Data unchained

SpaRTaN -MacSeNet
ITN Workshop
18.11.2016

BIBLIO THÈQUE
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data Unchained - Program

1. Research data – introductory *quiz*
2. Collaborative working with code:

   versioning, branching and metadata
3. Publishing your code:

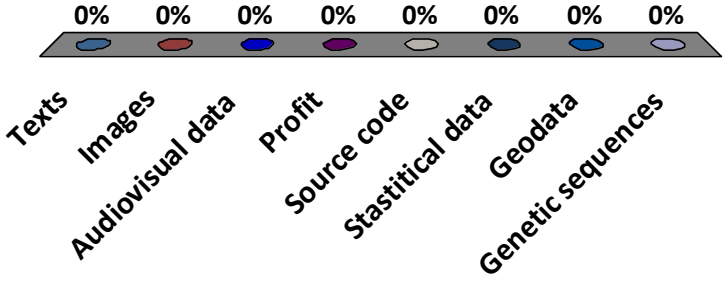
   from GitHub to Zenodo
4. Disseminating your code:

   licences, citation and publication
5. Data Management Plans

1. Research data introductory *quiz*

What would you associate research data with specifically?

A. Texts
B. Images
C. Audiovisual data
D. Profit
E. Source code
F. Stastitical data
G. Geodata
H. Genetic sequences



Definition

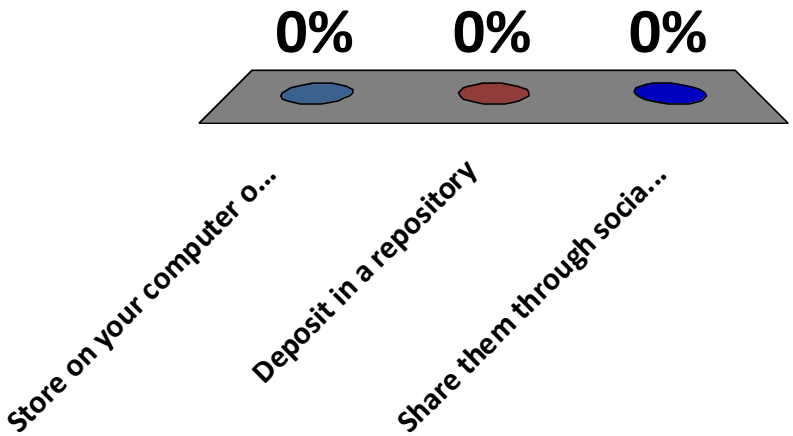
Several definitions are possible based on specific fields, institutions and organizations.

Research data are defined as **factual records** (numbers, texts, images and sounds), which are used as **principal sources for scientific research** and which are often recognized by the scientific community as being **necessary to validate research results**.

Organization for Economic Cooperation and Development (OECD)

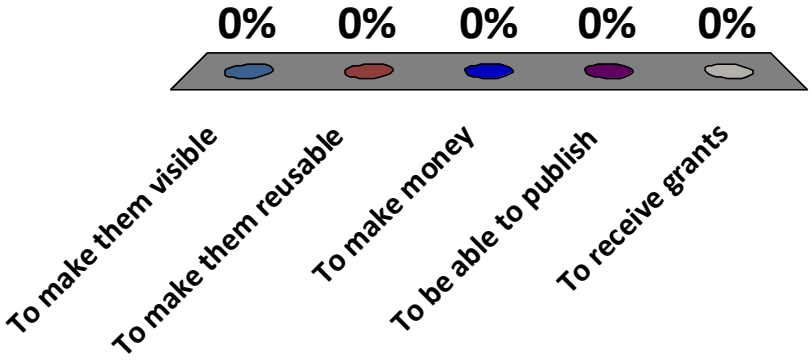
What do you usually do with your research data?

A. Store on your computer or cloud service
B. Deposit in a repository
C. Share them through social networks



Why would you make your research data available in a repository?

A. To make them visible
B. To make them reusable
C. To make money
D. To be able to publish
E. To receive grants



Data sharing

Human Genome Project (1996)

The National Human Genome Research Institute's (NHGRI) policy for release and deposition of **DNA sequence data** was devised to make sequence data available to the research community **as soon as possible for free, unfettered use**.

Data were to be deposited in a public database within 24 hours of generating a sequence assembly of 2 kb or larger. Data release according to this practice is far more rapid than the standard scientific practice of releasing data only upon publication.

