

Writing a reproducible PhD thesis using GitHub, R, and Docker



Nami Sunami

Data Steward | TU Eindhoven

n.sunami@tue.nl





**Join Our
Community**



openscience@tue.nl



sites.google.com/view/osceindhoven



First: Lunch & Menti



menti.com/al94p5cm6iwg

Writing a reproducible PhD thesis using GitHub, R, and Docker

Disclaimer:
I'm just sharing my journey
(not giving advice as a data steward)

**Why did I want to make my
thesis reproducible?**

A problem using Google Docs

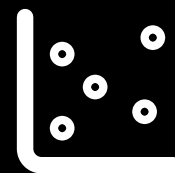


R

Tables



Plots



Stats



```
t = -3.7671, df = 18.332, p-value = 0.001374
```




R



Copy -paste



Google Doc



R



Google Doc



R



...

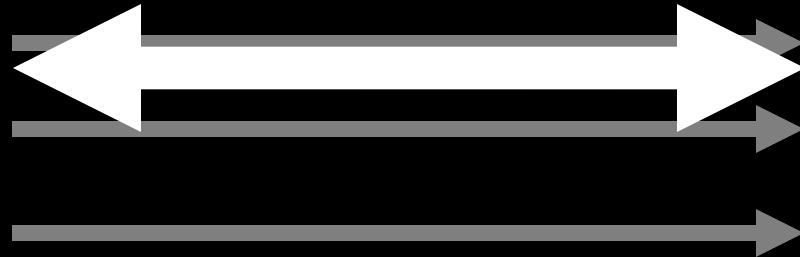


Google Doc


Are they synced?



R



Google Doc

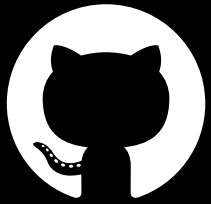
A photograph of a dark, weathered gravestone in a cemetery. The stone is rectangular with a pointed top and shows signs of age and decay. It is set in a patch of dry grass and weeds. In the background, another similar gravestone is visible, slightly out of focus. The overall lighting is somewhat dim, suggesting an overcast day or a shaded area.

Cause of death:
**A thousand
copy and pastes**

**Open Science is about
transparency.**

**A self-reproducing manuscript
achieves this goal.**

Also, it's cool



GitHub

I started my GitHub repo

github.com/nsunami/dissertation

R Markdown to create a document from R



Participants

```
I recruited `r s2_attention_APA$total`  
participants from Prolific in total (Age: `r  
s2_age_mean$msd`; `r s2_gender_id$Female`  
women, `r s2_gender_id$Male` men, and `r  
s2_gender_id$Other` not identifying as a  
woman or man).
```

5.1.2 Participants

I recruited 426 participants from Prolific in total (Age: $M = 24.92$, $SD = 7.33$; 133 women, 287 men, and 6 not identifying as a woman or man).

When did I start my GitHub repo?



When did I start my GitHub repo?



Regret 1

**Not starting
early & small**

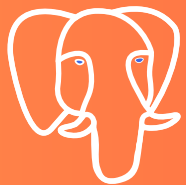
No good git history
for the proposal

Less time to
develop "muscles"
for the workflow

Where did I store my data?

A

**Managed
database**



B

**GitHub
Repo**



C

**Data
Repository**

zenodo

My inner Data Steward is not proud!

B

**GitHub
Repo**



Regret 2

Not storing data in a data repository

Bloating the git
repo size

Binary file & version
control don't go well

Not FAIR

Regret 3

Not committing **early**,
committing **often**,
with **meaningful**
commit messages

Nowadays, I try to follow Conventional Commits
conventionalcommits.org

Commits on Jan 25, 2021

prereg



nsunami committed



on Jan 25, 2021

prereg



nsunami committed



on Jan 25, 2021

prereg



nsunami committed



on Jan 25, 2021

**Nevertheless, I was
able to render the
thesis & graduate**

PLAYING ALONE, FEELING CONNECTED: DO SINGLE-PLAYER
VIDEO GAMES WITH SOCIAL SURROGATES REPLENISH
BELONGING AFTER SOCIAL REJECTION?

by
Naoyuki Sunami

A dissertation submitted to the Faculty of the University of Delaware in partial
fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

