Communication / Dissemination / Exploitation

Communication:
Promote your action and results

Inform, promote and communicate your activities and results

Reaching multiple audiences
Citizens, the media, stakeholders

How?
- Having a well-designed strategy
- Conveying clear messages
- Using the right media channels

When?
From the start of the action until the end

Why?
- Engage with stakeholders
- Attract the best experts to your team
- Generate market demand
- Raise awareness of how public money is spent
- Show the success of European collaboration
- **Legal obligation: Article 38.1 of the Grant Agreement**

Dissemination:
Make your results public

Open Science: knowledge and results (free of charge) for others to use

Only to scientists?
Not only but also to others that can learn from the results: authorities, industry, policymakers, sectors of interest, civil society

How?
Publishing your results on:
- Scientific magazines
- Scientific and/or targeted conferences
- Databases

When?
At any time, and as soon as the action has results

Why?
- Maximise results' impact
- Allow other researchers to go a step forward
- Contribute to the advancement of the state of the art
- Make scientific results a common good
- **Legal obligation: Article 29 of the Grant Agreement**

Exploitation:
Make concrete use of results

Commercial, Societal, Political Purposes

Only by researchers?
Not only, but also:
- Industry including SMEs
- Those that can make good use of them: authorities, industrial authorities, policymakers, sectors of interest, civil society

How?
- Creating roadmaps, prototypes, softwares
- Sharing knowledge, skills, data

When?
Towards the end and beyond, as soon as the action has exploitable results

Why?
- Lead to new legislation or recommendations
- For the benefit of innovation, the economy and the society
- Help to tackle a problem and respond to an existing demand

Legal obligation: Article 28 of the Grant Agreement

Source: https://rea.ec.europa.eu/system/files/2021-08/quick-guide_diss-expl_en.pdf

Importance of sharing (!)

1976 – Experiments on supercooled water (cooled far below its freezing point) showed a **critical point** at −20°C: its structure fluctuates widely between high- and low-density forms

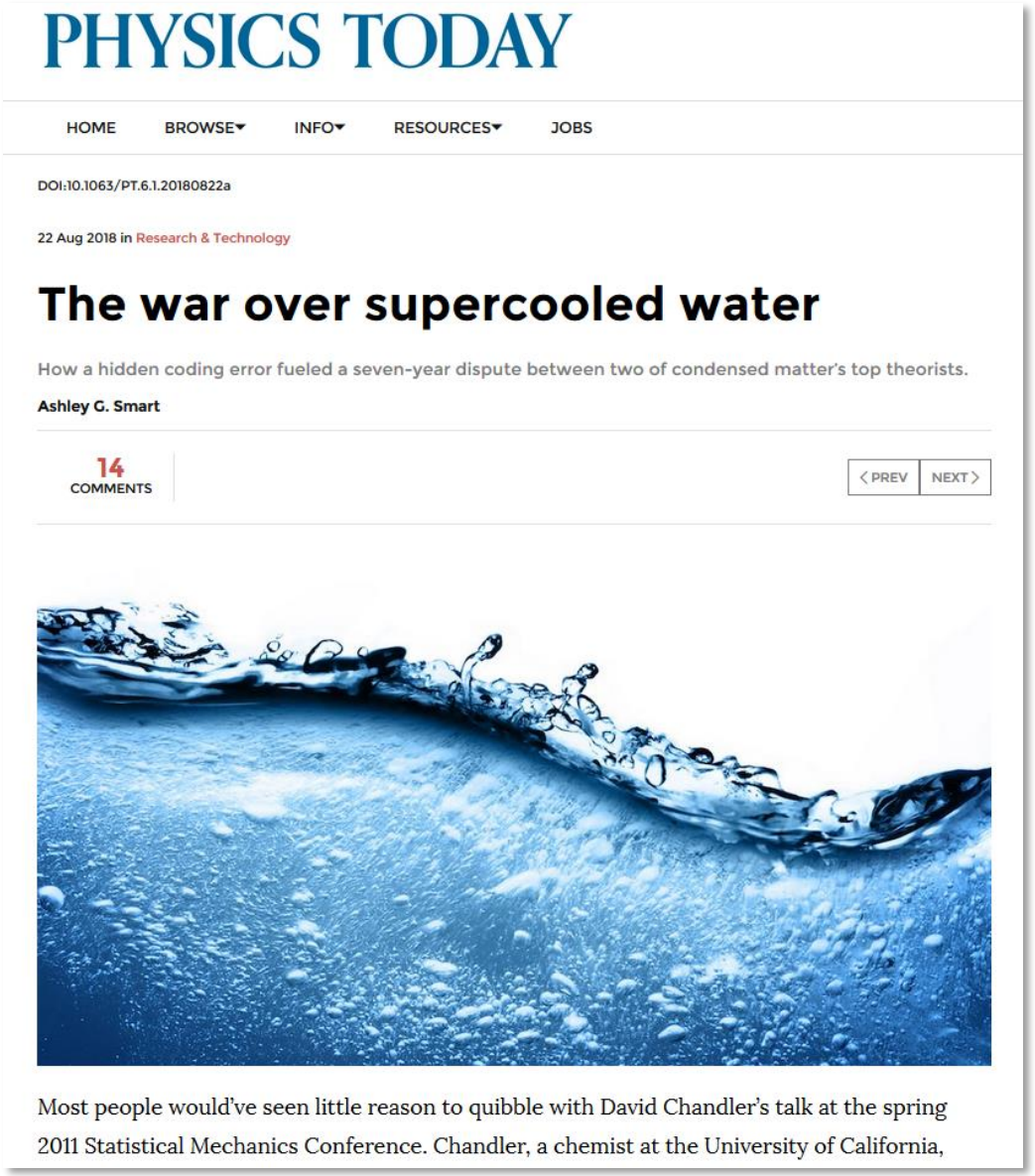
2011 – Seeking a unified theory of water, simulations on supercooled water by two world-leading groups revealed:

- Chandler et al.: **no critical point** (resembles ordinary water)
- Debenedetti et al.: **critical point** (morphs between two forms)

2014 – Debenedetti et al. **published their code** openly

2016 – At first, Chandler et al. only shared data, then revealed where to find its code and, after lot of reverse engineering …

2018 – … the trouble stemmed from an algorithmic trick the Chandler's team used to speed up their code!



PHYSICS TODAY

HOME BROWSE INFO RESOURCES JOBS

DOI:10.1063/PT.6.1.20180822a

22 Aug 2018 in Research & Technology

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top theorists.

Ashley G. Smart

14 COMMENTS

PREV NEXT

Most people would've seen little reason to quibble with David Chandler's talk at the spring 2011 Statistical Mechanics Conference. Chandler, a chemist at the University of California,

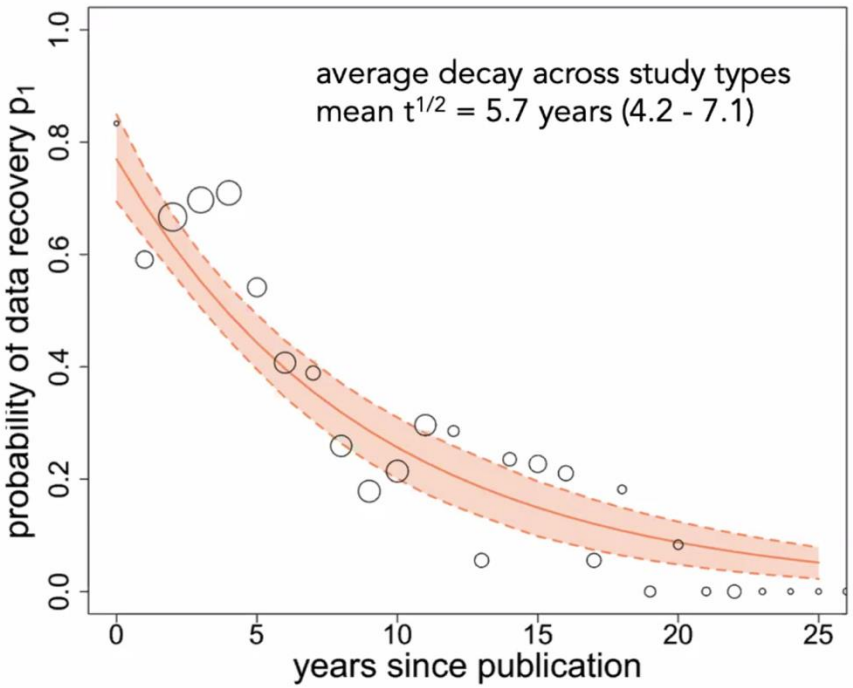
DOI: 10.1063/PT.6.1.20180822a

Importance of raw data

EDITORIAL

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa



Data & analysis script availability (prevalence estimates)

	Data	Analysis scripts
Psychology (2014-2017)[1]	2% [1-4%]*	1% [0-1%]
Social Sciences (2014-2017)[2]	7% [2-13%]	1% [0-3%]

*[95% confidence intervals]

[1]Hardwicke et al. (2021)
[2]Hardwicke et al. (2020)

Data availability on request (selected studies)

	Data shared
141 articles published in four major APA journals (2004)[3]	27%
516 ecology articles published (1991-2011)[4]	20%
111 most highly-cited psychology & psychiatry articles (2006-2016)[5]	14%

[3]Wicherts et al. (2006)
[4]Vines et al. (2014)
[5]Hardwicke & Ioannidis (2018)

