SciDataMAS: LLM-Driven MAS for Scientific Data Management (Student Abstract)

Alexander Sachuk, Vyacheslav Chukanov, Ekaterina Pchitskaya

Peter the Great St. Petersburg Polytechnic University
sachuk.as.bsns@gmail.com, kauter1989@gmail.com, katrin.pchitskaya@gmail.com

Abstract

The management and annotation of complex, multi-modal scientific data remains a major obstacle for AI-driven research due to poor reusability and scalability of current solutions. We propose SciDataMAS, a novel LLM-powered multi-agent system (MAS), which automate scientific data management through a structured data lake with provenance-based organization and an adaptive metadata taxonomy. The system uses specialized workflows for automated dataset creation, data insertion and retrieval. Experiments show the system's proficiency, with modern LLMs like GPT-5 successfully generating rich metadata schemas and filling them with high accuracy. This work provides a foundational step towards fully automated, reusable, and scalable scientific data organization which may lead to generation and accumulation by scientific community well annotated AI-ready datasets.

Code —
https://github.com/Biomed-imaging-lab/SciData-MAS

Introduction

The accumulation of interpretable, well-annotated data of specific modalities from various fields is a necessary part of training AI models and new discoveries in different fields (Bajcsy et al. 2025). While existing solutions usually focus on structuring some specific data modalities (Sarkans et al. 2021), which are good for a specific experiment in a small volume, they lack of general, scalable principles for accumulation big amounts of structured data across diverse modalities. This limits the creation of large-scale datasets necessary for advanced AI methods, that can lead to breakthrough research results in different areas (Lu et al. 2024). In this work, we present an LLM-based approach featuring a novel scientific data management paradigm. Our solution enables the generation of adaptable metadata schemas and a multi-agent system for semi-automated data management, supporting reusable and scalable data organization.

Methodology

Scientific data management paradigm

To structure scientific data, we propose to use data lake where datasets are partitioned by source and modality. Data is categorized by provenance into three types: "**external data**" (from third sources), "**real data**" (from inner laboratories' experiments), and "**software created**" data (synthetic/processed). This structure improves metadata accuracy, consistency and traceability (Fig. 1, top). A core challenge in scientific data is schema adaptability. We address this with a time-based metadata taxonomy: **pre-capturing metadata**, describes the object's origin before digitization (e.g., provenance, setup). These properties cannot be retroactively added and require careful design; **pro-capturing metadata**, describes properties derived from the digital object (e.g., features, analysis). These can be added to the schema and populated for existing data at any time (Fig. 1, down). Finaly, we propose a universal, modality-agnostic metadata designing schema based on the data generation pipeline: "**object metadata**" - describing research object;

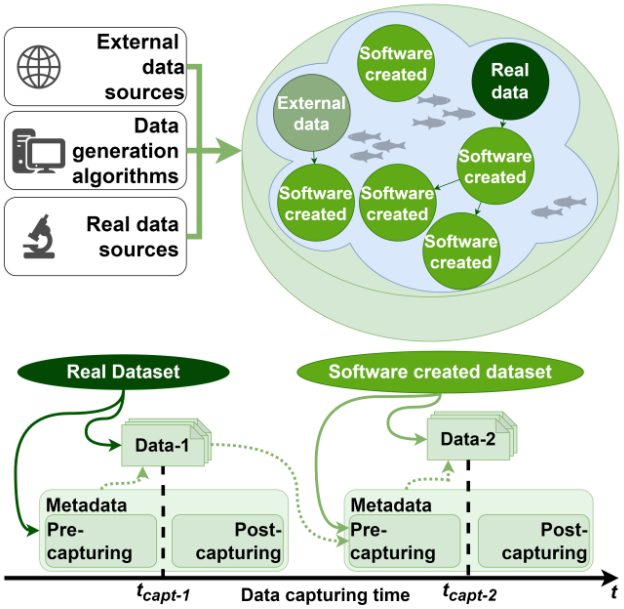

Figure 1: Proposed scientific (meta)data lake organization and SciDataMAS – multiagent system for automatic working with data lake. Top: data lake inner structure organization; bottom: the inner structure in datasets.

