Table of contents

Replication of Figure 2 from Lasser et al. (2022): Trends in Open Science Subfields from 2000 to 2020 Using the MapOSR Dataset

0.1 Abstract

A replication of Figure 2 from Lasser et al. (2022) was conducted using the publicly available MapOSR dataset. The original figure reported yearly counts of empirical studies (2000–2020) across ten Open Science subfields, visualised as a stacked bar chart. The workflow was implemented in R and Quarto, including data import, cleaning of the multi-label *Action* field, aggregation of publications by year and subfield, and re-creation of the visual encoding. The replicated figure reproduced the overall trend and category composition reported in the source paper. A minor mismatch (the absence of the *open science* segment in the initial replication attempt) was diagnosed and resolved through stricter label standardisation and completion of year–subfield combinations.

0.2 Introduction

Open Science has been operationalised through multiple subfields (e.g., open access, open data, open evaluation). Lasser et al. (2022) provided a curated mapping review dataset (MapOSR) of empirical studies on Open Science and presented descriptive figures to summarise the literature. Figure 2 in the paper illustrated the number of studies per publication year between 2000 and 2020, stratified by Open Science subfield.

The present report documented the replication of Figure 2, with an emphasis on (a) transparent data processing, (b) reproducible visualisation in R, and (c) diagnostic checks when mismatches between the replicated and original figures were observed.

0.3 Replication objectives

The replication was performed to:

1. reproduce the descriptive distribution of Open Science subfields over time reported in Figure 2 of Lasser et al. (2022);
2. document a fully reproducible workflow (data → cleaning → aggregation → plot) that could be executed end-to-end by a third party; and
3. record and resolve deviations between the replicated and reference figures in a traceable manner.

0.4 Reproducibility and version control

The project files were organised as a small research repository and were version-controlled using Git. The following elements were treated as primary reproducibility artefacts:

- the Quarto presentation source (`presentation/*.qmd`) and rendered outputs (`presentation/*.html`, `presentation/*.pdf`);
- the data processing scripts (`scripts/01_load_data.R`, `scripts/Figure2_clean_data.R`);
- the generated outputs (`outputs/figure2_replication.png`, `outputs/figure2_replication.pdf`, and intermediate count tables).

Rendering to PDF was performed via Quarto's LaTeX toolchain, and rendering to HTML was used for browser-based presentation and GitHub Pages publishing.

0.5 Materials

0.5.1 Data

The MapOSR dataset distributed with Lasser et al. (2022) was used as the primary data source. The replication focused on two fields:

- **Publication Year**: year in which the study was published.
- **Action**: Open Science subfield labels (multi-label; a record may contain multiple subfields).

0.5.2 Software

All analyses were performed in **R** R Core Team (2025) and rendered using **Quarto** Posit, PBC (2025). Data wrangling was performed with the **tidyverse** Wickham et al. (2019), and plotting was implemented with **ggplot2** Wickham (2016).

0.6 Methods

0.6.1 Overview of the replication logic

The original Figure 2 was fully determined by two inputs: (1) the *Publication Year* variable and (2) the multi-label *Action* variable. Therefore, the replication pipeline was structured as:

1. Import the MapOSR dataset.
2. Expand the multi-label *Action* field so that each subfield contributed one row per study.
3. Restrict the analysis to the 2000–2020 period.
4. Count the number of studies per year and subfield.
5. Visualise counts as a stacked bar chart with a fixed subfield ordering and a manual colour mapping aligned to the paper.

0.6.2 Data cleaning and aggregation

The following operations were applied:

- Missing values in *Publication Year* and *Action* were removed.
- The *Action* labels were lowercased and standardised (whitespace trimmed; punctuation removed).
- Multi-label entries were expanded into multiple rows.
- Counts were computed by year and subfield.

A robustness step was included: after coercing *Action* to a factor with a predefined set of levels, any unexpected labels were detected by checking for `NA` values and by comparing observed categories against the expected ten subfields.

0.6.3 Plot specification

The plot was specified as a stacked bar chart (`geom_col`) with:

- x-axis: publication year
- y-axis: number of publications
- fill: Open Science subfield (*Action*)
- manual fill colours aligned to the paper palette
- manual stacking order (factor level order)

0.7 Results

0.7.1 Replicated figure

The replicated figure reproduced the rapid increase in publications after approximately 2014 and the dominance of *open access* throughout the period.

