

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

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*.



IS THERE A
REPRODUCIBILITY
CRISIS?

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



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

doi.org/10.1038/nchem.2017

Published: 23 July 2014

Reproducibility

Bruce C. Gibb ✉

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

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 story¹ of a p 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

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 science headlines. Although much of the criticism is directed 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 article the inability of people compounds ac

ANITA BANDROWSKI, UNIVERSITY OF C

Danheiser is the editor-in-chief of the unconventional verified the experiments of all the papers it has pub 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

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

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



Annual Review of Chemical and Biomolecular Engineering

Does Chemical Engineering Research Have a Reproducibility Problem?

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

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, / 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>

Copyright © 2019 by Annual Reviews. All rights reserved

ANNUAL REVIEWS CONNECT

- www.annualreviews.org
- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

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 p tens of thousands of papers. We do not claim that research on MC (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 DURRANT | 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

LATEST

POPULAR



Peer review requires revisions



Salt crystal grows legs to avoid slippery surface



Is there life on Venus?



Human chemical communication



Ig Nobels feature knives created from human poo and vibrating worms



Trust in peer review VOICE OF THE ROYAL SOCIETY OF CHEMISTRY

Sign up to email newsletters

Get the latest chemistry research.

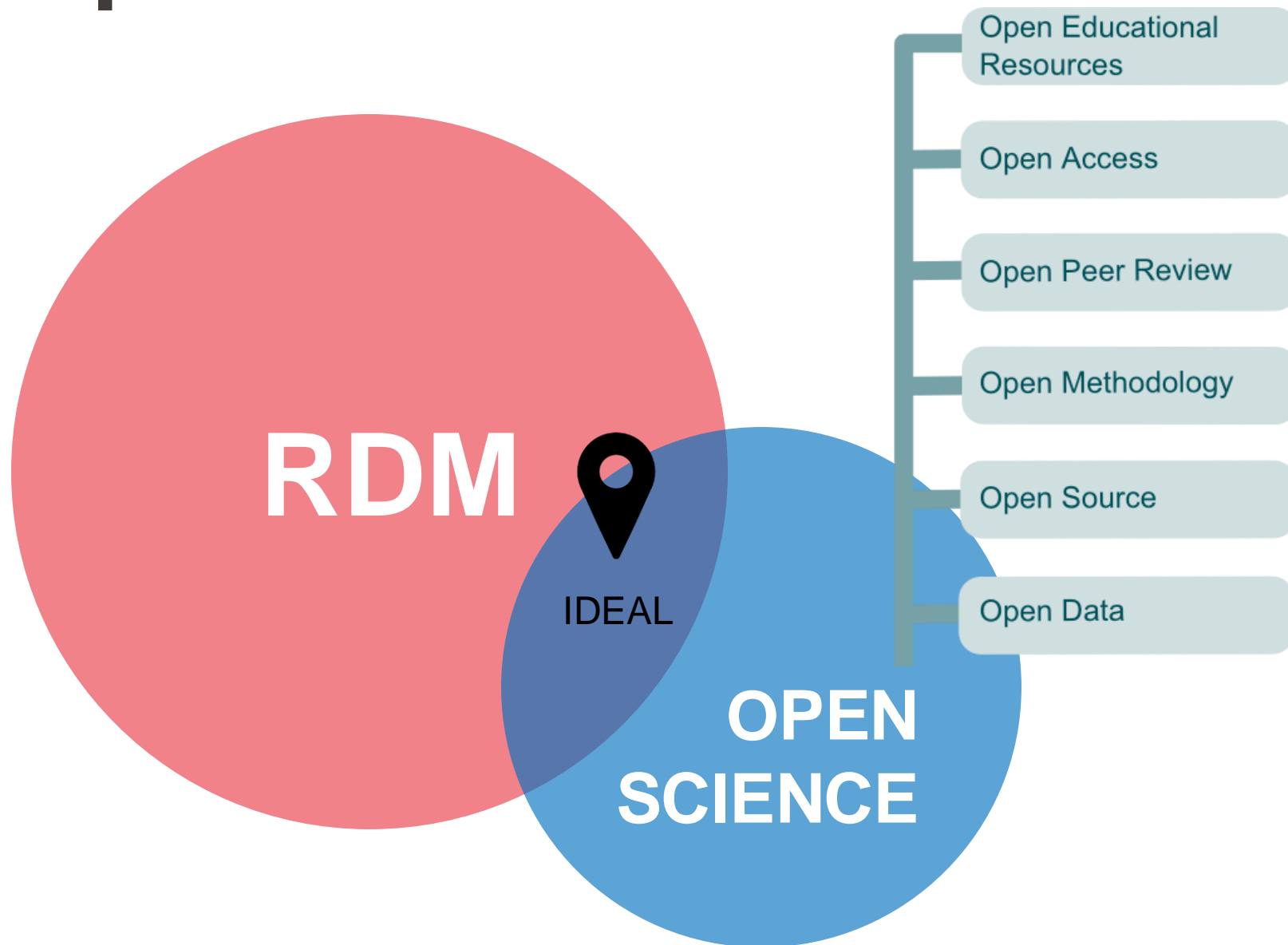


Research reproducibility: some recommendations

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 \Leftrightarrow Open Science?



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

True or False?

Institutional policies

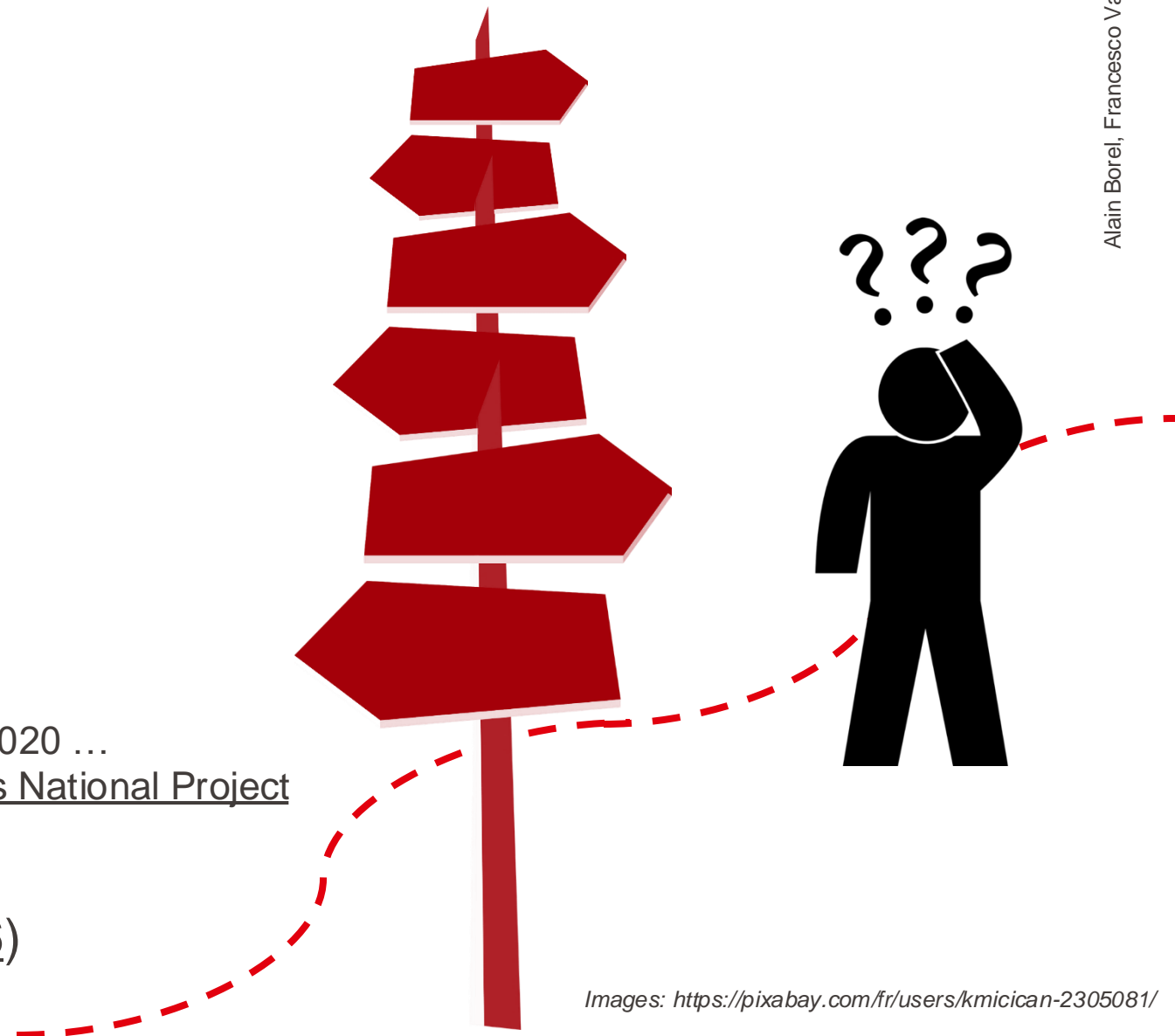
- ETHZ Guidelines for RDM
- Humboldt-Universität zu Berlin
- MIT
- TU Delft
- UNIGE
- University of Cambridge
- University of Edinburgh
- University of Oxford

