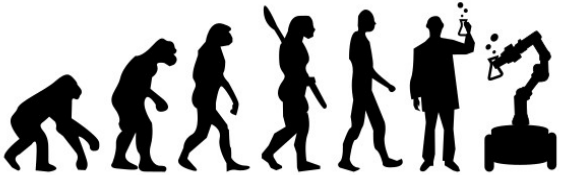


Automated and data-driven chemistry

LECTURE 8 :
Data management

Pascal Miéville
Stefano Di Leone
Edy Mariano
Jean-Charles Cousty

01 Data, files and types
02 Softwares
03 Data management
04 Data analysis / treatment
05 Project session

1. Data, files and types

Information vs data

Data : raw result (text, numbers, images ...), objective and without context.

Example : Temperatures in degrees Celsius: 20, 25, 18, 22, 28.

Information : processed data, contextualize and subjective

Example : Yesterday's temperature readings in degrees Celsius in Room 272 were : minimum temperature was 18°C, maximum temperature was 28°C, average temperature was 22.6°C

Data type for instruments

Types of data :

Spectroscopic Data (UV, IR, NMR)
Mass Spectrometry
Gas Chromatography (GC)
Liquid Chromatography (LC)
Titration Data
pH and Ion Concentration Data
Kinetic Data
...

→ Spectrums, Points, Text,…
→ (x,y) values , (x,t) values, (x,y,z) values, more complex (JCAMP) ...

```
##TITLE= ACET.SPOROPO.TYPHA/CH-CORR
##JCAMP-DX= 6.00   $$ ISAS JCAMP-DX program (Version 2.0 beta)
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= NTUPLES
##NUM DIM= 2
##ORIGIN= Dr. J. Lambert, ISAS Dortmund, Germany
##OWNER= COPYRIGHT (C) 1999 by ISAS Dortmund, Germany
##.SOLVENT NAME= CDCL3
##SPECTROMETER/DATA SYSTEM= JEOL GX 400
##.FIELD= 9.1
##.OBSERVE FREQUENCY=  100.40
##.OBSERVE NUCLEUS= ^13C
##.PULSE SEQUENCE= HETCOR
s1=        0.0
s3=        0.0
s4=         14
p1=       10.500
p2=       21.000
p3=       28.700
rd=     1000.000
id1=       0.114
d1=        0.114
d2=       10.000
NS=         2720
##.ACQUISITION SCHEME= NOT PHASE SENSITIVE
##.DIGITISER RES= 16
##.ZERO FILL=      0
##RESOLUTION=    5.87
##DATA PROCESSING= Ordinates are scaled between -32767 and +32767.
##NTUPLES= nD NMR SPECTRUM
##VAR_NAME= FREQUENCY1,    FREQUENCY2,    SPECTRUM
##SYMBOL=   F1,            F2,            Y
##.NUCLEUS= 1H,            13C
##VAR_TYPE= INDEPENDENT,   INDEPENDENT,   DEPENDENT
##VAR_FORM= AFFN,          ASDF,          ASDF
##VAR_DIM=  64,            1024,          1024
##UNITS=    HZ,            HZ,            POWER
##FIRST=       4370.000    ,  24038.50    ,
##LAST=        0.000000    ,  0.000000    ,
##MIN=         0.000000    ,  0.000000    ,       19135
##MAX=         4370.000    ,  24038.50    ,  626984008
##FACTOR=      1.000000    ,  23.49804    ,  19134.62
##PAGE= F1=    4370.000
##FIRST=   4370.000      ,  24038.50    ,    7806924
##DATA TABLE= (F2++(Y..Y)), PROFILE
1023D08K41k58M49o1169J9017j51Q78n32106L46n92M47o8p3k4212K98j00K1J77M63m19
999F78j5914j15M3n0oN14k68n61P80m17j20L85m10J74j90K2K4k46K65K52o6m46K91J45
974D85k74nR2J5J38k28N2J36P90185P2n66J23j25J82k97pL92L43k55k86O23n51k11P5o
948C22M62m99J63o3k75K65K52j94k7J0Q5j71M83175M12m73j27J53k05R9L69p0185j1M57
...
...
...
42A99J74L2J38K49o63J89K20j04k22j2M78m45j0J88118nN26K99o88M26m78j47O29k82P5
17D48P7j30m6p2J57J56J91p9196M83Lo75K47K99k97K64k89
0C41
##END NTUPLES= nD NMR SPECTRUM
##END=
```

Structure

Data can be structured

Predefined format

Databases, spreadsheets, tables with defined fields values

Consistent and predictable

→ easier to analyze and predict

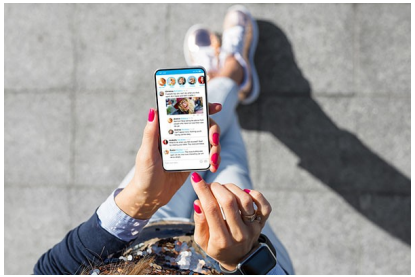
Data can be unstructured

Lack a specific format or organization

Text, images, audio, video, social media posts, emails

Variable and unpredictable

→ easier to construct



What data types do you know ?