DOI: 10.1186/s13041-020-0552-2

Importance of raw data

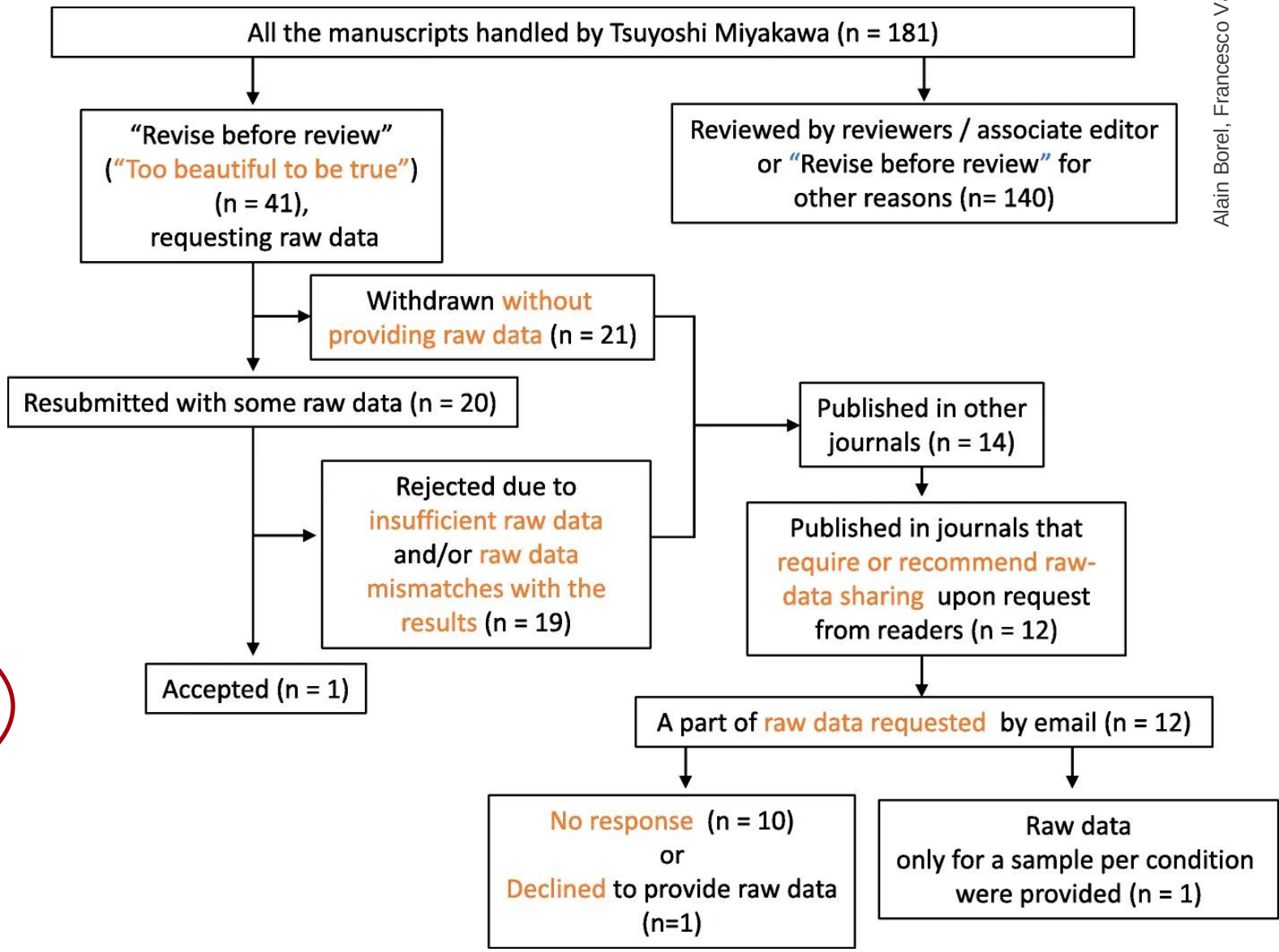
Alain Borel, Francesco Varrato

EDITORIAL

No raw data, no science: another possible source of the reproducibility crisis

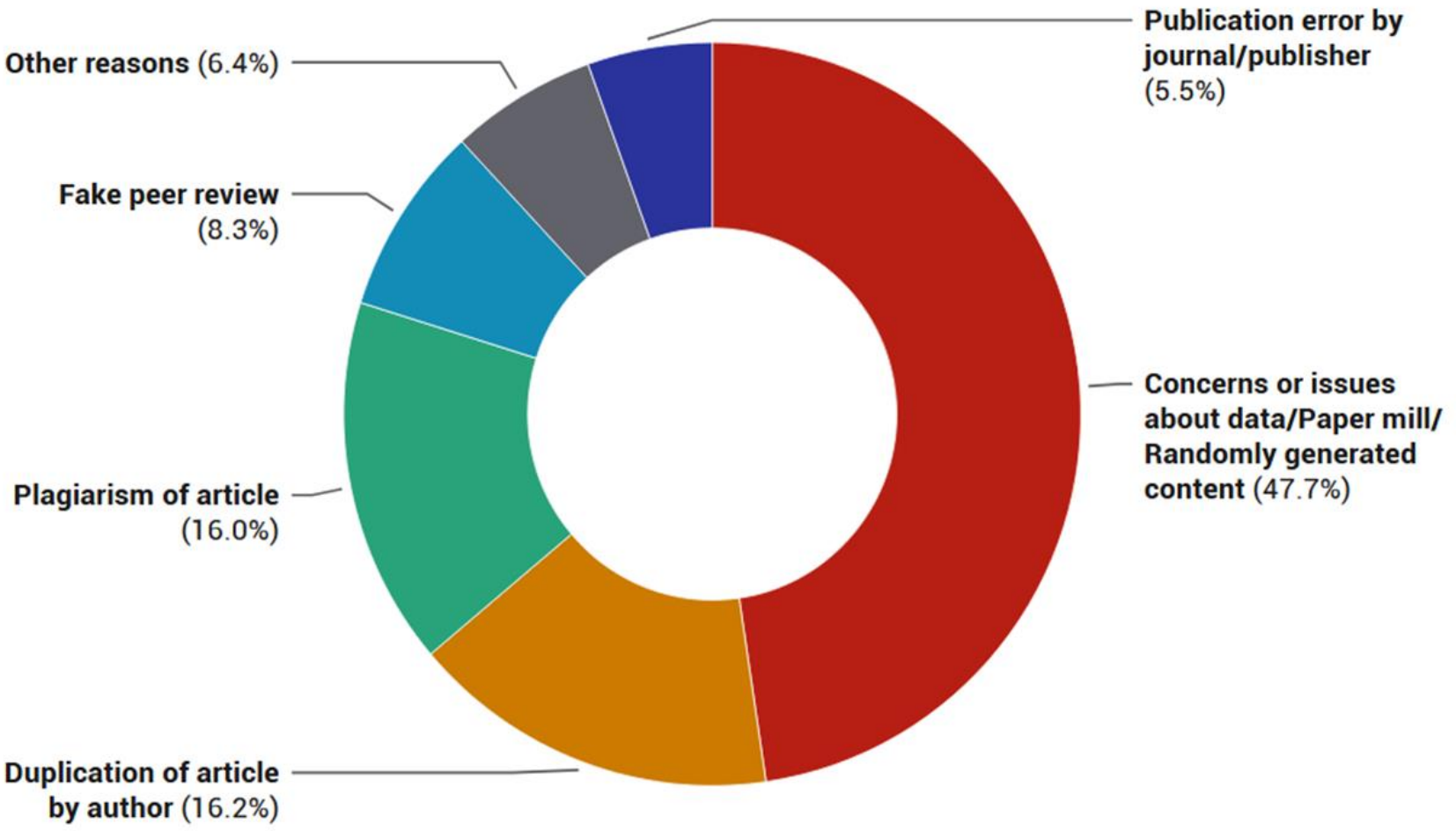
Tsuyoshi Miyakawa

- Lack of raw data: another possible **cause of irreproducibility**
- Many researchers did **not provide the raw data**
- Data fabrication: raw data may **not even exist** in some cases
- Good faith: the insufficiency or mismatch between raw data and results ~~can be **honest mistakes**~~
- Systematic review and meta-analysis: estimated that 1.97% of authors admitted to have **fabricated, falsified, or modified data** or results at least once […] the admission rate was 14.12% for falsification when asked about the colleagues



DOI: 10.1186/s13041-020-0552-2

Reasons for research paper retractions

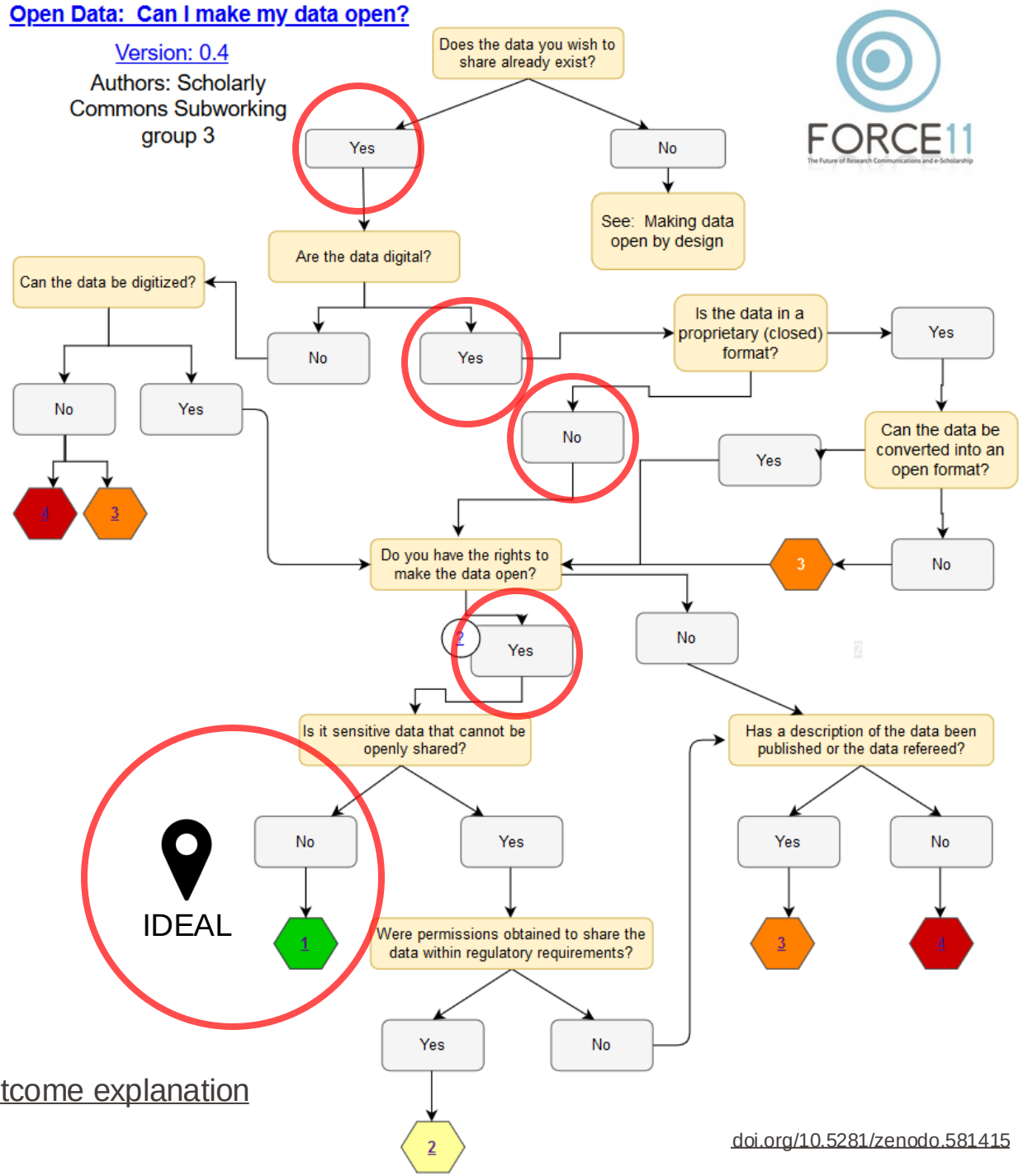


Sources:
https://theconversation.com/the-publish-or-perish-mentality-is-fuelling-research-paper-retractions-and-undermining-science-238983 || http://retractiondatabase.org/RetractionSearch.aspx? || https://www.datawrapper.de/_/eFeUA/

Open Data Decision Tree

go.epfl.ch/OD_Tree

Sources: QR Code generator library: Project Nayuki



Outcome explanation

doi.org/10.5281/zenodo.581415

Legal constraints for open data publication

Alain Borel, Francesco Varrato

SENSITIVE

- Tests on animals / humans
- Handle personal data
  - Federal Act on Data Protection (FADP), Human Research Act (HRA), GDPR
  - name, identification number, location data, online identifier, ...
  - factors specific to physical, physiological, genetic, mental, economic, cultural or social identity

→ *check the EPFL Human Research Ethics Committee* (AREC + HREC)

COMMERCIAL

- **Data from 3rd party** sources? (e.g. commercial datasets, research cooperations, etc.)