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/kmicican-2305081/>



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

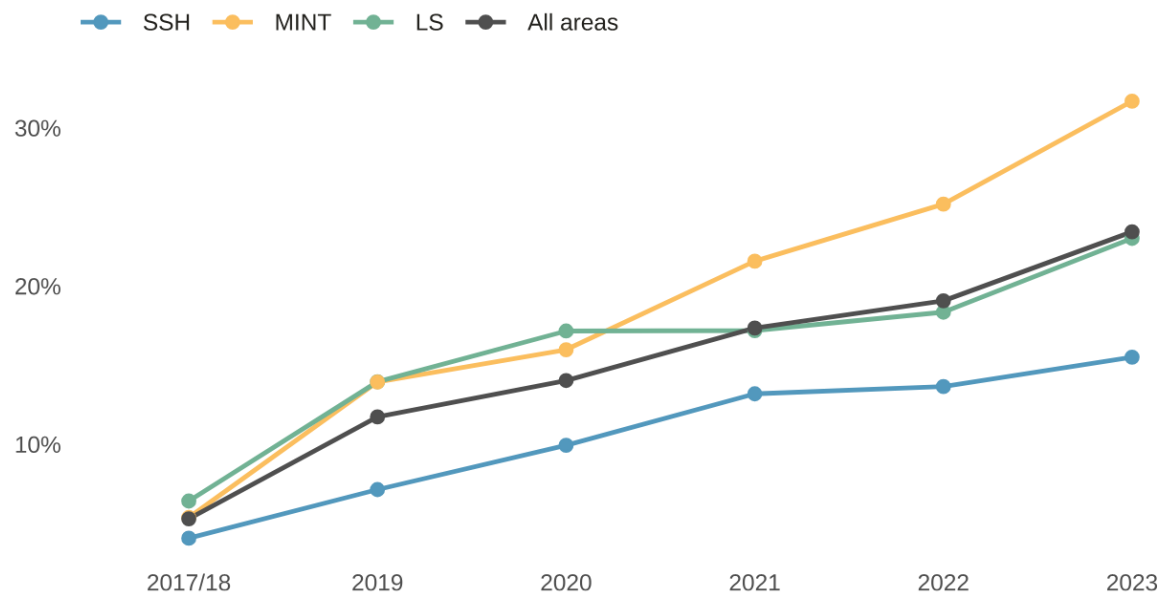
SERI guide: DMP as PDF on same platform as other project docs



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

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>

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



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.

FAIR data vs. Supplementary Information

- **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.



S.I.
human ? machine ?



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?



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

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)

**METADATA IS A
LOVE NOTE
TO THE FUTURE**

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!

DOI: 10.1038/d41586-019-03240-w

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 Castelvocchi

Twitter Facebook Email



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

Data Management Plan

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 Management Plan









Living document & Research roadmap

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

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
- **Compliance:** 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
- ...

go.epfl.ch/rdm-guide

 <p>EPFL DMP</p>	 <p>SNSF DMP (with examples)</p>	 <p>SNSF DMP (only guiding questions)</p>
 <p>ERC DMP</p>	 <p>H2020 DMP</p>	 <p>MSCA DMP (also Horizon Europe H2021-2027)</p>
 <p>EPFL RDM Strategy</p>		
 <p>NCCR RDM Strategy (series 5)</p>		

Other resources

DMP Templates

DMP templates provided by a funder or research organisations, available on DMP OPIDOR. You can download these templates and related guidances, create a plan from these templates.

