Hands-on with Research Data Management in Chemistry
- Day 3

Alain Borel

Francesco Varrato

Research Data Team, EPFL Library

EDCH-RDM, Feb. 2024

Alain Borel, Francesco Varrato

Program

	2024 – 02 – 21	2024 – 02 – 22	2024 – 02 – 29
Morning 9:00-12:30	1. [1h] **Theory 1**	5. [1h] **Tools 2**	9. [1h] **Theory 3**
	2. [3h] ***Patiny***	6. [3h] ***Dürr***	10. [3h] **Theory 4**
LUNCH BREAK			
Afternoon 13:15-17:00	3. [1h] **Tools 1**	7. [3h] **Hands-on: Data workflow**	11. [3h] **Hands-on: Project / Report**
	4. [3h] **Theory 2**	8. [1h] **Feedback 1**	12. [1h] **Feedback 2**

Alain Borel, Francesco Varrato

File formats: room for improvement

Unlicensed Published by De Gruyter July 11, 2022

Data format standards in analytical chemistry

David Rauh, Claudia Blankenburg, Tillmann G. Fischer, Nicole Jung, Stefan Kuhn, Ulrich Schatzschneider, Tobias Schulze and Steffen Neumann

From the journal Pure and Applied Chemistry

https://doi.org/10.1515/pac-2021-3101

Source: Rauh & al. "Data format standards in analytical chemistry" Pure and Applied Chemistry, vol. 94, no. 6, 2022, pp. 725-736. https://doi.org/10.1515/pac-2021-3101

"Generic data formats like AnIML and JCAMP-DX have been used for many applications. Special formats for some analytical methods are already accepted, like mzML for mass spectrometry or nmrML and NMReDATA for NMR spectroscopy data. Other methods still **lack common standards** for data."

"Only a **joint effort** of chemists, instrument and software vendors, publishers and infrastructure maintainers can make sure that the analytical **data will be of value in the future**"

Alain Borel, Francesco Varrato

File formats: what's the point

Standardized, **open** & **widely used** formats to:

... work on **multiplatform** / multi-OS

... **collaborate** with more people

... avoid **licensing** problems

... maximize future research **reusability**

... be **independent** of a particular software / company

Examples of generic Open data formats

PDF/A: ISO standard, archiving, no ciphers, included fonts, ...

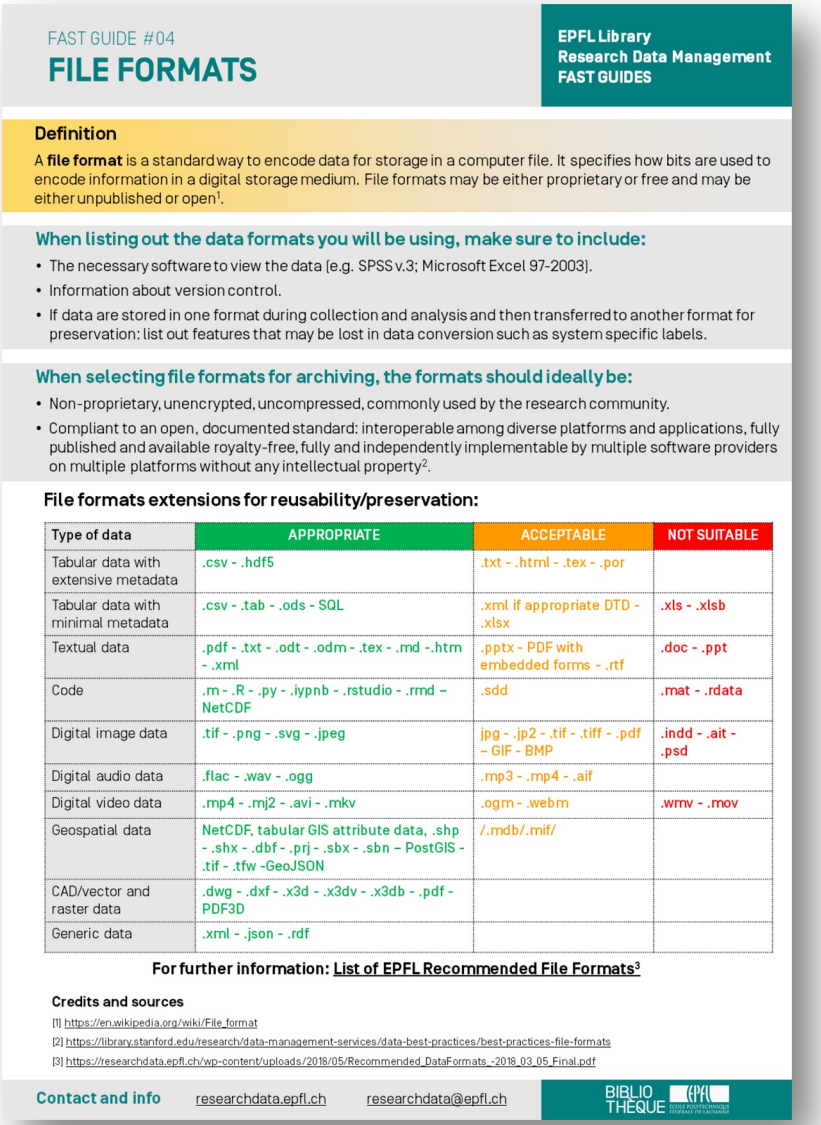
CSV: apt for tables, extensible with CSV on the Web