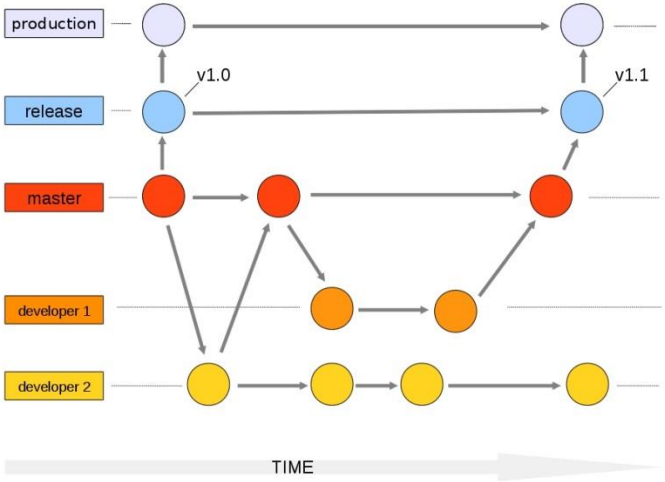
To know more about this project : Smith, D., & Carrano, A. (1996). International large-scale sequencing meeting. *Human Genome News*, *6*(7). Available at: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml

2. Collaborative working with code: versioning, branching and metadata

Git : Good practice for code management

Versioning

- Go back (to previous commits or versions)
- Keep track of evolutions (diff)



Source: https://github.com/hltfbk/Excitement-Open-Platform/wiki/Developers

Decentralized collaboration

- Work in parallel (repository, branches)
- Every repro is a master
- Merge work with others

Why GitHub today?

So why?

- Free for open projects
- Extremely popular
- Available to everyone
  (c4science.ch is restricted
  to Swiss academic community)

⇒ We will work on the basis of the GitHub repository
of the popular Scala language developped et EPFL.

GitHub

Exercise

Learn how to clone a code and how to create metadata on your own new version of a repository on GitHub

1. Register (if necessary)
2. Select official Scala repository
3. Check the reuse licence
4. Fork
5. Select the branch 2.13x
6. Create a new file for metadata (e.g. metadata.md)

How to create metadata?

Exercise

Use the Dublin Core Element Schema (dces)

1. http://dublincore.org/documents/dces/
2. Fill up the 15 fields

- Format: in a new text file in the repository. Enter one field per line, starting with the field name (e.g. «title: »)

- Help: have a look at the scala repository, wikipedia, webpages, etc

.

- Some thoughts:
  - Is it a good idea to include the institution in the «creators» or not?
  - How to describe the current branch and version?
  - What format should you use for the date?
  - Are there other ambiguous points?

Metadata : Beyond Dublin Core

Dublin Core Basics

- 15 fields
- simple and easy
- not perfect.

Dublin Core Metadata Initiative®

- http://dublincore.org/documents/dcmi-terms/
- Refined tems (>50)
  - relation -> isPartOf, hasVersion
  - coverage -> temopral, spatial
- And Schemes
  - subjects: LCSH (Thesaurus)
  - languages: RFC1766 (en, de, fr)
  - Date: ISO8601 (YYYY-MM-DD)

And more - disciplinary specific metadata formats:
See http://rd-alliance.github.io/metadata-directory/

3. Publishing your code: from GitHub to Zenodo

Research data publishing

Data repository

Why

Long-term accessibility and **preservation**
Increased **discoverability** and **reuse** of data

How to choose

- Data sharing practices within your community
- Repository's specific features
  a) "try to find the one offering the best combination of ease-of-deposit, community uptake, accessibility, discoverability, value-added curation, preservation infrastructure, organizational persistence, and support for the data formats and standards you use"
  b) disciplinary/generic; economic model behind; proposed licenses; partnerships with publishers; …
- Check Re3data.org
- Ask your librarian ☺

Comply with requirements and policies

- **Funders**
- **Publishers**
- **Institution**

Source: Goodman, A., Pepe, A., Blocker, A. W., et al. PLoS Computational Biology 10 (2014)

From Github to Zenodo: Why & How

Why

Preserve

Code repositories (included GitHub) are not intended nor equipped for long-term storage and preservation.
Data repositories, like Zenodo, are!

