Organizing your data and software with a reproducible project workflow

DCC Spring Training Days 2024 April 25th 2024

Renate Mattiszik, Saxion UAS Meron Vermaas, VU Stephanie van de Sandt, VU

What does your living room look like?



Forgemind ArchiMedia / CC BY 2.0 / Flickr



https://www.wallpaperflare.com/garbage-inside-a-room-abandoned-abando ned-building-decay-dirty-wallpaper-arqkg

How is your Desktop organized?



Anastasia Shuraeva: A White Wooden Closet of a Woman. Pexels



Ron Lach: Photo of an Untidy and Messy White Wooden Closet. Pexels

How many research data / software files do you have?



This file is made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication



This image was originally posted to Flickr by pasukaru76 at https://www.flickr.com/photos/38451115@N04/9824401426



Do you know where the data from your first work is located right now?



Tuxyso / Wikimedia Commons / CC-BY-SA-4.0



This work is in the public domain.

Planning our Workshop

https://edu.nl/h89dk

How we will run the workshop today with you



PRESENT

Organizing and structuring a project

1. Save your project in a single folder.



- description
- manuscript



Credit: Scriberia for The Turing Way

KÖMPENDIUM (3)? **2**. 1. ŵ 14x 1x 4x **1**x 3. 4. 5. (...) 880

Credit: Karthik Ram



2. Devise a logical system of sub-folders.







Automatically generated Human written Read-only





Automatically generated Human written Read-only





Automatically generated Human written Read-only





3. Introduce the project (or sub-folder) in a README.

What is the project about? • Who are involved? • How was the data collected?

• When & where is the research conducted?

Other inspiration...

- Abbreviations
- Folder structure
- File explanations
- References & links

4. Use interoperable file types.

191-	Å	Cut	Cal	bri (Body)		-]	12	v	A'	Α'	-	
Paste	10	Format	в	1	U + 1	Ξ		0	*	A	*	*	-
810	÷	× v	ſs										
		A			В					С			

1	Letter	Number	
2	A	124	
3	В	125	
4	С	126	
5	A	127	
6	В	128	
7	С	129	
8	A	130	
9	В	131	
10	С	132	
11	A	133	
12			

Bestandsnaam:	datafile_1.xlsx
Opslaan als:	CSV UTF-8 (door komma's gescheiden) (*.csv)

nagement	dataset.csv - Edited
Letter;Number	
A;124	
B;125	
C;126	
A;127	
B;128	
C;129	
A;130	
B;131	
C;132	
A;133	
D;124	
E;4005	



5. Use descriptive and logical file names.

File names should be

- Machine readable
- Human readable
- Logically sortable

Credit: Jenny Bryan

File names should be

Machine readable

Human readable

Logically sortable

Credit: Jenny Bryan

Avoid spaces, accents, odd punctuation...

Use delimiters (hyphens, underscores) for easy separation of elements Use informative elements so that files can be easily found



Credit: Allison Horst

File names should be

- Machine readable
- Human readable
- Logically sortable

- 01_marsh
- 01_marsh
- 02_pre-d
- 02_pre-d
- 03_dea-w
- 03_dea-w
- 04_explo
- 04_explo
- 90_limma
- 90_limma
- Malastila
- Makefile
- figure
- helper01
- helper02
- helper03
- helper04
- tmp.txt

Which set of file(name)s do you want at 3 AM before a deadline?

Use informative elements

and shake and	
nal-data.md	01.md
nal-data.r	01.r
dea-filtering.md	02.md
dea-filtering.r	02.r
vith-limma-voom.md	03.md
vith-limma-voom.r	03.r
ore-dea-results.md	04.md
ore-dea-results.r	04.r
a-model-term-name-fiasco.md	90.md
a-model-term-name-fiasco.r	90.r
9	Makefile
	figure
l_load-counts.r	helper01
2_load-exp-des.r	helper02
<pre>3_load-focus-statinf.r</pre>	helper03
<pre>4_extract-and-tidy.r</pre>	helper04
	tmp.txt



File names should be

- Machine readable
- Human readable
- Logically sortable



Name LEICA_Dmelanogaster_dpp-RNAi_20200822_03.zip LEICA_Dmelanogaster_dpp-RNAi_20200822_02.zip LEICA_Dmelanogaster_dpp-RNAi_20200822_01.zip LEICA_Dmelanogaster_dpp-RNAi_20200815_04.zip LEICA_Dmelanogaster_dpp-RNAi_20200815_03.zip LEICA_Dmelanogaster_dpp-RNAi_20200815_02.zip LEICA_Dmelanogaster_dpp-RNAi_20200815_01.zip LEICA_Dmelanogaster_ctrl_20200823_02.zip LEICA_Dmelanogaster_ctrl_20200823_01.zip LEICA_Dmelanogaster_ctrl_20200815_03.zip LEICA_Dmelanogaster_ctrl_20200815_02.zip LEICA_Dmelanogaster_ctrl_20200815_01.zip



6. Make your data tidy.

TIDY DATA is a standard way of mapping the meaning of a dataset to its structure.

