Which platform to publish my research data / code on?

Francesco Varrato
EPFL Library

Love Data Week 2022
Feb. 17, 2022
Love Data Week

EPFL 2022 edition

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

go.epfl.ch/quiz

go.epfl.ch/ldw22
Content

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

Specific questions from you

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

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:
To avoid this… among other things

"Data are available in the supplement."

It's literally SCREENSHOTS of their Excel spreadsheet in a Word document.

WHAT??

Source: https://twitter.com/robertnulrich/status/1490567709188182016
Pain points / Needs

- 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

(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 30.1 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

(A) Dissemination as publication

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)

Others can see it

All can see it

VS.
(B) Data / Code etc.

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

DATASET

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

What is needed to reproduce the results?
(B) Not everything is... ... publishable

BlueBrain/Search: v0.1.0 — COVID-19 paper

Stanislav Schmidt; Francesco Casalegno; Pierre-Alexandre Fonta; Jan; EmilieDeb; Mohameth François SY; Eugenia Oshurko; Anna; antWey; Didac

Blue Brain text mining toolbox for semantic search and structured information extraction

- 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

doi.org/10.5281/zenodo.4563998

ord.fkp.jku.at/dataset/151027b
(B) Not everything is... ... publishable

BlueBrain/Search: v0.1.0 — COVID-19 paper

Stanislav Schmidt; Francesco Casalegno; Pierre-Alexandre Fonta; Jan; EmileDeb; Mohameth François SY; Eugenia Oshurko; Anna; antibeY; Didac.

Blue Brain text mining toolbox for semantic search and structured information extraction

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

Steve Berry, www.flickr.com CC BY-NC-SA 2.0
(B) **Constraints to dataset publication**

- Tests on **animals** / humans
- **Personal data** ([Federal Act on Data Protection](https://www.admin.ch/lv/fr/lv/index.html), [Human Research Act](https://www.admin.ch/lv/fr/lv/index.html), [GDPR](https://www.admin.ch/))
  
  → check the EPFL [Human Research Ethics Committee](https://www.epfl.ch/ethics/) (AREC + HREC forms)

---

- **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](https://www.epfl.ch/tto/) (Technology Transfer Office) ... Choose the **data license** + tell in the DMP!
(C) **Platforms:** Where would/do you publish a dataset?

Whiteboard
Publication statements from 60 real DMPs

C) Francesco Varrato
(C) Platforms: Definitions and... nuances

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

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

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

- 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
(C) Code Repository

- Especially code
- Publication: not always
- Provide a DOI: no
- Collaborative: usually
- Institutional or commercial
- Non-profit or for-profit
- Usually generic
... 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](https://en.wikipedia.org/wiki/Data_paper))
### Communities

<table>
<thead>
<tr>
<th>Community</th>
<th>Title</th>
<th>View</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPFL École Polytechnique Fédérale de Lausanne</td>
<td>EPFL</td>
<td>View</td>
<td>Curated by: rdms@epfl</td>
</tr>
<tr>
<td>CHILI Lab, EPFL</td>
<td>CHILI Lab, EPFL</td>
<td>View</td>
<td>Shared Resources for Computer Human Interaction for Learning and Instruction Lab at EPFL, Lausanne, Switzerland. Curated by: stamatilov</td>
</tr>
<tr>
<td>From Neuroscience to Machine Learning</td>
<td>From Neuroscience to Machine Learning</td>
<td>View</td>
<td>Presentations of the Human Brain Project SP4 / CPDoS 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:骑马马</td>
</tr>
</tbody>
</table>

### License

- **Access right**: Open Access
- **License**: Creative Commons Attribution 4.0 International
- **Grants**
  - European Commission (EU)
(C) Demo: Github → Zenodo

Authorize application

Zenodo by @zenodo would like permission to access your account

Review permission

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

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.

Repositories

- Octocat/Hello-World

Github – Referencing and citing content
(C) Dataset dissemination: How to **choose** a platform

1. Listed on [re3data.org](http://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](http://tldrlegal.com), [dmlawtool.ccdigitallaw.ch](http://dmlawtool.ccdigitallaw.ch))
- You can choose **both** a GENERIC & field-SPECIFIC data repository
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

2022 Francesco Varrato

This is an open-access document distributed under the terms of the Creative Commons Attribution License
Love Data Week

EPFL 2022 edition

go.epfl.ch/ldw22

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

go.epfl.ch/quiz