SVG: web friendly, native multiplatform support

SQL (databases communication, Postgresql, PostGIS)

MySQL or MariaDB (supported by the EPFL central IT)

HDF5 (flexible, wide compatibility, Python, R, Matlab, ...)



FAST GUIDE #04

FILE FORMATS

EPFL Library
Research Data Management
FAST GUIDES

Definition

A **file format** is a standard way to encode data for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free and may be either unpublished or open[1].

When listing out the data formats you will be using, make sure to include:

- The necessary software to view the data (e.g. SPSS v.3; Microsoft Excel 97-2003).
- Information about version control.
- If data are stored in one format during collection and analysis and then transferred to another format for preservation: list out features that may be lost in data conversion such as system specific labels.

When selecting file formats for archiving, the formats should ideally be:

- Non-proprietary, unencrypted, uncompressed, commonly used by the research community.
- Compliant to an open, documented standard: interoperable among diverse platforms and applications, fully published and available royalty-free, fully and independently implementable by multiple software providers on multiple platforms without any intellectual property[2].

File formats extensions for reusability/preservation:

Type of data	APPROPRIATE	ACCEPTABLE	NOT SUITABLE
Tabular data with extensive metadata	.csv - .hdf5	.txt - .html - .tex - .por	
Tabular data with minimal metadata	.csv - .tab - .ods - SQL	.xml if appropriate DTD - .xlsx	.xls - .xlsb
Textual data	.pdf - .txt - .odt - .odm - .tex - .md -.htm - .xml	.pptx - PDF with embedded forms - .rtf	.doc - .ppt
Code	.m - .R - .py - .iypnb - .rstudio - .rmd – NetCDF	.sdd	.mat - .rdata
Digital image data	.tif - .png - .svg - .jpeg	jpg - .jp2 - .tif - .tiff - .pdf – GIF - BMP	.indd - .ait - .psd
Digital audio data	.flac - .wav - .ogg	.mp3 - .mp4 - .aif	
Digital video data	.mp4 - .mj2 - .avi - .mkv	.ogm - .webm	.wmv - .mov
Geospatial data	NetCDF, tabular GIS attribute data, .shp - .shx - .dbf - .prj - .sbx - .sbn – PostGIS - .tif - .tfw -GeoJSON	/.mdb/.mif/	
CAD/vector and raster data	.dwg - .dxf - .x3d - .x3dv - .x3db - .pdf - PDF3D		
Generic data	.xml - .json - .rdf		

For further information: List of EPFL Recommended File Formats[3]

Credits and sources

[1] https://en.wikipedia.org/wiki/File_format
[2] https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats
[3] https://researchdata.epfl.ch/wp-content/uploads/2018/05/Recommended_DataFormats_-2018_03_05_Final.pdf

Contact and info researchdata.epfl.ch researchdata@epfl.ch

BIBLIOTHEQUE

Download our Fast Guide on File Formats ☺

Alain Borel, Francesco Varrato

Commonly used file formats in chemistry

File Extension	MIME Type	Proper Name	Description	Type
cdx	chemical/x-cdx	ChemDraw eXchange file		chemical structures
cif	chemical/x-cif	Crystallographic Information File, Crystallographic Information Framework	Promulgated by the International Union of Crystallography	chemical structures
mol	chemical/x-mdl-molfile	MDL Molfile (developed into SDL)		chemical structures
dx, jdx		UV/VIS, NIR, IR, fluorescence, Raman spectra in JCAMP-DX 424/500 standard		spectroscopy

And at least 150 less frequent formats just for basic spectroscopy, chemical structures and X-ray diffraction

Alain Borel, Francesco Varrato

Open format files: a good example

Research data is open by default, since 2017

SUMMARY

[...] Musical scores will be stored in MusicXML or MIDI format [...]

…

4. INCREASE DATA RE-USE

[...] format converters will be employed (or implemented) to keep copies of MusicXML in other formats, such as MEI or Humdrum (which have been stable for a long time already).

