From Analog Records to Computational Research Data: Building the AI-Ready Lab Notebook

Joel Pepper
Department of Computer Science
Drexel University
Philadelphia, PA USA
0000-0002-1601-8729

Zach Siapno
Department of Computer Science
Drexel University
Philadelphia, PA USA
zs394@drexel.edu

Jacob Furst
Department of Chemistry
University of Central Florida
Orlando, FL USA
jacob.furst@ucf.edu

Fernando Uribe-Romo
Department of Chemistry
University of Central Florida
Orlando, FL USA
0000-0003-0212-0295

David Breen
Department of Computer Science
Drexel University
Philadelphia, PA USA
0000-0002-1376-5008

Jane Greenberg
Department of Information Science
Drexel University
Philadelphia, PA USA
0000-0001-7819-5360

***Abstract*—Scientific laboratory notebooks, particularly those in analog, handwritten form, represent a significant yet underutilized data source for computational studies. This paper reports on our research to further develop a pipeline for transforming analog lab notebooks to AI-Ready digital archives. The research is conducted within the framework for Computational Archival Science (CAS), extending CAS principles, drawing from archival practice and computational thinking. We provide background context on laboratory notebook history and current day use, explore CAS as a framework for study, followed by our research goals and methods. Automated extraction results for table records found in the notebooks have an error rate under 5% on a per cell basis. The framework, methods, and our findings seek to advance pipelines for making analog records, both historical and current, accessible and curated for computational research. The findings presented underscore both the accelerating pace of extraction technologies and the importance of more structured, consistent analog documentation practices to support computational transformation and AI-readiness. The conclusion summarizes results and identifies next steps.**

***Index Terms*—Computational archival science, AI-ready data, lab notebooks, digital collections**

I. Introduction

Artificial intelligence (AI) has accelerated the pace of scientific discovery in unprecedented ways, fueled by access to massive and ever-expanding reservoirs of digital data. However, within the context of historical and archival documents, vast quantities of data remain inaccessible for use in AI and machine learning due to never having been converted to a machine-readable digital representation. Analog, paper laboratory notebooks are a class of documents particularly at risk of being lost to time in private archives, with many labs keeping their notebooks indefinitely, but with few actually extracting the contents to an "AI-ready" format [1]. This is of significant concern, given that these resources contain a wealth of potentially important information on experimental design, outcomes and student learning.

This predicament, along with the reality that science builds on previous research, underscores the urgency to advance methods and systems that allow these resources to be integrated into AI pipelines. Innovations in automated lab notebook analysis will provide far greater utilization of previously learned, but not yet deployed, knowledge in future scientific studies. Here, we report on our on-going AI-ready analog lab notebook initiative, involving researchers at the Metadata Research Center, Drexel University, and the Reticular Synthesis and Materials Design Lab (RSMDL), University of Central Florida, as part of the NSF-HDR Institute for Data Driven Dynamical Design. The lab notebooks under study come from the RSMDL's research on designing and synthesizing metal organic framework (MOF) and covalent organic framework (COF) crystals. Example pages from several of these notebooks are shown in Figure 1.

This paper focuses on work in progress that transforms handwritten laboratory notebooks into AI-ready, machine-readable datasets by applying OCR-based text extraction, error pattern analysis, and data structuring methods. We also describe how the recently released document content extraction tool olmOCR 2 [2] represents a significant increase in performance and accuracy over programs previously used in our pipeline. The sections that follow provide a history of laboratory notebooks, describe their current use, and offer Computational Archival Science (CAS) as a framework for their study, followed by our research goals, methods, and results. A brief discussion explores our objectives and implications, then a conclusion summarizes results and identifies next steps.

National Science Foundation, Institute for Data-Driven Dynamical Design, OAC-2118201.

