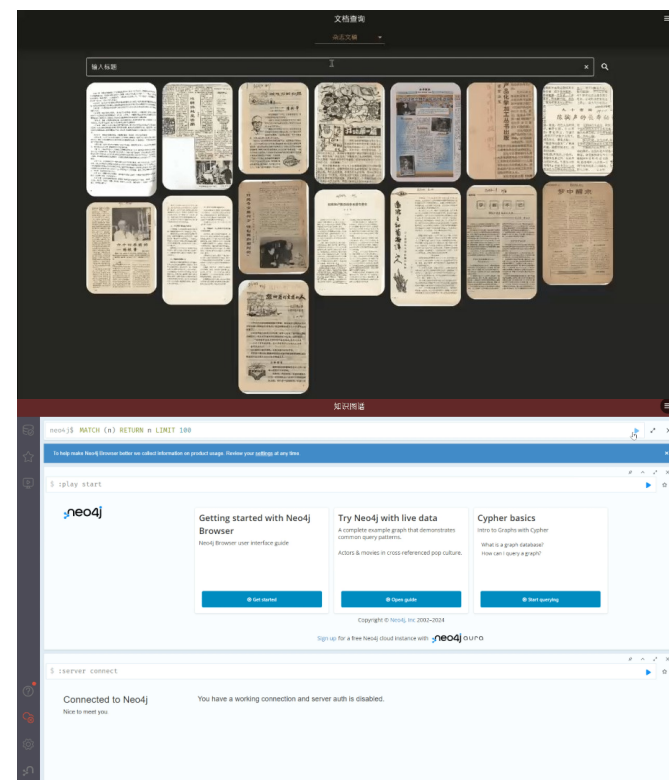
Organise scientists' archives from different sources, provide multiple visualisation solutions for valuable archival content, and support future researchers in scientifically exploring and analysing these archives.



Timeline, Photo wall




GIS Map, Story Map



Search Function, KG

Conclusion

- With the advent of the AGI era, the workflow for managing archives and records has undergone a profound transformative.
 - By embedding human expertise into the key stages of data processing and knowledge generation, the proposed model emphasises the accuracy, evidentiary value, and historical contextuality of archival content. Professional judgment and manual review by archival staff play a critical role, ensuring that the generated knowledge aligns with archival standards and retains its cultural and historical authenticity.
 - Ensuring the accuracy and trust of archive documents requires robust validation mechanisms, where human oversight plays a critical role in verifying the outputs generated by automated systems.
- 

References

Rolan, Gregory, et al. "More human than human? Artificial intelligence in the archive." *Archives and Manuscripts* 47.2 (2019): 179-203.

Emmert-Streib, Frank. "Is ChatGPT the way toward artificial general intelligence." *Discover Artificial Intelligence* 4.1 (2024): 1-8.

Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." *arXiv preprint arXiv:2311.02462* (2023).

Naveed, Humza, et al. "A comprehensive overview of large language models." *arXiv preprint arXiv:2307.06435* (2023).

Bukhari, Syed Saqib, et al. "anyocr: An open-source ocr system for historical archives." *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE, 2017.

References

Koch, Inês, et al. "Knowledge graph implementation of archival descriptions through CIDOC-CRM." International conference on theory and practice of digital libraries. Cham: Springer International Publishing, 2019.

Philips, James P., and Nasseh Tabrizi. "Historical document processing: historical document processing: a survey of techniques, tools, and trends." arXiv preprint arXiv:2002.06300 (2020).

Link: <https://iiif.io>

Link: <https://archivesspace.org>

SAA. Describing Archives: A Content Standard. Society of American Archivists' Technical Subcommittee, https://saa-ts-dacs.github.io/dacs/06_part_I/02_chapter_01.html (2022).

Omrani, Pouria, et al. "Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement." 2024 10th International Conference on Web Research (ICWR). IEEE, 2024.