EPFL

Bibliothèque

Which platform to publish my research data / code on?

Francesco Varrato

EPFL Library

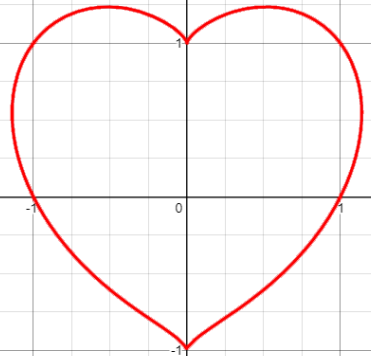
Love Data Week 2022

Feb. 17, 2022

École polytechnique fédérale de Lausanne

Love Data Week

EPFL 2♡22 edition



go.epfl.ch/quiz

Fri. Feb. 18 | 10:00-12:00
Online speed-dating with EPFL Data Champions

go.epfl.ch/ldw22

LDW22

Bibliothèque

Content

- Context (15’-20’)
- Demo: Zenodo (~10’)
- Demo: Github → Zenodo (~5’)
- Q&A (10’-15’)

Specific questions from you

Francesco Varrato

Do you have a question you would like us to address during the training?

Licenses

Do you have a question you would like us to address during the training?

Can you kindly elaborate the concept of archive and when to publish on archive? (arXiv.org)

Look for it:

Why?

To avoid this… among other things

Source: https://twitter.com/robertnulrich/status/1490567709188182016

Francesco Varrato

Pain points / Needs

Francesco Varrato

- Scholarly **Communication** shifts from articles/papers to richer content.
- **Dissemination** of data and code is crucial, but also in its infancy.

Which platform is best suited to
disseminate
research data or code?

Keywords:
(A) Disseminate / Communicate
(B) Data / Code
(C) Platform

(A) 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

(A) 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

(A) Dissemination as publication

Francesco Varrato

Sharing

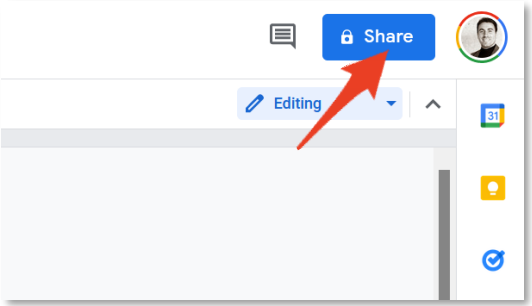
Making data available to others, also during the research, no data curation implied

Publishing

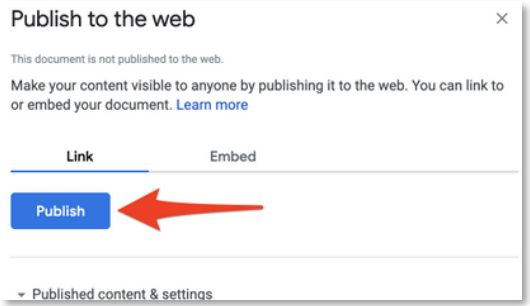
Publicly sharing a dataset as scholarly product: curated, preserved, *publicly* findable and accessible (DOI)

All can see it

Others can see it



VS.



(B) Data / Code etc.

Francesco Varrato

DATA
Factual records: numerical scores, textual records, images, sounds, protocols, **source code**, ...

RESEARCH DATA
Data used as primary sources for scientific research, and commonly accepted in the scientific community to validate research findings (OECD)

RESEARCH DATA MANAGEMENT (RDM)
The care and maintenance of research data during the research cycle (UC Berkeley Library)

OPEN RESEARCH DATA (ORD)
Research data that managed during its **lifecycle** to comply with the **FAIR** principles

RDM and ORD include legal & ethical aspects

(B) What to publish?

