AI-Ready Data: Structured Processing of Laboratory Notebook Tables

Zach Siapno[1], Joel Pepper[1], Jane Greenberg[1], David E. Breen[1], Jacob Furst[2], Fernando Uribe-Romo[2]
[1]Drexel University, [2]University of Central Florida

NSF-OAC# 2118201

Motivation

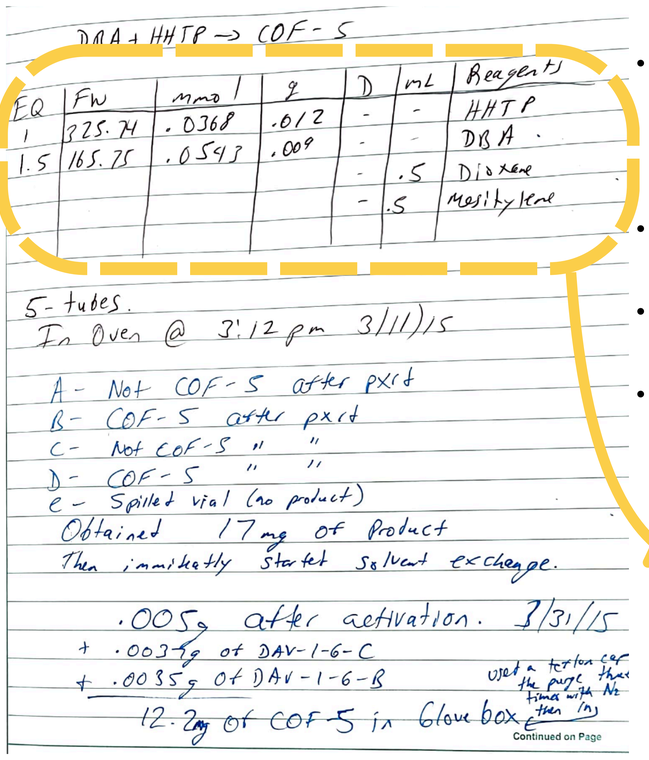
- A significant amount of invaluable data is buried in analog chemistry lab notebooks, which remains largely inaccessible without digitization.
- Research is needed to computationally examine digitized notebooks detailing synthesis experiments for metal-organic frameworks (MOFs), compounds with applications as conducting solids and supercapacitors.
- There is also a need to apply AI to lab notebook data in order to extract important information about chemistry processes, and lab technician training and experience.

Goals

Overall Goal: Develop methods for converting scanned handwritten lab notebooks into computer-accessible data for scientific analysis and AI use.

- Apply Optical Character Recognition (OCR) to extract data from the notebooks to support computational analysis of the information archived in these analog artifacts.
- Identify error patterns and factors that noticeably impact the accuracy of OCR outputs.
- Implement code-based solutions to clean and transform raw OCR outputs into structured Pandas Dataframe tables, enabling querying and organization.
- Contribute to the long-term improvement, reusability, and overall accessibility of analog historical lab data.

Methods and Procedures


DMA + HHTP → COF-5

EQ	FW	mmol	g	D	mL	Reagents
1	325.74	.0368	.012	-	-	HHTP
1.5	165.75	.0543	.009	-	-	DBA .
				-	.5	Dioxane
				-	.5	Mesitylene

5-tubes.
In Oven @ 3:12 pm 3/11/15

A- Not COF-5 after pxrd
B- COF-5 after pxrd
C- Not COF-5 " "
D- COF-5 " "
E- Spilled vial (no product)
Obtained 17 mg of Product
Then immediately started solvent exchange.

.005g after activation. 3/31/15
+ .003g of DAV-1-6-C
+ .0035g of DAV-1-6-B
12.2mg of COF-5 in Glove box
used a torlon cap the purge three times with N2 then in)
Continued on Page

- Over 100 tables of OCR'ed handwritten raw data were manually examined and analyzed.
- Common OCR error types (e.g. character misinterpretations, extra whitespace, duplicated values, etc.) were identified and classified.
- Segment the notebook's table data into readable strings through Python code.
- Attempt to remove as many OCR-related errors as possible from the initial OCR data.
- Cleaned data is reintroduced into a Pandas database for further validation and testing.

```
Raw OCR Output:
 EQ,FW,Mmo 1,9,D,ML,Reagents
1.0,325.74,. 0368,.0/2,-,-,HHTP
1.5,165.75,. 0543,",009",-,-,DBA
,,,,-,0.5,Dioxine
,,,,-,0.5,Mesitylene
,,,,,,
```