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Visualization credit: Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

-HADLEY WICKHAM



sensor_id	Temp day 1	Sun day 1	Temp day 2	Sun day 2
NY_01	26	189	25	594
Amst_01	12	2,8	16	0,7
NY_02	30	254	15	



Sensor_id	Temp day 1	Sun day 1	Temp day 2	Sun day 2
NY_01	26	189	25	594
Amst_01	12	2,8	16	0,7
NY_02	30	254	15	



An example of untidy data: Weather measurements

Sensor_id	Temp day 1	Sun day 1	Temp day 2	Sun day 2
NY_01	26	189	25	594
Amst_01	12	2,8	16	0,7
NY_02	30	254	15	Ę

Values in column names



An example of untidy data: Weather measurements



Sensor_idTemp day 1Sun dNY_0126Amst_0112NY_0230

tiple observations in a single row						
day 1	Ten	np day 2	Sun day 2			
189		25	59	94		
2,8		16	0	,7		
254		15		5		



Sensor_id	Element	Day	Measurements
Amst_01	Temp	1	
Amst_01	Sun	1	-
Amst_01	Temp	2	
Amst_01	Sun	2	
NY_01	Temp	1	
NY_01	Sun	1	-
NY_01	Temp	2	
NY_01	Sun	2	Ļ
NY_02	Temp	1	
NY_02	Sun	1	
NY_02	Temp	2	
NY_02	Sun	2	

Readme: Temp = Celsius Sun = minutes of sunshine/day



sensor_id	Day	Temp	Sun
Amst_01		1 12	16
Amst_01		2 16	4
NY_01		1 -3	18
NY_01		2	59
NY_02		1	25
NY_02		2 15	30

Readme: Temp = Celsius Sun = minutes of sunshine/day



Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length Leaf wid day 3 day 3
A1_14	control	120	21	. 122	2 23	124
A1_18	control	132	23	135	5 25	138
A1 21	control	121	10	122	20	125
AT_ZT	Control	1.21	TO	D 133	20	122
A2_09	UV	109	29) 114	۶ 1	. 115
A3_02	UV	125	25	5 127	, 27	129
A2 10		120	12	122	2 1/	136
<u></u>	0.	130	12	. 100	, тч	130





An example of untidy data: plant measurements					Values in column names		
Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length Leaf wie day 3 day 3	
A1_14	control	120	21	122	2 23	124	
A1_18	control	132	23	135	5 25	138	
A1_21	control	131	18	133	3 20	135	
A2_09	UV	109	29	114	۶ 1	. 115	
A3 02	UV	125	25	127	, 27	129	
	UV	130	12	133	3 14	. 136	




An exampl	e of untidy data:	plant measuremer		Multiple observations in a single				
Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length Leaf wid day 3 day 3		
A1_14	control	120	21	122	23	124		
A1_18	control	132	23	135	25	138		
A1 21	control	131	18	133	20	135		
A2 09	UV	109	29	114	31	115		
Δ3 02	UV	125	25	127	27	129		
A3_10	UV	123	12	133	14	136		









Same data, now in a tidy format

Plant_no	Treatment	Element	Day	Measurement		
A1_14	control	Stem length	1	120		
A1_14	control	Leaf width	1	21		
A1_14	control	Stem length	2	122		
41_14	control	Leaf width	2	23		
41_14	control	Stem length	3	124	+ 23 more r	
41_14	control	Leaf width	3	25		
41_18	control	Stem length	1	132		
41_18	control	Leaf width	1	23		
41_18	control	Stem length	2	135		
41_18	control	Leaf width	2	25		
41_18	control	Stem length	3	138		
41_18	control	Leaf width	3	27	1 00 mara m	
A1 21	control	Stom longth	1	121	+ 23 1101910	

The standard structure of tidy data means that "tidy datasets are all alike..."

"...but every messy dataset is messy in its own way." -HADLEY WICKHAM

Visualization credit: Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst







Save your project in a single folder 1. 2. Devise a logical system of sub-folders 3. Introduce the project in a README 4. Use interoperable file types 5. Use descriptive and logical file names Make your data tidy 6.

