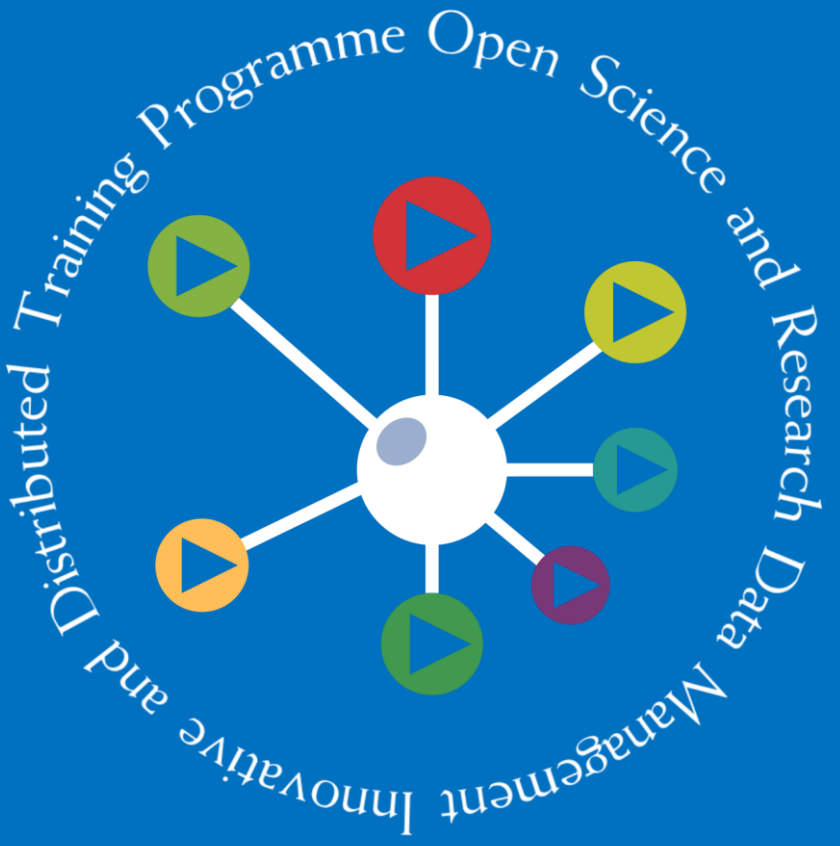
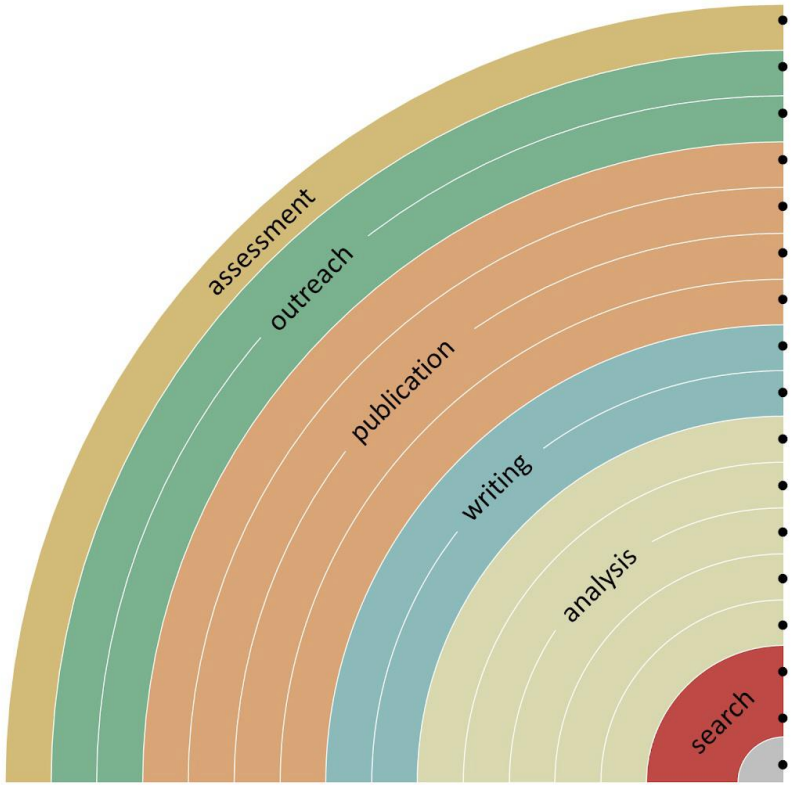
Open Science Platforms

This module is part of the training session “Train for trainers” within project TrainRDM



University POLITEHNICA of Bucharest
National Institute of Research and Development in Informatics (ICI Bucharest)

From Open Science practices to platforms



- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, 'green' or 'gold'
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

DOI: 10.5281/zenodo.1147025

Rainbow of open science practices (available on Zenodo in different formats, including as editable slide:10.5281/zenodo.1147025)

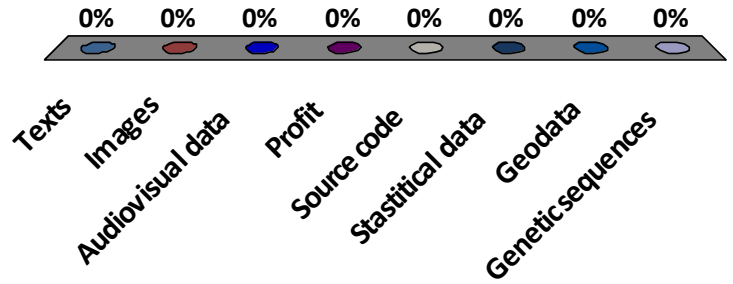
Tools and platforms

- There are many tools and platforms that support Open Science practices
  - Which tools and platforms to use (or advise) depends on factors such as: whether the tool is available (either free of at low cost or licensed to your institution), whether it works in your browser or for your operating system, whether it is available in your language, and whether it meets your security and privacy requirements.
  - In addition to these more technical criteria, consider whether a tool fits with the way you work.
- What is the open science practice you'd like to implement?
- Then explore which tools/platforms are available, which ones the people in your community use, and why (ask around!). Then make your own decision. Don't be afraid to experiment and try out something new!
  - Does it work well with other tools and platforms that you use?
  - Do the people you collaborate with use the same tool for the same practice, or at least one that is compatible with the one you use?
  - Do you need to invest a lot of time into learning the new tool, and if so, is that worth it for you?
  - Do you have support (either in real life or online) that can help you learn to use the tool?

1. Research data

What would you associate research data with specifically?

- Texts
- Images
- Audiovisual data
- Profit
- Source code
- Stastitical data
- Geodata
- Genetic sequences



Definition

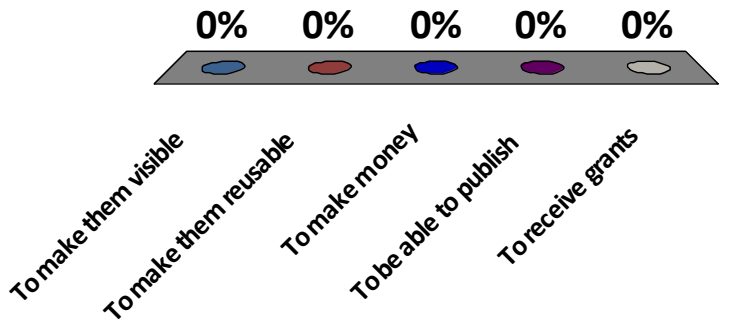
- Several definitions are possible based on specific fields, institutions and organizations.

- 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)

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



Limited incentives to give evidence against yourself

- We know that no one wants to incriminate themselves, and also that no one is infallible.
  - The Fifth Amendment to the United States Constitution includes a clause that no one “shall be compelled in any criminal case to be a witness against [them]sel[ves]”. (Edited to gender-neutral language.)
  - To “plead the fifth” means that someone chooses not to give evidence that there might have been something wrong in their past behaviour. They have the right to remain silent.
- Putting your code and data online can be very revealing and intimidating, and it is part of the human condition to be nervous of being judged by others.
- Although there is no law governing the communication of reproducible research - unless you commit explicit fraud in your work - sharing errors that you find in your work is heavily disincentivised.

Table 1 Descriptive statistics for the six groups of journal articles compared in our analyses

From: Pandemic publishing poses a new COVID-19 challenge

	COVID-19	Ebola	Cardiovascular disease	2019 COVID-19-publishing journals	2020 COVID-19-publishing journals (excluding COVID-19 records)[a]	2020 COVID-19-publishing journals (including all records)[a]
Total records	7,155	333	27,702	99,147	111,331	117,644
Total journal articles	4,403	164	20,080	79,588	94,952	98,858
Total journal articles with dates	2,113	48	13,117	56,465	65,032	66,758
Median days to acceptance [interquartile range; range]	6 [12; 134]	15 [45; 136]	102 [93; 1,053]	93 [100; 1,074]	84 [103; 1,089]	82 [103; 1,089]
Accepted within 7 days [*N*]	59% [1,250]	38% [18]	3% [374]	2% [1,386]	3% [2,113]	5% [3,138]
Accepted within 30 days [*N*]	93% [1,970]	71% [34]	9% [1,158]	13% [7,324]	18% [11,396]	20% [13,020]
Accepted within 100 days [*N*]	99% [2,099]	92% [44]	49% [6,465]	54% [30,536]	58% [37,972]	59% [39,698]

[a]Note that columns 2–5 report our results based on PubMed searches as specified above. For our analysis of 2020 records of journals that published COVID-19 articles, reported in columns 6–7, we validated the PubMed records against LitCovid. As a result, the number of COVID-19 records differs from those reported in column 2.

Nature Human Behaviour, vol. 4, pp. 666–669, (2020)

Dynamics of the COVID -19 Related Publications

Dynamics of the COVID -19 Related Publications

Abstract:

Background: This study aims to analyze the dynamics of the published articles and preprints of Covid-19 related literature from different scientific databases and sharing platforms.

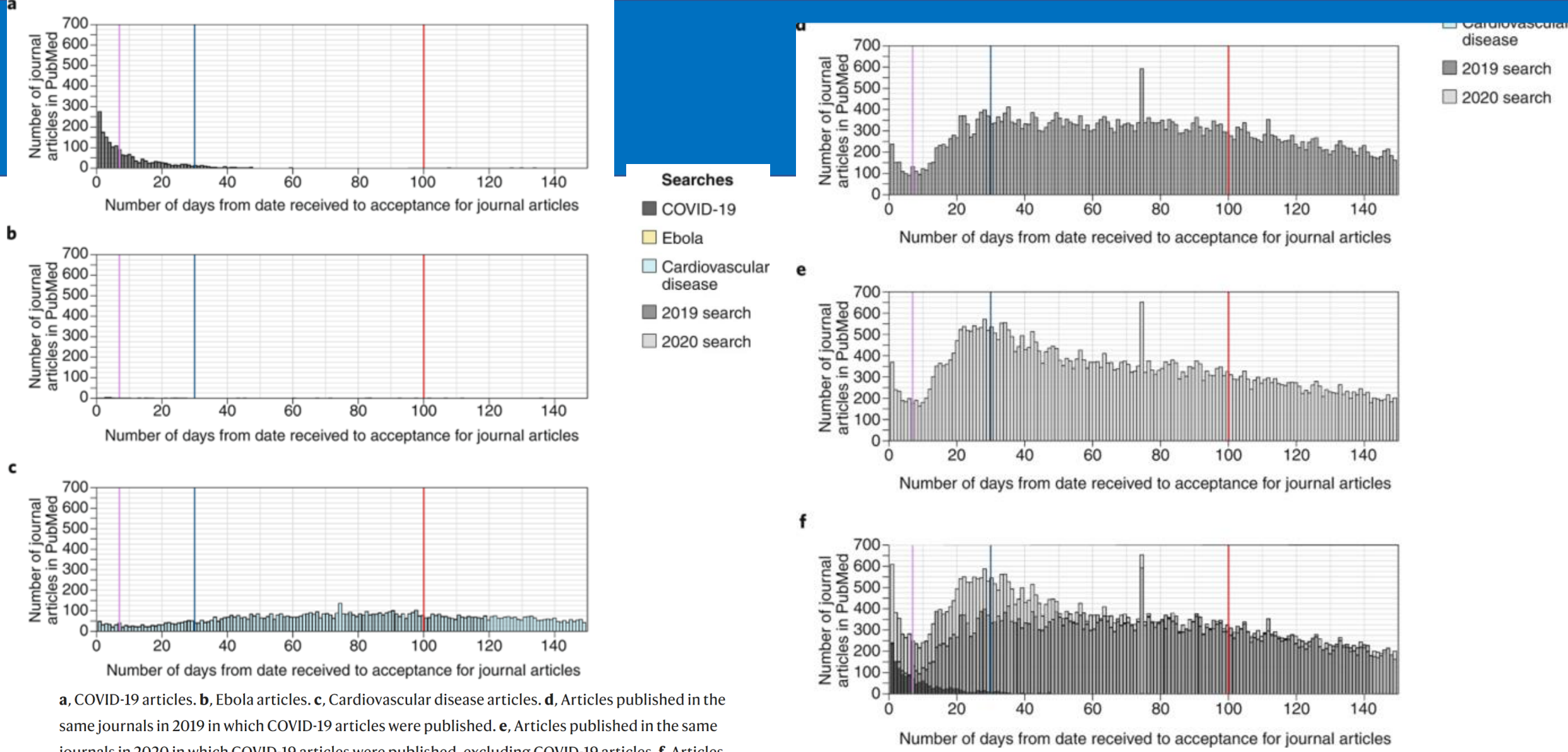
Methods: The PubMed, Elsevier, and Research Gate (RG) databases were under consideration in this study over a specific time. Analyses were carried out on the number of publications as (a) function of time (day), (b) journals and (c) authors. Doubling time of the number of publications was analyzed for PubMed "all articles" and Elsevier published articles. Analyzed databases were (1A) PubMed "all articles" (01/12/2019-12/06/2020) (1B) PubMed Review articles (01/12/2019-2/5/2020) and (1C) PubMed Clinical Trials (01/01/2020-30/06/2020) (2) Elsevier all publications (01/12/2019- 25/05/2020) (3) RG (Article, Pre Print, Technical Report) (15/04/2020 – 30/4/2020).

Findings: Total publications in the observation period for PubMed, Elsevier, and RG were 23000, 5898 and 5393 respectively. The average number of publications/day for PubMed, Elsevier and RG were 70.0 ±128.6, 77.6±125.3 and 255.6±205.8 respectively. PubMed shows an avalanche in the number of publication around May 10, number of publications jumped from 6.0±8.4/day to 282.5±110.3/day. The average doubling time for PubMed, Elsevier, and RG was 10.3±4 days, 20.6 days, and 2.3±2.0 days respectively. In PubMed average articles/journal was 5.2±10.3 and top 20 authors representing 935 articles are of Chinese descent. The average number of publications per author for PubMed, Elsevier, and RG was 1.2±1.4, 1.3±0.9, and 1.1±0.4 respectively. Subgroup analysis, PubMed review articles mean and median review time for each article were <0|17±17|77> and 13.9 days respectively; and reducing at a rate of -0.21 days (count)/day.

Interpretation: Although the disease has been known for around 6 months, the number of publications related to the Covid-19 until now is huge and growing very fast with time. It is essential to rationalize the publications scientifically by the researchers, authors, reviewers, and publishing houses.

Findings Total publications in the observation period for PubMed, Elsevier, and RG were 23000, 5898 and 5393 respectively. The average number of publications/day for PubMed, Elsevier and RG were 70.0 ±128.6, 77.6±125.3 and 255.6±205.8 respectively. PubMed shows an avalanche in the number of publication around May 10, number of publications jumped from 6.0±8.4/day to 282.5±110.3/day. The average doubling time for PubMed, Elsevier, and RG was 10.3±4 days, 20.6 days, and 2.3±2.0 days respectively.

doi: https://doi.org/10.1101/2020.08.05.237313



a, COVID-19 articles. **b**, Ebola articles. **c**, Cardiovascular disease articles. **d**, Articles published in the same journals in 2019 in which COVID-19 articles were published. **e**, Articles published in the same journals in 2020 in which COVID-19 articles were published, excluding COVID-19 articles. **f**, Articles published in the same journals in 2020 in which COVID-19 articles were published, including COVID-19 articles. In all panels, the purple line represents 7 days, the blue line 30 days and the red line 100 days from time of article receipt.

Nature Human Behaviour, **volume 4**, pp. 666–669(2020)

Publications related to Covid-19 in 2020 until the 17th week

J Clin Orthop Trauma. 2020 May; 11(Suppl 3): S304–S306. doi: 10.1016/j.jcot.2020.(

Question

- Would you consider publishing your research data?
  - after completion of the project for which the data were collected
  - After „traditional“ publication of the results
  - at a later date
  - never

Data Storage and Organisation

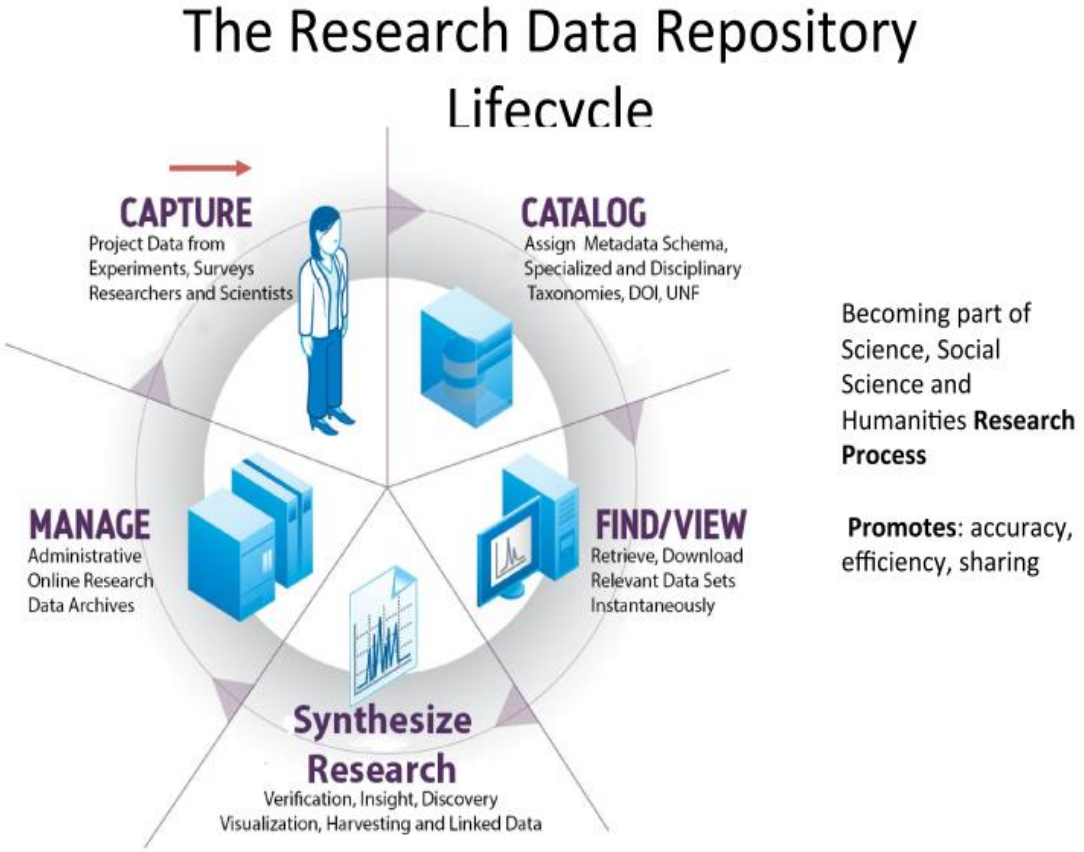
- Data loss can be catastrophic for your research project and can happen often.
  - You can prevent data loss by picking suitable storage solutions and backing your data up frequently.
- Where to Store Data
  - Most institutions will provide a *network drive* that you can use to store data.
  - *Portable storage media* such as memory sticks (USB sticks) are more risky and vulnerable to loss and damage.
  - *Cloud storage* provides a convenient way to store, backup and retrieve data. You should check terms of use before using them for your research data.

Documentation and Metadata

- Having data available is of no use if it cannot be understood. Without metadata to provide provenance and context, the data can’t be used effectively.
  - For example, a table of numbers is useless if no headings describe what the columns/rows contain.
  - You should ensure that open datasets include consistent metadata, that is information about the data so that the data is fully described.
  - This requires that information accompanying data is captured in documentation and metadata.
- Documentation
  - Documentation provides context for your work. It allows your collaborators, colleagues and future you to understand what has been done and why.
- Metadata
  - Metadata is information about the data, descriptors that facilitate cataloguing data and data discovery. Often, metadata are intended for machine reading.