Data types depends of the language :

Float : 26.5
Integer: 30
Text: "This is a sample text."
Boolean: True

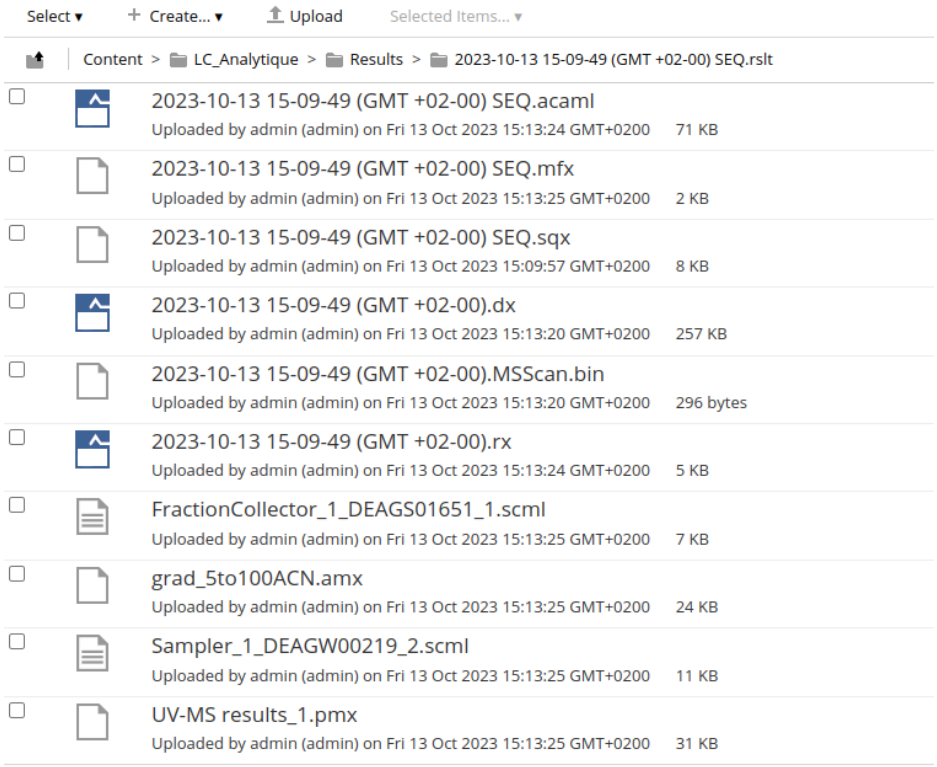
Concatenate multiple types of data :

Tuple : (1,2)
Array : my_array = [10, 20, "apple", True]
Dictionary : my_dict = {"name": "John", "age": 30, "city": "New York"}

File formats for instruments

.csv
.txt
.pdf
.xml
.json
.docx
.doc
.xlsx
.pptx
.yaml
.py
.sh
…….

.jpg
.png
.jpeg
.bin
.dx
.asm
.pmx
.amx
….



A result can be a mix of different file formats

Formats

Most instruments have their proprietary file format

→ Try to have an open format that you can analyze and from which you can view your data

Agilent : PDF report → ASM format

Bruker : PDF spectrum → Jcamp dx

Chemspeed : Log file .txt → csv, Database

Formats

Prefer descriptive and open formats to gather your data :
.xml
.json
.adf
.asm
…

Avoid **if possible** immutable (impossible to modify after creation) formats :
.pdf
.png
.jpeg

```
{
  "worksteps": [
    {
      "ID": "00100",
      "Name": "methyl hexanoate -ref_1d_1H_zg",
      "State": "Submitted",
      "WorkstepType": "EXPERIMENT",
      "ResourceName": "NMR-SPECT",
      "MethodID": "ref_1d_1H_zg",
      "Parameters": [
        {
          "ID": "SOLVENT",
          "DataType": "String",
          "Data": "CDCL3"
        },
        {
          "ID": "TUBEID",
          "DataType": "String",
          "Data": "SRE00020#NMR#A1"
        }
      ]
    }
  ]
}
```

2. Softwares

At a glance

Commercial softwares

Advantages :