Template Name	Organisation Name	Organisation Type	Description	Last Updated	Download	Actions
EPFL SNSF	EPFL - Ecole Polytechnique Fédérale de Lausanne	Institution	This template was co-written by EPFL Library and ETH Library in the scope of the DCM project. The current document is the EPFL version 5.0, revised in July 2019 by the EPFL Library Research Data team. ETH version is available from their own website. For further help, personal feedback or comments, you can contact us at researchdata@epfl.ch .	26-09-2019		
EPFL SNSF with examples hosted on web site	EPFL - Ecole Polytechnique Fédérale de Lausanne	Institution	This template was co-written by EPFL Library and ETH Library in the scope of the DCM project. The current document is the EPFL version 5.0, revised in July 2019 by the EPFL Library Research Data team. ETH version is available from their own website. For further help, personal feedback or comments, you can contact us at researchdata@epfl.ch . Example answers for this template are hosted on a dedicated EPFL web page rather than embedded in DMP OPIDOR.	26-09-2019		

dmp.opidor.fr/public_templates

ARGOS

Welcome to ARGOS
Create, Link, Share Data Management Plans

CREATE NEW DMP
Click on the button to create a new Dataset Description (DMP) plan. You will be prompted to enter the name of the dataset and the name of the DMP.

ADD A DATASET DESCRIPTION INTO AN EXISTING DMP
This button allows you to describe additional Dataset Descriptions (DMPs) associated with an existing DMP, providing you the necessary information for their creation.

0 My Datasets | 0 My Drafts | 0 My Drafts | 0 My Drafts | DRAFT DMPs | VIEW ALL

argos.openaire.eu

DMPTONLINE

EPFL - Ecole Polytechnique Fédérale de Lausanne

Templates

If you wish to add an organisational template for a Data Management Plan, use the 'create template' button. You can create more than one template if desired e.g. one for researchers and one for PhD students. Your template will be presented to users within your organisation when no further template apply. If you want to add questions to further templates use the 'customise template' options below.

Own Templates

Template Name	Description	Status	Edited Date	Actions
Customise template do not use		Unpublished	02-22-2022	Actions
EPFL DMP	This is the School of Architecture, Civil and Environmental Engineering (SACE) at EPFL data management plan template. It should be used for SACE internal funding schemes and other research. <p>Please reserve sufficient time to create your data management plan. Select EPFL Guidance to help you answer all the questions. You may use a collaborative approach and work on it together with your colleagues or ask for feedback from various research departments, e.g. ICT, when finished, do not forget to ask for feedback from the EPFL Library Data team (click on the "ask for feedback" button). Your feedback will significantly improve your data management plan and help you to meet all potential requirements.</p>	Published	01-29-2022	Actions
EPFL DMP template	This DMP template has been designed to be applicable to any EPFL research project that produces, collects or processes research data, wherever that is locally (EPFL premises) or elsewhere (cloud or other external storage). It is a generic DMP template and has not been tailored to a specific research project. <p>You should develop a single DMP for your project to cover the overall approach, however, where there are specific needs for a particular dataset, you can create a customised version.</p>	Published	02-22-2022	Actions

dmponline.dcc.ac.uk

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

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



1. Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

Questions you might want to consider:

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

B I U 1 ≡ Ω ↶ ↷ ✂ 📄 🖨

Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated data sets. (This relates to the [FAIR Data Principles F2, I3, R1 & R1.2.](#))

1.2 How will the data be collected, observed or generated?

1.3 What documentation and metadata will you provide with the data?

2. Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

2.2 How will data access and security be managed?

2.3 How will you handle copyright and Intellectual Property Rights issues?

3. Data storage and preservation

3.1 How will your data be stored and backed-up during the research?

3.2 What is your data preservation plan?

4. Data sharing and reuse

4.1 How and where will the data be shared?

4.2 Are there any necessary limitations to protect sensitive data?

4.3 All digital repositories I will choose are conform to the FAIR Data Principles.

4.4 I will choose digital repositories maintained by a non-profit organisation.

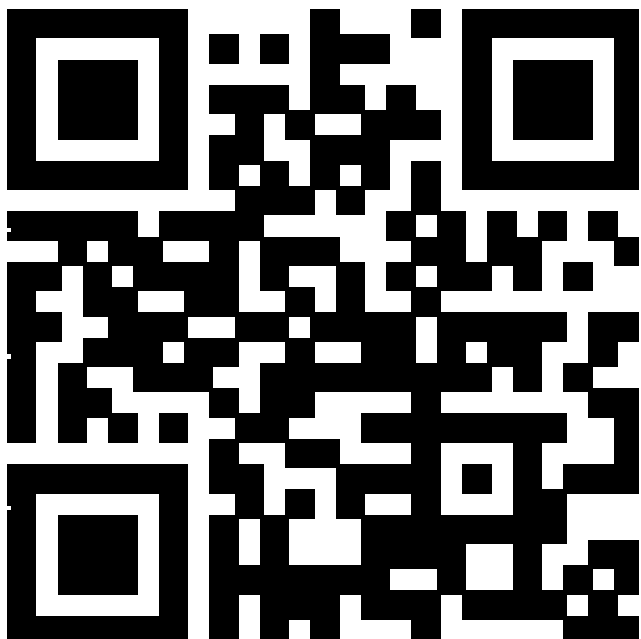
Exercise [~ 15']: SNSF DMP

[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



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 \(EPFL\)](#)

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



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>

Data or Metadata?