Data repository : the what

- IT infrastructure (cloud based/online) set up to manage, share, access, maintain, and archive datasets.
- An application database specialized in storing metadata of data files/datasets/databases.
- Differs from publication repository mainly in its ability to:
  - Store metadata at different level/hierarchy.
  - Store and ingest data files in various formats for long-term preservation

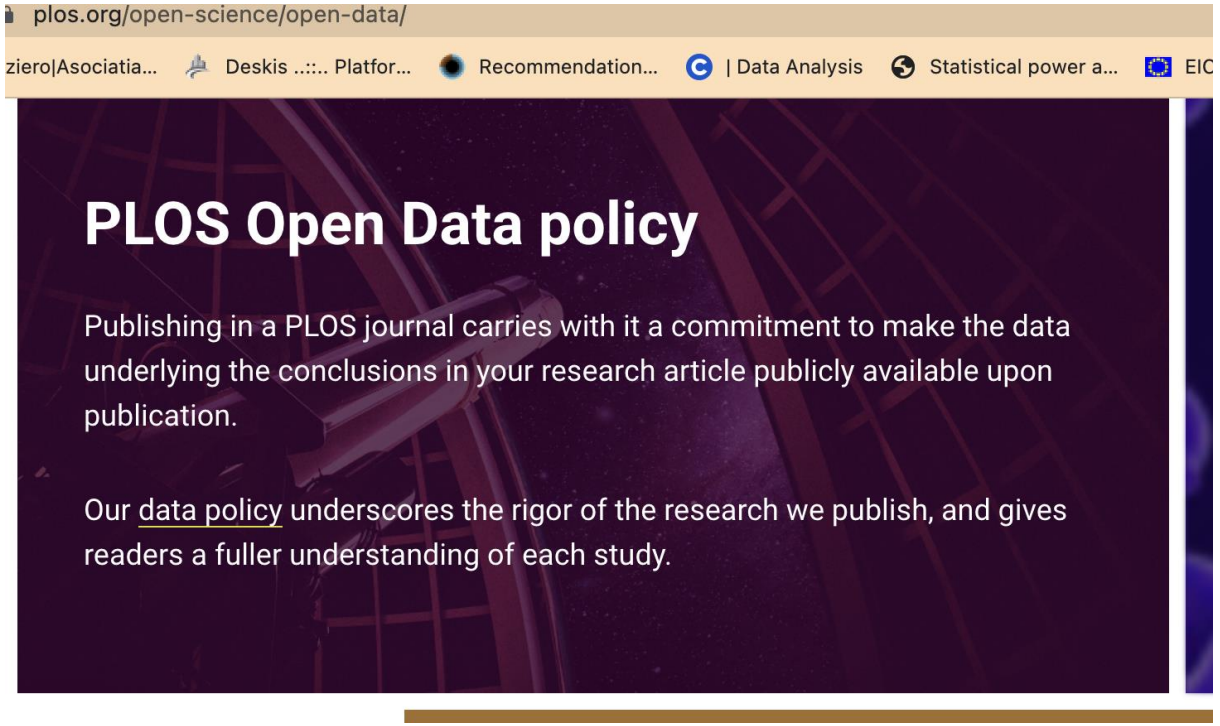


http://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml

Data publishing

- More and more funders expect the data produced in research projects they finance to be findable, accessible and as open as possible.

- How to make research data accessible? Publishing data as supplemental material associated with a research article, typically with the data files hosted by the publisher of the article.



What are Data Repositories?

- Databases in which objects can be
  - archived,
  - documented,
  - published.
  - Search tools allow the objects to be retrieved.

- Types of repositories:
  - discipline-specific
  - generic
  - institutional
  - media-specific (texts, research data, movies, etc.)

Using Repositories

A repository makes it possible to ...

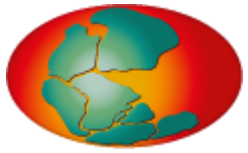
- … archive data safely over the long term.
- ... hold data and metadata together.
- ... share data (restricted user group).
- ... search for data yourself.
- … keep track of the state of research.
- ... include data from other authors in your own work.
- … publish data

- *Research data is often exchanged in informal ways with colleagues.*
- *However, officially published data brings many benefits to both data sharer and data re-user.*

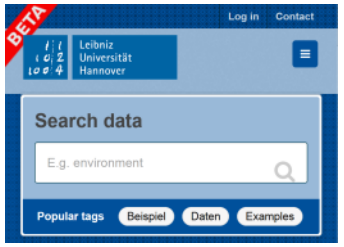
Examples of Repositories

discipline-specific

gesis
Leibniz-Institut
für Sozialwissenschaften



institutional



generic
(interdisciplinary)

zenodo

data journals

nature.com
SCIENTIFIC DATA

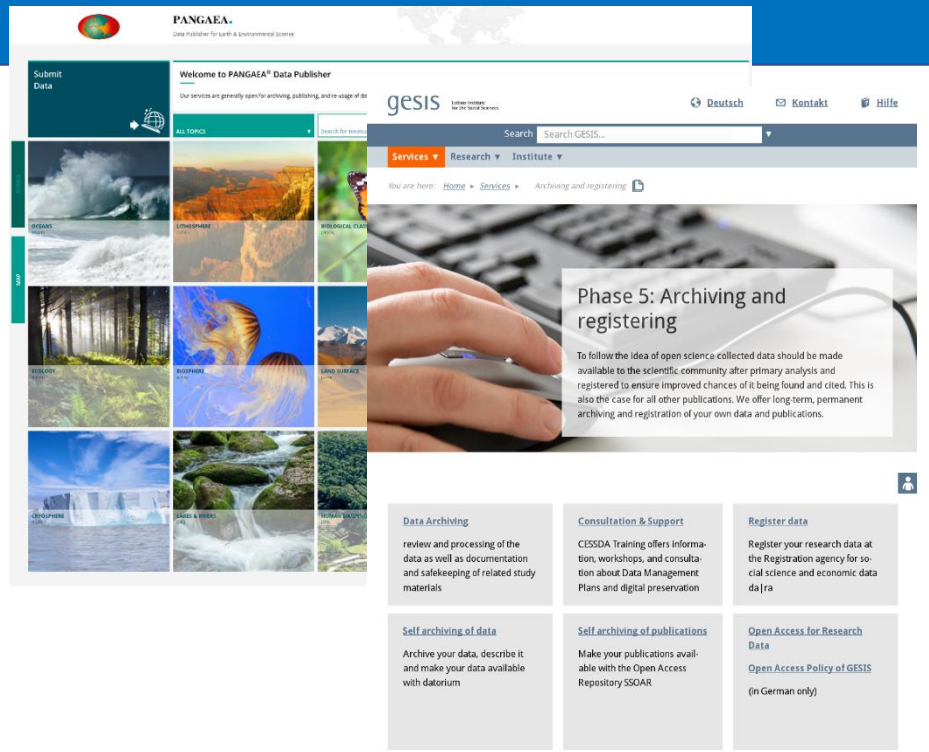
media-specific

TIB AV-PORTAL

GitHub

Disciplinary / Domain Repositories

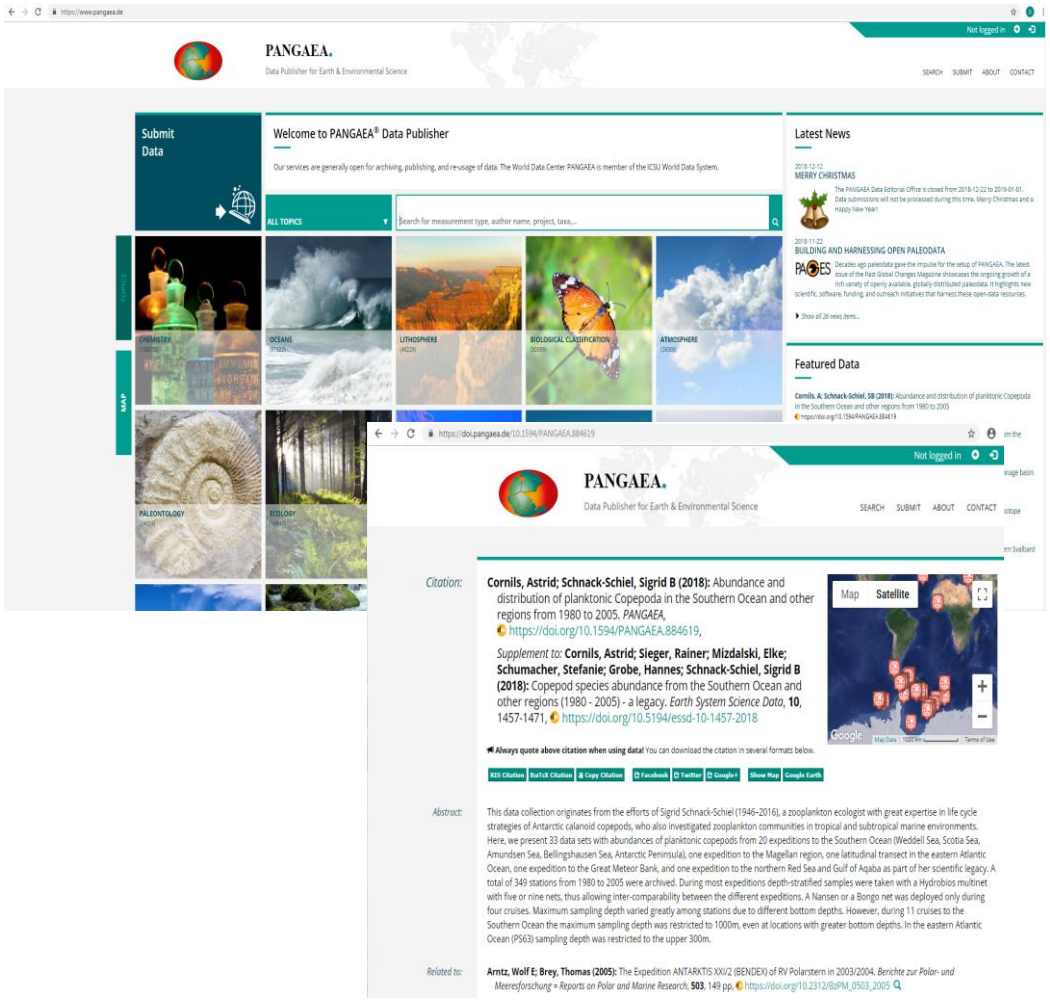
- for one scientific discipline / domain
- usually fixed metadata schema
- a central place to look for data in a domain
- Use it, if there is one for your domain!



Examples:

- Pangaea – Geodata repository - https://www.pangaea.de/
- Nomad Repository - Novel Materials Discovery Repository - http://nomad-repository.eu/cms/

A Discipline-Specific Repository: Pangaea.de



- https://www.pangaea.de
- data publisher for earth & environmental science
- DOIs for datasets
- quality control by professional data curators
- disciplinary metadata (parameters, spatial coverage, events)
- free of charge for data submitters
  (funded by the European Commission, BMBF, DFG, International Ocean Discovery Program (IODP))
- hosted in Germany
  (Alfred Wegener Institute & University of Bremen)

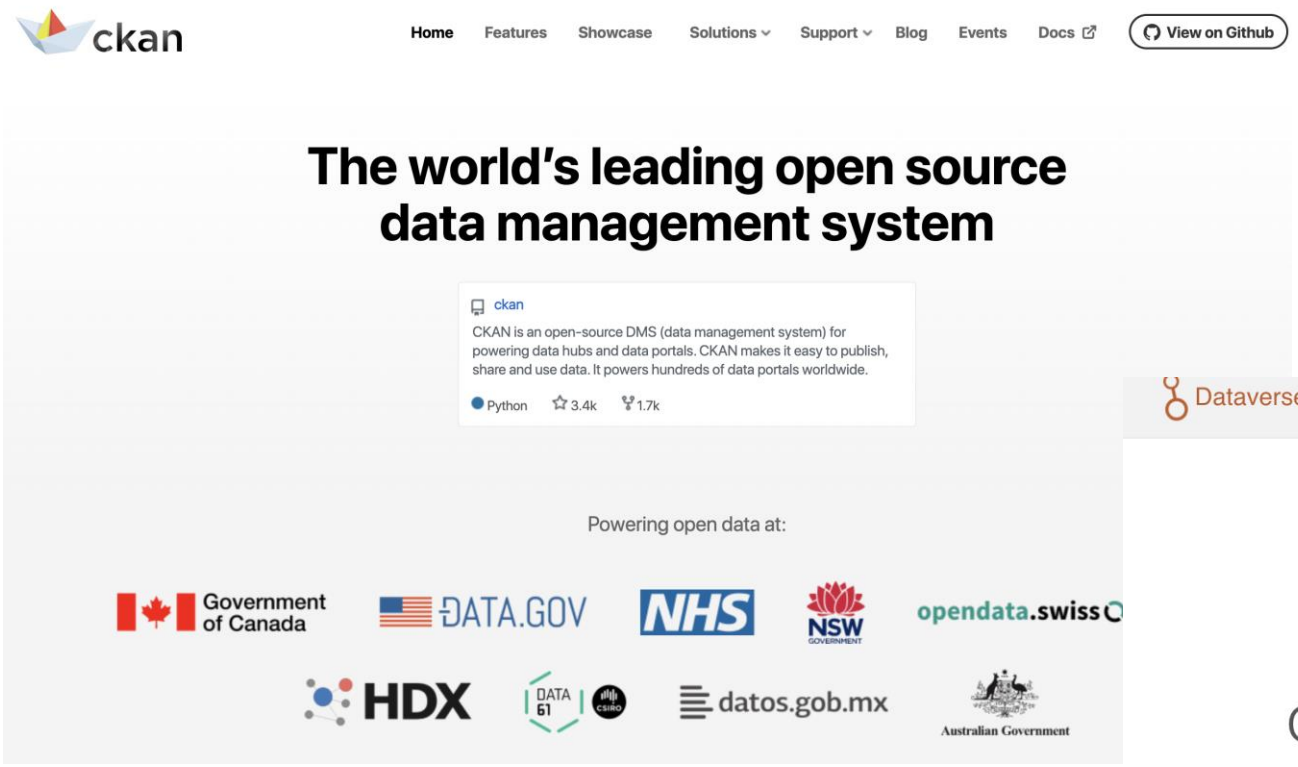
Generic and Institutional Repositories

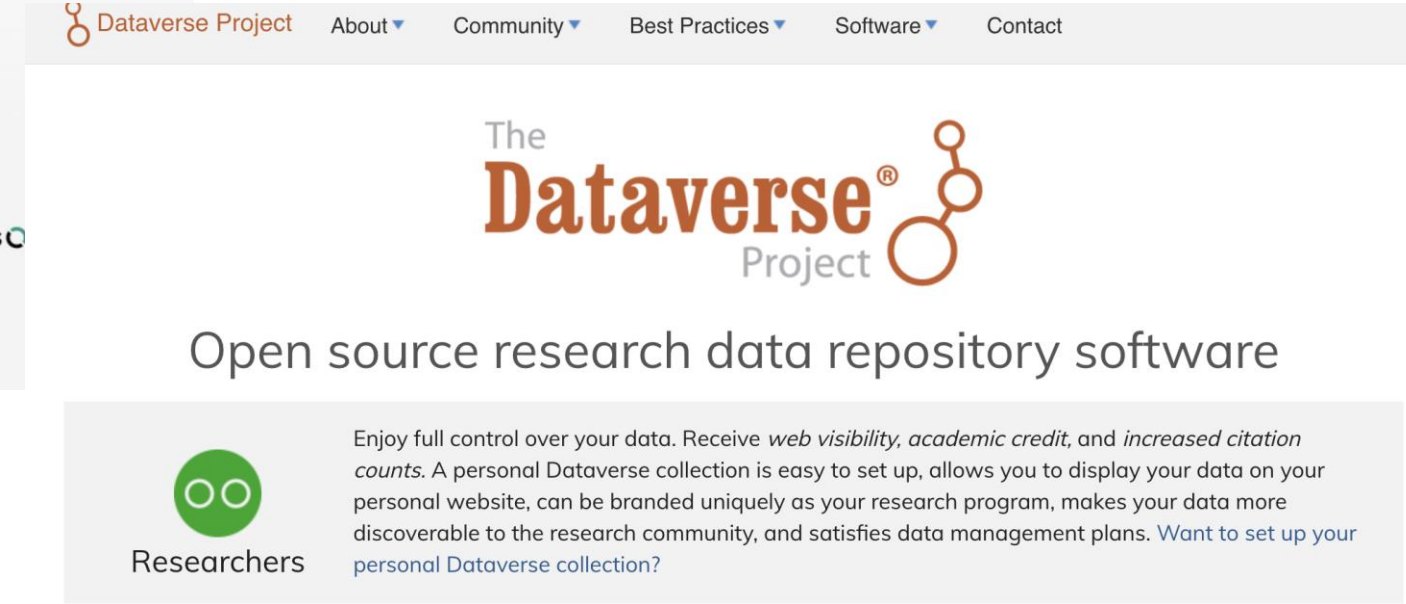
- open for various subject-specific data
- metadata schemata usually more general (DataCite, Dublin Core)
- for domains without subject-specific repositories

- Generic: examples of public repositories
  - Zenodo (CERN) - https://zenodo.org/
  - Figshare - https://figshare.com/

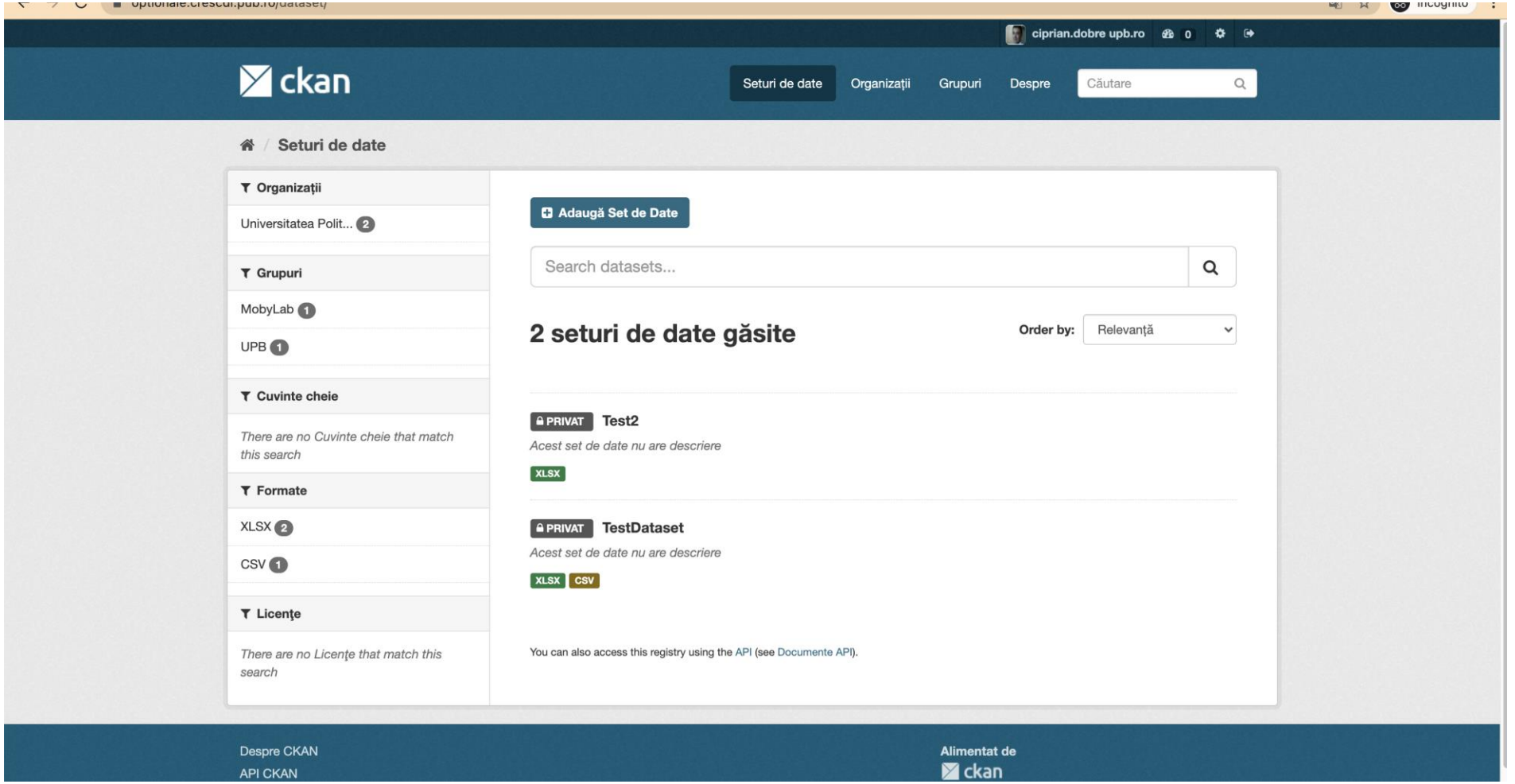
- Institutional: e.g. for a university
  - pure research data repositories, e.g.
    - OpARA – the TU Dresden repository - https://opara.zih.tu-dresden.de/
    - TUdatalib (TU Darmstadt) - https://tudatalib.ulb.tu-darmstadt.de/
  - repositories for both documents and research data, e.g.
    - RWTH Publications - http://publications.rwth-aachen.de/
    - KITopen - https://www.bibliothek.kit.edu/cms/kitopen.php
    - DepositOnce (TU Berlin) - https://depositonce.tu-berlin.de/

Institutional repositories





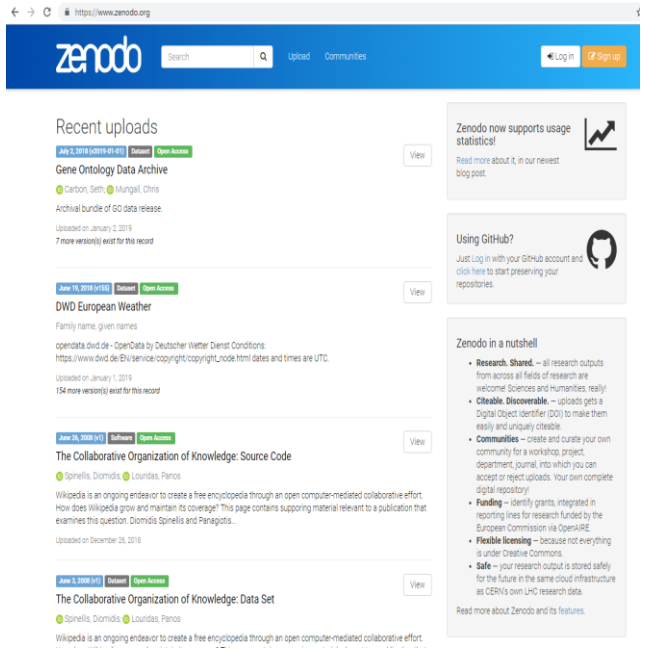
Example at UPB...



