Demystifying data management: example from physics and chemistry

Petr Štěpánek

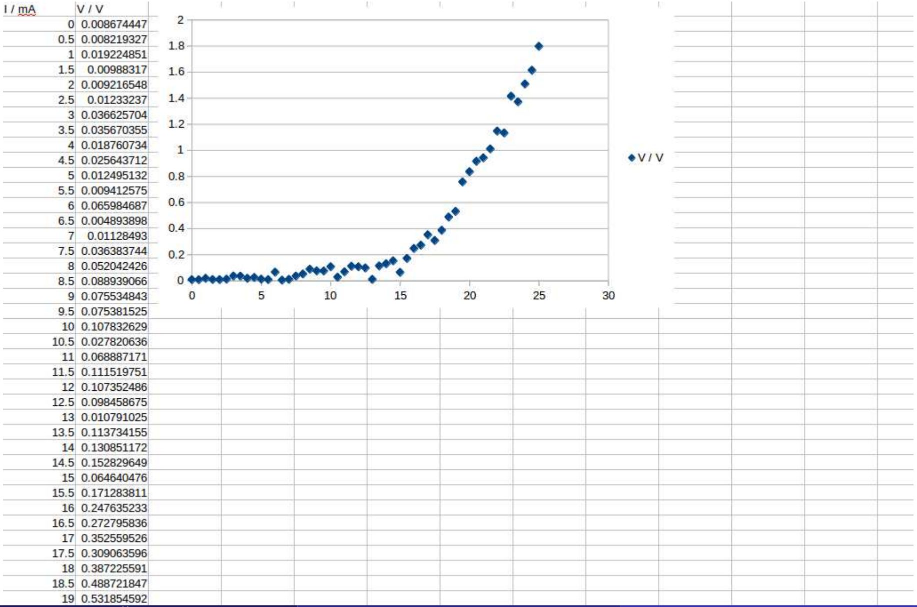
University of Oulu
Oulu, Finland

16 June 2021

Background

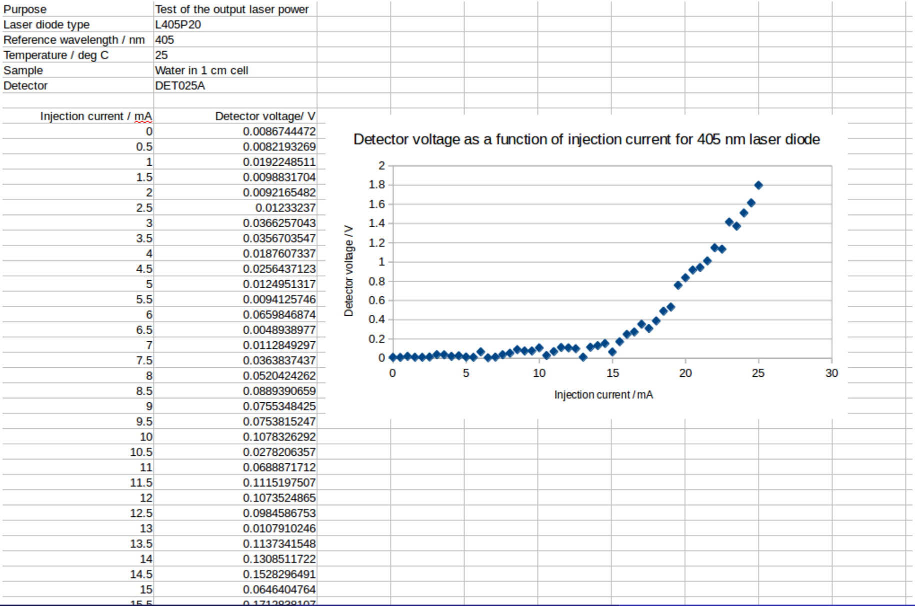
- NMR Research Unit
- Molecular spectroscopy and materials research
- Faculty of Science

What is good data?



