AI-Ready Data: Knowledge Extraction from Archival Lab Notebooks

IEEE Big Data 2024, CAS Workshop

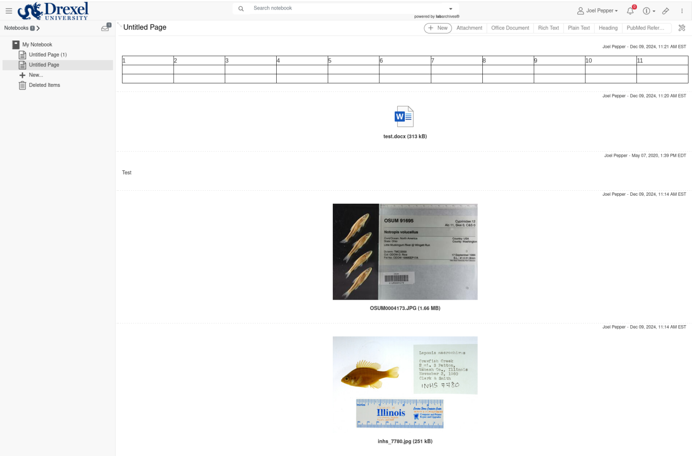
Joel Pepper[1], Elizabeth Jones[4], Xintong Zhao[2], Jacob Furst[3], Kyle Langlois[3], Fernando Uribe-Romo[3], David Breen[1], Jane Greenberg[2]

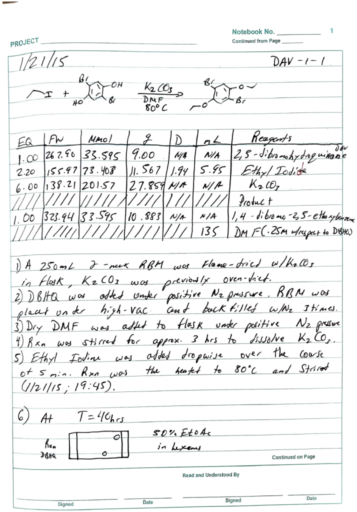
[1]Drexel University, Department of Computer Science
[2]Drexel University, Department of Information Science
[3]University of Central Florida, Department of Chemistry
[4]Northeastern University, Department of Computer Science

December 17, 2024

Drexel University Metadata Research Center College of Computing & Informatics

Background




Notebook No. 1
Continued from Page
PROJECT

1/21/15 DAV-1-1

K_2CO_3 / DMF / 80°C

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_2CO_3
////	////	////	////	///	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	////	////	////	///	135	DMF (.25M w/respect to DBHQ)

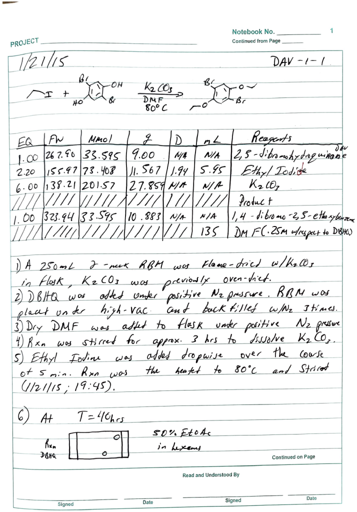
1) A 250mL 2-neck RBM was flame-dried w/ K_2CO_3 in flask, K_2CO_3 was previously oven-dried.
2) DBHQ was added under positive N_2 pressure. RBM was placed under high-vac and back filled w/ N_2 3 times.
3) Dry DMF was added to flask under positive N_2 pressure
4) Rxn was stirred for approx. 3 hrs to dissolve K_2CO_3.
5) Ethyl Iodine 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
Rxn DBHQ
50% EtOAc in hexanes

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. 1

Continued from Page

1/21/15 DAV-1-1

K_2CO_3 / DMF / 80° C

EQ	FW	MMol	g	D	mL	Reagents
1.00	267.90	33.595	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_2CO_3
////	////	////	////	///	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	/////	/////	/////	///	135	DMF (.25M w/respect to DBHQ)

1) A 250 mL 2-neck RBM was flame-dried w/ K_2CO_3 in flask, K_2CO_3 was previously oven-dried.
2) DBHQ was added under positive N_2 pressure. RBM was placed under high-vac and back filled w/ N_2 3 times.
3) Dry DMF was added to flask under positive N_2 pressure
4) Rxn was stirred for approx. 3 hrs to dissolve K_2CO_3.
5) Ethyl Iodine was added dropwise over the course of 5 min. Rxn was the heated to 80°C and stirred (1/21/15; 19:45).

