Making the intangible tangible: From lab notebook to digital insight

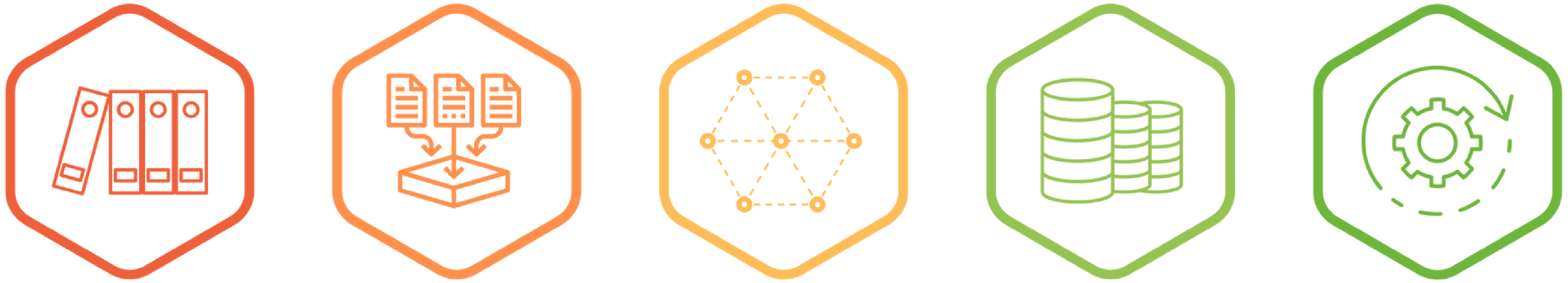
About Me

Senior Research Assistant at the University of Southampton

Background in developing predictive anylytics for chemical systems

Co-director of Data Revival

Data Revival



A quick journey through history

Prior Research

- Digital chemistry
- Electronic lab notebooks
- Research Data Management

Always fixing the future?

- Digitalisation has to be done correctly
- The ELNs don’t solve all our problems
- How much data exists on things that didn’t work?

Polymer informatics

- Work to build a predictive analytics platform
- Very little digital data
- Goldmine of knowledge buried in lab notebooks

Where did we start?

- 3000/4000 Chemistry Lab Notebooks
- 2000 chemist years worth of work

Practical

- Free up cupboard space
- Remove fire/flood risk
- Combine all the records together

Compliance

- H&S
- Longevity of records
- Traceability

Productivity

- Integration
- Data mining
- AI/ML

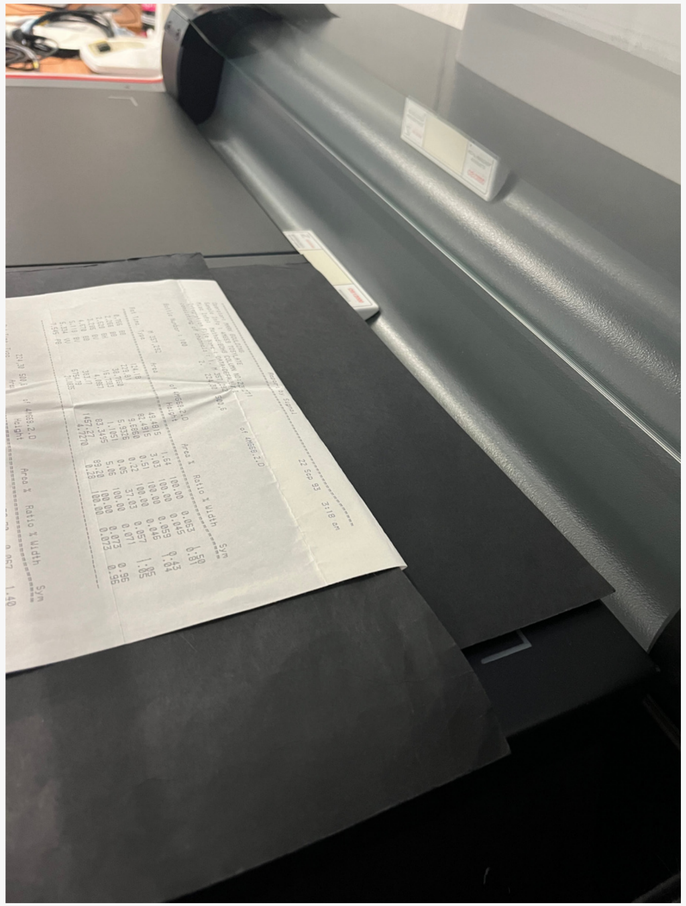
Why scan?

Professional scanning equipment

- Scanning takes a LONG time

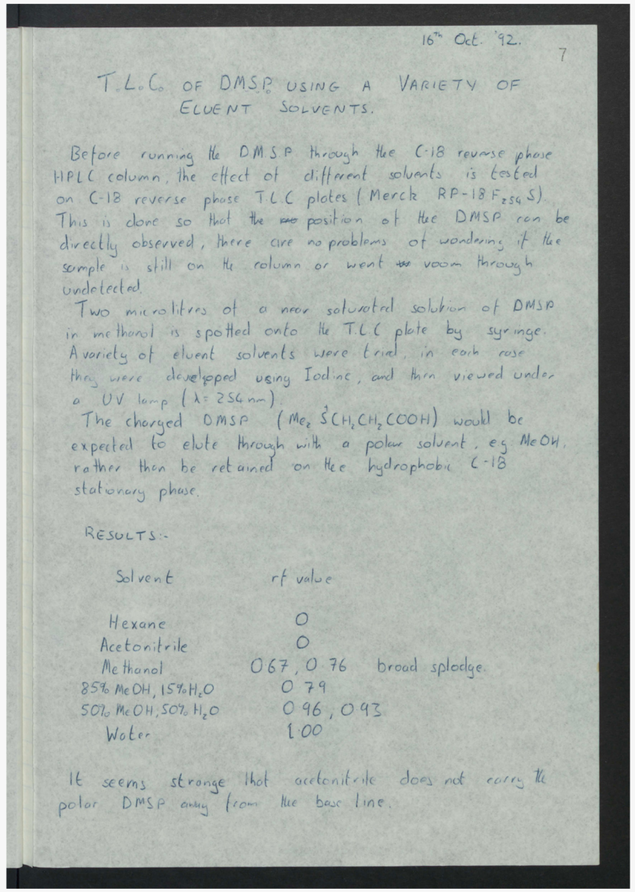
The difficulty even with scanners

- The process



Examples of scans

- Scans using professional scanning equipment



16th Oct. '92. 7

T.L.C. OF DMSP USING A VARIETY OF ELUENT SOLVENTS.

Before running the DMSP through the C-18 reverse phase HPLC column, the effect of different solvents is tested on C-18 reverse phase T.L.C plates (Merck RP-18 $F_{254}S$). This is done so that the position of the DMSP can be directly observed, there are no problems of wondering if the sample is still on the column or went through undetected.