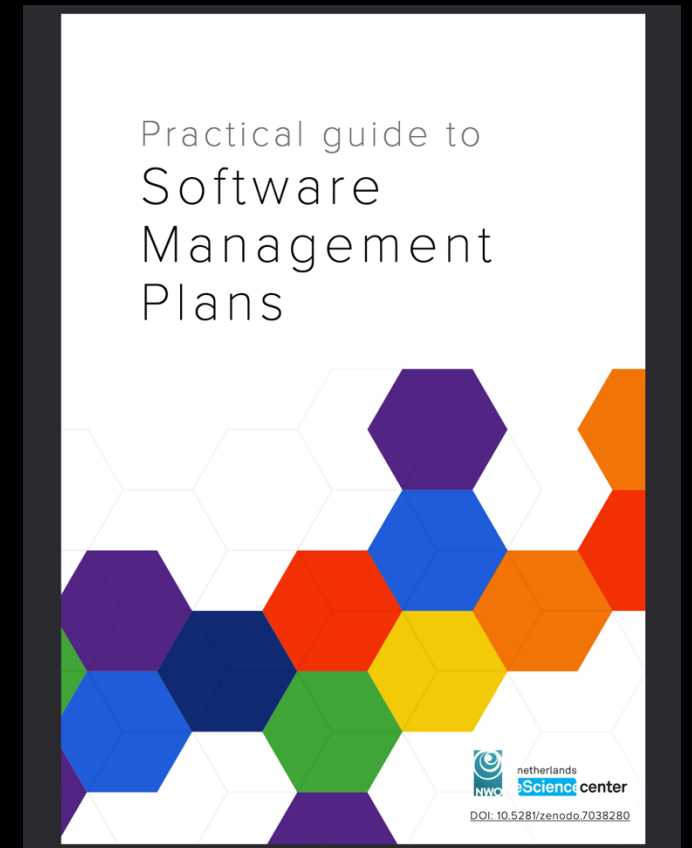
Summer 2021

**And I thought I was done with my
dissertation**

Until last year...

At a meeting about Software Management Plan

I remembered about my dissertation



**“It’s been a while.
Does my dissertation still run?”**



...nope.

What things were broken?

