AI-Ready Data: Knowledge Extraction from Archival Lab Notebooks

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

Elizabeth Jones
Department of Computer Science
Northeastern University
Boston, MA USA
jones.eliza@northeastern.edu

Xintong Zhao
Department of Information Science
Drexel University
Philadelphia, PA USA
xz485@drexel.edu

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

Kyle Langlois
Department of Chemistry
University of Central Florida
Orlando, FL USA
kylerlanglois@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

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825206

***Abstract*—Collections of analog lab notebooks are an invaluable source of data about research conditions, steps, and outcomes, and in aggregate have the potential to provide new insights into the successes, failures and pedagogy of research laboratories. Unfortunately, these artifacts are increasingly at risk of being lost from the historical scientific record, given limited archiving and an absence of computational and AI readiness. This paper reports on research addressing this challenge by testing mechanisms for transforming digital scans of analog lab notebooks into AI-ready data resources. The research being pursued is framed by the field of computational archival science (CAS) and the aim to utilize analog, research lab notebook data for scientific study. The paper presents background context on archival lab notebooks and CAS, discusses MOF (metal organic frameworks) and COF (covalent organic frameworks) synthesis – the scientific domain of the lab notebooks under study, and details our research methods. We demonstrate a promising approach that automatically segments pages into discrete entry types, extracts the contents of those entries, refines the output and assesses the automated results. These efforts represent a first step towards developing a framework for both improving the usability of archival lab notebooks, and enabling their contents to be used in subsequent scientific inquiry.**

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

I. Introduction

Over the past two decades, traditional paper-based laboratory (lab) notebooks have been increasingly replaced by electronic lab notebooks (ELNs). This transition is significant, as ELNs streamline workflows by allowing real-time data entry, integration with lab instruments, and direct linking to files and databases [1]. Moreover, they help standardize recording processes, minimize the risk of data loss, and enable data sharing and collaborative research [2]. Given these benefits, scientific lab managers have increasingly prioritized ELN adoption and implementation, relegating analog lab notebooks to archival status with limited access. ELN prioritization makes sense in our highly computational, data-driven world; however, this transition places paper-based lab notebooks into the category of "data at risk" [3], [4], as valuable data is lost from the record of scientific knowledge [1].

Research is needed to address the challenge of converting analog, archival lab notebooks into computationally ready data resources. We are investigating approaches to address these issues, as part of the NSF supported Institute for Data Driven Dynamical Design (ID4). Our research group includes chemists at the Reticular Synthesis and Materials Design Lab (RSMDL), University of Central Florida, who focus on MOFs/COFs; and computer/information/data scientists at the Metadata Research Center, Drexel University. The research team is collaboratively working with digital, scanned images of archival lab notebooks from the RSMDL that record MOF and COF research (e.g. Figure 1). Our efforts to convert the data recorded in these notebooks into computationally ready data is grounded in computational archival science (CAS) [5], which targets the application of computational techniques to extract, process, and examine data recorded in archival artifacts for scholarly and scientific study.