6) At T = 40 hrs

Rxn DBHQ 50% EtOAc in hexanes

Continued on Page

Read and Understood By

Signed Date Signed Date

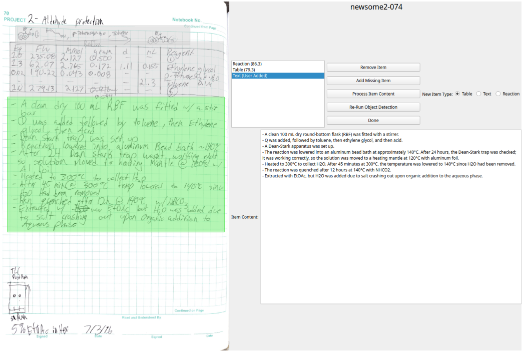
Introduction

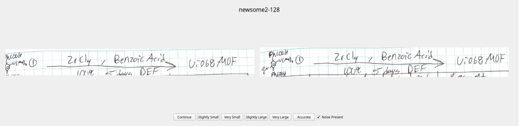
- Paper-based lab notebooks becoming "data at risk" [3, 4]
- Collections of notebooks may have the potential to provide new insights into the successes, failures and pedagogy of research labs
- Research is needed to address challenge of converting analog lab notebooks into computationally ready resources

[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





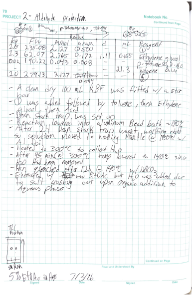
[a] Sourced from U of Central Florida Reticular Synthesis and Materials Design Lab (RSMDL)

[b] Main focus of this talk

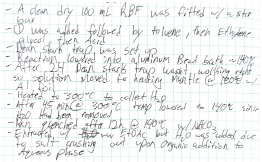
Methodology Overview

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



Object Detection →

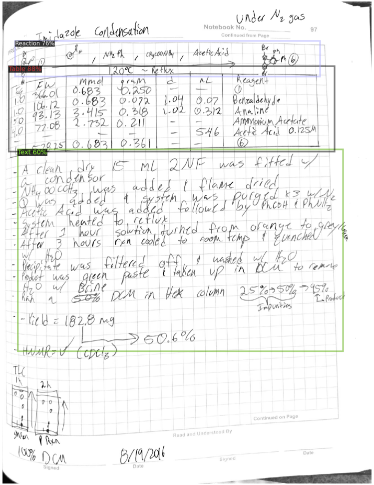


Entry Specific Processing | *Optical Character Recognition* ↓

```
- 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 NHCO2.
- Extracted with EtOAc, but H2O was added due to salt crashing out upon
  organic addition to the aqueous phase.
```

Segmentation

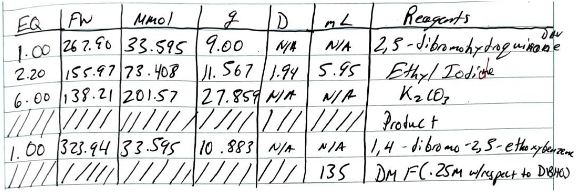
- Make use of the Detectron2 object detection platform [14]
- Three entry types in model: text, tables and chemical reactions