"The SNSF expects that researchers share at least the data underlying their publications, but only to the extent to make the published results reproducible"

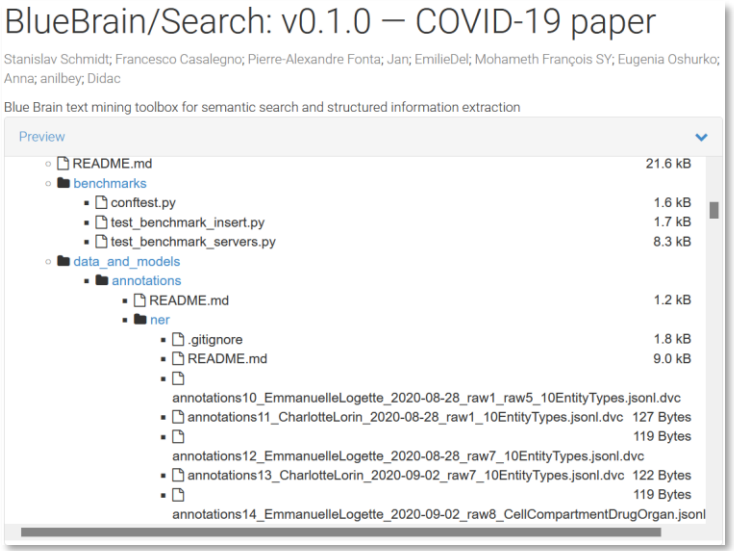
SNSF DMP Guidelines

What is needed to reproduce the results?

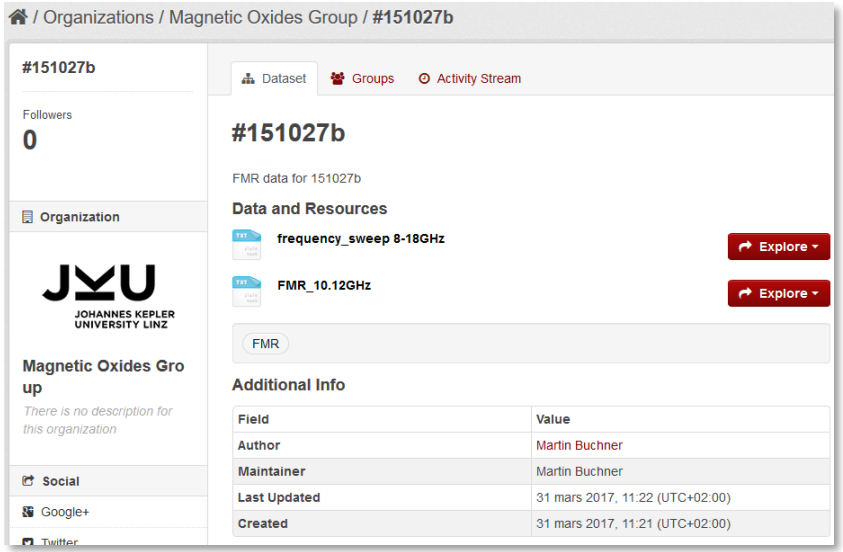
DATASET

+ Raw data
+ Processed data
+ Code
+ Documentation
  - README
  - Protocols
  - parameter files
  - DMP
  - …

(B) Not everything is… … publishable



doi.org/10.5281/zenodo.4563998

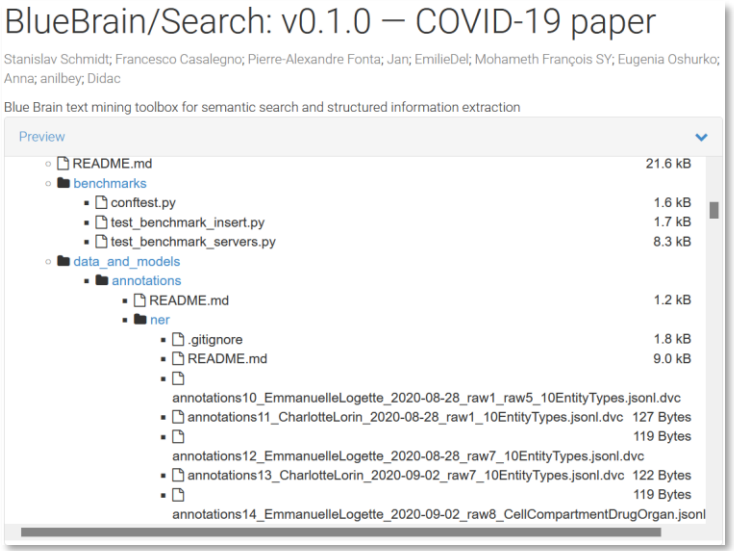


ord.fkp.jku.at/dataset/151027b

- Documentation
- Consistent structure and naming
- Code
- Did I say documentation?