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

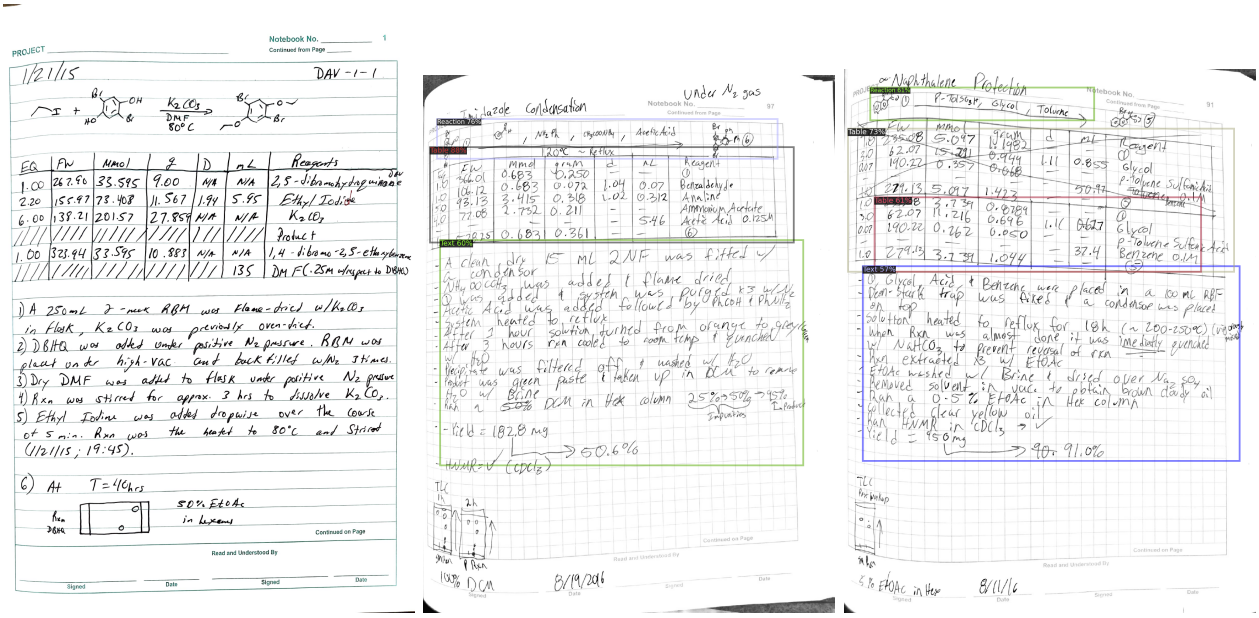

Fig. 1. Pages from organic synthesis of intermediates for MOFs in the notebooks. (Left) A plain, unprocessed page. (Center) A page with fairly accurate automated segmentations overlaid. (Right) A page containing a crossed out table that causes segmentation issues.

This paper reports on initial steps and results of this computationally ready lab notebook project. The research is motivated by the desire to identify various scientific research trends and outcomes that are present, but currently undisclosed, in the raw data within the notebooks. For example, we wish to answer certain questions about methods that lead to successful, as well as unsuccessful, results. Additionally, we would like to understand how the competency and abilities of lab technicians evolve over time in order to improve training methods for future technicians.

The remaining sections are organized in the following way. A background section provides important context for the association between archival lab notebooks and the field of CAS, and for MOF and COF research. Next, project goals and objectives are presented, followed by the results, and then a discussion of the results' implications and future work for the project. The paper wraps up with a conclusion that summarizes key findings and next steps.

II. Background Context

A. Lab Notebooks & Archival Computational Science

Lab notebooks are a fundamental component of scientific research, as they offer a record of scientific activities [6], [7]. Chemistry laboratories record data and metadata of synthesis experiments and chemical reactions in these notebooks. Due to the nature of chemical synthesis research, chemistry lab notebooks are traditionally physical books made of chemical-resistant paper, for record the synthesis experiment. Lab notebooks are generally curated by the researchers involved (e.g., a student and the principal investigator(s)), and stored as archival records once a researcher has left the group. As archival artifacts, lab notebooks reflect archival principles of *original order*, documenting the order in which the researchers pursued their steps, and a collective *provenance* in terms of the researchers involved [8], [9].

The analysis of (semi) unstructured documents is a multi-faceted problem and an active area of research [10]. Handwritten lab notebooks, with the inclusion of unstructured tabular data, diagrams, and reactions, bring challenges beyond those covered in much of the literature. The digitization of analog lab notebooks, with the goal of seeking further computational use of these artifacts, presents a critical link to the field of computational archival science. Integrating archival science with computational methods supports the management, preservation, and accessibility of historical analog records. These combined approaches also provides an essential framework for leveraging machine learning (ML) and other computational approaches that allow researchers to parse and study archival artifacts in new ways. Overall, the CAS rubric, bolstered by computational thinking [5], [11], [12], informs our work on the AI-readiness of lab notebooks to study the development of MOFs and COFs.