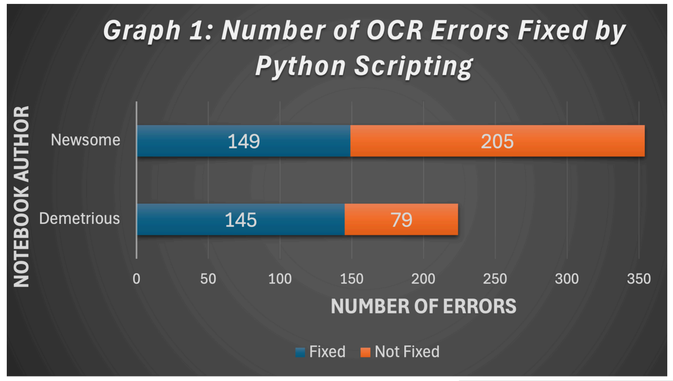
```
Cleaned Text:
 EQ,FW,mmol,g,D,mL,Reagents
1.0,325.74,.0368,.012,-,-,HHTP
1.5,165.75,.0543,.009,-,-,DBA .
-,-,-,-,-,0.5,Dioxine
-,-,-,-,-,0.5,Mesitylene
-,-,-,-,-,-,-
```

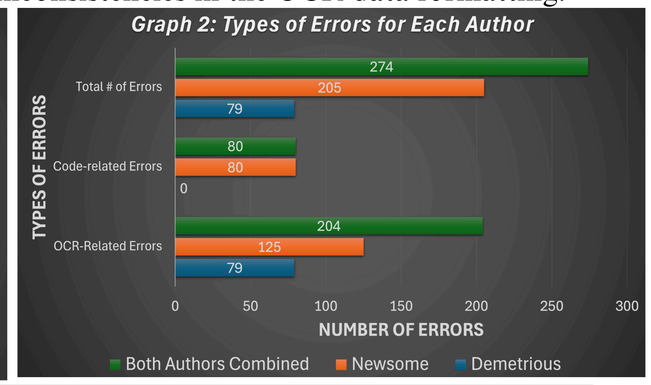
Handwritten tables from Demetrious's Notebooks are converted to Raw OCR data (bottom left), which is then converted into a cleaned string to be organized into a Pandas dataframe (top right).

Reagents	EQ	FW	mmol	g	D	mL	calculated_gram	passed_check	solvent
HHTP	1.0	325.74	0.0368	0.012	–	–	0.012	True	False
DBA .	1.5	165.75	0.0543	0.009	–	–	0.009	True	False
Dioxine	–	–	–	–	–	0.5	None	False	True
Mesitylene	–	–	–	–	–	0.5	None	False	True
–	–	–	–	–	–	–	None	False	True

Results

- The code ran through two of the lab notebooks from two different authors, with 100 total tables being tested: 50 entries from Demetrious and 50 entries from Newsome.
- 294 out of 578 errors were fixed; 284 errors lingered post-processing and examined if they were a result of lapses in the Python script or the original OCR output as shown in Table 2.
- Gained insight into OCR's limitations from Newsome's entries (excessive columns) that complicate data extraction; this resulted in Newsome entries having more errors and needing additional care and code iterations due to inconsistencies in the OCR data formatting.





Notebook	Total Errors	Fixed	% Fix Rate	OCR-Related Errors	Code-Related Errors
Demetrious	224	145	64.73%	79	0
Newsome	354	149	42.09%	125	80
Total	578	294	50.87%	204	80

Future Work

- In the future, it is important to test out current code function on a different set of data tables from different authors, as currently the program only works flawlessly on Demetrious' data entries but has some inaccuracies on Newsome's tables.
- More improvements can be made to address cases where the OCR falsely creates additional columns such as in every single one of Newsome's raw OCR outputs. Notably, correction rate for Newsome (42.09%) is way less efficient than my correction rate for Demetrious (64.73%).

References

[1] Pepper et al., 2024, "AI-Ready Data: Knowledge Extraction from Archival Lab Notebooks"
[2] Chenet al., 2011, "Table Detection in Noisy Off-Line Handwritten Documents"
[3] Dozias et al., 2018, "Smart Pens to Assist Fibre Optic Sensors Research"
[4] Franco-Gaona et al., 2024, "Towards the Automatic Extraction and Annotation of Information Elements from Handwriting Notes"
[5] Gaona et al., 2020, "Extracting Information Objects from Handwriting Laboratory Notes"
[6] Weir et al., 2021, "Chempix: Automated Recognition of Hand-Drawn Hydrocarbon Structures Using Deep Learning"

We acknowledge the original authors of the lab notebooks: Wesley Newsome, and Demitrious Vazquez-Molina