- No DOI
- No documentation
- A couple of txt files with mysterious labels
- No much else

(B) Not everything is…
… publishable



doi.org/10.5281/zenodo.4563998

Steve Berry, www.flickr.com CC BY-NC-SA 2.0

- Documentation
- Consistent structure and naming
- Code
- Did I say documentation?

- No
- No
- A s
- No much else

(B) Constraints to dataset publication

Francesco Varrato

SENSITIVE

- Tests on **animals** / humans
- **Personal data** (Federal Act on Data Protection, Human Research Act, GDPR)

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

COMMERCIAL

- **3rd party data** (e.g. commercial datasets, research cooperations, etc.)

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

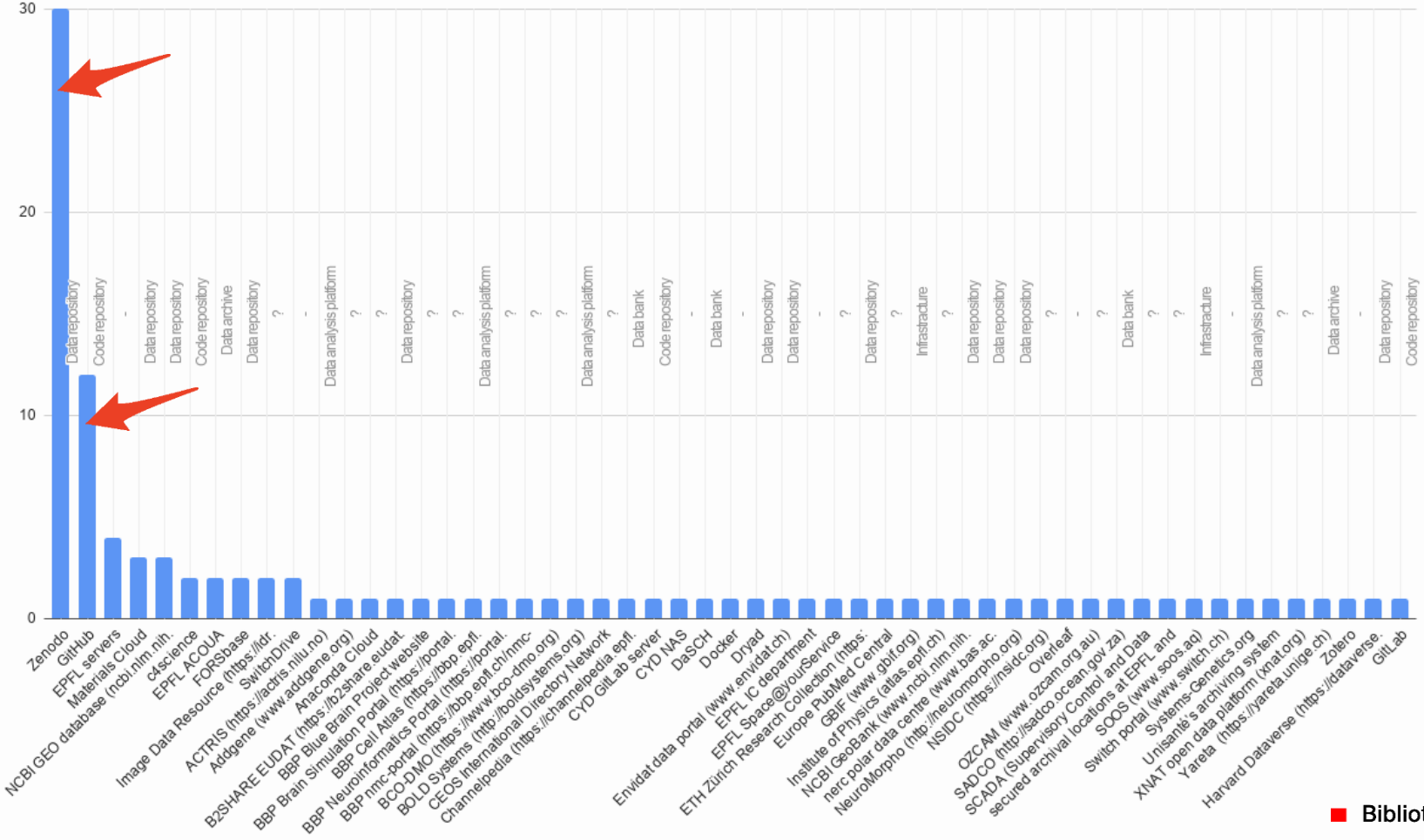
- Potential **commercial / industrial** exploitation or patents