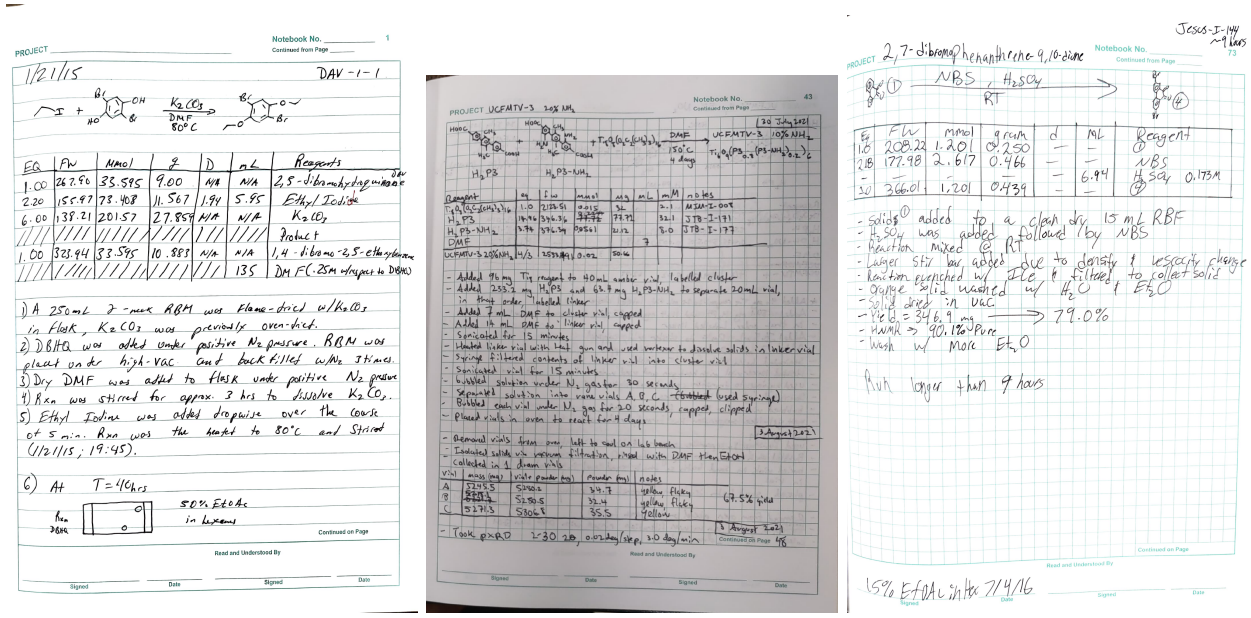


Fig. 1. Pages from organic synthesis of intermediates for MOF's in the notebooks.

II. Analog Lab Notebooks and CAS Opportunities

A. Analog Lab Notebooks: Foundations and Current Trends

The origins of scientific research notebooks can be traced back to the earliest traditions of scientific inquiry, when philosophers such as Aristotle engaged in systematic observation and documentation of the natural world [3]. Early note-taking practices are interconnected with the history of learning, when Renaissance humanists emphasized collecting, organizing, and preserving knowledge on paper to support long-term study and discovery [4]. With the Scientific Revolution, record-keeping became more formalized as scientists like Francis Bacon articulated empirical and inductive methods that relied on systematic documentation as the foundation for reproducible inquiry [5]. Robert Boyle's detailed laboratory notes, later rediscovered and analyzed by historians, reveal the emergence of a documentary infrastructure for science—what Rheinberger [6] describes as literary technologies of experiment.

By the eighteenth century, Antoine Lavoisier's meticulous notebooks, analyzed in Holmes's *Lavoisier and the Chemistry of Life* (1985), exemplified how structured records underpinned the standardization and communication of experimental work [7]. The *Reworking the Bench* collection [8] further demonstrates that such notebooks not only captured results but also reflected evolving conventions of scientific writing and record-keeping, helping to establish reproducible, traceable, and ultimately standardized practices. We might even look at some of these developments as precursor to the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles that guide today's scientific record-keeping [9].

Today, the laboratory notebook remains a fundamental component of scientific research, providing a record of scientific activity and discovery [8], [10]. While many laboratories have adopted *electronic lab notebooks* (ELN's), these are not always suitable for wet labs—particularly in chemistry, where researchers routinely handle hazardous materials while wearing gloves, masks, and other protective gear. The continued reliance on analog lab notebooks underscores both their practicality and their archival importance. Consequently, chemists continue to rely on traditional, paper-based lab notebooks made of chemical-resistant paper to record synthesis experiments and other research details.

However, the persistence of paper notebooks presents challenges for research environments increasingly shaped by digital workflows and AI. Although electronic systems can enhance data management and integration, they often require financial investment, technical infrastructure, and training resources that small academic laboratories may lack. Despite these constraints, many scientists in wet-lab disciplines express growing interest in leveraging AI for data analysis, discovery, and workflow optimization. Realizing this potential requires new methods for digitizing and structuring handwritten records, so that they can be computationally processed and reused. Developing AI-ready pipelines for analog laboratory notebooks is therefore an essential step towards bridging the gap between traditional experimental practice and data-driven research. Such efforts would not only enable the reuse of legacy data, but would also ensure that the scientific record remains accessible, verifiable, and interoperable within the evolving landscape of computational research.