By EPFL *Digital and Cognitive Musicology Lab*

Alain Borel, Francesco Varrato

Open format files: a not-so-good example (1/2)

Measurement parameters (1)	1) Characterization of organo- and biocatalytic reactions 2) Study of compatibility of catalysts and design of one-pot process (in batch mode)	1) Conversion, enantiomeric and diastereomeric excess 2) Substrate loading, weight	1) Sample balance 2) NMR/GC/-HPLC, balance	1) .jpg 2) .xlsx 3) .docx	1) Origin 2) Chemoffice, Mestrenova, MS Office
Measurement parameters (2)		1) Fluid velocities 2) Liquids contact angles		1) mpeg4	1) MatLab
Measurement parameters (3)	1) Flow and phase characterization 2) Reaction performance incl. separation & recovery	1) Photograph/movie of samples, density 2) Concentration, weight	1) Sample, balance/syringe 2) NMR/GC/-HPLC, balance	1) .jpg/.avi 2) .xlsx	1) Origin 2) Chemoffice, Mestrenova
Measurement parameters (4)	Protein quantification assays, concentration, weight	Concentrations	Sample balance, UV/vis spectrophotometer	.TIF, .xlsx, .doc	Excel, Word
Sample description (1)	Chemicals, solvents, organocatalysts, enzymes, products	See under Measurement parameters and Experimental results	Chemical provider, own work /preparation	.txt, .jpg, .pdf, .docx	1) Origin 2) Chemoffice, Mestrenova, MS Office
Sample description (3)	Chemicals, solvents, solvent mixtures, reaction mixtures	Not relevant or given in Measurement parameters and Experimental results	Chemical provider, partner (UniBi), own work/preparation	.txt, .jpg, .pdf	Chemwatch
Sample description (4)	Chemicals, reaction mixtures, protein mixtures	Characterization of chemicals and enzymes	NMR and MS analysis 2)SDS-PAGE and native-PAGE	.TIF, .pdf, .xls, .doc, .cdx	Adobe acrobat, Kaleidagraph, Excel, Word, Chemoffice, Mestrenova

2.4 Reuse of data
"Data will become available for re-use immediately after publication at 4TU.Centre for Research Data"

Source: **Appendix 1 : detailed overview of data types and formats,** a project of Eindhoven Technology University & 3 partners (ERC project archive)

Open format files : a not-so-good example (2/2)

Tabulated data

XLSX and XLS? Why? Why BOTH?
Spreadsheet are not inherently bad, but people rely on them too much. Excel is the worst offender in that category despite:

- Known bugs that can affect statistical analyses
- Poor default settings for graphs that are difficult if not impossible to change
- Limited number of rows in XLS…

Origin, Kaleidagraph & friends: 100% closed if not handled carefully.

Video data

MP4 is an ISO standard, good.
AVI on the other hand only defines part of the file: the actual video and sound inside the files could be using just about any encoding scheme.

Chemical structures

ChemDraw CDX format isn't too bad, at least the vendor is providing technical documentation so it can be converted using a number of tools (OpenBabel for example)

