

AI-Ready Data: Knowledge Extraction from Archival Lab Notebooks

IEEE Big Data 2024, CAS Workshop

Joel Pepper¹, Elizabeth Jones⁴, Xintong Zhao², Jacob Furst³, Kyle Langlois³,
Fernando Uribe-Romo³, David Breen¹, Jane Greenberg²

¹Drexel University, Department of Computer Science

²Drexel University, Department of Information Science

³University of Central Florida, Department of Chemistry

⁴Northeastern University, Department of Computer Science

December 17, 2024

Background

Drexel UNIVERSITY | Search notebook | Joel Pepper | Untitled Page

My Notebook | Untitled Page (1) | Untitled Page | New... | Deleted Items

Joel Pepper - Dec 09, 2024, 11:21 AM EST

1	2	3	4	5	6	7	8	9	10	11
---	---	---	---	---	---	---	---	---	----	----


Joel Pepper - Dec 09, 2024, 11:29 AM EST

test.docx (313 KB)

Joel Pepper - May 07, 2020, 1:39 PM EST


Test

Joel Pepper - Dec 09, 2024, 11:14 AM EST



OGL88004173.JPG (1.44 KB)

Joel Pepper - Dec 09, 2024, 11:14 AM EST



aha_7190.jpg (291 KB)

PROJECT _____ | Notebook No. _____ | Continued from Page _____


1/2/15 | DAY - 1 - 1

CC1=CC=C(C=C1)O + Oc1ccc(O)c(O)c1 $\xrightarrow[\text{DMF, 80}^\circ\text{C}]{\text{K}_2\text{CO}_3}$ CC1=CC=C(C=C1)Oc2ccc(O)c(O)c2

EQ	FW	MMOL	g	D	mL	Reagents
1.00	267.96	33.585	9.00	N/A	N/A	2,5-dibromohydroquinone
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K ₂ CO ₃
///	///	///	///	///	///	Product
1.00	323.94	33.585	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
///	///	///	///	///	///	DMF (-25M of react to DMF)

- A 250 mL 2-neck RBM was flame-dried w/K₂CO₃ in flask, K₂CO₃ was previously oven-dried.
- DBHQ was added under positive N₂ pressure. RBM was placed under high-VAC and back filled w/N₂ 3 times.
- Dry DMF was added to flask under positive N₂ pressure.
- Rxn was stirred for approx. 3 hrs to dissolve K₂CO₃.
- Ethyl Iodide was added dropwise over the course of 5 min. Rxn was then heated to 80°C and stirred (1/2/15; 19:45).

6) At T=4hrs

Rxn  50% EtOAc in hexane

Continued on Page _____

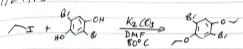
Read and Understood By _____ | Signed _____ | Date _____ | Signed _____ | Date _____

Background

- ▶ Gloves, chemicals, nature of work, complex reactions and custom diagrams make switch to digital notebooks unfeasible for chemists
- ▶ Notes recorded on special, chemical resistant paper
- ▶ Manual logging of paper notes as faithful digital copies extremely time intensive

PROJECT _____ Notebook No. _____
Continued from Page _____


1/21/15 DAY - 1 - 1



EQ	FW	MMol	g	D	mL	Reagents
1.00	267.96	33.585	9.00	N/A	N/A	2,5-dibromohydroquinone
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K ₂ CO ₃
///	///	///	///	///	///	Product
1.00	323.94	33.585	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
///	///	///	///	///	///	DMF (25mL reagent to DMF)

- 1) A 250mL 2-neck RBM was flame-dried w/K₂CO₃ in flask, K₂CO₃ was previously oven-dried.
- 2) BHTQ was added under positive N₂ pressure. RBM was placed under high-VAC and back filled with N₂ 3 times.
- 3) Dry DMF was added to flask under positive N₂ pressure.
- 4) Rxn was stirred for approx. 3 hrs to dissolve K₂CO₃.
- 5) Ethyl Iodide was added dropwise over the course of 5 min. Rxn was then heated to 80°C and stirred (1/21/15; 19:45).