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

(C) Platforms: Where would/do you publish a dataset?

Whiteboard

(C) Publication statements from 60 real DMPs



(C) Platforms: Definitions and… nuances

Francesco Varrato

go.epfl.ch/FORCE21epfl

01 - DEFINITIONS

02 - DATA REPOSITORY

03 - DATA ARCHIVE

04 - DATA BANK

~~05 - CLOUD STORAGE~~

06 - CODE REPOSITORY

~~07 - ELN/LIMS~~

~~08 - DATA ANALYSIS PLATFORM~~

~~09 - IT INFRASTRUCTURE~~

10 - OTHER

	A	B	C	D	E	F	G	H
1	TYPE	SHORT DEFINITION (main use)	SHORT DEFINITION (secondary use)	Data (D) / Code (C) / Both (B)	SOME FEATURES	DATA "TEMPERATURE"	RETRIEVAL FREQUENCY	PROVIDI (DOI, e
2	02 - Data repository	Store Retrieve Preserve Discover	Publish, store	B	Online infrastructure where researchers can submit their data. Allows to manage, share, and access datasets for the long-term. Can be generic (discipline-agnostic) or specialized (dicipline-specific): specialized data repositories often integrate tools and services useful for a discipline. The main purpose is to preserve data to make it reusable for future research. Data repositories may have specific requirements concerning: research domain; data re-use and access; file format; data structure; types of metadata. They can also have restrictions on who can deposit data, based on: funding; academic qualification; quality of data. Institutional data repositories may collect a university or consortium of	Warm / Cold	Medium	Mostly yes
3	03 - Data archive	Store Retrieve	-	D	Storage infrastructure enabling the long-term retention and reusability of data. Provides secure locations for	Cold	Low	Mostly no
	04 - Data bank							

	A	B	C	D	E	F	G	H	I	J
1		GENERAL						License & Pricing		
2	Name	URL	Domain / Technique	Managing institution	Country	Type	Type 2	Profit/ Non-4-profit	License	Pricing
4	Zenodo	https://zenodo.org/	Generic	CERN & OpenAIRE	EU/CH	02 - Data repository		N4P	Many (> 400, from SPDX.org)	free
5	Figshare	https://figshare.com/	Generic	Digital Science	UK/self-hosting	02 - Data repository		P	CC	free
6	Dryad	https://datadryad.org/stash	Generic	Springer Nature	USA	02 - Data repository		N4P	CC	multiple
7	Dataverse	https://dataverse.harvard.edu/	Generic	Harvard University	USA/international	02 - Data repository		N4P	CC0	free
8	FORSbase	l.ch/	Social sciences	UNIL	CH	repository		N4P	?	free
9	IEEE Dataport	dataport.org/	IEEE scope	IEEE	USA	repository		N4P	CC	multiple
10	MaterialsCloud	https://www.materialscloud.org/	Materials Science, Chemistry, Physics	MARVEL NCCR	CH	02 - Data repository		N4P	CC-BY-SA	only, see https://archive.materialsclou
11	DaSCH	https://dasch.swiss/	humanities	Center for the Humanities, unibas	CH	02 - Data repository		N4P	CC	free ???
12	Mendeley Data	https://data.mendeley.com/	Generic	Elsevier	NL	02 - Data repository		P	CC	free
13	EnviDat	t.ch/	Environment	WSL (ETH-Domain)	CH	repository		N4P	Many	free
14	European Nucleotide Archive (EBI ENA)	.uk/ena/browser/home	nucleotide sequencing information	EMBL-EBI	UK/EU/International	02 - Data repository		N4P	CC	free
15	NCBI GEO	m.nih.gov/gds/	expression	NCBI	USA	repository		N4P	Many	free
16	Neuromorpho	neuromorpho.org	curated neurons	George Mason University et al.	USA	02 - Data repository		N4P	Other	free
17	NCBI dbGap	http://www.ncbi.nlm.nih.gov/gap	curated genomes and phenomes	NLM	USA	02 - Data repository		N4P	PublicDomain, Other	free
18	Openneuro.org	openneuro.org	Neuroimaging Data	Standford	USA	02 - Data repository		N4P	CC0, ODC, CC	free
19	PRIDE EMBL	.uk/pride/	Proteomics data biological molecules,	EMBL-EBI	EU	repository		N4P	CC	free

