Github for Data and Data Science

Alex Sanchez-Pla

Genetics, Microbiology and Statistics Department (UB)

Statistics and Bioinformatics Unit (VHIR)

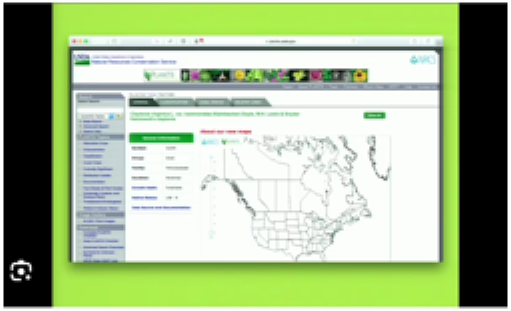
2023-07-06

Outline

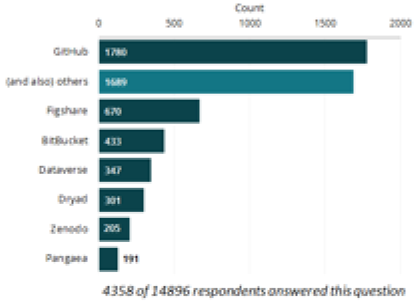
1. Introduction: On Open Science and its Tools
2. Data Sharing with GitHub
3. Open Data Repositories Projects
4. Pros & Cons of some approaches
5. Summary

Introduction

Is GitHub good for data sharing?

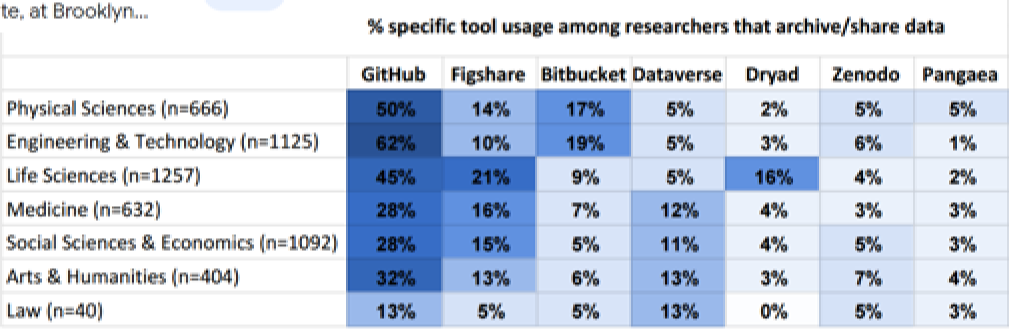


The Use of GitHub, an Open Source Code and Data Sharing Website, at Brooklyn...



4358 of 14896 respondents answered this question

Figure 3 – Survey results: tools used for archiving and sharing data & code

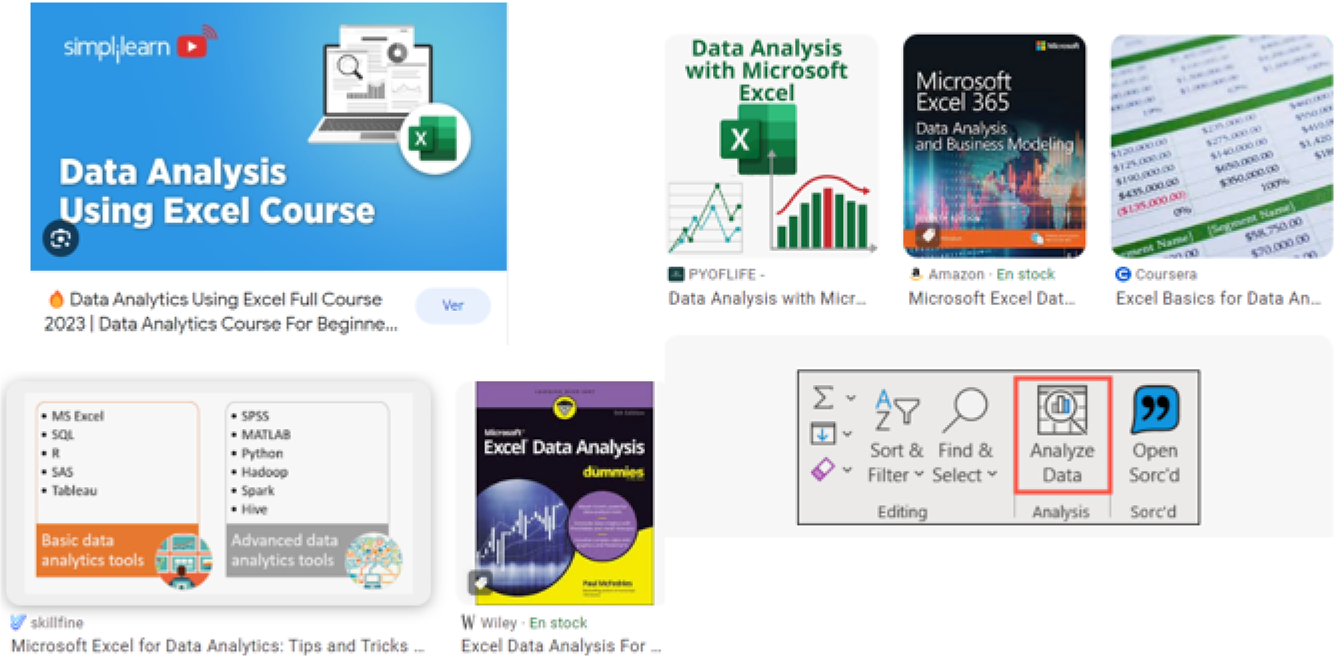


% specific tool usage among researchers that archive/share data

	GitHub	Figshare	Bitbucket	Dataverse	Dryad	Zenodo	Pangaea
Physical Sciences (n=666)	50%	14%	17%	5%	2%	5%	5%
Engineering & Technology (n=1125)	62%	10%	19%	5%	3%	6%	1%
Life Sciences (n=1257)	45%	21%	9%	5%	16%	4%	2%
Medicine (n=632)	28%	16%	7%	12%	4%	3%	3%
Social Sciences & Economics (n=1092)	28%	15%	5%	11%	4%	5%	3%
Arts & Humanities (n=404)	32%	13%	6%	13%	3%	7%	4%
Law (n=40)	13%	5%	5%	13%	0%	5%	3%