A

**Git-ignoring
files that I
should not**

B

**Could not
install
dependencies**

C

**Issues about
LaTeX & PDF
rendering**

All of these were broken

A

**Git-ignoring
files that I
should not**

B

**Could not
install
dependencies**

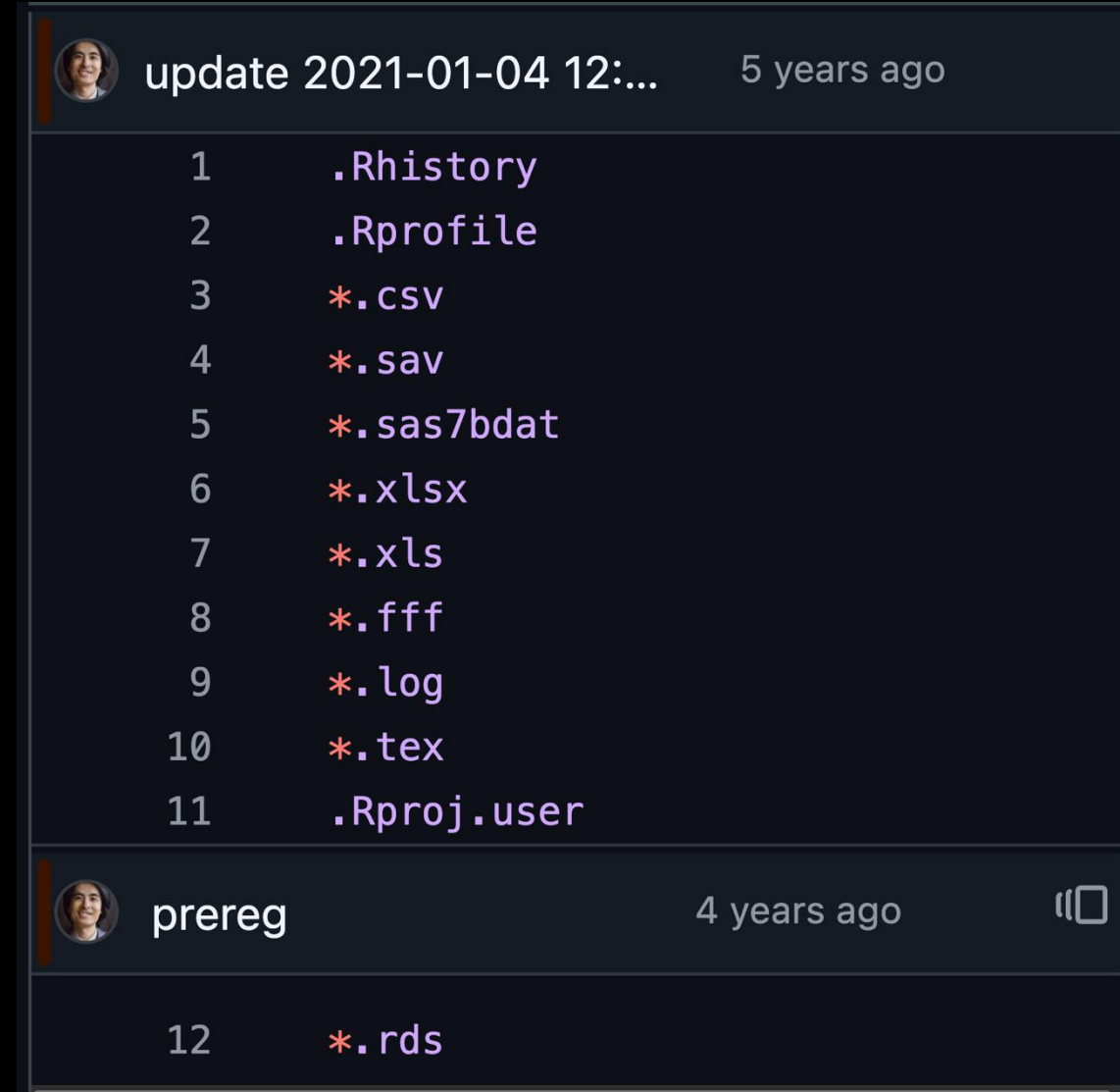
C

**Issues about
LaTeX & PDF
rendering**

Regret 4

Git-ignoring things
that I should not

Treating the git repo as a
scratchpad / working
directory



```
update 2021-01-04 12:... 5 years ago
1      .Rhistory
2      .Rprofile
3      *.csv
4      *.sav
5      *.sas7bdat
6      *.xlsx
7      *.xls
8      *.fff
9      *.log
10     *.tex
11     .Rproj.user

prereg 4 years ago
12     *.rds
```

Next broken thing:
Dependencies

How many packages of dependencies did my dissertation have?

A Less than 50

C 101-200

B 51-100


D More than 200

D More than 200

Total dependencies:

242

packages



Cause of death:

**Random
packages**

Regret 4

**Not keeping track of
dependencies**

**(& not using a
dependency manager)**

*Later learning renv would have been a good idea

Two types of dependencies

R packages



**System
Requirements
(OS)**

 clang++

 lib-png

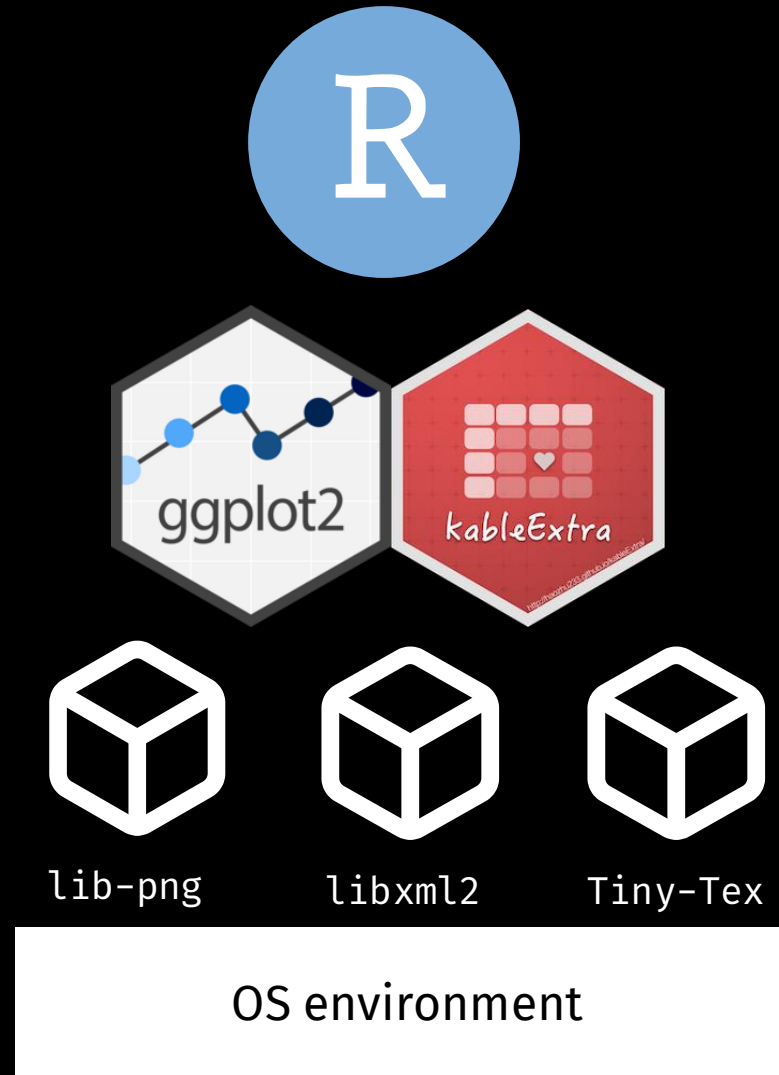
 libxml2

`pak::pkg_deps` can help navigating dependencies

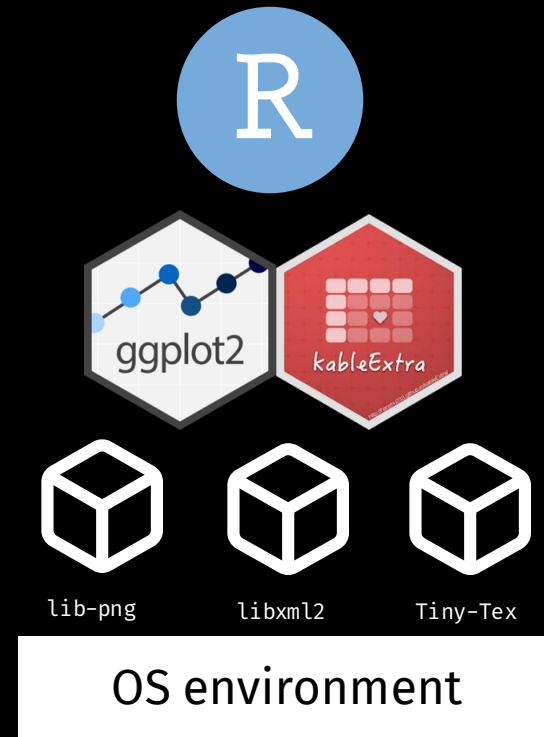
**Re-installing 242 packages
means re-installing their
system requirements**

**Another thing broken:
To render a PDF, I also
needed LaTeX, another
system dependency.**

**I wanted to
snapshot all**



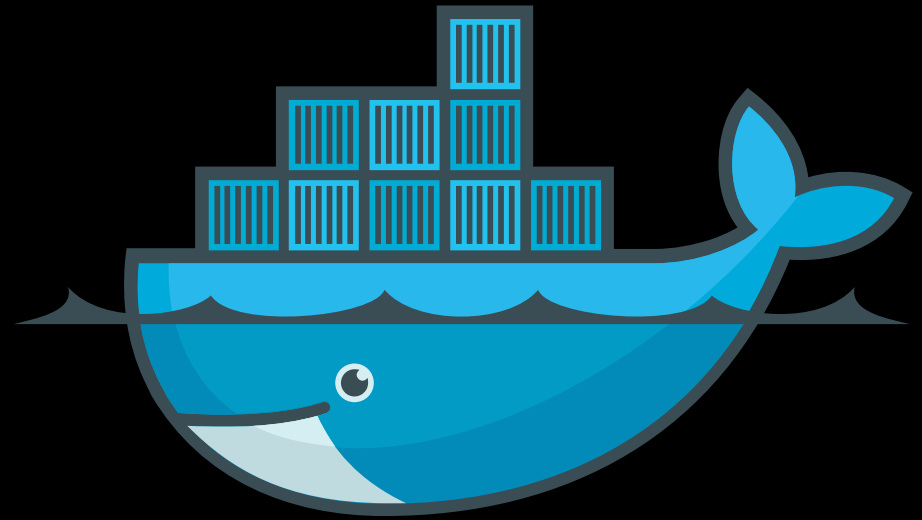
You can create a docker image to do just that



Regret

**Not using Docker
from the start**

(In fact, I finished
Dockerizing this weekend.)



How to reproduce my thesis



```
git clone https://github.com/nsunami/dissertation.git  
cd dissertation  
docker compose up
```

And it will be in the docs folder.

Regrets Recap

1. Not starting early & small
2. Not using a data repository
3. Using bad commit practices
4. Git-ignoring willy-nilly
5. Letting 240+ dependencies run wild
6. Not using Docker earlier

Let's discuss!

Questions

**What are your experiences
working towards a
reproducible paper?**

What are your challenges?

Join Our Community



✉ openscience@tue.nl

🌐 [sites.google.com/view/osceindhov
en](https://sites.google.com/view/osceindhoven)



Footnotes

Icon Sources

- Lucide.dev
- SVGrepo.com

Image Source

- Unsplash

This work is marked with CC0 1.0