For your convenience: one possible format

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

Documentation all along

See paper copy + on Moodle go.epfl.ch/ChE-601

Alain Borel, Francesco Varrato

Research reproducibility (issue)

“There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.”

Enrico Fermi

But you can have results that were unexpected, or that had barely achieved *statistical significance*.

Source: Nature 2016

Research reproducibility

Discussing/addressing the issue

Two projects of the *Open Science Framework*:

- Reproducibility Project: Cancer Biology
- Reproducibility Project: Psychology

Some readings:

- Implementing Reproducible Research
- Reproducible research with R and Rstudio

- *Reproducibility in chemistry research* *(literature review)*

EDITORIAL

Nature Methods (2014)

The measure of reproducibility

A clear idea of the performance—the strengths but also the limits—of biological research methods is critical for generating reliable data that others are able to reproduce.

Science & Society (2017)

Science & Society

EMBO reports

The reproducibility "crisis"

Reaction to replication crisis should not stifle innovation

Philip Hunter

EDITORIAL

Science (2018)

Progress on reproducibility

Ideas supported by well-defined and clearly described methods and evidence are one of the cornerstones of science. After several publications indicated that a substantial number of scientific reports may not be readily reproducible, the scientific community and public began engaging in discussions about mechanisms to measure and enhance the reproducibility of scientific projects. In this context, several innovative been utilized and extended in published studies from several other laboratories. This case reinforces the notion that reproducibility, certainly in cancer biology, is quite nuanced, and considerable care must be taken in evaluating both initial reports and reported attempts at extension and replication. Clear description of experimental details is essential to facilitate these efforts. The increased use of preprint servers such as bioRxiv by the biological and

Editor-in-Chief.

ACS Publications
Most Trusted. Most Cited. Most Read.

J. Chem. Inf. Model. 2020

RETURN TO ISSUE | < PREV **EDITORIAL** NEXT >

Editorial: Method and Data Sharing and Reproducibility of Scientific Results

Kenneth M. Merz Jr.*, Rommie Amaro, Zoe Cournia, Matthias Rarey, Thereza Soares, Alexander Tropsha, Habibah A. Wahab, and Renxiao Wang

Research reproducibility (Horror stories)

doi.org/10.1038/nchem.2017

Published: 23 July 2014

Reproducibility

Bruce C. Gibb

Nature Chemistry **6**, 653–654(2014) | Ci

84 Accesses | **4** Citations | **4** Altmetri

Bruce Gibb looks back at some ex
laboratory and suggests ways in v
maximized.

I recently came across a wonderful
the behind-the-scenes story[1] of a pu
first thought – a fast, loose and per
academia. This will be good.

Alas, that part of my brain was sligh
believe he described, a familiar tale
garnering hard cash and able-bodie
equipment purchases; of riding wa
keep going. Reading it was time wel
many personal memories.

Although Deville's blog post did no
hoped, he did briefly discuss one to

www.chemistryworld.com/news/chemistrys-reproducibility-crisis-that-youve-probably-never-heard-of/4011693.article

Taking on chemistry's reproducibility problem

BY DALMEET SINGH CHAWLA | 20 MARCH 2017

Efforts to get to grips with the problem has
technologies are now being brought to be

Not a week passes without reproducibility in scienc
headlines. Although much of the criticism is directe
psychology, many of the same problems also perva

A survey of over 1500 scientists conducted by *Nat*
researchers think that science faces a reproducibilit
faith in published literature in their field – with che
confident despite reporting the most difficulty repl
work. Although this observation seems contradicto
chemists are more often looking to repeat experime
synthetic organic chemist at the Massachusetts Inst

Chemical journal artic
the inability of peo
compounds ac

ANITA BANDROWSKI, UNIVERSITY OF C

Danheiser is the editor-in-chief of the unconventio
verified the experiments of all the papers it has put
journal does this by having the research replicated
publishing them – a practice that is almost unheard
field (the exception being a few brief instances in h
for reproducibility in the lab of one of the journal's
students and postdoctoral researchers working und

doi.org/10.1146/annurev-chembioeng-060718-030323

ANNUAL REVIEWS

Annual Review of Chemical and Biomolecular Engineering

Does Chemical Engineering Research Have a Reproducibility Problem?

Rebecca Han, Krista S. Walton, and David S. Sho

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, A
Georgia 30332–0100, USA; email: david.sholl@chbe.gatech.edu

Annu. Rev. Chem. Biomol. Eng. 2019. 10:43–57

First published as a Review in Advance on March 27, 2019

The *Annual Review of Chemical and Biomolecular Engineering* is online at chembioeng.annualreviews.org

https://doi.org/10.1146/annurev-chembioeng-060718-030323

Keywords

reproducibility, materials chemistry, adsorption, crystal structures, metal-organic frameworks

Abstract

Concerns have been raised in multiple scientific fields in recent ye
the reproducibility of published results. Systematic efforts to exami
sue have been undertaken in biomedicine and psychology, but less
about this important issue in the materials-oriented research that u
much of modern chemical engineering. Here, we relate a dramatic
episode from our own institution to illustrate the implications of
ing reproducible research and describe two case studies based on
analysis to provide concrete information on the reproducibility o
materials-oriented research. The two case studies deal with the pro
metal-organic frameworks (MOFs), a class of materials that have g
tens of thousands of papers. We do not claim that research on MO
(or more) reproducible than other subfields; rather, we argue that
acteristics of this subfield are common to many areas of materials

www.chemistryworld.com/news/taking-on-chemistrys-reproducibility-problem/3006991.article

Computational chemistry faces a coding crisis

BY JAMIE DURRANI | 1 JULY 2020

SOURCE: © ROYAL SOCIETY OF CHEMISTRY; ELEMENTS © SHUTTERSTOCK

In October last year, a team of natural product chemists discovered a glitch in a widely used piece of NMR software. Buried deep inside the code was a simple file sorting issue, which on certain operating systems led to incorrect values being predicted for chemical shifts. The finding cast uncertainty over results published in more than 150 scientific papers over a five year period.

Ten years is a long time in this field in terms of architecture developments, compiler developments, all sorts of developments

LYNN KAMERLIN, UPPSALA UNIVERSITY

This is not the first time that an error in a piece of software code has cast a shadow over computational research, these sorts of issues are actually surprisingly common. In one famous case, a coding error was at the heart of a seven-year dispute between some of the world's top theoretical chemists, who were trying to model the phases of supercooled water. And recently, an algorithm used in older versions of the popular molecular dynamics software Gromacs was found to introduce order of magnitude mistakes during simulations.

Ideally, code will be well documented and publicly available, allowing researchers to

Alain Borel, Francesco Varrato

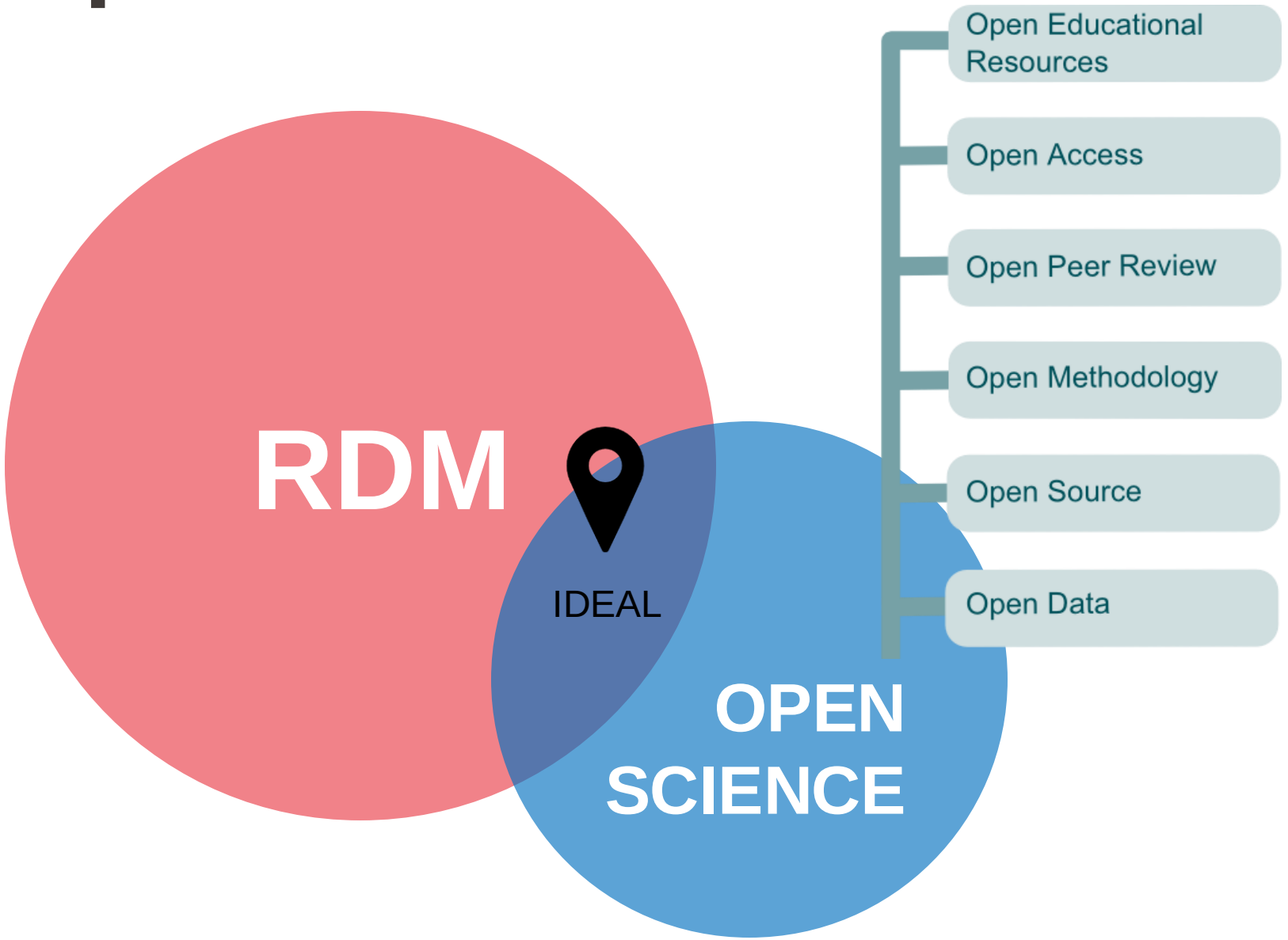
Research reproducibility: some recommendations

analytical chemistry | **Reproducibility and Replicability** >

1. Are the data and analysis laid out with sufficient transparency and clarity that the results can be checked?
2. If checked, do the data and analysis offered in support of the result in fact support that result?
3. If the data and analysis are shown to support the original result, can the result reported be found again in the specific study context investigated?
4. Finally, can the result reported or the inference drawn be found again in a broader set of study contexts?

Source: https://doi.org/10.1021/acs.analchem.9b02719

RDM ⇔ Open Science?



Good data management practices apply also to **non**-Open Data and **non-**Open-Source research

Openness

“As open as *necessary*, as restricted as *possible*”

True or False?

RDM policies, guidelines ...