B. MOFs and COFs Synthesis

Metal-organic frameworks (MOFs) and Covalent Organic Frameworks (COFs) are a class of crystals made from the self-assembly of molecular building units into predetermined molecular architectures that exhibit unique properties, such as high porosity and chemical tunability. MOFs have attracted attention because of their potential use in advanced technologies, such as gas storage and separation, carbon and water capture, drug delivery, catalysis, and optoelectronic applications [13].

Researchers in the RSMDL (Reticular Synthesis and Materials Design Lab) are trained in organic synthesis to prepare the molecular components of MOFs and COFs, and also on their

solid-state crystallizations. Molecular synthesis refers to using multistep organic synthesis methods to prepare molecular compounds that are the starting materials for MOFs/COFs, as well as intermediates. Solid state crystallizations refer to the preparation of MOFs with targeted crystal morphologies like bulk microcrystalline powders, single crystals, thin films or grafted crystals.

Members of the RSMDL have traditionally recorded their research in paper-based notebooks. As with many labs, the RSMDL has increasingly adopted computational tools and workflows, and while analog lab notebooks are still used, research data documenting experiment processes are stored in other mechanisms. A significant challenge remains, however, in that there is a historical collection of lab notebooks that students have left, and which could inform scientific research more easily if they were computationally ready. The noted challenges in the RSMDL are a microcosm for the needs of other laboratories with historical collections. Even fairly recent collections of analog lab notebooks could benefit from the application of CAS approaches. It is this recognition that motivates the partnership underlying this research and the overall goal and objectives addressed in the next section.

III. Goals and Objectives

The overall goal of our work is to advance approaches for making the information contained in analog lab notebooks AI-ready. These approaches will facilitate the answering of scientific questions. Specific objectives include:

- converting scanned notebook pages into a digitized, structured form,
- designing vectorized/graph-based, machine learning-compatible representations of notebook contents,
- extracting information elements from the scans in these data representations, and
- performing document classification and clustering analysis to answer scientific questions based on the extracted elements.

IV. Methodology

This section presents our methods, covering input data, object detection, data extraction, and our approach to data analysis and assessment. A review of relevant data extraction techniques is also provided.

A. Input Data

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. Our input data came from two complete notebooks from two distinct authors at the RSMDL. Lab notebook pages were photographed using a camera held above the page, and are aligned and cropped in variable ways. One of the notebooks contains 101 pages, and the other contains 186.

Each notebook is assigned a three letter code that refers to the initials of a student-researcher, followed by a roman numeral that indicates the volume number, increasing in chronological order. Each synthesis data sheet is labeled with the lab notebook name and volume, followed by the page number, and an optional letter counter. Each synthesis data sheet contains the date the experiment was run, a sketch of the chemical reaction performed, a table of stoichiometries, and a section with notes and observations. In the stoichiometry table, the utilized chemicals are indicated by their name or by an identifier that links to the chemical reaction (especially when names of chemicals are too long or obscure), and by the amount utilized (in mass, mol, or volumes). Data such as theoretical yield and initial concentration of substrate is also indicated.

B. Object Detection

Information found on notebook pages includes text, tables, diagrams, equations, chemical reactions, and more depending on the procedure being performed. After reviewing the notebooks in our study, we determined there to be three major entry types worth focusing on: text, tables and chemical reactions. Some calculations and diagrams found in the notebooks were excluded, as these entry types were infrequently present and would likely be difficult to leverage in machine learning analysis.