DATA = Input Values



METADATA = Info about data

- **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](https://www.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>

Metadata: Data

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



In a client database, **client**, **brand** and **fabrication country** are your metadata.

The data is not in bold

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

```
{
  "metadata": {
    "access_right_category": "success",
    "doi": "10.5281/zenodo.1345472",
    "description": "<p>In this paper, we prove fixed point theorem and common fixed point theorem for two pairs of set valued mappings in \ud835\udd3a metric spaces. Further, the famous Banach's contraction principle and some of its generalizations and variants are realizable as special cases of our results.</p>",
    "license": {
      "id": "CC-BY-4.0"
    },
    "title": "Common Fixed Point Theorem for Two Pairs of Set Valued Mappings in \ud835\udd3a - Metric Space",
    "journal": {
      "volume": "03",
      "issue": "08",
      "pages": "309-313",
      "title": "RESEARCH REVIEW International Journal of Multidisciplinary"
    },
    [...]
  }
}
```

Machine readable

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: <input type="text"/>				
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](#).

Dataset description (metadata)

<ul style="list-style-type: none"> ▪ Identifier ▪ Title ▪ Creator ▪ Subject ▪ Description 	<ul style="list-style-type: none"> ▪ Publisher ▪ Date ▪ Formats ▪ Rights ▪ ...
----------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------

Basic & General schemas

- Dublin Core
- DataCite
- ...

Disciplinary schemas

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

Data Repository integration (example)

July 27, 2017

Dataset Open Access

0.48 Angstrom 3,5-dinitrobenzoic acid (3,5-DNBA) C2/c polymorph single crystal X-ray diffraction data set recorded at Diamond Light Source I19-1

Saunders, Lucy; Nowell, Harriott; Winter, Graeme

Versions

Version 4	Jul 27, 2017
10.5281/zenodo.1036416	
Version 3	Jul 27, 2017
10.5281/zenodo.1036405	
Version 2	Jul 27, 2017
10.5281/zenodo.1036299	
Version 1	Jul 27, 2017
10.5281/zenodo.835537	

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.835536](https://doi.org/10.5281/zenodo.835536). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

External resources

Indexed in



Communities



Keywords and subjects

x-ray diffraction diamond light source i19-1
chemical crystallography

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

EPFL Metadata standards




Clear AllRegistry: StandardQuery string: chemistry

<1234>

Displaying 1 to 30 of 112.

MISFISHIE



R

Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments

MISFISHIE is the Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments. This specification details the minimum information that should be provided when publishing, making public, ...


Life ScienceEvidenceExpression...All+8 more tags

Related Standards1

Implementing Databases3

Endorsing Policies0

Chemistry vocabulary



R

Controlled vocabulary used for indexing bibliographical records dealing with chemistry in the PASCAL database (1972-2015). It is aligned with the terms of the ChEBI (Chemical Entities of Biological Interest), RXNO (name reacti...


Organic Ch...Inorganic ...Not applic...+9 more tags

Related Standards12

Implementing Databases0

Endorsing Policies0

CML



R

Chemical Markup Language

CML (Chemical Markup Language) is an XML language designed to hold most of the central concepts in chemistry. It was the first language to be developed and plays the same role for chemistry as MathML for mathematics and GML f...


Biochemist...ChemistryMolecular...Mathemati...All+5 more tags

Related Standards3

Implementing Databases3

Endorsing Policies0

CSMD



R

Core Scientific MetaData model

Capturing high level information about scientific studies and the data that they produce, the CSMD is developed to support data collected within a facility's scientific workflow. However the model is also designed to be generic acro...