Cite

By archiving your codes and softwares in Zenodo you will get a DOI

Discover

Data discoverability enhances the visibility and the impact of your research, and enables a broad dissemination of your research outputs

How

Exercise

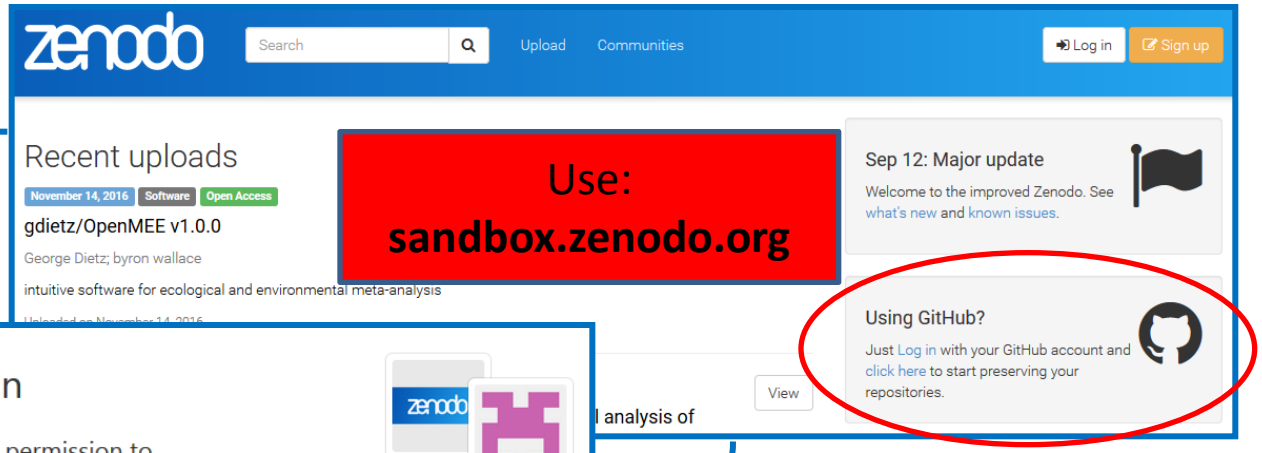
!!! Every (trusted) data repository helps you in making your data safely stored, easily citeable, and discoverable !!!

Capture and publish your GitHub repository in Zenodo

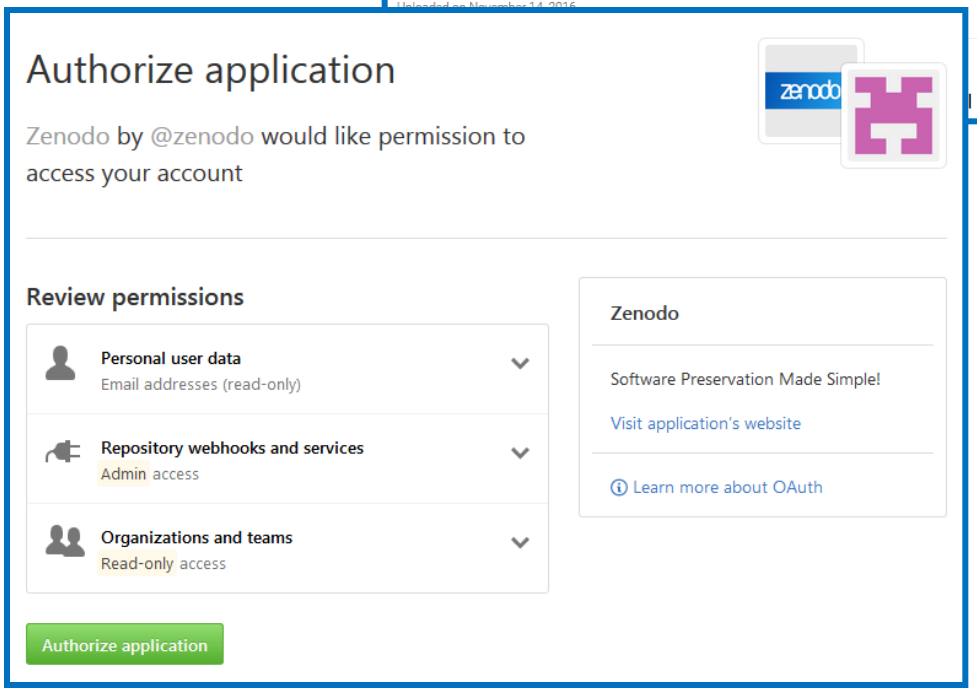
Exercise

When you write your software, you can **make the work you share on GitHub citable**: **archive** one of your **GitHub repositories in Zenodo** and **assign it a DOI**

Preliminary steps



Use:
sandbox.zenodo.org



Capture and publish your GitHub repository in Zenodo

Exercise

zenodo

Search

Upload Communities

lorenza.salvatori@epfl.ch

Home / Account / GitHub

Settings

- Profile
- Change password
- Linked accounts
- Applications
- Shared links
- GitHub

GitHub Repositories

(updated 2 minutes 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.