Table 1: specific tool usage for sharing data & code across disciplines

Is Excel the option to analyze data?


simplilearn
Data Analysis Using Excel Course
Data Analysis with Microsoft Excel
Microsoft Excel 365 Data Analysis and Business Modeling
MS Excel
SQL
R
SAS
Tableau
Basic data analytics tools
SPSS
MATLAB
Python
Hadoop
Spark
Hive
Advanced data analytics tools
Excel Data Analysis
dummies
Sort & Filter
Find & Select
Analyze Data
Open Sorc'd
Editing
Analysis
Sorc'd
Ver

Data Analytics Using Excel Full Course 2023 | Data Analytics Course For Beginne...

PYOFLIFE -
Data Analysis with Micr...

Amazon · En stock
Microsoft Excel Dat...

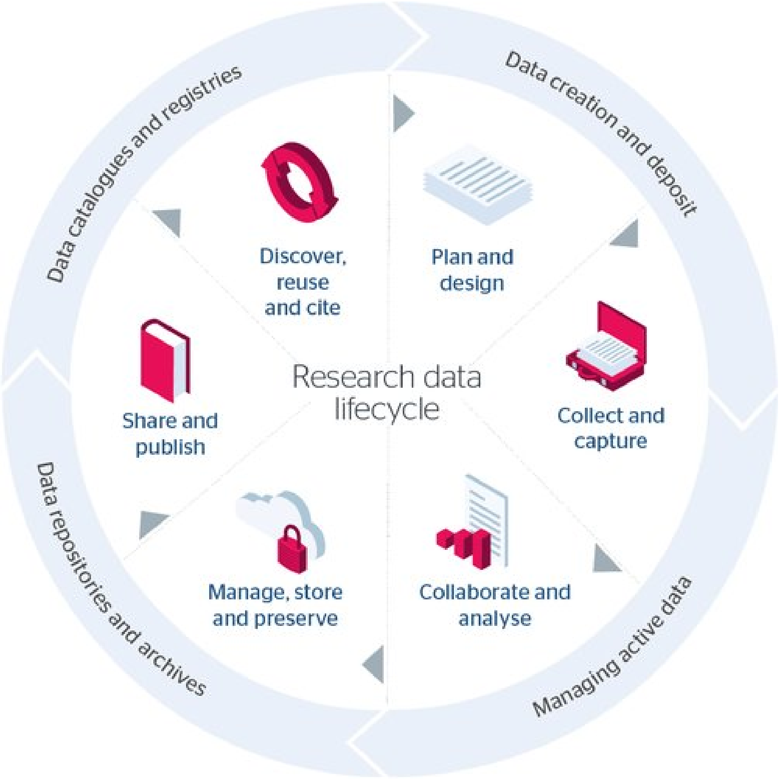
Coursera
Excel Basics for Data An...

skillfine
Microsoft Excel for Data Analytics: Tips and Tricks ...

Wiley · En stock
Excel Data Analysis For ...

Before we start ...

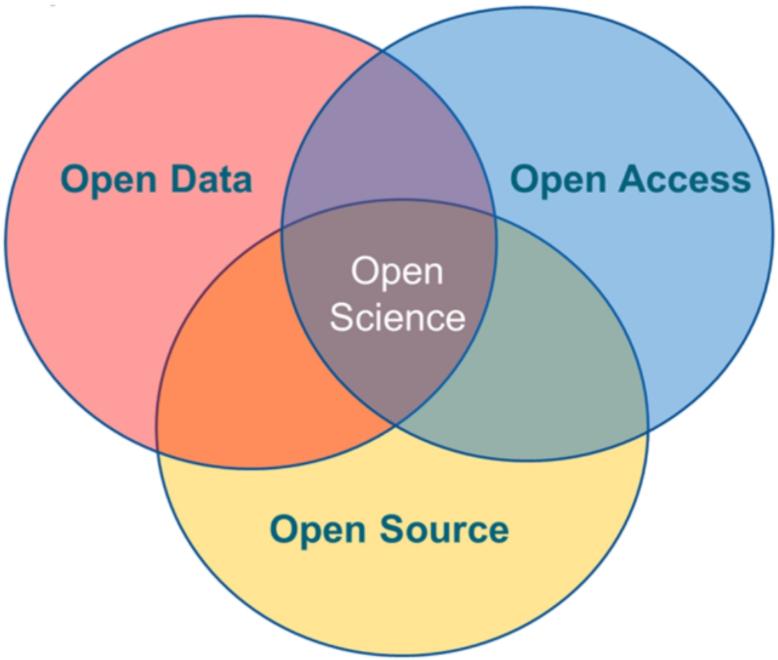
- Data Sharing has become relevant, not only because it is useful and meaningful, but as part of the increasing importance of *Open Science*.
- Concepts such as FAIR, Data Management (Plans), Reproducibility or Repositories, are pervading the (Data) Scientist's vocabulary at different speeds with different degrees of assumption.
- Let's make a quick review