Institutional policies

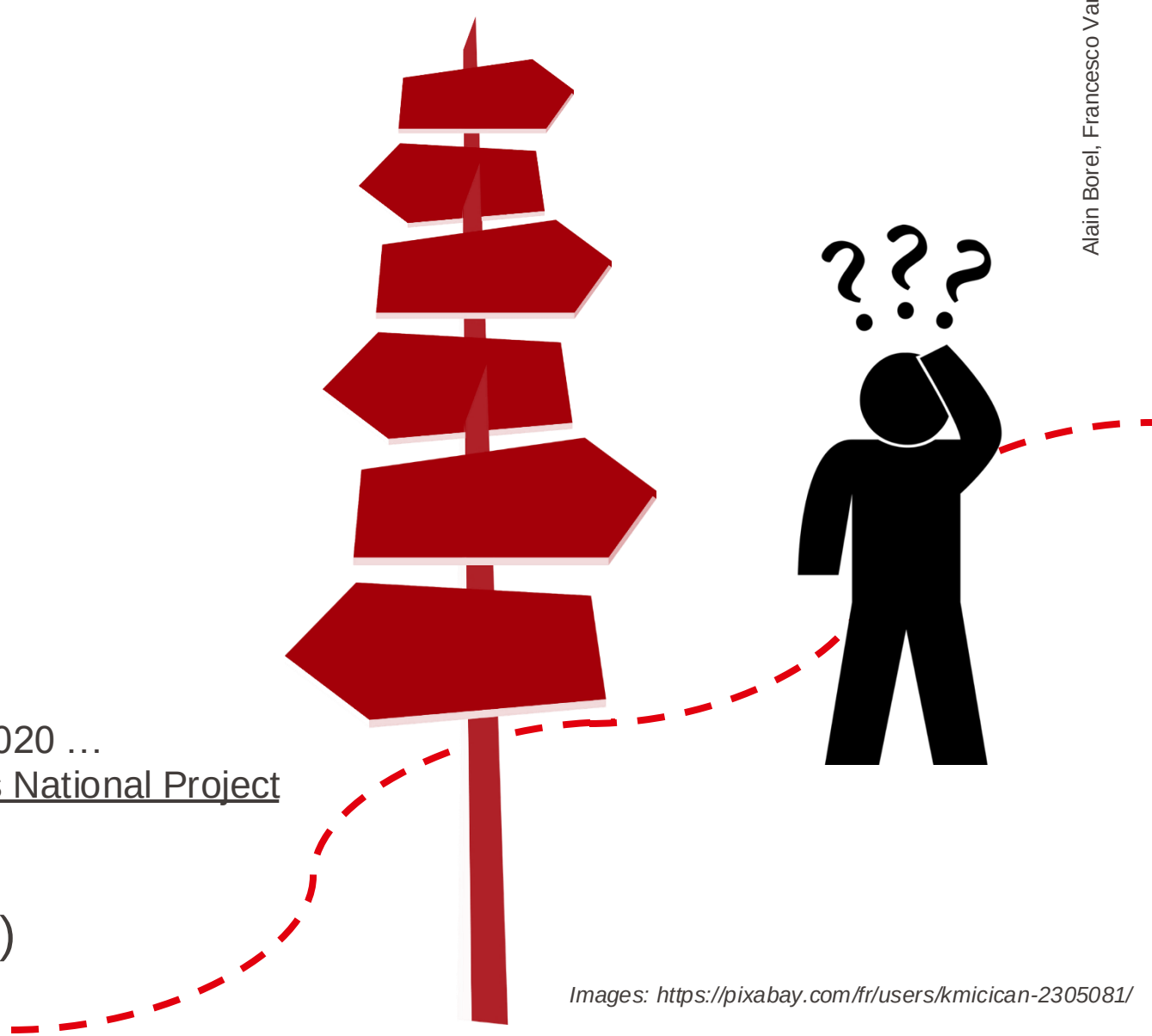
- ETHZ Guidelines for RDM
- Humboldt-Universität zu Berlin
- MIT
- TU Delft
- UNIGE
- University of Cambridge
- University of Edinburgh
- University of Oxford

General guidelines

- SNSF Open Research Data policy
- EC Data Management manual for Horizon 2020 …
- Digital Lifecycle Management (DLCM) Swiss National Project

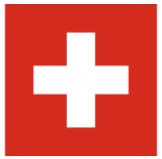
EPFL compliance guide 2019 (p.36)

No unique Data Policy



Images: https://pixabay.com/fr/users/kmician-2305081/

Funders' requirement



SNSF (Ambizione, NCCR, …)

and SNSF ERC "replacement grants"

- Researchers must share (at least) the data underlying their publications
- Mandatory DMP to obtain funding

ERC (MSCA, Horizon Europe 2021-2027, ...)

- The research data is **open by default** (also metadata)
- Mandatory DMP to obtain funding

SERI guide: DMP as PDF on same platform as other project docs

+

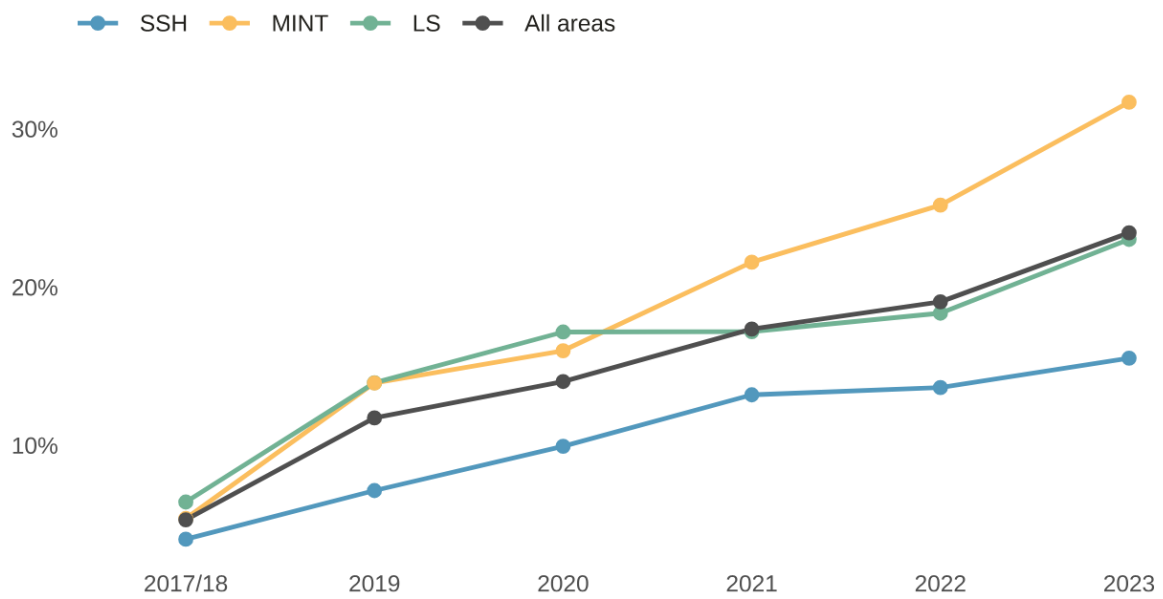
EPFL (internal projects) + **U.S. Federal Grants** + **Wellcome Trust** + **Defitech** + **AXA** Research Fund + **Chinese government** + **Ligue contre le cancer** + …

Open Research Data

Alain Borel, Francesco Varrato

The share of completed grants that declare a dataset is increasing

Interactive content



The year refers to the date when the grant ended. The number of grants each year in this dataset was: 2694 in 2017/2018, 1975 in 2019, 1523 in 2020, 2050 in 2021, 1781 in 2022, 1548 in 2023. 2017 only includes grants that ended after the new ORD regulations were put in place (October 2017), and is therefore combined with 2018.

SNSF report 2017-2018

- ***16%** applicants requested ORD funds*
- ***0.2%** annual costs budgeted for ORD*
- ***21%** applications budgeted > 10k CHF*
- ***45% did not mention a data repository***

SNSF news August 2024

- **~30% of "hard science" grants completed in 2023 have published at least 1 dataset**

https://data.snf.ch/stories/open-research-data-2023-en.html

Alain Borel, Francesco Varrato

Publishers' requirement

Many journals require authors to
publish the data underlying the published results

- Examples of journals policies on data / code publication
- A list of journal open-data policies
- A list of Publisher Data Availability Policies (See: ACS, RSC)

Not just supplementary material.

It's an academic output!

~

OPEN
DATA

Open Data logo by the EPFL Research Data Library Team:
- *https://pixabay.com/fr/donn%C3%A9es-ouvertes-base-de-donn%C3%A9es-1518223/*
- *Open Sans: https://fonts.google.com/specimen/Open+Sans?selection.family=Open+Sans*

Publishers' requirement: Examples

Science
All data used in the analysis must be available to any researcher for purposes of reproducing or extending the analysis. Data must be available in the paper or deposited in a community special-purpose repository or a general-purpose repository [source]

SpringerNature
[…] authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission [source]

ACS
Where ethically and legally feasible, all ACS journals strongly encourage authors to make the research data underlying their articles publicly available at the time of publication […] To ensure data accessibility, we encourage the use of open licenses for reuse of data, such as Creative Commons CC BY [source]

… etc.

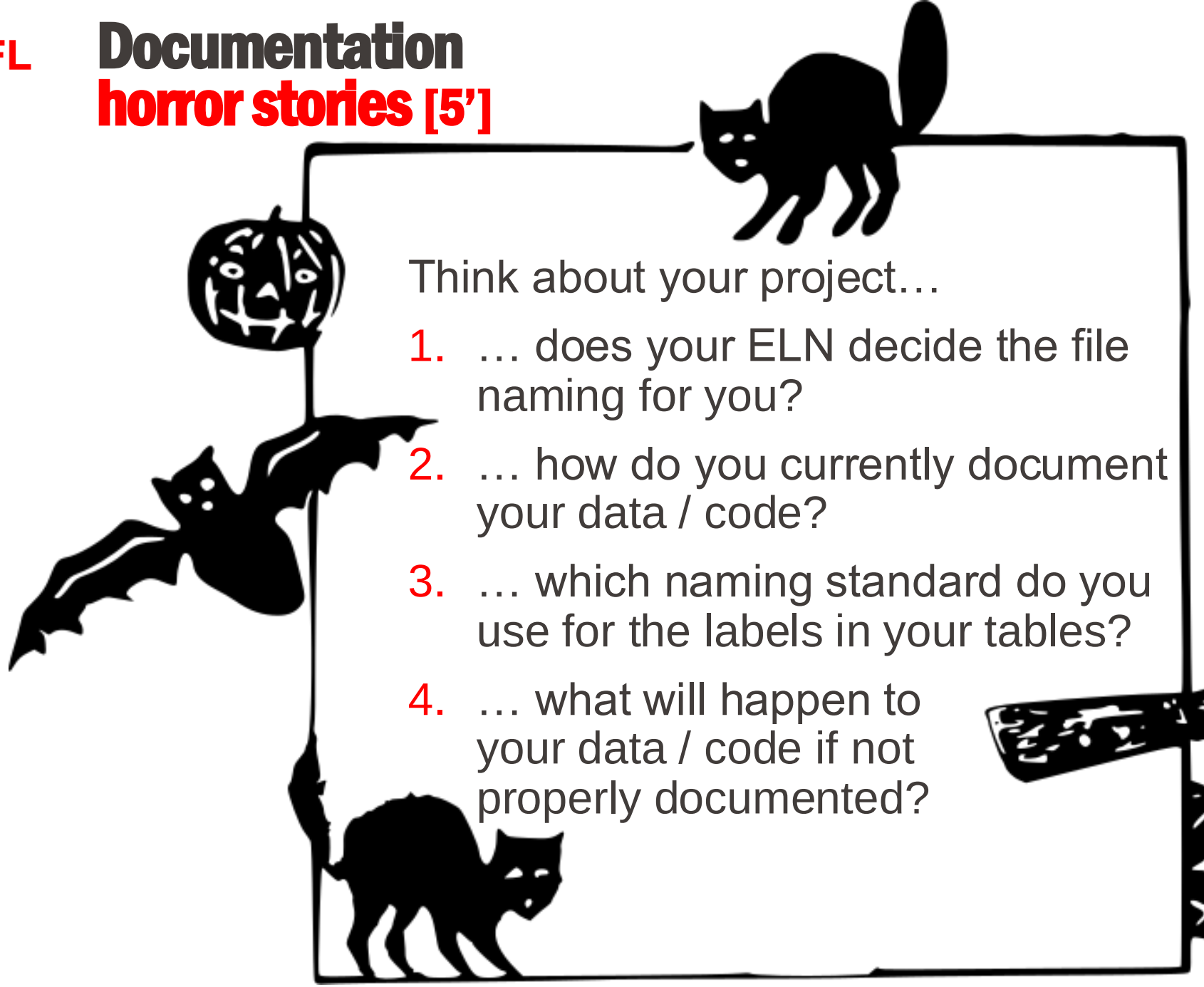
Alain Borel, Francesco Varrato

FAIR data vs. Supplementary Information

Alain Borel, Francesco Varrato

	S.I.	
	human ?	machine ?
▪ ***F* indable** Data and metadata are easy to find by both humans & computers.	✓	✗
▪ ***A* ccessible** Machines & humans can readily access or download (meta)data.	✓	±
▪ ***I* nteroperable** Data from different datasets are ready to be exchanged or combined.	±	✗
▪ ***R* eusable** (Meta)data are easily replicated / combined in future research.	±	✗

Documentation horror stories [5']



Think about your project…

1. … does your ELN decide the file naming for you?
2. … how do you currently document your data / code?
3. … which naming standard do you use for the labels in your tables?
4. … what will happen to your data / code if not properly documented?