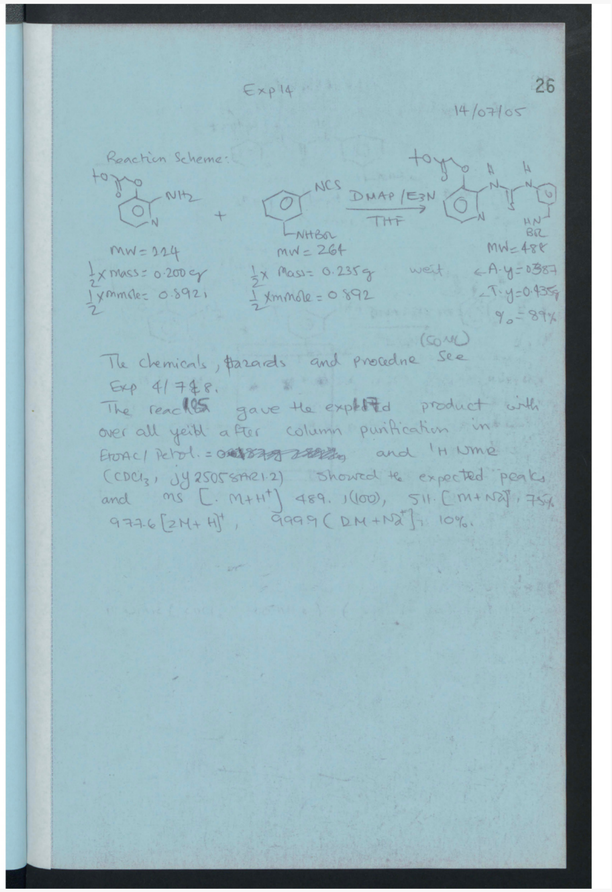
Two microlitres of a near saturated solution of DMSP in methanol is spotted onto the T.L.C plate by syringe. A variety of eluent solvents were tried, in each case they were developped using Iodine, and then viewed under a UV lamp (λ = 254 nm).

The charged DMSP ($Me_2S^+CH_2CH_2COOH$) would be expected to elute through with a polar solvent, e.g. MeOH, rather than be retained on the hydrophobic C-18 stationary phase.

RESULTS:-

Solvent	rf value
Hexane	0
Acetonitrile	0
Methanol	0.67, 0.76 broad spledge
85% MeOH, 15% H_2O	0.79
50% MeOH, 50% H_2O	0.96, 0.93
Water	1.00

It seems strange that acetonitrile does not carry the polar DMSP away from the base line.



Exp 14 26

14/07/05

Reaction Scheme:

MW = 324 MW = 264 MW = 488

½ x mass = 0.200 gr ½ x Mass = 0.235 g west ← A: y = 0.357

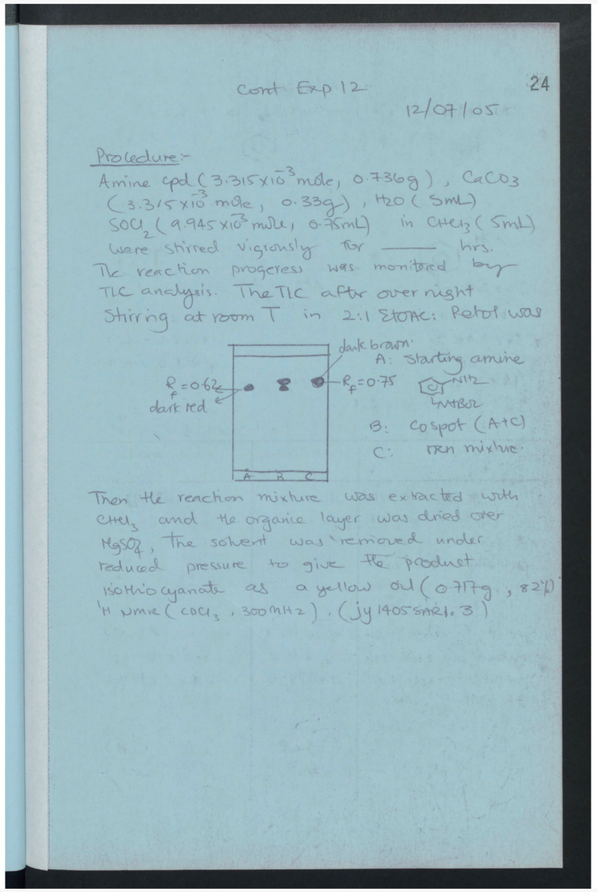
½ x mmole = 0.892 ½ x mmole = 0.892 T: y = 0.435 g

% = 89%

(50 ml)

The chemicals, hazards and procedure See Exp 417 & 8.

The reaction gave the expected product with over all yeild after column purification in EtOAc/Petrol. = 0 and ^{1}H NMR (CDCl$_3$, jy2505snh12) showed the expected peaks and MS [. M+H$^+$] 489. 1(100), 511 [. M+Na$^+$], 75% 977.6 [2M+H]$^+$, 999.9 (DM+Na$^+$], 10%.



Cont Exp 12 24

12/07/05

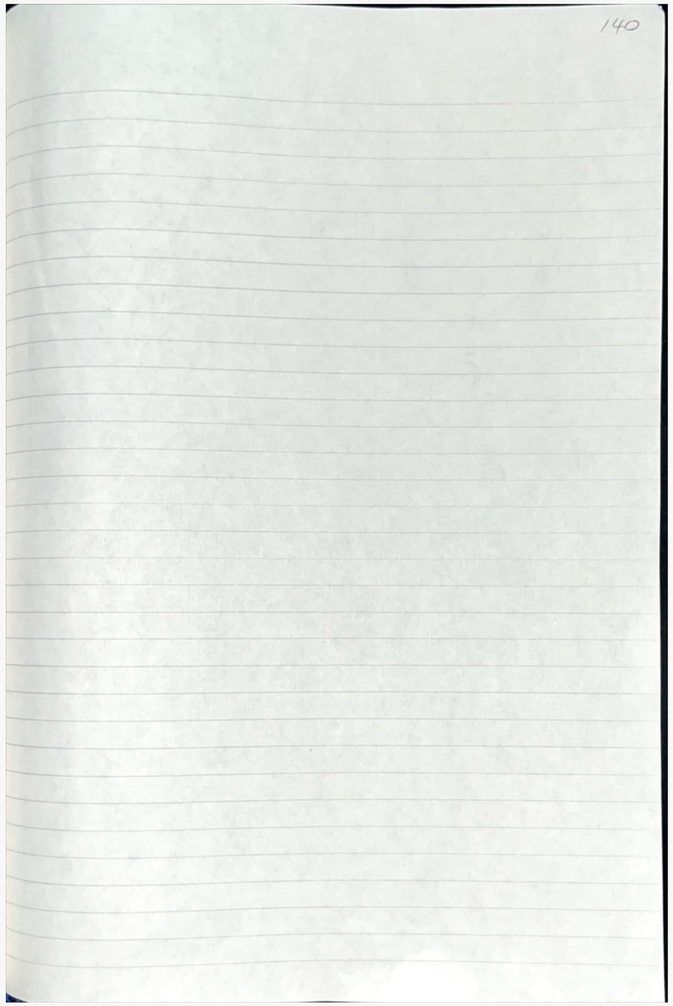
Procedure:-

Amine cpd (3.315×10^{-3} mole, 0.736g), $CaCO_3$ (3.315×10^{-3} mole, 0.33g), H_2O (5mL) $SOCl_2$ (9.945×10^{-3} mole, 0.75mL) in CH_2Cl_2 (5mL) were stirred vigiously for ____ hrs. Tlc reaction progress was monitored by TLC analysis. The TLC after over night stiring at room T in 2:1 EtOAc: Petrol was

Then the reaction mixture was extracted with CH_2Cl_2 and the organic layer was dried over $MgSO_4$, the solvent was removed under reduced pressure to give the product isothiocyanate as a yellow oil (0.717g, 82%) ^{1}H NMR (CDCl$_3$, 300MHz), (jy1405snh1.3)