Text documents

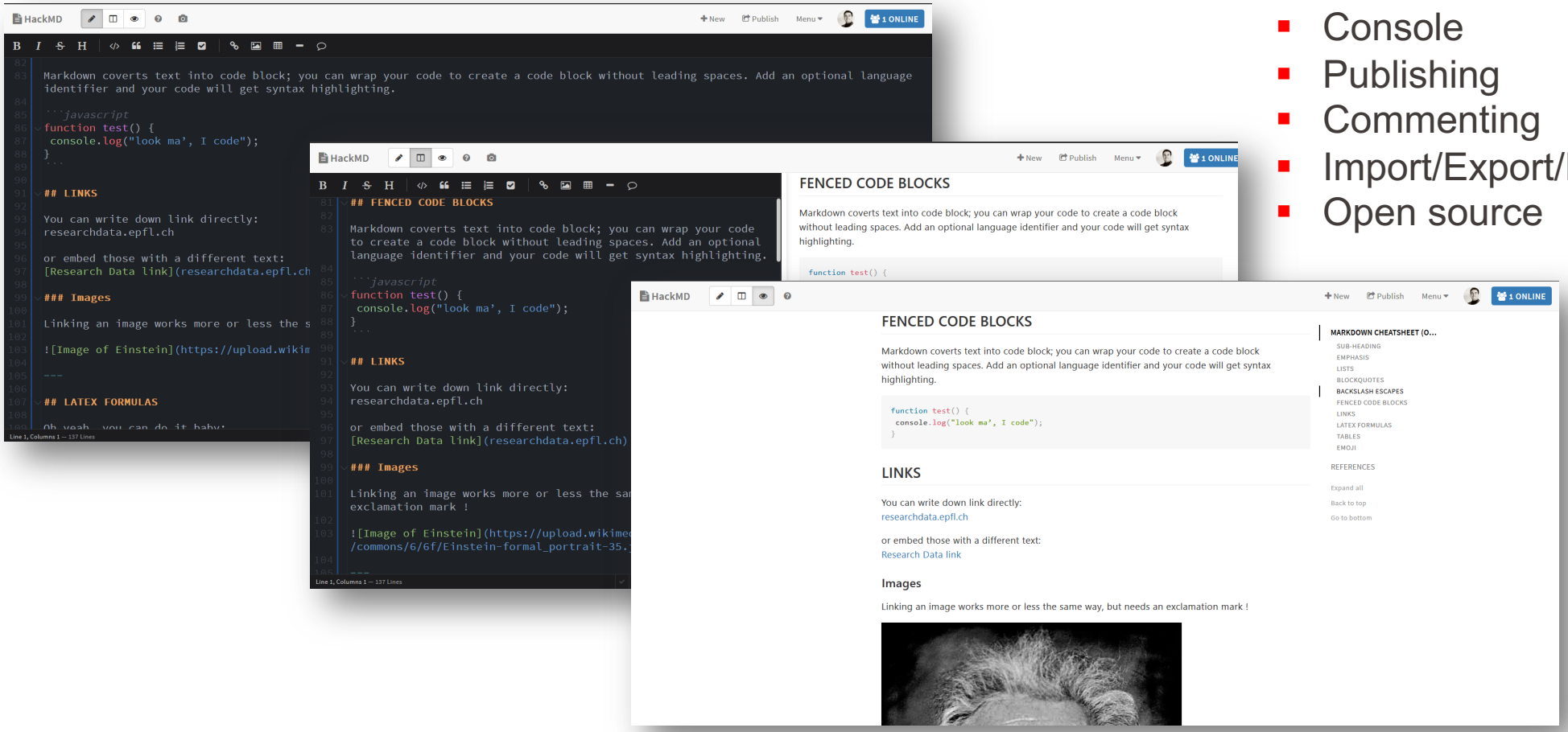
DOCX: editable (i.e. re-usable) but never properly documented⇨conversions often slightly flawed
PDF: well-documented, at least one variant is an ISO standard… great to read but difficult to re-use

NOTE: making the data public does not make it FAIR.
Using open formats does **not** imply that data in that format is public

Alain Borel, Francesco Varrato

Open software – Example (hackMD.io)

1. **Writing** collaborative tool: hackMD.io
2. **Markdown** cheat/sheet: go.epfl.ch/bsX



Why hackMD?

- Multilanguage
- Access management
- WYSIWYG → YSWYG
- Console
- Publishing
- Commenting
- Import/Export/Download
- Open source

Open Science-friendly software… ?

Office 365	OnlyOffice	Framaforms
Google Docs	LibreOffice	Qualtrics
Authorea	LaTeX	REDCap
Word	OverLeaf	SurveyMonkey
Writer	GitHub	Google Forms
Miro	GitLab	LimeSurvey
Openboard.ch	Doodle	CATM
Zotero	Framadate	Nvivo
		…

Alain Borel, Francesco Varrato

Open Science-friendly = Open Source ?

Office 365	**OnlyOffice**	**Framaforms**
Google Docs	**LibreOffice**	Qualtrics
Authorea	**LaTeX**	REDCap
Word	OverLeaf	SurveyMonkey
Writer	GitHub	Google Forms
Miro	**GitLab**	**LimeSurvey**
Openboard.ch	Doodle	**CATMA**
Zotero	**Framadate**	Nvivo
		…

Alain Borel, Francesco Varrato

Discussion: If your storage breaks…

☹️ Should you **be sad?**

 **How much** data would you lose?

 Could you **rescue** some data?

 How do you **prevent** it?

Image: pixabay.com/fr/images/search/hazard

Alain Borel, Francesco Varrato

Storage ≠ Back-up

1. Active *vs.* Cold

Active: Data to be *accessed* right away during the research
Cold: Data with *low-frequency* access, not requiring fast access

2. Backup *vs.* Archiving

Backup: Copies of original files made before the original is overwritten
Archiving: Copies of files for space management and *long-term* retention