B. Computational Archival Science

Computational Archival Science (CAS) provides a theoretical and methodological foundation for the efforts to make scientific lab notebooks AI-ready. Marciano and colleagues [11] describe CAS as an interdisciplinary field that bridges archival

science, computer science, and domain research to develop scalable, automated, and reproducible processes for managing complex cultural and scientific data. The field is grounded in computational thinking, integrating computational methods, such as machine learning, natural language processing, and data mining, into the curation, analysis, and preservation of archival materials. These resources reflect the longstanding history of documentation, tracking the life cycle of a research experiment. The archival context is emphasized with the tie-in of archival practice and principles, such as *original order* and *provenance* [12], [13].

CAS brings forward the analog-to-digital transformation of physical, handwritten research artifact into a computationally tractable and semantically enriched data object. Analog lab notebooks may be considered archival records, as they document the research procedures, specifically observations, experimental steps, measurements, and outcomes, while capturing the researcher's overall activity. The application of CAS methodologies allow for the preservation of this embedded structure, while linking these artifacts to the broader computational ecosystem, thus supporting a new level of study. Perhaps the biggest challenge with analyzing paper lab notebooks is the structuring procedures, or lack thereof, of these analog resources. Some laboratory managers provide very clear and strict guidelines for notebook composition, while others are more free form, with only a general rubric of what is contained inside. Advancing AI-ready lab notebook research in the context of CAS could have implications for how laboratory managers guide their teams in the future, particularly as they seek to integrate AI and data-driven technologies into long-term research practices and record-keeping systems.

III. Goals and Objectives

The overall goal of our work is to advance methods for converting scanned handwritten lab notebooks into structured, computer-accessible data for scientific analysis and AI use. Specific objectives include:

- Extract data from the notebooks to support computational analysis of the information archived in these analog artifacts,
- Identify common error patterns in the extraction outputs and develop a postprocessing pipeline to correct them, and
- Contribute to the long-term overall accessibility and reuse of analog historical lab data.

A. Data Repository

The RSMDL possesses an archive of lab notebooks that was started in 2013. The collection consists of about 80 lab notebooks, which document the work performed by graduate and undergraduate students and postdocs. Eleven of these notebooks from four distinct authors have been scanned thus far. Lab notebook pages were photographed using a camera held above the page, and are aligned and cropped in somewhat variable ways. All notebooks cover the synthesis of MOF crystals, following a variety of different procedures.

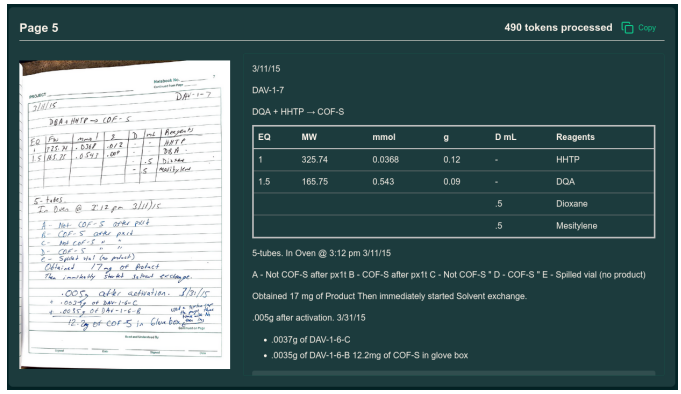


Fig. 2. olmOCR 2 output.

B. Image Segmentation

Information found on notebook pages includes text, tables, diagrams, equations, chemical reactions, and more depending on the procedure being performed. Keeping with our goal of creating machine-learning-compatible representations of notebook contents, we have focused primarily on the tables and text blocks within the notebooks.

In order to perform automated page segmentation, we have been employing the Detectron2 object detection platform [14]. This platform, first released in 2019, does provide some concrete benefits over newer, larger models despite its age. Specifically, it is relatively easy to train on a custom dataset, it is relatively lightweight and easy to run locally on a standard desktop computer, and when only bounding boxes are needed it is sufficiently performant for typical use cases. Most of the newer object detection models that have supplanted Detectron2 focus on generating segmentation masks on large, landscape-style images, and are either only provided as cloud-based commercial products or are too large to run outside of AI data centers. For these reasons, we continued to utilize a custom-trained Detectron2 Faster R-CNN model for page segmentation.

