



Information Research – Vol. 30 No. iConf (2025)

# Innovative practice of archival data development workflow in the AGI era: a case study of scientist archives project

Yaming Fu, Jie Song, Xinran Zhang, and Jingyun Bi

DOI: <https://doi.org/10.47989/ir30iConf47335>

## Abstract

**Introduction.** The advent of large language models (LLMs) presents a transformative opportunity for the field of archival science, offering advanced capabilities in intelligent information processing, semantic search, and more. These innovations address critical challenges posed by the exponential growth in archival materials and the increasing demand for efficient data analysis. Traditional archival workflows, often rely on manual description and optical character recognition (OCR), struggle with the complexities of unstructured digital data, especially in the context of digitized historical archives and manuscripts.

**Method.** This paper explores a novel archival workflow through the case study of the scientist archives project, integrating human-machine collaboration and leveraging technologies such as open-sourced archival databases, IIF-supported OCR environments, and advanced LLMs, including classic retrieval-augmented generation (Classic RAG) and graph-based retrieval-augmented generation (GraphRAG).

**Analysis.** A Scientist Archives project is analysed, which utilises AGI era technologies to mine and manage the archives.

**Results.** By embracing these technologies, the proposed approach seeks to revolutionize archival management, enhancing both efficiency and the depth of content revelation in the AGI era.

**Conclusion(s).** This study contributes to the ongoing discourse on the intelligent transformation of archival practices, providing a roadmap for future archival data mining and management.

## Introduction

Since 2018, the archival community has actively discussed the impact of artificial intelligence and digital technologies, and research on the digitalization and intelligent transformation of archival management has attracted much attention (Rolan, 2021). By late 2022, the release of ChatGPT marked a significant milestone in the development of large language models (LLMs), fueling growing discussions and heightened expectations for the era of artificial general intelligence (AGI) (Emmert-Streib, 2024). As a core and controversial concept in the field of computer research, AGI refers to the AI systems that exhibit comparable intelligence to human intelligence in a wide range of tasks and situations, and such systems usually have general task-solving capabilities and continuous self-learning capabilities and ‘*exhibit brain-like intelligence characteristics in cognition and decision-making*’ (Morris, 2023). LLM shows versatility in natural language processing, provides a window for the public to approach AGI, and serves as an important milestone on the road to AGI. As a generative natural language processing model for large-scale corpus training, ‘*LLM has the functions of automatic text generation, intelligent information processing, semantic search and recognition, intelligent image generation, etc.*’ (Naveed, 2023), which provides a promising research direction for archival data mining and archival management, and triggers deep thinking on a more ‘*intelligent*’ transformation of archival work.

The surge in archival materials, the growing demand for efficient methods of analysing archival data and in-depth archival mining have brought unprecedented challenges to traditional archival workflow. Manual description processes are time-consuming, labour-intensive, and prone to errors, particularly when dealing with large volumes of archival materials. Optical character recognition (OCR) software is widely used to extract text from digital images, facilitating the creation of searchable text files (Bukhari, 2017). However, digital archival data is often unstructured, complicating retrieval and utilization. This is particularly true for digitized historical archives and manuscripts, which are difficult to recognize, analyse, and describe. Moreover, traditional workflows fall short in granular content revelation, failing to meet users' diverse information needs in the new era.

This project uses the archival materials of Chen Taosheng (In Chinese: 陈陶声), the founder and pioneer of modern industrial microbiology in China, as a case study to explore a novel archival workflow that integrates human-machine collaborative models in managing scientist archives. The proposed workflow leverages advanced technologies, including an open-sourced archival database, the international image interoperability framework (IIIF) supported work environment, and LLMs, particularly classic retrieval-augmented generation (Classic RAG) and graph-based retrieval-augmented generation (GraphRAG). The project aims to revolutionize archival workflows, enhancing the overall efficiency and quality of archival management in the era of AGI.