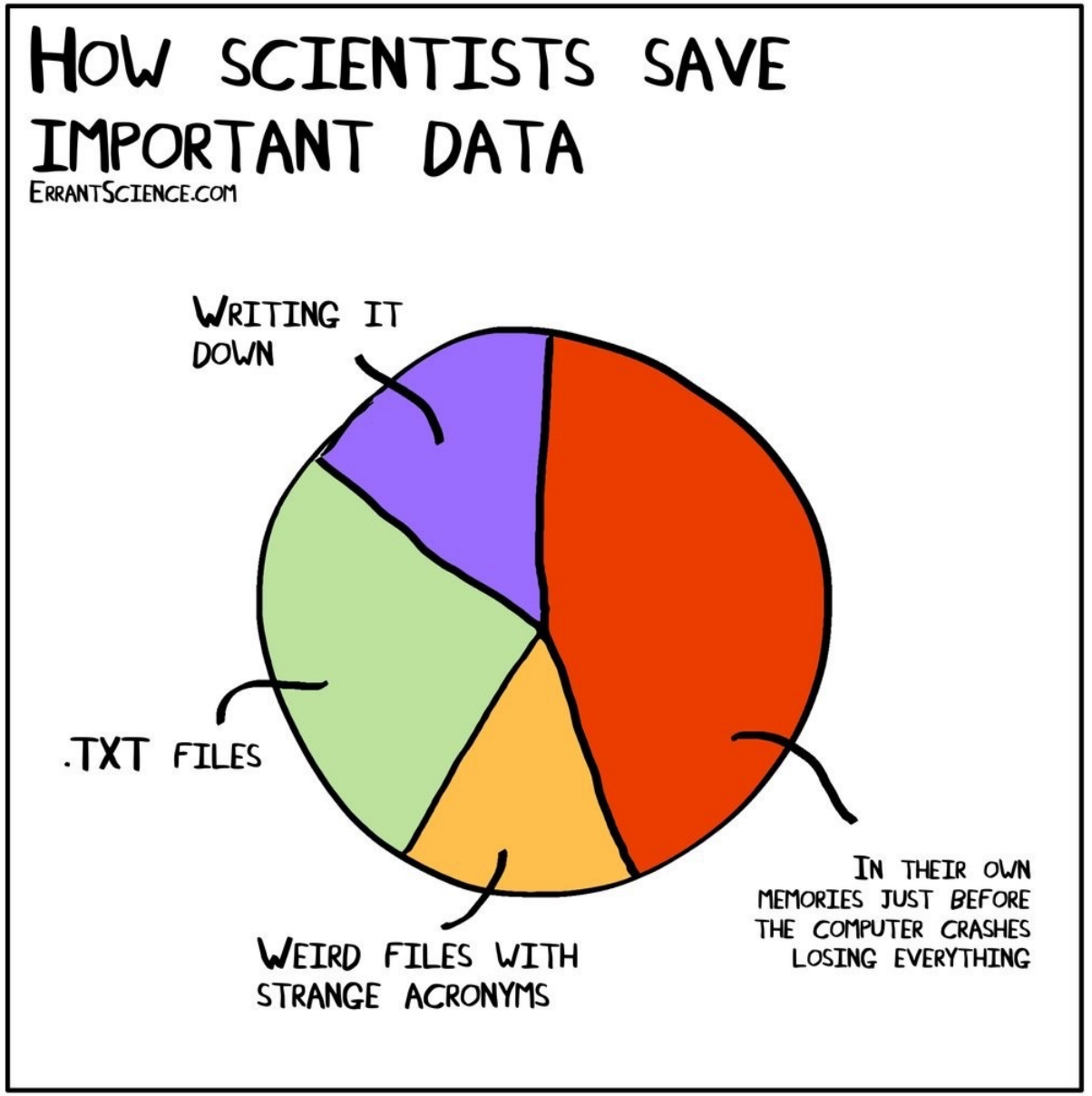
3. Preservation *vs.* Publication

Preservation: Archiving *after* data reformatting, conversion, metadata and data rescue
Publication: Preserved, findable (DOI), *publicly* accessible, scholarly product

4. Security *vs.* Privacy

Security: Protection against the unauthorized *access* of data
Privacy: Related to any rights on controlling and/or using *personal* data

Reliable storage options (?)



Alain Borel, Francesco Varrato

Alain Borel, Francesco Varrato

Active storage solutions

Just a few on EPFL servers

EPFL

DSI central NAS file storage

MyNAS individual storage for students, staff and guests

Atempo Lina service to backup workstations / personal computers

rsync utility to transfer / synchronize data across computer systems

eln.epfl.ch chemistry-oriented, web-based ELN with archiving support

SLIMS life sciences-oriented, ELN + LIMS

gitlab.epfl.ch version control, open alternative to GitHub using EPFL servers

SCITAS' Work storage and c4science storage for coders (w/ LFS)

SWITCHDrive hosted by SWITCH (CH)

MS OneDrive hosted on Microsoft servers (last one in CH for EPFL)

gdrive.epfl.ch hosted on Google (Alphabet) servers

Reduced comparative table: www.epfl.ch/campus/services/en/it-services/storage-of-documents

Active storage solutions

External / Other

- Backups disks / NAS
- Private cloud (Box? Dropbox? kDrive? Tresorit? …)
- Research facilities (SDSC? ETHZ partners?)
- Code repository (Github?)