01_DEFINITIONS | 02_DATA_REPO | 03_DATA_ARCHIVES | 04_DATA_BANKS | 05_CLOUD_STORAGE | 06_CODE_REPO | 07_ELN-LIMS | 08_ANALYSIS_PLATF

EPFL-centered: surveys + interviews + DMP reviews

(C) Data Repositories

- Both data & code
- Publication: yes
- Provide a DOI: Usually
- Collaborative: not really
- Institutional or commercial
- Non-profit or for-profit
- Discipline-specific or generic

Name
Zenodo
Figshare
Dryad
Dataverse
FORSbase
IEEE Dataport
MaterialsCloud
DaSCH
Mendeley Data
EnviDat
European Nucleotide Archive (EBI ENA)
NCBI GEO
Neuromorpho
NCBI dbGap
Openneuro.org
PRIDE EMBL
PDB Protein Database
The International Arctic Systems for Observing the Atmosphere (IASOA)
NOAA
UCI Machine Learning Repository
PhysioNet
Environmental Data Initiative
ERIC Open
The Cambridge Structural Database (CSD)
NCBI SRA
ProteomeXchange
Image Data Resource (IDR/OMERO)
Copernicus
ScienceBase-Catalog
European Centre for Medium-Range Weather Forecasts (ECMWF)
NASA
ETH Research Collection
BORIS-Portal

(C) Data Archives

Name
OLOS (DLCM)
Data Reception and Archive at the University of Bern
EPFL ACOUA
ETH Data Archive
European Genome-phenome Archive (EGA)

- Both data & code
- Publication: not always
- Provide a DOI: Sometimes
- Collaborative: not really
- Usually institutional
- Usually non-profit
- Usually generic

… concept of archive and when to publish on archive?

(C) Data Banks

- Especially data
- Publication: not always
- Provide a DOI: not really
- Collaborative: might be
- Usually institutional
- Usually non-profit
- Usually discipline-specific

Name
opendata.swiss
EPFL Infoscience
DBnomics
Svalbard Integrated Arctic Earth Observing System (SIOS)
EarthCube
DataLakes
Channelpedia
SwissPalm
Portail cartographique du canton de Vaud
GeoVITe
ASIT
INSPIRE Geoportal
GéoportaiL
Geospatial Platform
NGA Raster Roam
Living Archive ENAC
The International Arctic Systems for Observing the Atmosphere (IASOA)

(C) Code Repository

- Especially code
- Publication: not always
- Provide a DOI: no
- Collaborative: usually
- Institutional or commercial
- Non-profit or for-profit
- Usually generic

Name
GitHub
GitLab
GitLab EPFL
c4Science
framagit
Bitbucket
Renku
NOTO EPFL

(C) Other

?

Name
Swiss Data Cube (SDC)
Papers with Code
Research Data Alliance
arXiv Code & Data
protocols.io
notion
Hydrological Processes journal (Wiley Publishers)
frontiers
Lab website
Microsoft OneNote
google sites
binder
Qiita
DVC
Snakemake
Aiida
Jupyter Lab
Makefiles

… concept of archive and when to publish on archive?