→ *check out the **contract** for data usage / sharing ... Or make one!*

- Want to potentially submit a **patent**?

→ *check the TTO (Technology Transfer Office) ... Choose the **data license** + tell in the DMP!*

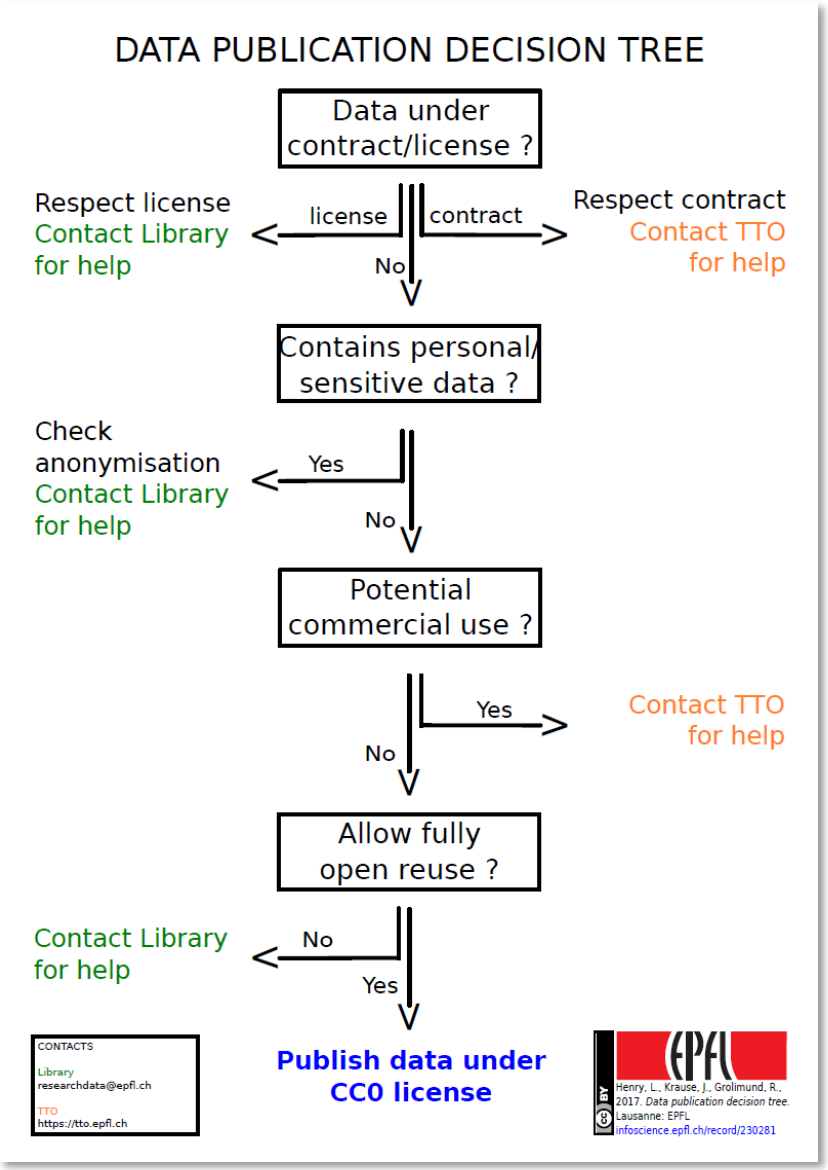
Live poll: Who owns my data/code?

- ❑ **Me**
- ❑ **Thesis supervisor / PI**
- ❑ **Publisher / Platform** (once published)
- ❑ **3rd party** (obtained from provider)
- ❑ **EPFL** (ETH Domain)
- ❑ **I don't know**

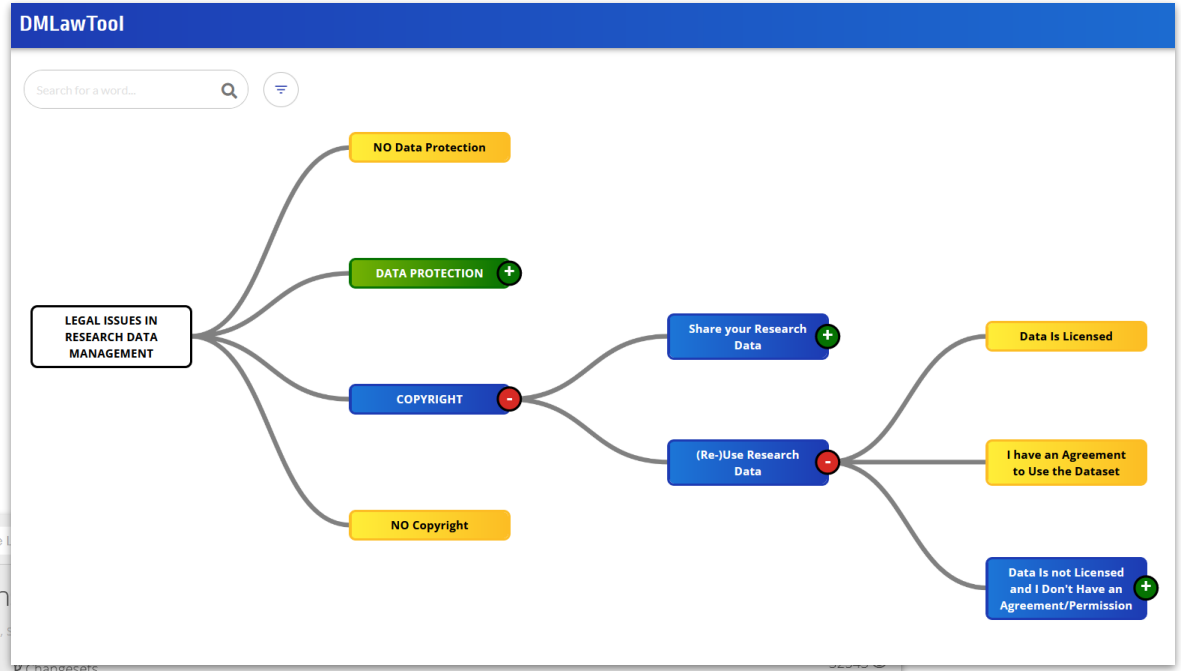
NOTE: ownership ≠ authorship ≠ licensee

Potential commercial use?

