

Towards Reducing Researchers' Burden in Open Science: An AI-Driven Framework for Automated Metadata Entry

Masaharu Hayashi¹ , Makoto Asaoka¹ , Masashi Kawai¹ , and Mikiko Tanifuji¹ 

¹ National Institute of Informatics

Abstract

To facilitate low-burden open science, we propose an AI-driven framework for institutional repositories. By integrating external APIs with large language models (LLMs) and vision-language models (VLMs), the system automatically generates metadata from DOIs or PDFs, including image-based summaries. This approach minimizes manual input while ensuring quality through provenance tracking, thereby streamlining the path to immediate open access.

1. Background

The global shift toward open science requires researchers to make publicly funded results openly accessible [1]. In Japan, the publication of academic papers and evidence data resulting from public funding has become mandatory. As reliance on Article Processing Charges (APCs) for open access becomes financially unsustainable, open access with institutional repositories is increasingly vital. To implement open science in practice, self-archiving must be streamlined through technical solutions [2]. This study presents a proof of concept (PoC) for an AI-driven framework that simplifies repository registration while maintaining high metadata quality.

2. Methods

The framework automates metadata entry for academic papers and evidence data using a DOI or a PDF (Figure 1). For academic papers, it prioritizes metadata from external APIs, using an LLM as a fallback for PDF extraction. For evidence data, basic metadata are inherited from the academic paper metadata, while summaries are generated from images and captions. A VLM describes visual elements, which are integrated into a coherent summary by an LLM. To ensure integrity, the LLM evaluates generated fields for consistency [3] prior to researcher review. The framework maintains provenance by recording whether each field was AI-generated, corrected, or verified. Additionally, the framework uses the Open Policy Finder to provide information about the version of the paper that researchers can deposit in an institutional repository, and exports metadata using the JPCOAR schema.

3. Results

This framework automates metadata generation from DOIs or PDFs, transforming deposits into a simple verification task. Testing on 100 papers yielded high F1 scores:

Poster

Available Aug 01, 2026

Correspondence to
Masaharu Hayashi
mhaya@nii.ac.jp



Copyright © 2026 Hayashi *et al.*
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

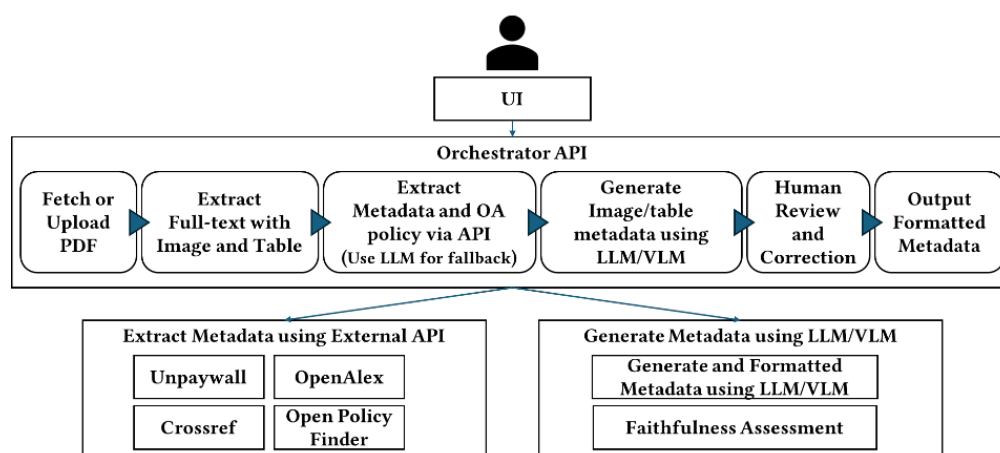


Figure 1. PoC Architecture for Repository Deposit using AI.

0.989 (DOI), 0.860 (title), and 0.803 (creator). The system leverages a complementary approach where APIs provide canonical data and the LLM extracts verbatim strings (e.g., ROR identifiers and keywords) from PDFs. This allows researchers to focus on low-confidence, AI-flagged entries. Additionally, by extracting data-level metadata from images and tables, the framework significantly minimizes the manual effort required for separate dataset registration.

4. Conclusion

This study demonstrates a low-burden repository workflow by integrating external APIs with LLM/VLM-based generation. While automated quality control requires further refinement, this PoC confirms AI's potential to mitigate administrative overhead. Future work will focus on optimizing human–AI collaboration to achieve a fully streamlined and sustainable scholarly communication environment.

References

- [1] U.S. Office of Science and Technology Policy, “Ensuring Free, Immediate, and Equitable Access to Federally Funded Research,” 2022.
- [2] L. Rothfritz, L. Matthias, H. Pampel, and M. Wrzesinski, “Current challenges and future directions for institutional repositories: A systematic literature review,” *J. Assoc. Inf. Sci. Technol.*, vol. 77, 2026, doi: 10.1002/asi.70016.
- [3] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. in Proc. EMNLP 2023. doi: 10.18653/v1/2023.emnlp-main.153.