## Problem statement

Traditional archival workflows are insufficient to meet the demands of modern archives management. This is not only due to the unreliable manual processing and cataloguing but also the fragmented use of independent tools for different archival tasks. These tools include but are not limited to ‘*metadata management tool, OCR tool, ontology design tool and knowledge graph tool*’ (Koch, 2019; Philips, 2020). In addition, traditional workflows lack the granularity needed to reveal the full depth of archival content, but in the meantime, the authenticity and security of the archival data remain, which is key to archival management. Therefore, the research problems of this project are:

First, design and implement an IIIF-based OCR environment to support archival transcription, recognition, comparison, and description. IIIF, as ‘*a set of open standards for delivering high-quality, attributed digital objects online at scale*’ (Link: <https://iiif.io>), could provide standardized image

processing protocols, enhancing image-formatted archival data management, improving the efficiency, and ensuring accuracy of archival data.

Second, formulate a workflow for managing scientists' archives, organize scientists' archives, photos, literary works, monographs, patents, and relevant documents, index them in the open-sourced archival workspace ArchivesSpace, and integrate them with LLM tools to facilitate efficient access and utilization.

Third, 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 utilization.

With the above three research goals, the project aims to develop an AGI-Era Archival workflow solution, an efficient and reliable framework to promote the intelligent processing of archival description, analysis, development, and utilization.

## Methodology

### Research process

This work originated in early 2023 with the call to collect, organize and preserve the archival materials of scientists, especially the ones that were kept by individuals and the ones that were in bad condition. The aim was to collect, digitise, analyse, and make use of the scattered and deteriorating archival materials related to the lives and work of scientists, as well as to excavate the values hidden in the archives with available technologies. In October 2023, a project member interviewed a close friend of Chen Taosheng, a founder and pioneer of modern industrial microbiology in China and a key figure in establishing the department of bioengineering at the predecessor of Shanghai University. During this interview, the project member obtained valuable family archival materials. Subsequently, project members visited the Shanghai municipal archives, and the Shanghai University archives, acquiring more materials related to Chen Taosheng. With the number and types of archival materials growing, an urgent need has emerged to design an efficient workflow to manage, describe, interpret, and analyse those archival materials safely without damaging the original files' quality and meaning. Thus, the project takes Chen Taosheng's archival materials as a case study to explore the development of an archival workflow for archival data in the AGI era.

An innovative archival development workflow in the AGI era should enhance efficiency while maintaining data authenticity, accuracy, and security. This project, an AI-integrated working environment based on the IIIF framework and embedded with OCR tools, was designed to facilitate processing image-based archival data. It also utilized LLM technology to assist in analysing massive archival raw images and their corresponding texts. Classic RAG and GraphRAG were built to realize the 'chat' function with the scientist and form scientists' knowledge graphs, scientists' footprint maps, and scientists' chronology with the help of artificial intelligence, thereby preserving and extending their research contributions. Through this integrated platform, tasks such as archival recognition, classification, transcription, description, comparison, and analysis were completed in a unified environment, significantly improving the efficiency of archival development while ensuring data accuracy.

### Data and ethical statement

All archival materials utilized in this project were obtained with permission from four sources, including:

(1) the close friend of Chen Taosheng: The materials are mainly published magazines and newspapers related to the scientist, essays published or mentioned by Chen Taosheng himself, life photos of Chen Taosheng, and his unpublished poetry collection 'Guanwei collection' (in Chinese: 观微集).

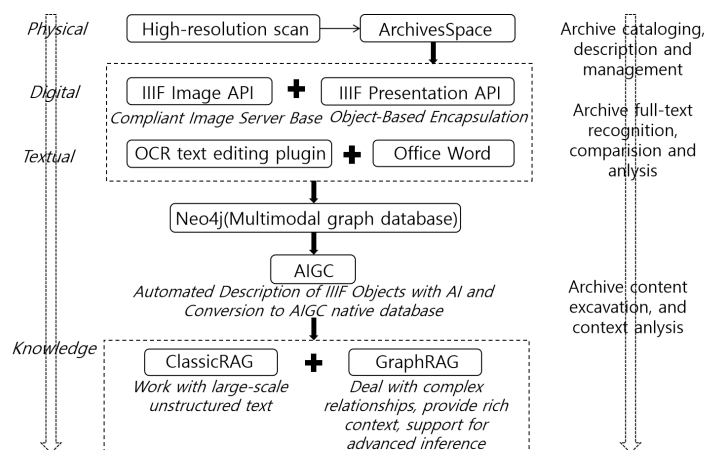
- (2) Shanghai municipal archives: the materials contain his views on scientific research, his career and work proposal and its reply, his work introduction letters, and materials for selecting advanced individuals in the science field.
- (3) Archives of Shanghai University: the materials include the forms filled in by him during his work at the predecessor of Shanghai University, his thoughts and opinions, the relevant school records, and his relevant letter of appointment.
- (4) Open data from Shanghai Library: records of other celebrities and scientists related to Chen Taosheng; records of the institutions and places that relate to him.