EPFL-DSI storage (evolving !)

	Collaborative	Online archive	Raw
Performance	Very good	Good	Good
Replication	✓	✓	✗
Snapshots	✓	✓	✗
Protocols	NFS, SMB/CIFS, WebDAV (optional)	NFS, WebDAV (optional)	S
Backups	Optional	Optional	✗
Price (Old)	CHF 165 /TB /year	CHF 110 /TB /year	CHF 55 /TB /year

Storage ≠ Backup

More info at the EPFL File Storage page or contacting 1234@epfl.ch

Alain Borel, Francesco Varrato

...RCP storage

	NASRCP
PERFORMANCE	Very good
REPLICATION	✓
SNAPSHOTS	✓
PROTOCOLS	NFS, SMB/CIFS
BACKUPS	No
PRICE	**CHF 27/TB /year**

More info at the EPFL File Storage page or contacting 1234@epfl.ch

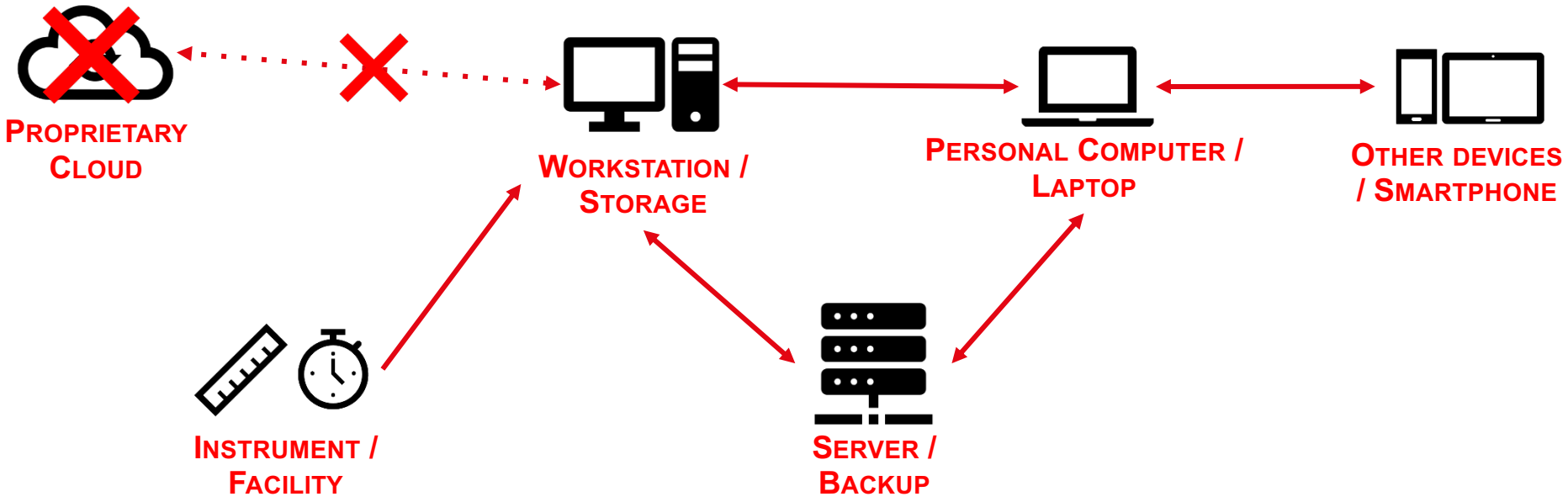
File sync (… not as 'convenient' as the cloud, but …)

Alain Borel, Francesco Varrato

Back-up oriented

Time Machine | Atempo Lina | + | FreeFileSync | Rsync | Syncthing | … | OwnCloud

Example:



Cold storage

EPFL

ACOUA (ACademic OUtput Archive), EPFL data preservation tool:
Data curation + Long-term integrity + Security

Stockage Object S3 based on AWS protocol for large data sets, archiving and backups