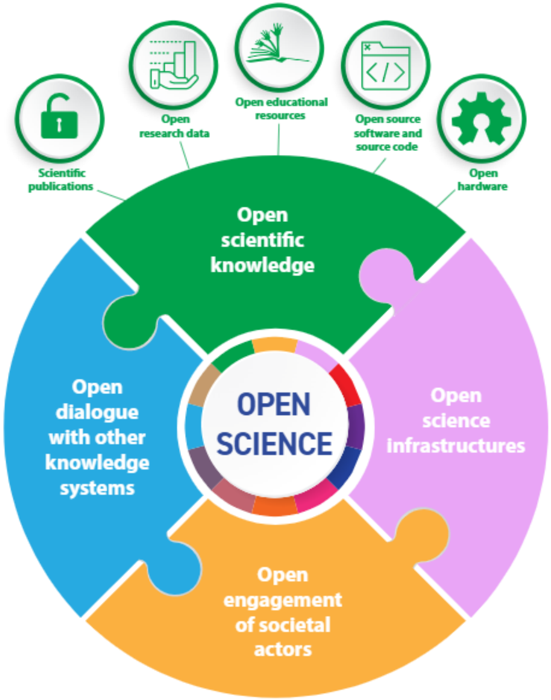


https://beta.jisc.ac.uk/guides/research-data-management-toolkit

Open Science

- Open science refers to a movement and set of practices aimed at making scientific research and its outputs more accessible, transparent, and collaborative.
- It emphasizes the free sharing of research data, methods, and findings with the scientific community and the general public.




Scientific publications
Open research data
Open educational resources
Open source software and source code
Open hardware
Open scientific knowledge
Open dialogue with other knowledge systems
OPEN SCIENCE
Open science infrastructures
Open engagement of societal actors

Research Data management

- *Research Data Management* concerns the organisation of data, from
  - its **entry** to the research cycle
  - through to the **dissemination** and **archiving** of valuable results.
- It aims to ensure reliable **verification of results**, and
- It permits new and innovative research **built on existing information** .

D | C | C

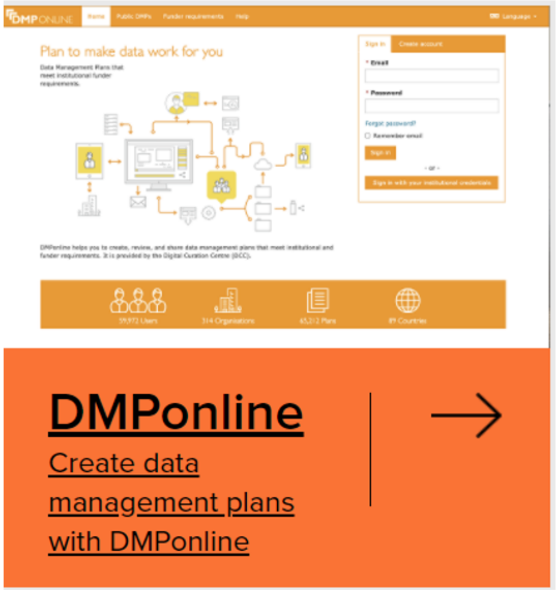
The Digital Curation Center

Because Good Research requires good Data

Data Management Plans

- The planning on how data is going to be collected, processed, analyzed, shared and preserved is stated in the Research's Project Data Management Plan (DMP)

- DMPs have become an important step of any research process: from PhD theses to European Projects, all are required to prepare, follow and provide a Data Management Plan.



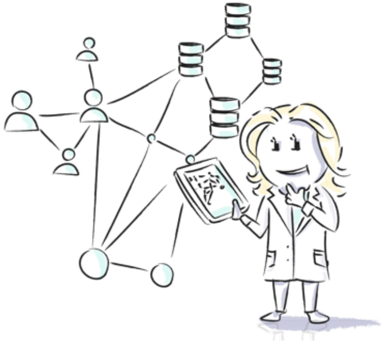
Plan to make data work for you

Data Management Plans that meet institutional funder requirements.