Public Repositories: Zenodo, Figshare

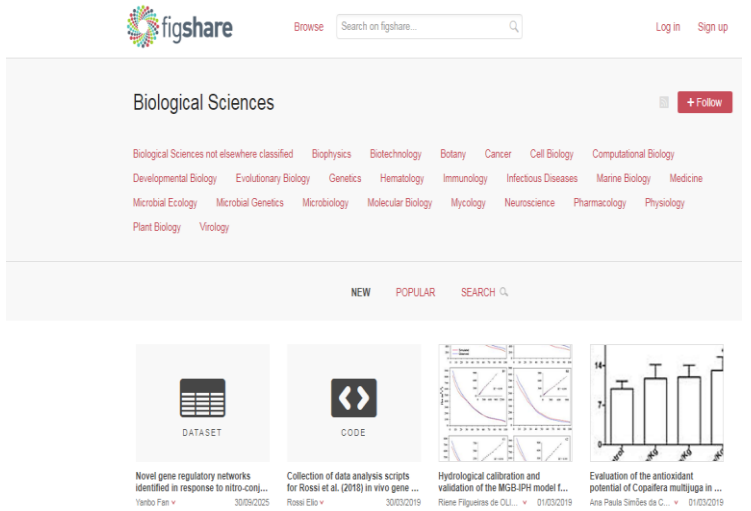
Zenodo.org

- public repository
- integration with GitHub



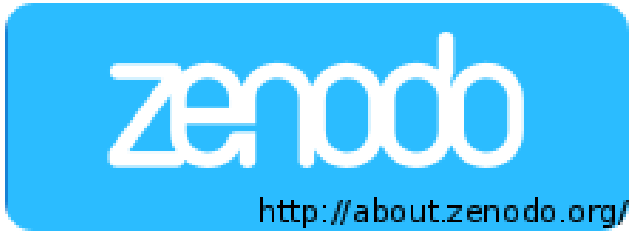
Figshare.com

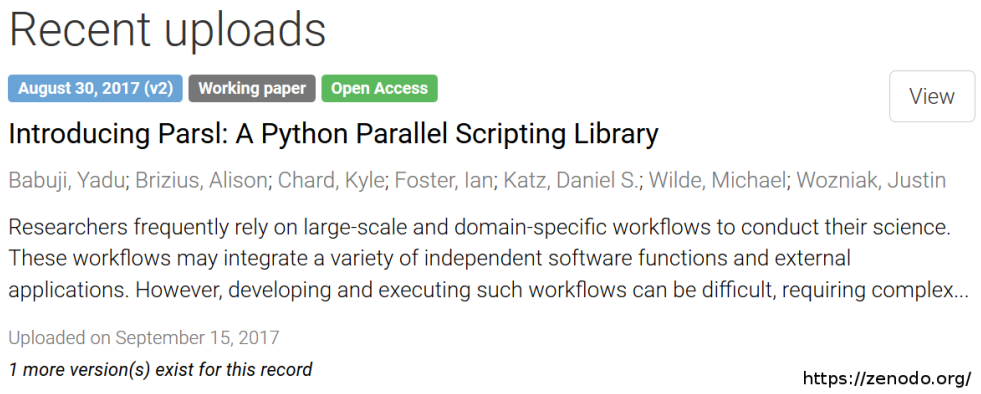
- public and institutional offer
- DOIs



Zenodo – One for All and Everything

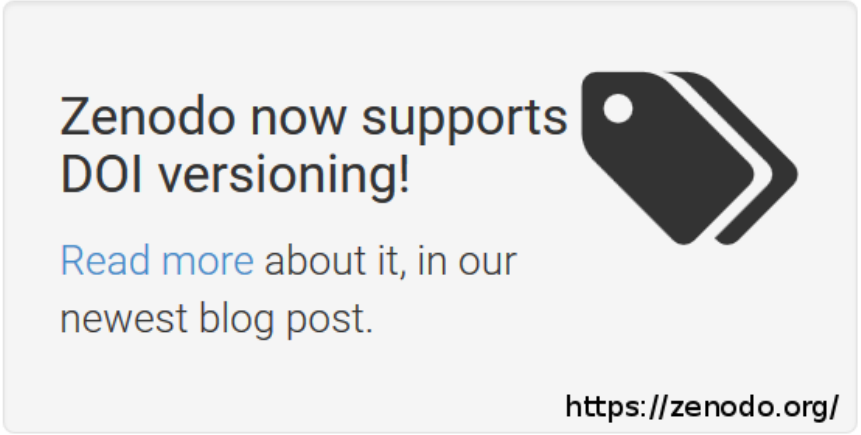
- Zenodo is funded by OpenAire, CERN and the EU Horizon2020 programme.
- open to **all** fields of research
- open to **all** types of research data
- free service
- File size up to 50 Gigabyte

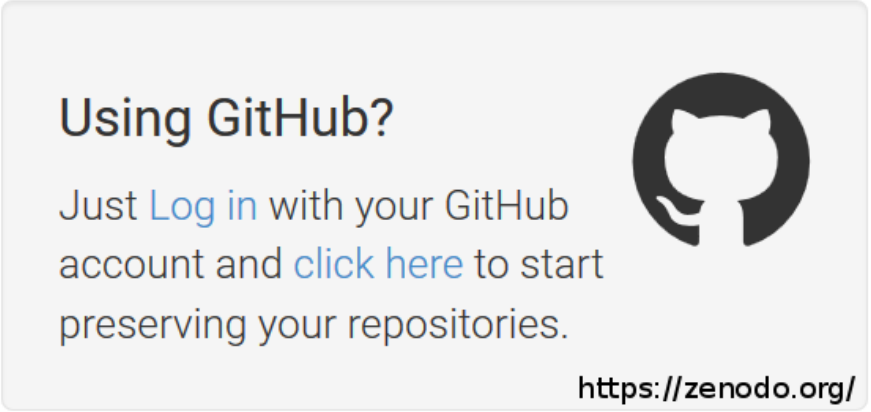




Zenodo – Special Features

- DOI-Versioning:
- edit/update data after publication
- quote a specific version or all versions of a dataset
- In combination with GitHub:
- archiving of software for the evaluation of data
- also with versioning and citable link/identifier





Figshare – a Service

- Figshare is a commercial offer for scientists and institutions.
- open to all fields of research
- open to all types of research data
- Freemium model
- File size up to 5 Gigabyte

figshare

https://figshare.com/

COMBINE Archive Show Case

Version 2 24.10.2017, 10:39 by martin scharm, Vasundra Touré, Matthias König

A fully featured COMBINE archive of syncytial mitotic cycles in Drosophila embryos, including (i) the original publication, (ii) model code, (iii) simulation descriptions, and (iv) simulation results.

A COMBINE archive is a standardised container for data files related to a simulation study in computational biology. The archive can be used for both, educational and research purposes. Anyone may reuse, extend and update the archive to make it a valuable resource for the scientific community.

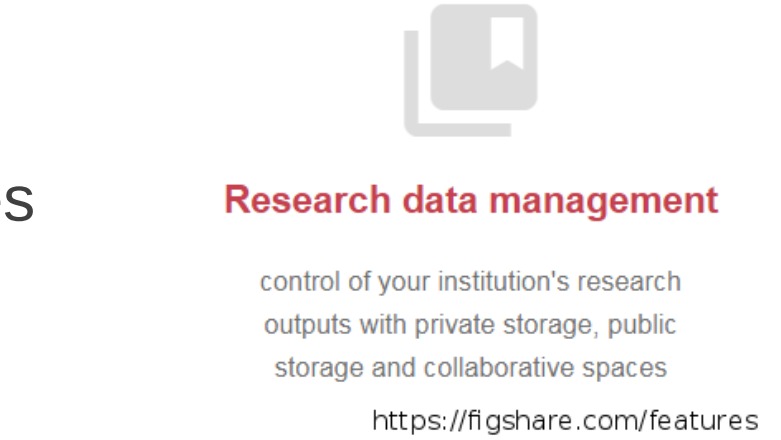
https://figshare.com/articles/COMBINE_Archive_Show_Case/3427271

Figshare – Special Features

Services:

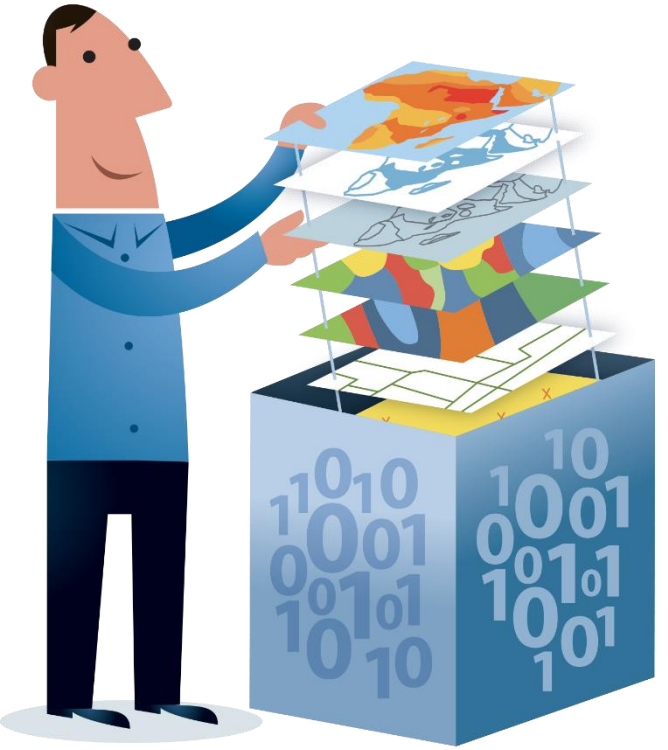
- offers repositories and other services for universities
- cooperates with publishers, e.g. PLOS or also:

https://springernature.figshare.com



Task: Zenodo versus Figshare

- Please compare the offer of Zenodo and Figshare.
- Which provider would you confide your data to?



www.digitalbevaring.dk

Architecture

Class	Feature	DSpace	CKAN	Figshare	Zenodo	Dataverse
Architecture	Deployment	Installation package	Installation package	Service	Service	Installation package
	Storage location	Local or remote	Local or remote	Remote	Remote	Local or remote
	Maintenance costs	Infrastructure management	Infrastructure management	Monthly fee	Monthly fee	e-mail based-free of cost
	Open Source	√	√	×	×	√
	Platform customization	√	√	×	Community policies	√
	Embargo period	√	Private storage	Private storage	√	√
	Content versioning	×	√	×	×	√
	Pre-reserving DOI	√	×	√	√	√

Architecture…

Class	Feature	DSpace	CKAN	Figshare	Zenodo	Dataverse
Metadata	Required fields	Title, Date of issue	Title	Author, title, categories description	Type, DOI, author, title, description	Title, Author, Description, Contact Email, Subject, and DOI
	Exporting schemas	Any pre-loaded schema	×	DC	DC, MARCXML	XML
	Schema flexibility	Flexible	Flexible	Fixed	Fixed	Flexible
	Validation	√	×	×	√	√
	Versioning	×	√	×	×	√

Architecture…

Class	Feature	DSpace	CKAN	Figshare	Zenodo	Dataverse
Dissemination	API	√	√	√	√	√
	OAI-PMH Compliance	√	With ckanext-harvest installer	√	√	√
	Faceted search	√	√	√	√	√

Architecture

- Most of the mentioned software are open source based and have given some flexibility to the users.
- Speedy and simple deployment of the used software is a crucial part for the implementation.
- Open-source software can be installed in house whereas platforms like Figshare and Zenodo are to be installed and implemented by the help of the developer.
- Dspace, Dataverse & CKAN have better control in the recorded data as they are open source.

Metadata

- Zenodo and Figshare software are able to export records that comply with established metadata schemas (Dublin Core and MARC-XML respectively).
- DSpace goes further by exporting DIPs (Dissemination Information Package) that include METS metadata records, thus enabling the ingestion of these packages into a long-term preservation workflow.
- Although CKAN and Dataverse metadata records do not follow any standard schema, the platform allows the inclusion of a dictionary of key-value pairs that can be used to record domain specific metadata as a complement to generic metadata descriptions.

Repositories: How do I find them?

- This will help you to find a suitable repository to ...

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

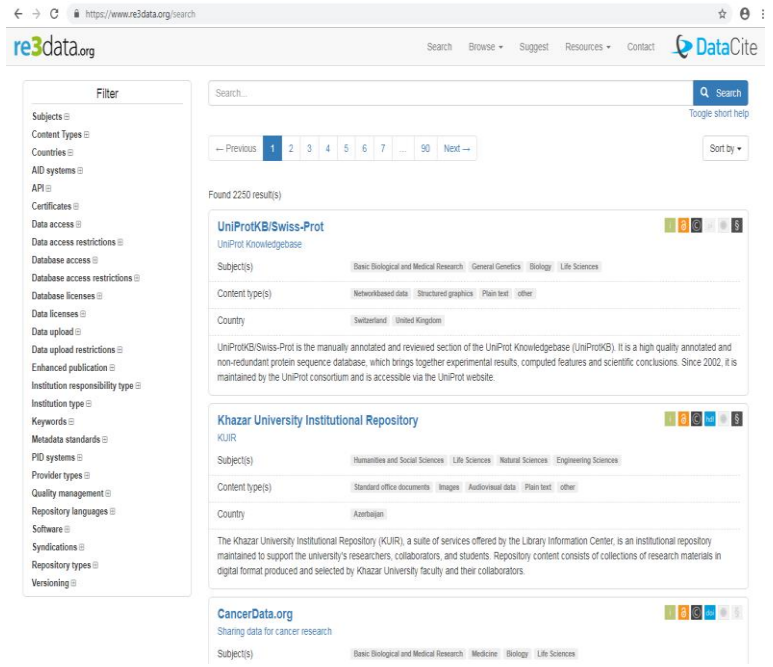
- ... archive or publish your own data.
- ... search for interesting data of other researchers.
- Registry of Research Data Repositories = re³data

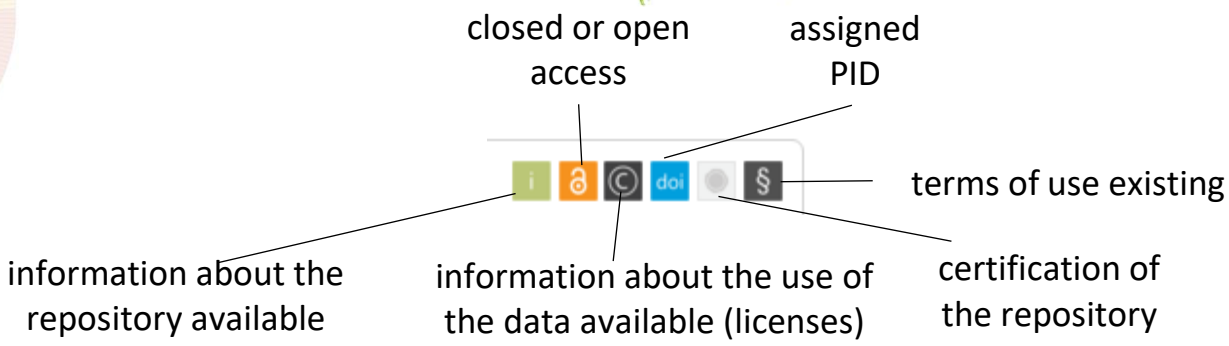
Search filters (selection):

- scientific discipline
- countries
- content types
- licenses
- identifiers
- terms of use
- data access

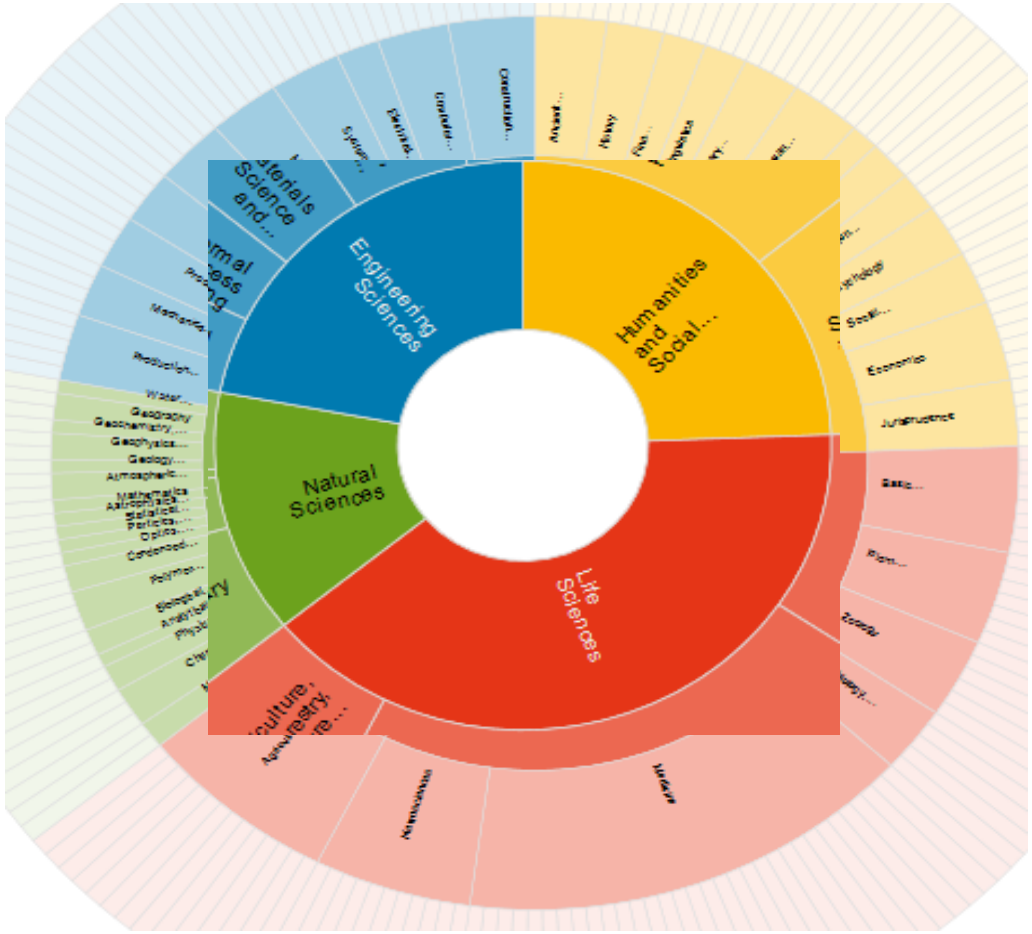
re3data.org
Registry of Institutional and Discipline-Specific Repositories

- lists 2000+ repositories
- searchable and filterable according to many criteria
- allows you to browse by subject, country, and content type
- graphical categorization of the most important features