Alain Borel, Francesco Varrato

Alain Borel, Francesco Varrato

Documentation

WHAT?

- **Description** of your data / code
- **Planned** before starting the data collection

HOW?

- Project-dependent
- **Consistent** documentation methodology

WHY?

- Data more understandable **for yourself**
- Data more understandable **for others**
- Saves time upon publication

Ways to document data

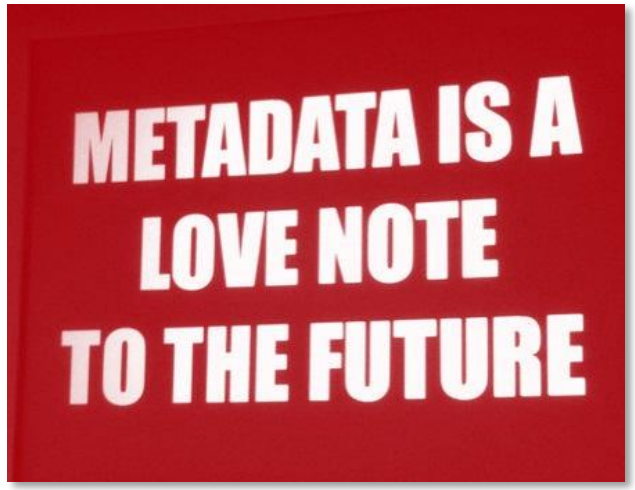
Alain Borel, Francesco Varrato

Documentation

- **README** files
- Laboratory **notebooks**
- Experimental **protocols**
- Software **parameters**
- **Output / Log** files
- **DMP**
- **RDM Strategy**

Metadata

- **In-file** metadata (ex. *.docx* author, creation date, file tagging,…)
- **Naming** convention (for files & folders & database objects)
- Folders **structure**
- Database **schema**
- **Version**
- Data **dictionaries**
- **Codebooks**
- Metadata **standards**
- Metadata **vocabularies**
- **Discovery** metadata (ex. publication keywords)



Importance of metadata (!)

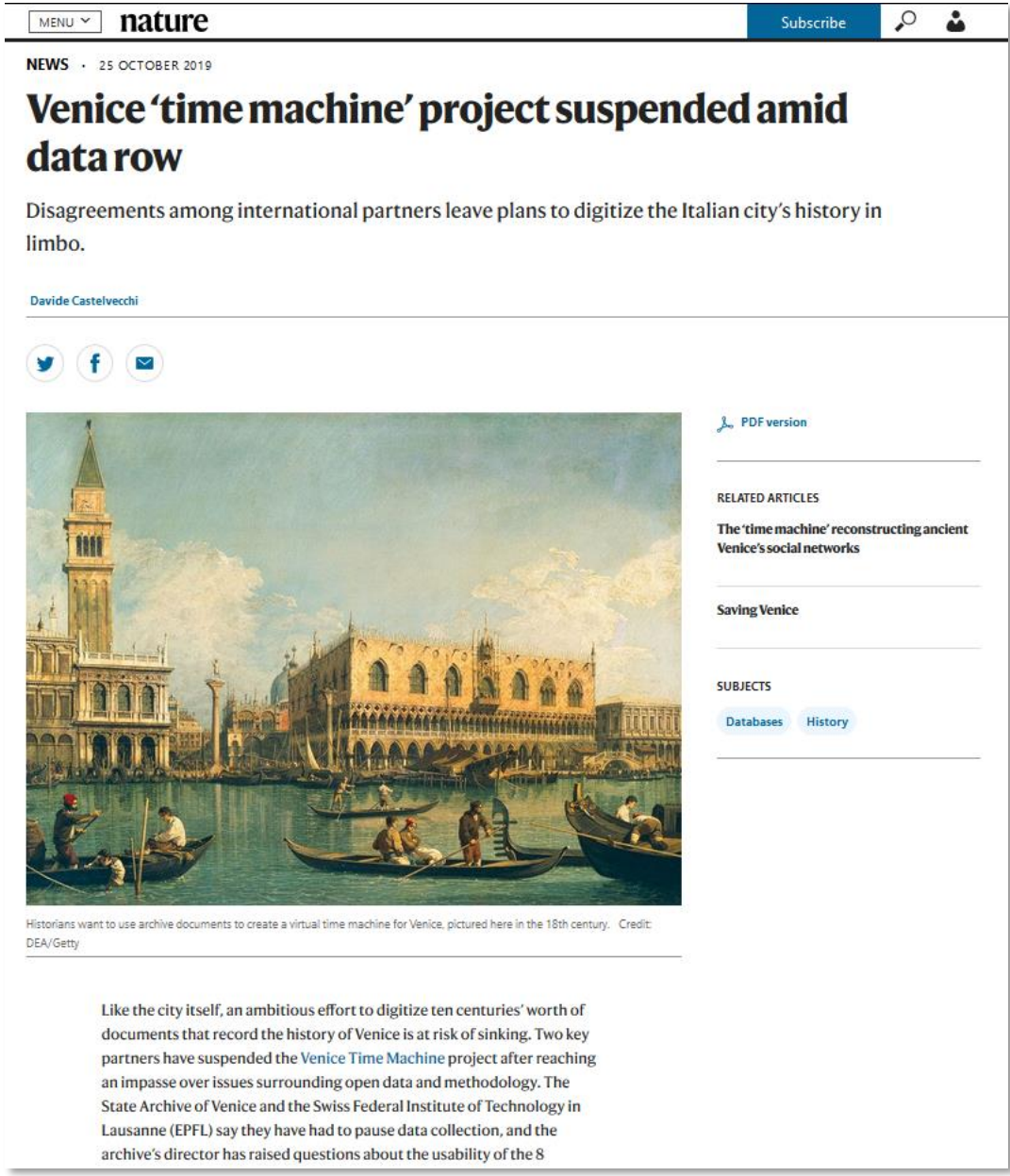
2012 – Project of officially **launched**:
Venice's State Archive + Ca' Foscari Univ. + EPFL (DHLAB)

2014 – Non-binding agreement signed. But … didn't specify the licensing that would regulate researchers' use of the digitized data

2017 – At stake: 1,000 years of records in dynamic digital form: special high-speed scanners, thousands HD images per hour

2019 – Allegedly, the digitization of ~190,000 documents (**8 TB**) **didn't follow a common metadata policy**: archival-science guidelines (require records of provenance for each document)

2019 – … **data collection has been paused, amid doubts on the usability of the data already collected!**



MENU nature Subscribe

NEWS · 25 OCTOBER 2019

Venice 'time machine' project suspended amid data row

Disagreements among international partners leave plans to digitize the Italian city's history in limbo.

Davide Castelvecchi

PDF version

RELATED ARTICLES

The 'time machine' reconstructing ancient Venice's social networks

Saving Venice

SUBJECTS

Databases History

Historians want to use archive documents to create a virtual time machine for Venice, pictured here in the 18th century. Credit: DEA/Getty

Like the city itself, an ambitious effort to digitize ten centuries' worth of documents that record the history of Venice is at risk of sinking. Two key partners have suspended the Venice Time Machine project after reaching an impasse over issues surrounding open data and methodology. The State Archive of Venice and the Swiss Federal Institute of Technology in Lausanne (EPFL) say they have had to pause data collection, and the archive's director has raised questions about the usability of the 8

DOI: 10.1038/d41586-019-03240-w

DMP as documentation: What is it?

Data **M**anagement **P**lan

Living document & Research **roadmap**

Describes

- strategy to manage data
- actions to take
- needed resources

(time, money, people, tools)

(Minimum) Content of a DMP

- Institution and contacts
- Data collection and documentation
- Ethics, legal and security issues
- Resources and responsibilities
- Data storage and preservation
- Data sharing and reuse

What is a DMP?

Data **M**anagement **P**lan

Living document & Research **roadmap**

Do you/your thesis supervisor(s) have a DMP for your project?

Alain Borel, Francesco Varrato

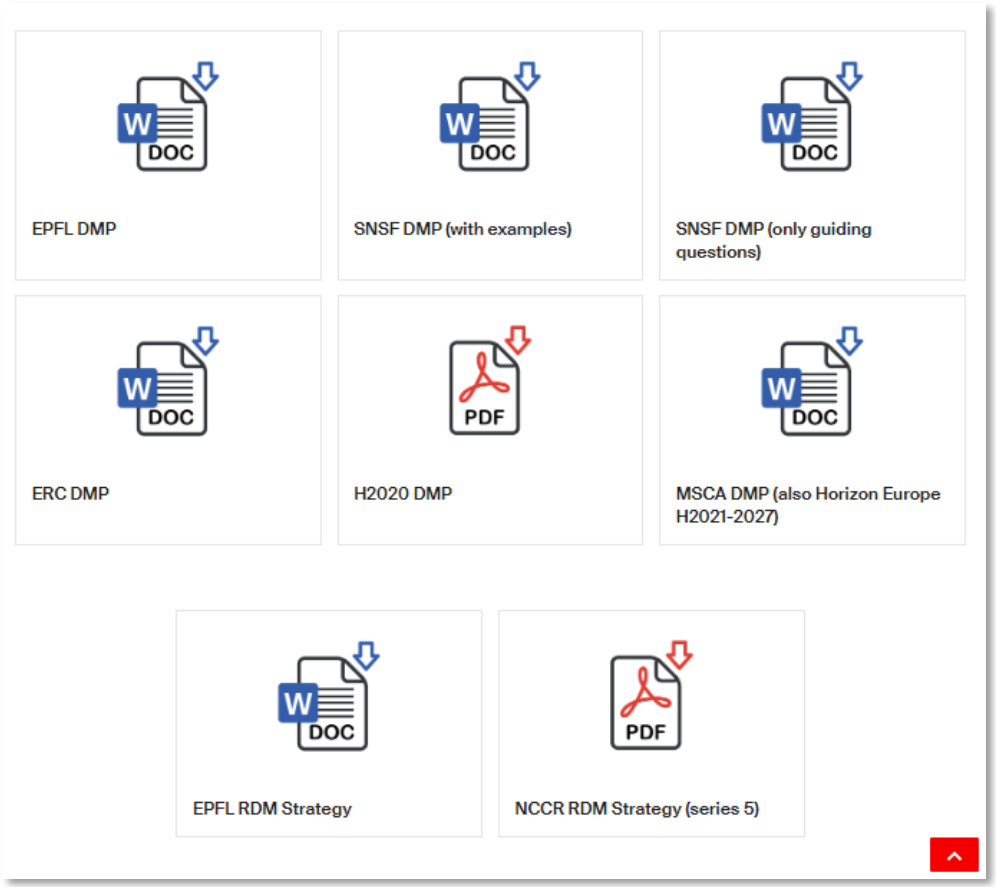
Why a DMP?