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

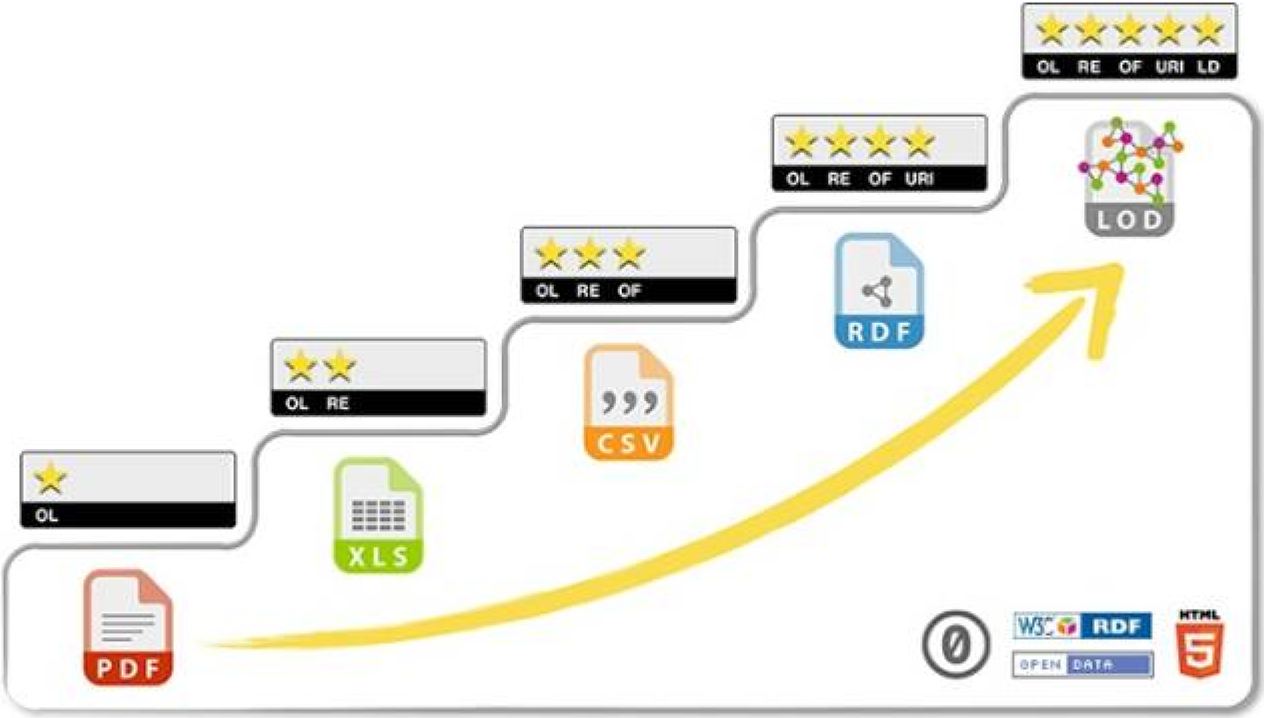
DMPonline

Create data management plans with DMPonline



FAIR Data

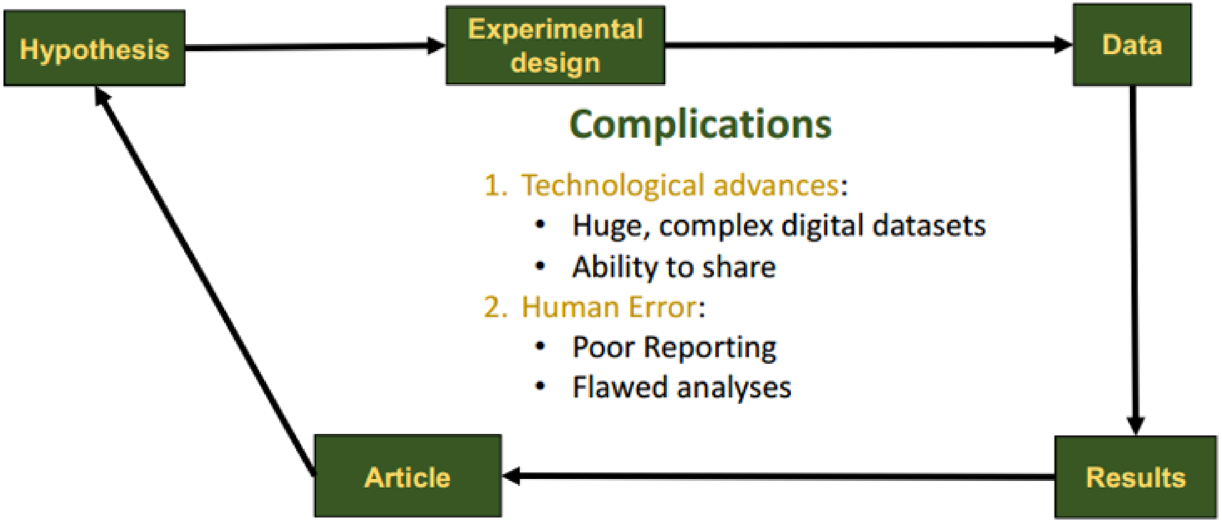
- One of the goals of DMP is to ensure that data generated by the project is **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.
- But FAIR is not an state, it's a degree. do



https://5stardata.info/en/

Reproducibility issues

- The research cycle may flow erratically due to distinct issues.

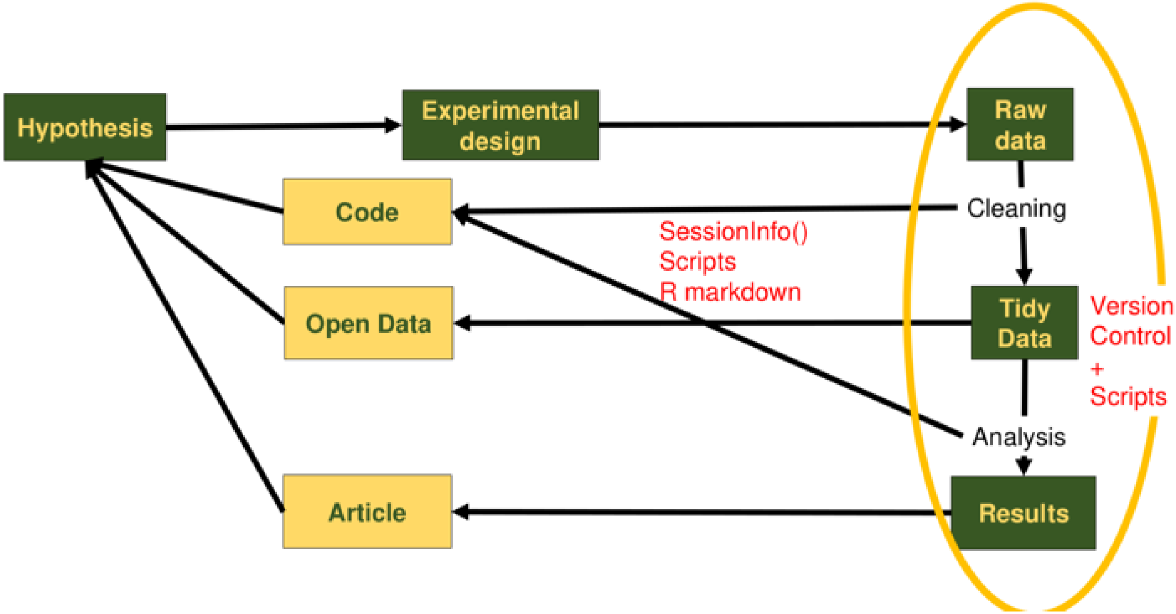


Complications

1. Technological advances:
   - Huge, complex digital datasets
   - Ability to share
2. Human Error:
   - Poor Reporting
   - Flawed analyses

Approaches to enhance reproducibility

- Many issues on reproducibility can be managed using tools such as literate programming and control version.