[14]: Wu, Detectron2

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



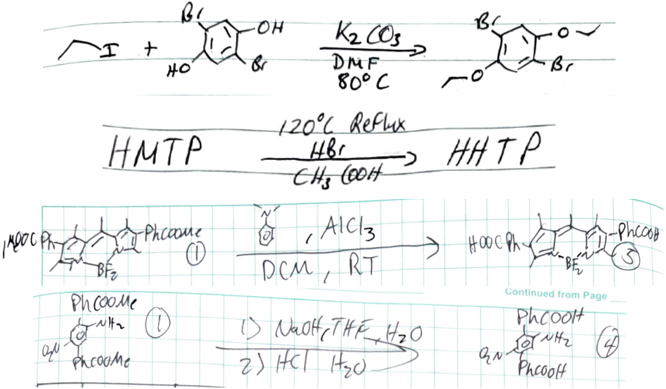
EQ	FW	MMol	g	D	mL	Reagents
1.00	267.90	33.595	9.00	N/A	N/A	2,5-dibromohydroquinone DAV
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K_2CO_3
////	////	/////	////	///	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	////	/////	////	///	135	DMF (.25M w/respect to DBHQ)

EQ,FW,MMOl,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/

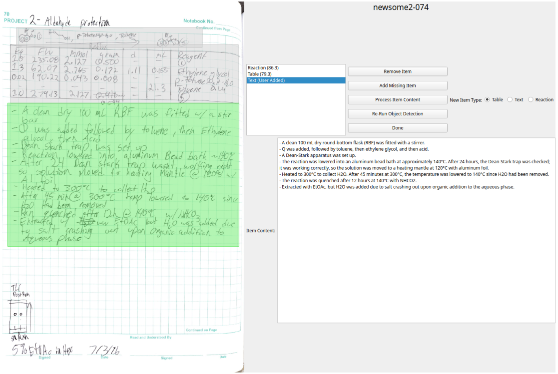
Content Extraction – Reactions

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



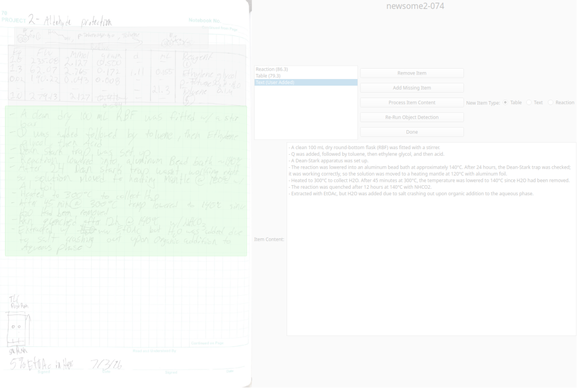
Manual Review – Analysis

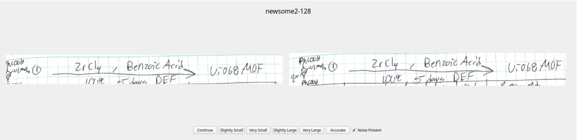
- Two interfaces for assessing and improving automated segmentation accuracy
- Refinement interface used to redraw, remove and add bounding boxes