You have no repositories on GitHub.
Go to GitHub and create your first or click Sync-button to synchronize latest changes from GitHub.

Capture and publish your GitHub repository in Zenodo

Exercise

Create a new release

Fill in the release notes

Publish your release

Creating a new release triggers Zenodo into archiving your repository

Go to Upload

Select your Upload

...and edit metadata

Use the metadata

Source: https://guides.github.com/activities/citable-code/

4. Disseminating your code: licences, citation and publication

Licences

They cover code, data , text and multimedia:

- CC-By.
  - Reuse, adapt, publish dervivatives.
  - Obligation : cite the creator(s).
- Additional options :
  - non commercial, share alike, no derivs
  - resulting in total 6 CC-By licences.
- CC0 : almost public domain.

Other licenses are specific for code:

- GNU-GPL : Open Software.
- Apache2.0 : smaller codes, libraries.
  - Premissive.
  - No share alike clause.
  - Preservation of copyright notice.
- BSD-3clause - similar.

Research data citation

DOI = Digital Object Identifier
… is a **unique** & **persistent** identifier

It concerns datasets, but not only (online documents like papers).
Once a DOI is assigned to a dataset, this dataset is **published**.
It is given by **repositories** (like Zenodo).

Advantages?
It gives credits to data producers.
Contract with DataCite: the provider (Zenodo) has to maintain access to the dataset.

Alternatives to DOI?
Handle System (system used by Datacite), Persistent URLs, ARKs, etc.

Research data publishing

Data

in a repository

Papers in journals

Link to your dataset
"Data availability statement"
(compliance with publishers' policies)

Data papers in (data) journals

"The data paper is a 'data-publishing' product […] that may appear in a data journal or any other journal. […] Data articles focus on making data discoverable, interpretable, and reusable, rather than testing hypotheses or presenting new interpretations (by contrast with traditional journal articles)".

Peer review

Papers in Software journals

"…which accept **submissions that are primarily about the software**, and not necessarily on new algorithms or new science. There is an expectation that the use of the software will enable new research to be carried out".

Sources: Austin, C.C., Bloom, T., Dallmeier-Tiessen, S. et al. Int J Digit Libr (2016) ; https://www.software.ac.uk/resources/guides/which-journals-should-i-publish-my-software

5. Data Management Plan (DMP)

Data Management Plan (DMP)

Description of the data
- 1.1 Type of study
- 1.2 Type of data
- 1.3 Format and scale of the data

File naming
Versioning

Data collection / generation
- 2.1 Methodologies for data collection / generation
- 2.2 Data quality and standards

Metadata & documentation

Data management, documentation and curation
- 3.1 Managing, storing and curating data
- 3.2 Metadata standards and data documentation
- 3.3 Data preservation strategy and standards

Long term preservation strategies

Data security and confidentiality
- 4.1 Formal information/data security standards
- 4.2 Main risks to data security

Data confidentiality

Data sharing and access
- 5.1 Suitability for sharing
- 5.2 Discovery by potential users of the research data
- 5.3 Governance of access
- 5.4 The study team's exclusive use of the data
- 5.5 Restrictions or delays to sharing
- 5.6 Regulation of responsibilities of users

Sharing strategies (repositories, journals)

Licences

Helpful tool to manage your data

Some funders require a DMP

Credits

All the icons are published under CC BY 3.0 license on thenounproject.com:

User attention by TMD

Repository by lastspark

Law by Marie Van den Broeck

Peer Evaluation By Duke Innovation Co-Lab

Friends by Les vieux garçons

Email by Lorena Salagre

DATA UNCHAINED

EPFL Library Team
Noémi Cobolet
Jan Krause
Raphaël Rey
Lorenza Salvatori
questions.bib@epfl.ch

library.epfl.ch
facebook.com/EPFL.library
youtube.com/epfllibrary
@EPFLlibrary