The types of archival data include original archival files, old photographs, manuscripts, newspaper clippings, journal articles clippings. These materials were acquired with explicit agreements permitting their digitization and use only for academic purposes.

The subsequent data processing and experiments were conducted within the university laboratory under the oversight and management of the university's office of scientific research management, safeguarding the integrity and authenticity of the archival materials throughout the project.

## Research process and methods

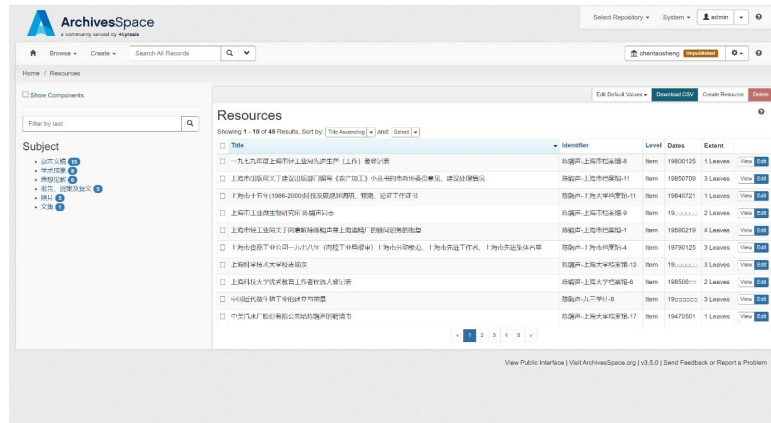
In the scientist archival data development workflow project, the main research path is ‘physical form of archives—digital archival images—textual archival data—archival knowledge’ (see Fig. 1).



**Figure 1.** Research process and methods

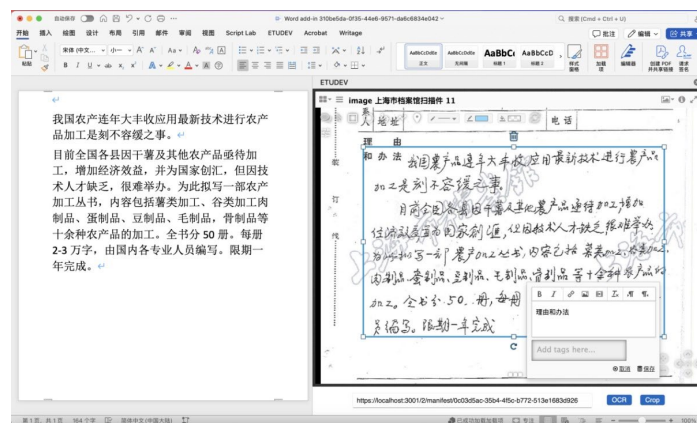
The workflow began with digitizing physical archives and records, utilizing high-resolution scanners to convert archival documents, photographs, and other archival materials into digital formats. These digitalized files were then organized and catalogued using the open-sourced archival management system, ArchivesSpace, where structured metadata and standard archival descriptions to enhance data retrieval and operability (see Figure 2). During this process, raw images were integrated through IIIF objects and recorded as digital object instances corresponding to physical items.

‘ArchivesSpace is an open-source, browser-based archival information management software application’ (Link: <https://archivesspace.org>). 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 archive metadata. DACS is a foundational statement of content standards and principles for archival organization and description within the United States archival community. It includes descriptions of archival material and authoritative records representing the people and organizations that created them (SAA, 2022). DACS contains 25 archival elements, many encoded with multiple EAD tags.



**Figure 2.** ArchivesSpace archival management system

Then IIIF was integrated with the OCR text editing plugin in the Office Word environment (see Figure 3) to realize a commonly used archival work environment and the seamless linking of raw images with their corresponding OCR texts, ensuring comprehensive documentation and accessibility.