- nothing to code

- small implementation

- gain time

- don’t have to reinvent the wheel yourself

Drawbacks :

- price

- not always a good fit for your demands

- may be proprietary / priced to get some data

- often complicated to modify (except if you have a contract stipulating it)

- code invisible (in most cases)

Open-source softwares

Advantages :

- community based

- modification possible (you can even code it yourself)

- less “hidden doors” → you can look at the code if you don’t trust it

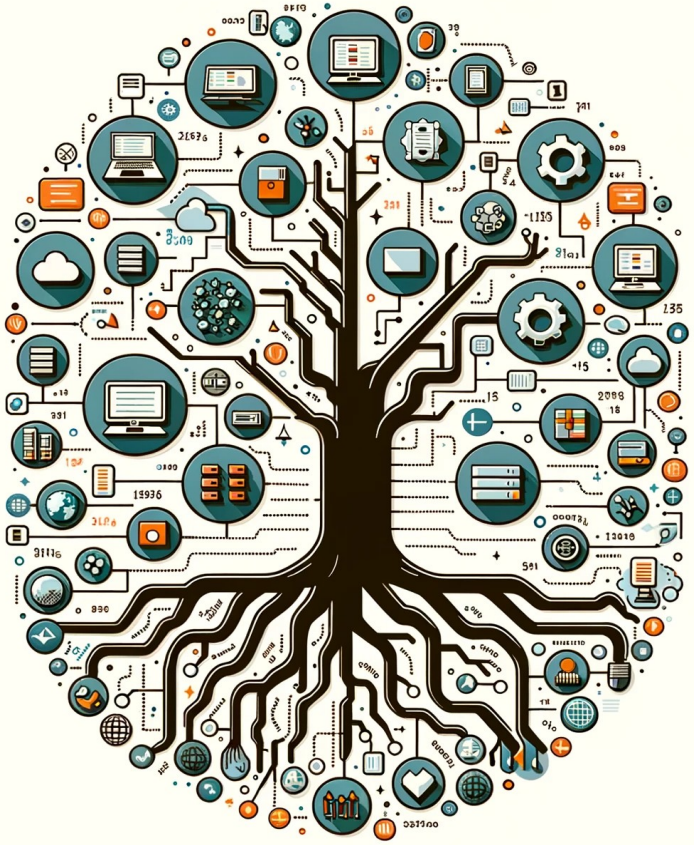
- PRICE (often support is priced)

Drawbacks :

- little code / implementation to be done yourself (depends on the soft / script to implement)

- could become complex to maintain

- modifications take time



3. Data management

Data is value

Data is money

→ Twitter (now X) sold for 44 billion dollars
→ Instagram sold for 1 billion in 2010

What is sold is a database of client + a way of advertising to a large audience at once

Data management

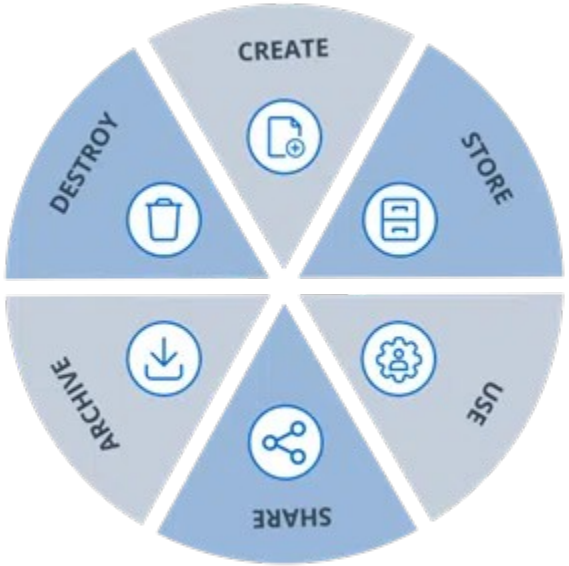
Proper data management ensures data accuracy, security, and accessibility.

Poor data management led to data breaches, files loss, production decrease ...

Ryuk example
→ Phishing
→ encrypted data and computers
→ weeks of paralysis for the entire IT system and the company

Data management lifecycle:

Creation / Collection: Gathering and creating data from various sources.
Storage: Safely storing data in appropriate formats.
Processing: Transforming data into useful information.
Analysis: Extracting insights from data.
Sharing: Distribute data to other users.
Archiving: Long-term storage of historical data.
Deletion: Removing data that is no longer needed.



Data integrity

Data integrity: Accuracy and consistency of trustworthy data throughout its lifecycle. Data documentation has to exist.