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





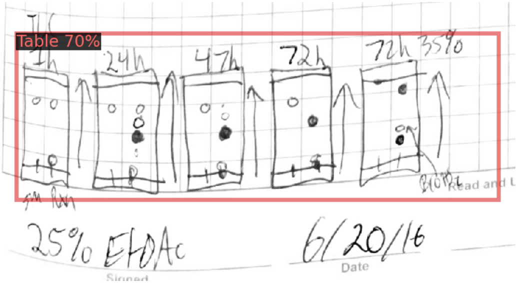
Results

- 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

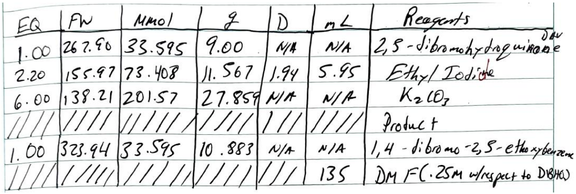
Results

- 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

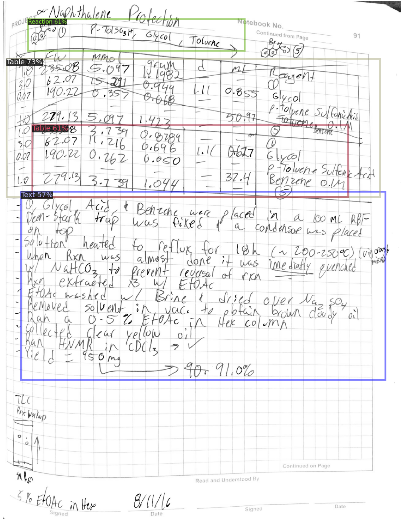


EQ	FW	Mmol	g	D	mL	Reagents
1.00	267.90	33.595	9.00	N/A	N/A	2,5-dibromohydroquinone DAV
2.20	155.97	73.408	11.567	1.94	5.95	Ethyl Iodide
6.00	138.21	201.57	27.859	N/A	N/A	K_2CO_3
////	////	/////	////	///	////	Product
1.00	323.94	33.595	10.883	N/A	N/A	1,4-dibromo-2,5-ethoxybenzene
////	/////	/////	////	///	135	DMF(.25M w/respect to DBHQ)

EQ,FW,MMOl,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)

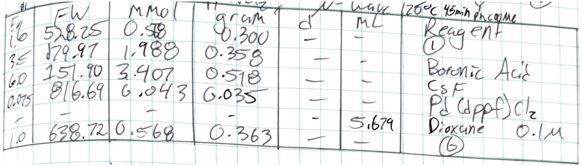
Results

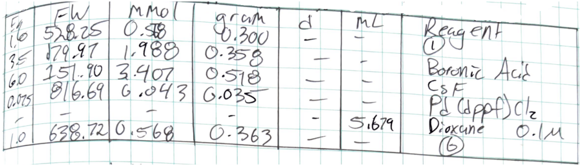
- 154 pages for the testing set and manual review
- 78.8% of entries have accurate automated bounding boxes
- 15.6% of entries have nontrivial noise within their bounding boxes
- There are some experiment-specific diagrams that Detectron2 interpreted as tables
- Table style varied between the two authors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



Future Work

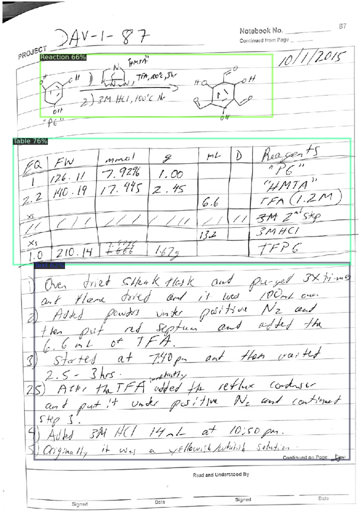
- Automated de-noising of entries
- Further investigate viability of chemical parsing tools
- Create vectorized/graph-based representation of entries
- Analyze the collection to answer scientific questions about experimental outcomes and pedagogy





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


PROJECT DAV-1-87 Notebook No. 87

Continued from Page

10/1/2015

1) HMTA, TFA, 100°C, 3hr; 2) 3M HCl, 100°C 1hr

"PG"

EQ	FW	mmol	g	mL	D	Reagents
1	126.11	7.9296	1.00			"PG"
2.2	140.19	17.445	2.45			"HMTA"
xs				6.6		TFA (1.2M)
//	///	///	///	//	//	3M 2nd step
xs				13.2		3M HCl
1.0	210.14	[illegible]	1.67g			TFPG

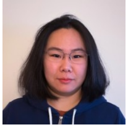
1) Oven dried Schlenk flask and purged 3X times and flame dried and it was 100mL over.
2) Added powders under positive N_2 and then put red septum and added the 6.6 mL of TFA.
3) Started at 7:40 pm and then waited 2.5 - 3 hrs.
2.5) After the TFA immediately added the reflux condenser and put it under positive N_2 and continued step 3.
4) Added 3M HCl 14 mL at 10:50 pm.
5) Originally it was a yellowish solution.

Read and Understood By

Signed Date Signed Date

Questions?

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

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)

UCF UNIVERSITY OF CENTRAL FLORIDA

Northeastern University

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)

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