In order to perform automated page segmentation, we make use of the Detectron2 object detection platform [14], specifically its implementation of the Faster R-CNN model. We hand labeled a training set from the two notebooks under study, which consisted of 101 instances of tables, 101 instances of reactions, and 153 instances of text blocks. The model was trained with a learning rate of 0.02 and a batch size of 128 for 1,000 epochs. 20% of images in the training set were reserved for validation, on which the final loss value was 0.107. A review of model performance can be found in Section V-A.

C. Data Extraction

With bounding boxes determined for entries on a given page, the next step is to extract and digitize their contents.

1) Text and Tables: While digitizing handwritten text is a very active area of research [15], [16], open-source, locally executable tools lag behind commercial tools in accuracy and features [17]. Two open-source programs to come out of academia are OrigamiNet [18] and DAN [19], with cloud-based offerings being provided by Amazon, Google, Microsoft and OpenAI. Of particular relevance to this project is the processing of tables, which requires borders to be interpreted as separators (i.e. commas) in the output.

A commercial, cloud based software called Handwriting OCR is available which is capable of handling tabular inputs, as well as plain text, with an accuracy comparable to the other major cloud providers [20]. We are currently using this software to extract the contents of text and tables. An API is provided to upload documents, with flags to process the uploaded material as text or as a table. Each entry on each page is cropped down to its bounding box and is uploaded as a separate document to ensure that the various page elements

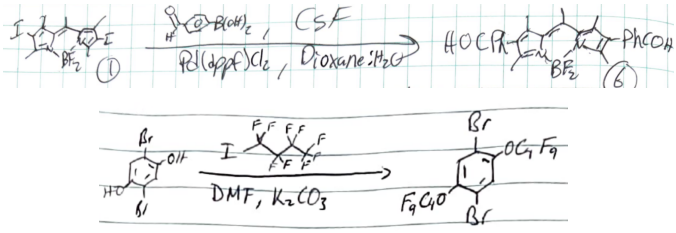

Fig. 2. Examples of synthesis equations found in the notebooks.

do not impact each other. The element is returned either as plain text or a spreadsheet file that can be easily converted to a basic comma separated value (CSV) representation.

2) Reactions: Several recent papers have been published on optically recognizing hand drawn organic molecules. The oldest presents a tool called ChemPix specifically targeting hand drawn inputs, but its trained model seems to have never been released [21]. This project also only targeted fairly simple hydrocarbon structures, and would likely not be able to identify many of the molecules in our notebooks. The second work on this topic is called Img2Mol. While it can handle more complex molecules, it is not tailor-made for hand-drawn inputs and its performance on them is likely not high enough to be useful [22]. The final and most recent two approaches are the "new" DECIMER model [23] and ChemReco [24]. ChemReco only targets molecules comprised exclusively of carbon, hydrogen and oxygen, but DECIMER is potentially capable of handling the molecules involved in MOF synthesis.

Attempting to extract the full equations remains as future work, due to the need to further segment the quite variable structure of the graphical reaction equations before feeding anything to optical recognition tools. As shown in Figure 2, the reagents and solvents noted on the reaction arrow can themselves be quite complex, and they, as well as the reactants and products, can contain chemical shorthands that may or may not be recognized by optical tools.

D. Data Analysis and Assessment

Pursuant to creating a robust foundation on which to perform subsequent data processing and ML analysis, we have designed two interfaces to aid in assessing and improving automated segmentation accuracy. These interfaces also help assess the existing archival challenges that limit or confound the automated extraction of accurate information from the notebooks. The interfaces, one for refining segmentations and one for assessing them, are shown in Figure 3. Statistics and discussion based on these data are presented in Section V.

The refinement interface allows a user to redraw bounding boxes, remove spurious ones, and add missing ones. Entries are listed with their object detection confidence scores, along with their automatically extracted contents. Users can simply pick which entry to refine, and redraw its bounding box with the mouse. The extracted contents can also be manually refined. Both content versions are saved in a database for subsequent analysis and potential archiving.