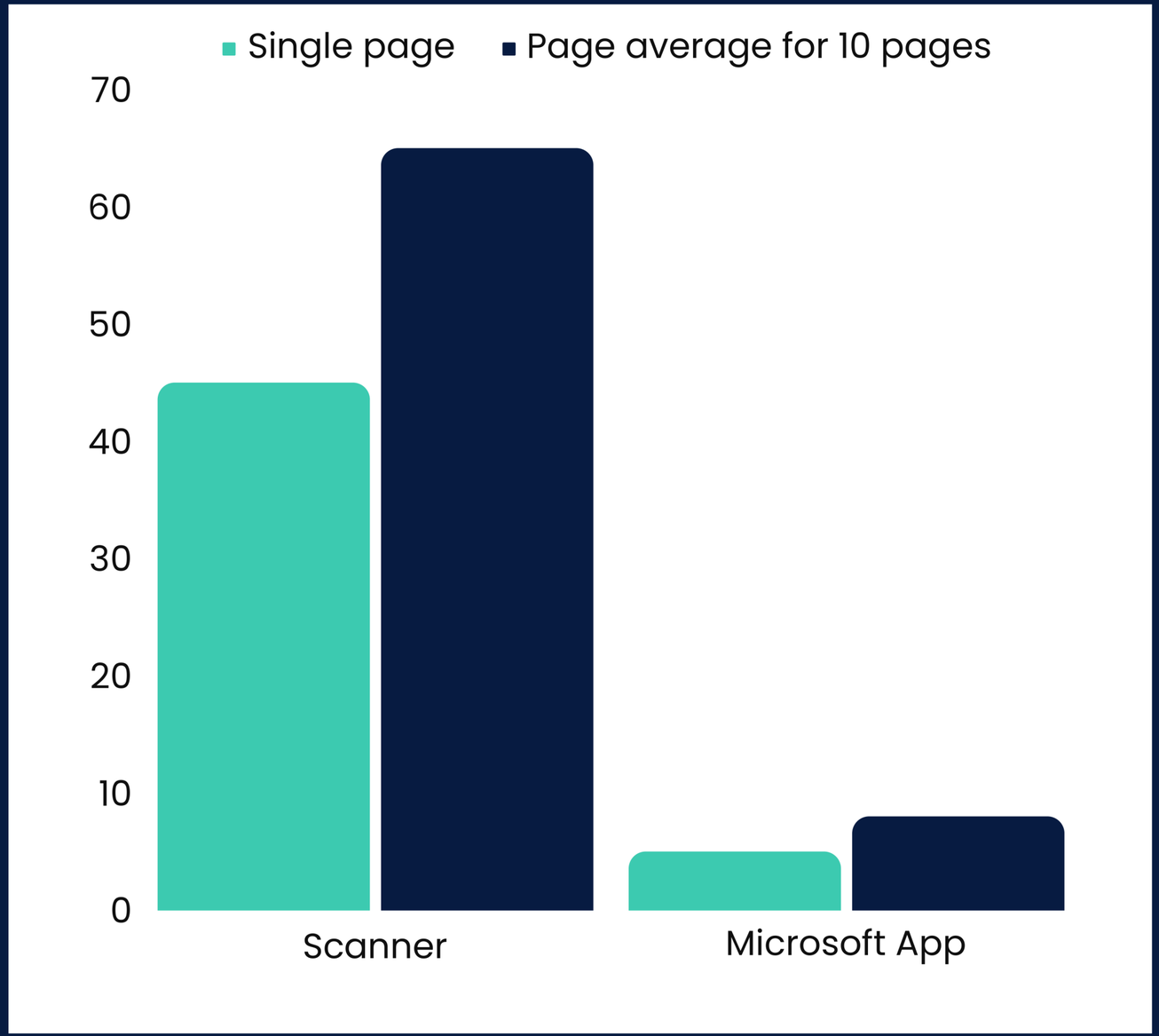
Moving away from real scanners

- Microsoft scanning app



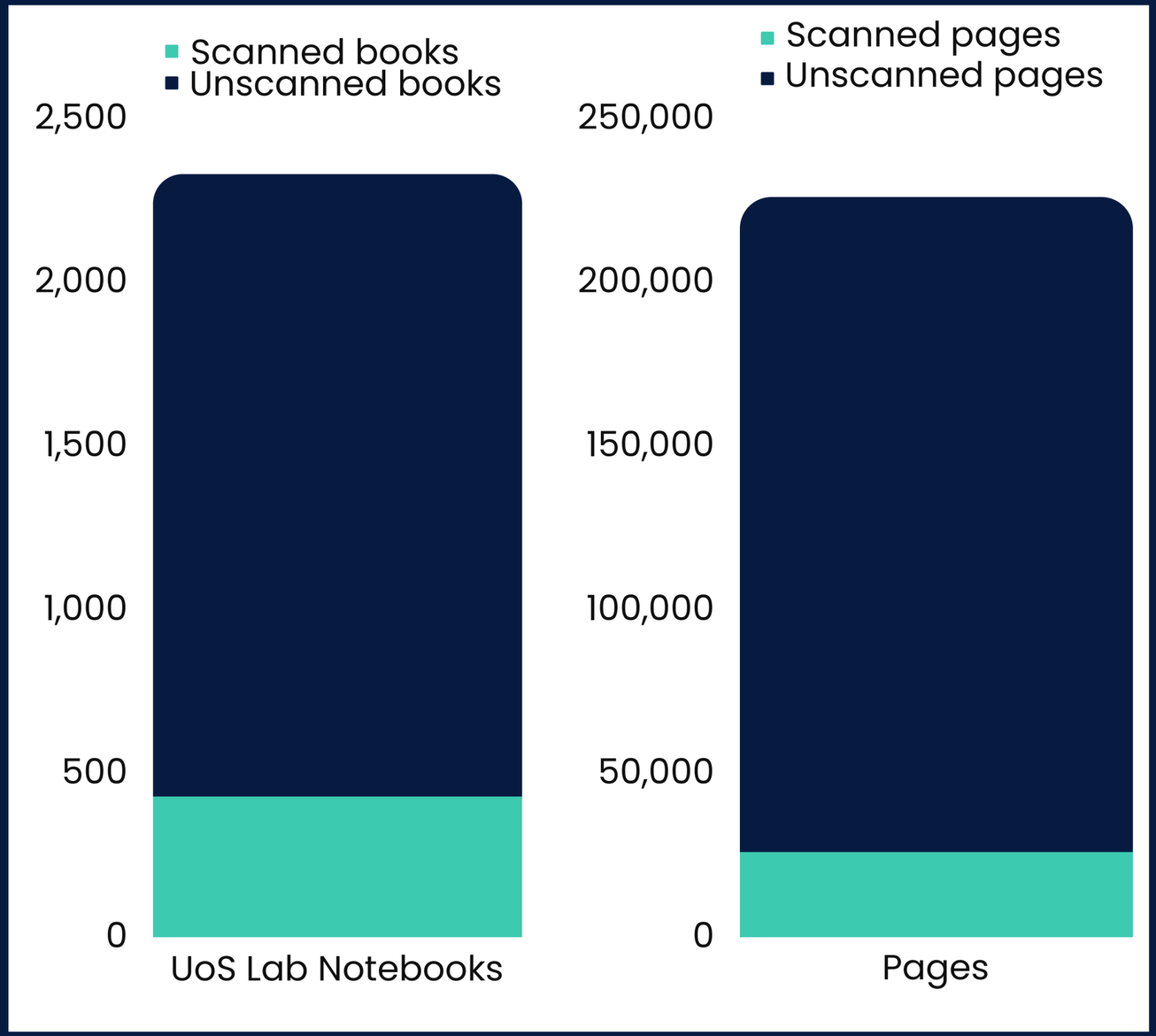
Time considerations

Using the app is much much faster than the scanner



How much has been scanned

Have I mentioned scanning is a long process?