I / mA	V / V
0	0.008674447
0.5	0.008219327
1	0.019224851
1.5	0.00988317
2	0.009216548
2.5	0.01233237
3	0.036625704
3.5	0.035670355
4	0.018760734
4.5	0.025643712
5	0.012495132
5.5	0.009412575
6	0.065984687
6.5	0.004893898
7	0.01128493
7.5	0.036383744
8	0.052042426
8.5	0.088939066
9	0.075534843
9.5	0.075381525
10	0.107832629
10.5	0.027820636
11	0.068887171
11.5	0.111519751
12	0.107352486
12.5	0.098458675
13	0.010791025
13.5	0.113734155
14	0.130851172
14.5	0.152829649
15	0.064640476
15.5	0.171283811
16	0.247635233
16.5	0.272795836
17	0.352559526
17.5	0.309063596
18	0.387225591
18.5	0.488721847
19	0.531854592

What is good data?



Purpose	Test of the output laser power
Laser diode type	L405P20
Reference wavelength / nm	405
Temperature / deg C	25
Sample	Water in 1 cm cell
Detector	DET025A

Injection current / mA	Detector voltage/ V
0	0.0086744472
0.5	0.0082193269
1	0.0192248511
1.5	0.0098831704
2	0.0092165482
2.5	0.01233237
3	0.0366257043
3.5	0.0356703547
4	0.0187607337
4.5	0.0256437123
5	0.0124951317
5.5	0.0094125746
6	0.0659846874
6.5	0.0048938977
7	0.0112849297
7.5	0.0363837437
8	0.0520424262
8.5	0.0889390659
9	0.0755348425
9.5	0.0753815247
10	0.1078326292
10.5	0.0278206357
11	0.0688871712
11.5	0.1115197507
12	0.1073524865
12.5	0.0984586753
13	0.0107910246
13.5	0.1137341548
14	0.1308511722
14.5	0.1528296491
15	0.0646404764
15.5	0.1712828107

Detector voltage as a function of injection current for 405 nm laser diode

The problem: insufficient details

Mindset

1. What was the last time you read an article with a description of an experiment and found it difficult to replicate because it was not well described?
2. Try to avoid that

The basic principle

1. Record everything
2. What, why, when, by whom, what were the conditions
3. It is better to have too much than too little
4. Store it in a findable way already during the production

What data do we have?

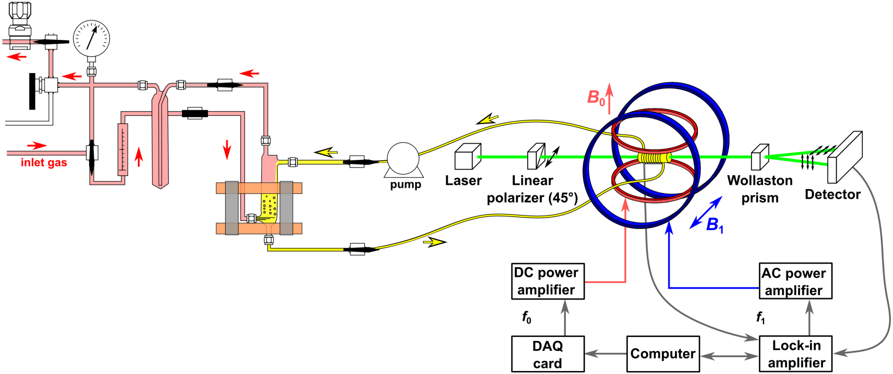
Type of data we produce:

1. Instrumental measurements
    - *e.g.*, ordered sets of values $[f, t, V, I, ...]$
    - Experiment description
    - Experiment settings (parameters)
2. Results of the theoretical modelling

- Case study: Measurement of spectroscopic effects using a custom-built instrument