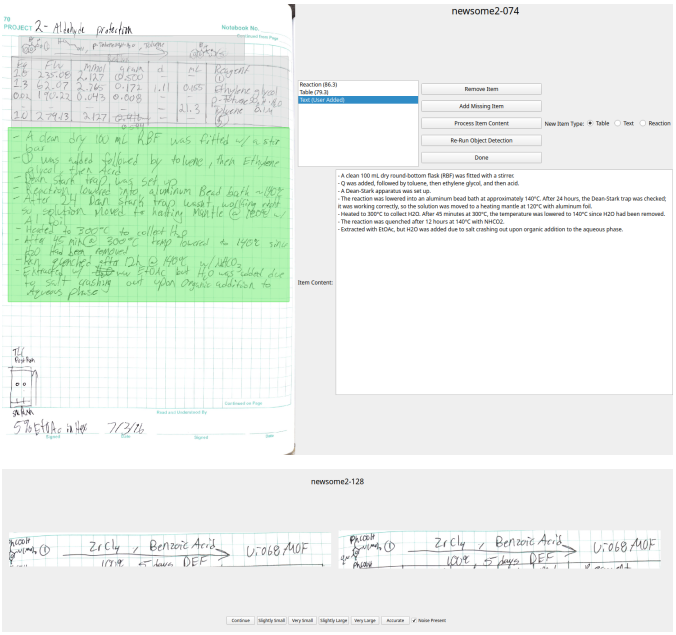

Fig. 3. The analysis and refinement interface (top) and bounding box accuracy assessment interface (bottom).

The analysis interface, intended to be run after using the segmentation refinement interface on all pages, shows both the original and user-drawn versions of each entry (where applicable). Its purpose is to allow a user to determine, based on pertinent, cropped information or unrelated, added information, if the automatically drawn bounding box was only slightly too large/small and still likely sufficient for accurate content extraction, or far too large/small and very likely to cause processing errors. If an entry was added by the user, deleted by the user, or the user did not redraw the bounding box, then it is skipped in the interface and automatically recorded as deleted, added or accurate in the database. There is an additional flag for noise, i.e. unrelated overlapping objects, falling inevitably within the bounding box of the entry. An example of this is shown in Figure 4. Discussions on the implications of this issue are in Section V-B, and potential automated solutions to it are proposed in Section VI.

V. Results

A. Object Detection Performance

The training data spanned 133 of the 287 total notebook pages, leaving 154 pages for the testing set and manual review. The true instance counts for each type of entry in these pages is 123 tables, 131 text blocks and 124 chemical reactions, for a total of 378 entries. The cutoff confidence score from the object detection model for including an object was set to 50%. The performance of the model is detailed in Table I, with specific quality labels for segmentations in the first section and more general quality levels in the second.

The model produced 41 truly perfect bounding boxes that could not be improved upon. A total of 248 segmentations

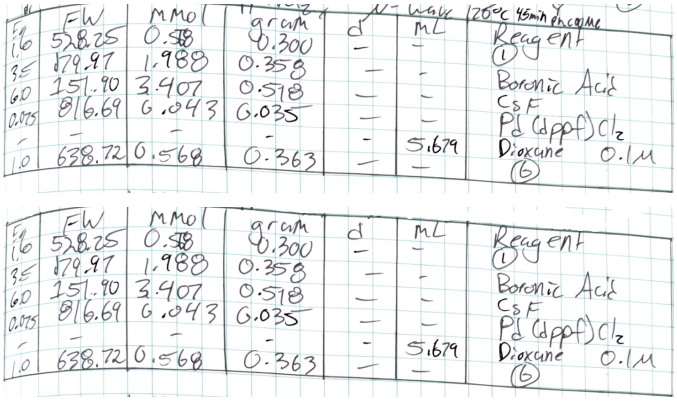

Fig. 4. (Top) An example of a table entry which, if cropped to capture all its content, inevitably contains noise from the preceding reaction entry. (Bottom) The entry with its noise manually removed.

were only slightly too small or slightly too large, and captured the entries contained within well enough that their use would likely cause minimal or no issues further down the analysis pipeline. 30 segmentations were either far too small or far too large, and do not accurately capture the entries. 53 of the bounding boxes it produced were for nonexistent entries, and it missed 50 entries entirely. Notably, all of the 50 missed entries were text blocks, no reactions or tables were missed.