Increasing research reproducibility

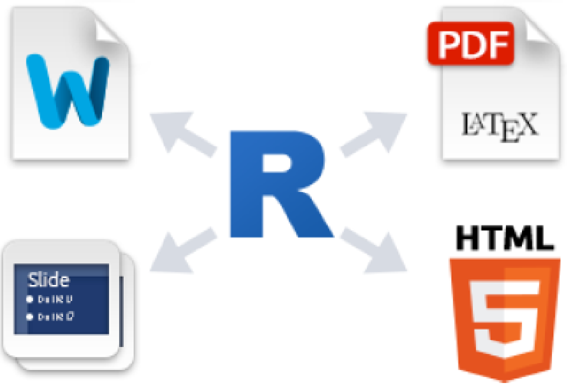
Literate Programming

- (R)markdown in quarto

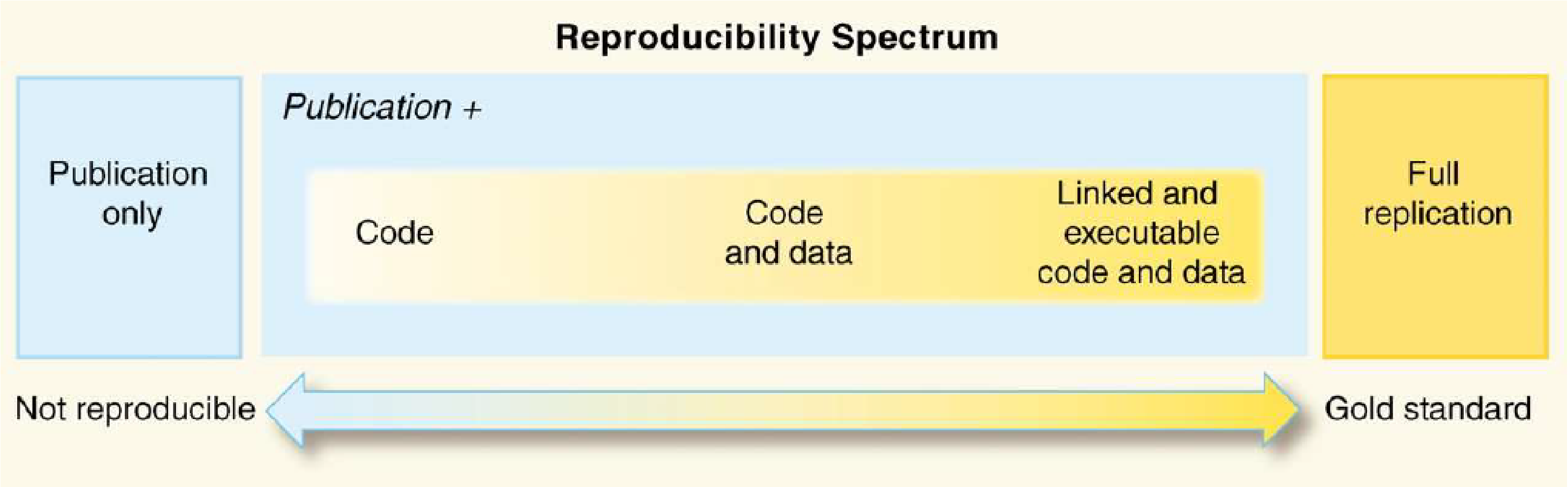
Version Control Systems

- GitHub

Both available within RStudio (posit) with a high level of integration



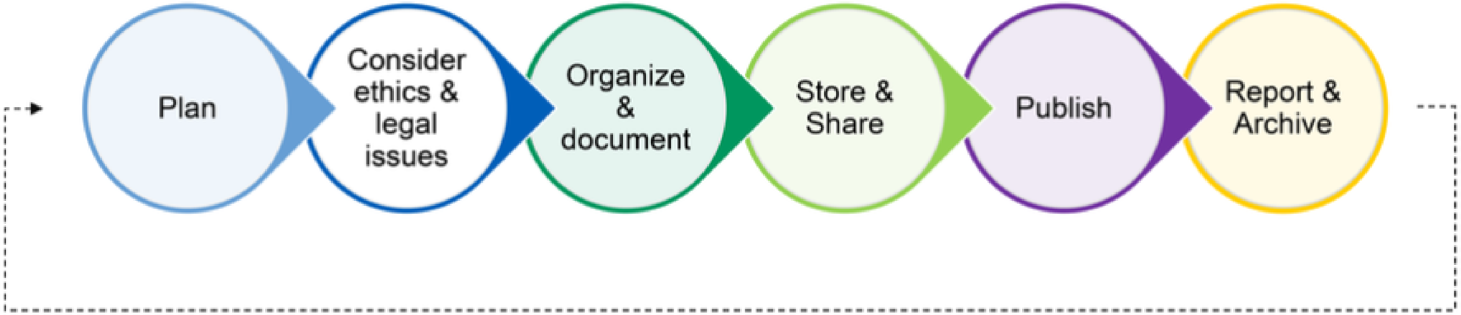
The Reproducibility Spectrum



"Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

What about data?

- We have focused on how to enhance reproducibility,
- But, typically the research cycle focuses more on **data**
- Data ...
  - Has to be acquired.
  - Has to be **processed** and **analyzed**
  - In a reproducible way, which means it has to be **stored**.
  - Has to be **preserved**.
  - Has to be **published / shared** in a FAIR way.



https://www.aalto.fi/en/services/introduction-to-research-data-management

Store/Share/Publish/Preserve Data

Along the life cycle the data undergoes related but distinct processes.

