



Eötvös Loránd University
Faculty of Education and Psychology

CharmEU open science training

-

Open data and version control

Dr. Tamás Nagy
ELTE Eötvös Loránd University
@nagyt

Introduction



- Tamás Nagy
- Assistant professor at ELTE Eötvös Loránd University, Budapest.
- Original training in Psychology.
- Engaged with data science, and data analysis issues and open science.



01

ONE

Open data, FAIR principles



02

TWO

Best practices for sharing FAIR data



03

THREE

Where and how to share data



04

FOUR

Version control



05

FIVE

Q&A session

Why share data?

- Universism: prove by evidence and not by authority (Merton, 1942).
- This also means that everyone has to be able to verify the evidence, based on data.
- By that time paper was the main information carrier, thus data sharing was cumbersome.
- Later came the computers, digital data, and the internet.
- Science was still stuck in paper-based information sharing until recently.
- Scientists should embrace the possibility to share data (as it was originally intended).
- Sharing data makes questionable research practices detectable, thus improves the reliability of findings.
- Sharing data facilitates good practices in data management.

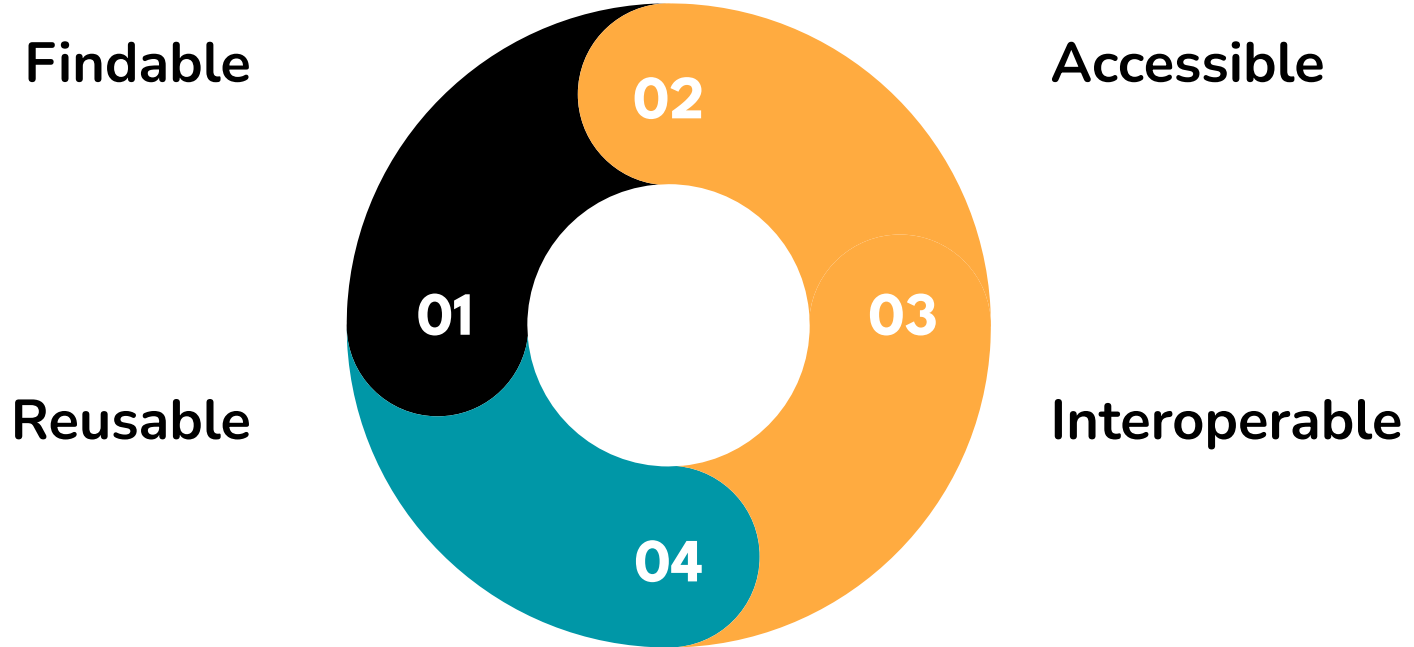
Teaching students to share their data

- Teaching students to share data (and study materials) should be a priority!
- Not only a good practice, but also nurtures a general open science mindset.
- It makes students better collaborators on research projects.
- Teach easy to use tools, so students without special skills can use them.

How to share data?