Recently, a significant leap forward in document content extraction has been made with the release of the olmOCR 2 toolkit. Developed by the Allen Institute for AI, the first version of this toolkit was released in February 2025 [15], with an updated second version being released on October 22, 2025 [2]. olmOCR is a vision language model that integrates document segmentation quite effectively into the content extraction process, and fully eliminates the need to run separate object detection and optical character recognition tools. It is open source and designed to be run on the end user's hardware, with the stipulation that the minimum hardware requirement of a 16GB graphics card. This is significantly less accessible than the hardware requirements for something like Detectron2, but with most research institutions having access to at least one such card, it is still a lean enough model that many laboratories should be able to deploy it without the need for expensive cloud computing resources.

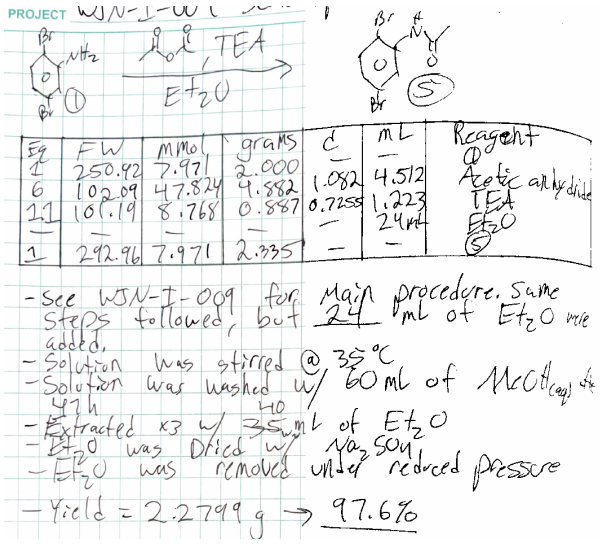


Fig. 3. Grid removal via thresholding.

olmOCR 2 provides a significant increase in accuracy compared to our previous computational tools, but given how recently olmOCR 2 was released, we have not yet fully integrated it into our workflow. The performance of both olmOCR 2 and our existing workflow will be discussed in Section III-F.

C. Content Extraction

In tandem with Detectron2, we used a commercial, cloud based software system called Handwriting OCR to extract the contents of segmented text blocks and tables [16]. Prior to the release of olmOCR, this software was the only one we tested that could extract the contents of handwritten tables into comma-separated-value format, something required for this project. While Handwriting OCR can extract content from full pages, we found it most accurate on pre-cropped images of the segmented page features. This is the operating mode by which the data described in Section III-F was generated.

olmOCR 2 is much more capable of processing full page inputs and accurately differentiating where one element ends and the next begins. Its output takes the form of a single HTML document per input page, which can then be further parsed to separate out the table elements and text blocks. An example of its output is shown in Figure 2.

D. Image Preprocessing

When working with Handwriting OCR, it was found that the lining and grids on the notebook pages were interfering with row and column generation for the extracted tables. Via basic thresholding on the hue and color intensity of pixels, the lines were removed from all the table images as a preprocessing step before running the OCR. The results of this are shown in Figure 3. olmOCR 2, however, does not seem impacted by any linings or grids, so this process is likely unnecessary moving forward.

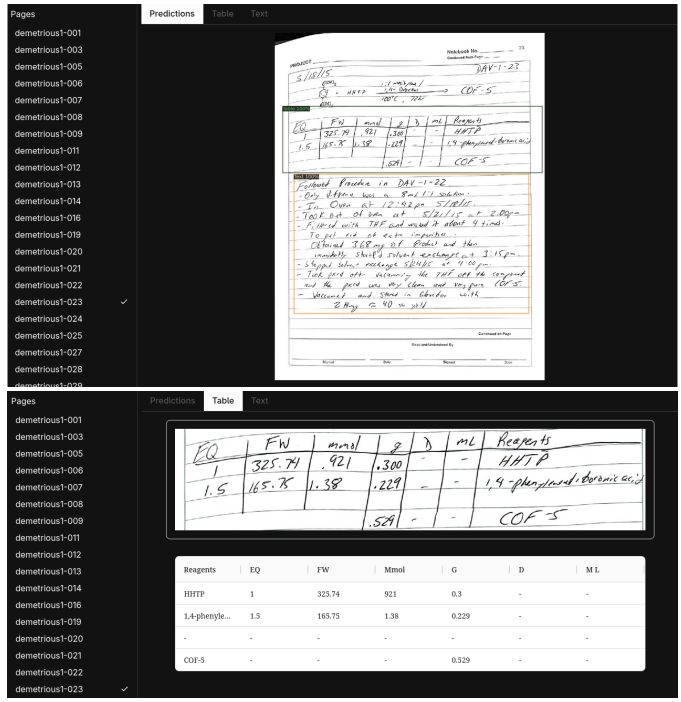


Fig. 4. ELN style web interface.