- **Plan**: future needs (material, software, HR ...)
- **Science**: impact, better reproducibility, posterity
- **Data reuse**: better use of public funds
- **Openness**: impact, transparency, accountability
- **Visibility**: citations, collaborations, career
- **Compliancy**: law (ex. GDPR), funders (ex. SNSF)
- **Efficiency**: faster research for your lab and beyond
- **Risk reduction**: data loss, privacy, patents, …
- **Modernity**: world-scale digital research, big data
- …

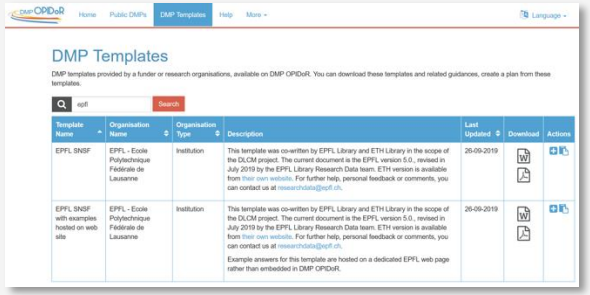
Alain Borel, Francesco Varrato

Some resources

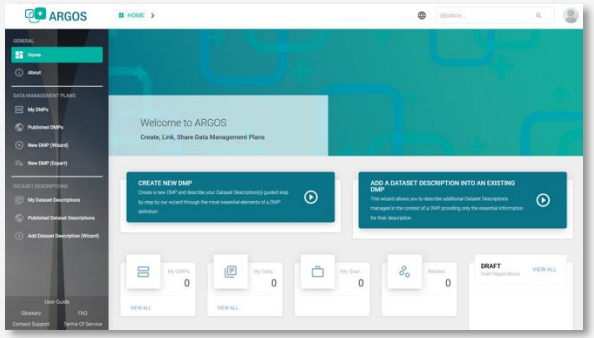
go.epfl.ch/rdm-guide



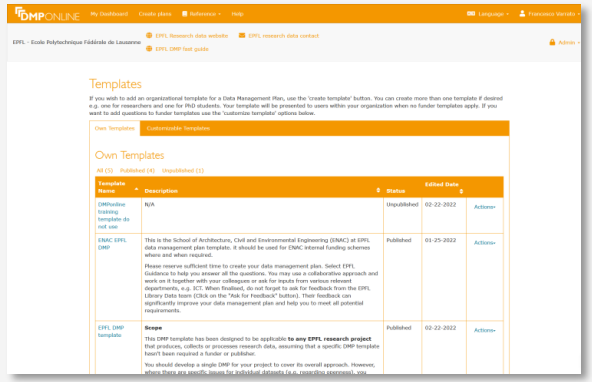
Other resources



dmp.opidor.fr/public_templates



argos.openaire.eu

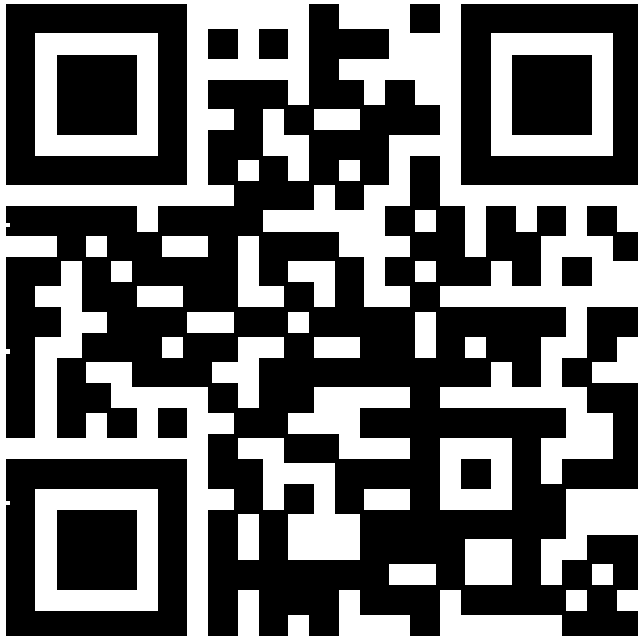


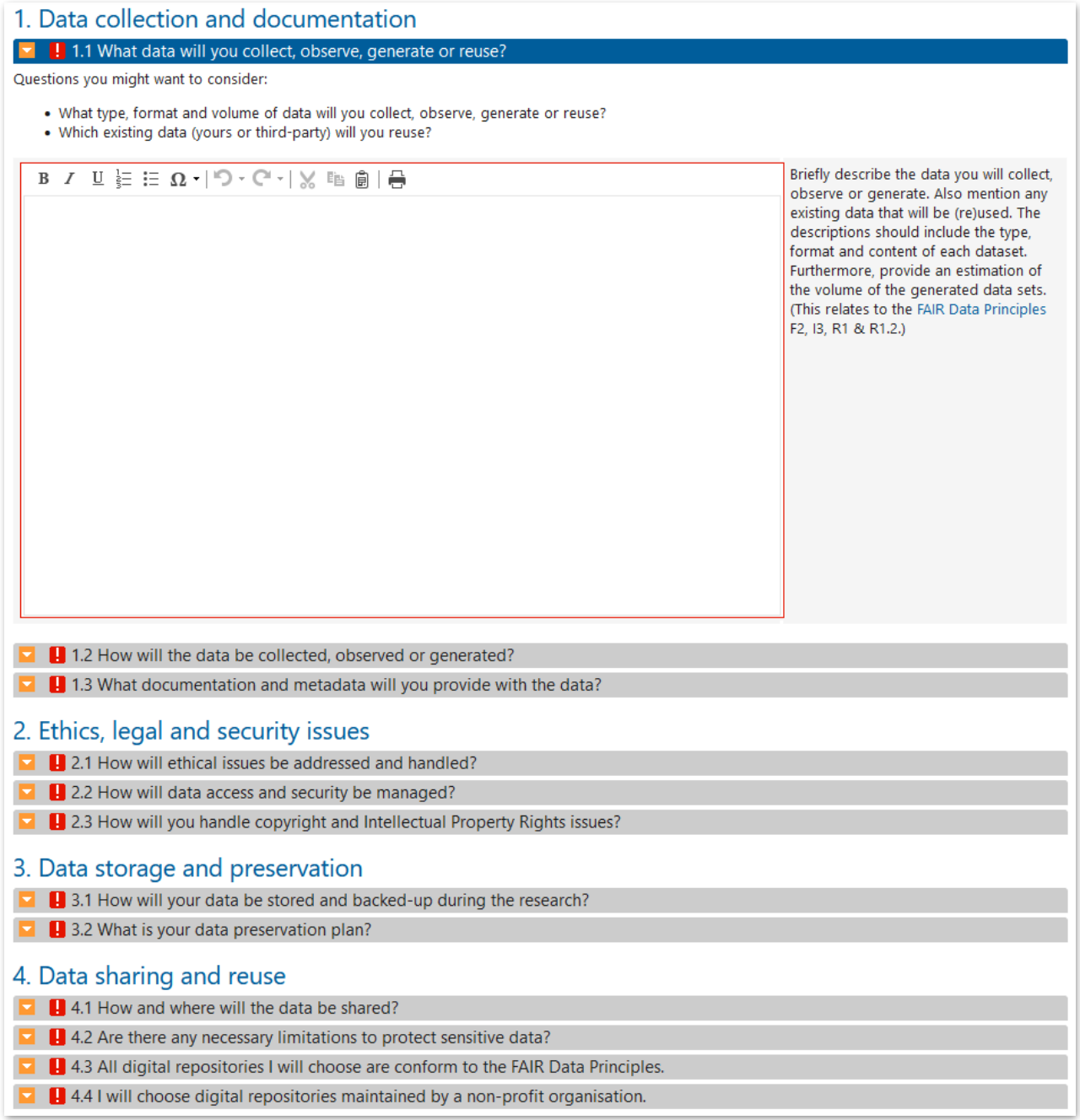
dmponline.dcc.ac.uk

DMP example: SNSF

- **Online** form
- Accessible from **PI** (Principal Investigator)
- **Can be modified** at each moment
- Should be **updated** when funding ends

go.epfl.ch/**dmp-snsf**





Exercise [~15']: SNSF DMP

Alain Borel, Francesco Varrato

[5'] Write section *1.1 of DMP* (about your actual project)
[5'] Live peer-review (by couples)
[5'] Common feedback

go.epfl.ch/**dmp-snsf**

1.1 What data will you collect, observe, generate or re-use?

- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?

Data categorization: Example

Following data are generated in order to investigate neural processing that produce behavior.

[A] New experimental data (format, size)

1.	Cortical Imaging Data:	.mat, 20 TB
2.	Behavioral Filming Data 1:	.tif, 20 TB
3.	Behavioral Filming Data 2:	.avi, 10 TB
4.	Behavioral Task Data:	.txt, 500 GB
5.	Behavioral Log Data:	.bin, 1 TB
6.	Optical Control Data:	.txt, 1 TB
7.	Experiment Log:	.xlsx, 1 GB
8.	Histology Data:	.tiff, 1 TB
9.	Electrophysiology Data:	.mat, 5 TB

[B] Analyzed data (format, reuse, origin, size)

1.	Imaging Data (A1):	.mat, 10 TB
2.	Filming Data (A2, A3):	.bin, 5 TB
3.	Filming Data (A2, A3):	.mat, 5 TB
4.	Behavioral Data (A4, A5):	.mat, 500 GB
5.	Histology Data (A8):	.mat, 500 GB
6.	EPhys Data (A9):	.mat, 500 GB

Source: DMP draft by Keita Tamura, Marie Curie fellowship application

RDM strategy

DMP is for 1 project

RDM Strategy is for > 1 projects

- Personal RDM strategy
- Research group strategy
- Research collaboration strategy

DMP templates:

- go.epfl.ch/rdm-guide
- argos.openaire.eu

RDM Strategy templates:

- NCCR RDM Strategy (SNSF)
- RDM Strategy (EPFL)

Alain Borel, Francesco Varrato

README as documentation

A README provides info about data file(s) and enables reusability

README content:

- General information
- Data and file overview
- Sharing and access information
- Methodological information

Best practices

- Write the README as a plain **text file** (open format)
- Follow existing conventions…. Smartly ☺
- 1 README per data **folder** (whenever possible)
- Name the README in accordance with described **files**
- Use the same **template** for multiple READMEs
- Use standardized **date** formats [*W3C/ISO 8601 date standard*]:

  YYYY-MM-DD or YYYY-MM-DD-hh:mm:ss
- Write for human readers (does not replace metadata)

README vademecum and template

go.epfl.ch/rdm-readme

Alain Borel, Francesco Varrato

README example/template (* recommended)

Alain Borel, Francesco Varrato

Dataset title

General information or Introduction section

- author(s) info (name, affiliation, persistent id) *
- date of collection *
- geolocation data
- funding or sponsorship info *

Sharing / Access information / License

- licenses *
- terms of use *
- citation instructions *
- links to related publication(s)
- links to related datasets or code
- url in repository
- persistent identifiers

Data and file(s) overview

- files and folders structure description *
- file formats *
- additional related data
- data dictionary or data codebook *
- original source if any
- dataset version, update description/changelog

Methodological info / Preparation

- link to publications used as base for methods
- methods for processing data *
- technical requirements: instruments, software, parameters, calibration data *
- people involved in experiments, surveys, analysis
- quality assurance process applied
- standards applied

More details: https://go.epfl.ch/rdm-readme-template

Alain Borel, Francesco Varrato

Data or Metadata?