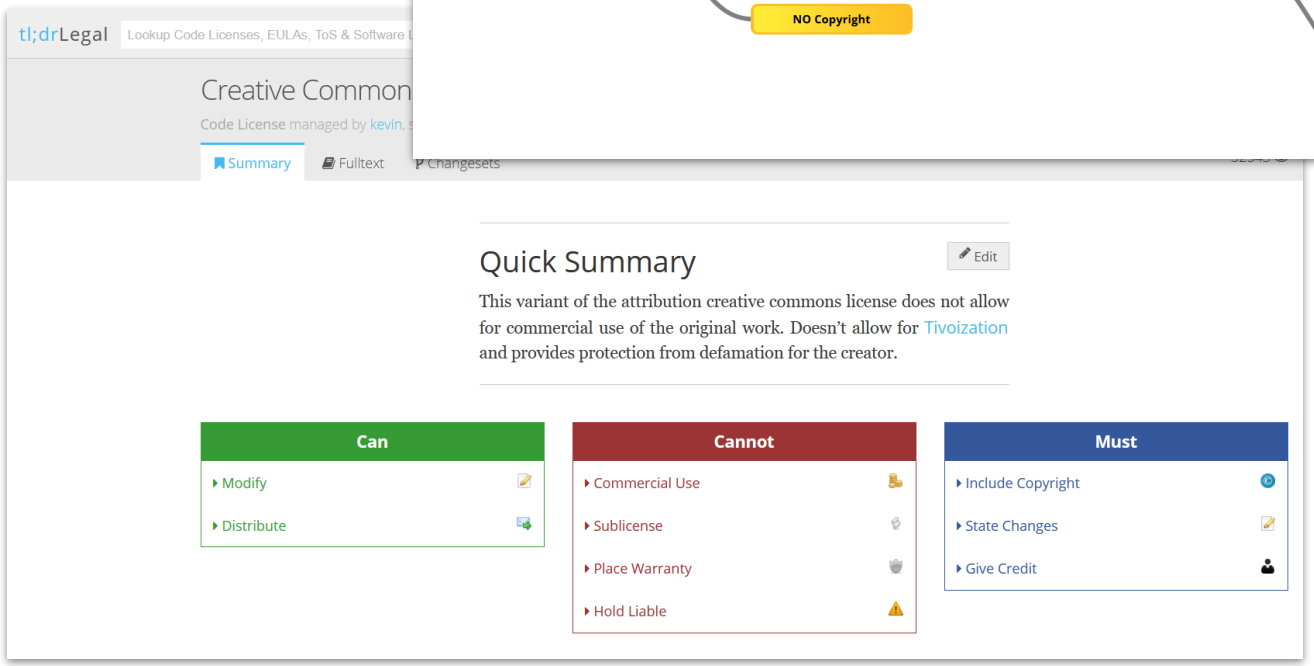
infoscience.epfl.ch/record/230281



DMLawTool



tldrlegal.com

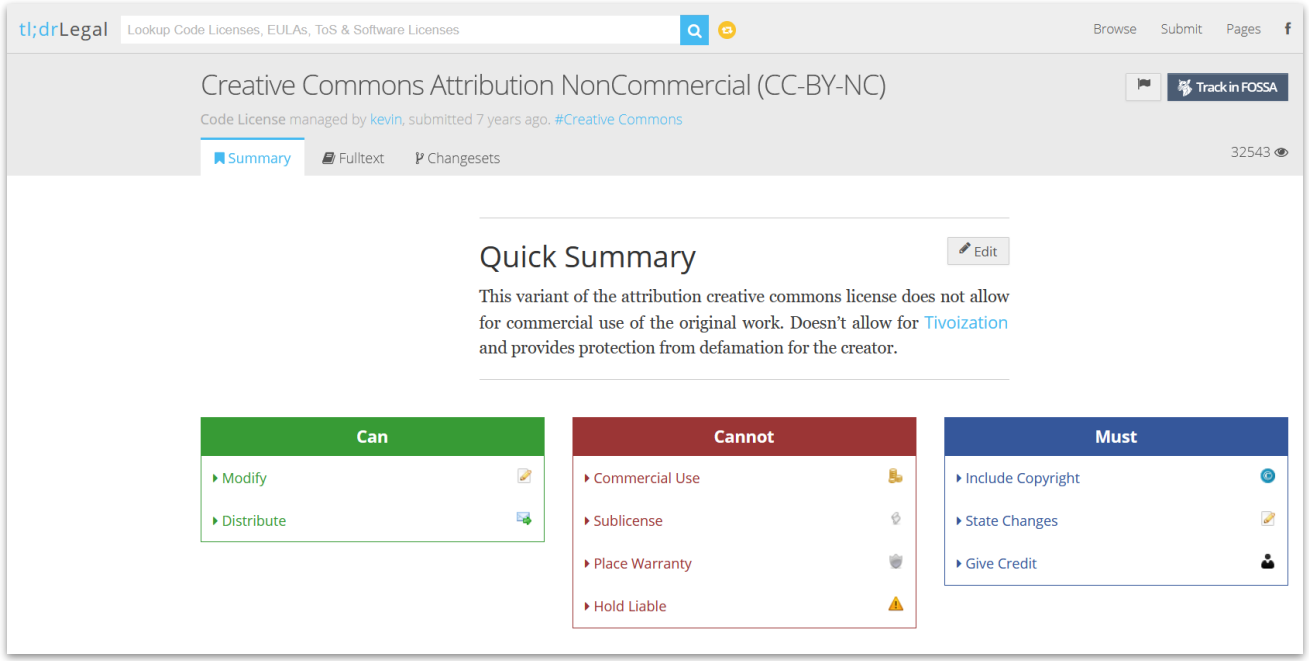


Let's find out [5']

Check out these licenses:

- ❑ **CC0**
- ❑ **CC BY-4.0**
- ❑ **MIT**
- ❑ **CC BY-NC-ND**

tldrlegal.com



Data / Code licences

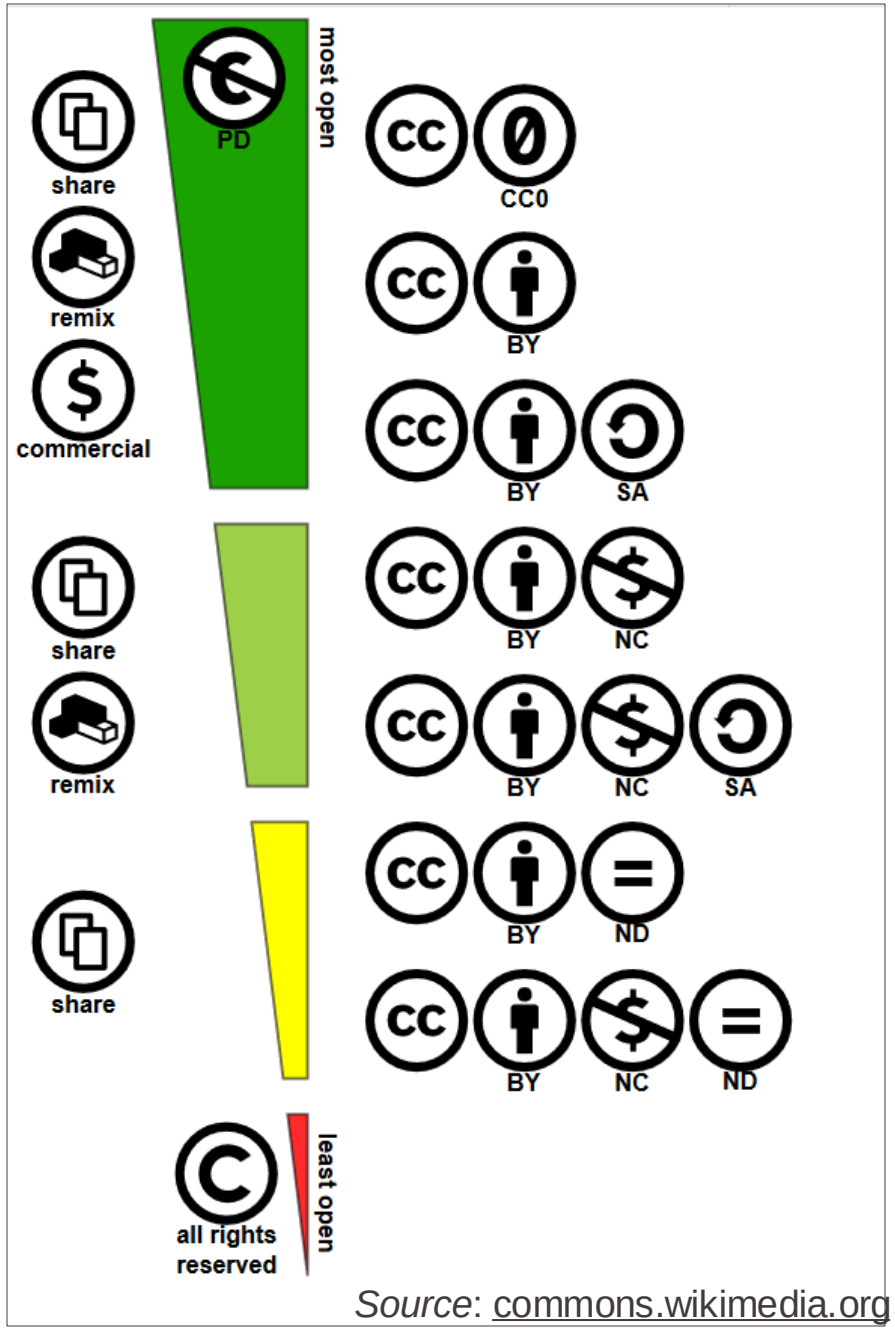
COMMONS LICENSES FOR
DATA, **TEXT** & **MULTIMEDIA**

Creative Commons:

- Enforced by the author
- Check platform's policy
- On datasets (no data points)

The **Open Data Commons** can be a viable option

The 96/9/EC Directive protects only vs. "substantial" copies of datasets



Source: commons.wikimedia.org

FastGuide on Data & Code Licensing ☺ Research Data Rights Management Guide (ANDS)

Data / Code licences

MORE SPECIFIC FOR
CODE

- **GNU-GPL** (Open Software)
- **Apache2.0** (smaller codes, libraries)
  - Permissive
  - No share-alike clause
  - Preservation of copyright notice
- **BSD-3clause** – Similar
- **MIT License** – GPL compatible

FastGuide on Data & Code Licensing ☺ Research Data Rights Management Guide (ANDS)

Importance of licensing

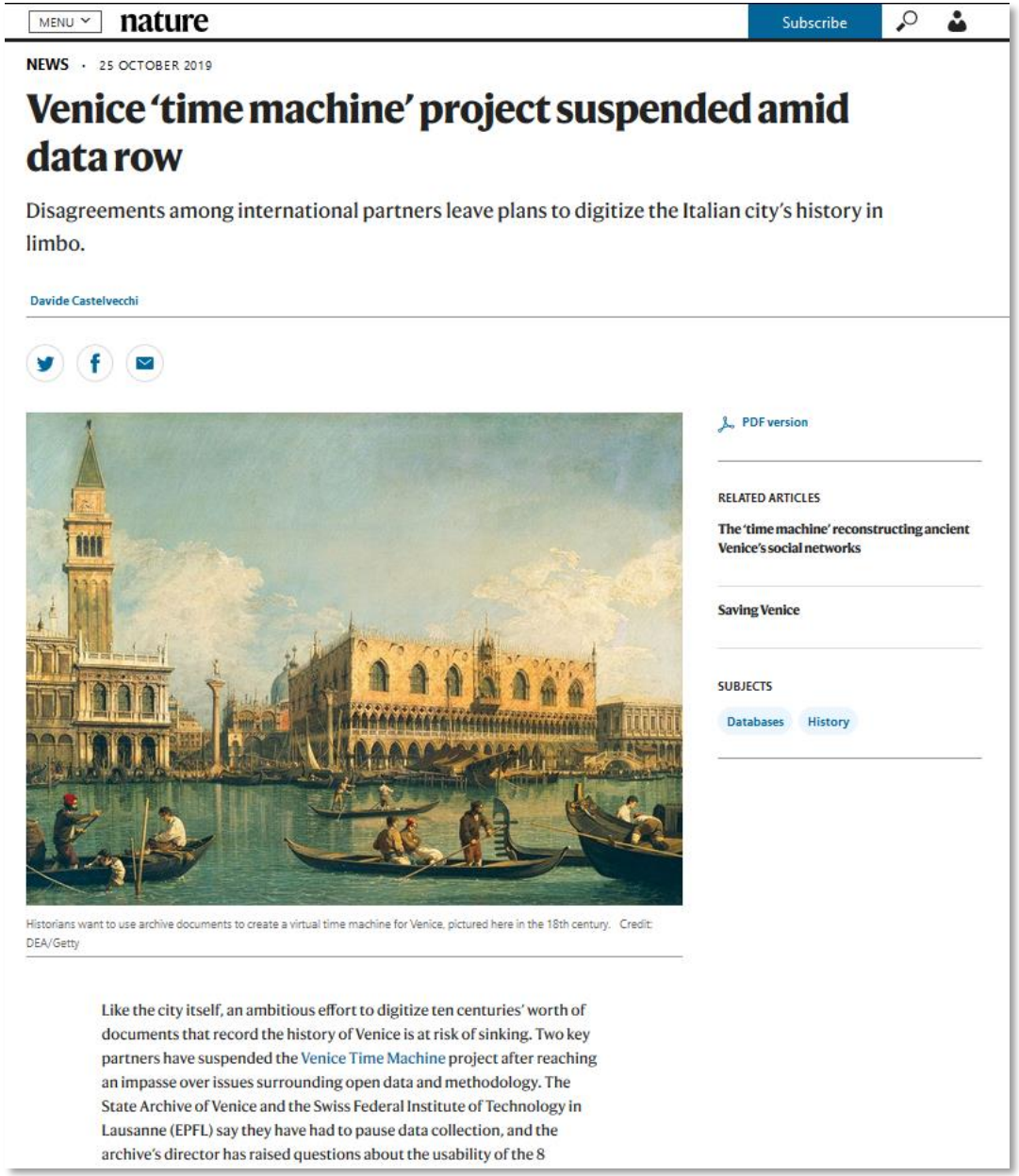
2012 – Project of officially **launched**:
Venice's State Archive + Ca' Foscari Univ. + EPFL (DHLAB)

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

Discussion: sensitive data in research?

Alain Borel, Francesco Varrato

- Do you collect, process or store data which is**... sensitive?**
- Do you collect, process or store information on**... identifiable persons?**
- If yes, how do you inform persons/subjects on what you will be doing?