Excluding the erroneous segmentations, 298 out of 378 entries (78.8%) have automated bounding boxes that accurately capture their contents. Of the 53 erroneous detections, 36 (67.9%) had a confidence score below 60%, indicating the cutoff threshold may be too low.

B. Data Quality

During the results assessment process, some data quality issues became apparent that may contribute to difficulties in notebook archiving and analysis:

1) Noise: On some pages, the notebook's author recorded reactions, tables and/or text extremely close to one another. In those cases, reactions even cut into tables, as in Figure 4. While not a problem for a human reading these notebooks, this is a serious issue for content extraction. When feeding entries, particularly tables, into additional processing software, this noise will be included in the output unless it is somehow removed, one possible avenue for doing so is discussed in Section VII. Of the 378 testing set entries, 59 (15.6%) have nontrivial noise within their bounding boxes.

2) Page Alignment: The notebooks were captured by placing them under a camera held by a stand, and the placement, alignment and flatness of the pages can vary significantly. This generally did not cause problems for object detection, and does not seem to directly impact the OCR software, but it does exacerbate the previous noise problem, and could impact other analysis tasks depending on one's research objective(s). The general variability of page captures can be seen in Figure 1, and the flatness issue in particular is quite notable in Figure 4.

TABLE I
OBJECT DETECTION RESULTS

Bounding Box Quality	Count
Perfect	41
Erroneous	53
Missed	50
Slightly Small	176
Slightly Large	81
Very Small	27
Very Large	3
Acceptable Quality	298
Unacceptable Quality	80
Erroneously Labeled	53

3) TLC Diagrams: There are some experiment-specific thin layer chromatography (TLC) diagrams found in our notebooks. They are used to monitor the course of chemical reactions, and are not currently included in the object detection model. Some are reminiscent of tables, as shown in Figure 5. Converting any diagram in these notebooks to a machine-readable representation would likely require a bespoke ML model, which is why they are not included in our current study. However, for the purpose of better distinguishing them from tables, it may be prudent to add them in as a fourth class.

4) Table Style Variations: A more minor complexity in the data is the authors drawing tables using distinct styles. The primary difference that impacts analysis is that one author drew diagonal slashes in empty table records, and the other drew hyphens. The diagonal slashes are interpreted as 1's by the OCR software in some cases, which needs to be accounted for in subsequent processing.

5) Corrections and Amendments: The final major data extraction complication observed in the notebooks is author correction and amendments. These include entries as small as one crossed out number with the replacement written next to it, up to entire pages crossed out. An example of this is shown on the right in Figure 1, where an entire table has been crossed out and rewritten directly below. This is not a situation the object detection model could handle, and it segmented both tables together, as well as the second by itself.

VI. DISCUSSION

Lab notebooks offer unique challenges in terms of document analysis and archiving. Their structure is more variable than most types of paper records, in addition to being hand-written with many types of diagrams, figures, and error corrections. The work completed thus far represents the first phase of building a robust analysis pipeline for this type of document, with ongoing and future work consisting of structuring the unstructured notebook data for analysis, and performing said analysis with the objective of answering open questions about experimental quality and student learning. The next step in data processing is to develop an automated technique for denoising entries. One possible pathway to removing unwanted entries captured in the bounding boxes of neighboring entries is by passing them through an autoencoder trained to remove

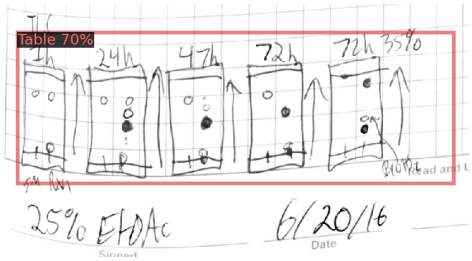

Fig. 5. An example of thin layer chromatography (TLC) diagrams used to monitor reactions, an entry outside the scope of the current segmentation model, being misinterpreted as a table.

the noise as a preprocessing step before running the optical recognition tools [25].