What have we learnt from all of this?

Our understanding of what is needed is evolving

Images

- The quality doesn’t need to be as high as you think
- Straight and flat is better, but not essential
- Bleed through is a killer
- A LOT of lab books have barely anything in

Storage

- Large long term volume storage essential
- Smaller, fast access storage is required if you want to do ML
- Integration with your tech stack is essential

Destructive or not?

- The human and machine readability fo the scans is high enough to warrant destroying paper versions
- However, destructive scanning may not bring huge speeds increases with high volumes of inserts

Proper meta data records when scanning

- Standardise naming conventions
- Meta-data for book identifier and page number

Define label classes

- What is important to you?
- What isn't important to you?
- What are the downstream tasks?
- Are you integrating with databases/ELNs?
- Are you combining datasets?

Modelling

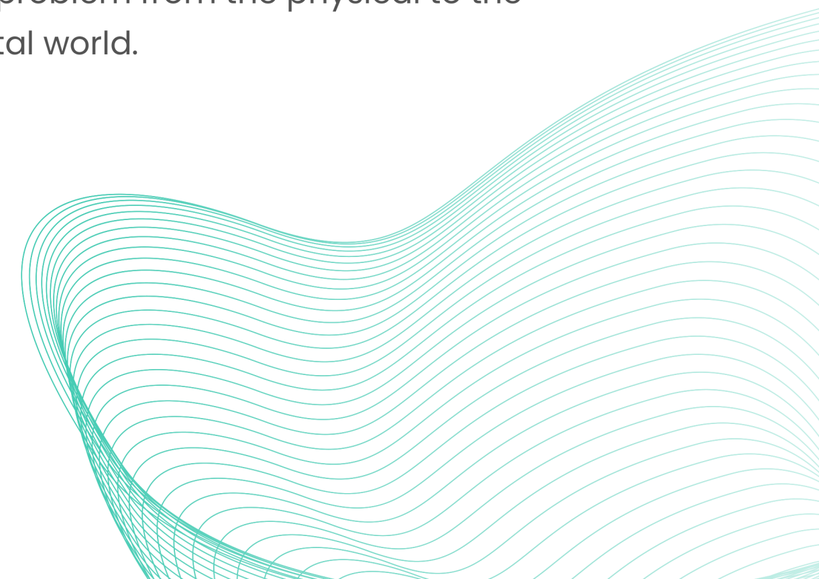
- Various models for various tasks
- Segmentation
- OCR
- OCSR
- Data structuring

Connection with other resources

- Does the strcuture need to be consistant with downstream sources?
- Are we combining data across recording mediums?
- Are databases being connected to or built?

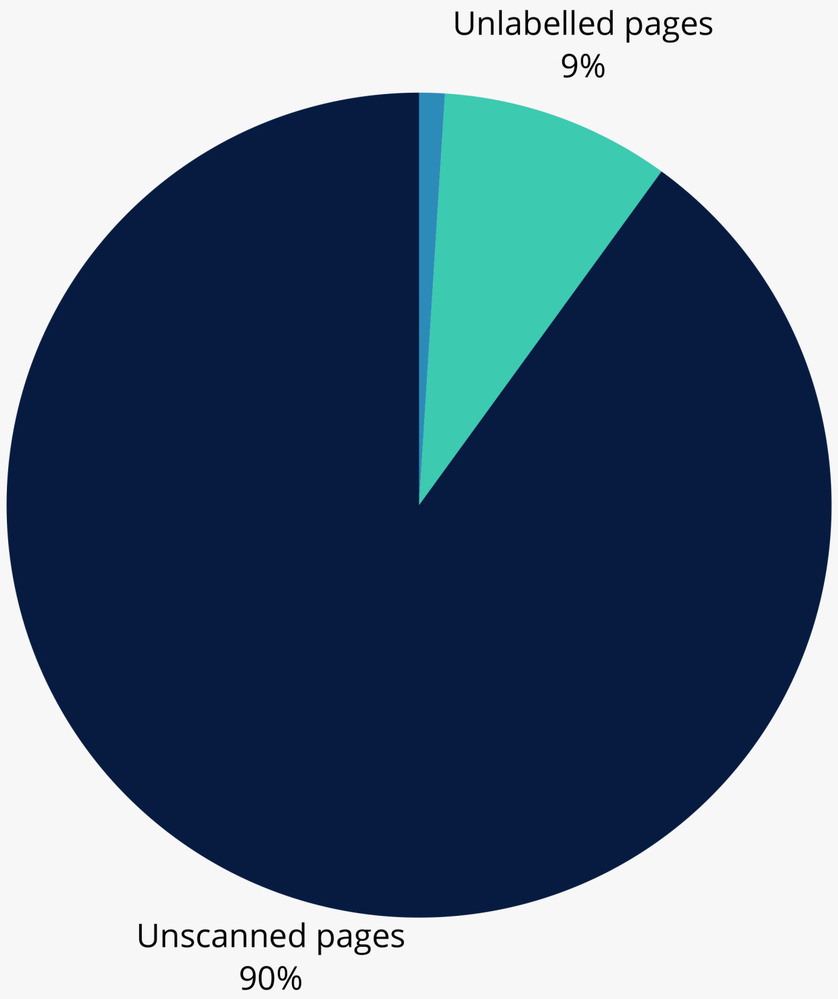
Scanning isn't enough

Simply digitising the material just moves the problem from the physical to the digital world.