+ EXTERNAL

OLOS is the data repository of UniGe

SWISSUbase data repository for SSH projects

… re3data.org

use **go.epfl.ch/datarepo (more on this later)**

Alain Borel, Francesco Varrato

ELN/LIMS

ELN = ELECTRONIC LAB NOTEBOOK

LIMS = LAB INFORMATION MGMT SYSTEM

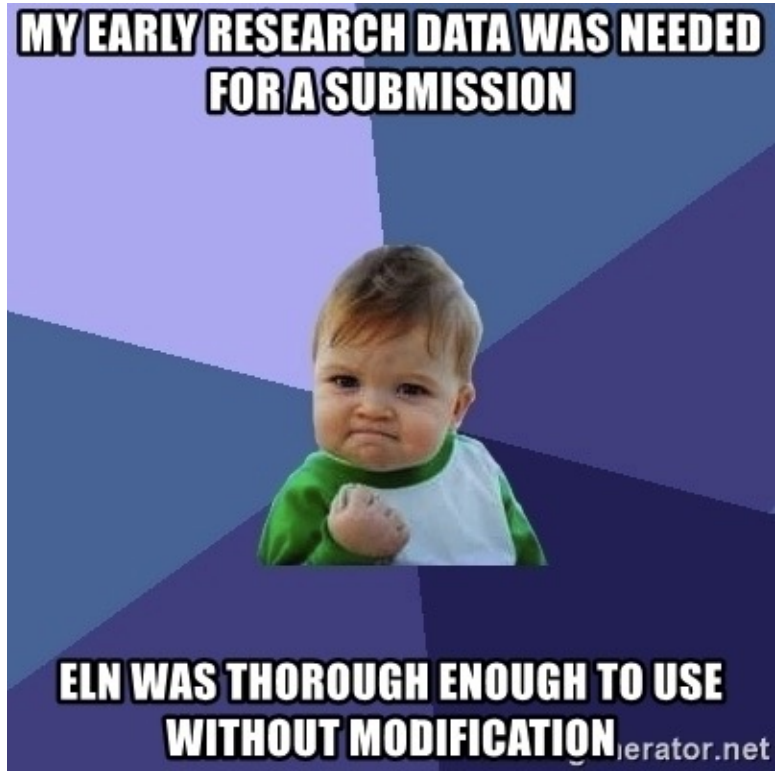


Image by kaidran

Bundling: data + notes + logs + code + scripts + stock management + …

ELN, LIMS & more …

… to rule them all (**?**)

Image by Peter J. Yost, Wikipedia

ELN, LIMS & more: a non-authoritative typology

ELN
Electronic Lab Notebook

Replacement for paper laboratory notebooks

- Allows sharing and rights management
- Allows searching
- Easier to include digital data
- …

LIMS
Laboratory Information management system

Supports laboratory operations

- Tracks workflows (samples, methods, etc.) and data
- Easier to include digital data
- Integrates with instruments
- …

SDMS
Scientific Data Management System

Manages documents and data

- Captures, indexes and archives data produced by instruments and software (incl. ELNs and LIMS…)

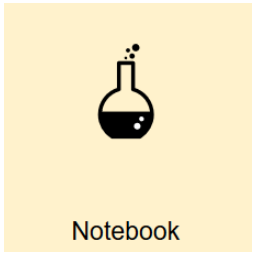
For more details and other definitions: https://www.limswiki.org/index.php/LIMSWiki:Glossary

Alain Borel, Francesco Varrato

A few ELN/LIMS examples

SLIMS

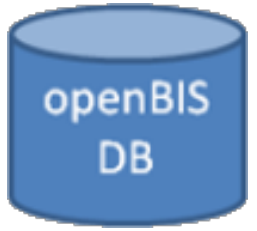
ELN EPFL



Chemotion

ResearchSpace

openBIS



EPFL

ETHZ

More about Rspace:
https://www.youtube.com/watch?v=bhRExXIGxek
More about Chemotion:
https://doi.org/10.5281/zenodo.7634481

(Download our Fast Guide on Electronic Lab Notebooks ☺)

Others: Harvard Medical School
comparison table (2021)

TU Darmstadt
ELN Finder (2023)

Alain Borel, Francesco Varrato

Reproducibility: how ELNs can help

1985: George P. Smith reports an original phage display method: inserting genes in the DNA for a specific phage protein, allowing the phage to infect and reproduce in bacteria