**Figure 3.** IIIF integrated word environment with OCR plugin

Archival images (unstructured data) are transformed into structured and semi-structured data through the IIIF-based OCR tool. Structured data refers to tabular data that can be directly stored in a structured database, while semi-structured data refers to data in a loose key value format such as JSONL format. Traditional manual description tasks are thus converted into data mapping tasks that can be automatically completed by LLMs. Once a high-quality electronic archival database is constructed, it can serve as a private AIGC database for Classic RAG or GraphRAG needed by LLMs, targeting at content mining, content analysis, and other tasks that can reveal the value of archives.

## Findings

### Large archival digital images solutions

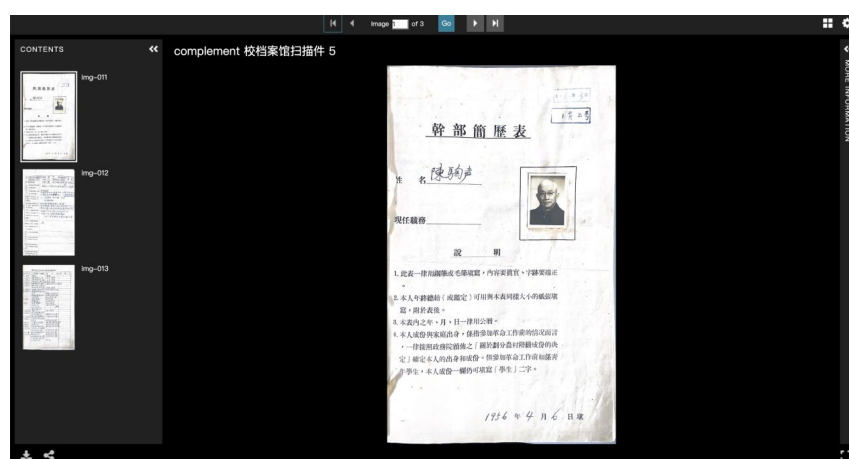
Using scanned archival images as archival data posed several technical challenges, including handling large quantities of images, managing complex image-text interactions, ensuring description accuracy, and maintaining archival security when employing LLMs. To address these challenges, the project adopted the IIIF framework as a standard, breaking down image development into three independent yet progressive workflows.



First, the project implemented an IIIF Image API as a compliant image server base (see Figure 4 and Figure 5). The Image API defined pixel-level operations for images, forming the foundation for dynamically presenting archival images, particularly high-definition images. It also enabled real-time acquisition and manipulation of finely indexed images. The project incorporated a series of engineering technologies, including container orchestration and serverless computing, to ensure system performance meets business needs when handling large quantities of high-definition images.

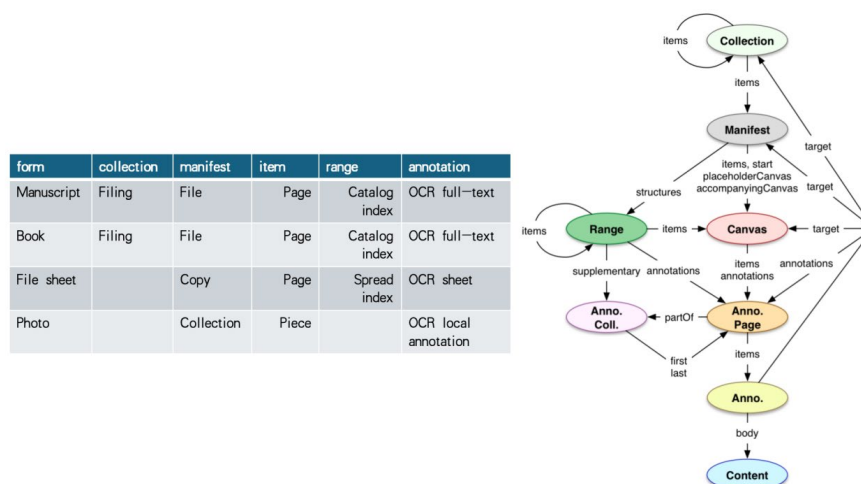


**Figure 4.** IIIF Image API



**Figure 5.** An example of IIIF Image API

Second, the project utilized the IIIF presentation API for object-based encapsulation and OCR Text integration. The project leveraged the flexible definition methods of the presentation API to combine individual images into logically related image objects and described metadata in multiple languages at different levels (see Figure 6). Given the diverse layout of images, the project incorporated the ability to select appropriate areas freely and to apply different OCR models for handwritten and printed texts. The project developed an OCR text editing plugin that combines IIIF and office word, enhancing annotation flexibility and solving the problem of standardized storage of annotation results, enabling quick associations between texts and images during retrieval.