Discussion [5’]

Distinction

Personal data

Information that relates to an identified or identifiable individual.

Not only data that can directly identify a person (e.g., name) is considered personal data, but also **data that can make a person identifiable through the combination of data** (e.g., combining age, e-mail address and information related to usage of social networks may allow identifying a person).

Consider that information together with the means reasonably likely to be used by either you or any other person to identify that individual.

Sensitive data

Information that includes but is not limited to religious, ideological, political or trade union-related views or activities; **health, genetic, biometric**, or concerning the intimate sphere or the racial origin; ethnic data or social security measures; administrative or criminal proceedings and sanctions.

See also **Art. 3**.c of the FADP Swiss law.

Personal / Sensitive data processing

Any operation with personal data [...] in particular

- the collection
- storage
- use
- revision
- disclosure
- archiving
- or destruction of data

Swiss *Federal Act on Data Protection* (FADP)
(*Loi sur la Protection des Données* LPD), Art. 5

Protection: Personal data must be protected **against unauthorised processing** through adequate technical and organisational measures

FADP Art. 7-9

Disclosure: Making personal data **accessible**, for example:

- by permitting access
- transmission
- or publication

FADP Art. 5e

Alain Borel, Francesco Varrato

Collecting consent (online & offline)

Art. 7 Consent

[1] Research involving human beings may only be carried out if, in accordance with the provisions of this Act, the persons concerned have given their informed consent or, after being duly informed, have not exercised their right to dissent.

[2] The persons concerned may withhold or revoke their consent at any time, without stating their reasons.

Human Research Act, Art. 7

The consent must be:

- **Simple**
- **Understandable**
- **Adapted** to the subject (child, teenager...)

HRA, Art. 21-22

Alain Borel, Francesco Varrato

HREC Review Procedures

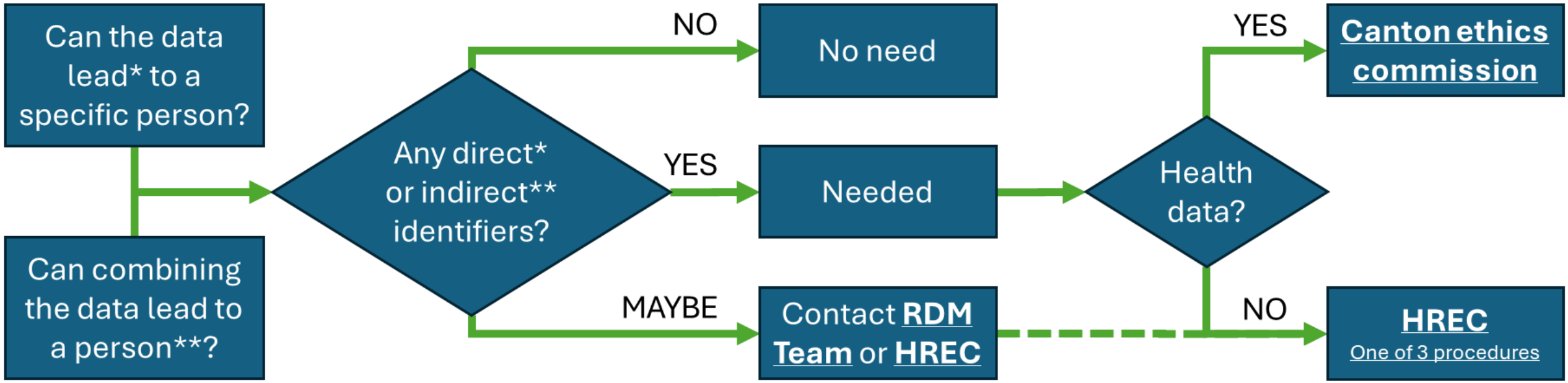
Standard Procedure	Simplified Procedure	Presidential Decision
▪ Projects with higher risks ▪ For example, involvement of vulnerable participants, sensitive data, physical risks, data protection risks	▪ Projects with lower risks For example, collection of non-sensitive personal data, testing of prototype ▪ Modification of authorized research projects, if they raise minor/ specific ethical, scientific or legal issues ▪ Sub-projects covered by an already approved general protocol	▪ Further use of data obtained with informed consent ▪ Research projects that do not raise specific ethical, scientific or legal issues

Source: three different review procedures on ReO webpage, https://drive.google.com/file/d/104o9imHvJ3tLp6c8gtQPZV7qvZJBPYuN/view

FastGuide on Personal Data Management and Data Masking ☺

Do you even need an Ethics Review?

(EPFL example decision tree)



*** Direct identifiers**: 1. Name; 2. Civil Identification Number; 3. Passport number; 4. Driver's license number; 5. Address details; 6. Email Address; 7. Phone number; 8. Fax Number; 9. Bank Account; 10. Vehicle identifiers and serial numbers, including license plate numbers; 11. Social Security Number; 12. Health Card Number; 13. Medical Record Number; 14. Device identifier and serial number; 15. Biometric identification codes, including fingerprints and voice prints, etc.; 16. Full face picture images and any other; comparable pairs of images; 17. Genetic information about a person; 18. Account number, certificate number; or license number; 19. Internet Protocol (IP) address number; 20. Web Universal Resource Locators (URLs)

**** Indirect identifiers**: 1. Gender; 2. Date of birth or age; 3. Date of event (e.g. admission, surgery, discharge, visit-related date); 4. Geographic range (e.g., zip code, building name, region); 5. Ethnic origin; 6. Nationality, place of origin; 7. Language; 8. Aboriginal Identity; 9. Visible minority status ; 10. Job title, work unit, department and other occupational information; 11. Marital Status; 12. Education level; 13. Years of schooling; 14. Total revenue; 15. Religious beliefs

Know more: HREC Review Procedures. Diagram source: EPFL Sr. Legal Counsel Valentin Conrad; HREC Procedures (Simplified, Standard, Presidential Decision)

Data masking techniques

Alain Borel, Francesco Varrato

Pseudonymization
(working data, reversible)

- **PSEUDONYMIZATION**
  Replace data by identifiers. The key is kept separately & securely
- **ENCRYPTION**
  Encrypt the data & keep the key secure. Also for long-term preservation, not data publishing

Some tools:
- **R package:** sdcMicro
- **Java application:** ARX Data Anonymization Tool
- **Java application:** ARGUS
- **Platform:** Amnesia

Anonymization
(published data, irreversible)

- **GENERALIZATION**
  Diminish granularity by generalizing the variables. Appropriate for data too specific or unique records
- **SUPPRESSION**
  Suppress data or part of the outlier records. Appropriate for processing identifiers
- **ADD FAKE DATA**
  To prevent the identification of specific records, add fake data while preserving correlations
- **SHUFFLE**
  Shuffle data over one / several columns without compromising the utility of the data

(Other: Differential Privacy, T-closeness, …)

Images:
- https://www.flaticon.com/packs/general
- https://www.flaticon.com/packs/hawcons-documents-filled

Deletion of identifying data

name	gender	city	age	disease
KELLER Anna	f	Basel	32	no diabetes
BRUNNER Emilia	f	Basel	37	diabetes 2
DURANT Pierre	f	Basel	44	no diabetes
GRAF Julia	f	Basel	45	diabetes 2
GERBER Fritz	m	Basel	20	diabetes 1
FISCHER Urs	m	Basel	23	diabetes 1
WYSS Emilien	m	Geneva	24	no diabetes
STEINER Leo	m	Geneva	28	no diabetes
ROTH Christian	m	Geneva	42	no diabetes
WYSS Rudolf	m	Geneva	48	diabetes 2

	name	gender	city	age	disease
0	*	f	Basel	30 - 39	no diabetes
1	*	f	Basel	30 - 39	diabetes 2
2	*	f	Basel	40 - 49	no diabetes
3	*	f	Basel	40 - 49	diabetes 2
4	*	m	Basel	20 - 29	diabetes 1
5	*	m	Basel	20 - 29	diabetes 1
6	*	m	Geneva	20 - 29	no diabetes
7	*	m	Geneva	20 - 29	no diabetes
8	*	m	Geneva	40 - 49	no diabetes
9	*	m	Geneva	40 - 49	diabetes 2

K-anonymity 2

Alain Borel, Francesco Varrato

Deletion of identifying data

name	gender	city	age	disease
KELLER Anna	f	Basel	32	no diabetes
BRUNNER Emilia	f	Basel	37	diabetes 2
DURANT Pierre	f	Basel	44	no diabetes
GRAF Julia	f	Basel	45	diabetes 2
GERBER Fritz	m	Basel	20	diabetes 1
FISCHER Urs	m	Basel	23	diabetes 1
WYSS Emilien	m	Geneva	24	no diabetes
STEINER Leo	m	Geneva	28	no diabetes
ROTH Christian	m	Geneva	42	no diabetes
WYSS Rudolf	m	Geneva	48	diabetes 2

	name	gender	city	age	disease
0	*	f	*	30 - 39	no diabetes
1	*	f	*	30 - 39	diabetes 2
2	*	f	*	40 - 49	no diabetes
3	*	f	*	40 - 49	diabetes 2
4	*	m	*	20 - 29	diabetes 1
5	*	m	*	20 - 29	diabetes 1
6	*	m	*	20 - 29	no diabetes
7	*	m	*	20 - 29	no diabetes
8	*	m	*	40 - 49	no diabetes
9	*	m	*	40 - 49	diabetes 2

L-diversity 2

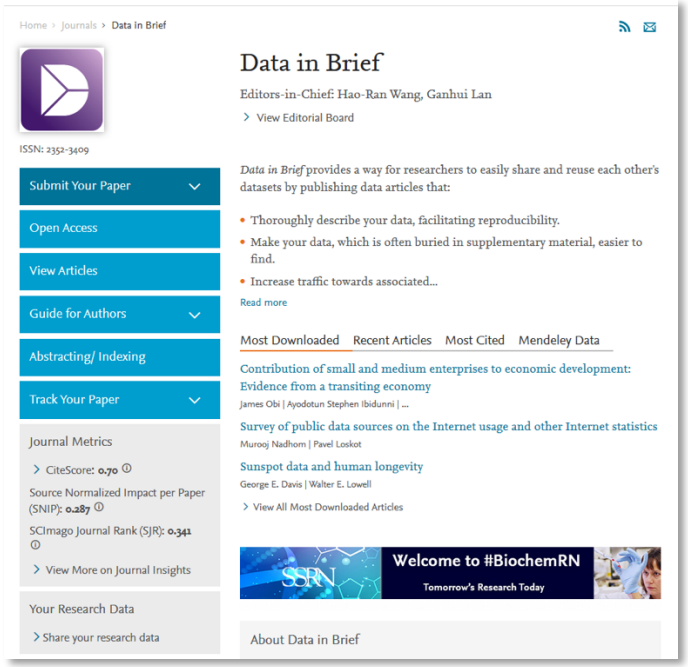
Alain Borel, Francesco Varrato

Data publication: journal pathway

Alain Borel, Francesco Varrato

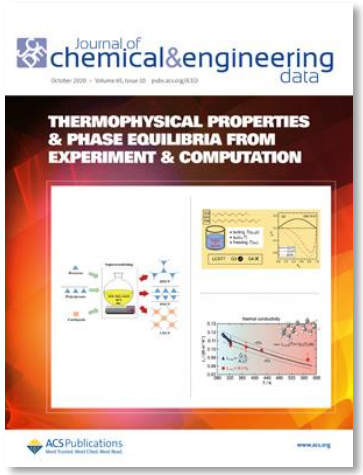
Data Papers

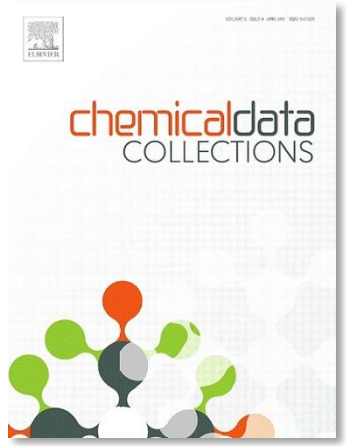
A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data (GBIF, 2019)



Data Journals

Data papers are supported by many journals, some of which are "pure", i.e. they are dedicated to publish data papers only, while others – the majority – are "mixed", i.e. they publish a number of articles types including data papers. (Wikipedia, 01.04.2019)





F1000Research
Open for Science

Data / Code journals

go.epfl.ch/datajournals

etc.