Re3data – Registry of Research Data Repositories



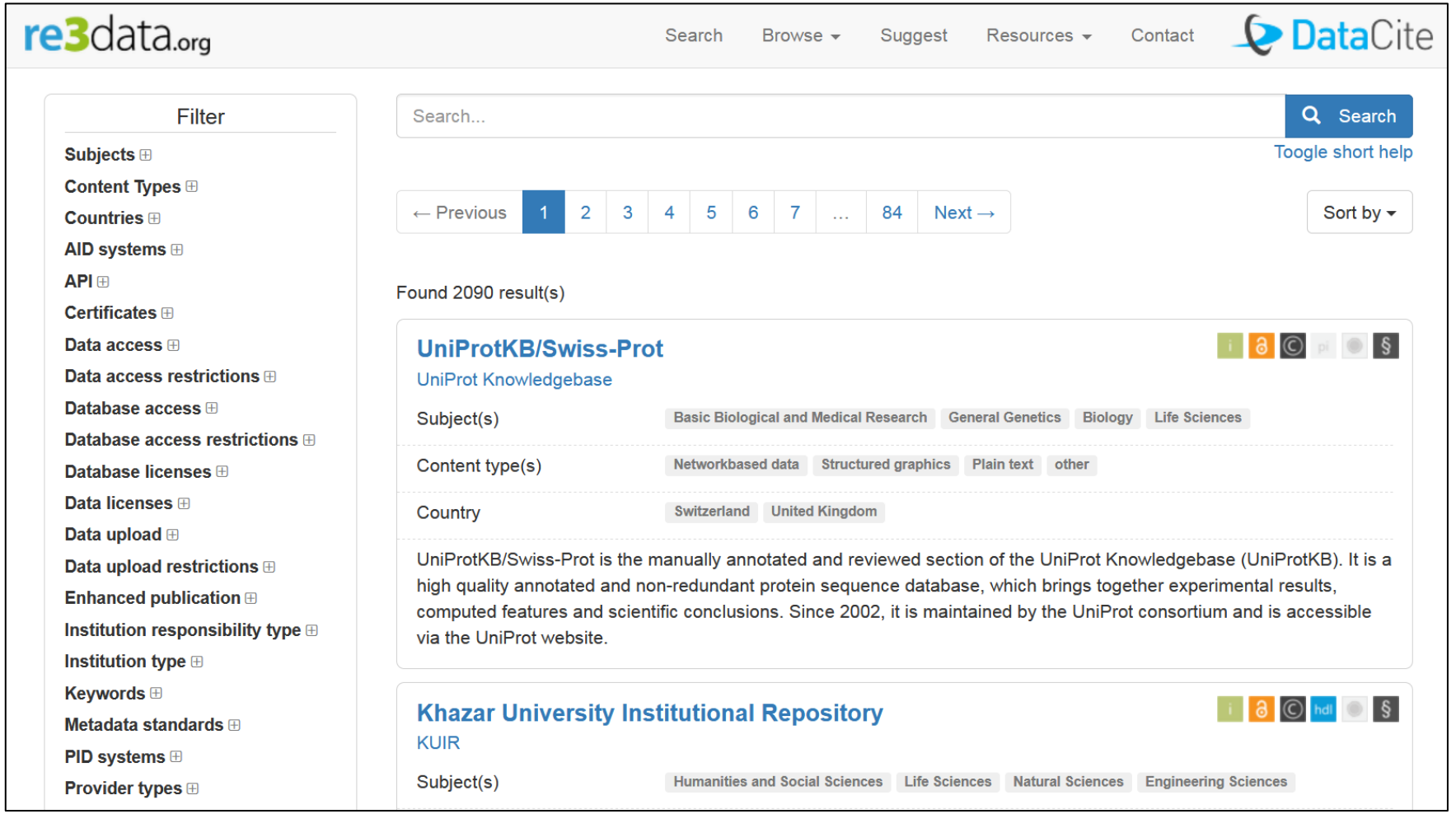
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

https://www.re3data.org/

[Bild] http://www.forschungsdaten.org/index.php/Datei:Re3data_icons.jpg

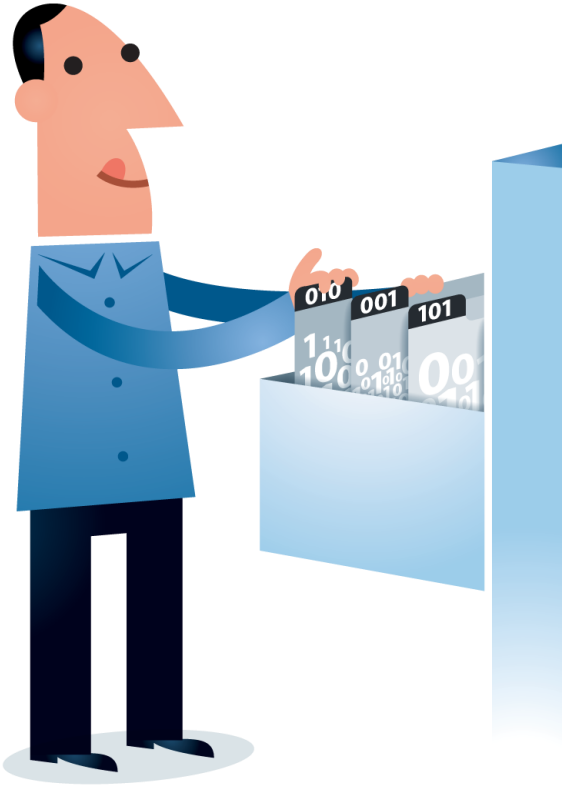
How do I find a suitable repository?

Registry of Research Data Repositories https://www.re3data.org/search



Task: re3data.org

- Please browse www.re3data.org for relevant repositories for your own research project.



www.digitalbevaring.dk

Exercise: Finding a Repository

- Search the Registry of Research Data Repositories (https://www.re3data.org/) for a suitable repository for your research data.

  10 minutes

  - How many repositories do you find for your subject?
  - Use the filter criteria (e.g. repository type, country, data upload restrictions) to further limit your result.
  - Take a look at your favorites on the detail page. What do you find regarding metadata standards, data upload and data access? Which persistent identifier system is used?

- Discussion: What did you find? What do you look for in particular when selecting a repository?

  10 minutes

Repositories: How do I choose?

Criteria:

- visibility in your community
- long-term availability
- accessibility
- persistent identifiers (e.g. DOI, URN, ARK)
- licenses offered (e.g. open access)
- review process

Discipline-specific or institutional repository?

What Criteria Do I Use to Select a Repository?

Is the repository certified? (e.g. Core Trust Seal)?

Are persistent identifiers assigned? (e.g. DOI)?

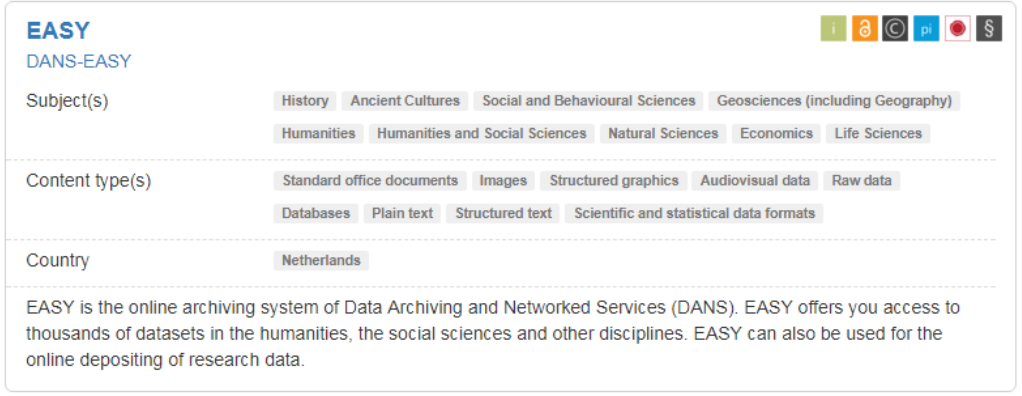
How is access to the data regulated (open, restricted)?

What are the terms of use/license for the data (e.g. Creative Commons licenses)?

Does the repository have its own policy?

How to select a data repository?

- We recommend to use a data subject / domain-oriented data repository, if available
- Check if the repository meets specific data needs, e.g. supported formats, open and restricted access mix.
- Is a persistent and unique global identifier assigned for sustainable citations and links to specific researchers and grants?
- Look for the "Trusted Digital Repository" certification with an explicit ambition to keep data available for the long term.



Icons to specify open access licenses, PIDs, certificates...

Task

- Discuss with your neighbor: What criteria would you use to select a repository? What is particularly important to you?
- Consider criteria such as costs, storage location, reliability, visibility, citability, storage duration, support offered, etc.

Task

- Do you know of a recognized repository for research data in your discipline?
- Is there an institutional repository in your institution?
- Would publishing on Figshare or Zenodo be an alternative for you?
- Search on re3data.org for possible additional repositories for your research data.
- What criteria are important to you when selecting a repository? Which repository would you select?

Steps To Share Your Data

- Step 1: Select what data you want to share
  - Not all data can be made openly available, due to ethical and commercial concerns, and you may decide that some of your intermediate data is too large to share. As such, you first need to decide which data you need to share for others to be able to reproduce your research.
- Step 2: Choose a data repository or other sharing platform
  - Data should be shared in a formal, open, and indexed data repository where possible so that it will be accessible in the long run. Suitable data repositories by subject, content type or location can be found at Re3data.org, and in FAIRsharing where you can also see which standards (metadata and identifier) the repositories implement and which journal/publisher recommend them. If possible use a repository that assigns a DOI, a digital object identifier, to make it easier for others to cite your data.
- Step 3: Choose a licence and link to your paper and code
  - So that others know what they can do with your data, you need to apply a licence to your data. The most commonly used licences are Creative Commons, Open Government Licence, or an Open Data Commons Attribution License. To get maximum value from data sharing, make sure that your paper and code both link to your data, and vice versa, to allow others understand your project better.
- Step 4: Upload your data and documentation
  - In line with the FAIR principles, upload the data in open formats as much as possible and include sufficient documentation and metadata so that someone else can understand your data. It is also essential to think about the file formats in which the information is provided. Data should be presented in structured and standardised formats to support interoperability, traceability, and effective reuse. In many cases, this will include providing data in multiple, standardized formats, so that it can be processed by computers and used by people.

What Is the Right Place for My Research Data?

- Criteria:
  - secure storage of data
  - assignment of a persistent identifier
  - landing page with descriptive metadata
  - accessible and visible for all relevant users
- Ideal would be a place …
  - … that is well-known within the relevant community.
  - … that offers the possibility to describe the data with relevant properties (subject-specific metadata).
  - … where the data is made available to the (relevant) public in a form that can be easily found and quoted.

Differentiation between Repositories and Digital Preservation

Repositories

- document servers to make research data accessible to third party users
- access to the data can be restricted by rights and role management
- hosting usually through universities or other academic institutions
- usually free access
- description of the data by metadata, including a persistent identifier.

Digital Preservation

- time-unlimited storage
- two safety aspects:

1) Substance conservation
   - Strategy: Bitstream-Preservation
2) Preservation of usability
   - Some files require special hardware or software with which they can be read.
   - changing file formats over time
   - Strategies: migration or emulation
   - Data can often only be understood when their context is known.
   - Strategies: Persistent Identifiers, Semantic Web Technologies

- Digital preservation platforms: e.g. Rosetta, Archivematica

Which Data Should Be Stored in the Repository?

- According to visibility
  - data intended for publication
  - data for exchange within a project / research group / Collaborative Research Centre
  - internal data not intended for publication
- According to processing stage
  - raw data
  - analyzed and aggregated data
  - visualized data (plots)

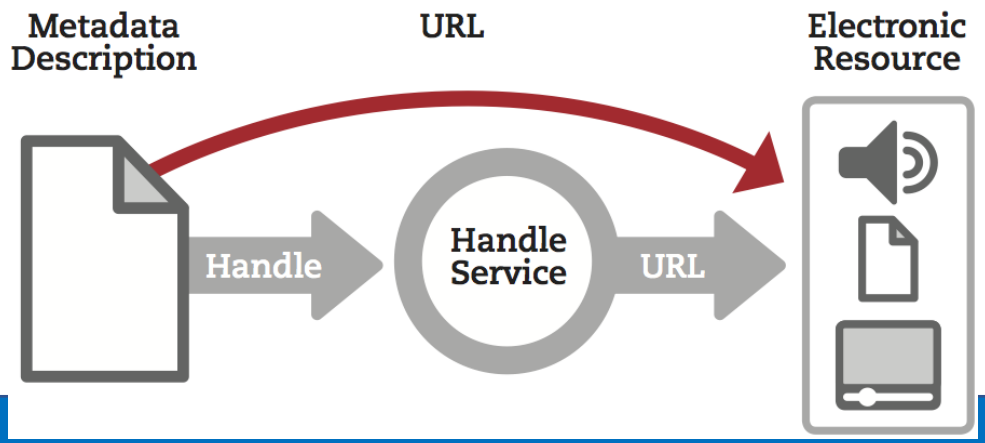
Attach persistent identifiers (PID)

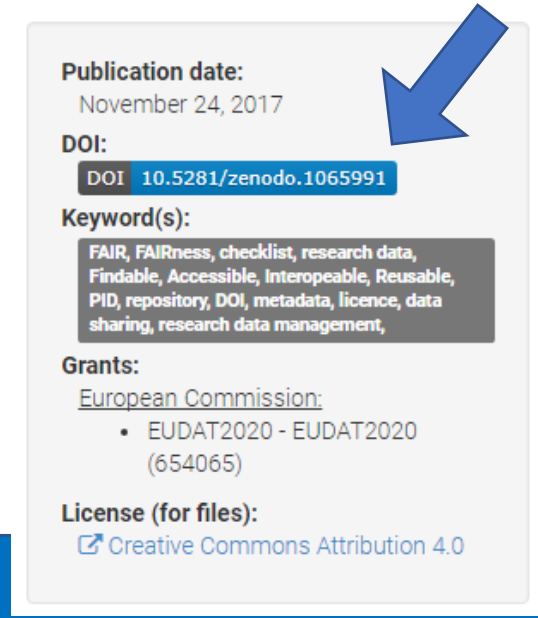
A PID is a long-term reference to a document, file, or other object

PIDs come in various forms, e.g. ARK, DOI, URN, PURL, Handles ...

They are usually actionable, meaning they can be typed in the web browser to access them

Many deposits will be allocated directly to the repository





Code ↔ Data ↔ Paper

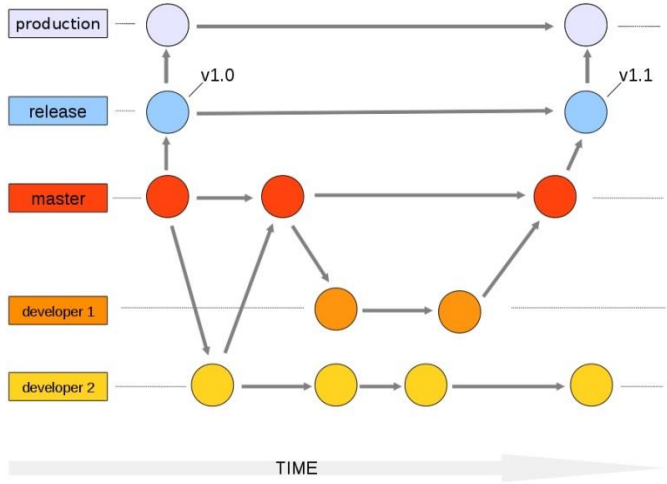
2. Research code

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)

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



Git & GitHub

- Git is a **distributed version-control system** for tracking changes in source code during software development
  - Written to manage development of Linux

- GitHub is a **social network for software development**
  - Git repository hosting service + extra features
  - GitHub provides a web-based interface on top of git.
    - Collaboration features: access control, wikis, issues, projects, …
  - Connected with **Zenodo**:
    - Got a DOI from a GitHub repo

How to create metadata?

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: »)

   - 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?

Citation metadata for Data Repositories. Key:

Statistical power and underpowered statistics...
www.refsmmat.com/statistics/power.html

Citation Metadata	Dublin Core[a]	Schema.org[b]	DataCite[c]	DATS[d]
Dataset Identifier	identifier	@id*	identifier	identifier
Title	title	name	title	title
Creator**	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	datePublished	publicationYear	date
Version	*not available*	version	version	version
Type	type	type	resourceTypeGeneral	type

[a]Dublin Core Metadata Element Set (https://dublincore.org/documents/dces/);

[b]Dataset - Schema.org (https://schema.org/Dataset);

[c]DataCite Metadata Working Group[21];

[d]Gonzalez-Beltran & Rocca-Serra[22,23];

*name of ID field depends on schema.org serialization format, it is **@id** for JSON-LD;

Metadata can use any standard...

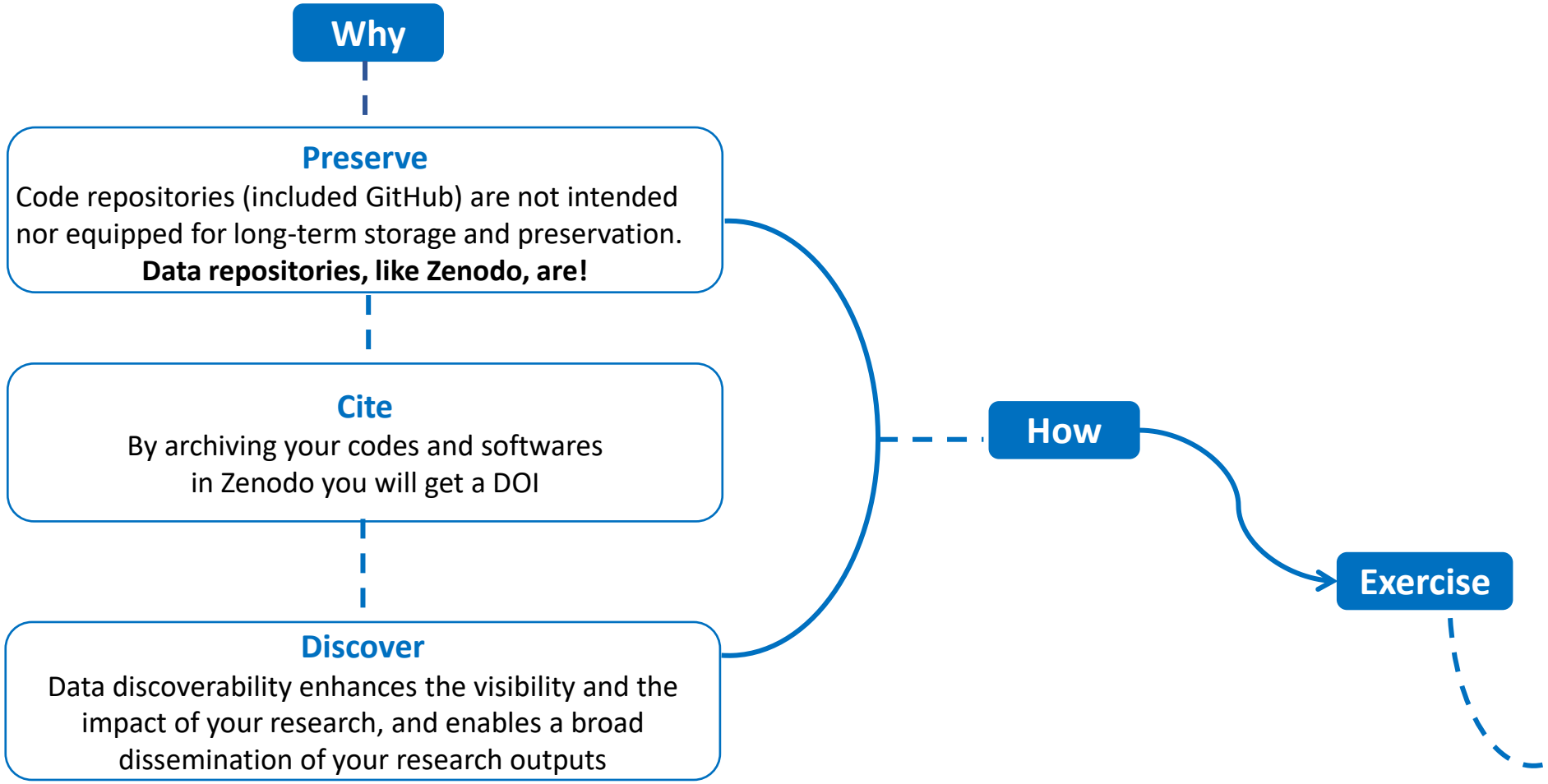

```
{
"login": "torvalds",
"id": 1024025,
"avatar_url": "https://avatars.githubusercontent.com/u/1024025?v=3",
"gravatar_id": "",
"url": "https://api.github.com/users/torvalds",
"html_url": "https://github.com/torvalds",
"followers_url": "https://api.github.com/users/torvalds/followers",
"following_url": "https://api.github.com/users/torvalds/following/other_user",
"gists_url": "https://api.github.com/users/torvalds/gists/gist_id",
"starred_url": "https://api.github.com/users/torvalds/starred/owner/repo",
"subscriptions_url": "https://api.github.com/users/torvalds/subscriptions",
"organizations_url": "https://api.github.com/users/torvalds/orgs",
"repos_url": "https://api.github.com/users/torvalds/repos",
"events_url": "https://api.github.com/users/torvalds/events/privacy",
"received_events_url": "https://api.github.com/users/torvalds/received_events",
"type": "User",
"site_admin": false,
"name": "Linus Torvalds",
"company": "Linux Foundation",
"blog": null,
"location": "Portland, OR",
"email": null,
"hireable": null,
"bio": null,
"public_repos": 2,
"public_gists": 0,
"followers": 31456,
"following": 0,
"created_at": "2011-09-03T15:26:22Z",
"updated_at": "2015-06-11T00:46:13Z",
}
```


Figure User meta information example. Display the meta information of a Github user by accessing the Github API with url="https://api.github.com/users/Torvalds"

Important discovery metadata for Data Repositories. Key:

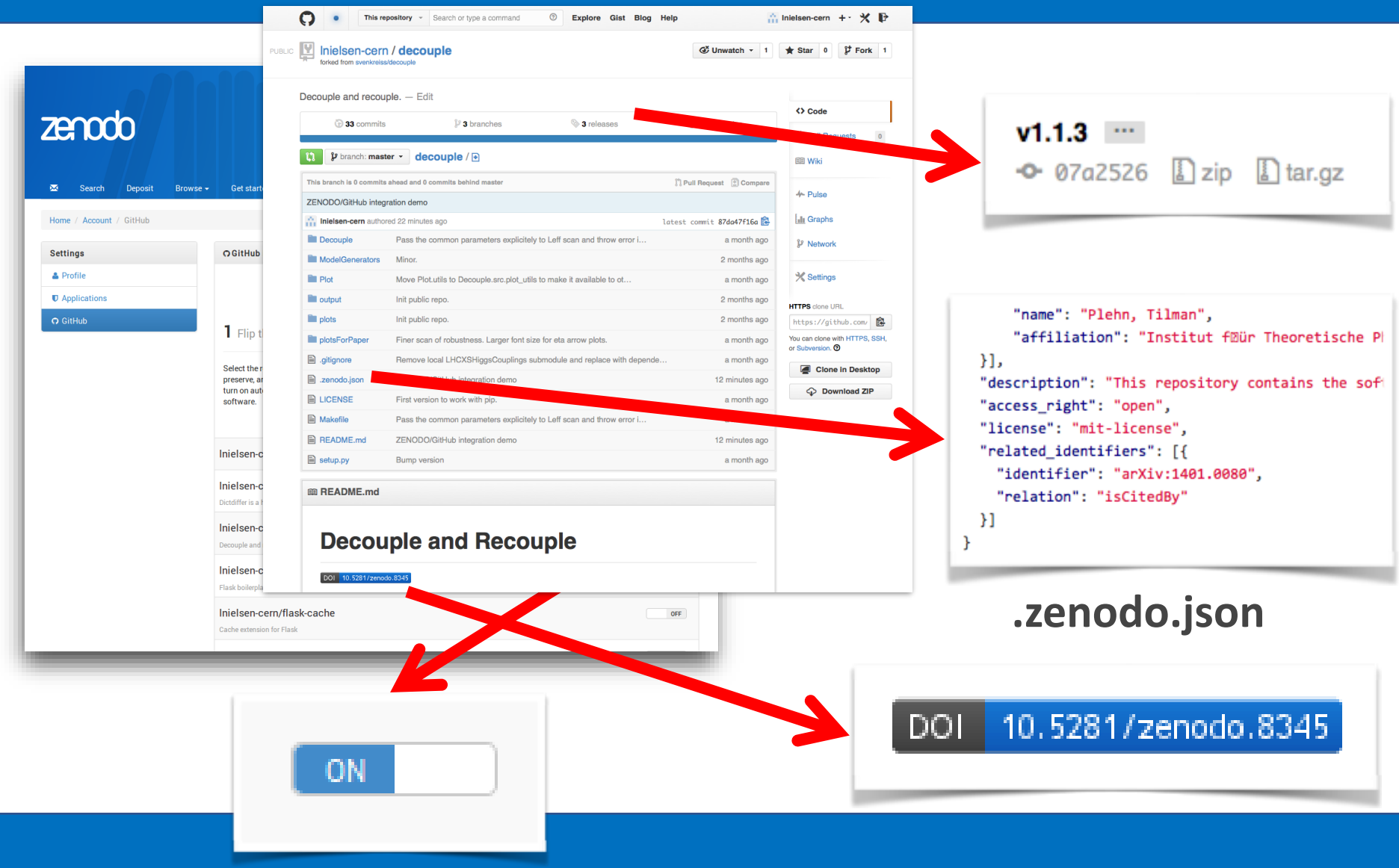
Discovery Metadata	Dublin Core	Schema.org	DataCite	DATS
Description	description	description	description	dataType dimension Material...*
Keywords	subject	keywords	subject	keywords
License	license	license	rights	license
Related Dataset**	isPartOf isVersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related Publication***	bibliographicCitation	citation	relatedIdentifier	publication

From GitHub to Zenodo... Why & How



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

Zenodo – GitHub bridge



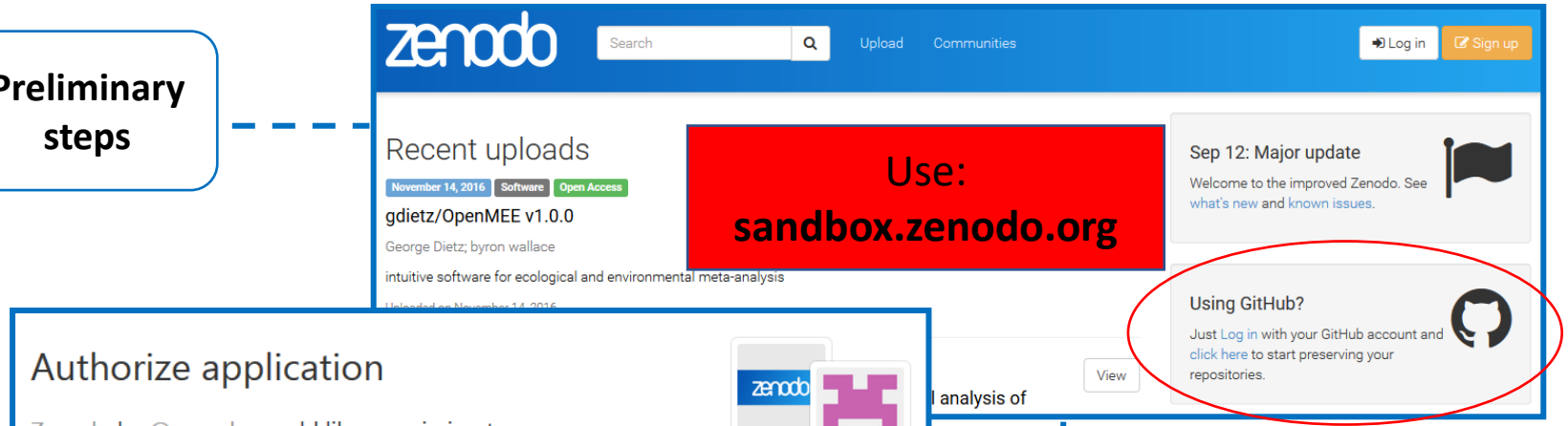
.zenodo.json

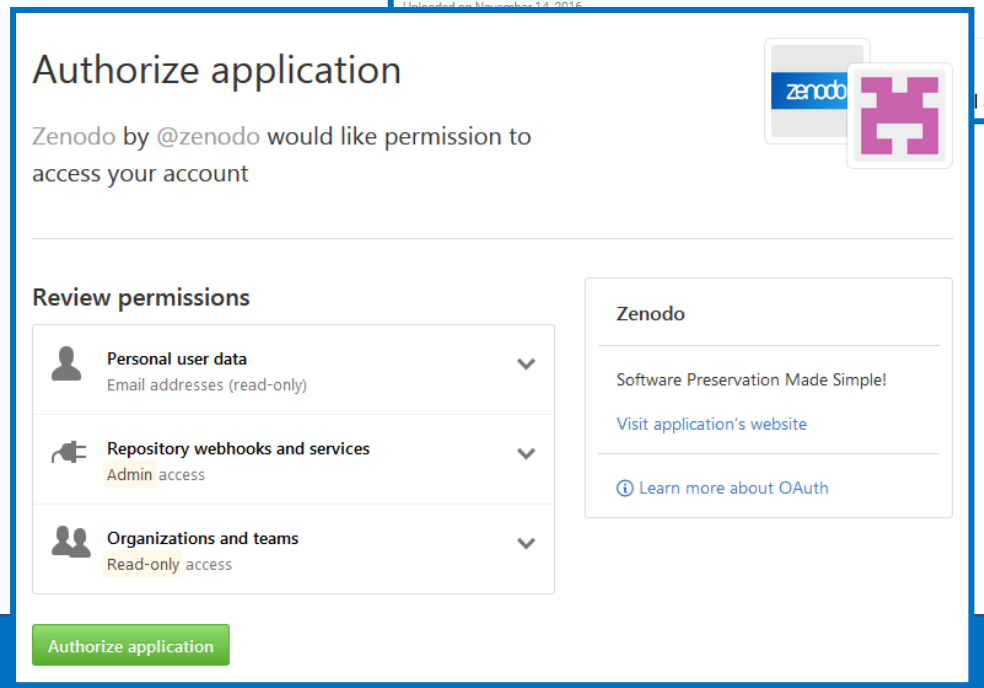
Capture and publish your GitHub repository in Zenodo

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**

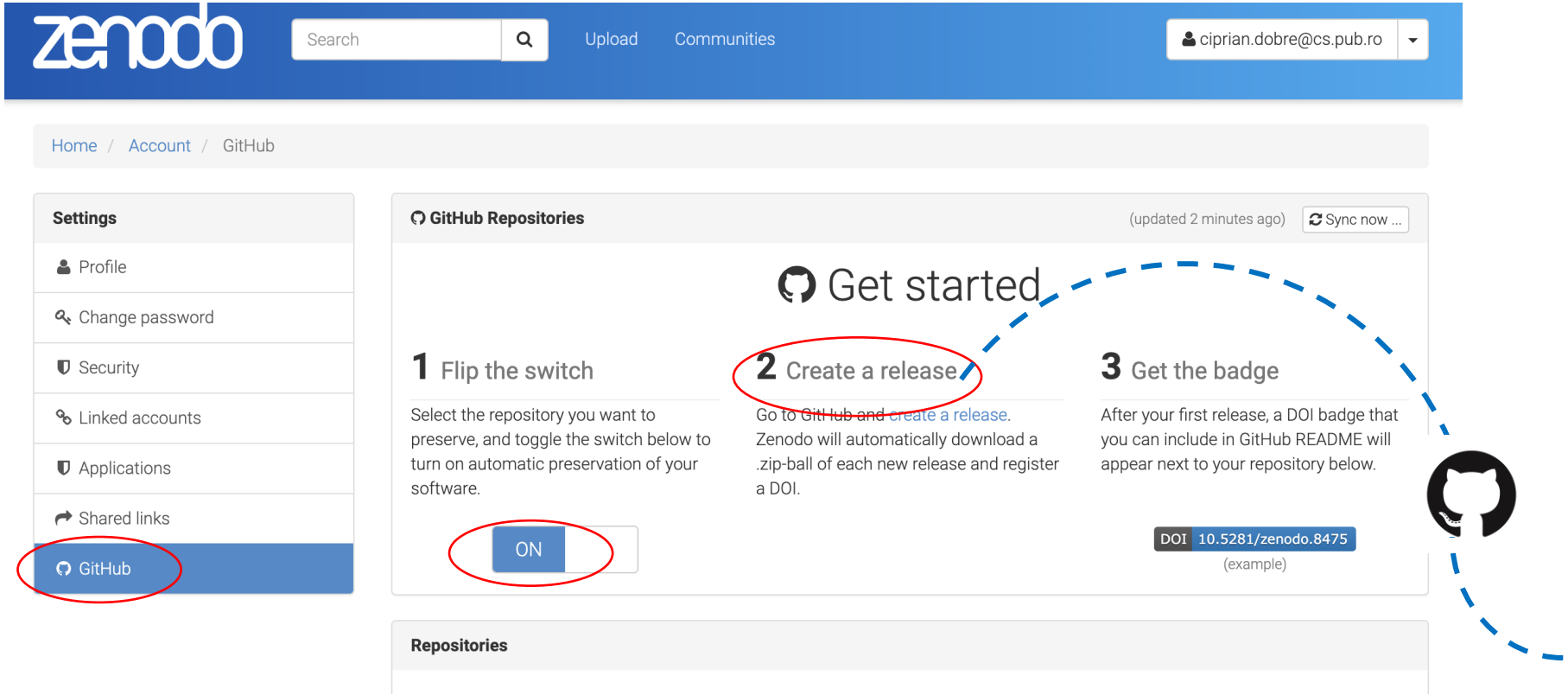
Preliminary steps





Capture and publish your GitHub repository in Zenodo

Exercise



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/

3. Reproducible Environments

Reproducible Environments

- Every computer has its unique computational environment consisting of its operating system, installed software, versions of installed software packages, and other features that we will describe later.
- Suppose a research project is carried out on one computer but transferred to a different computer. In order for research to be reproducible, the computational environment that it was conducted in must be captured in such a way that others can replicate it.

```
a = 1
b = 5
print(a/b)
```

One divided by five is `0.2`, and this is what is printed if the script is run using Python 3. However, if a slightly older version of Python, such as Python 2, is used, the result printed is `0`. This is because integer division is applied to integers in Python 2, but (normal) division is applied to all types, including integers, in Python 3.

Capturing Computational Environments

What is reproduced? \ Interaction style	Graphical	Command line
Software & versions	binder	CONDA
Entire system		docker

Open Notebooks

- Electronic Lab Notebooks (ELNs) enable researchers to organize and store experimental procedures, protocols, plans, notes, data, and even unfiltered interpretations using their computer or mobile device.
- They are a digital analogue to the paper notebook most researchers keep.
- ELNs can offer several advantages over the traditional paper notebook in documenting research during the active phase of a project, including; searchability within and across notebooks, secure storage with multiple redundancies, remote access to notebooks, and the ability to easily share notebooks among team members and collaborators.

Jupyter Notebooks

Reproducible Open Science

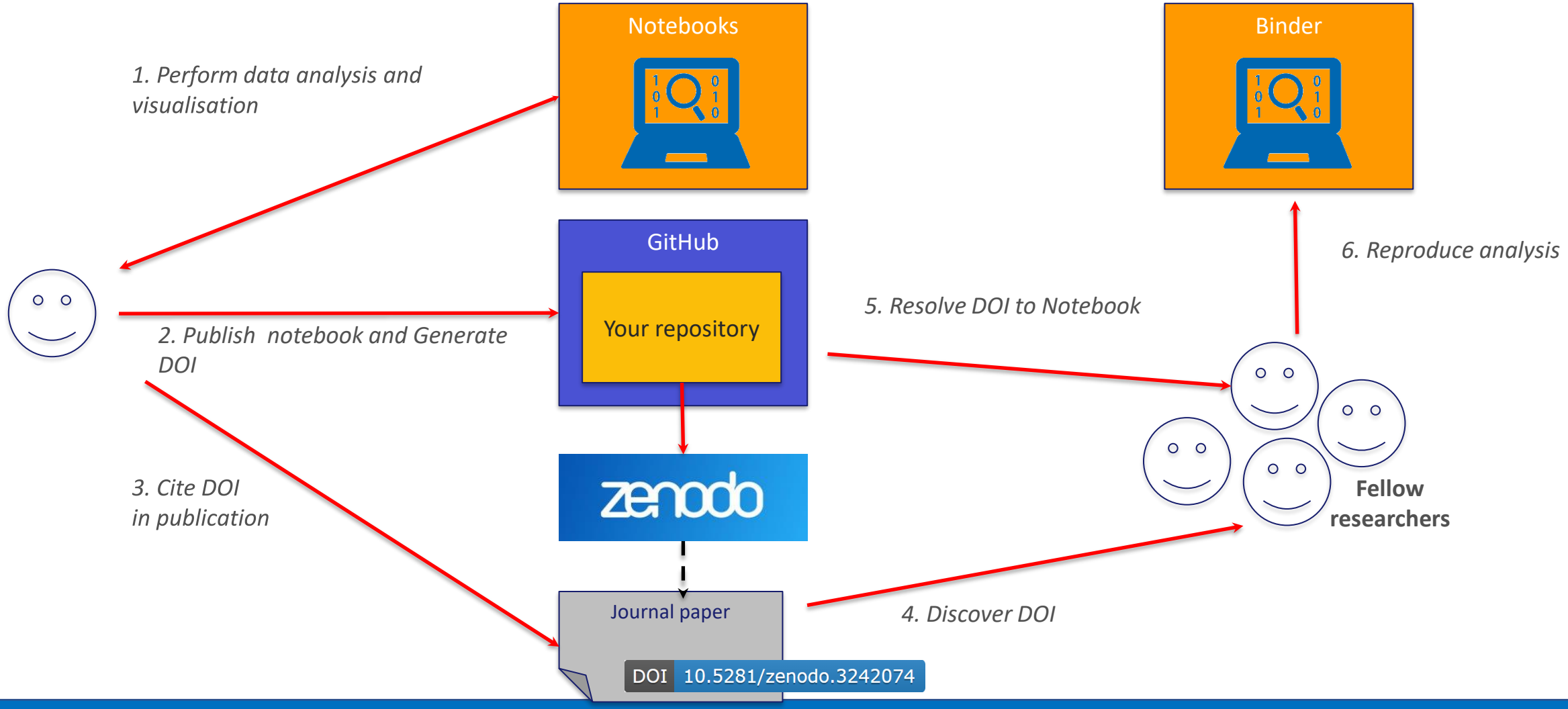
- The computational tools to solve a problem
  - Python, R, Julia, and wide ecosystem of libraries and tools for science
- An interface to facilitate coding/creating
  - Jupyter
- A way to communicate your work
  - Notebooks
- Leverage on the EGI Cloud to scale-up the resources
- A way to share your work
  - GitHub, Zenodo or other similar repositories
- A way to pack it all for replication
  - Docker (used by Binder)
- A way to persistently identify it
  - DOIs (Digital Object Identifiers)

binder

jupyterhub

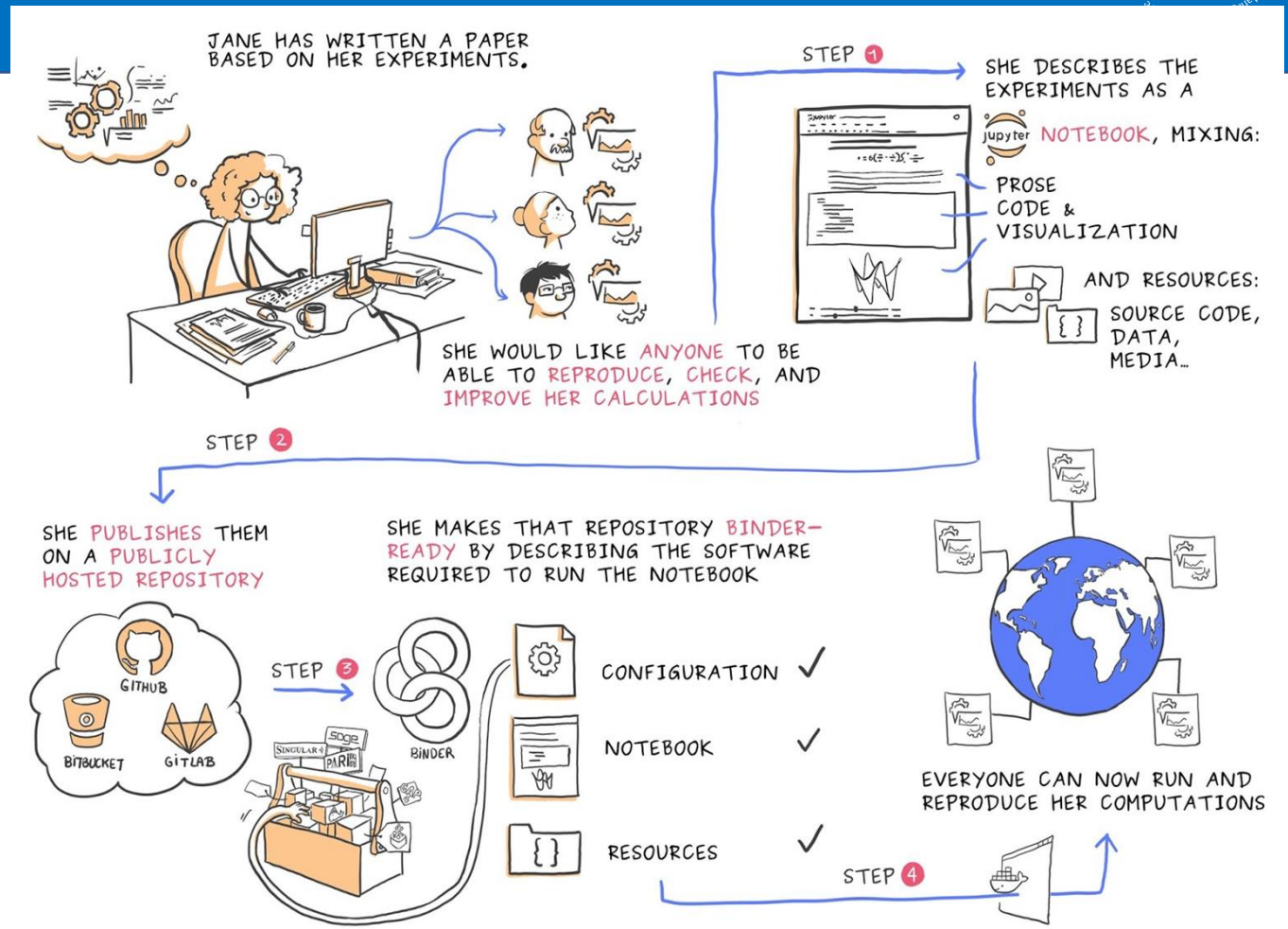
zenodo

Implement Open Science



Binder

- An open-source web application to turn repositories in interactive notebooks
- It uses Modern technology in cloud orchestration (Kubernetes), interactive computing (Jupyter), scientific computing (the open-science ecosystem)



Credit: Juliette Taka https://twitter.com/JulietteTaka

What does Binder do?