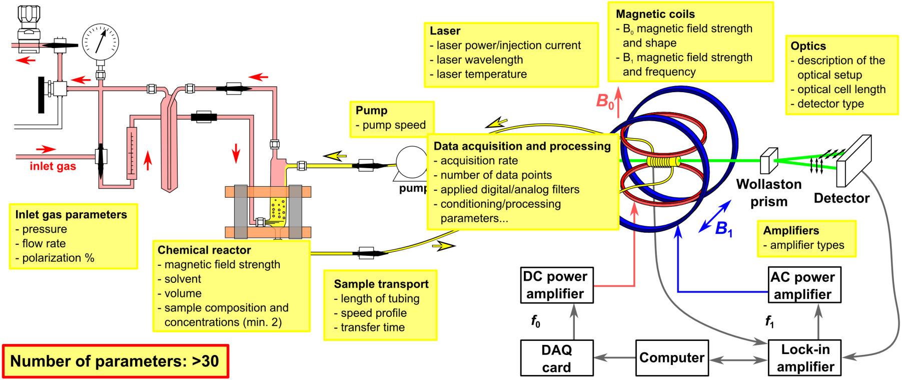
What data do we collect?

Main challenge: a custom-built instrument with many variables



What data do we collect?

Main challenge: a custom-built instrument with many variables



Laser
- laser power/injection current
- laser wavelength
- laser temperature

Magnetic coils
- B_0 magnetic field strength and shape
- B_1 magnetic field strength and frequency

Optics
- description of the optical setup
- optical cell length
- detector type

Pump
- pump speed

Data acquisition and processing
- acquisition rate
- number of data points
- applied digital/analog filters
- conditioning/processing parameters...

Inlet gas parameters
- pressure
- flow rate
- polarization %

Chemical reactor
- magnetic field strength
- solvent
- volume
- sample composition and concentrations (min. 2)

Sample transport
- length of tubing
- speed profile
- transfer time

Amplifiers
- amplifier types

Number of parameters: >30

Handling the data: Experimental settings

Description of the experimental parameters

- Digital (can be automatized)
  - Magnetic fields and its shape
  - Detection frequencies
  - Laser intensity
  - Laser temperature
  - ...
- Analog (depend on user)
  - Concentrations
  - Flow rates
  - Solvents
  - Physical properties (tubing diameters)
  - Optical setup
  - Instrument models
  - ...

General considerations

Data stored in a permanent hierarchy /measurements/YYYMMDD/
Each directory contains

1. A text file (.odt) with the experiment description
   - What and why is being measured
   - Hypotheses tested
   - Methodology
   - Description of the experimental parameters
2. Actual data measurement file(s). Each file contains
   - What and why is being measured
   - A header with description of the experimental parameters
   - Actual data ([f,t, V, ...] in binary format)

$\Rightarrow$ Redundancy in parameters description

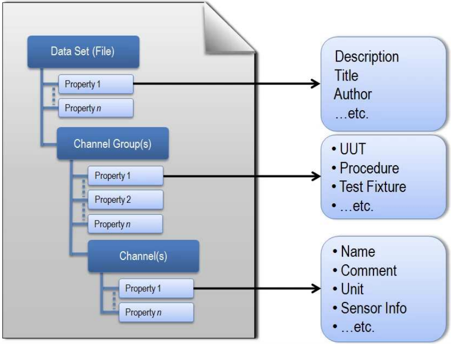
3. Processed data (.ods)

- Check of data completeness

Additional documentation: Photography

Handling the data: Instrumental measurements

Data stored in a file structure (TDMS)



https://www.ni.com/fi-fi/support/documentation/supplemental/06/the-ni-tdms-file-format.html

- Proprietary format (National Instruments); plugings available (Excel, MATLAB, Python, ...)
- Mainly optimized for recording of continuous data, but customizable
- Each channel automatically has a property field

Handling the data: Lifetime of the data

- Production of data
    - Record data itself
    - Describe the data and parameters (in the file and externally)
    - Backups
- Publication
    - (Format conversion + validity check)
    - Complete metadata (authors, keywords, what kind of data)
        - Record metadata (15 minutes in QVAIN)
    - Cite in the article
    - Deposit (in repository) (15 minutes, depending on the level of data reorganization)

Take-home suggestions

- Try imagining what would someone need to replicate your experiment
- Try to set up your experiment assuming no prior knowledge
- Show your results to someone not familiar with your project