UoS lab notebook labelling

Label Type
Date
Diagram
Equation
Graph
Molecule
Page Number
Table
Text

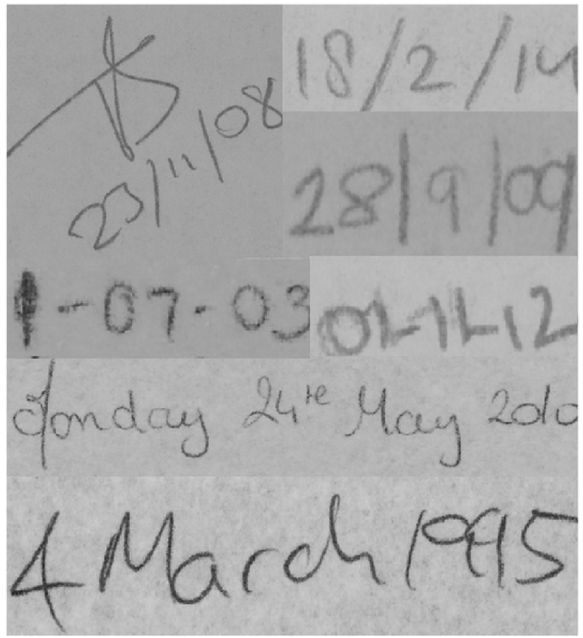


Segmentation

The results from 430 books

Label Type	Count
Date	30,541

17,217 pages containing dates

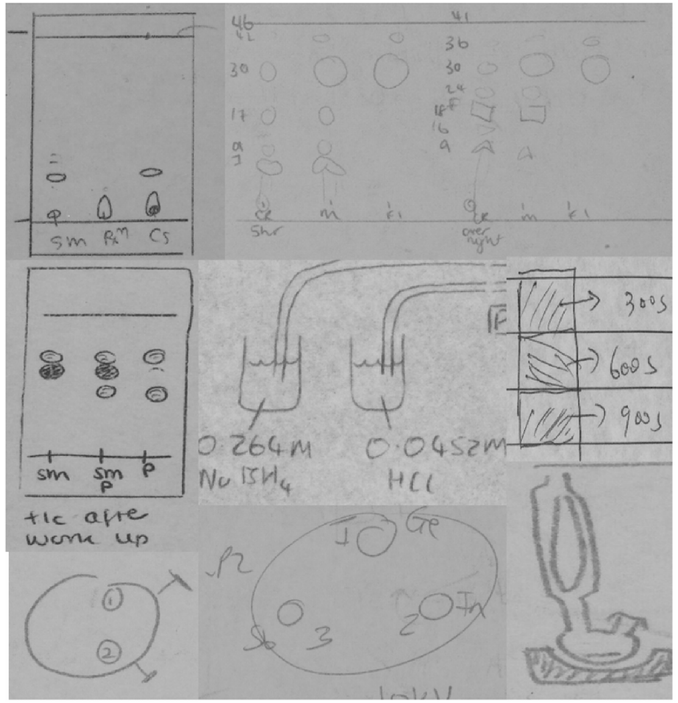


Segmentation

The results from 430 books

Label Type	Count
Diagram	7348

3971 pages containing diagrams

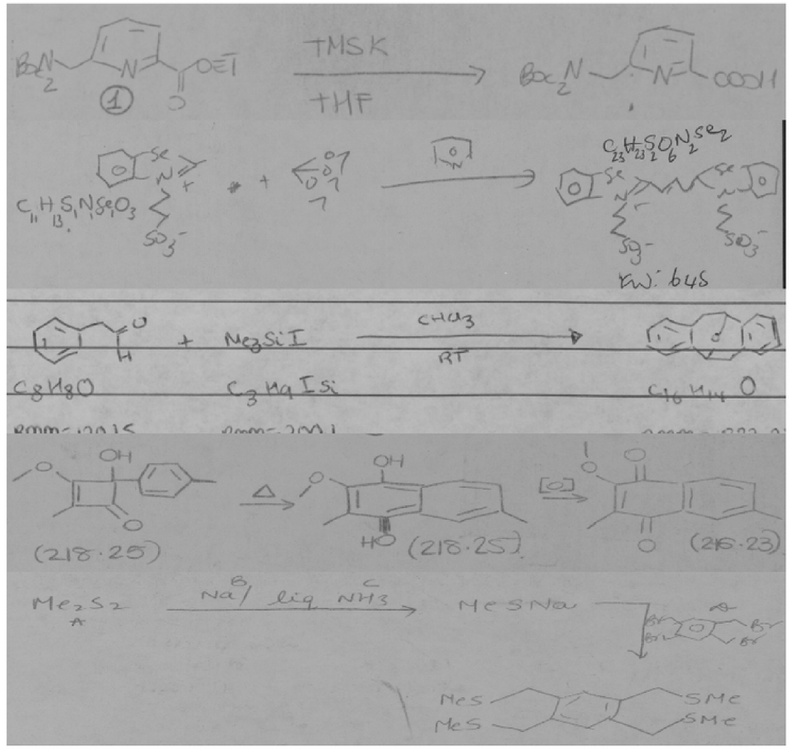


Segmentation

The results from 430 books

Label Type	Count
Equation	6714

5,617 pages containing equations

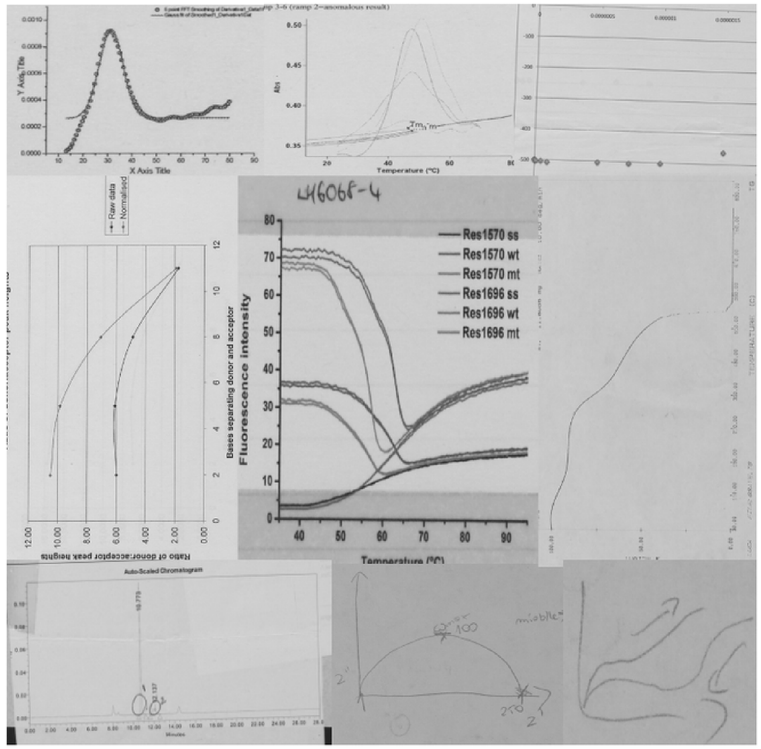


Segmentation

The results from 430 books

Label Type	Count
Graph	2755

1449 pages containing graphs

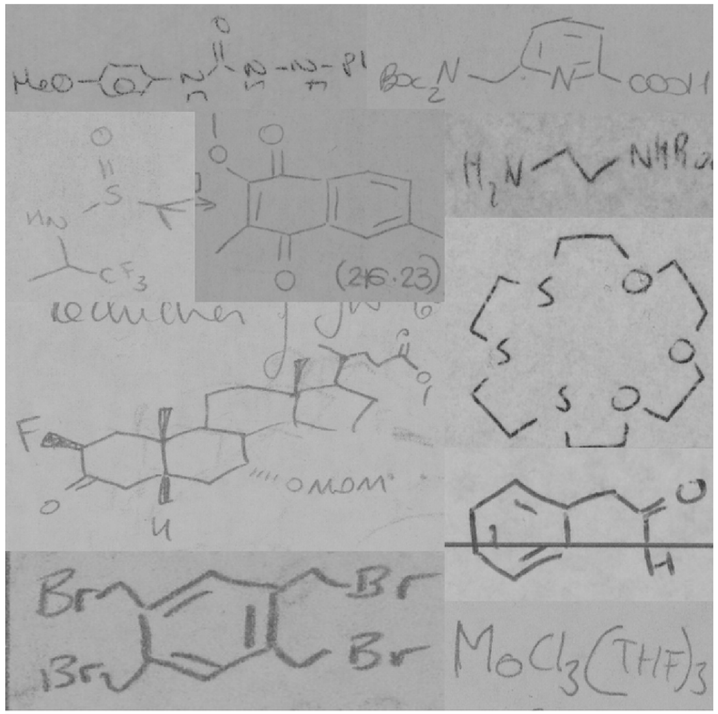


Segmentation

The results from 430 books

Label Type	Count