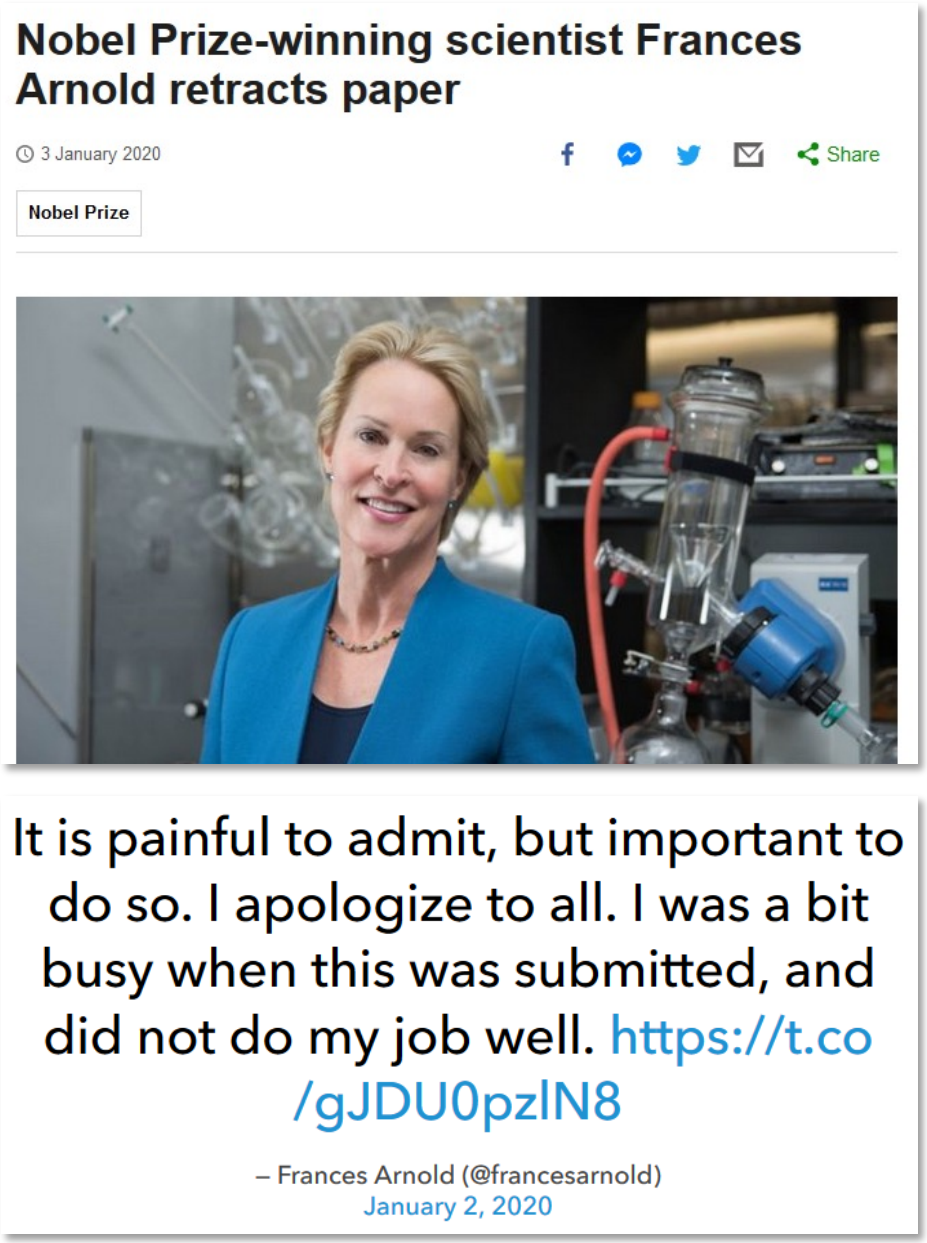
1993: Frances H. Arnold conducts the first directed evolution of enzymes

2016: The €1m Millennium Technology Prize is awarded to Prof Arnold for her pioneering work on "directed evolution"

2018: Nobel: Prof Arnold shares the award with George P. Smith and Gregory Winter for their research on enzymes

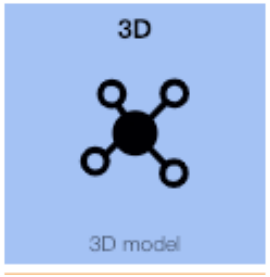
2019: Cho, Jia & Arnold publish a *Science* report on how to apply the appropriate evolutionary pressure

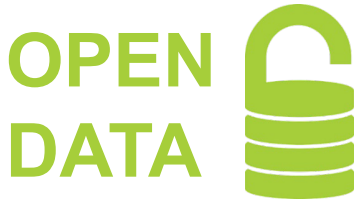
2020: Retraction: efforts to reproduce the report's work and consequent **examination of the first author's lab notebook** revealed missing entries and raw data for key experiments.



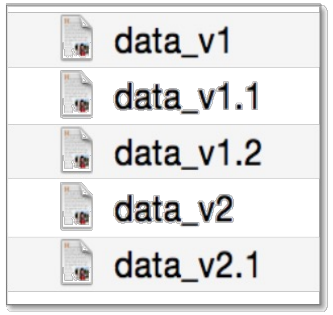
Sources: www.bbc.com/news/world-us-canada-50989423, www.chemistryworld.com/opinion/frances-arnolds-retraction-and-the-case-for-slow-science/4010994.article

Recap





Example: FR3S.140623.129C.2653.W.JPG



ACTIVITIES	COLLEAGUE / PARTNER	TOOLS	TO-DO
FUNDING PLANNING			
CREATION			
ETHICAL CLEARANCE			
ACQUISITION			
STORING			
ANALYSIS			
LEGAL CLEARANCE			
SHARING			
PUBLISHING			
ARCHIVING			

Alain Borel, Francesco Varrato