FAIR data



Findable Data

- Data are findable if it can be easily located and accessed by users or systems.
- This means that it is **indexed** relevant by search engines.
- **Persistent identifiers** (PIDs) — like DOI — are keys for indexing, as it means that independent of the storage space.
- Several sites offer generating a DOI for free (e.g. osf.io, zenodo.org).
- Some journals publish datasets (e.g., Journal of Open Psychology Data).

Accessible Data

- Data should be accessible to **everyone** at any time.
- This means that **no burdens** are posed by the owner (or storage space) to get the data.
- It also means that the storage space is maintained in the foreseeable future.
- Scientific data **repositories** (e.g., osf, zenodo, figshare) make it easy to store and share data, and provide long-term maintenance.

Interoperable Data

- Interoperability means that data is usable **irrespective of the software or operating** system of the user.
- Data should be in an **open format**, i.e., users should be able to use the data with any software (e.g., csv, xml, json, API endpoint, etc.).
- Proprietary software often use their own data format which limits accessibility.
- There are also **metadata standards**, that are ensuring that data is human and machine readable (e.g., Dublin Core).

Reusable Data

- The ability of data to be easily reused for different purposes **beyond the original research context** (e.g., reproduction, meta-analysis, mega-analysis, etc.).
- Data should be well-documented, structured, and prepared in a way that allows **other researchers** to understand, interpret, and effectively utilize the data.
- It is a good practice to make the **language of the dataset** and documentation accessible for those not speaking the local language by providing translations.
- Data licencing.

Data dictionary/ Code book

1. Information about the **variables** in the dataset.
2. Information about the **data processing**.
3. Information about the **data collection procedure**.

This data frame contains 456 observations (rows), each representing a movie, and 27 variables (columns):

1. **title**: Title of movie
2. **audience_score**: Audience score on Rotten Tomatoes (response variable)
3. **type**: Type of movie (Documentary, Feature Film, TV Movie)
4. **genre**: Genre of movie (Action & Adventure, Comedy, Documentary, Drama, Horror, Mystery & Suspense, Other)
5. **runtime**: Runtime of movie (in minutes)
6. **year**: Year the movie is released
7. **mpaa_rating**: MPAA rating of the movie (G, PG, PG-13, R, Unrated)
8. **studio**: Studio that produced the movie

Best practices for sharing research data

- Digitally shareable data, accessible to everyone.
- In a repository with a **permanent identifier** and **time stamp** (e.g., OSF).
- Includes a **codebook** (or data dictionary) to help other researchers understand the structure of the data and the content of the variables.
- A **license** that specifies that others can copy and distribute the data.
- + Compliance with data management principles and rules (anonymisation, transparency to participants).
- Open Data is linked in the published journal article.



What to share?

1. Raw data
 - Data generated at the **time of recording**, unchanged, without modification, deletion or aggregation.
2. And/or processed data
 - Square format: each **observation is one row**, each **variable one column**.
 - Each observation unit is assigned a **unique identifier**.
3. Code book on variables
 - Information on **each variable** (including unit of measurement).
 - Information on **how the variables are calculated** (e.g. aggregates).
 - Information on the design of the study and **how the data were collected**.
4. Precise description of how the data were **transferred** from the raw form to the processed form (1->2), if possible with program code.

Example sentences for consent forms for data sharing

Type of sharing	Example sentence
Data sharing part in the IRB submission and consent form	<i>The data collected during the research will be published in a non-personally identifiable, anonymised form at [repository name].</i>
Research data and materials are available on a public repository.	<i>All data and materials have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].</i>
Anonymised research data and materials are available on public repositories.	<i>Anonymized data and materials have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].</i>
The data of the analysis are available on a public repository.	<i>The data used for the analyses have been made publicly available at the [repository name] and can be accessed at [persistent URL or DOI].</i>

Recap

- Findible
- Accessible
- Interoperable
- Reusable

Where and how to share data



Sharing data on OSF (Open Science Foundation)

- URL: osf.io
- Stored by the Center for Open Science (COS).
- Can select data center in Europe (or other region).
- Has insurance to keep the data online for 50 years.
- Storage limit per project: 5 Gb



Sharing data on zenodo

- URL: zenodo.org
- Hosted and maintained by CERN
- Each upload has a 50 Gb / 100 files limit
- Publication related data and materials
- Built in version control
- Integrations with github