(C) Dataset as classic publication

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)

JOSS
The Journal of Open Source Software

protocols.io

SCIENTIFIC DATA

Papers With Code

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

etc.

(C) Demo: Zenodo

Communities

epfl

Showing 0 to 10 out of 14 communities.

Sort by

EPFL - École Polytechnique Fédérale de Lausanne View

Curated by: rdmepfl

CHILI Lab, EPFL View

Shared Resources for Computer Human Interaction for Learning and Instruction Lab at EPFL, Lausanne, Switzerland.

Curated by: StianHaklev

From Neuroscience to Machine Learning View

Presentations of the Human Brain Project SP4 / CPD5 workshop "From Neuroscience to Machine Learning" held at the European Institute for Theoretical Neuroscience (EITN) at Paris. Workshop Abstract: The workshop aims to bring together researchers from...

Curated by:

Digital Object Identifier e.g. 10.1234/foo.bar

Reserve DOI

Publication date *

Title *

Authors * Family name, given names | Affiliation | ORCID (e.g.: 0000-0002-1825-0097

+ Add another author

Description *

Source

Required.

Version

Keywords

Access right *
- Open Access
- Embargoed Access
- Restricted Access
- Closed Access

Licenses

License * Creative Commons Attribution 4.0 International

Grants European Commission (EU) | Start typing a grant number, name or abbreviation...

+ Add another grant

Related identifiers e.g. 10.1234/foobar.56789 | N/A

Optional. Resource type of the related identifier.

(C) Demo: Github → Zenodo

Francesco Varrato

zenodo
Search
Upload
Communities
Log in
Sign up

Log in to accoun
Log in with GitHub
Log in with ORCID
— OR —
Email Address
Password
Log In

Authorize application

Zenodo by @zenodo would like permission to access your account

Review permi
Personal
Email addr
Reposito
Admin acc
Organiza
Read-only

Home / Account / GitHub

Settings
Profile
Change password
Linked accounts
Applications
Shared links
GitHub

GitHub Repositories
(updated 11 seconds ago)
Sync now ...

Get started

1 Flip the switch
Select the repository you want to preserve, and toggle the switch below to turn on automatic preservation of your software.
ON

2 Create a release
Go to GitHub and create a release. Zenodo will automatically download a .zip-ball of each new release and register a DOI.

3 Get the badge
After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.
DOI 10.5281/zenodo.8475
(example)

Repositories
If your organization's repositories do not show up in the list, please ensure you have enabled third-party access to the Zenodo application. Private repositories are not supported.

octocat/Hello-World
ON

Github – Referencing and citing content

(C) Dataset dissemination: How to choose a platform

1. Listed on re3data.org: for peace of mind
2. DOI or another 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**
- A website is **not a data repository**
- An archive might or not be a **dissemination platform**
- Check data repos. recommended **by your funder** (ex. SNSF, ERC)
- Check **licenses** (ex. tldrlegal.com, dmlawtool.ccdigitallaw.ch)
- You can choose **both a GENERIC & field-SPECIFIC data repository**

Just ask

Colleague(s)

- Group / Faculty data manager
- NCCR data official / Dept. scientific advisor / …
- **go.epfl.ch/datachampions**

researchdata@epfl.ch

Questions

go.epfl.ch/training

go.epfl.ch/rdm

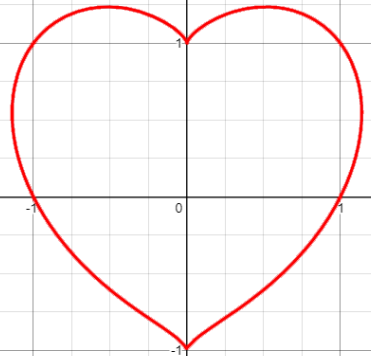
Contacts & Credits

francesco.varrato@epfl.ch

researchdata@epfl.ch

Love Data Week

EPFL 2♡22 edition



go.epfl.ch/quiz

Fri. Feb. 18 | 10:00-12:00
Online speed-dating with EPFL Data Champions

go.epfl.ch/ldw22

LDW22

Bibliothèque