DATA = Input Values

METADATA = Info about data

Metadata

Alain Borel, Francesco Varrato

- **Is** typed and formatted (not just free text)
- **Is** both machine-readable AND human-readable

Metadata is **structured** information **associated** with an object for **purposes** of discovery, description, use, management, and preservation.

NISO (2008) framework.niso.org/24.html

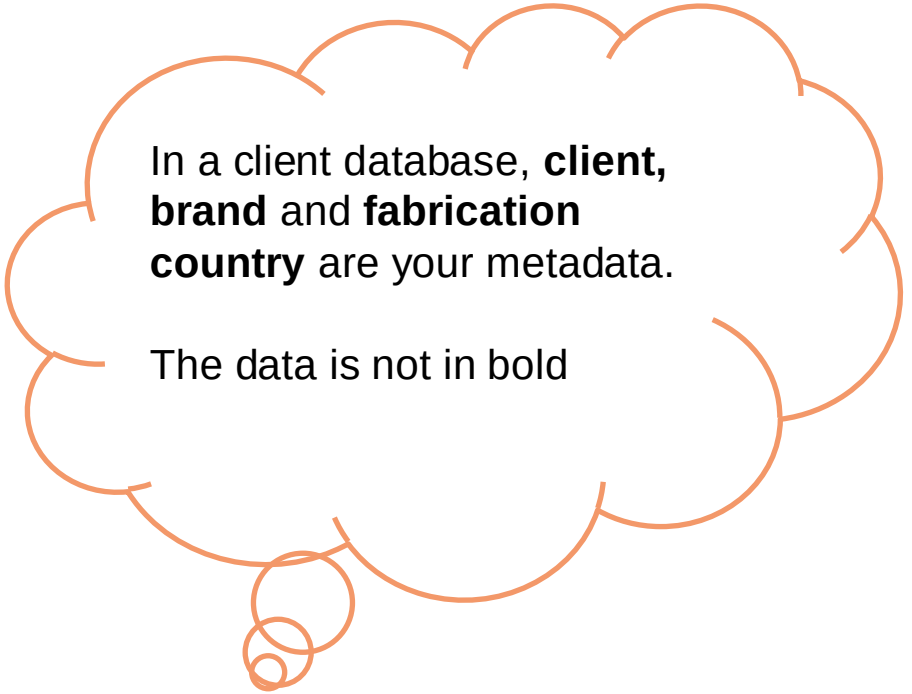
- Ex. associated to research data and code
- Supporting the research data **lifecycle**

(*) Taken from Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., … Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology, 25(8), 887-893. https://doi.org/10.1038/nbt1329

Data or Metadata?

Metadata: Data

- ***Brand*: Sainsbury's**
- ***Fabrication Country*: Italy**
- ***Fat in grammes*: 1.1**
- ***Pasta Type*: Fusilli**
- ***Cooking Time in minutes*: 12**



Client	Brand	Fabrication Country
Sainsbury's	Sainsbury's	Italy
Coop	Barilla	Italy

Metadata Example

- Title, author, date, DOI, format, version, …

Publication date:
August 7, 2018

DOI:
DOI 10.5281/zenodo.1345472

Keyword(s):
Fixed point, complete metric space, set-valued mapping, G – metric space

Published in:
RESEARCH REVIEW International Journal of Multidisciplinary: 03 pp. 309-313.

License (for files):
Creative Commons Attribution 4.0

Human readable

- Info stored in repository's internal database

Machine readable

Alain Borel, Francesco Varrato

Data dictionaries / Codebooks

Explain variables used in a dataset, within a table

Sheet_1

Source: https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Example of content:

variable name, variable label, variable definition, units of measure, allowed ranges, value code, missing data, etc.

Discover more on how to **Create a Codebook** on the Data Documentation Initiative (DDI) Alliance website.

Metadata standards

Dataset description (metadata)

▪ Identifier ▪ Title ▪ Creator ▪ Subject ▪ Description	▪ Publisher ▪ Date ▪ Formats ▪ Rights ▪ …

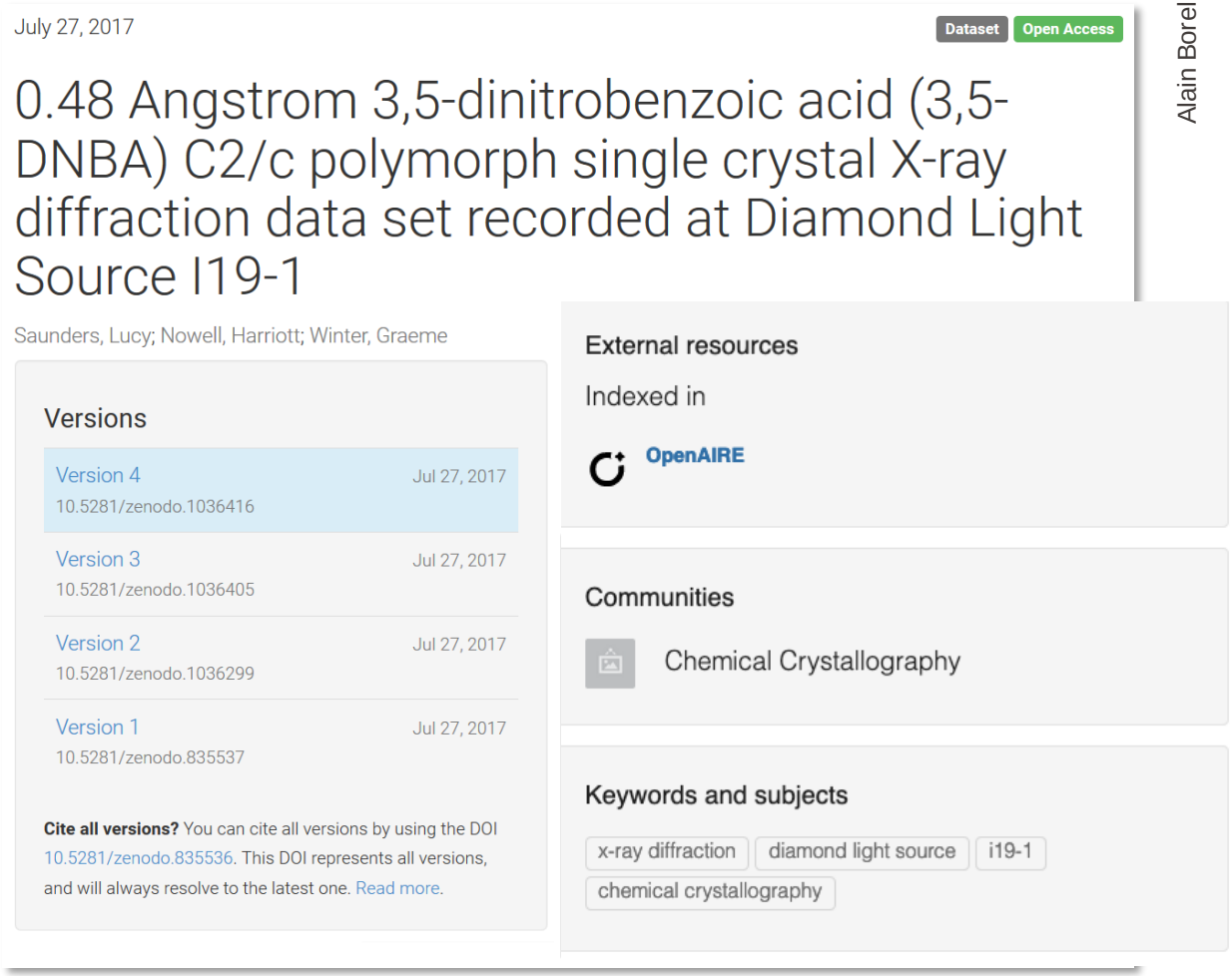
Basic & General schemas

- Dublin Core
- DataCite
- …

Disciplinary schemas

- Digital Curation Centre
- Linked Open Vocabularies (LOV)
- Fairsharing
- …

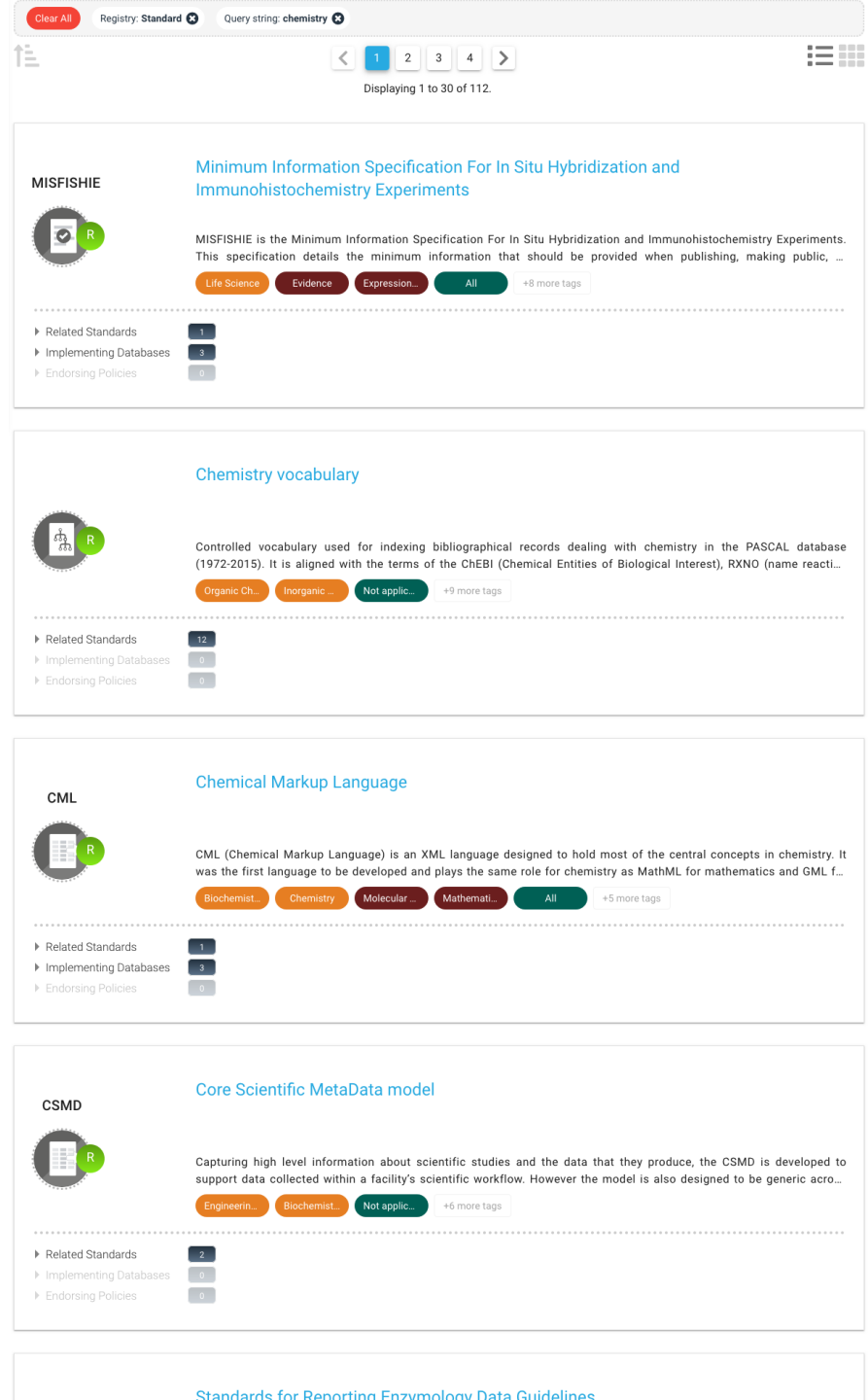
Data Repository integration (example)