1) Data Post Processing: Regardless of which method is used for content extraction, some number of errors will be present. However, with well structured, consistent lab notebook elements, the bulk of these errors can be caught and flagged for manual review or even corrected automatically. Considering the tables from the RSMDL notebooks, the laboratory sets strict standards for how tables should be written. For each active reagent in a reaction, students are to write the formula weight (FW), moles (mmol), grams and equivalent value (Eq). Using the redundancies in this information, two main checks can be performed:

1) Within rows: $\mathrm{FW} * \mathrm{mmol} = \mathrm{grams}$, and
2) Between columns:

$$\mathrm{mmol}_i/\mathrm{Eq}_i = \mathrm{mmol}_j/\mathrm{Eq}_j$$

for rows i and j.

Additional automated corrections include:

- Removing any empty rows or columns and realigning elements,
- Checking and correcting column headers (e.g. FW being read as MW), and
- Validating that all columns except the Reagent column only contain numbers.

E. ELN Style Visualization

A primary goal for this work is to simplify the task of creating "AI-Ready" representations of archival lab notebook contents. Pursuant to this, following the post processing of the extracted content, an archive of the notebook is generated that

contains all page contents in a JSON file, as well as images of all the segmentations and cropped page elements. We have also developed a mock ELN interface capable of visualizing this archive, as shown in Figure 4.

F. Results

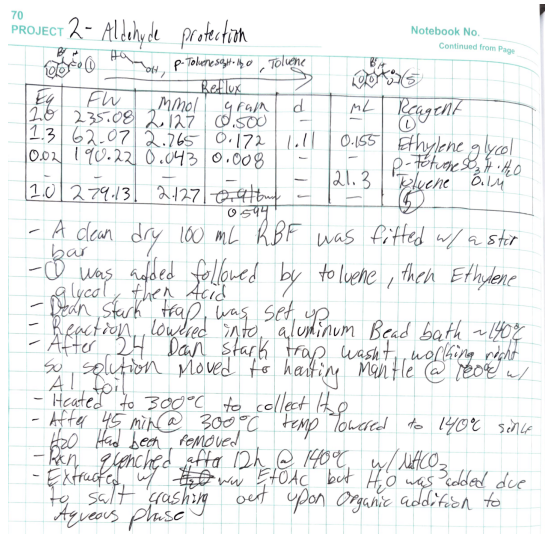
For notebook tables, results evaluation is based on the per cell accuracy of the final extraction output. The accuracy of each procedure was evaluated on thirty experiment pages from a single notebook author. Including the table headers, this sampling of pages contains 1,092 cells. For the process involving Detectron2 and Handwriting OCR, the output directly from the OCR program resulted in 199 incorrect fields for an error rate of 18.22%. olmOCR 2, when fed images directly without any segmentation, resulted in only 61 incorrect fields (5.59% error rate). When the post processing described in Section III-D1 is applied, the Detectron2/Handwriting OCR workflow error rate drops to 62 incorrect field (4.76% error rate), and olmOCR 2 drops to 46 (4.21% error rate).

G. Work in Progress

With the core automated archival aspects of this project completed, we are now working on converting the contents of text blocks into a tokenized list of actions as shown in Figure 5. This work is an extension of previous work using large language models to similarly tokenize methodology sections from MOF research papers [17]. Tokenizing protocols in this manner can greatly improve ML analyses of their content [18], thus increasing the data's "AI-readiness" and the potential for scientific knowledge discovery. Once tokenized and rectified with the quantitative data found in the tables, we plan to begin analyzing these notebook records more deeply in order to gain novel insights into MOF experimental design and the pedagogy of MOF synthesis.