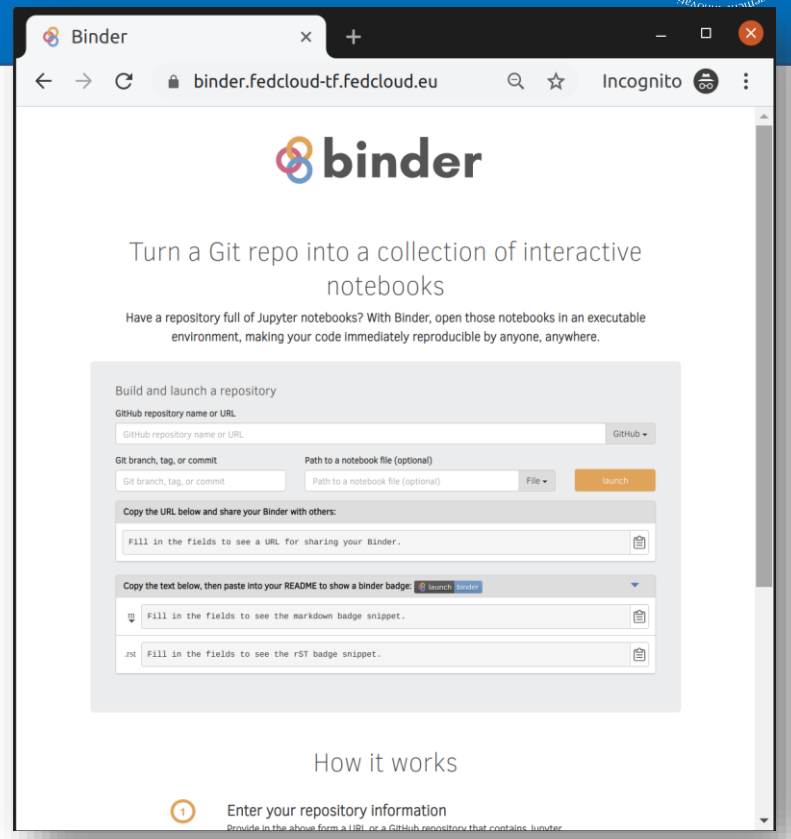
Pulls code from repository

GitHub

repo2docker: Creates reproducible containers from repositories

jupyterhub: Generates user sessions that serve these containers

Kubernetes: manages the computing infrastructure

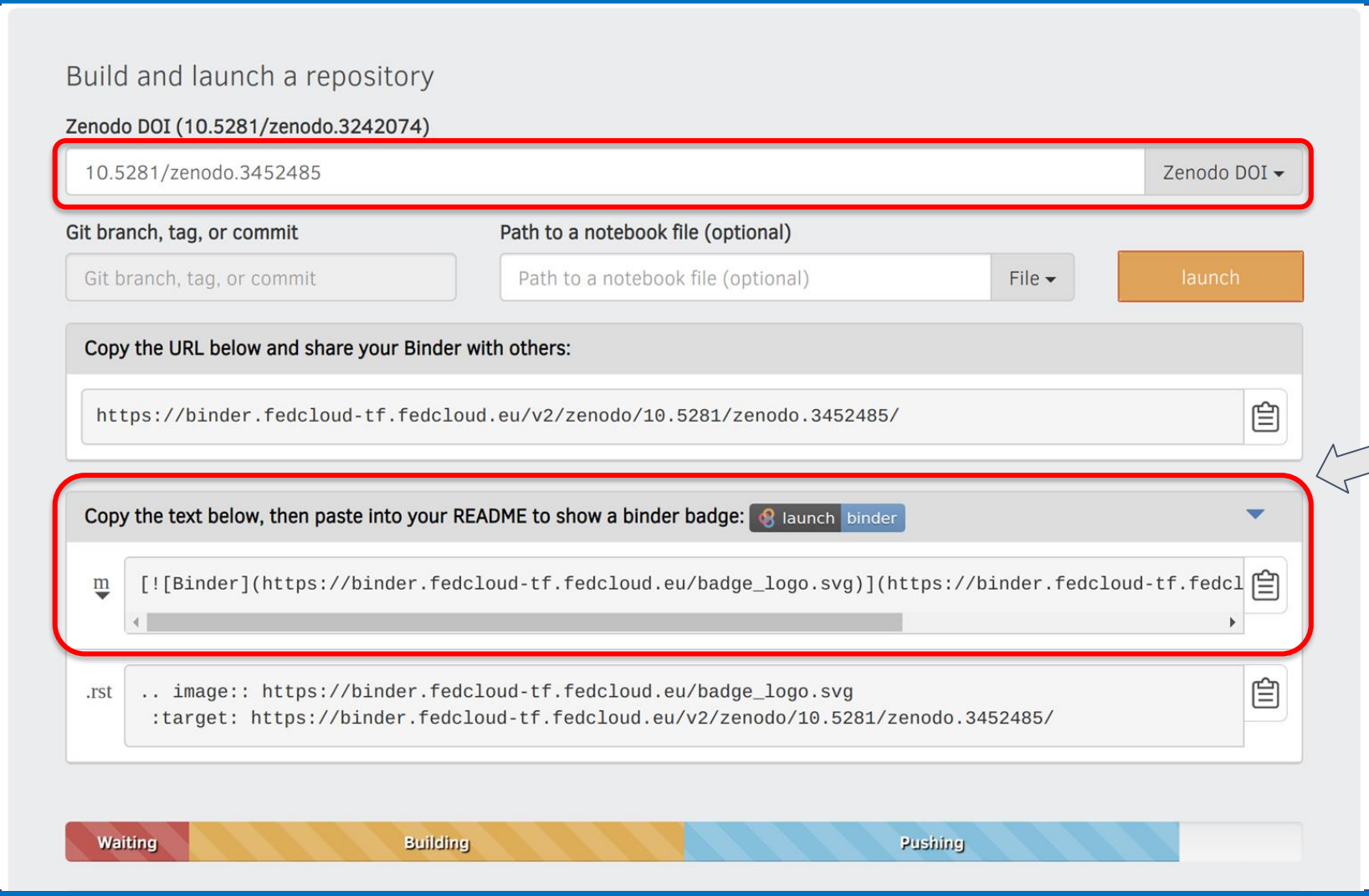


EGI Binder:
Provides an interface to create, share, and share the sessions

Compute Resources

- BinderHub is a cloud-based technology that can launch a repository of code (from GitHub, GitLab, and others) in a browser window such that the code can be executed and interacted with.
  - A unique URL is generated allowing the interactive code to be easily shared.
- The purpose of these Binder instances is to promote reproducibility in research projects by encouraging researchers to document their software dependencies and produce fun, interactive environments!
- Binder, as a user interface, is useful for reproducibility because the code needs to be version controlled and the computational environment needs to be documented in order to benefit from the functionality of Binder.
- Each change to the code repository forces a new build of the Binder instance. This acts as a proxy for continuous integration of the computational environment as the Binder instance will break if the configuration file is not updated.
- BinderHub relies on different tools and resources in order to create and launch the Binder instances.

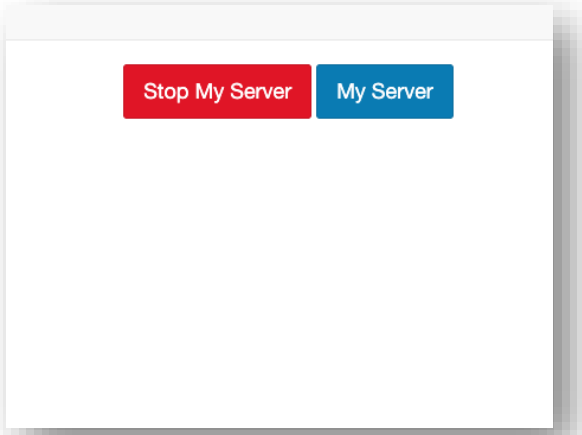
Reproduce your analysis with Binder



WARNING:
Registration of the DOI can take up to 10 minutes
Binder has to wait until registration is finished.

To check whether the DOI is resolved, please use the website: https://dx.doi.org/

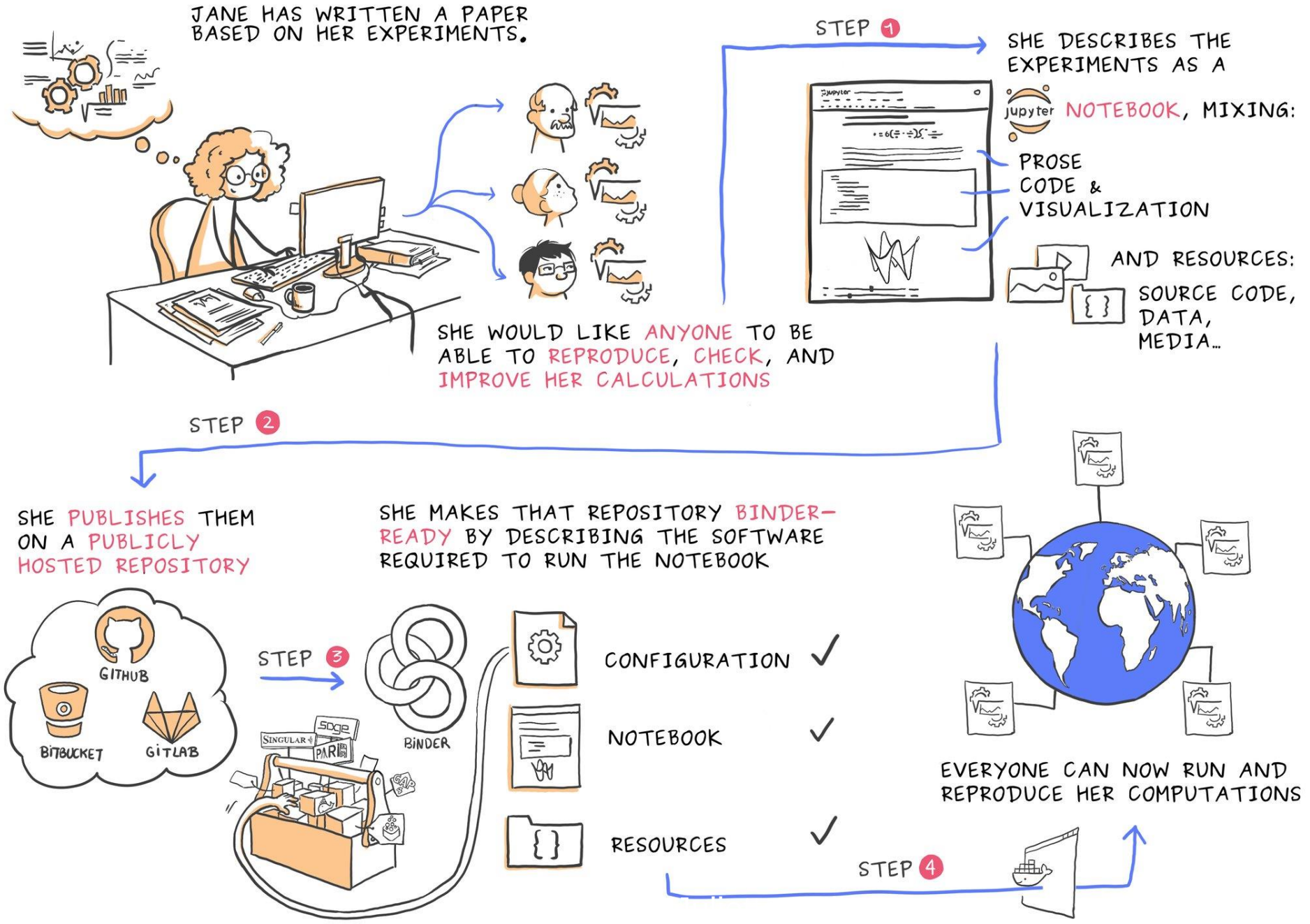
Create a LaunchBinder button in your repo



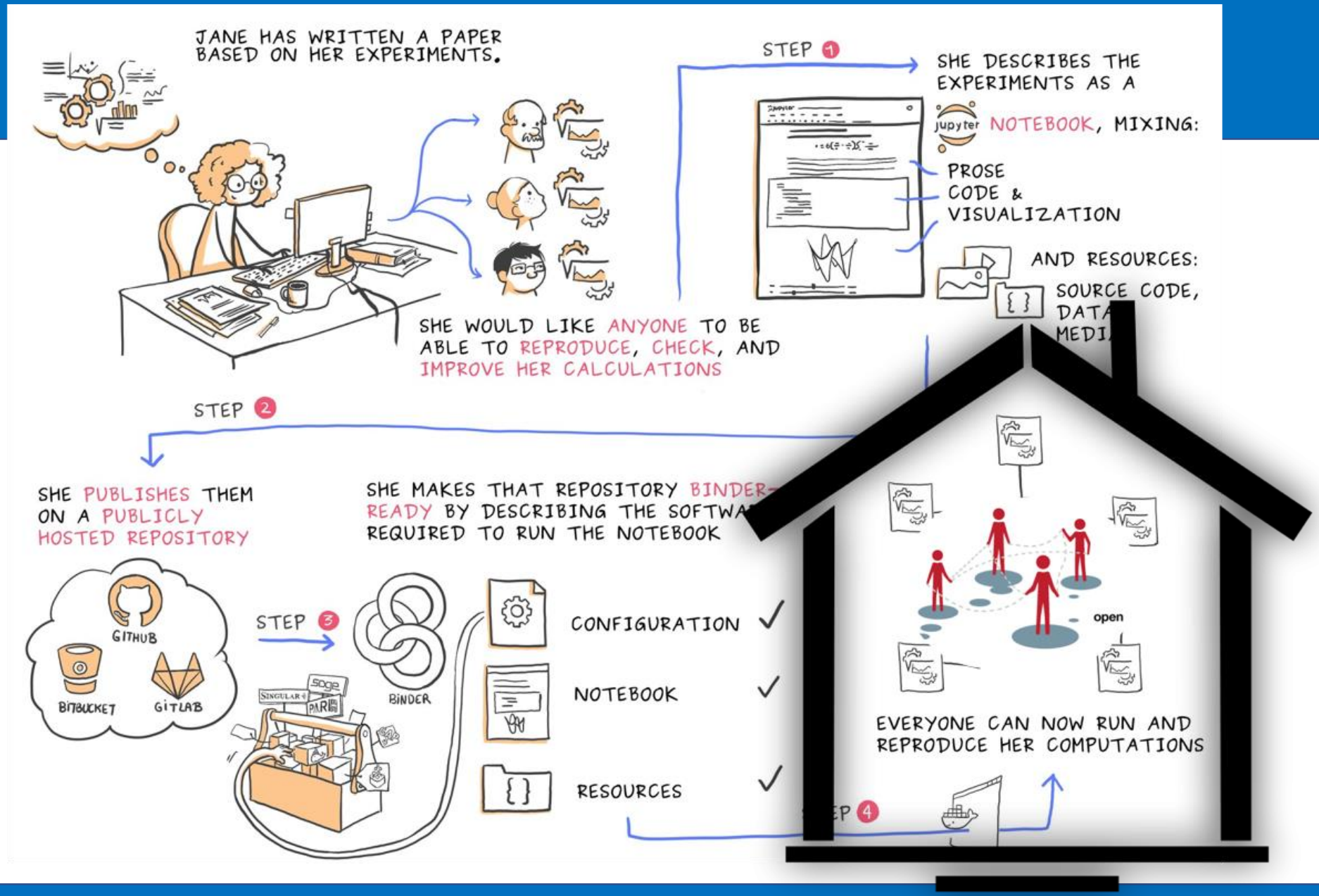
Go to https://cs3.fedcloud-tf.fedcloud.eu/hub/home
and stop your server before retry!

What happens when a Binder link is clicked?

1. The link to the repository is resolved by BinderHub.
2. BinderHub searches for a Docker image relating to the provided reference (for example, git commit hash, branch or tag).
3. **If a Docker image is not found**, BinderHub requests resources from the Kubernetes cluster to run repo2docker to do the following:
    1. Fetch the repository,
    2. Build a Docker image containing the software requested in the configuration file,
    3. Push that image to the Docker registry.
4. BinderHub sends the Docker image to JupyterHub.
5. JupyterHub requests resources from the Kubernetes cluster to serve the Docker image.
6. JupyterHub connects the user's browser to the running Docker environment.
7. JupyterHub monitors the container for activity and destroys it after a period of inactivity.



https://twitter.com/JulietteTaka/status/1082735653929000960



Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960 and Sarah Gibson

4. Collaborative platforms

Open Science Framework - OSF

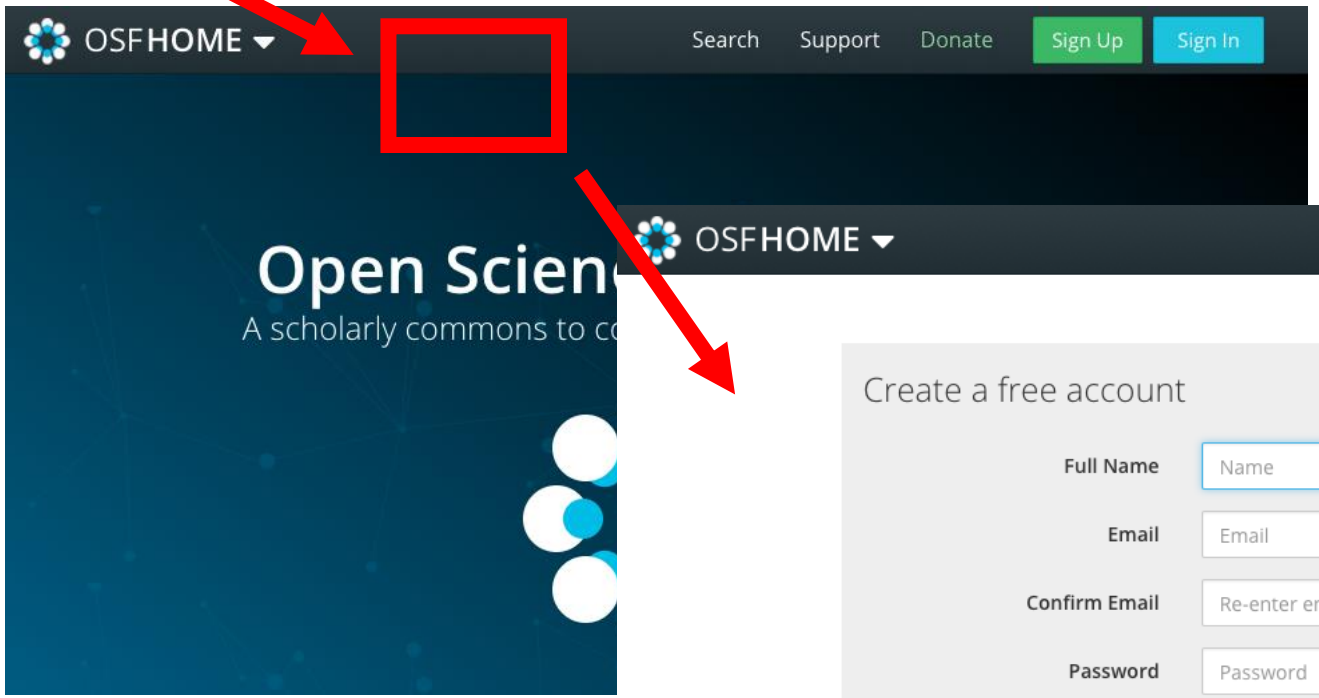
- Collaborative OS platform that brands itself as "a scholarly commons to connect the entire research cycle"
- OSF enables researchers to work on projects privately with a limited number of collaborators and make any part or the whole of their project public.
- It connects directly with many other collaborative systems like dropbox, GitHub and Google Docs, and can be used to store and archive research data, protocols, and materials.