Molecule	17387

6759 pages containing molecules

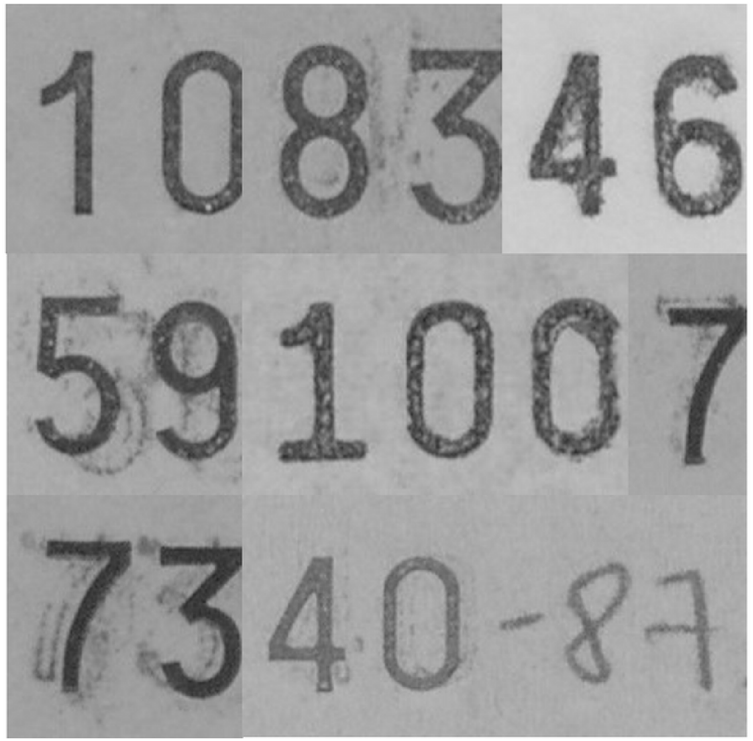


Segmentation

The results from 430 books

Label Type	Count
Page number	24698

22545 pages containing page numbers

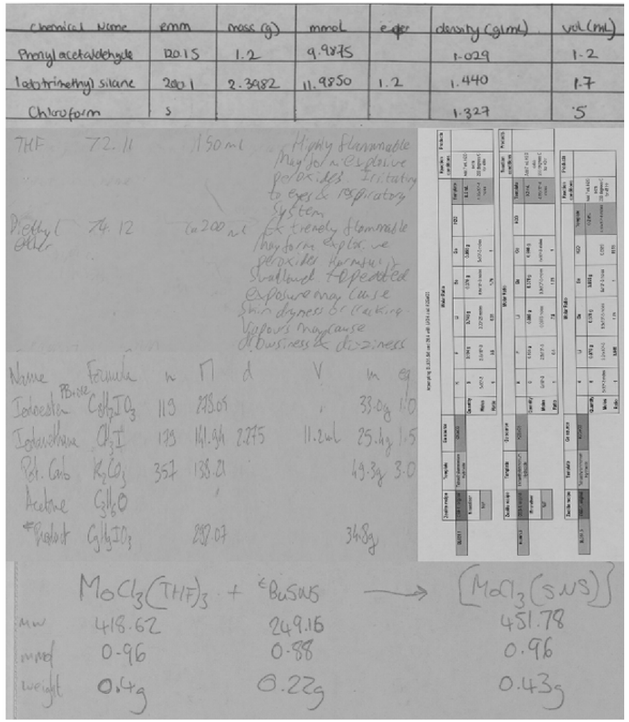


Segmentation

The results from 430 books

Label Type	Count
Table	19623

13601 pages containing tables

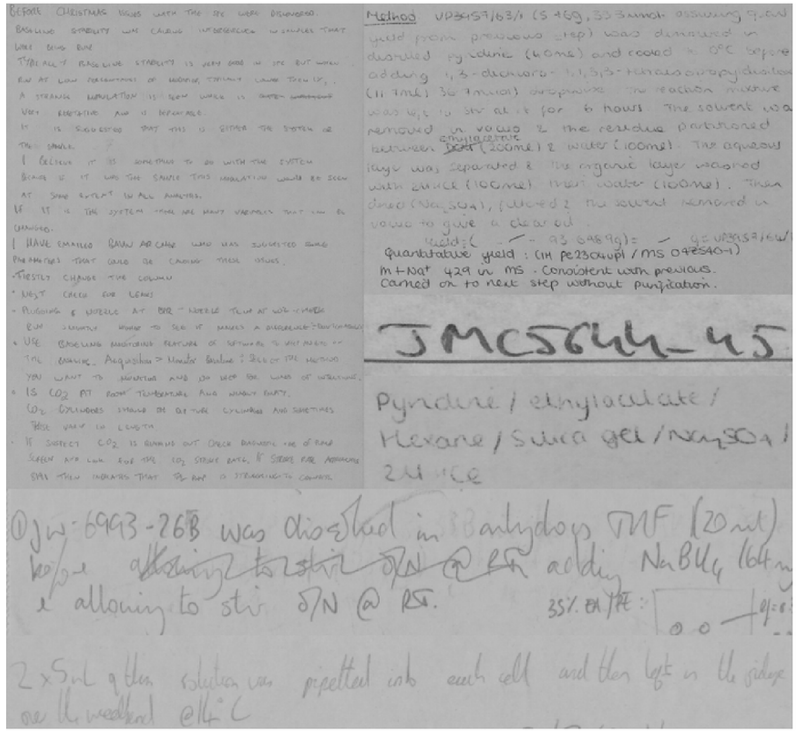


Segmentation

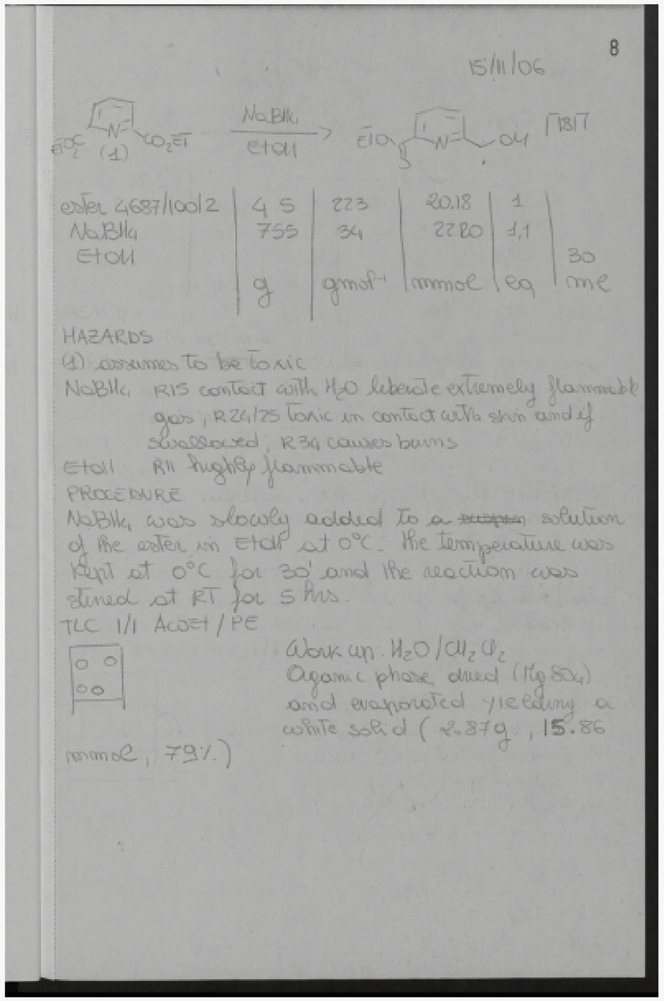
The results from 430 books

Label Type	Count
Text	119752

24066 pages containing text



Segementation overview



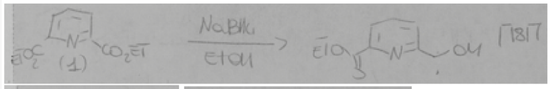
Page Number

8

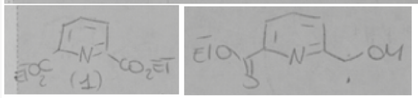
Date

15/11/06

Equation



Molecules



Table

ester 4687/1001/2 NaBH4 EtOH	4.5 755 g	223 34 gmol⁻¹	20.18 2280 mmol	1 1.1 eq	30 ml