Exercise Present:

https://dcc-training-lab.github.io/project-mana gement/lessons/present.html#exercise





PAST

TATELLE PRESERVED AND ADDRESS OF ADDRESS OF ADDRESS ADDRES

Recording project history



Organise versions

"notFinal.doc" by Jorge Cham, https://www.phdcomics.com

"FINAL".doc













FINAL_rev.6.COMMENTS.doc

FINAL_rev.8.comments5. CORRECTIONS.doc

Track changes





FINAL_rev.18.comments7. FINAL_rev.22.comments49. corrections9.MORE.30.doc corrections.10.#@\$%WHYDID ICOMETOGRADSCHOOL????.doc

WWW. PHDCOMICS. COM



Revert Time



Jorge Cham, https://www.phdcomics.com

WWW. PHDCOMICS. COM

Painless less painful collaboration

"notFinal.doc" by Jorge Cham, https://www.phdcomics.com







Git vs GitHub



1. It is a software

2. It is installed locally on the system

3. It is a command line tool

4. It is a tool to manage different versions of edits, made to files in a git repository

It provides functionalities like Version Contr System Source Code Management

https://andersenlab.org/dry-guide/2022-03-09/img/git_v_github.png

	GitHub
	1. It is a service
	2. It is hosted on Web
	3. It provides a graphical interface
	4. It is a space to upload a copy of the Git repository
rol	5. It provides functionalities of Git like VCS, Source Code Management as well as adding few of its own features



- 1. It is a software
- 2. It is installed locally on the system
- 3. It is a command line tool

4. It is a tool to manage different versions of edits, made to files in a git repository

5. It provides functionalities like Version Control System Source Code Management

But what is git exactly?!

Imagine Git like a Photographer



That allows you to capture the moment



4

Staging in Git



By default captures everything in the git project.

Git add specific files



Move objects you want to capture into the focus

Adding Moments to an Album

"Family Christmas 2023″





Tracking Changes



Family Christmas 2023



Family Christmas 2023+10 seconds



Restoring older versions



Family Christmas 2023

Family Christmas 2023+10 seconds



Word Documents and Git

- Git works great for code because it compares text;
- Also works with .csv data
- Can also be used for markdown text files (e.g. LaTeX);
- Doesn't work great for Microsoft Word/Excel documents



VIE	w	PICTURE	TOOLS MAT DI	TABLE TOOL	S YOUT				Kul	? kdawa	📧 🗕 🗗 ala, Imran F. 🔹	×
¶ •	Aa T	BbCcDd Normal	AaBb Heading 1	AaBbCcD Heading 2	• AaBbCcE ¶ List Bullet	AaBbCo Quote	AaBbCcl Subtitle	AAE Title	AaBbCcDdEe ¶ Caption	-	H Find → Cac Replace	
5		Styles							Г <u>а</u>	Editing	^	

This is the place for your 'elevator pitch.' If you only had a few seconds to pitch your products or services to someone, what





- 1. It is a service
- 2. It is hosted on Web
- 3. It provides a graphical interface
- 4. It is a space to upload a copy of the **Git** repository

5. It provides functionalities of Git like VCS, Source Code Management as well as adding few of its own features

But what is GitHub exactly?!

GitHub - Why?



Jorge Cham, https://www.phdcomics.com

WWW. PHDCOMICS. COM

Collaborative Aspect



Cloud Aspect





PAST

TATELLE PRESERVED AND ADDRESS OF ADDRESS OF ADDRESS ADDRES

Recording project history





*) Thanks to: A , but not A

time goes on...







1. Keep a changelog

Changelog All notable changes to this project will be documented in this file.

The format is based on [Keep a Changelog](https://keepachangelog.com/en/1.0.0/ and this project adheres to [Semantic Versioning](https://semver.org/spec/v2.0

[Unreleased]

[1.0.0] - 2017-06-20

Added

- Version navigation.
- Links to latest released version in previous versions.
- "Why keep a changelog?" section.
- "Who needs a changelog?" section.
- "How do I make a changelog?" section.

New visual identity by [@tylerfortune8](https://github.com/tylerfortune8).





Mow do I make a good changelog?

Credit: keepachangelog.com



Guiding Principles

- Changelogs are for humans, not machines.
- There should be an entry for every single version.
- The same types of changes should be grouped.
- Versions and sections should be linkable.
- The latest version comes first.
- The release date of each version is displayed.
 - Mention whether you follow Semantic Versioning.