1. **Data Storing**: Securely storing research data in a suitable storage infrastructure or systems.
2. **Data Sharing**: Making research data available to other researchers or interested parties, typically within the research community.
3. **Data Publishing**: Formal process of making research data publicly available beyond the research community, promoting transparency and reproducibility.
4. **Data Preserving**: Long-term retention and maintenance of research data to ensure ongoing accessibility, integrity, and usability for future access and potential reuse.

Where can all this be done?

Along the life cycle the data undergoes related but distinct processes.

1. **Data Storing**: Cloud storage platforms [*Amazon S3, Google Cloud Storage, or Microsoft Azure*] or Institutional data storage infrastructures.

2. **Data Sharing**: Discipline specific rpositories [*GenBank, Dryad*], Institutional infrastructures [*CORA.RDR (CSUC)*], Data Sharing platforms [*Research Gate, Mendeley Data, **GitHub***].

3. **Data Publishing**: Journals, List of Data Journals in Zenodo, Discipline Specific Repositories [*Gene Expression Omnibus (GEO*], General-purpose data repositories [*Dryad, ZENODO*].

4. **Data Preserving**: Trusted digital repositories, Data Archiving services, Institutional Data Repositories.

So what till now?

- Open Science emphasizes that a series of (methodological) procedures are adopted,
- In order to warrant the free sharing of
  - research data,
  - methods, and
  - findings

  with the scientific community and the general public.
- Such procedures include
  - Appropriate Data Management
  - The highest level of reproducibility
  - Aiming at the highest level of Data FAIRness
  - At all levels where data exists:
    - Storage
    - Sharing
    - Publication
    - Preservation

And where to now?

- In order to implement better open science workflows,
- It seems, not only reasonable, but advisable to share data and code together.

- A state of the art code sharing system is the version control system git and the associated web platforms GitHub or Gitlab.

- The question we (still) have to answer is:

  - *how good is github at sharing data*
  - *is it as good as it is for code?*

- The short answer is

  - Github can be used for sharing data in a FAIR way.
  - It has however some limitations (more than with code).

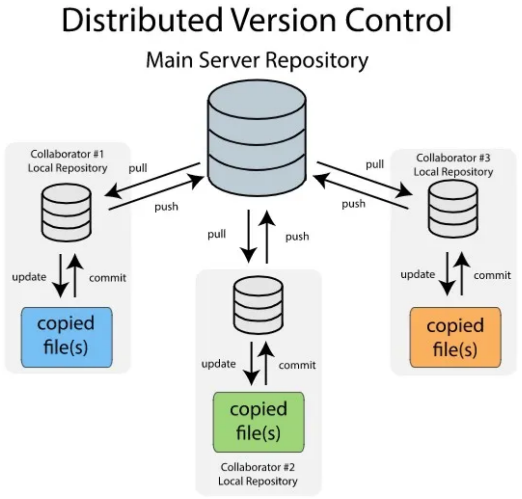
- Let's see how this can be done.

Save and Share Data with Github

Control Version Systems and Git

- A **Control Version System** is a software tool that helps track and manage changes to source code or files,
- It allows multiple people to collaborate on a project,
- It keeps a record of revisions and modifications over time, eventaually allowing to change between versions.

- **Git** is a *distributed CVS* that enables developers to track changes to files and collaborate on projects efficiently.
- It provides features like *branching*, *merging*, and *version history*.
- Installed locally, works as a Command Local Tool (CLT), used from a console.



A Distributed Version Control System. Each collaborator has a local copy of the repository, so no Internet connection is required.

GitHub

- **GitHub** is a service for hosting git repositories in the web.
- It offers all of git's functionality, adding a number of its own features such as *bug tracking*, *task management*, and *per-project wikis*.
- It can be used from the terminal, but also from a Graophical User Interface.
- Full integration with Rstudio (Posit) through Rstudio projects.

GitHub crash course

- In order to start using GitHub there are (only) a few things to learn.

1. How to create a new repository

2. How to synch your work with it

   `pull`, `stage`, `commit`, `push`.

3. Anything else

- We will work in "demo mode" but there are many good tutorials out there:
  - GitHub cheatsheet
  - Using Git and GitHub with RStudio: : CHEATSHEET

Live Demo

- Create an account (and Synch it with Rstudio)
- Create new repository in GitHub
- Clone the repository as a new Rstudio project
- Populate the repository with code and data
- Stage changes: commit, push, iterate

- *Recover past status*

- Random, but important concepts

    - username.github.io, organizations, github pages

- More advanced concepts

    - Branches, Forks
    - Automatization with github actions.

- Document the repository: Wikis, GitHub pages,

Document & Share Data with GitHub

- GitHub offers a straightforward way to share data:

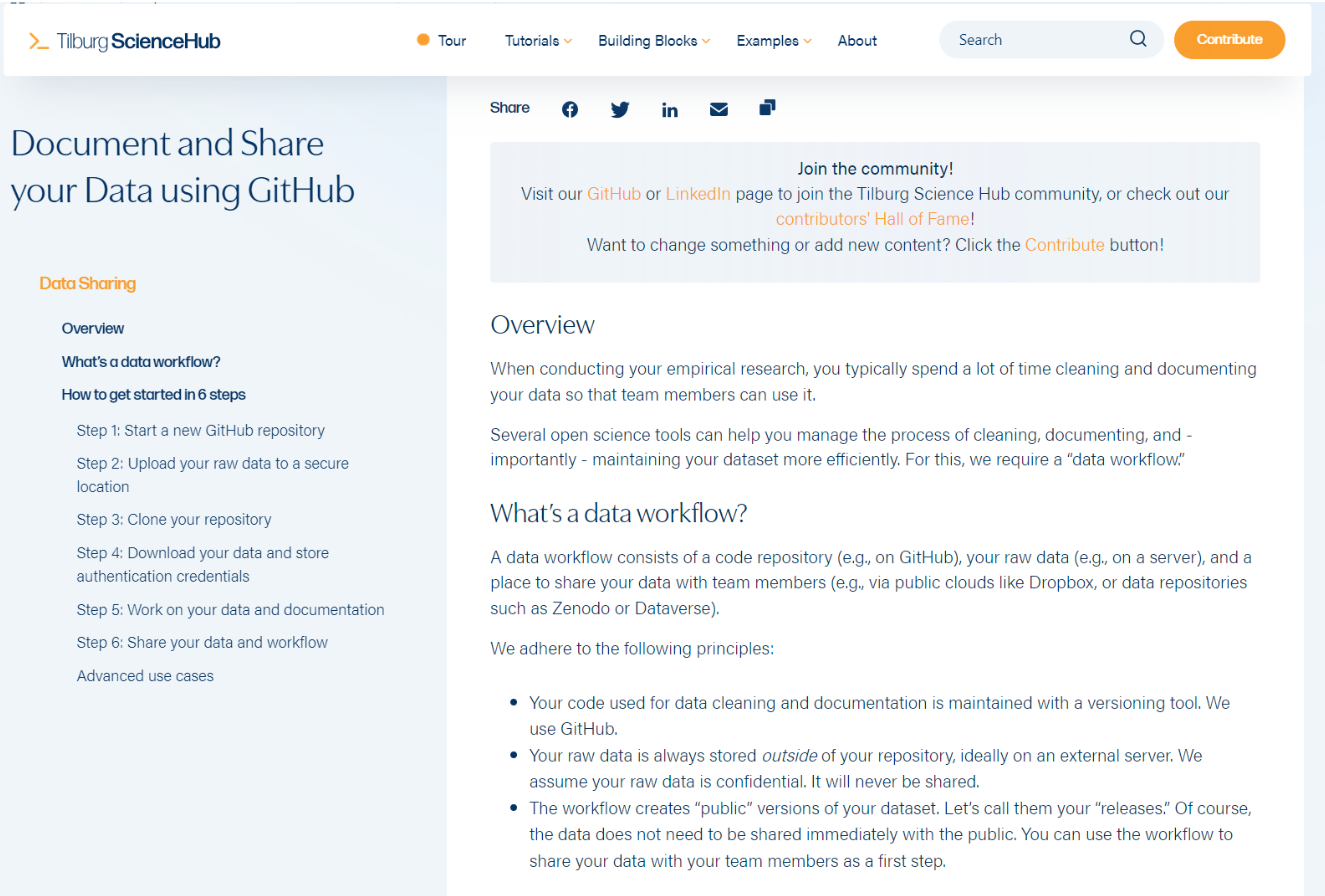
1. Create a repository for your data, Decide its structure but it should probably contain.
    - Data folder
    - Metadata information
    - Maintenance code [Optional: Start with a private repository]
2. Populate the repository with the data, code and documents you intend to share. [Optional but adviced: Document the repository with Markdown (e.g. in README.MD) or HTML (e.g. index.html)]
3. Commit your changes and push. [Iterate this step as needed]

4. Activate githubpages. Check the page visibility and complete the process by making the repository public.

5. Improve your data through *community error checking*:

    - If someone notices an error they can suggest a change with a `pull request`.
    - The owner of the data set can then decide whether or not to accept the change.

Data repositories without data



Document and Share your Data using GitHub

Join the community!
Visit our GitHub or LinkedIn page to join the Tilburg Science Hub community, or check out our contributors' Hall of Fame!
Want to change something or add new content? Click the Contribute button!

Overview

When conducting your empirical research, you typically spend a lot of time cleaning and documenting your data so that team members can use it.

Several open science tools can help you manage the process of cleaning, documenting, and - importantly - maintaining your dataset more efficiently. For this, we require a "data workflow."

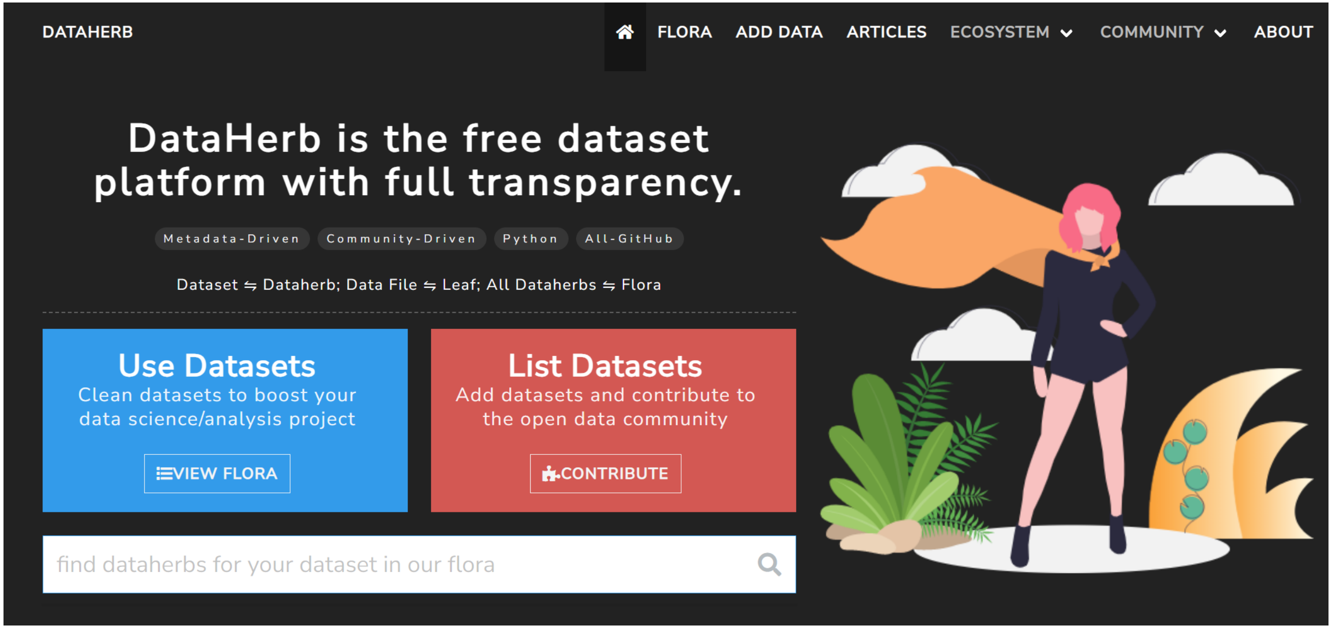
What's a data workflow?