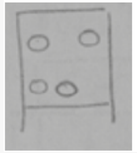
Text

HAZARDS
(1) assumes to be toxic
NaBH4 R15 contact with H2O liberate extremely flammable gas; R24/25 Toxic in contact with skin and if swallowed; R34 causes burns
EtOH R11 highly flammable
PROCEDURE
NaBH4 was slowly added to a ~~suspension~~ solution of the ester in EtOH at 0°C. The temperature was kept at 0°C for 30' and the reaction was stirred at RT for 5 hrs.
TLC 1/1 AcOEt/PE

TLC 1/1 AcOEt/PE

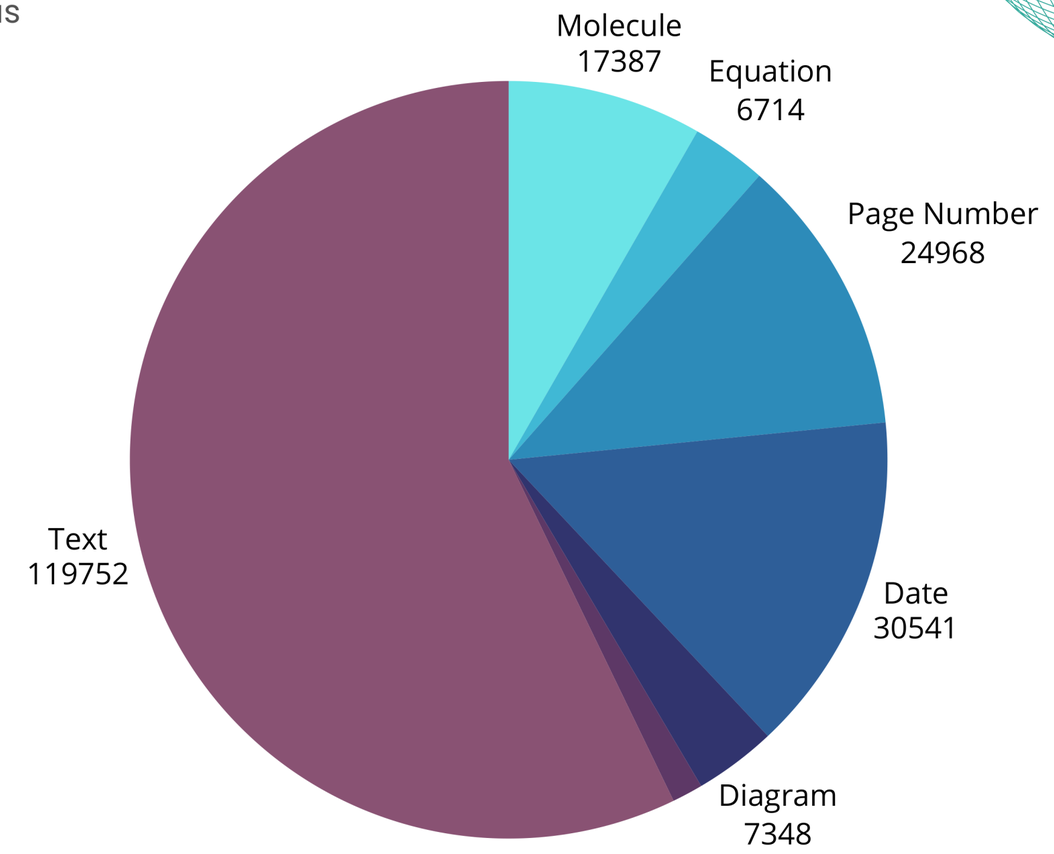
Work up: H2O/CH2Cl2
Organic phase dried (MgSO4) and evaporated yielding a white solid (2.87g, 15.86 mmol, 79%)

Diagram



UoS Lab Notebooks

- 430 lab notebooks scanned
- 25,715 pages
- 230,000 segmented datas



OCR

- Conversion of handwritten text into machine readable tokens

OCSR

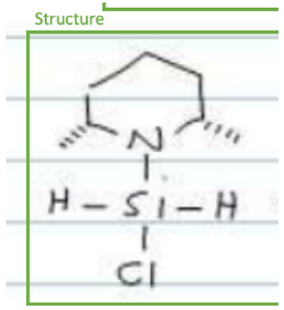
- Conversion of hand drawn chemical structures into SMILES/INCHI

Integration

- Structuring and integration of extracted data

Now for the interesting part

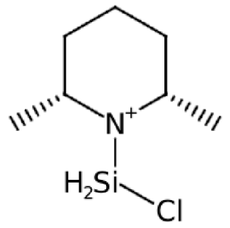
Results - Handwritten documents

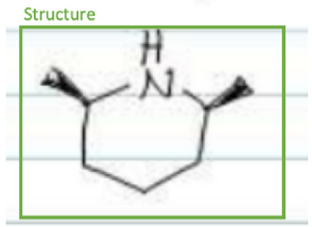


DR OCSR

C[C@@H]1CCC[C@H](C)[N+]1[SiH2]Cl

CDK depiction

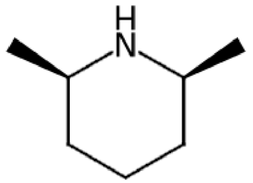


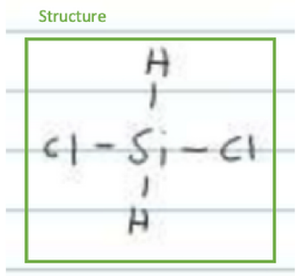


DR OCSR

C[C@@H]1CCC[C@H](C)N1

CDK depiction

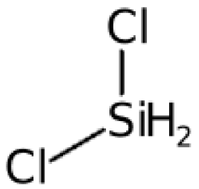




DR OCSR

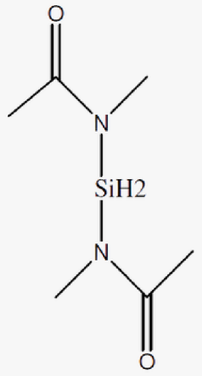
Cl[SiH2]Cl

CDK depiction



Results - Digital documents

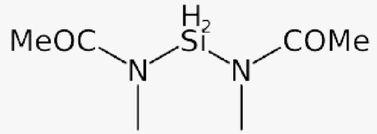
We have a very high conversion rate for digitally created structures.