Source: https://doi.org/10.5281/zenodo.835536

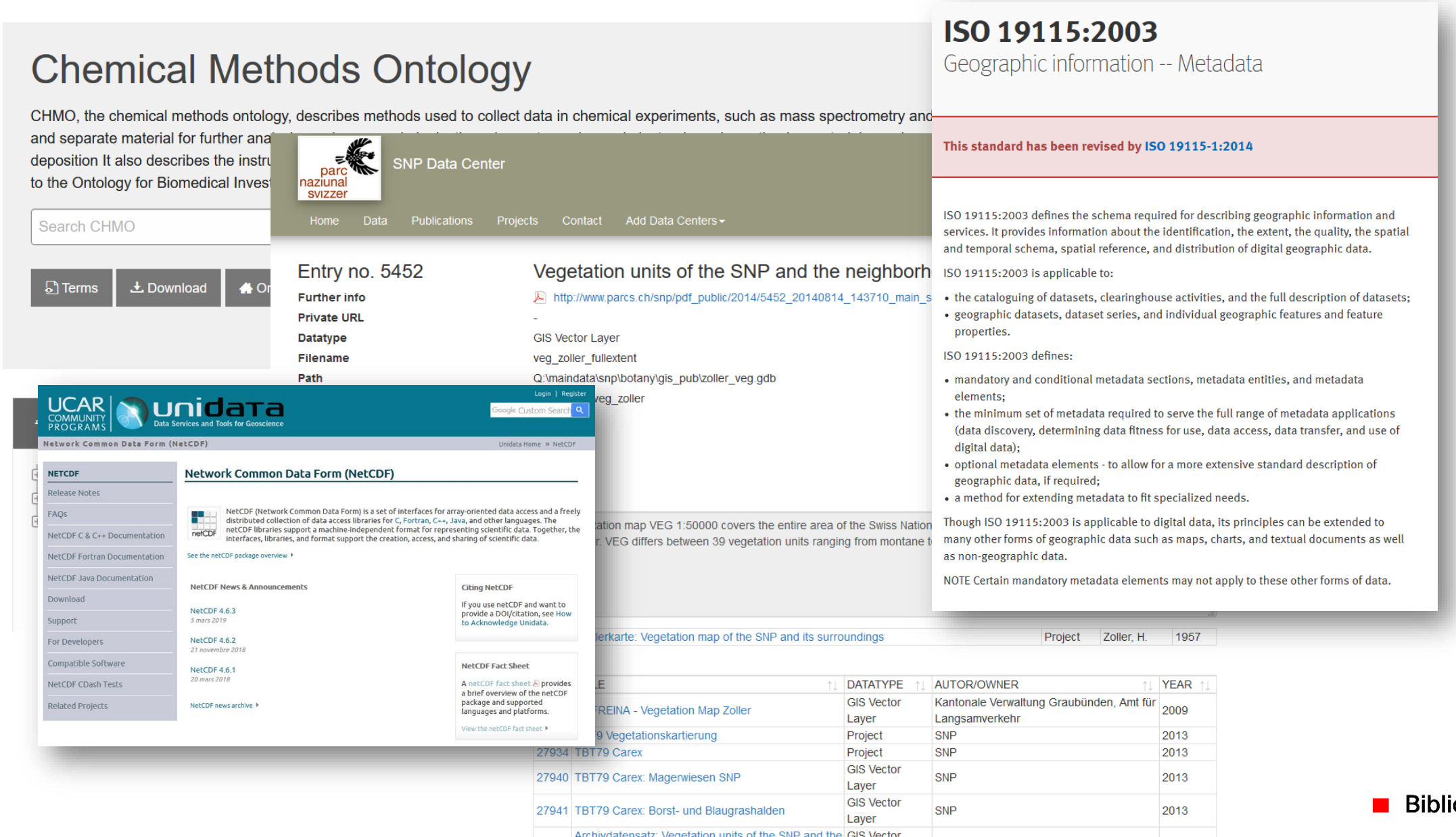
Metadata standards

FAIRsharing.org
standards, databases, policies



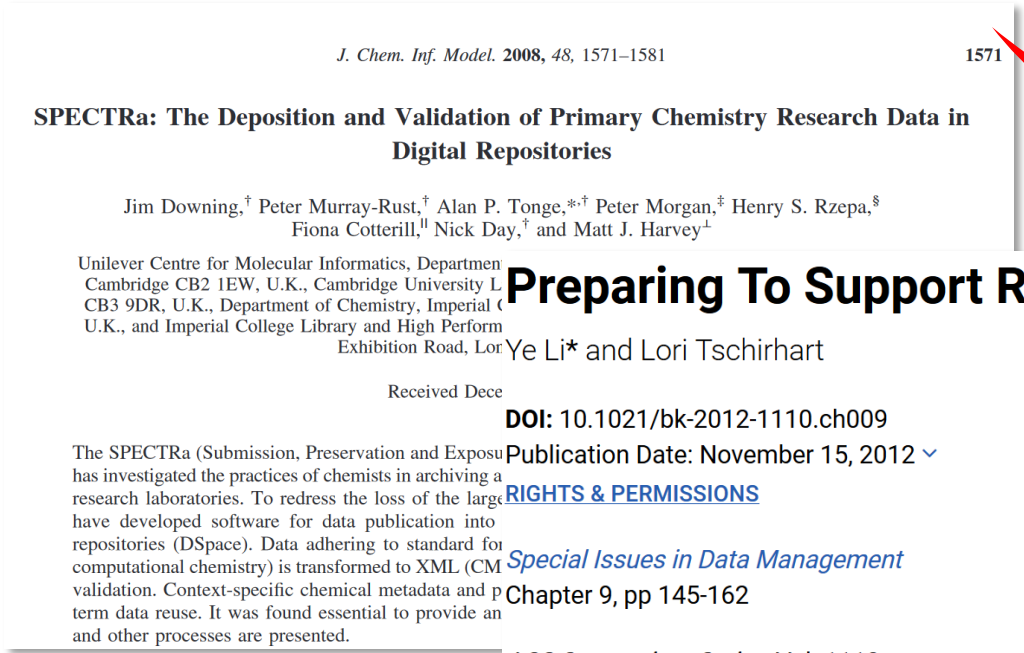
Metadata standards usage (examples)

Alain Borel, Francesco Varrato



Standards: From awareness to practice

Alain Borel, Francesco Varrato



J. Chem. Inf. Model. 2008, 48, 1571–1581 1571

SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories

Jim Downing,† Peter Murray-Rust,† Alan P. Tonge,*,† Peter Morgan,‡ Henry S. Rzepa,§ Fiona Cotterill,|| Nick Day,† and Matt J. Harvey⊥

Unilever Centre for Molecular Informatics, Departmen
Cambridge CB2 1EW, U.K., Cambridge University L
CB3 9DR, U.K., Department of Chemistry, Imperial
U.K., and Imperial College Library and High Perform
Exhibition Road, Lor

Received Dece

The SPECTRa (Submission, Preservation and Exposu
has investigated the practices of chemists in archiving a
research laboratories. To redress the loss of the large
have developed software for data publication into
repositories (DSpace). Data adhering to standard for
computational chemistry) is transformed to XML (CM
validation. Context-specific chemical metadata and p
term data reuse. It was found essential to provide an
and other processes are presented.

doi.org/10.1021/ci7004737

Preparing To Support Research Data Sharing

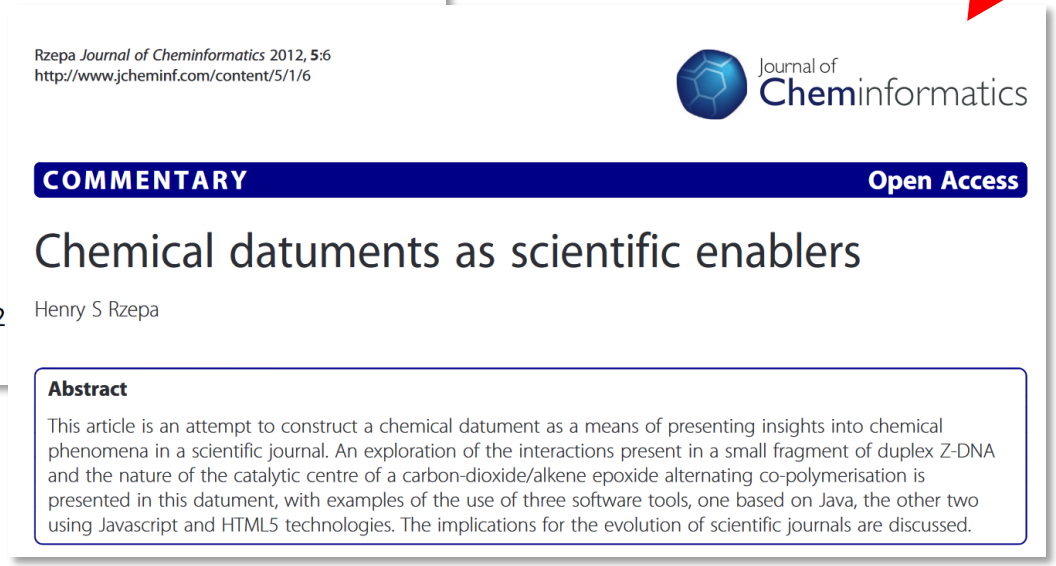
Ye Li* and Lori Tschirhart

DOI: 10.1021/bk-2012-1110.ch009
Publication Date: November 15, 2012

Special Issues in Data Management
Chapter 9, pp 145-162

ACS Symposium Series, Vol. 1110
ISBN13: 9780841227125 **eISBN:** 9780841227132

doi.org/10.1021/bk-2012-1110.ch009



Rzepa *Journal of Cheminformatics* 2012, **5**:6
http://www.jcheminf.com/content/5/1/6

Journal of Cheminformatics

COMMENTARY **Open Access**

Chemical datuments as scientific enablers

Henry S Rzepa

Abstract

This article is an attempt to construct a chemical datument as a means of presenting insights into chemical phenomena in a scientific journal. An exploration of the interactions present in a small fragment of duplex Z-DNA and the nature of the catalytic centre of a carbon-dioxide/alkene epoxide alternating co-polymerisation is presented in this datument, with examples of the use of three software tools, one based on Java, the other two using Javascript and HTML5 technologies. The implications for the evolution of scientific journals are discussed.

doi.org/10.1186/1758-2946-5-6

Ex.: "*Datument […] refers to a data-rich document […] that describes a story of chemical research in a manner which allows the data underpinning the discourse to be provided as an integral part of that story*."

Files or Database?