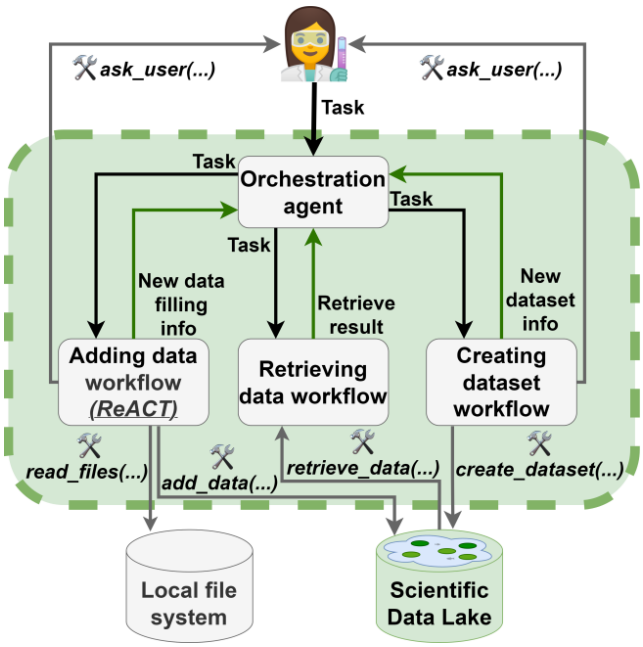

Figure 2: MAS structure with tools and environments.

"**experimental metadata**"- describing all experiment design aspects; "**hardware metadata**" - describing the digitization equipment parameters.

MAS for (semi-)automatic data management

While LLM and multi-agent systems (MAS) are advancing in general data management (Zhou et al. 2025), they often overlook the specific needs of scientific data. To operationalize the proposed data storage framework, we developed a SciDataMAS: the ReAct-based (Yao et al. 2022) multi-agent system that orchestrates various specialized goal-oriented workflows for data management tasks: *dataset creation with metadata generation, data insertion* and *data retrieval* (Fig. 2).

Experiments

A prototype data lake with local file storage and CSV-based metadata was implemented. The evaluation assessed MAS performance on two core operations: "**dataset creation**" - metadata schema generation for describing hippocampal neuron activity data, "**data upload**" - ability to add fluorescence microscopy data to a dataset with extracting metadata from specified files (6 direct fields) and infer others via basic operations (6 derived fields). All experiments were conducted over 16 independent runs using "Mistral" (Mistral-medium-2508), GPT-4o and GPT-5 (mini) models. Execution time, token count, the total and non-string (machine-processable) number of generated fields was recorded. Accuracy was measured by the count of correctly vs. incorrectly filled metadata.

Model	Fields created	Non-string	Tokens (10^3)	Time (sec)
Gpt-4o	18.8 ± 2.9	10.3 ± 2.4	4.3 ± 0.3	27.7 ± 6
Gpt-5	70.5 ± 18.6	39.8 ± 11	17 ± 1.1	151 ± 32
Mistral	44.8 ± 14.9	28.43 ± 10	10.8 ± 4	45 ± 16

Table 1: Metrics of datasets metadata generation

Model	Correct fields	Incorrect fields	Tokens (10^3)	Time (sec)
Gpt-4o	10.9 ± 1.9	1.1 ± 1.9	11.7 ± 2	34 ± 7.8
Gpt-5	10.3 ± 2	2.1 ± 2	17 ± 3	139 ± 111
Mistral	8.8 ± 1.8	3.5 ± 1.8	23 ± 7.8	$\mathbf{30 \pm 3.8}$

Table 2: Automatic metadata filling accuracy evaluation

Results

Results in Table 1 demonstrate that proposed MAS with modern LLMs like GPT-5 and Mistral can design rich metadata schemas. Additionally, results in Table 2 shows the system's ability to fill metadata with high accuracy

Conclusion

Our results demonstrate that the proposed solution successfully enables structured storage and automated description of experimental data. This work highlights the potential of LLM-assisted systems in scientific data management, though further validation is needed - particularly in scaling the approach and evaluating its performance across diverse domains. We believe this direction merits significant attention, as it addresses critical challenges in data reuse and interdisciplinary collaboration (Bajcsy et al. 2025).

Acknowledgments

The work was supported by the Foundation for Scientific and Technological Development of SPBPU.

References

Bajcsy, P.; Bhattiprolu, S.; Börner, K.; et al. 2025. Enabling Global Image Data Sharing in the Life Sciences. *Nature Methods*, 22(4): 672–676.

Lu, Y.; Wang, H.; Zhang, L.; et al. 2024. Unleashing the Power of AI in Science: Key Considerations for Materials Data Preparation. *Scientific Data*, 11(1): 1039.

Sarkans, U.; Chiu, W.; Collinson, L.; et al. 2021. REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nature Methods*, 18(12): 1418–1422.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629.

Zhou, X.; He, J.; Zhou, W.; et al. 2025. A Survey of LLM × DATA. arXiv:2505.18458.