Engineerin...Biochemist...Not applic...+6 more tags

Related Standards2

Implementing Databases0

Endorsing Policies0

Standards for Reporting Enzymology Data Guidelines

Metadata standards usage (examples)

Chemical Methods Ontology

CHMO, the chemical methods ontology, describes methods used to collect data in chemical experiments, such as mass spectrometry and separate material for further analysis and deposition. It also describes the instructions to the Ontology for Biomedical Investigations.

[Terms](#)
[Download](#)
[Or](#)


SNP Data Center

Home Data Publications Projects Contact Add Data Centers ▾

Entry no. 5452

Further info

Private URL

Datatype

Filename

Path

Vegetation units of the SNP and the neighboring

http://www.parcs.ch/snp/pdf_public/2014/5452_20140814_143710_main_s

-

GIS Vector Layer

veg_zoller_fulltext

Q:\maindata\snp\botany\gis_pub\zoller_veg.gdb

veg_zoller

UCAR COMMUNITY PROGRAMS | unidata Data Services and Tools for Geoscience

Network Common Data Form (NetCDF)

NetCDF

Release Notes

FAQs

NetCDF C & C++ Documentation

NetCDF Fortran Documentation

NetCDF Java Documentation

Download

Support

For Developers

Compatible Software

NetCDF CDash Tests

Related Projects

NetCDF (Network Common Data Form) is a set of interfaces for array-oriented data access and a freely distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data.

See the netCDF package overview ▸

NetCDF News & Announcements

NetCDF 4.6.3
5 mars 2019

NetCDF 4.6.2
21 novembre 2018

NetCDF 4.6.1
20 mars 2018

NetCDF news archive ▸

Citing NetCDF

If you use netCDF and want to provide a DOI/citation, see How to Acknowledge Unidata.

NetCDF Fact Sheet

A netCDF fact sheet provides a brief overview of the netCDF package and supported languages and platforms.

View the netCDF fact sheet ▸

ISO 19115:2003

Geographic information -- Metadata

This standard has been revised by ISO 19115-1:2014

ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.

ISO 19115:2003 is applicable to:

- the cataloguing of datasets, clearinghouse activities, and the full description of datasets;
- geographic datasets, dataset series, and individual geographic features and feature properties.

ISO 19115:2003 defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data);
- optional metadata elements - to allow for a more extensive standard description of geographic data, if required;
- a method for extending metadata to fit specialized needs.

Though ISO 19115:2003 is applicable to digital data, its principles can be extended to many other forms of geographic data such as maps, charts, and textual documents as well as non-geographic data.

NOTE Certain mandatory metadata elements may not apply to these other forms of data.

Vegetation map of the SNP and its surroundings

Project Zoller, H. 1957

↑↓	DATATYPE	↑↓	AUTOR/OWNER	↑↓	YEAR	↑↓
REINA - Vegetation Map Zoller	GIS Vector Layer		Kantonale Verwaltung Graubünden, Amt für Langsamverkehr		2009	
9 Vegetationskartierung	Project		SNP		2013	
27934 TBT79 Carex	Project		SNP		2013	
27940 TBT79 Carex: Magerwiesen SNP	GIS Vector Layer		SNP		2013	
27941 TBT79 Carex: Borst- und Blaugrashalden	GIS Vector Layer		SNP		2013	
Archivdatensatz: Vegetation units of the SNP and the	GIS Vector					

Standards: From awareness to practice

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, Department
Cambridge CB2 1EW, U.K., Cambridge University L
CB3 9DR, U.K., Department of Chemistry, Imperial C
U.K., and Imperial College Library and High Perform
Exhibition Road, Lor

Ye Li* and Lori Tschirhart

Received Dece

DOI: 10.1021/bk-2012-1110.ch009

Publication Date: November 15, 2012 ✓

[RIGHTS & PERMISSIONS](#)

[Special Issues in Data Management](#)

Chapter 9, pp 145-162

ACS Symposium Series, Vol. 1110

ISBN13: 9780841227125 eISBN: 9780841227132

Copyright © 2012 American Chemical Society

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

The SPECTRa (Submission, Preservation and Exposure) 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

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



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.”