File management

- File / folder organization
- File / folder naming
- File / folder versioning system
- File / folder access rights management

Database management

- Data model / Data dictionary
- Metadata design / standards
- Administrative data / logs
- User rights management
- Database administrator

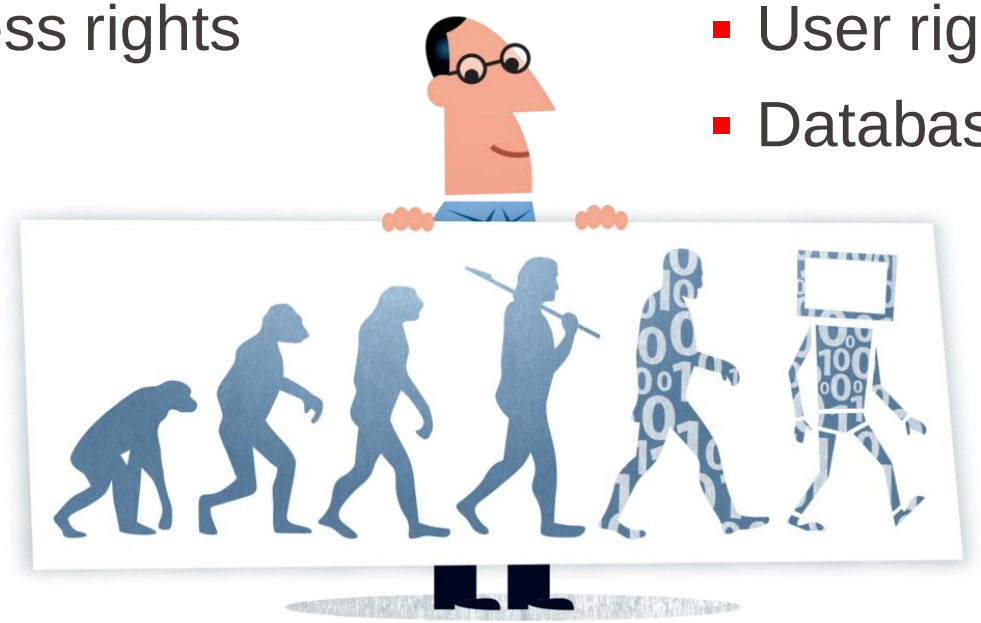


Image source: Digitalbevaring.dk (CC BY 2.5 DK)

Alain Borel, Francesco Varrato

Dataset or Database?

Database:
Organized collection of data stored as multiple datasets

Dataset:
Data+Code+Metadata associated with a unique body of work

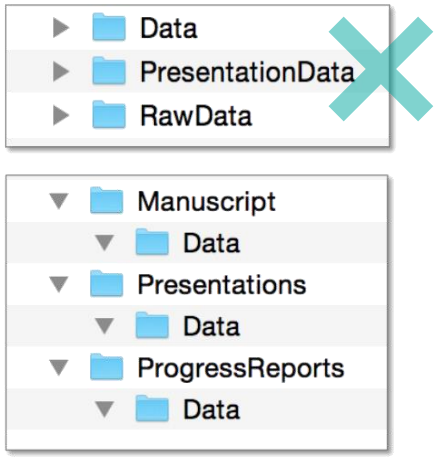
Definition source: U.S.G.S.

Alain Borel, Francesco Varrato

Dataset organization

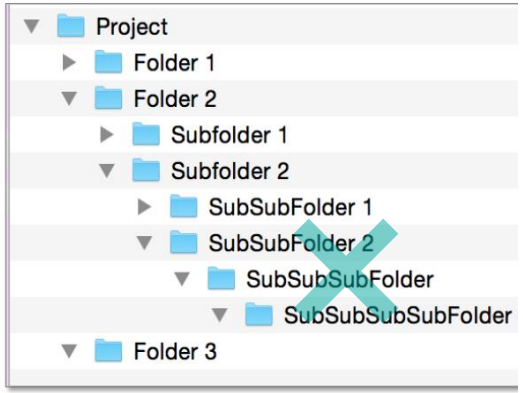
Try to **avoid** …

overlapping categories



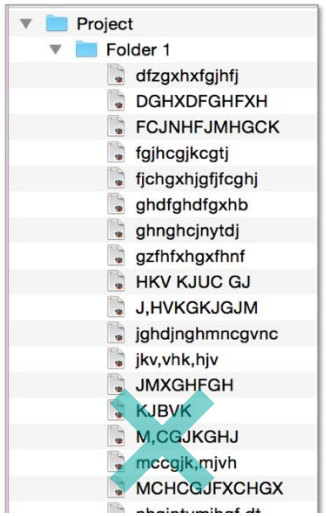
Rule of thumb:
"***sure of the right subdirectory***"

too deep structures



Rule of thumb:
"***no more than 3 clicks***"

too crowded folders



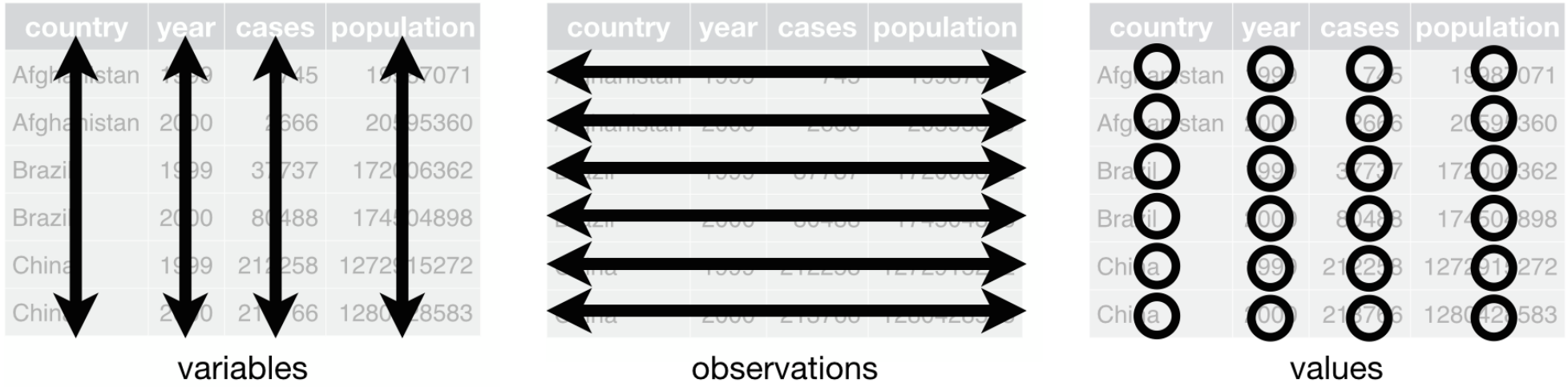
Rule of thumb:
"***fit in one screen***"

Check out:
https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming
https://libraries.mit.edu/data-management/files/2014/05/file-organization-july2014.pdf

Data in spreadsheets?

Tidy format

- Always prefer plain-text formats
- Don’t mix data and analysis
- Don’t use color coding or other features that don’t translate to plain text



Data validation service goodtables.io

Alain Borel, Francesco Varrato

Example : Organization starter template

Alain Borel, Francesco Varrato

RO = Read-only, never changes
HW = Human writable, you make changes here
PG = Project generated, your analysis script that update code here

```
├── .gitignore
├── CITATION.md
├── LICENSE.md
├── README.md
├── requirements.txt
├── bin                 <- Compiled and external code, ignored by git (PG)
│   └── external        <- Any external source code, ignored by git (RO)
├── config              <- Configuration files (HW)
├── data                <- All project data, ignored by git
│   ├── processed       <- The final, canonical data sets for modeling. (PG)
│   ├── raw             <- The original, immutable data dump. (RO)
│   └── temp            <- Intermediate data that has been transformed. (PG)
├── docs                <- Documentation notebook for users (HW)
│   ├── manuscript      <- Manuscript source, e.g., LaTeX, Markdown, etc. (HW)
│   └── reports         <- Other project reports and notebooks (e.g. Jupyter, .Rmd) (HW)
├── results
│   ├── figures         <- Figures for the manuscript or reports (PG)
│   └── output          <- Other output for the manuscript or reports (PG)
└── src                 <- Source code for this project (HW)
```

Check out Starter template 1 or Starter template 2

Importance of naming (!)

2014 – "*Room-Temperature Cu(II)-Catalyzed Chemo- and Regioselective Ortho-Nitration of Arenes via C–H Functionalization*" gets published, co-signed by three chemists

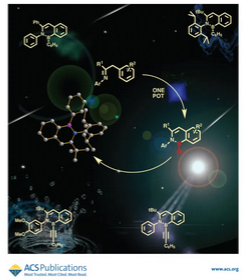
2015 – Article retracted (and resubmitted!) by its authors

- incorrect files used in preparing the published paper
- the student kept similar files names for both the updated and un-updated files
- by mistake, the student provided the un-updated file (incorrect file) that led to error in the yield

Moral (?) – It's the student's fault … or is it?

Chem paper fails to catalyze when wrong files are "inadvertently used"

Three chemists at the Indian Institute of Technology Guwahati in India have retracted a paper from the *Journal of Organic Chemistry* because the "incorrect files were inadvertently used."



The article, "Room-Temperature Cu(II)-Catalyzed Chemo- and Regioselective Ortho-Nitration of Arenes via C–H Functionalization," described a protocol to perform nitration — the addition of nitro groups onto an organic compound — using an inexpensive copper catalyst.

All three authors signed the one-sentence notice:

> *This article was retracted by the authors when it was discovered that incorrect files were inadvertently used in the preparation of the published paper.*

retractionwatch.com/2015/04/29/chem-paper-fails-to-catalyze-when-wrong-files-are-inadvertently-used/

Alain Borel, Francesco Varrato

Discussion: File naming

1. **6 months from now**, will you recognize what your files contain?
2. **What information** needs to be contained in your file names?
3. **What would you change** in the following names?

My passwords.doc	My data.xls
IMPORTANT.doc	My study.doc
My Thesis final final.doc	Doc.1.doc
My Thesis version 12.doc	New doc.doc
Data 01/08/2016.xls	Int 1 (2).doc

What is the content of the file?

FR3S_140623_129C_2653_W.jpg

What is the content of the file?

1. **Study site.** Indicated by the name, ex. FR3, FR7, FR9.
2. **Depth of the water.** Indicated by S (shallow), M (middle), or D (deep).
3. **Date.** Indicated by YYMMDD.
4. **Tile number.** Indicated on the tile.
5. **Tile treatment.** Indicated by C (caged) or U (uncaged).
6. **Number assigned to photo by camera.**
7. **Whether the post-removal photo was of the entire tile or a tile section.** Indicated by W (whole area), A (upper right), B (lower right), C (lower left), or D (upper left).

FR3S_140623_129C_2653_W.jpg

What is the content of the file?

The researchers wanted to track several things about the tiles:

1. **Study site.** Indicated by the name, ex. FR3, FR7, FR9.
2. **Depth of the water.** Indicated by S (shallow), M (middle), or D (deep).
3. **Date.** Indicated by YYMMDD.
4. **Tile number.** Indicated on the tile.
5. **Tile treatment.** Indicated by C (caged) or U (uncaged).
6. **Number assigned to photo by camera.**
7. **Whether the post-removal photo was of the entire tile or a tile section.** Indicated by W (whole area), A (upper right), B (lower right), C (lower left), or D (upper left).

FR3S_140623_129C_2653_W.jpg

This was image 2653 of whole, uncovered tile 129 from study site 3 in shallow water, taken on June 23, 2014.