Types of changes

- Added for new features.
- Changed for changes in existing functionality.
- Deprecated for soon-to-be removed features.
- Removed for now removed features.
- Fixed for any bug fixes.
- Security in case of vulnerabilities.

2. Formalize updates using semantic versioning

Semantic versioning



semver.org/spec/v2.0.0.html





semver.org/spec/v2.0.0.html



1.2.0

Your first version



semver.org/spec/v2.0.0.html



Your first version



semver.org/spec/v2.0.0.html


Changelog All notable changes to this project will be documented in this file.

The format is based on [Keep a Changelog](https://keepachangelog.com/en/1.0.0/ and this project adheres to [Semantic Versioning](https://semver.org/spec/v2.6

[Unreleased]

[1.0.0] - 2017-06-20

Added

New visual identity by [@tylerfortune8](https://github.com/tylerfortune8).

- Version navigation.

- Links to latest released version in previous versions.

"Why keep a changelog?" section.

- "Who needs a changelog?" section.

"How do I make a changelog?" section.





"FINAL".doc













FINAL_rev.6.COMMENTS.doc

FINAL_rev.8.comments5. CORRECTIONS.doc









FINAL_rev.18.comments7. FINAL_rev.22.comments49. corrections9.MORE.30.doc corrections.10.#@\$%WHYDID ICOMETOGRADSCHOOL????.doc

WWW. PHDCOMICS. COM

Exercise Past

https://dcc-training-lab.github.io/projectmanagement/lessons/past.html



Publishing and archiving your project for future use

FUTURE

Step 1: control how your work will be used by others (and yourself!)

Why should I choose a license?







Your project ...

Open license



Restrictive license







Sensitive information?!



The Turing Way







Example for Permissive Licenses for Data



ttribution

Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you

No Derivative Works

Others can only copy, distribute, display or perform verbatim copies of your work

Share Alike

Others can distribute your work only under a license identical to the one you have chosen for your work

Non-Commercial

Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.

Restricted vs Permissive Software











How to find the right license?

Ask a data steward or privacy champion if in doubt.

A license finder may help you to get a good idea:

- https://choosealicense.com
- https://data.europa.eu/elearning/en/module4/#/id/co-01



2. Maintain your living project on GitHub.

A pull request

Update session-05.mc	d #6	
I'l Open XinyanFan-hub wants to merge 1	commit into DCC-training-lab:main from XinyanFan-	hub:main 🖸
Q Conversation 0 -O- Commits 1	Checks Files changed Files changed Files chan	
XinyanFan-hub commented on Apr 2	20	First-time contributor
update the instructor		
-o- 😭 Update session-05.md …		Verified c075715
Add more commits by pushing to the ma	ain branch on XinyanFan-hub/2023-spring-training-da	ays.
This branch has not been of No deployments	deployed	

Inviting collaborators in 'settings' > 'collaborators'

诊 General	Who has access	
Access A: Collaborators D: Moderation options	PUBLIC REPOSITORYImageDIRECT ACCESSRThis repository is public and visible to anyone.3 have access to this repository. 2 collaborators. 1 invitation.Manage	
	Manage access	Add people
 Actions Webhooks Environments Codespaces 	Select all Q Find a collaborator	Type -
 Pages Security O Code security and analysis 	 meronvermaas Collaborator Max Paulus paulmaxus • Collaborator 	Remove
Deploy keys		

Showcasing



gs	\$
onvermaas:main.	b.github.io/project-ma
d	
Update future.md add pptx presentation used in movie	
project archive transcript first draftCreate .gitignorefix typoUpdate README.md	ring-Training Latest
Update preparation.md	d
data and software v	Active
data and software v	Active

3. Archive the project for posterity on Zenodo.

Why archiving?



A release in GitHub

Releases Tags	Draft a new release	e Q Find a release
o days ago	2023-DCC-Spring-Training-Days	Ø Ü
> vansteph > 2023 ≻ c19735b ⊘	This is the material as used in our 2023 edition of this DCC spring training	
Compare +	• Assets 2	
	Source code (zip)	5 days ago
	Source code (tar.gz)	5 days ago
pr 17	Zenodo Release of 2022 training material	0 Û
vansteph		
> 2022.1 - 6f511a2 ⊘	This is the material of the DCC spring trainings days in 2022	
Compare 👻		
	► Assets 2	

... can be automatically archived in Zenodo!