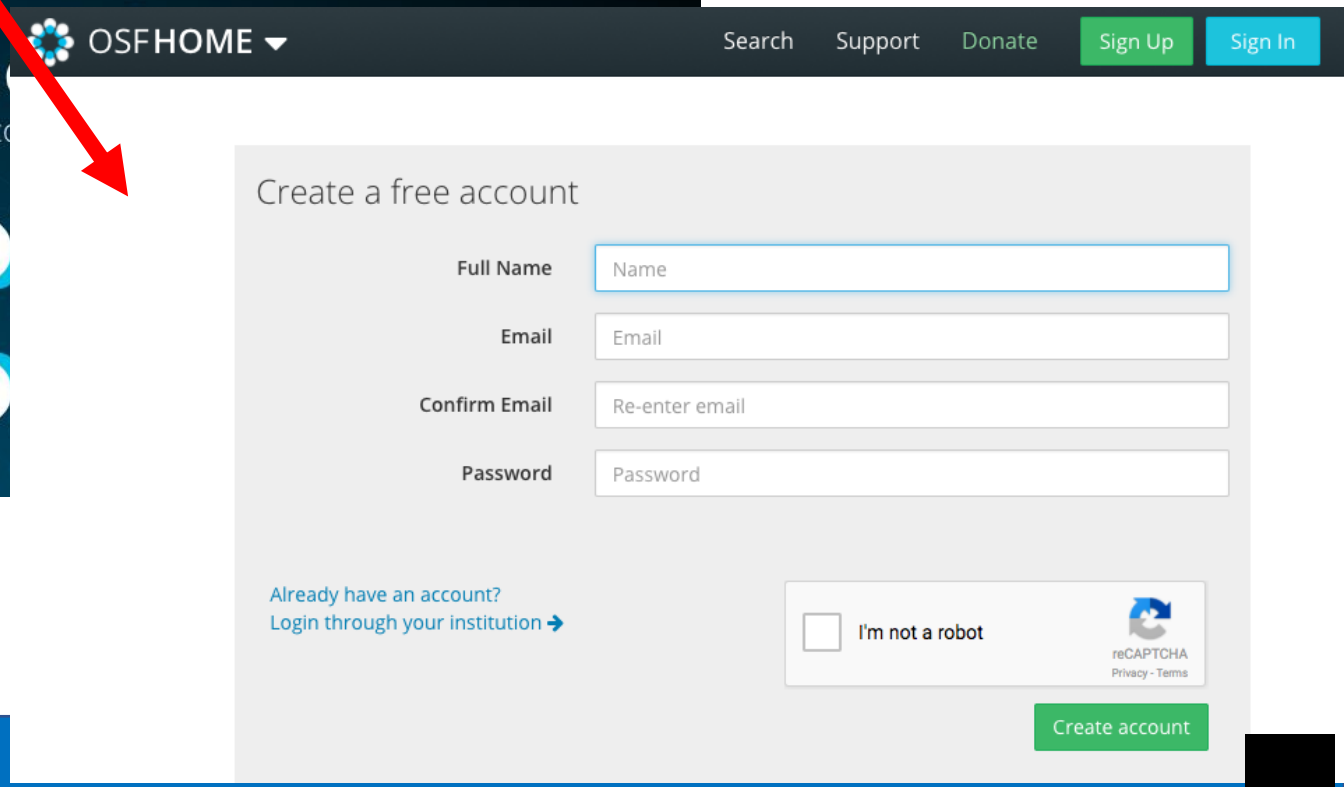
Tool: Open Science Framework

Topic	OSF feature
Metadata	Wiki
Hierarchical organization	Components
Storage	OSF and Add-ons
Access	Contributors
Versioning	Built in version control

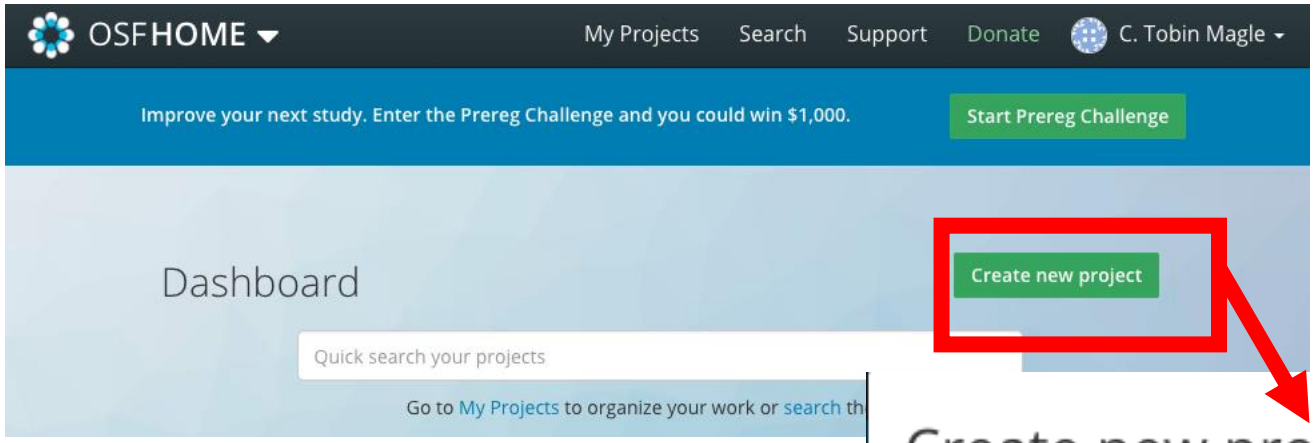
Creating an account

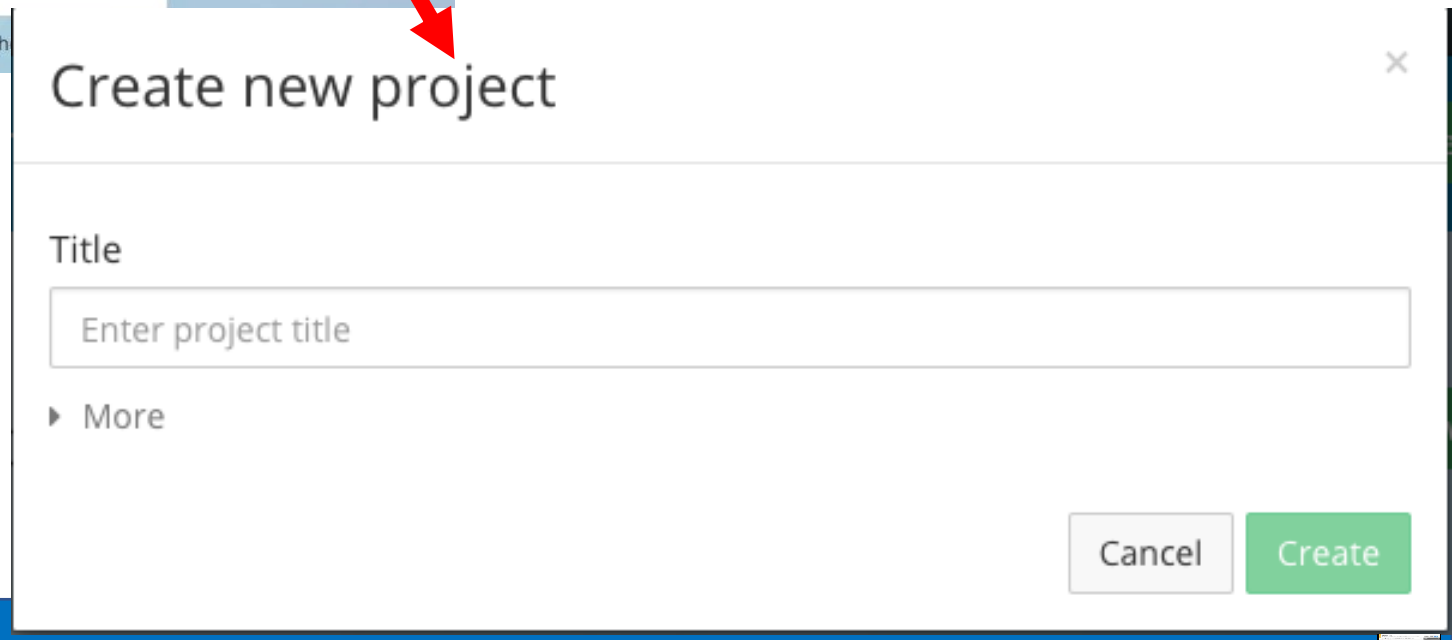
Go to http://osf.io





Add a Project

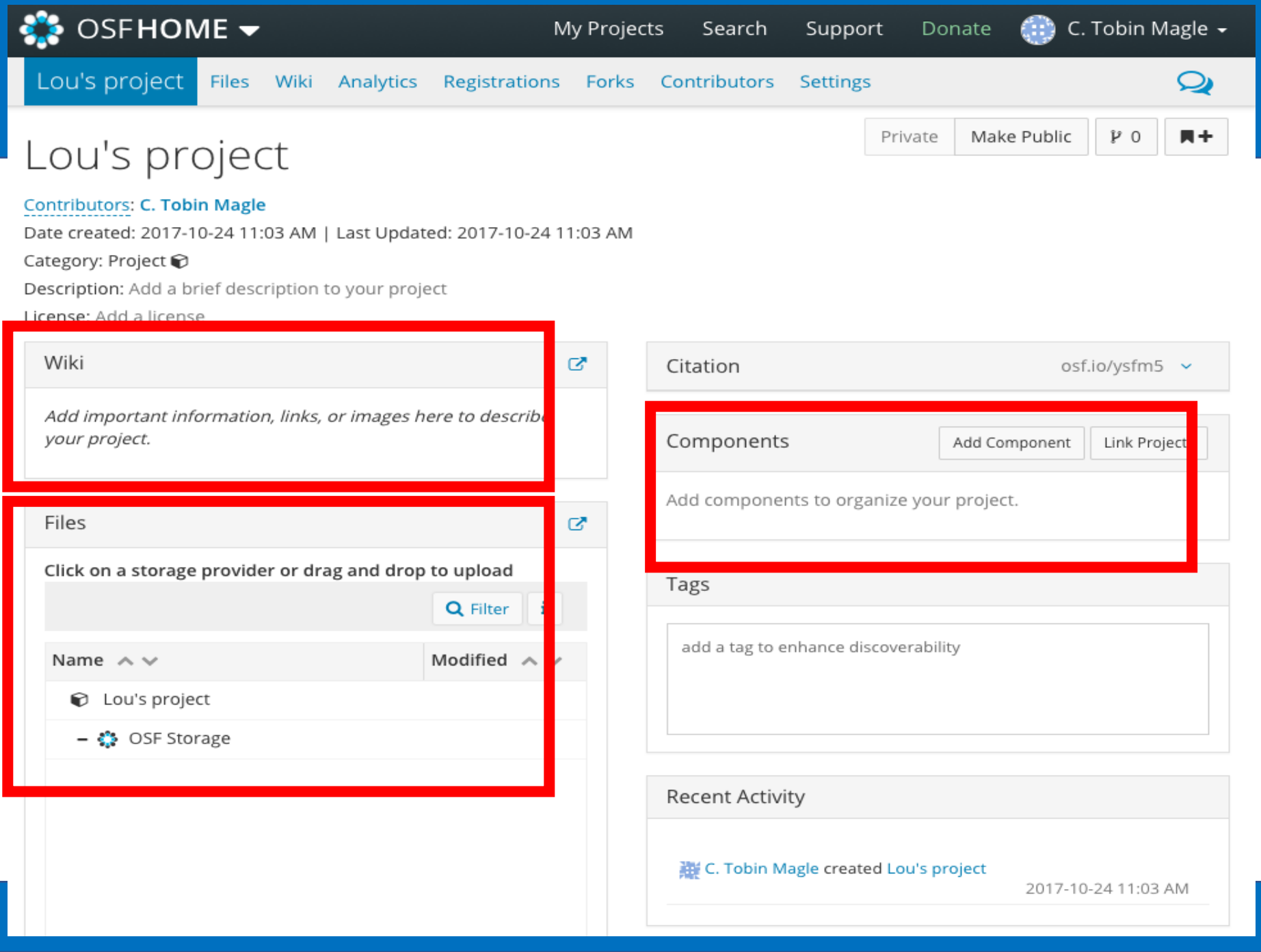




Exercise

Lou is a first year graduate student working on a project in a biomedical research laboratory. He's trying to decipher data left by a former post doc as a start for his thesis project.

- **Download** Lou's files: http://tinyurl.com/hvna4mg
- Create an OSF account
- Create a project called "Lou's project"


OSF HOME
My Projects
Search
Support
Donate
C. Tobin Magle
Lou's project
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Private
Make Public
0
Lou's project
Contributors: C. Tobin Magle
Date created: 2017-10-24 11:03 AM | Last Updated: 2017-10-24 11:03 AM
Category: Project
Description: Add a brief description to your project
License: Add a license
Wiki
Add important information, links, or images here to describ
your project.
Citation
osf.io/ysfm5
Components
Add Component
Link Project
Add components to organize your project.
Files
Click on a storage provider or drag and drop to upload
Filter
Name
Modified
Lou's project
OSF Storage
Tags
add a tag to enhance discoverability
Recent Activity
C. Tobin Magle created Lou's project
2017-10-24 11:03 AM

Context: Wiki

- Describe the project
- Evolves during project
  - Progress report
  - Goals
- Code book:
  - ID systems (for records)
  - Variable systems

Estimating the Reproducibility of Psychological Science

Open Science Collaboration

Abstract: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects (Mr = .197, SD = .257) were half the magnitude of original effects (Mr = .403, SD = .188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($p < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Citation: Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716

Contents

Summary Report: Read the *Science* article and supplementary material summarizing the results of the Reproducibility Project: Psychology. Or, read the Green OA version with supplementary information in the same file.

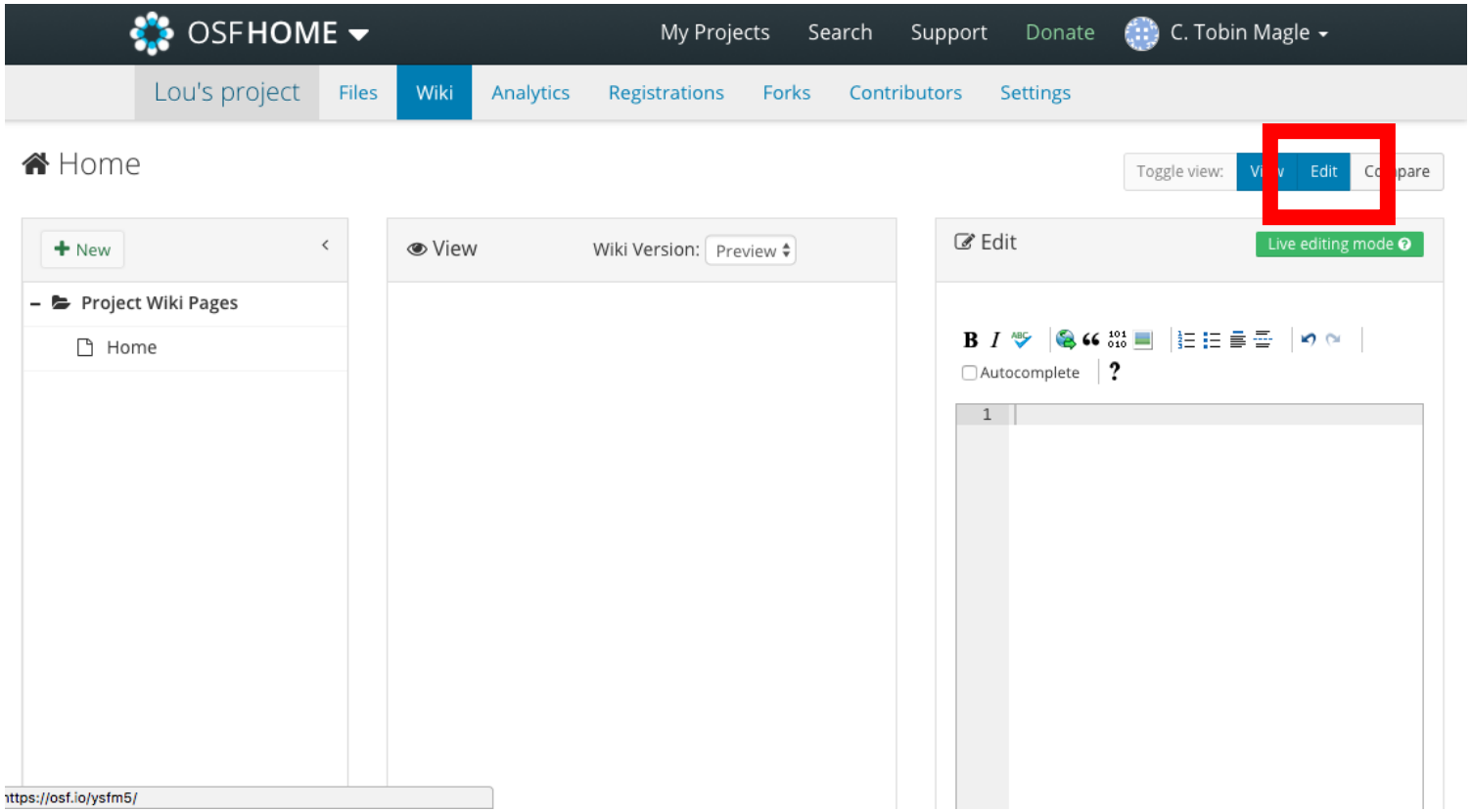
Supplement only. Supplementary materials to "Estimating the Reproducibility of Psychological Science." Includes additional graphs and details on analyses.

Replicated Studies: Explore the preregistrations, materials, data, and result reports of the individual replication projects.

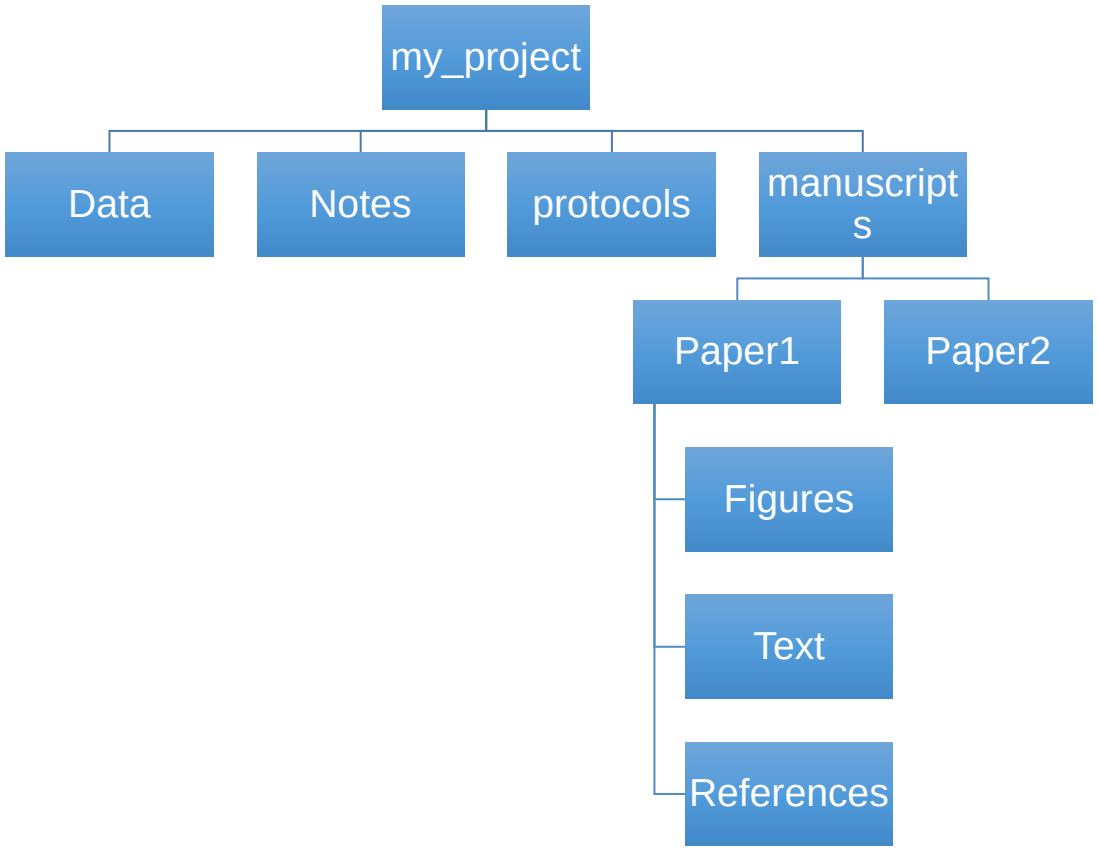
Guide to Analyses: Reproduce the analyses of the individual projects and the aggregate results.

RPP Process: Learn more about the design, management, and operation of this large-scale crowdsourced project.