Use it for consistent file versioning

Sort

…
FR3S_140623_129C_**2651**_W.jpg
FR3S_140623_129C_**2652**_W.jpg
FR3S_140623_129C_**2653**_W.jpg
FR3S_140623_129C_**2654**_W.jpg
FR3S_140623_129C_**2655**_W.jpg
…

Distinguish

…
FR3S_140623_129C_2653_**A**.jpg
FR3S_140623_129C_2653_**B**.jpg
FR3S_140623_129C_2653_**C**.jpg
FR3S_140623_129C_2653_**D**.jpg
FR3S_140623_129C_2653_**W**.jpg
…

Separate

…\CURRENT

FR3S_**140623**_129C_2655_W.jpg

…\OLD

FR3S_**140622**_129C_2645_W.jpg
FR3S_**140621**_129C_2638_W.jpg
FR3S_**140620**_129C_2631_W.jpg

FR3S_140623_129C_2653_W.jpg

Align it with a data codebook

Sheet_1

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

FR3S_140623_129C_2653_W.jpg

Explain variables within a table!

Standardize it for **both** files & folders

Limit name to 32 characters	32CharactersLooksExactlyLikeThis.csv
Use leading zeros for multi-digit versions	NO ProjID_1.csv ProjID_12.csv YES ProjID_01.csv ProjID_12.csv
Use _ or - instead of spaces	NO Proj ID 1.csv YES Proj_ID_01.csv YES Proj-ID-01.csv
Avoid special characters: ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “	NO name&date@location.doc
Use standardized date formats: YYYYMMDD or YYMMDD.	ProjID_01_20180305.csv
Use only one period for the file extension	NO name.date.doc NO name_date..doc YES name_date.doc
Use specific file names to avoid conflicting naming	NO MyData.csv YES ProjID_data.csv

Document it in a README file

[…]

1. **Study site.** Indicated by the name, ex. FR3, FR7, FR9.
2. **Depth of the water.** Indicated by S (shallow), M (middle), or D (deep).
3. **Date.** Indicated by YYMMDD.
4. **Tile number.** Indicated on the tile.
5. **Tile treatment.** Indicated by C (caged) or U (uncaged).
6. **Number assigned to photo by camera.**
7. **Whether the post-removal photo was of the entire tile or a tile section.** Indicated by W (whole area), A (upper right), B (lower right), C (lower left), or D (upper left).

[…]

Info about data

+

Reusability **instructions**

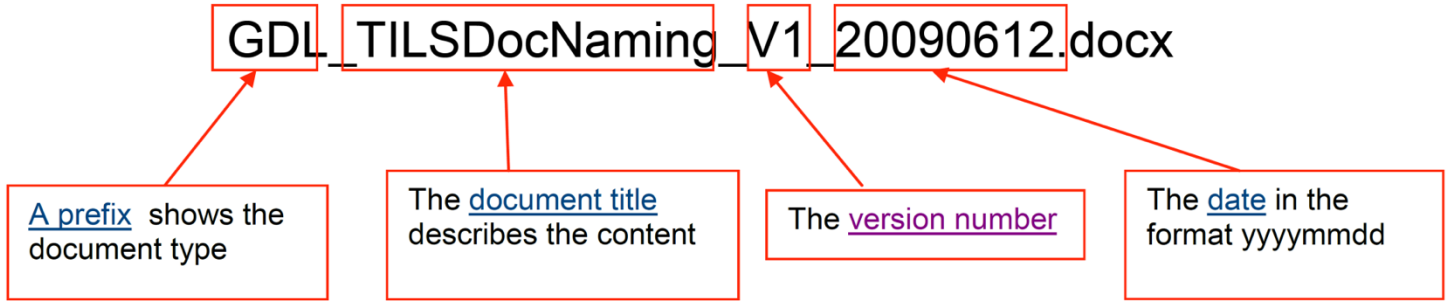
=

Recipe!

Naming convention(s)

Alain Borel, Francesco Varrato

TILS Document Naming Convention (just an example)



Some renaming tools

- Bulk Rename Utility (Win; free)
- PowerRename (part or PowerToys, Win)
- Renamer 4 (Mac)
- Ant Renamer (Win; open source)

Main takeaways

Organize your data with a naming convention to:

- **Know the content** of files without opening them
- **Version** your files
- **Harmonize practices** in your lab
- Make your research more easily
  - **Understandable**
  - **Reusable**
- ~~Cook pasta better~~ ;-)

File versioning

Sort

…
FR3S_140623_129C_**2651**_W.jpg
FR3S_140623_129C_**2652**_W.jpg
FR3S_140623_129C_**2653**_W.jpg
FR3S_140623_129C_**2654**_W.jpg
FR3S_140623_129C_**2655**_W.jpg
…

Distinguish

…
FR3S_140623_129C_2653_**A**.jpg
FR3S_140623_129C_2653_**B**.jpg
FR3S_140623_129C_2653_**C**.jpg
FR3S_140623_129C_2653_**D**.jpg
FR3S_140623_129C_2653_**W**.jpg
…

Separate

…\CURRENT

FR3S_**140623**_129C_2655_W.jpg

…\OLD

FR3S_**140622**_129C_2645_W.jpg
FR3S_**140621**_129C_2638_W.jpg
FR3S_**140620**_129C_2631_W.jpg

Versioning solutions

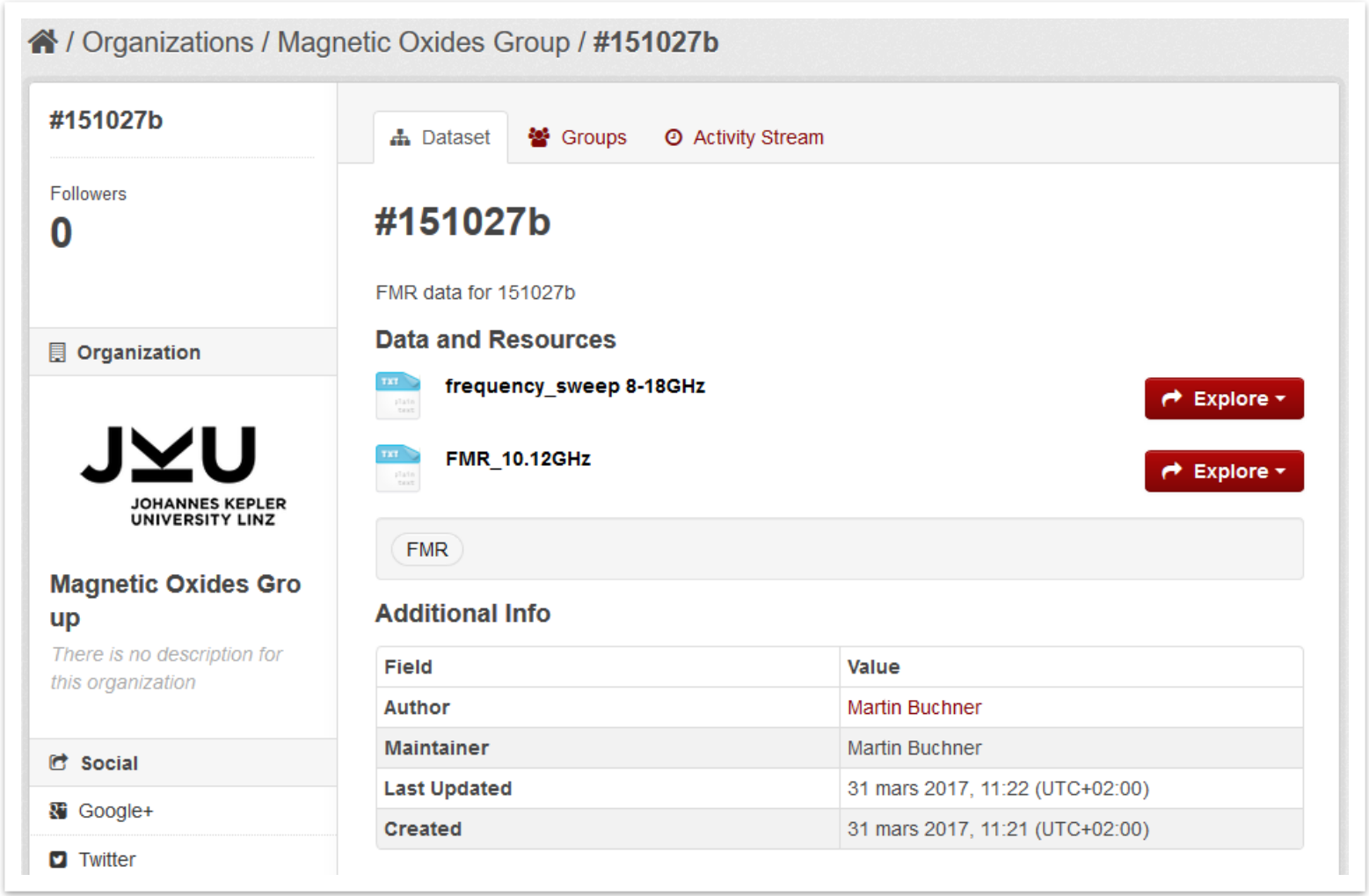
EPFL

Git Project | Git LFS | GitHub | EPFL GitLab | C4Science | TortoiseGit | git-annex

Alain Borel, Francesco Varrato

Git

Discussion: what information is missing?

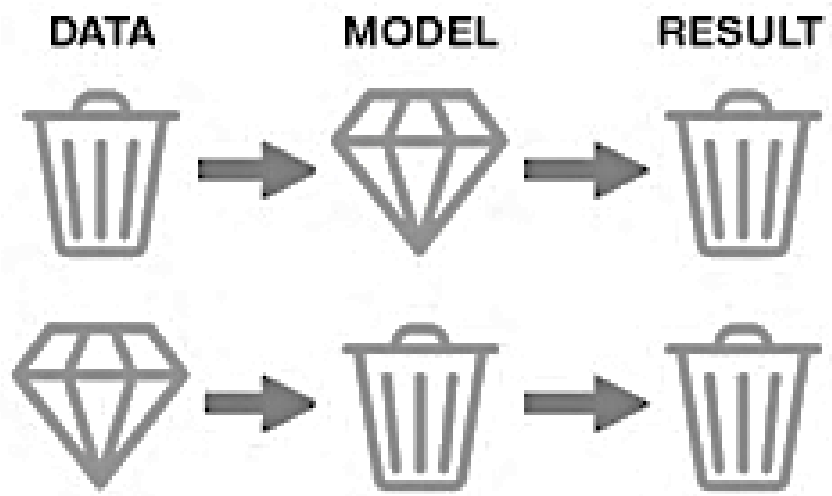


go.epfl.ch/btF

Source: QR Code generator library of the Project Nayuki.

Data cleaning: to be documented, too

GARBAGE IN ➔ GARBAGE OUT



"60% data scientists say they spend the most time cleaning and organizing data"

Crowdflower 2016 Datascience report

"in a data analysis project, data cleansing of poor quality data can take up to 80% of the total effort"

Cost-Benefit analysis for FAIR research data, EC report 2018

Image source:
thedailyomnivore.net/2015/12/02/garbage-in-garbage-out

Data cleaning: to be documented, too

WHEN

- **Preprocessing**, as 1st step (if applicable)
- **Quality assurance** processes

WHY

- **Data ready** for analysis / sharing / publishing / preservation / …
- **Compliance**

HOW

- **Transform** / Reformat / Clean / Merge / Reconciliate data
- **Detect errors** / aberrations
- **Define expected quality** / criteria in a policy (completeness, consistency, accuracy, …)
- **Implement quality** control with human / machine protocols / procedures

Ex. of tools: OpenRefine (free, open-source tool for working with messy data)

Are you using it? Or some other way to keep track?

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

Documentation all along

Alain Borel, Francesco Varrato