6) At T=4hrs

 50% EtOAc in hexanes

Read and Understood By _____
Signed _____ Date _____ Signed _____ Date _____

Introduction

- ▶ Paper-based lab notebooks becoming “data at risk” [3, 4]
- ▶ Collections of notebooks may have the potential to provide new insights into the successes, failures and pedagogy of research labs
- ▶ Research is needed to address challenge of converting analog lab notebooks into computationally ready resources

[3]: Thompson, Data-at-risk predicament

[4]: Mayernik, Risk assessment for scientific data

Note: Citation numbers match those in our manuscript



Introduction

- ▶ We are investigating how to extract and structure the information contained in analog lab notebooks^a, in order to make them “AI-ready”
- ▶ Notebooks come from metal/covalent organic framework (MOF/COF) synthesis experiments
- ▶ 3 main goals:
 1. Automatically extract contents of pages^b
 2. Create a vectorized/graph-based, machine learning-compatible representation of contents
 3. Perform document classification and clustering analysis to answer scientific questions

^a Sourced from U of Central Florida Reticular Synthesis and Materials Design Lab (RSMDL)

^b Main focus of this talk

The image shows a digital lab notebook interface for 'newsome2-074'. On the left, a handwritten page is displayed with a table and text. The table has columns for 'Time', 'Temp', 'pH', 'Conc', 'Notes', and 'Remarks'. The text below the table describes a procedure: 'A clean dry 100 mL RBF was fitted w a stir bar...'. On the right, a digital interface shows a 'Reaction (M, S)' section with a table for 'Table (7/3/26)'. Below the table are buttons for 'Remove Item', 'Add Missing Item', 'Process Item Content', 'Re-Run Object Detection', and 'Done'. A 'View Content:' section below the buttons displays a list of extracted text items, including: 'A clean 100 mL dry round bottom flask (RBF) was fitted with a stir bar...'. At the bottom left of the interface, a small diagram shows a flask with a stir bar and the text '5. The stirrer is in the 7/3/26'.

The image shows a digital lab notebook interface for 'newsome2-128'. On the left, a handwritten page is displayed with a chemical reaction scheme: 'ZrCl₄ + Benzoic Acid → U-668-MOF'. On the right, a digital interface shows a 'Reaction (M, S)' section with a table for 'Table (7/3/26)'. Below the table are buttons for 'Remove Item', 'Add Missing Item', 'Process Item Content', 'Re-Run Object Detection', and 'Done'. A 'View Content:' section below the buttons displays a list of extracted text items, including: 'A clean 100 mL dry round bottom flask (RBF) was fitted with a stir bar...'. At the bottom left of the interface, a small diagram shows a flask with a stir bar and the text '5. The stirrer is in the 7/3/26'.

Methodology Overview

General content extraction workflow:

1. Segment pages into discrete entries
2. Extract contents from entries individually
3. Process output to improve accuracy if necessary (work in progress)
4. Build database, manually review results, add additional metadata

Project 2: Alkyl halide synthesis

Time	Temp	Pressure	Notes
13	250°C	0.127	0.0250
14	250°C	0.127	0.0250
15	250°C	0.127	0.0250
16	250°C	0.127	0.0250

5/11/2024 7/3/24

Object
Detection

- A clean dry 100 mL RBF was fitted w/ a stirrer
 - Q was added followed by toluene, then Ethylene glycol, then acid
 - Dean Stark trap was set up
 - Reaction lowered into aluminum bead bath ~140°C
 - After 24 hours Dean Stark trap wasn't working right so solution moved to heating mantle @ 120°C
 - Heated to 300°C to collect H₂O
 - After 45 min @ 300°C trap lowered to 140°C since H₂O had been removed
 - Reaction quenched after 12 hours @ 140°C w/ NH₄CO₂
 - Extracted with EtOAc but H₂O was added due to salt crashing out upon organic addition to aqueous phase

Entry
Specific
Processing

Optical
Character
Recognition

- A clean 100 mL dry round-bottom flask (RBF) was fitted with a stirrer.
- Q 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 140°C.
- After 24 hours, the Dean-Stark trap was checked; it was working correctly, so the solution was moved to a heating mantle at 120°C with aluminum foil.
- Heated to 300°C to collect H₂O.
- After 45 minutes at 300°C, the temperature was lowered to 140°C since H₂O had been removed.
- The reaction was quenched after 12 hours at 140°C with NH₄CO₂.
- Extracted with EtOAc, but H₂O was added due to salt crashing out upon organic addition to the aqueous phase.