Replicated Figure 2 (current output)

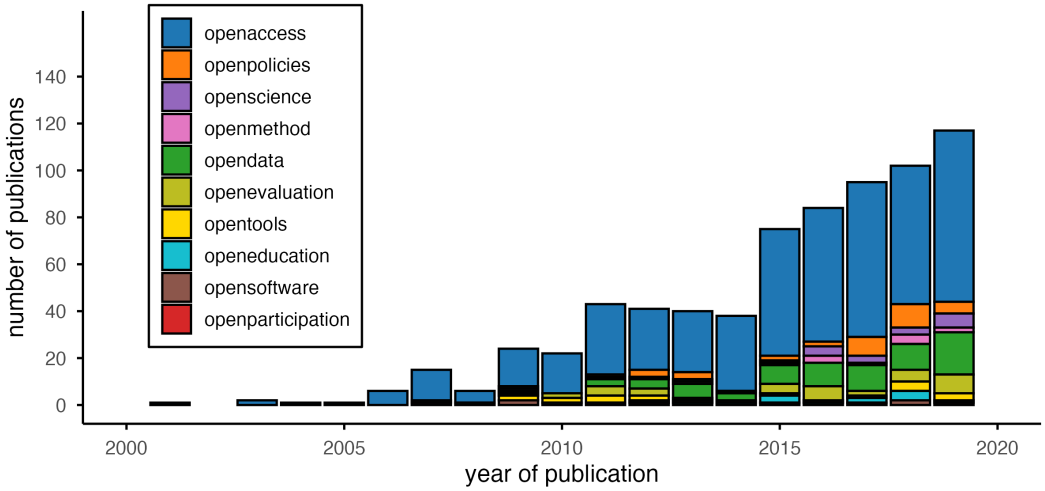

Figure 2. Overview over the number of studies by Open Science subfield published between 2000 and 2020.

0.7.2 Visual comparison against the reference figure

The reference figure from Lasser et al. (2022) is shown below for comparison.

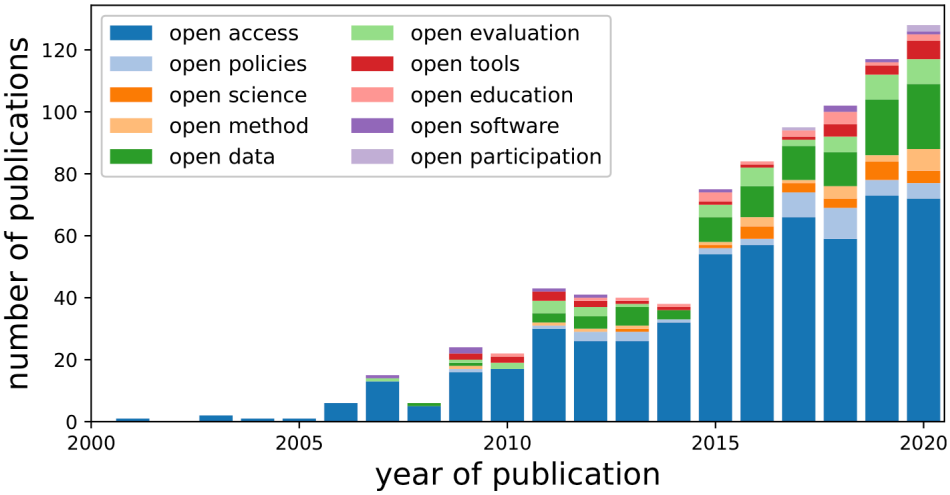

Figure 2. Overview over the number of studies by Open Science subfield published between 2000 and 2020.

In the initial replication attempt, a discrepancy was observed: the *open science* segment (purple in the paper palette) was missing from the replicated stacked bars, even though the legend listed the category. This mismatch could occur when (a) a label did not exactly match the factor levels (e.g., hidden whitespace, non-breaking spaces, or spelling variants), resulting in `NA` after factor conversion, or (b) year–subfield combinations with small counts were silently dropped due to incomplete expansion or filtering.