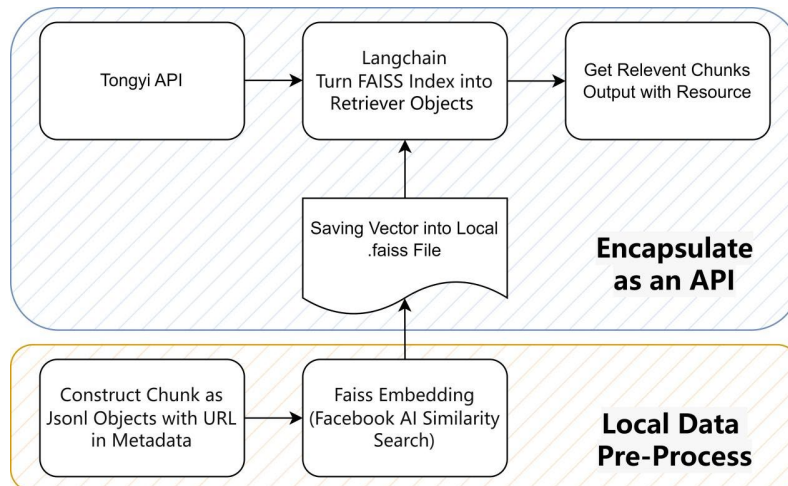
**Figure 6.** Flexible definition methods of the presentation API

This solution preserves the originality of archival materials, respects the existing form of archival materials, gives them the possibility of being used as ‘data’ and ‘evidence’, and ensures that archival materials can be preserved for a long time.

### Archival data AIGC knowledge base and LLM solutions

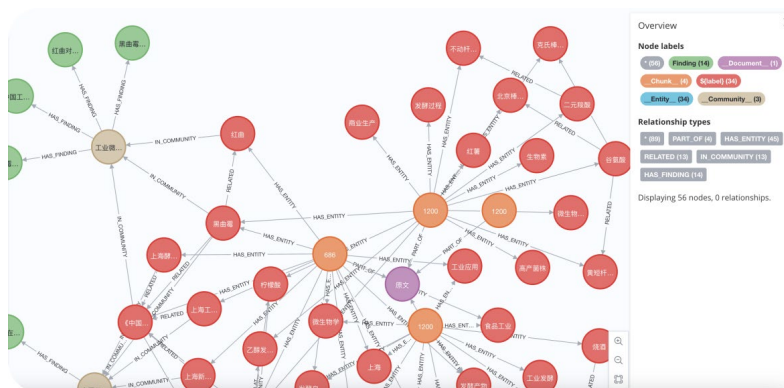
By leveraging LLMs' language reconstruction and mining capabilities, the project significantly enhanced the explanatory power of archival data, establishing a new paradigm for archival utilization within the new technical environment. ‘RAG (retrieval enhanced generation) technology is evolving, including the Classic RAG and Graph RAG, representing a significant advance in information retrieval and knowledge representation’ (Omrani,2024). This project explored the two approaches and built a chat function allowing users to ‘chat’ with the scientist (see Figure 9).

The Classic RAG system (see Figure 7) first converted documents into Markdown format through OCR. Using the hierarchical structure of this format, the system employed an overlapping sliding window method to segment text while maintaining semantic integrity. Each segmented text block was then converted into a JSONL object containing metadata such as text content and URLs. These objects are then converted into embedding vectors stored in a vector database. During retrieval, the system augmented the context window by locating similar text in the database through nearest neighbour search.