Following the completion of all data extraction work, we aim to develop a structured graph representation for the notebook data, and use that representation to perform document classification/clustering experiments with the goal of answering scientific questions about the experimental procedures. The objectives of this portion of the project partially align with aspects of the text document clustering literature [26], as we are aiming to discern meaningful patterns within the textual records of experimental procedures. To achieve this, both table data and unstructured textual data, such as synthesis procedure descriptions, will need to be integrated into an ML analysis compatible representation. Currently, we are working to develop a knowledge extraction pipeline to identify important entities (i.e. key words and phrases) based on the text found in peer-reviewed MOF and COF research articles. We are working to construct key phrases from the literature into triple pairs, where two entities are linked by their relation. By running this pipeline, we aim to transform information buried in un-queryable data into structured, machine-readable data. These triplets can greatly enhance the content quality of knowledge graphs [27], and enable better scientific knowledge discovery.

Chief among the scientific questions that we are seeking to answer is what patterns in experimental procedures might distinguish successful synthesis experiments from unsuccessful. What constitutes "success" varies depending on the procedure being performed, but the yield of final product is a common metric. In addition, these notebooks may contain insights on student learning. These pedagogic questions regarding student-researcher training include:

- What is the progress of training in crystal engineering of MOFs, COFs, and related materials?
- When is a student properly trained at chemical syntheses of both organic and solid-state compounds?
- Are there any data-based hints that can be utilized to predict the progress of future researchers?

Unsupervised clustering of experimental records may also lead to the discovery of patterns not yet postulated explicitly.

VII. Conclusion

The overall goal of our work is to advance approaches for making the information contained in analog lab notebooks AI-ready, which will facilitate the answering of scientific questions. We have developed a process to extract the contents of scanned lab notebook pages, analyzed the results and presented potential challenges with data quality and archiving. This initial research effort helps to frame next steps. Each component of our work to date has been conducted not only to make archived lab notebooks AI-ready, but also to begin the process of building a pipeline for scientific inquiry based on the extracted data.

A longer term goal of our project is to apply our extraction and analysis approach to the full collection of digitized lab notebooks at the RSMDL, University of Central Florida, as well as other labs that have digitized copies of analog MOF and COF notebooks. The methods and pipeline we are working to develop may then be further adapted to prepare AI-ready data from digitized lab notebooks capturing research across other areas of experimental chemistry. Finally, the research presented in this paper can help mitigate loss of historical scientific artifacts and make sure that AI operations are informed by a more complete record of research.

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] C. L. Bird, C. Willoughby, and J. G. Frey, "Laboratory notebooks in the digital era: the role of elns in record keeping for chemistry and other sciences," *Chemical Society Reviews*, vol. 42, no. 20, pp. 8157–8175, 2013.

[2] S. G. Higgins, A. A. Nogiwa-Valdez, and M. M. Stevens, "Considerations for implementing electronic laboratory notebooks in an academic research environment," *Nature Protocols*, vol. 17, no. 2, pp. 179–189, 2022.

[3] C. A. Thompson, W. D. Robertson, and J. Greenberg, "Where have all the scientific data gone? LIS perspective on the data-at-risk predicament," *College & Research Libraries*, vol. 75, no. 6, pp. 842–861, 2014.