IV. Discussion

We have begun to address our overarching research goal, which is to advance methods for converting scanned handwritten lab notebooks into structured, computer-accessible data for scientific analysis and AI use. The results reported in this paper, and our ongoing research in this area, demonstrate that laboratory notebooks, which have been overlooked in AI workflows, can be transformed into AI-ready resources. We have developed a workflow that takes an image of a lab notebook page and outputs a digital record that can be displayed in a form equivalent to ELN records. Part of our work identified a number of common errors resulting from the OCR of notebook pages, and developed a postprocessing program to check, correct and/or flag anomalies and mistakes in table records. Additionally, the recent release of olmOCR 2 shows promise for greatly accelerating our current work and many other archival tasks of this nature. While this research is in early stages, particularly our use of olmOCR which still needs to be fully integrated into our existing codebase, this



OCR

– A clean 100 mL dry round-bottom flask (RBF) was fitted with a stirrer.
– 1 was added, followed by toluene, then ethylene glycol, and then acid.
– A Dean-Stark apparatus was set up.
– The reaction was lowered into an aluminum bead bath at approximately 140C.
– After 24 hours, the Dean-Stark trap was checked; it was working correctly, so the solution was moved to a heating mantle at 120C with aluminum foil.
– Heated to 300C to collect H2O.
– After 45 minutes at 300 C, the temperature was lowered to 140C since H2O had been removed.
– The reaction was quenched after 12 hours at 140 C with NHCO2.
– Extracted with EtOAc, but H2O was added due to salt crashing out upon organic addition to the aqueous phase.

Structure Contents

Action	Entity	Duration	Temperature	Size
dried	RBF			100 mL
fitted	stirrer			
added	1			
added	toluene			
added	ethylele glycol			
added	acid			
set up	Dean-Stark apparatus			
lowered	aluminum bead bath		140C	
checked	Dean-Stark trap	24 hours		
moved	heating mantle		120C	
heated			300C	
lowered		45 minutes	140C	
quenched	NHCO2	12 hours		
extracted	EtOAc			

Fig. 5. The process of converting a synthesis protocol into a tokenized list of actions. Actions are highlighted in yellow, entities/objects involved in given action are highlighted in red, durations are highlighted in green, and temperatures are highlighted in blue.

work represents an important step towards improving the long-term overall accessibility and reuse of analog historical lab data.

The pace of technological advancement in document processing tools has had a direct impact on our work. The release of olmOCR 2, together with rapid developments in vision language models and extraction workflows [19], [20], underscores how quickly the research landscape is changing. Bulk processing of paper lab notebooks into highly accurate digital representations as if they were an ELN, previously significantly challenging for smaller research laboratories, is now feasible due to vision language models and other recent advancements [2], [21]. Our research highlights the reality that AI-ready data workflows need to remain flexible, given the continued rapid evolution of tools and associated technologies.

In assessing the implications of our work, it is important to also acknowledge that AI-ready data does not rely solely on the robustness of digital archival tools, but also on documentation practice and recording notes in a legible and discernible structure. Scientists need to keep AI-readiness in mind when recording their experiments in analog notebooks. This is especially relevant in chemistry laboratories or other similar settings, where there remains a dependency on analog notebooks. The greater the consistency, readability, simplicity, and accuracy of the initial pages being input into an extraction pipeline, the greater the intrinsic value they hold for scientific study and archival study. These fundamental practices are important for generating reliable digital and computationally ready data that can support AI and reproducibility.

V. Conclusion

In this paper we demonstrate a pipeline to convert scans of archival paper lab notebooks into a digital representation akin to that provided by ELNs. This pipeline is enabled by software toolkits that can segment and OCR the handwritten notebook pages, including Detectron2, Handwriting OCR, and more recently olmOCR 2. We then refine the output of these extraction tools via automated correction of common error patterns found in the digitized records, generate an archive of the data, and present the contents of the archive in an ELN-style web interface. Next steps include fully integrating olmOCR 2 into our workflow and testing it with a greater variety of lab notebooks, further refining our extraction pipeline, and continuing to work towards using the data within these notebooks as a resource to answer scientific questions. Additionally, we intend to quantify findings about notebook design and quality in a framework that would seed a dialog on how to best structure the analog lab notebooks that wetlabs continue to produce, so that they are more AI-ready in the future.

Acknowledgment

This work was supported by the National Science Foundation through the Institute for Data-Driven Dynamical Design, OAC-2118201. We acknowledge the original authors of the lab notebooks: Wesley Newsome, and Dimitrious Vazquez-Molina.