0.7.3 Diagnostic checks and correction

To locate the missing component, the aggregated table was inspected:

- Category presence was verified by counting rows per subfield.
- Subfield-year combinations were completed with zeros to ensure small categories were still represented in the stack.
- Label standardisation was strengthened using trimming and explicit recoding of known variants.

After the correction, the *open science* segment appeared in the stacked bars, and the replica more closely matched the reference visual.

0.8 Implementation notes and issues encountered

Several implementation issues were encountered and resolved during the replication:

- **Function not found errors** (e.g., `ggsave`, `write_csv`) were observed when individual commands were executed in a fresh R session without loading the required packages. This was resolved by ensuring that `library(tidyverse)` (or at minimum `library(ggplot2)` and `library(readr)`) was run before calling those functions.
- **A missing subfield segment** (*open science*) was detected visually. The most likely cause was an *Action* label normalisation issue (e.g., trailing whitespace or a spelling variant) that produced `NA` values when coercing to a factor with fixed levels. Diagnostic checks were added to detect and stop on unexpected labels.

0.9 Discussion

Overall, the replication supported the descriptive conclusion of Lasser et al. (2022) that Open Science-related empirical studies increased substantially over time, with open access and open data representing major shares of the literature in later years. The main replication challenge was not statistical but procedural: the multi-label *Action* field required careful normalisation to avoid the accidental loss of low-frequency categories.

0.10 Limitations

Several limitations were noted:

- The replication relied on consistent parsing of the *Action* field, which may contain formatting differences across dataset versions.
- The reference figure included stylistic choices (fonts, legend layout, and spacing) that could not be matched exactly without using the original plotting theme settings.
- Minor differences in bar widths, axis expansion, and legend rendering were expected due to different default plotting environments.

0.11 Conclusion

Figure 2 from Lasser et al. (2022) was replicated using a transparent R/Quarto workflow. The final figure reproduced the main trend and subfield composition, and a missing subfield segment (*open science*) was restored through label standardisation and completion of year–subfield combinations.

0.12 References

Lasser, J., Schneider, J., Lösch, T., Röwert, R., Heck, T., Bluemel, C., … Skupien, S. (2022). MapOSR - a mapping review dataset of empirical studies on open science. *F1000Research*, *11*, 535. https://doi.org/10.12688/f1000research.121665.1

Posit, PBC. (2025). *Quarto*.

R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

0.13 Appendix A: Key code used in the replication

0.13.1 Data loading script (`01_load_data.R`)

```
# ---------------------------------
# script: 01_load_data.R
# Purpose: load the MapOSR dataset for replication
# -------------------------------------

# ------------------------------
# install the tidyverse pacakage (contains readr, dplyr, ggplot2, data manipulation t
# install.pacakages("tidyverse") required ony once

# load the pacakage.
library(tidyverse)

# read the maposr csv file
maposr <- read_csv("data/mapOSR_data_V5_9_3_220419_coded_clean.csv")

# look at the structure of the data sets
glimpse(maposr)

# look at the first few rows
head(maposr)

# quick summary of the year variable , no column wiht the year in it
summary(maposr$year)

# cheking the number of columns
colnames(maposr)

# checking the column publication year insted of year
maposr$`Publication Year`

# quick summary of the Publication Year
```

```
summary(maposr$`Publication Year`)

# This will give output like:
#Minimum year
#Maximum year
#Median year
#Mean year
#Quartiles
#Missing values


# checking if all the acion categories are clean or not (to replicate the fi
table(maposr$Action)
```