[4] M. S. Mayernik, K. Breseman, R. R. Downs, R. Duerr, A. Garretson, C.-Y. S. Hou, E. D. G. Initiative *et al.*, "Risk assessment for scientific data," *Data Science Journal*, vol. 19, pp. 10–10, 2020.

[5] 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.

[6] 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.

[7] F. L. Holmes, J. Renn, and H.-J. Rheinberger, *Reworking the bench: Research notebooks in the history of science*. Springer, 2003.

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

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

[10] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *Journal of Big Data*, vol. 11, no. 1, p. 92, Jul. 2024. [Online]. Available: https://doi.org/10.1186/s40537-024-00948-z

[11] N. Payne, "Stirring the cauldron: Redefining computational archival science (CAS) for the big data domain," in *Proc. IEEE International Conference on Big Data*, 2018, pp. 2743–2752.
[12] R. Marciano, "AFTERWORD: Towards a new discipline of computational archival science (CAS)," *Digital Humanities Research— Volume 2*, p. 205, 2021.
[13] O. M. Yaghi, M. J. Kalmutzki, and C. S. Diercks, *Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks*. John Wiley & Sons, Aug. 2019.
[14] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.
[15] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access*, vol. 8, pp. 142 642–142 668, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9151144
[16] W. AlKendi, F. Gechter, L. Heyberger, and C. Guyeux, "Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey," *Journal of Imaging*, vol. 10, no. 1, p. 18, Jan. 2024. [Online]. Available: https://www.mdpi.com/2313-433X/10/1/18
[17] M. S. Islam, M. K. B. Doumbouya, C. D. Manning, and C. Piech, "Handwritten Code Recognition for Pen-and-Paper CS Education," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, ser. L@S '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 200–210. [Online]. Available: https://dl.acm.org/doi/10.1145/3657604.3662027
[18] M. Yousef and T. E. Bishop, "OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 14 698–14 707. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9157803
[19] D. Coquenet, C. Chatelain, and T. Paquet, "DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8227–8243, Jul. 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10013687
[20] "Handwriting OCR." [Online]. Available: https://www.handwritingocr.com/
[21] H. Weir, K. Thompson, A. Woodward, B. Choi, A. Braun, and T. J. Martínez, "ChemPix: Automated recognition of hand-drawn hydrocarbon structures using deep learning," *Chemical Science*, vol. 12, no. 31, pp. 10 622–10 633, Sep. 2021. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc02957f
[22] D.-A. Clevert, T. Le, R. Winter, and F. Montanari, "Img2Mol – accurate SMILES recognition from molecular graphical depictions," *Chemical Science*, vol. 12, no. 42, pp. 14 174–14 181, Jul. 2021. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc01839f
[23] K. Rajan, H. O. Brinkhaus, A. Zielesny, and C. Steinbeck, "Advancements in hand-drawn chemical structure recognition through an enhanced DECIMER architecture," *Journal of Cheminformatics*, vol. 16, no. 1, p. 78, Jul. 2024. [Online]. Available: https://doi.org/10.1186/s13321-024-00872-7
[24] H. Ouyang, W. Liu, J. Tao, Y. Luo, W. Zhang, J. Zhou, S. Geng, and C. Zhang, "ChemReco: Automated recognition of hand-drawn carbon–hydrogen–oxygen structures using deep learning," *Scientific Reports*, vol. 14, no. 1, p. 17126, Jul. 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-67496-7
[25] A. Creswell and A. A. Bharath, "Denoising Adversarial Autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 968–984, Apr. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8438540/?arnumber=8438540
[26] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, p. 664, Feb. 2021. [Online]. Available: http://ijece.iaescore.com/index.php/IJECE/article/view/22867
[27] 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.
[28] X. Zhao, K. Langlois, J. Furst, Y. An, X. Hu, D. G. Gualdron, F. Uribe-Romo, and J. Greenberg, "Research evolution of metal organic frameworks: A scientometric approach with human-in-the-loop," *Journal of Data and Information Science*, 2024.