Segmentation

- ▶ Make use of the Detectron2 object detection platform [14]
- ▶ Three entry types in model: text, tables and chemical reactions

[14]: Wu, Detectron2

Under N₂ gas

Titration Condensation

Notebook No. 97

Time	FW	MMol	gram	d	nL	Reagent
1:00	466.01	0.683	0.250	-	-	①
1:06:12	466.01	0.683	0.072	1.04	0.07	Benzaldehyde
1:10:13	466.01	3.415	0.318	1.02	0.312	Aniline
1:10:40	466.01	2.752	0.211	-	-	Ammonium Acetate
1:10:40	466.01	-	-	-	546	Acetic Acid 0.125M
1:10:40	466.01	0.683	0.361	-	-	②

A clean, dry 15 mL 2NF was fitted w/ condenser
 100 mg CDCl₃ was added & flask dried
 ① was added & system was purged w/ N₂
 Acetic Acid was added followed by PhCOCl & PhNH₂
 system heated to reflux
 After 3 hour solution turned from orange to grey
 After 3 hours rxn cooled to room temp & quenched
 The precipitate was filtered off & washed w/ H₂O
 solvent was green paste & taken up in DCM to reanalyze
 100% w/ Brine
 had a 50% DCM in Hex column 25% → 50% → 45%
 solutions
 - Yield = 182.8 mg
 Yield = 50.6%
 Purified (CDCl₃)

TLC
 100% DCM
 8/19/2021

Content Extraction – Tables & Text

- ▶ Digitizing handwritten text primarily a cloud-based task
- ▶ Need table processing, one provider of which is software called Handwriting OCR [20]
- ▶ Each entry is uploaded as a separate document, and returned either as plain text or a spreadsheet file

EQ	FW	MMOI	g	D	mL	Reagents
1.00	267.96	33.595	9.00	N/A	N/A	2,5-dibromohydroquinone ^{DAV}
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K ₂ CO ₃
////	////	////	////	////	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	////	////	////	////	135	DM F(.25M w/respect to DIBHG)

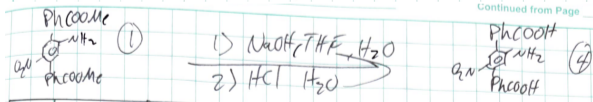
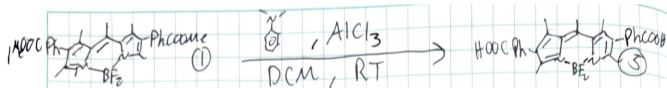
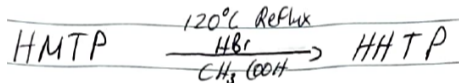
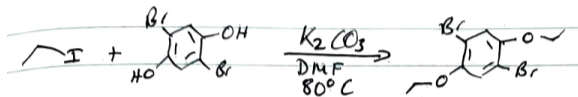


EQ,FW,MMOI,2,D,ML,Reagents
 1,267.9,33.595,9,N/A,N/A,"2,5 - dibromohydroquinone DAV"
 2.2,155.97,73.408,11.567,1.94,5.95,Ethyl Iodine
 6,138.21,201.57,27.859,N/A,N/A,K2LO3
 1/11,////,11/11,,111,1/11,Product
 1,323.94,33.595,10.883,N/A,N/A,"1,4 - dibromo-2,5- ethoxy benzene"
 11/1,11111,,1111,,135,DM F(.25M w/respect to DIBHG)