JOURNAL	OA TYPE	FOCUS	DISCIPLINE(S)	APC (* = Waiver policy)	EPFL LIBRARY AGREEMENT
Computational Engineering and Physical Modeling (Pouyan Press, Iran)	Diamond	Code	Electrical engineering, Electronics, Nuclear engineering, Computer engineering	0	
Digital Humanities Quarterly (Alliance of Digital Humanities Organizations, Netherlands)	Diamond	Code	Humanities, Philology, Linguistics, Mass media	0	
Earth system science data (Copernicus Publications, Germany)	Diamond	Data	Environmental science, Earth science, Geology	0	
Image Processing On Line (Image Processing On Line, France)	Diamond	Code	Mathematics, Computer science	0	
International Journal of Data and Network Science (Growing Science, Canada)	Diamond	Both	Social Sciences, Management, Industrial management	0	
Journal of Data and Information Science (Sciendo, Poland)	Diamond	Code	Technology, Industrial engineering, Management engineering, Information technology, Mathematics, Computer science	0	
Journal of data science (School of Statistics, Renmin University of China, China)	Diamond	Code	Mathematics, Computer science, Mathematical statistics	0	
Journal of Statistics and Data Science Education (Taylor & Francis Group, USA)	Diamond	Code	Mathematics, Mathematical statistics, Education	0	
Journal of Open Source Software (Journal of Open Source Software, USA)	Diamond	Code	Science, Computer Science	0	

Alain Borel, Francesco Varrato

Alain Borel, Francesco Varrato

Data publication

Data papers can be highly cited (Scopus)

3,667 documents found

Analyze results

All | Export | Download | Citation overview | ••• More | Show all abstracts | Sort by Cited by (highest)

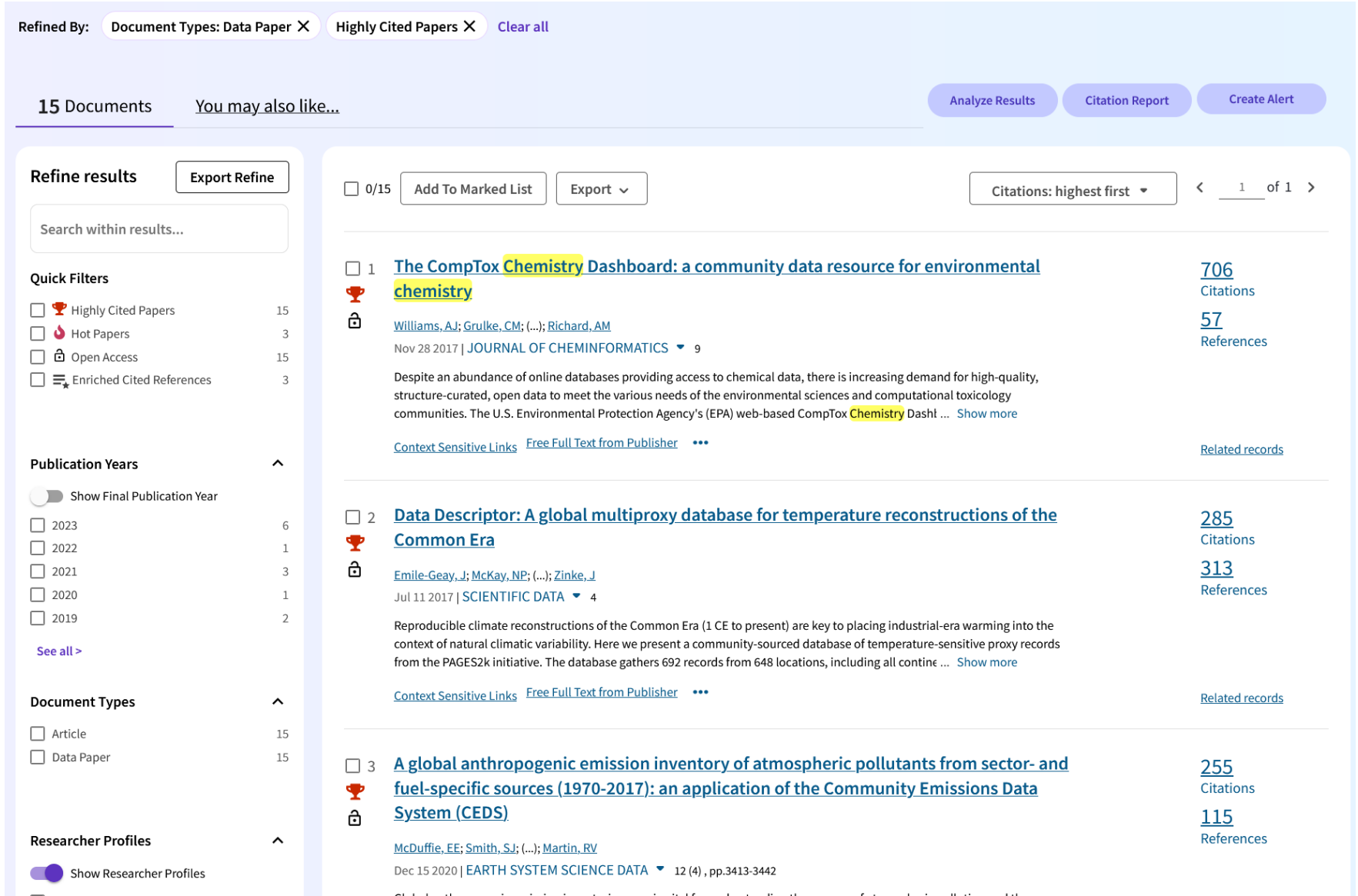
	Document title	Authors	Source	Year	Citations
1	Data Paper • *Open access* **ERA5-Land: A state-of-the-art global reanalysis dataset for land applications** Show abstract | Full text at EPFL Library | View at Publisher | Related documents	Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., ... Buontempo, C., Thépaut, J.-N.	Earth System Science Data, 13(9), pp. 4349–4383	2021	987
2	Data Paper • *Open access* **Global Carbon Budget 2021** Show abstract | Full text at EPFL Library | View at Publisher | Related documents	Friedlingstein, P., Jones, M.W., O'Sullivan, M., ...Zaehle, S., Zeng, J.	Earth System Science Data, 14(4), pp. 1917–2005	2022	590
3	Data Paper • *Open access* **China CO_2 emission accounts 2016–2017** Show abstract | Full text at EPFL Library | View at Publisher | Related documents	Shan, Y., Huang, Q., Guan, D., Hubacek, K.	Scientific Data, 7(1), 54	2020	513

Advanced search on Scopus with query: ALL (chemistry) AND (LIMIT-TO (DOCTYPE , "dp"))

Data publication

Alain Borel, Francesco Varrato

Data papers can be highly cited (WoS)



Alain Borel, Francesco Varrato

Data repositories: Publication and/or Preservation

Search here...

PLATFORM (*= Institutional)	TYPE(S) (* = For-Profit)	RECOMMENDED BY SNSF/EU	DISCIPLINE(S)	HOSTING	DOI	MAX SIZE
ACOUA* EPFL	Archive		All	CH	1	10 TB
ArrayExpress	Data Repository	SNSF, EU	Genetics, Biology, Life Sciences	USA/EU/UK	0	N/A
BORIS Portal*	Archive, Data Repository	SNSF	All	CH	1	No limit
c4Science* EPFL	Code Repository		All	CH	0	N/A
CERN Open Data*	Data Repository		High Energy Physics, Condensed Matter Physics, Physics	EU	1	N/A
Channelpedia EPFL	Databank		Electrophysiology, Physiology	CH	0	N/A
Copernicus	Data Repository		Geosciences, Ecology, Atmospheric Science	EU	0	N/A
DaSCH	Data Repository	SNSF	Humanities, Social Sciences, Linguistics	CH	0	N/A
dbGap	Data Repository		Genetics	USA	0	N/A

go.epfl.ch/**datarepo**

+ SNSF Open Data criteria
+ Europe-approved repositories
+ List of Nature's Recommended Data Repositories per discipline
+ (Non-exhaustive) list of repositories approved by some publishers for hosting data alongside the articles

How to choose a data repository

1. Listed on re3data: for peace of mind
2. DOI or other Persistent IDentifier
3. Non-profit: SNSF doesn't reimburse
4. Good licenses choice: reuse & compliancy
5. Cross-linking: dataset / code ↔ article
6. Target public: field-specific *and/or* generic
7. Max upload: size matters

GitHub is **not a data repository**
Your own website is **not a data repository**
You can choose **both a GENERIC & field-SPECIFIC data repository**

Alain Borel, Francesco Varrato

Which data repository? Try re3data [5']

Apply filters: *subj.* (**Chemistry**) & *country* (**EU** + **CH**)

Found 3 result(s)

CARIBIC

Civil Aircraft for the Regular Investigation of the atmosphere Based on an Instrument Container

Subject(s): Atmospheric Science | Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy | Analytical Chemistry, Method Development | Geosciences (including Geography) | Geophysics and Geodesy | Chemistry

Content type(s): Plain text | Raw data | Scientific and statistical data formats | other

Country: Germany | United Kingdom | Switzerland | France | European Union | Netherlands | Sweden

CARIBIC is an innovative scientific project to study and monitor important chemical and physical processes in the Earth´s atmosphere. Detailed and extensive measurements are made c
scientific apparatus which are connected to an air and particle (aerosol) inlet underneath the aircraft. We use an Airbus A340-600 from Lufthansa since December 2004.