Sharing data on figshare

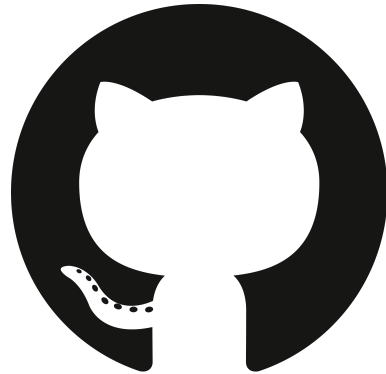
- URL: figshare.com
- Owned ultimately by Holtzbrinck Publishing Group.
- Each upload has a 20 Gb limit.
- Publication related data and materials.
- Built in version control.
- Integrations with Github.
- Has a paid version (figshare+ with options to share really large data)



Some easy to use research data repositories

Repository	Provider / Funder	Limits	Best for
osf.io	COS	5 Gb	Preregistered research projects
zenodo.org	CERN	50 Gb / 100 files	Publication related data and materials
figshare.com	Holtzbrinck Publishing Group	20 Gb	Publication related data and materials
github.com	Github / Microsoft	2 Gb	Storing and version controlling code

Version control



What is this and why is it important?

- Git and github is a version control system that helps you preserve and share different versions of our code and work together effectively.
- Versioning is essential in programming and data science projects, so it is imperative to learn how to use it.

Motivation



- Bad version control practices are common in academia:
 - Manuscripts
 - Data
 - Analysis code
- Manually version controlling files can be time consuming and tedious (e.g., keeping in mind who works on which version of the processed data).
- Small misunderstandings can lead to considerable setbacks.

cikk_20190625_rev_rev2_final_2.docx

Easiest good version control practices in the academy

Research materials: OSF, zenodo, figshare

Data: OSF, GitHub, Onedrive

Manuscript: Google drive (.docx), Office365

Analysis code: GitHub (GitLab, etc.)



Google Drive



Analysis code version control

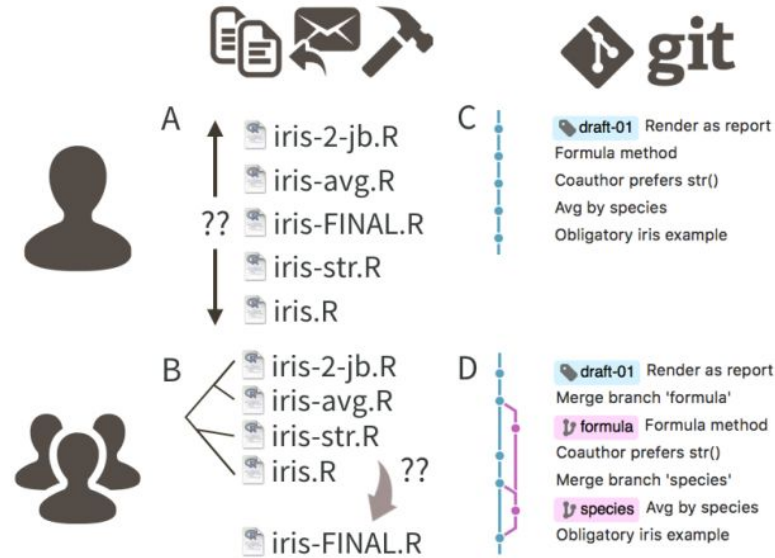


Figure 1: A: Solo work with DIY version control via filename. B: Collaborative work with DIY version control. C: Solo work with Git. D: Collaborative work with Git.

Installing git

<https://git-scm.com/downloads>

What is git?



- A local version control system
- All previous versions are available, but only the latest version is visible -> transparent
- It is easy to see what changes have taken place between the versions
- Optimized for program code or any text file
- Open source system, etc.

What is GitHub?



- Optimized for cloud-based code sharing
- R's infrastructure is heavily building on it
- Profit-oriented company, current owner is Microsoft
- Public projects are free + you can have an academic license
- There are several alternatives

Advantages of Gitgub



- Storage in the cloud
- Always available
- Transparent version structure
- RStudio integration



- Easy to share
- Report bugs / demands
- External collaborators / assistance
- Community
- Program code maintenance (eg automatic testing, etc.)





Eötvös Loránd University
Faculty of Education and Psychology

Git terminology

Using Git and GitHub

- Git is a system used primarily from the command line.
- There are several GUIs (Graphical User Interfaces; e.g., GitHub Desktop) that make most features accessible.
- GitHub is easy to use from a browser.

3-step saving process



1. **Save** the file to your computer.



2. **Commit** (== named version) tells git that this is a new important version of the file. (stage: select file for commit, diff: line-by-line differences).