[20]: <https://www.handwritingocr.com/>

Content Extraction – Reactions

- ▶ Tools to parse chemical equations this complicated do not currently exist
- ▶ Lacking this capability likely of negligible impact to main aim of our research



Continued from Page

Manual Review – Analysis

- ▶ Two interfaces for assessing and improving automated segmentation accuracy
- ▶ Refinement interface used to redraw, remove and add bounding boxes

newsome2-074

Reaction (M4)
Table (75)

Remove Item
Add Missing Item
Process Item Context
Re-Run Object Detection
Done

Item Context

Reaction (M4)
Table (75)
News Item Type: Table Text Reaction

A clean 100 mL dry round Bottom Flask (RBF) was fitted w/ a stir bar. It was added, followed by toluene, then Et₃N. Then starch trap, was set up. Reaction, loaded into aluminum Beal bath ~100°C. After 24 hrs starch trap was set, washing with no solution placed in heating mantle @ 100°C. Al₂O₃. Heated to 200°C to collect H₂O. After 45 min @ 200°C, the temperature was lowered to 140°C since H₂O had been removed. The reaction was quenched after 12 hours at 140°C with EtOAc. Extracted with EtOAc, but H₂O was added due to salt crashing out upon organic addition to the aqueous phase.

Item Context

newsome2-128

Reaction (M4)
Table (75)

Remove Item
Add Missing Item
Process Item Context
Re-Run Object Detection
Done

Item Context

Reaction (M4)
Table (75)
News Item Type: Table Text Reaction

A clean 100 mL dry round Bottom Flask (RBF) was fitted w/ a stir bar. It was added, followed by toluene, then Et₃N. Then starch trap, was set up. Reaction, loaded into aluminum Beal bath ~100°C. After 24 hrs starch trap was set, washing with no solution placed in heating mantle @ 100°C. Al₂O₃. Heated to 200°C to collect H₂O. After 45 min @ 200°C, the temperature was lowered to 140°C since H₂O had been removed. The reaction was quenched after 12 hours at 140°C with EtOAc. Extracted with EtOAc, but H₂O was added due to salt crashing out upon organic addition to the aqueous phase.

Item Context

Reaction (M4)
Table (75)
News Item Type: Table Text Reaction

A clean 100 mL dry round Bottom Flask (RBF) was fitted w/ a stir bar. It was added, followed by toluene, then Et₃N. Then starch trap, was set up. Reaction, loaded into aluminum Beal bath ~100°C. After 24 hrs starch trap was set, washing with no solution placed in heating mantle @ 100°C. Al₂O₃. Heated to 200°C to collect H₂O. After 45 min @ 200°C, the temperature was lowered to 140°C since H₂O had been removed. The reaction was quenched after 12 hours at 140°C with EtOAc. Extracted with EtOAc, but H₂O was added due to salt crashing out upon organic addition to the aqueous phase.

Item Context

Manual Review – Refinement

- ▶ Analysis interface used to determine if automatically drawn bounding box is “perfect,” only slightly too large/small, or far too large/small
- ▶ Additional flag for noise/unrelated artifacts within bounding box

newsome2-074

Reaction 08-24
Table 175-10

Time	Temp	pH	Conductivity	Turbidity	DO	Notes
2:30	2:12	0.250				
2:35	2:10	0.175	1.1	0.00	0.00	0.00
2:40	2:05	0.150	1.1	0.00	0.00	0.00
2:45	2:00	0.125	1.1	0.00	0.00	0.00

A clean dry 100 mL RBF was fitted with a stirrer. 10 mL was added followed by 10 mL of 10% NaOH. The solution was stirred for 10 minutes. The solution was then cooled to 10°C. The reaction was monitored for 12 hours at 10°C with 30-min intervals. The reaction was quenched after 12 hours at 10°C with 30% HCl. Extracted with EtOAc, 30% HCl was added due to salt crystallizing out upon organic addition to the aqueous phase.