A data workflow consists of a code repository (e.g., on GitHub), your raw data (e.g., on a server), and a place to share your data with team members (e.g., via public clouds like Dropbox, or data repositories such as Zenodo or Dataverse).

We adhere to the following principles:

- Your code used for data cleaning and documentation is maintained with a versioning tool. We use GitHub.
- Your raw data is always stored *outside* of your repository, ideally on an external server. We assume your raw data is confidential. It will never be shared.
- The workflow creates "public" versions of your dataset. Let's call them your "releases." Of course, the data does not need to be shared immediately with the public. You can use the workflow to share your data with your team members as a first step.

Open Data Repositories Projects



DataHerb is the free dataset platform with full transparency.

Dataset ⇋ Dataherb; Data File ⇋ Leaf; All Dataherbs ⇋ Flora

Use Datasets
Clean datasets to boost your data science/analysis project

List Datasets
Add datasets and contribute to the open data community

What are the Pros & Cons, if any?

What are the Pros & Cons, if any?

Pros	Cons
Easy and widely used platform	Limited storage space for large datasets
Version control for data	Limited control over access permissions
Collaboration and contribution	Potential privacy concerns for sensitive data
Issue tracking and project management	Learning curve for beginners
Integration with other tools and services	Reliance on an external platform

Summary

Summary

- Data Sharing is an essential part in Open Science
- This can be done in different ways and using GitHub is one of the options.
- ALthough it has some advantages such as the ease of use, and wide users platform.
- There are concerns about security and peristance, so that, the decision on how to share needs to be given a few thoughts.

Resources and References

Git and GitHub

- Happy Git and GitHub for the useR
- Using Git and GitHub with RStudio
- Git and GitHub
- Introduction to GitHub Actions to R users

Data Sharing with GitHub

- Document and Share your Data using GitHub
- GiHub and more: Sharing Data and Code

- Data on GitHub: The easy way to make your data available

- DAGHUb: Github for Data Science

- Democratic databases: science on GitHub (Nature article)

- Sharing code and data with github
- Streaming Data From APIs To Github Repositories
- Data Sharing, Distribution and Updating Using Social Coding Community Github
- The Use of GitHub, an Open Source Code and Data Sharing Website, at Brooklyn Botanic Garden

Open Science

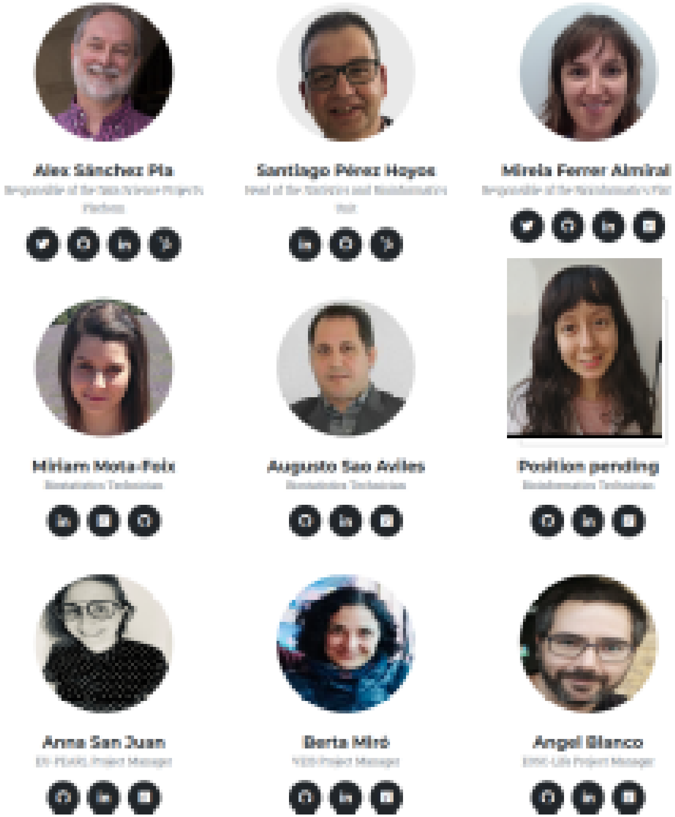
- Open Data: 5 stars deployment scheme
- Open Science Workshops (github).

Research Data Management

- CSUC: Guia per elaborar un pla de gestió de dades per a doctorands
- Mantra, online course on RDM
- Introduction to Research Data Management
- Research data management toolkit
- DCC - Digital Curation Center

Acknowledgements

Acknowledgements



Alex Sánchez Pla
Responsible of the Data Science Projects Platform

Santiago Pérez Hoyos
Head of the Statistics and Bioinformatics Unit

Mireia Ferrer Almirall
Responsible of the Bioinformatics Plat

Miriam Mota-Foix
Biostatistics Technician

Augusto Sao Aviles
Biostatistics Technician

Position pending
Bioinformatics Technician

Anna San Juan
EU-PEARL Project Manager

Berta Miró
[illegible] Project Manager

Angel Blanco
[illegible] Project Manager