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

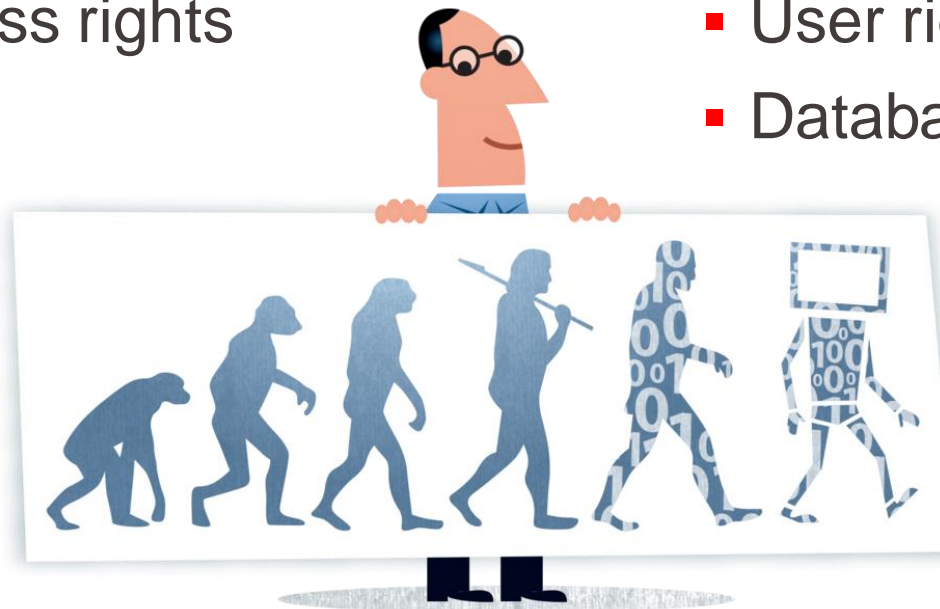


Image source: Digitalbevaring.dk ([CC BY 2.5 DK](https://creativecommons.org/licenses/by/2.5/dk/))

Dataset or Database?



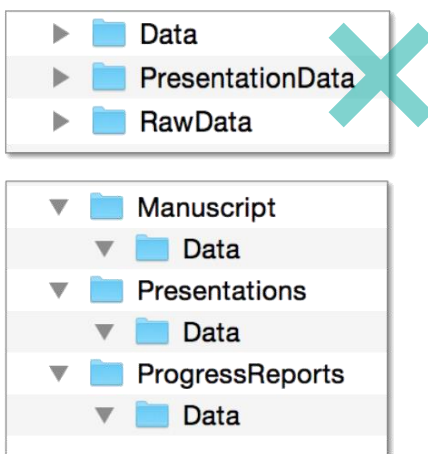
Database:
Organized collection of data
stored as multiple datasets



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

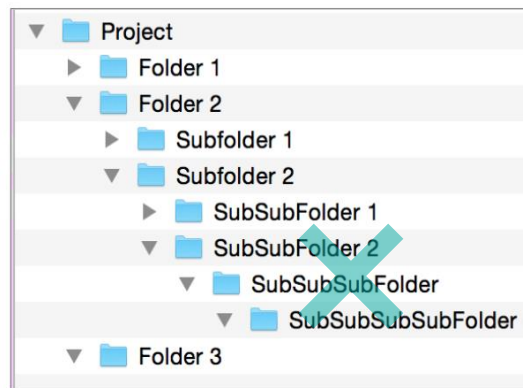
Try to **avoid** ...

overlapping categories



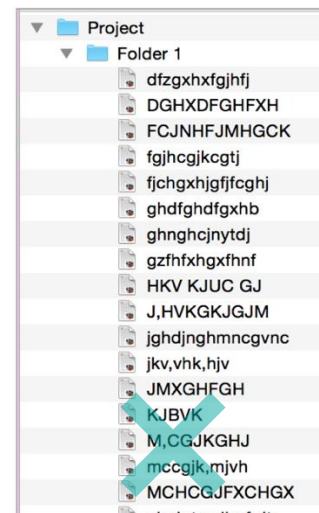
Rule of thumb:
“sure of the right subdirectory”

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?



- 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

country	year	cases	population
Afghanistan	1999	2775	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	216766	128042583

variables

country	year	cases	population
Afghanistan	1999	2775	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	216766	128042583

observations

country	year	cases	population
Afghanistan	1999	2775	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	216766	128042583

values

Example : Organization starter template

```

├── .gitignore
├── CITATION.md
├── LICENSE.md
├── README.md
├── requirements.txt
├── bin
│   └── external
├── config
├── data
│   ├── processed
│   ├── raw
│   └── temp
├── docs
│   ├── manuscript
│   └── reports
├── results
│   ├── figures
│   └── output
└── src

```

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

<- Compiled and external code, ignored by git (PG)
 <- Any external source code, ignored by git (RO)
 <- Configuration files (HW)
 <- All project data, ignored by git
 <- The final, canonical data sets for modeling. (PG)
 <- The original, immutable data dump. (RO)
 <- Intermediate data that has been transformed. (PG)
 <- Documentation notebook for users (HW)
 <- Manuscript source, e.g., LaTeX, Markdown, etc. (HW)
 <- Other project reports and notebooks (e.g. Jupyter, .Rmd) (HW)
 <- Figures for the manuscript or reports (PG)
 <- Other output for the manuscript or reports (PG)
 <- Source code for this project (HW)

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 file 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/

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).