1. Completeness: Data should be complete, meaning that all relevant information is included.

2. Consistency: Data should be consistent, meaning that it is uniform throughout the dataset.

3. Validity: Data should be valid and reliable, meaning that it conforms to predefined rules and constraints. → Data quality checks

Data safety

Data safety: Data protection from loss, corruption, or unauthorized access. It involves ensuring that data is available, reliable, and secure, and that it can be recovered.

Data backup: copies of data stored securely, tested regularly

Data recovery: restoring data from backup in the event of a disaster or system failure

Data security: protect data from unauthorized access, modification, or theft → encryption, tokens, access control and permissions...

4. Data analysis / treatment

At a glance

Raw or processed or both ?

Raw data is extracted directly from the equipment and can hence be large in size (several MB per sample)

Processed data is a subset / a processed set of the raw data

Solutions to gather data from the instrument :

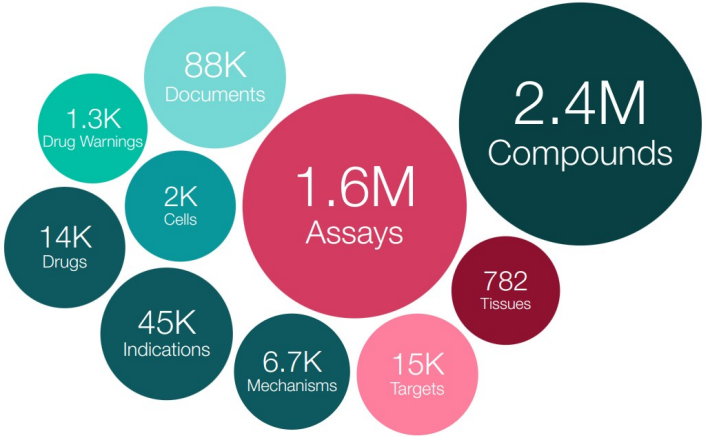
1- Keep only the raw data : example: spectrum NMR / dx file
2- Keep only the processed data : example : keep the peak list
3- Keep both

Think about your use case first to know what to keep !

At a laboratory level : tips you can use

Next week - Small demo of a Jupyter Notebook to do simple chemical formulas / reasearch …

Use of ChEMBL database : database of bioactive molecules with drug-like properties : https://www.ebi.ac.uk/chembl/g/#browse/compounds



Hands on session – Next week

Go here :
https://github.com/jccousty/AutomatedLabCourse/blob/main/target_CHEMBL205.csv.gz
Download the file

Go here:
https://github.com/jccousty/AutomatedLabCourse/blob/main/Exercice_data_analysis.ipynb
And click on the « Open in Colab » button
When launched click on the file icon on the left and then the upload to session storage icon and upload the file you just downloaded

Then you’re good to go !

5. Project session

For the next project session the groups have to :

a. identify all data that can be extracted from the automated workflow previously described for the project;
b. describe as much as possible the type of extracted data, suggest data and file formats;
c. propose a complete IT-structure for the project including data safety aspects.
Reports must be sent to teacher before the next lecture. An individual correction will be done and sent back with comments within one week.

Feel free to ask questions !!

Conclusion

Thank you for your attention.
Do you have any questions ?

Thanks to

Swiss CAT+ team & partners

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

ETH-RAT

Innosuisse

Gimp, Ubuntu Foundation

References

https://www2.chemie.uni-erlangen.de/projects/vsc/chemoinformatik/erlangen/daten/jcamp.html

https://www.google.com/url?sa=i&url=https%3A%2F%2Fstock.adobe.com%2Fsearch%3Fk%3Dhacker&psig=AOvVaw1sJDsrr0du1FKkIUnmZ8vt&ust=1699114082194000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCKi_yaybqIIDFQAAAAAdAAAAABAJ

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.bitlyft.com%2Fresources%2Fryuk-ransomware-zerologon-exploit&psig=AOvVaw3ZHnncIui4xYX50TpALQac&ust=1699114815909000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCODQmYqeqIIDFQAAAAAdAAAAABAE

https://blog.netwrix.com/2023/03/03/data-lifecyle-management/

https://www.google.com/url?sa=i&url=https%3A%2F%2Fblog.quest.com%2Fusing-database-schemas-in-sql-server%2F&psig=AOvVaw24N_O_JsF8PuHzRxmO7uM0&ust=1700576130641000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCPDznfPh0oIDFQAAAAAdAAAAABAE

https://www.google.com/url?sa=i&url=https%3A%2F%2Fpenntoday.upenn.edu%2Fnews%2Fannenberg-communication-what-makes-us-share-posts-social-media&psig=AOvVaw2ZFY2-JOxeDJjscRTevvkJ&ust=1700575923610000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCKjHuZDh0oIDFQAAAAAdAAAAABAE