**Figure 7.** Classic RAG

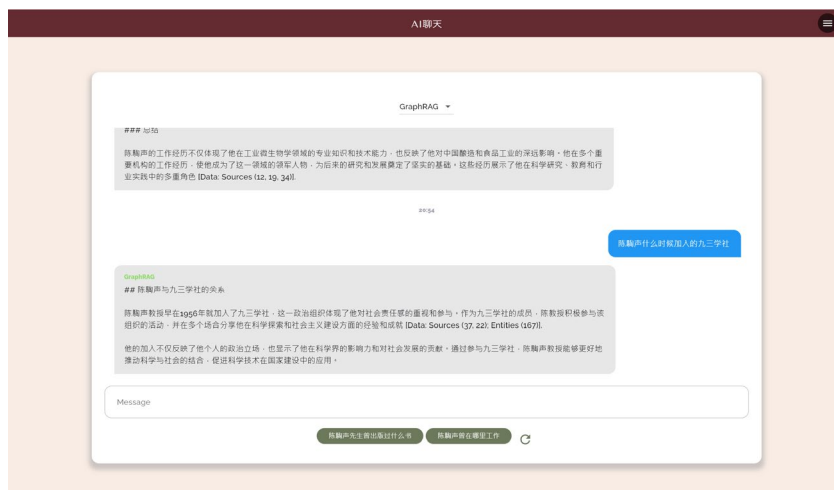
GraphRAG followed this process but introduced an additional layer of inference. The LLM first performed inference operations on each sentence in the dataset, extracting entities and relationships. Named entity recognition (NER) was used to identify entities within the text and determine the relationships between entities, as well as the strength of those relationships. Then, the extracted relationships were formed into a knowledge graph, which consisted of nodes (entities) and edges (relationships) (see Figure 8).



**Figure 8.** Knowledge graph based on GraphRAG

While the Classic RAG primarily relied on vector databases to process unstructured text, the Graph RAG established a super-large-scale vocabulary by introducing knowledge graphs that equated entities and relationships with words. This enabled a more accurate understanding of content and more accurate search results.





**Figure 9.** Chat interface of the project

## Towards a human-machine archival workflow

The Chen Taosheng archival data development project successfully implemented a systematic workflow by utilizing the open-source archival management system ArchivesSpace, along with an IIIF integrated word environment with OCR plugin, for comprehensive archival data cataloguing, describing, recognition, comparison, and management. Additionally, by extending the LLM with Classic RAG and GraphRAG, the project successfully enabled AIGC-generated texts to refer to original images. This technique not only effectively addressed potential hallucination issues in LLMs but also significantly enhanced the credibility of LLM responses by presenting the original images, thereby maximizing the utility and reliability of archival resources.

The integrated IIIF and OCR tool-based work environment ensured the automated generation of full-text data from archival materials and enhanced the efficiency and accuracy of the process through human-machine collaboration. The automation of repetitive tasks reduced the workload on archivists, allowing them to focus on more complex aspects of archival work. This hybrid approach not only improved efficiency but also reduced the risk of human error. The adoption of LLM supported the archival workflow in a good way by enabling the explanation ability of archival materials and enriching users' interaction with archives.

## Lessons and future research

The integration of AIGC texts with original images through IIIF objects effectively addressed the hallucination issue in LLMs, significantly enhancing the credibility of LLM responses. However, several challenges were encountered during the project experiment when interacting with the LLM. Firstly, when answering questions requiring large chunks of content, the original large chunks of content were often truncated (see Figure 10 as an example). This truncation hindered the LLM's ability to identify relationships between keywords, leading to responses that, although traceable, were often speculative and lacked credibility. To solve this issue, the size of the text chunks and the length of overlapping segments were typically increased to make the semantic relationships more complete. However, this approach proved insufficient for large-scale overviews unless the corpus was rich enough to provide summative content. This limitation underscored the need for GraphRAG.