Upload

Q



Featured communities



Transform to Open Science

Transform to OPen Science (TOPS) is a Source Science initiative. Within the Te initiative to spark change and inspire o

Curated by: nasatransformtoopen

Recent uploads

June 20, 2023 (v14) Dataset Open Access

Binary black-hole surrogate waveform catalog

Scott E. Field; Chad R. Galley; Jan S. Hesthaven; Jason Kaye; Manuel Tiglio; Jonathan Blackman; Béla Szilágyi; Mark A. Scheel; Daniel A. Hemberger; Patricia Schmidt; Rory Smith; Christian D. Ott; Michael Boyle; Lawrence E. Kidder; Harald P. Pfeiffer; Vijay Varma

This repository contains all publicly available numerical relativity surrogate data for waveforms produced by the Spectral Einstein Code. The base method for building surrogate models can be found in Field et al., PRX 4, 031006 (2014). Several numerical relativity surrogate models are currently...

Uploaded on June 20, 2023 13 more version(s) exist for this record

Communities		➔ Log in	🕼 Sign up
		Need help uploa	ading? Contact us
	Browse	🛓 New upload	
a \$40 million, 5-year mission, led by NASA's Scie OPS mission, NASA is designating 2023 as the Y	nce Mission Director /ear Of Open Science	rate's Open- e, a community	
pen science			8

View

Why use Zenodo?

- Safe your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- Trusted built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- Citeable every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- No waiting time Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- Open or closed Share e.g. anonymized

Obtain a DOI for your software for every release

🏝 Profile	
🔩 Change password	
Security	
� Linked accounts	
Applications	
Anared links	
🙃 GitHub	

Home / Account / GitHub

E README.md

DOI 10.5281/zenodo.8047483

Organizing your data and software with a reproducible project workflow



Digital Object Identifier (DOI)

June 16, 2023

Software Open Access

DCC-training-lab/project-management: 2023-DCC-Spring-Training-Days

Barbara Vreede; Stephanie van de Sandt; meronvermaas

This is the material as used in our 2023 edition of this DCC spring training

Preview	~
🖹 project-management-2023.zip	×
DCC-training-lab-project-management-c19735b	
• 🗅 .gitignore	12 Bytes
C LICENSE.md	114 Bytes
BEADME.md	3.5 kB
• 🖿 data	
 BREADME.md 	712 Bytes
 datafile_1.xlsx 	9.1 kB
 datafile_10.xlsx 	9.2 kB
 datafile_11.xlsx 	9.2 kB
 datafile_2.xlsx 	9.1 kB
 datafile_3.xlsx 	9.1 kB
 datafile_4.xlsx 	9.1 kB
 datafile_5.xlsx 	9.1 kB
 datafile_6.xlsx 	9.2 kB
 datafile_7.xlsx 	9.3 kB
 datafile_8.xlsx 	9.3 kB
 datafile_9.xlsx 	9.2 kB
 datafiles.zip 	79.6 kB
• 🖿 lessons	

Turriy.		Size		
OCC-training-	ab/project-management-2023.zip	9.7 MB	Preview	🛓 Download
nd5:773d16e0a	15d99d9251deef8e64f2fd43 🛛			
Citations O	0			×
Show only:	Literature (0) Dataset (0) Software (0) Ur	nknown (0)	Search	Q
Show only:	 Literature (0) Dataset (0) Software (0) Ur Citations to this version 	nknown (0)	Search	۹

License

Publication date:

May 12, 2021

DOI:

DOI 10.5281/zenodo.4753914

Keyword(s):

R package Natura 2000 habitat Flanders Belgium data

Related identifiers:

Supplement to https://github.com/inbo/n2khab/tree/v0.5.0

Communities:

Research Institute for Nature and Forest (INBO) Zenodo

License (for files): C GNU General Public License v3.0 only

Older versions

Versions

Version 0.5.0 10.5281/zenodo.4753914	May 12
Version 0.4.0 10.5281/zenodo.4531807	Feb 10
Version 0.3.1 10.5281/zenodo.4133524	Oct 26
Version 0.3.0 10.5281/zenodo.4096139	Oct 16
Version 0.2.0 10.5281/zenodo.3817690	May 8

View all 8 versions



Exercise

- 1. Add a license to the project
- 2. Connect your GitHub page to Zenodo SANDBOX.
- 3. Make a release from the GitHub main page:
- 4. Return to the GitHub page in your Zenodo profile and see what happened.



P.P.



Discussion





Answer these questions for yourself and share the answers with the group

What is the main take-away from this workshop?

What experience in your past has taught you an important lesson about project management?



Thank you!