7/2/16

newsome2-128

ZrCl₄ / Benzoic Acid → U:068, MOF
110°C, 45 hours, DEF

ZrCl₄ / Benzoic Acid → U:068, MOF
110°C, 45 hours, DEF

Continue | Slightly Smaller | Very Small | Slightly Larger | Very Large | Access | Hide Process

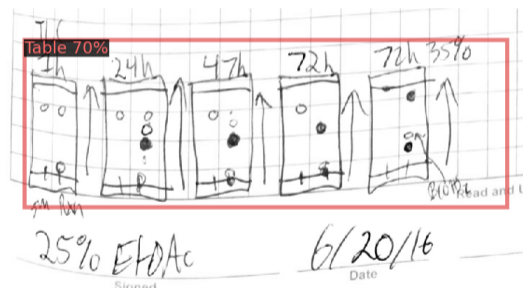
Results

- ▶ 154 pages for the testing set and manual review
- ▶ 78.8% of entries have accurate automated bounding boxes
- ▶ 15.6% of entries have nontrivial noise within their bounding boxes
- ▶ There are some experiment-specific diagrams that Detectron2 interpreted as tables
- ▶ Table style varied between the two authors
- ▶ Corrections in the notebooks very hard to automatically parse

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

Results

- ▶ 154 pages for the testing set and manual review
- ▶ 78.8% of entries have accurate automated bounding boxes
- ▶ 15.6% of entries have nontrivial noise within their bounding boxes
- ▶ There are some experiment-specific diagrams that Detectron2 interpreted as tables
- ▶ Table style varied between the two authors
- ▶ Corrections in the notebooks very hard to automatically parse



Results

- ▶ 154 pages for the testing set and manual review
- ▶ 78.8% of entries have accurate automated bounding boxes
- ▶ 15.6% of entries have nontrivial noise within their bounding boxes
- ▶ There are some experiment-specific diagrams that Detectron2 interpreted as tables
- ▶ Table style varied between the two authors
- ▶ Corrections in the notebooks very hard to automatically parse

EQ	FW	mmol	g	D	mL	Reagents
1.00	267.90	33.595	9.00	N/A	N/A	2,5-dibromohydroquinone ^{DAV}
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K ₂ CO ₃
////	////	////	////	////	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	////	////	////	////	135	DM F(.25M w/respect to DIBHG)



EQ,FW,MMOI,2,D,ML,Reagents

1,267.9,33.595,9,N/A,N/A,"2,5 - dibromohydroquinone DAV"

2,2,155.97,73.408,11.567,1.94,5.95,Ethyl Iodine

6,138.21,201.57,27.859,N/A,N/A,K2LO3

1/11,////,11/11,,111,1/11, Product

1,323.94,33.595,10.883,N/A,N/A,"1,4 - dibromo-2,5- ethoxy benzene"

11/1,11111,,1111,,135,DM F(.25M w/respect to DIBHG)

Results

- ▶ 154 pages for the testing set and manual review
- ▶ 78.8% of entries have accurate automated bounding boxes
- ▶ 15.6% of entries have nontrivial noise within their bounding boxes
- ▶ There are some experiment-specific diagrams that Detectron2 interpreted as tables
- ▶ Table style varied between the two authors
- ▶ Corrections in the notebooks very hard to automatically parse

Naphthalene Protection

Table No. 5734

Reagent: P-Toluene Sulfonic Acid, Glycol, Toluene

Time	Wt	mmol	g/mol	d	ml	Notes
1.0	279.13	5.097	1.422	-	50.97	Reaction
5.0	62.07	15.211	0.944	1.11	0.855	Glycol
0.07	190.22	0.357	0.668	-	-	p-Toluene Sulfonic Acid
1.0	279.13	5.097	1.422	-	50.97	Reaction
5.0	62.07	15.211	0.944	1.11	0.855	Glycol
0.07	190.22	0.357	0.668	-	-	p-Toluene Sulfonic Acid
1.0	279.13	5.097	1.422	-	50.97	Reaction

Notes:

- Glycol, Acid & Benzene were placed in a 100 ml RBF
- Reaction trap was fixed & a condenser was placed on top
- Solution heated to reflux for 18h (~200-250°C) (very hot)
- When rxn was almost done it was immediately quenched w/ NaHCO₃ to prevent reversal of rxn
- Rxn extracted 3x w/ EtOAc
- EtOAc washed w/ Brine & dried over Na₂SO₄
- Removed solvent in vac. to obtain brown cloudy oil
- Rxn a 0.5% EtOAc in Hex column
- Collected clear yellow oil
- Rxn H₁NMR in CDCl₃ ✓
- Yield = 950mg → 90% 91.0%

TLC for setup

5% EtOAc in Hex

8/11/16

Future Work

- ▶ Automated de-noising of entries
- ▶ Further investigate viability of chemical parsing tools
- ▶ Create vectorized/graph-based representation of entries
- ▶ Analyze the collection to answer scientific questions about experimental outcomes and pedagogy

Eq	FW	mmol	gram	d	ML	120°C 45min outcome
1.0	528.25	0.58	0.300	-	-	Reagent ①
3.5	179.97	1.988	0.358	-	-	Boric Acid
6.0	151.90	3.407	0.518	-	-	C ₅ F
0.025	816.69	6.043	0.035	-	-	Pd(dppf)Cl ₂
-	-	-	-	-	5.679	Dioxane 0.1M
1.0	638.72	0.568	0.363	-	-	⑥

Eq	FW	mmol	gram	d	ML	Reagent
1.0	528.25	0.58	0.300	-	-	①
3.5	179.97	1.988	0.358	-	-	Boric Acid
6.0	151.90	3.407	0.518	-	-	C ₅ F
0.025	816.69	6.043	0.035	-	-	Pd(dppf)Cl ₂
-	-	-	-	-	5.679	Dioxane 0.1M
1.0	638.72	0.568	0.363	-	-	⑥

Conclusions

- ▶ Overall goals:
 1. Make the information contained in analog lab notebooks AI-ready
 2. which will facilitate the answering of scientific questions.
- ▶ To date we have:
 1. Developed a process to extract contents of scanned lab notebook pages
 2. analyzed the results
 3. presented potential challenges with data quality and archiving. This initial research effort helps
- ▶ Next phase of work is developing ML compatible representation of data

Notebook No. _____ 87
Continued from Page _____

PROJECT DAV-1-87 10/11/2015

Reaction 658

Reaction 658

EA	FW	mmol	g	ml	D	Reagents
1	126.11	7.92%	1.00			"PG"
2,2	110.19	17.995	2.45			"HMIA"
xs				6.6		TFA (1.2M)
xs					11	3M 2 nd step
xs					132	3M HCl
1.0	210.14	1.00	147g			TFFG

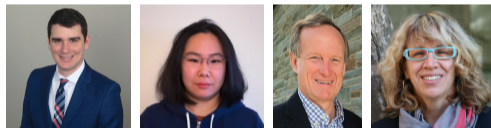
- 1) Open dried flask and pour out contents and then dried and it was 100% dry.
- 2) Add powder under positive N₂ and then put red septum and added the 6.6 mL of TFA.
- 3) Started at 7:40 pm and then waited 2.5 - 3 hrs.
- 25) After the TFA added the reflux condenser and put it under positive N₂ and continued step 3.
- 4) Added 3M HCl 14 mL at 10:50 pm.
- 5) Originally, it was a yellowish solution.

Continued on Page _____

Read and Understood By _____

Signed _____ Date _____ Signed _____ Date _____

Questions?



Drexel MRC: Joel Pepper (PhD student, jcp353@drexel.edu), Xintong Zhao (PhD student), David Breen (Professor, david@cs.drexel.edu), Jane Greenberg (Professor, jg3243@drexel.edu)

This work supported by the National Science Foundation "Institute for Data Driven Dynamical Design" under Grant No. OAC-2118201.



UNIVERSITY OF
CENTRAL FLORIDA



Elizabeth Jones (Summer REU, Northeastern University), Jacob Furst (PhD Student, U of Central Florida), Kyle Langlois (Student, U of Central Florida), Fernando Uribe-Romo (Professor, U of Central Florida)