Network for the Detection of Atmospheric Composition Change

NDACC

Subject(s): Geosciences (including Geography) | Particles, Nuclei and Fields | Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy | Atmospheric Science and Oceanography | Natural Sciences | Physics | Geophysics and Geodesy | Chemistry

Content type(s): Plain text | other

Country: United States | Germany | Belgium | Switzerland | European Union | International

The Network for the Detection of Atmospheric Composition Change (NDACC), a major contributor to the worldwide atmospheric research effort, consists of a set of globally distributed re
of atmospheric trace gases, particles, spectral UV radiation reaching the Earth's surface, and physical parameters, centered around the following priorities.

Rhea

Subject(s): Basic Biological and Medical Research | Bioinformatics and Theoretical Biology | Metabolism, Biochemistry and Genetics of Microorganisms | Microbiology, Virology and Immunology | Medicine | Chemistry | Natural Sciences

Content type(s): Standard office documents | Plain text | Scientific and statistical data formats | Structured graphics | Structured text | Software application

Country: European Union | Switzerland

Rhea is a freely available and comprehensive resource of expert-curated biochemical reactions. It has been designed to provide a non-redundant set of chemical transformations for appl
and metabolic network reconstruction. There are three types of reaction participants (reactants and products): Small molecules, Rhea polymers, Generic compounds. All three types of re
Biological Interest) which provides detailed information about structure, formula and charge. Rhea provides built-in validations that ensure both mass and charge balance of the reactions
classification (i.e. in the IntEnz and ENZYME databases), extending it with additional known reactions of biological interest. While the main focus of Rhea is enzyme-catalysed reactions,
"spontaneous") also are included.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search... Search

SNSF recommends using
re3data.org vertical search engine

Data repositories (Example)

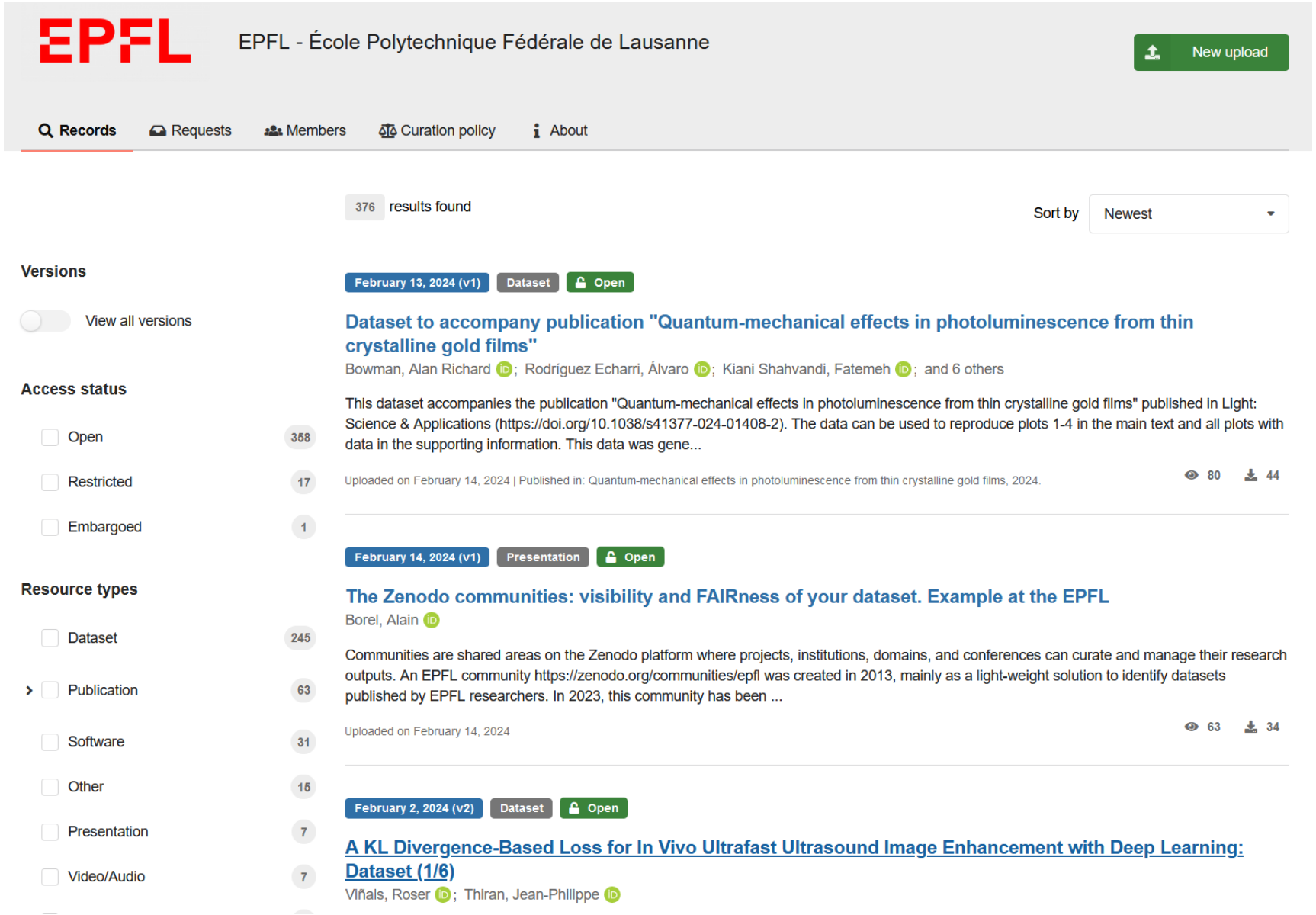
- Hosted by the CERN
- Free of charges
- Max 50GB/dataset
- Unlimited datasets
- Automated DOI assignment

- OpenAIRE integration (EC reporting)
- GitHub integration
- ORCID integration

- All file formats accepted
- Usage statistics interface
- OAI-PMH protocol (content harvesting)
- 18 petabytes disk cluster
- Each file has 2 replicas on different servers
- 2 independent MD5 checksums per file
- Metadata 12-hourly backup cycle
- ...

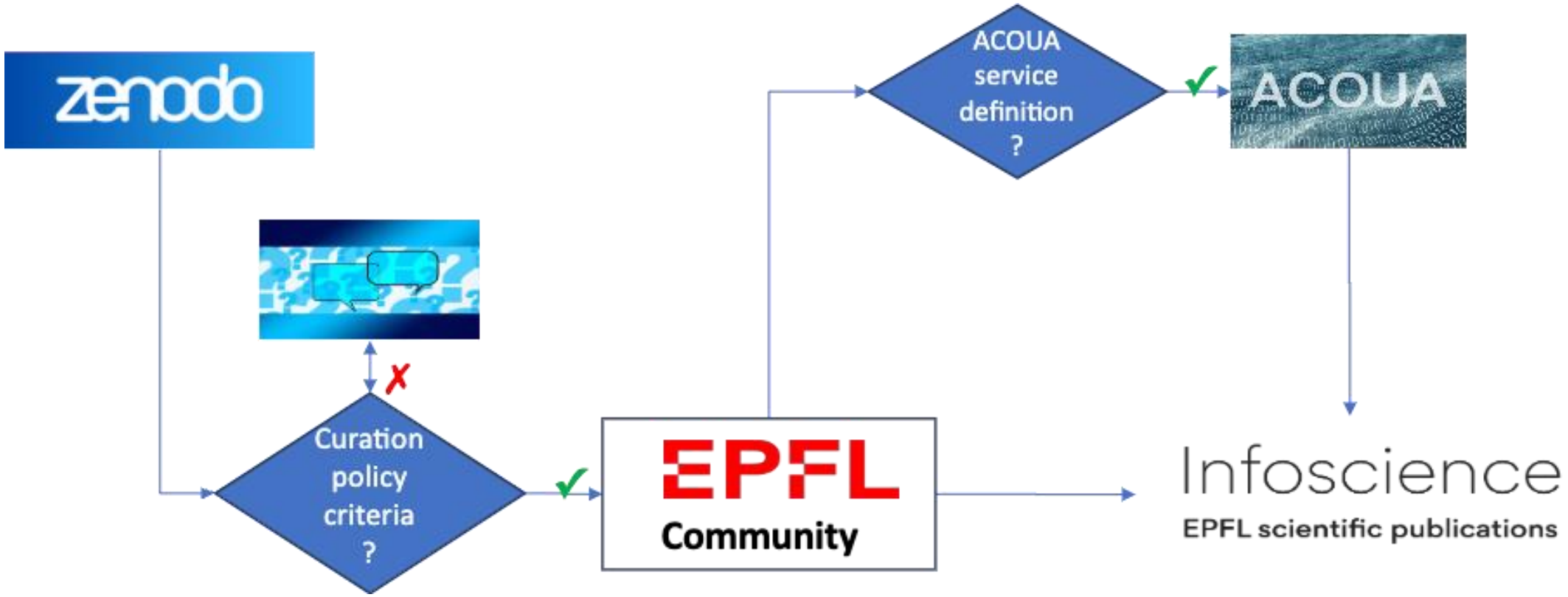
Zenodo.org has an EPFL Community!

Data repositories (Example)