为什么陈陶声说自己是编书而不是著书

ClassicRAG

回答: 根据提供的上下文, 没有明确提到陈陶声说过自己是编书而不是著书。上下文主要描述了陈陶声编写中文教科书的过程和他的努力, 但并没有涉及他具体如何定义自己的工作为“编书”还是“著书”。因此, 对于这个问题, 我不知道答案, 不会进行推测或编造。

图片: 在工业微生物的“海洋”中遨游

链接: <https://viewers-dev.etu.wiki/p/d6772ce5-41ce-423e-9c36-014c55e5915b/m3.html>

图片: 跟细菌打交道的人

链接: <https://viewers-dev.etu.wiki/p/06a4fb99-29df-455b-8208-80344d291ca2/m3.html>

**Figure 10.** Large chunks of content may be truncated in the answering the question.

Secondly, ambiguity in naming entities posed another significant issue. For example, ‘Mr. Chen’ and ‘Professor Chen’ in the literature both referred to Chen Taosheng, however, these designations were sometimes incorrectly identified as different entities when constructing knowledge graphs. This problem was further exacerbated by OCR errors (see Figure 11 as an example). To solve this problem, cluster-based entity disambiguation is required, i.e., identifying the real-world entity to which ambiguous referents were directed by leveraging the principle that the same referent typically appeared in approximate contexts. Clustering algorithms were then applied to disambiguate these entities.

陈陶声曾在哪里工作

GraphRAG

## 陈陶声的工作经历

陈陶声是中国著名的工业微生物学家。他的职业生涯涵盖了多个重要的工作单位。尤其是在酿造和微生物研究领域。他的工作经历不仅展示了他在科学研究方面的贡献。也反映了他对中国工业微生物学发展的影响。

### 主要工作单位

1. 山东得益糖厂

陈陶声在山东得益糖厂工作多年。这里是他职业发展的重要起点。在此期间, 他应用自己的知识改善了酒精生产工艺, 取得了显著成效 [Data: Entities (79); Relationships (60)]。

2. 山东得益酿造厂

在山东得益酿造厂, 陈陶声专注于酒精生产, 并在此过程中获得了工程师的职务。他的研究和实践为工厂的酒精生产带来了重要的技术进步 [Data: Entities (161); Relationships (93)]。

**Figure 11.** Ambiguity in naming entities.

Future directions for the project include considering the integration of multi-modal large language models (MLLMs) to deeply mine semantic information from raw archival images beyond text, creating innovative modes for multi-modal archival management and development. The project team also plan to continue to optimize the human-machine collaborative archival development models and technological applications. These efforts aim to explore more efficient data processing and archival management solutions, further enhancing the intelligence and processing capabilities of archival work. Such advancements were expected to not only increase the value and utilization efficiency of archival data but also provide new theoretical support and practical guidance for the development of the archival industry.

## Conclusion

This project demonstrated the potential of modern technology in the AGI era in archival management through a case study of the scientist Chen Taosheng's archival materials. By leveraging IIIF, OCR and LLMs technologies, and implementing human-machine collaborative models, an innovative workflow for native image archival data has been developed in this project. These advancements significantly enhanced the efficiency and accuracy of archival development and utilization, providing theoretical and practical guidance for the field.

As the AGI era unfolded, the workflow for managing archives and records underwent a transformative evolution. The collaboration between humans and machines proved crucial in

navigating the complexities of the AGI era. Human expertise in archival science ensures the contextual accuracy and ethical handling of records, while machine capabilities significantly expedite the processing and organization of vast amounts of data. This symbiotic relationship enhances the overall workflow, leveraging the strengths of both human intuition and machine precision. 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. By fostering a collaborative environment, archivists are able to release the full potential of AGI technologies, driving innovation in archival practices while maintaining the integrity and authenticity of archival records. This balanced approach not only preserves the historical and cultural significance of archives but also paves the way for more advanced and reliable archival management in the future.

## About the authors

**Dr. Yaming Fu** is a Lecturer at School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China. Her research interests include digital humanities, knowledge organization and archives informatization. E-mail: ymingfu@126.com

**Jie Song** is the Founder and CEO of Intellijourney Co., Ltd., Shanghai, China. His research interests include IIF and artificial intelligence. E-mail: joe.song@intellijourney.com

**Xinran Zhang** is an undergraduate at School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China. Her research interests include basic theory of archival science, digital humanities. E-mail: 1654770712@qq.com

**Jingyun Bi** is an undergraduate at Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. His research interests include application of generative AI in social data, and radio frequency surrogate models for automatic design of active and passive circuits. E-mail: [2537503782@qq.com](mailto:2537503782@qq.com)

## 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.
- 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://iif.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\\_1/02\\_chapter\\_01.html](https://saa-ts-dacs.github.io/dacs/06_part_1/02_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.

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).