Wiki page


OSF HOME
My Projects
Search
Support
Donate
C. Tobin Magle
Lou's project
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Home
Toggle view:
Edit
New
Project Wiki Pages
Home
View
Wiki Version:
Preview
Edit
Live editing mode
Autocomplete
1
https://osf.io/ysfm5/

Hierarchical Organization

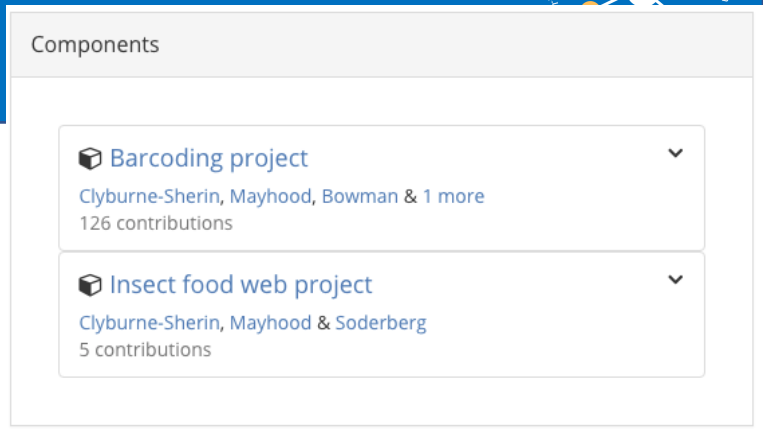


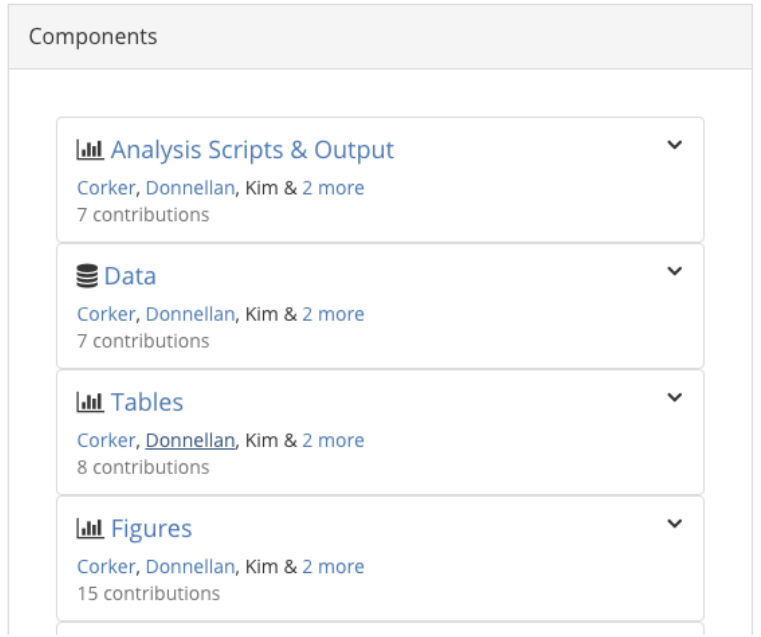
- “folders in folders”
- So you can find things!
- Agree on a standard with your collaborators

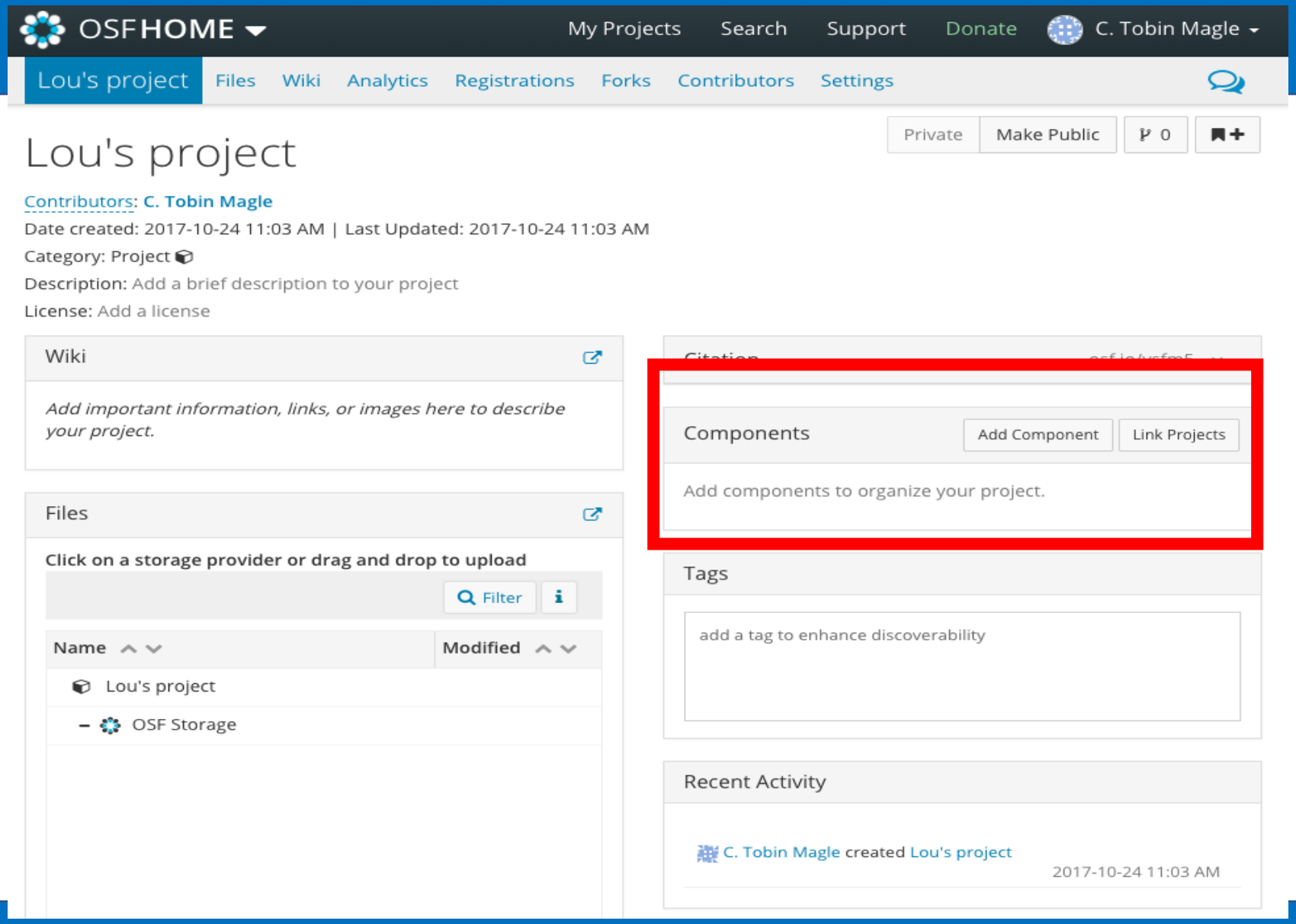
Components

- “Subprojects”

- Separate privacy settings, contributors, wiki, add-ons, and files.
- Examples:
  - Different projects: https://osf.io/82fba/
  - Clinical: https://osf.io/gq4mz/
  - Manuscript: https://osf.io/if7ug/
  - Collaboration: https://osf.io/ezcuj/

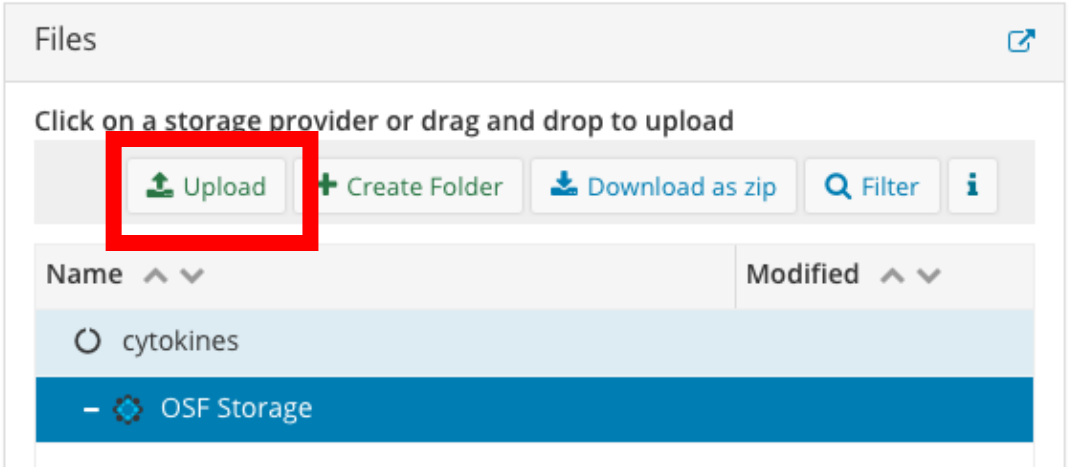




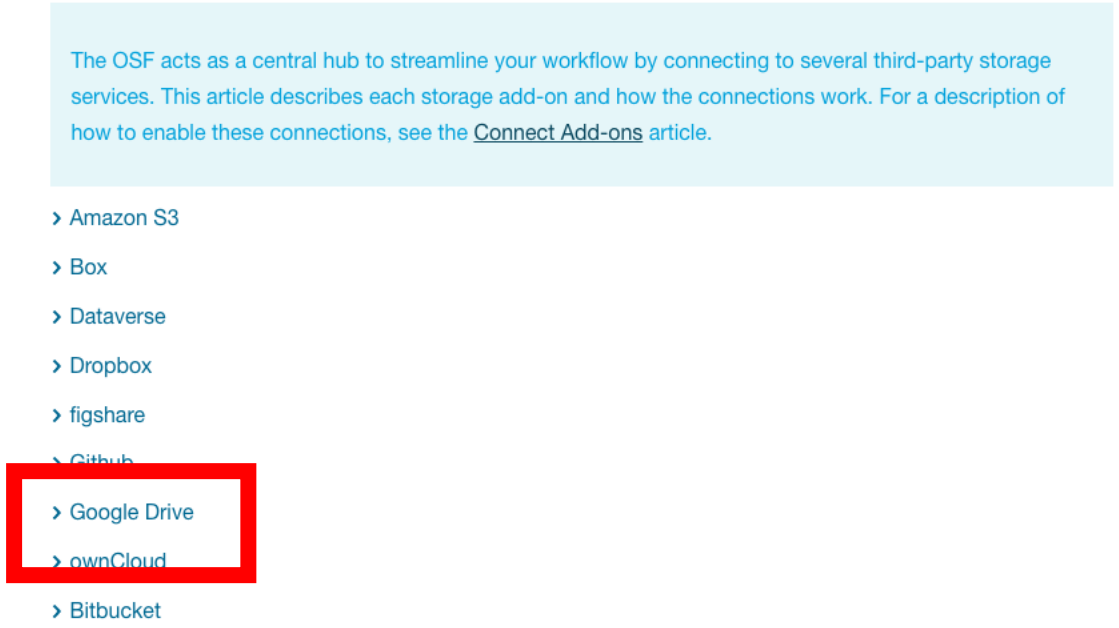

OSF HOME
My Projects
Search
Support
Donate
C. Tobin Magle
Lou's project
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Private
Make Public
0
Lou's project
Contributors: C. Tobin Magle
Date created: 2017-10-24 11:03 AM | Last Updated: 2017-10-24 11:03 AM
Category: Project
Description: Add a brief description to your project
License: Add a license
Wiki
Add important information, links, or images here to describe your project.
Files
Click on a storage provider or drag and drop to upload
Filter
Name
Modified
Lou's project
OSF Storage
Components
Add Component
Link Projects
Add components to organize your project.
Tags
add a tag to enhance discoverability
Recent Activity
C. Tobin Magle created Lou's project
2017-10-24 11:03 AM

Add files to OSF

- Click on OSF storage
- Click upload
- Select files from file browser



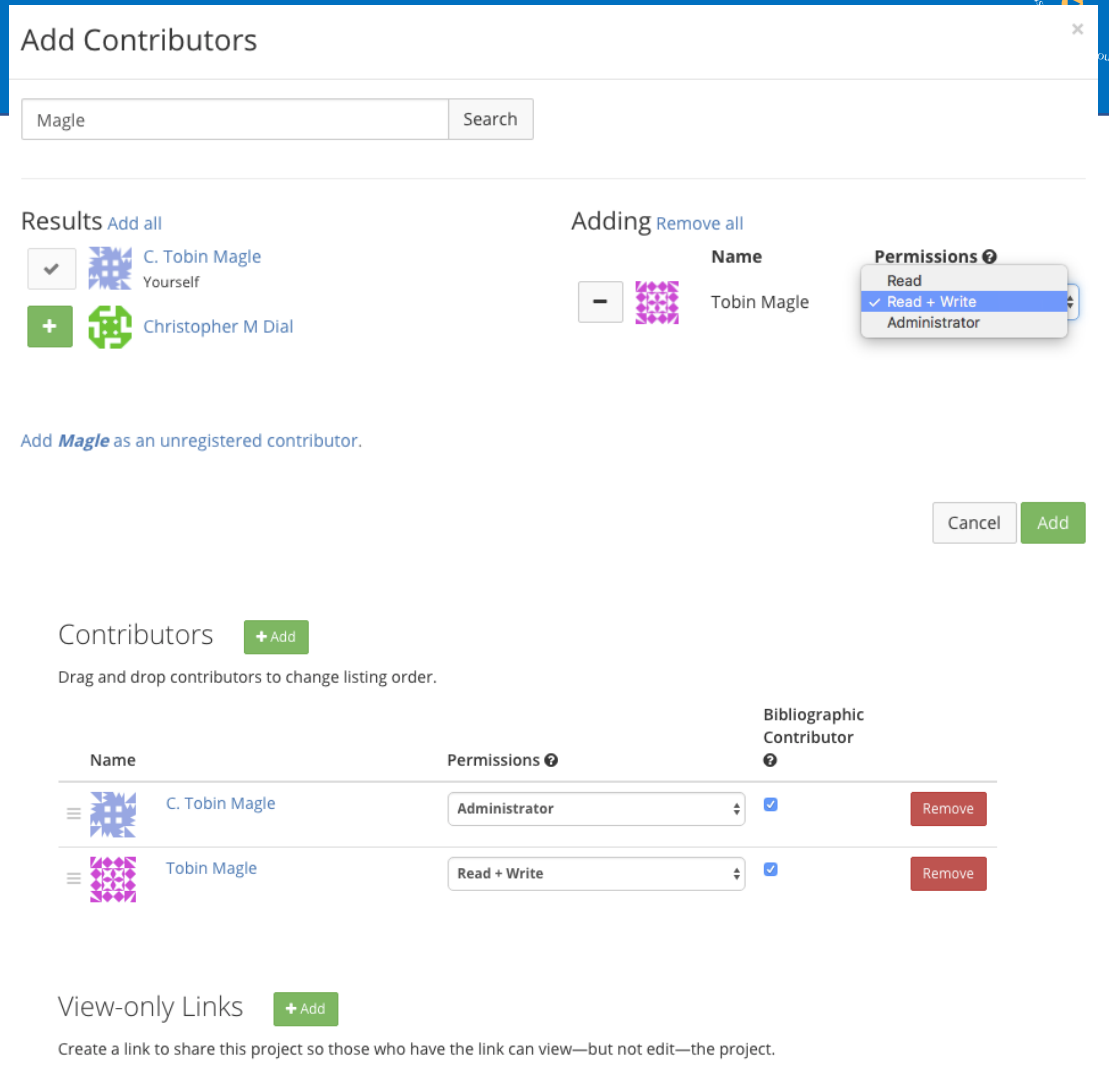
Storage Add ons



http://help.osf.io/m/addons/l/524149-storage-add-ons

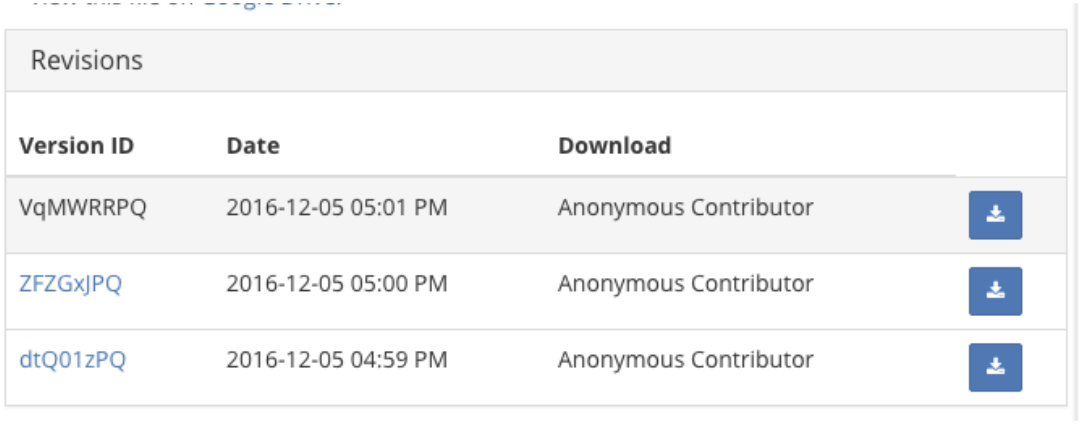
Contributors

- Control who can see what
  - Administrator
  - Read/Write
  - Read only
- Separate for each component

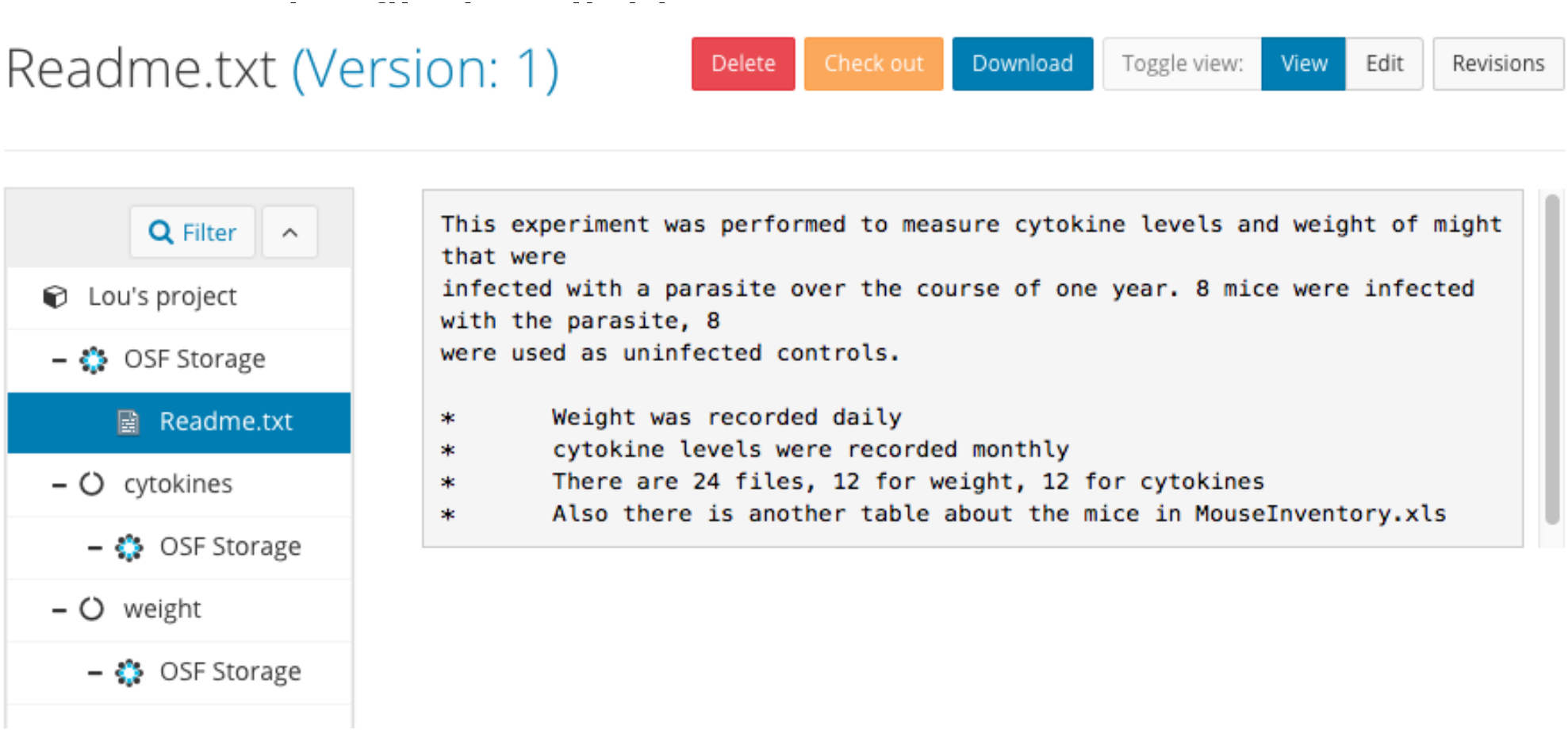


Version control

- Who did what when?
  - Native in OSF
  - GitHub Integration
- One file name,
  many versions



Version history



Other Collaborative Platforms...

- Collaborative writing platforms
  - Online tools and platforms for researchers to work together on documents in real-time, and so avoid the versioning-hell of emailing Word documents back and forth.
  - Platforms include Overleaf, Authorea, Fidus Writer, ShareLaTeX and Google Docs.
  - Many of these tools are based on proprietary technologies and some require payment for advanced features.
- Reference management & discovery
  - There are plenty of tools which enable groups to store and manage references.
  - Examples include Zotero, Citavi and CiteUlike. Mendeley incorporates a sharable reference manager, as well as a social network and article visualization tools. Relatedly, BibSonomy allows researchers to share bookmarks and lists of literature.
- Annotation and review
  - The power of the Web enables new modes of post-publication collaborative review through services like PubPeer and Academic Karma, as well as annotation tools like Hypothes.is and PaperHive.
- Academic social networks
  - Researchers have long made use of the Web for social networking - either via mainstream social networks like Twitter, Facebook and Linkedin or dedicated academic social networks like ResearchGate, Academia.edu and Loop.

The future... EOSC

Projects	Offer
EOSC Future	Operates the EOSC-Core, the set of internal services that facilitates EOSC operations upon which researcher services can rely and integrate with as appropriate: Resource Catalogue, Marketplace, AAI Federation, Order Management, Helpdesk, Monitoring, Accounting
C-SCALE	Services for the discovery, access and processing of EO/Copernicus data in EOSC.
DICE	Cutting-edge data management services (data archive, policies-based data storage, personal and project workspaces, data repository and data discovery services) with a significant amount of storage resources for EOSC users. https://marketplace.eosc-portal.eu/services?related_platforms=53
EGI-ACE	Delivers the EOSC Compute Platform and contributes to the EOSC Data Commons through a federation of cloud compute and storage facilities, Platform as a Service (PaaS), and data spaces with analytics tools and federated access services. https://marketplace.eosc-portal.eu/services?related_platforms=52
OpenAIRE Nexus	A framework of services for EOSC that can greatly assist in publishing research, monitoring its impact and helping to promote its discovery.
Reliance	Services to boost the discovery of and access to research data (including Copernicus data), improve the extraction of relevant information and manage the research lifecycle via research objects.
ARCHIVER	ARCHIVER (Archiving and Preservation for Research Environments) has spent over 3 years designing, prototyping and piloting innovative new services for the Long Term Digital Preservation (LTDP) of scientific datasets. During the project, multiple data intensive organizations representing several research domains (CERN, DESY, PIC and EMBL-EBI) have worked closely and collaboratively with suppliers (Arkivum and LIBNOVA) on the research and development of new services and solutions for scientific data preservation relevant for the European Open Science Cloud (EOSC). https://archiver-project.eu/

THANK YOU!

Follow us