Alain Borel, Francesco Varrato

Zenodo curation: FAIRness is a facilitator

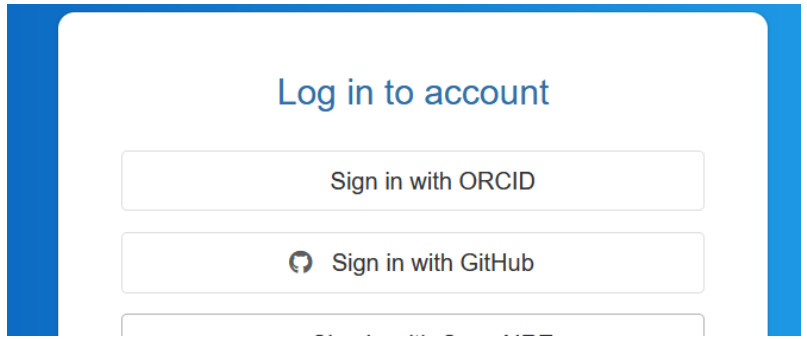


Alain Borel, Francesco Varrato

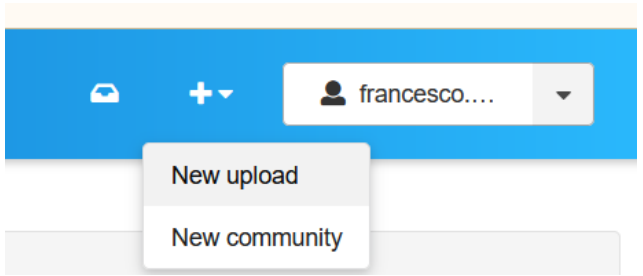
Hands-on: Zenodo sandbox

Alain Borel, Francesco Varrato

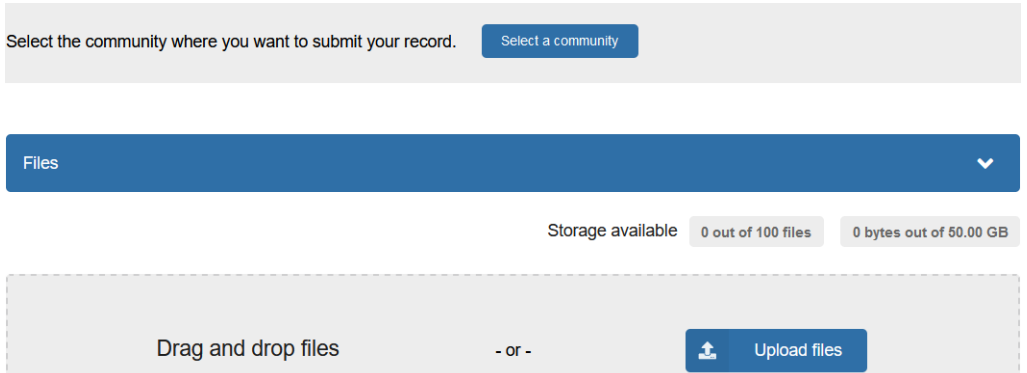
STEP 1
Login



STEP 2
Start upload



STEP 3
Upload dataset + Fill all useful fields



sandbox.zenodo.org

You can use a dataset from *Day 2* on **Moodle**

Publication ≠ Preservation

ACOUA: Long-term preservation

Why using it

- archive the entire dataset underlying a publication
- archive datasets of a finished research project
- archive datasets of a collaborator leaving EPFL
- get space for large datasets that need preservation
- preserve raw data useful during your research
- get expert support for data curation

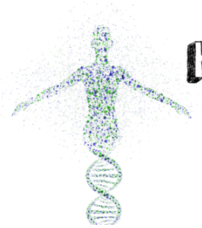
What you get

- **trustworthy**, safe and EPFL-backed environment
- **free** for EPFL researchers
- up to **10TB** per archived dataset
- help in **data curation** prior to archival
- periodic **integrity audits** of your datasets
- periodic **reports** on your preserved datasets
- **referral** of your datasets on Infoscience
- can **publish** your datasets on Zenodo (size limits)

go.epfl.ch/acoua

Data access sustainability

- More than 60% of links to astronomy datasets are **broken after 10 years**
- The bibliography of **1 out of every 5 is impacted** by this phenomenon



Human Genome Project

A "*good example of a large-scale research endeavour in which an openly accessible data repository is being used successfully*" [OECD]

REPORT | VOLUME 24, ISSUE 1, P94-97, JANUARY 06, 2014

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines • Arianne Y.K. Albert • Rose L. Andrew • ... Jean-Sébastien Moore • Sébastien Renaut • Diana J. Rennison • Show all authors

Open Archive • Published: December 19, 2013 • DOI: https://doi.org/10.1016/j.cub.2013.11.014 •

Check for updates

Highlights

- We examined the availability of data from 516 studies between 2 and 22 years old
- The odds of a data set being reported as extant fell by 17% per year
- Broken e-mails and obsolete storage devices were the main obstacles to data sharing
- Policies mandating data archiving at publication are clearly needed

Data degradation problem

Alain Borel, Francesco Varrato

CERN

A 2007 study showed that a bitrot error ratio of 10^{-7}(over 2 months)

Ex.: ~$10^{9} \cdot 10^{-7} = 10^{2}$ = 100 bytes of bitrot every 1GB (1024MB)

US FDA

In 2017 the agengy added data integrity requirements for the drugs industry (FDA 21 CFR, 11 & 211)

Data integrity failure (Possible causes)

- Processing
  CPU heat, encryption errors, ...
- Transfer
  Network failures, backup errors, ...
- Read / Write
  Single bits errors at RAM or ROM levels
- Storage
  Aging, background radiation, ...

Countermeasures (Data repositories)

- Redundant hardware
- Uninterruptible power supply
- Certain types of RAID arrays
- Radiation hardened chips
- Error-correcting memory
- Clustered file system
- File systems with block level checksums

Access sustainability without degradation

Alain Borel, Francesco Varrato

SNSF

Researchers must share data "*according to the FAIR Data Principles on* ***publicly accessible, digital repositories***."

ERC

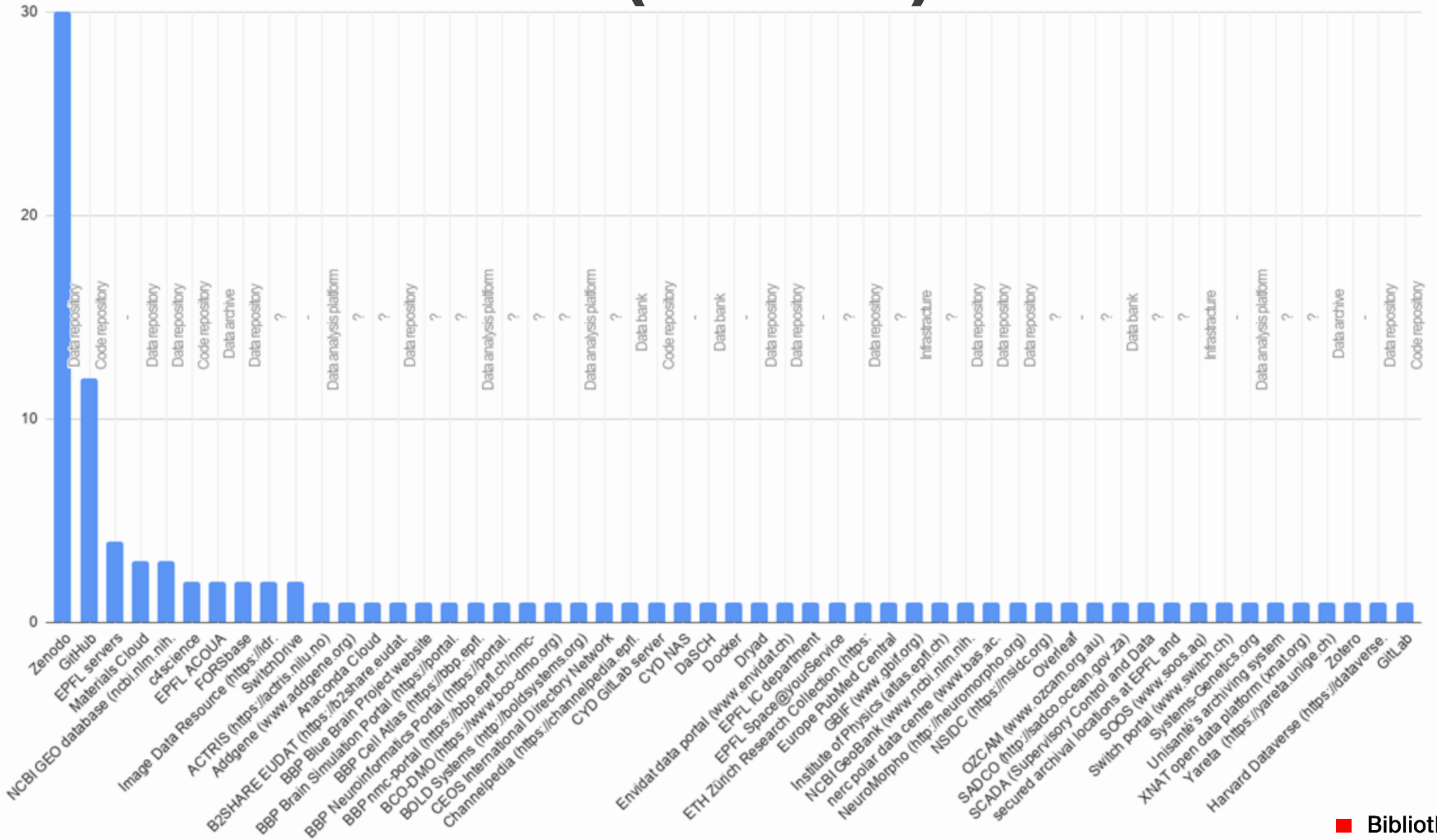
Researchers must "*deposit research data [...], including associated metadata,* ***in the repository as soon as possible***."

Back-up ≠ Publication ≠ Preservation

	BACKUP	PUBLICATION	LONG-TERM PRESERVATION
ACTIVE DATA	✓	✗	✗
DATA RECOVERY	✓	✓	✓
INTEGRITY (monitoring, repair, authenticity)	?	?	✓
APPRAISAL (what & for how long)	✗	✓ ✗	✓
PERMANENT IDENTIFIERS	✗	✓	✓ ✗
DESCRIPTION (metadata)	✗	✓	✓
RENDERABILITY (format migration, virtualization)	✗	?	✓

Alain Borel, Francesco Varrato

Preservation statements (from DMPs)



Alain Borel, Francesco Varrato

FAQ End of thesis

Where should I archive the data and code that support my thesis?

Recommendation

You are recommended (1, 2 and 3) to:

- archive the Research Data necessary to make your thesis reproducible
- whenever legally possible, provide the jury president and members with access to the archived datasets.

What

Research Data includes code and is defined as evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). **Datasets** can be composed by research data necessary to validate the findings of your thesis, such as:

- raw data
- pre-processed data
- processed data
- plots
- source code
- executables documentation (e.g., README files, protocols, parameter files, log files, etc.).

How

Research Data should be saved in a digital archive that allows for its **long-term preservation** and retrieval. To ensure that the datasets will remain usable in the future, **data curation** prior to archiving is recommended:

- cleaning the datasets
- documenting
- enriching metadata
- converting proprietary formats into open formats
- restructuring the dataset and naming etc.

Where

EPFL offers a **free archiving service** for Research Data: ACOUA, the ACademic OUtput Archive. For support and information, contact the Research Data team of EPFL Library.

How long

Archived datasets can be safely stored and retrieved for many years. A basic recommendation is to make datasets preserved for at least 10 years. Depending on the research funder (SNSF, ERC, etc.) there might be specific duration requirements.

If your datasets include personal data or health data (e.g. clinical trials), both the archiving and the associated retention duration can be legal requirements. For more information in such cases, contact the EPFL Human Research Ethics Committee (HREC).

www.epfl.ch/education/phd/regulations/internal-regulations/edoc-faq-end-of-thesis

How to cite data(sets)?

Same as any other citation:

- **Author(s)** of the dataset
- **Title** of the dataset / study
- **Year** of online publication
- **Publisher** responsible for distributing the dataset
- **Edition / Version** number associated with the dataset
- **Persistent identifier(s)** as URI, DOI, ORCID, ...
- Link to **related objects** (paper, poster, other datasets, code)
- Using citation management software (e.g. Zenodo, Mendeley…)

DC1 Data Citation Principles

Source: *www.force11.org/datacitation*

Exploring data citations (examples)

- Google Dataset Search
- OpenAIRE Graph

Source: *www.force11.org/datacitation*