References

[1] J. Pepper, E. Jones, X. Zhao, J. Furst, K. Langlois, F. Uribe-Romo, D. Breen, and J. Greenberg, "AI-ready data: Knowledge extraction from archival lab notebooks," in *IEEE International Conference on Big Data (BigData)*, 2024, pp. 2489–2495.

[2] J. Poznanski, L. Soldaini, and K. Lo, "olmocr 2: Unit test rewards for document ocr," 2025. [Online]. Available: https://arxiv.org/abs/2510.19817

[3] J. G. Lennox, "Aristotles biology," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Stanford, CA: Metaphysics Research Lab, Stanford University, 2006. [Online]. Available: https://plato.stanford.edu/entries/aristotle-biology/

[4] A. Blair, "The rise of note-taking in early modern europe," *Intellectual History Review*, vol. 20, no. 3, pp. 303–316, 2010.

[5] F. Bacon, *Novum organum*. Clarendon press, 1878.

[6] H.-J. Rheinberger, "Scrips and scribbles," *MLN*, vol. 118, no. 3, pp. 622–636, 2003.

[7] F. L. Holmes, *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity*. Madison, WI: University of Wisconsin Press, 2012. [Online]. Available: https://archive.org/details/lavoisierchemist0000holm

[8] F. L. Holmes, J. Renn, and H.-J. Rheinberger, *Reworking the Bench: Research Notebooks in the History of Science*. Springer, 2003.

[9] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[10] F. L. Holmes, "Laboratory notebooks: Can the daily record illuminate the broader picture?" *Proceedings of the American Philosophical Society*, vol. 134, no. 4, pp. 349–366, 1990.

[11] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad, "Archival records and training in the age of big data," in *Re-Envisioning the MLS: Perspectives on the future of library and information science education*. Emerald Publishing Limited, 2018, vol. 44, pp. 179–199.

[12] T. R. Schellenberg *et al.*, *Modern Archives*. University of Chicago Press Chicago, IL, 1956.

[13] T. R. Schellenberg, "Archival principles of arrangement," *The American Archivist*, vol. 24, no. 1, pp. 11–24, 1961.

[14] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[15] J. Poznanski, A. Rangapur, J. Borchardt, J. Dunkelberger, R. Huff, D. Lin, A. Rangapur, C. Wilhelm, K. Lo, and L. Soldaini, "olmocr: Unlocking trillions of tokens in pdfs with vision language models," 2025. [Online]. Available: https://arxiv.org/abs/2502.18443

[16] "Handwriting OCR." [Online]. Available: https://www.handwritingocr.com/

[17] X. Zhao, J. F. F. Rojas, J. Furst, K. Ardila, K. Langlois, Y. An, X. Hu, F. Uribe-Romo, D. Gomez-Gualdron, and J. Greenberg, "Expert-Guided LLM Approach for Sequence-Aware Extraction of MOF Synthesis," Aug. 2025. [Online]. Available: https://chemrxiv.org/engage/chemrxiv/article-details/689c179c728bf9025e39cb58

[18] A. Li, X. Wang, W. Wang, A. Zhang, and B. Li, "A survey of relation extraction of knowledge graphs," in *Web and Big Data: APWeb-WAIM 2019 International Workshops, KGMA and DSEA, Chengdu, China, August 1–3, 2019, Revised Selected Papers 3*. Springer, 2019, pp. 52–66.

[19] X. Tang, X. Li, Y. Ding, M. Song, and Y. Bu, "The pace of artificial intelligence innovations: Speed, talent, and trial-and-error," *Journal of Informetrics*, vol. 14, no. 4, p. 101094, 2020.

[20] Y. Qin, Z. Xu, X. Wang, and M. Skare, "Artificial intelligence and economic development: An evolutionary investigation and systematic review," *Journal of the Knowledge Economy*, vol. 15, no. 1, pp. 1736–1770, 2024.

[21] J. Zhang, Y. Liu, Z. Wu, G. Pang, Z. Ye, Y. Zhong, J. Ma, T. Wei, H. Xu, W. Chen, Z. Wang, Q. Ji, F. Zhou, Q. Zhang, Y. Hu, J. Liu, Z. Li, Z. Zhang, Q. Liu, and X. Bai, "MonkeyOCR v1.5 Technical Report: Unlocking Robust Document Parsing for Complex Patterns," Nov. 2025. [Online]. Available: https://arxiv.org/abs/2511.10390v2