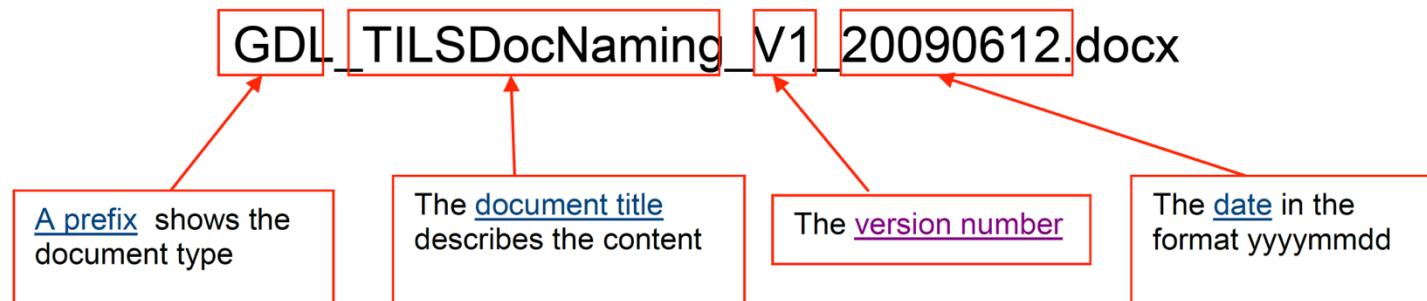
[...]



Info about data
+
Reusability **instructions**
=
Recipe!



TILS Document Naming Convention (just an example)



Some renaming tools

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

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 ;-)~~



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



Git Project



Git LFS



GitHub



EPFL GitLab



C4Science



TortoiseGit



git-annex

EPFL



Image by [Recoopre](#), CC-BY-SA-4.0 


Discussion: what information is missing?

Home / Organizations / Magnetic Oxides Group / #151027b

#151027b

Followers
0

Organization



Magnetic Oxides Group

There is no description for this organization

Social

Google+

Twitter

Dataset Groups Activity Stream

#151027b

FMR data for 151027b

Data and Resources

frequency_sweep 8-18GHz

Explore

FMR_10.12GHz

Explore

FMR

Additional Info

Field	Value
Author	Martin Buchner
Maintainer	Martin Buchner
Last Updated	31 mars 2017, 11:22 (UTC+02:00)
Created	31 mars 2017, 11:21 (UTC+02:00)

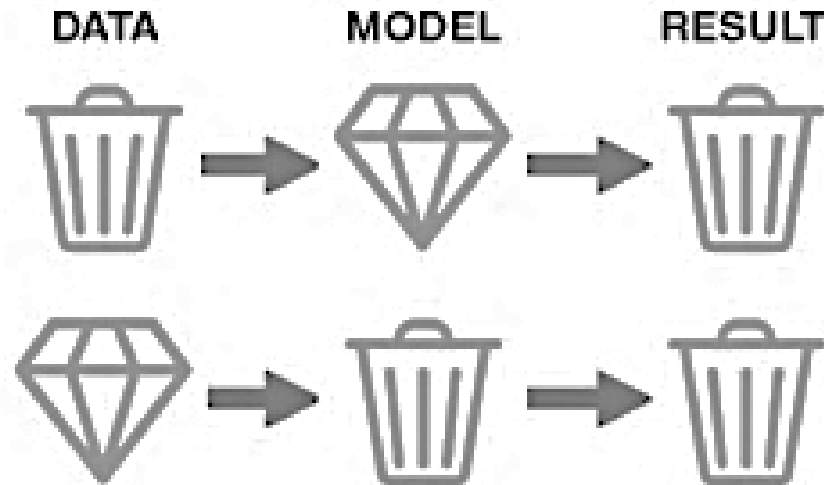
go.epfl.ch/btF



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”

Crowdfunder 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