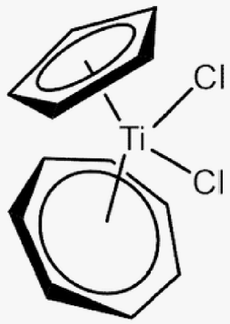


DR OCSR

CC(=O)N(C)[SiH2]N(C)C(C)=O

CDK depiction

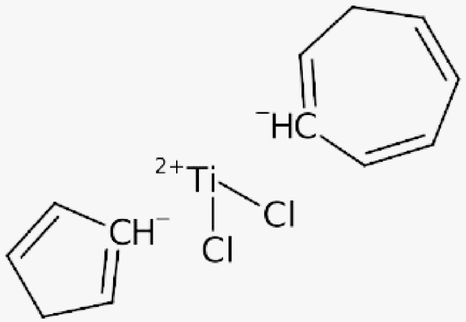




DR OCSR

C1=[CH-]C=CC1.[Ti+2](Cl)Cl.C1=[CH-]C=CC=CC1

CDK depiction



Structuring the data

Our preferred output is JSON


```
{
  "uniqueId": "xx",
  "fileName": "page_1.png",
  "Type": "image/png",
  "dataset": "nist-cmc",
  "confidenceScore": "High",
  "formFields": [
    {
      "fieldName": "additive",
      "value": "AG NO3 DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DIOXANE DEUTERIUM OXIDE ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL ETHANOL",
      "ocrConfidence": 0.9812571428571428
    },
    {
      "fieldName": "cmc-value",
      "value": "8.67 X10-3 M 8.81 X10-3 W 8.55 X10-3 M 8.51 X10-3 W 8.43 X10-3 M 8.47 X10-3 W 8.25 X10-3 M 2.330X10-1 D 8.081X10-3 M 2.340X10-1 D 8.116X10-3 M 2.36 X10-1 D 8.185×10-3 M 8.39 X10-3 W 8.1 ×10-3 M 2.324X10-1 D 8.061X10-3 M 8.3 X10-3 M 8.16 X10-3 M 8.27 X10-3 W 8.27 X10-3 M 8.15 X10-3 M 8.2 ×10-3 M 8.4 ×10-3 M 8.44 X10-3 W 8.23 X10-3 M 8.57 X10-3 W 8.4 X10-3 M 8.39 X10-3 M 8.88 X10-3 W 8.60 X10-3 M 9.10 X10-3 W 8.86 X10-3 M 9.61 X10-3 W 9.18 X10-3 M 9.95 X10-3 W 9.8 ×10-3 M 9.49 X10-3 M 9.61 X10-3 M 1.016×10-2 M 1.091×10-2 M 1.14 X10-2 M 5.0 ×10-3 M 6.73 X10-3 M 7.31 X10-3 M 9.03 X10-3 M 1.38 X10-2 M 2.10 X10-2 M 8.01 X10-3 M 8.06 X10-3 M 7.7 X10-3 M 8.5 X10-3 M 9.0 ×10-3 M 1.31 X10-2 M 2.12 X10-2 M 3.0 ×10-2 M 8.74 X10-3 M 1.05 X10-2 M 1.90 X10-2 M 8.05 X10-3 M 5.51 X10-3 M 7.33 X10-3 M 6.63 X10-3 M 6.33 X10-3 M 5.55 X10-3 M 5.50 X10-3 M 5.54 X10-3 M 5.67 X10-3 M 5.65 X10-3 M 8.5 X10-3 M 5.96 X10-3 M 1.067X10-2 M 6.33 X10-3 M 1.146X10-2 M 6.72 X10-3 M",
      "ocrConfidence": 0.9411028037383177
    },
    {
      "fieldName": "temp",
      "value": "10 10 10.0 15 15 20 20 25 25 25 25 25 25 25 25 25 25.0 25 25 25 30 30 35 35 35 40 40 45 45 50 50 55 55 55.0 55 60 65 70 35 15 15 15 15 15 25 25 25 25 25 25 25 25 35 35 35 25.0 5 10.0 10.0 10.0 10.0 10 15 20 20 20 25 25 30 30 35",
      "ocrConfidence": 0.9781408450704225
    },
    {
      "fieldName": "compound-info",
      "value": "COMPOUND NO = 1 MOL WGT 288.3 SODIUM DODECYL 1 SULFATE",
      "ocrConfidence": 0.9773636363636363
    }
  ],
  "tables": []
}
```

Mapping

What information makes it to the thesis or the literature? Can we create a database of things that don’t work?

Map where the information flows

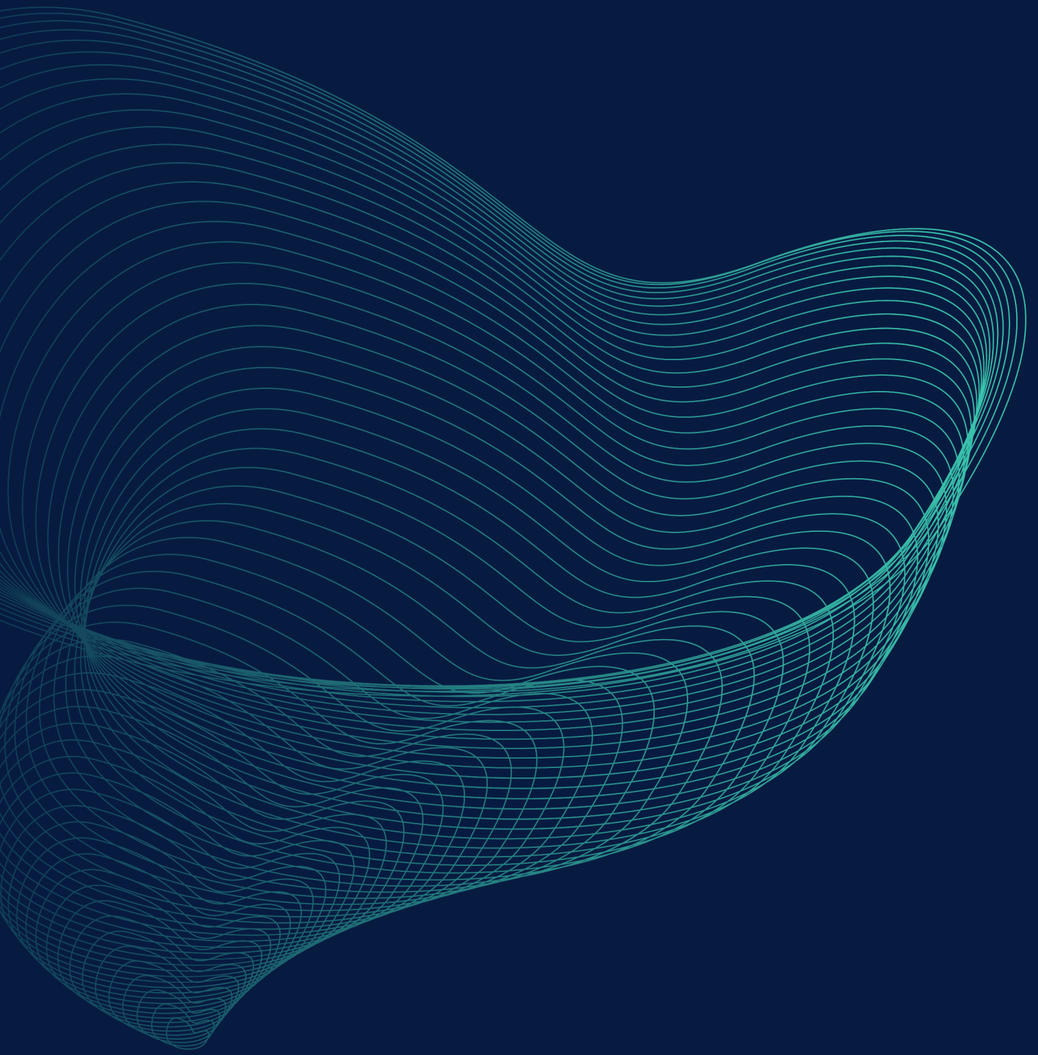
- Automatically extract meta data from lab books to track and map the thesis, literature and everything else

Analyse the full stack

- Extract information from the full stack of documents to create a web of knowledge
- Make all this information searchable, findable, accessible, open to AI etc

Create database of bad results

- What information makes it from the notebook all the way to the literature?
- Can we create a database of things that didn’t work?
- Would that be useful?



Thank you

Contact details

s.a.munday@soton.ac.uk
https://www.linkedin.com/in/samuel-munday/
www.data-revival.com