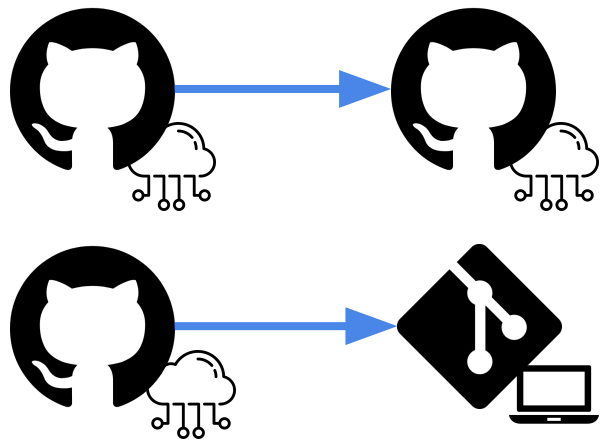
3. **Push** into the cloud (GitHub) puts the file.

Some terminologies are straightforward, some less so

- **Repository (repo)** == project.
 - Can be public (default) or private
- **Branch:** alternative realities of a repo (e.g. developer and public version).
- **Main:** main branch.
- **Head:** Which branch are we on?
- **Checkout:** Switch to another branch



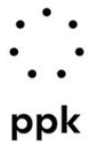
Connecting to a project



1. **Fork:** Copying another repo to yourself (keeping in touch with the original). Use it if you want to contribute to a project.
2. **Clone:** Copy a repo to your own machine (can be your own or someone else's). Use it if you want to have that project locally.
3. **Pull:** Retrieving changes within a repo (e.g. what someone else did).

Interactions between commits

- **Merge:** merge a branch into another (e.g. dev to main)
- **Pull request** to offer changes to your own branch / fork to another branch / fork
 - This has yet to be accepted by the owner, hence the name (but why not push request?). This is can be done multiple ways, e.g. through the [GitHub website](#)
- **Merge conflict** If one branchen in a file on the same line is different. If GitHub cannot resolve it automatically you will need to e.g. through the website ([resolve](#))



Eötvös Loránd University
Faculty of Education and Psychology

Practice

Downloading git

- Step 0: Install Git
<https://git-scm.com/downloads>
- Enable version control in RStudio
- GitHub registration on homepage
www.github.com
- Download Github desktop:
<https://desktop.github.com/download/>
- Log in to Github dektop

Practice

1. Create a new repo
2. Create and save a script
3. Stage and commit
4. Push changes to GitHub
5. Pull
6. Create a new file on a new branch
7. Pull request
8. Merge

Extra feature for researchers: OSF integration

The screenshot shows the OSF project page for "Introduction to R and the tidyverse - 2022-03". The page includes a navigation bar with "OSFHOME" and user "Tamas Nagy". The project title is "Introduction to R and the tidyverse - 2022-03" with a size of 10.8MB and status "Public".

Contributors: Tamas Nagy
Date created: 2022-03-03 12:06 PM | **Last Updated:** 2022-03-12 10:12 PM
Create DOI
Category: Project
Description: Add a brief description to your project
License: CC-BY Attribution 4.0 International

Wiki: Add important information, links, or images here to describe your project.

Files: Click on a storage provider or drag and drop to upload. A table of files is shown, with a red box highlighting the first three rows:

Name	Modified
Introduction to R and the tidyverse - 2022-03	
- GitHub: nthun/intro-to-R-2022-03 (master)	
.gitignore	

Other files listed include "intro-to-R-2022-03.Rproj", "Google Drive: slides", "Data science workflow: version control with git.slides" (2022-03-13 01:54 AM), "Introduction to R and the tidyverse.gslides" (2022-03-12 10:13 PM), "OSF Storage (Germany - Frankfurt)", and "cheat sheets".

Citation: (Empty field)

Components: Add components to organize your project. (Buttons: Add Component, Link Projects)

Tags: data analysis, data science, introduction, programming, R, reporting, visualization, workshop, wrangling, Add a tag

Recent Activity:

- nthun added file intro-to-R-2022-03.Rproj to GitHub repo nthun/intro-to-R-2022-03 in Introduction to R and the tidyverse - 2022-03 (2022-03-12 10:12 PM)
- nthun added file .gitignore to GitHub repo nthun/intro-to-R-2022-03 in Introduction to R and the tidyverse - 2022-03 (2022-03-12 10:12 PM)
- Tamas Nagy added tag data science to Introduction to R and the tidyverse - 2022-03

Academic licence for github

University researchers and doctoral student can get free premium features by applying for an academic license.

<https://docs.github.com/en/education/about-github-education/github-education-for-teachers/apply-to-github-education-as-a-teacher>



Eötvös Loránd University
Faculty of Education and Psychology

Thank you for your attention

Dr. Tamás Nagy
ELTE Eötvös Loránd University
@nagyt