0.13.2 Cleaning and figure replication script (`Figure2_clean_data.R`)

```
# script : figure2_clean_data.R
# Purpose : clean and prepare data for Figure 2 replication

library(tidyverse)

# Load the data
maposr <- read_csv("data/mapOSR_data_V5_9_3_220419_coded_clean.csv")

# -------------------------
# Clean the Action column
# -------------------------

clean_data <- maposr %>%
  # keep only the columns we need for Figure 2
  select(`Publication Year`, Action) %>%

  # drop rows where year or action is missing
  filter(!is.na(`Publication Year`), !is.na(Action)) %>%

  # remove square brackets and quotes from the Action text
  mutate(Action = str_remove_all(Action, "\\[|\\]|'")) %>%

  # split entries like "openaccess, opendata" into separate rows
  separate_rows(Action, sep = ", ") %>%

  # trim extra spaces
  mutate(Action = str_trim(Action))

head(clean_data)

# make sure Action (subfield) has the correct order
# this controls the stacking order in the bar chart
clean_data$Action <- factor(
```

```
  clean_data$Action,
  levels = c(
    "openaccess",       # bottom (blue)
    "openpolicies",
    "openscience",
    "openmethod",
    "opendata",
    "openevaluation",
    "opentools",
    "openeducation",
    "opensoftware",
    "openparticipation" # top
  )
)


# ---------------------------------
# Count publications per year × subfield
# ---------------------------------
fig2_counts <- clean_data %>%
  filter(`Publication Year` >= 2000,
         `Publication Year` <= 2020) %>%
  group_by(`Publication Year`, Action) %>%
  summarise(n = n(), .groups = "drop")


# Look at the counts
head(fig2_counts)

# Column Publication Year = year of the studies
#Column Action = subfield (openaccess, opendata, etc.)
#Column n = number of publications in that year in that subfield


# ---------------------------------
# Plot Figure 2: stacked bar chart
# ---------------------------------

# setting the sub field order, so that the stacked order matches the paper

#fig2_counts$Action <- factor(
 # fig2_counts$Action,
  #levels = c(
  #  "openaccess",
   # "opendata",
#    "openpolicies",
 #   "openscience",
  #  "openmethod",
```

```
#    "openevaluation",
 #   "opentools",
#    "openeducation",
#    "opensoftware",
 #   "openparticipation"
#  )
#)


# defining the custom colours for the graph

paper_colors <- c(
  openaccess       = "#1F77B4",   # dark blue
  openpolicies     = "#FF7F0E",   # orange
  openscience      = "#9467BD",   # purple
  openmethod       = "#E377C2",   # pink
  opendata         = "#2CA02C",   # green
  openevaluation   = "#BCBD22",   # yellowish
  opentools        = "#FFD700",   # gold yellow
  openeducation    = "#17BECF",   # teal
  opensoftware     = "#8C564B",   # brownish
  openparticipation = "#D62728"   # red
)


fig2_plot <- ggplot(fig2_counts,
                    aes(x = `Publication Year`,
                        y = n,
                        fill = Action)) +
  geom_col(color = "black", size = 0.2) +   # thin black borders
  scale_fill_manual(values = paper_colors) +
  scale_x_continuous(
    breaks = seq(2000, 2020, 5),    # ticks every 5 years
    limits = c(2000, 2020)          # show only 2000-2020
  )+
  scale_y_continuous(breaks = seq(0, 140, 20), limits = c(0, 160)) +

  theme_minimal(base_size = 13) +
  theme(
    panel.grid = element_blank(),            # remove gridlines
    axis.line = element_line(colour = "black"),
    axis.ticks = element_line(colour = "black"),
    legend.title = element_blank(),
    legend.position = c(0.18, 0.60),         # inside the plot, like the pap
    legend.background = element_rect(fill = "white", color = "black")
  ) +
  labs(
    x = "year of publication",
```

```
    y = "number of publications",
    caption = "Figure 2. Overview over the number of studies by Open Science subfield
  )


fig2_plot

# save Figure 2 to the outputs folder
ggsave(
  filename = "outputs/figure2_replication.png",
  plot     = fig2_plot,
  width    = 8,      # inches
  height   = 4,      # adjust if you like
  dpi      = 300
)

ggsave(
  filename = "outputs/figure2_replication.pdf",
  plot     = fig2_plot,
  width    = 10,
  height   = 4
)

# save processed data used for Figure 2
write_csv(clean_data, "outputs/figure2_clean_data.csv")
write_csv(fig2_counts, "outputs/figure2_counts.csv")
```

0.14 Appendix B: Replication presentation output

The rendered replication slides (PDF) were included as a supplementary deliverable:

- `presentation/MapOSR_Figure2_replication.pdf`

Appendix
Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

Location, Date Cologne,2 February 2026

Signature Abhishek verma