MapOSR Replication Project

Replication of Figure 2 — Studies by Open Science Subfield (2000–2020)

Abhishek Verma

What this presentation does

Replication target: Figure 2 (stacked bars: number of studies per year by Open Science subfield).

Data used: the provided MapOSR coded dataset (`mapOSR_data_V5_9_3_220419_coded_clean.csv`). The replication is feasible because the dataset already contains the two key variables needed:

- `Publication Year` = year of publication
- `Action` = Open Science subfield(s) (sometimes multiple per record, stored as a list-like string)

Folder structure (so this runs on your machine)

Expected project folders:

- `data/` → contains the CSV
- `outputs/` → will store the generated figure + intermediate CSVs

If `outputs/` does not exist, this script will create it.

```
# Create outputs folder if it doesn't exist
if (!dir.exists("outputs")) dir.create("outputs", recursive = TRUE)
```

Step 1 — Load libraries

We use **tidyverse** for data import, cleaning, and plotting.

```
library(tidyverse)
library(here)
```

Step 2 — Load the dataset

```
maposr <- read_csv(here("data/mapOSR_data_V5_9_3_220419_coded_clean.csv"))
```

Quick sanity checks:

```
#dim(maposr)
#glimpse(maposr)
```

Step 3 — Confirm the key variables exist

This replication depends on `Publication Year` and `Action`.

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2001    2014    2017    2016    2019    2021      10
```

```
table(is.na(maposr$`Publication Year`))
```

```
FALSE  TRUE
  685    10
```

```r
table(is.na(maposr$Action))
```

```
FALSE
  695
```

Step 4 — Why cleaning is needed for `Action`

In the dataset, `Action` can contain multiple subfields per study (example idea: `"['openaccess', 'opendata']"`).
To replicate Figure 2, we must:

1. remove the bracket/quote formatting
2. split multiple subfields into separate rows (one row = one subfield assignment)

Step 5 — Clean and reshape the data

```r
clean_data <- maposr %>%
  select(`Publication Year`, Action) %>%
  filter(!is.na(`Publication Year`), !is.na(Action)) %>%
  mutate(Action = str_remove_all(Action, "\\[|\\]|'")) %>%   # remove [ ] and '
  separate_rows(Action, sep = ", ") %>%                      # split multi-label cells
  mutate(Action = str_trim(Action))                           # trim spaces

head(clean_data)
```

```
# A tibble: 6 x 2
  `Publication Year` Action
               <dbl> <chr>
1               2015 openaccess
2               2013 openaccess
3               2013 openpolicies
4               2018 openaccess
5               2020 openaccess
6               2019 openaccess
```

Step 6 — Fix the stacking order (match paper)

Stacking order is controlled by the **factor levels** of `Action`.

```
clean_data$Action <- factor(
  clean_data$Action,
  levels = c(
    "openaccess",
    "openpolicies",
    "openscience",
    "openmethod",
    "opendata",
    "openevaluation",
    "opentools",
    "openeducation",
    "opensoftware",
    "openparticipation"
  )
)
```

Step 7 — Count publications per year × subfield (2000–2020)

```
fig2_counts <- clean_data %>%
  filter(`Publication Year` >= 2000, `Publication Year` <= 2020) %>%
  group_by(`Publication Year`, Action) %>%
  summarise(n = n(), .groups = "drop")

head(fig2_counts)
```

```
# A tibble: 6 x 3
  `Publication Year` Action         n
               <dbl> <fct>      <int>
1               2001 openaccess     1
2               2003 openaccess     2
3               2004 openaccess     1
```

```
4                2005 openaccess     1
5                2006 openaccess     6
6                2007 openaccess    13
```

Step 8 — Define the paper color palette

```
paper_colors <- c(
  openaccess         = "#1F77B4",
  openpolicies       = "#FF7F0E",
  openscience        = "#9467BD",
  openmethod         = "#E377C2",
  opendata           = "#2CA02C",
  openevaluation     = "#BCBD22",
  opentools          = "#FFD700",
  openeducation      = "#17BECF",
  opensoftware       = "#8C564B",
  openparticipation = "#D62728"
)
```

```
library(tibble)
library(dplyr)
library(knitr)

tibble(
  subfield = names(paper_colors),
  hex = unname(paper_colors)
) |>
  mutate(color = paste0(
    "<span style='display:inline-block;width:80px;height:20px;background:",
    hex,
    ";border:1px solid #444'></span>"
  )) |>
  select(subfield, color, hex) |>
  kable(escape = FALSE)
```

subfield	color	hex
openaccess		#1F77B4

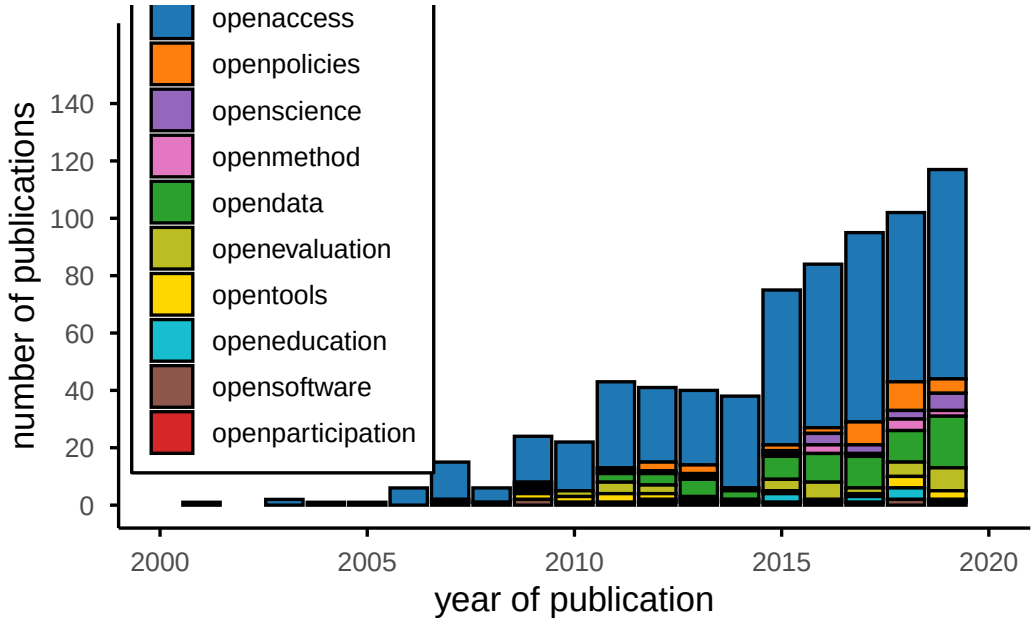
subfield	color	hex
openpolicies		#FF7F0E
openscience		#9467BD
open-method		#E377C2
opendata		#2CA02C
openevalua-tion		#BCBD22
opentools		#FFD700
openeduca-tion		#17BECF
opensoft-ware		#8C564B
openpartici-pation		#D62728

Step 9 — Plot Figure 2 (replication)

```
fig2_plot <- ggplot(fig2_counts,
                    aes(x = `Publication Year`, y = n, fill = Action)) +
  geom_col(color = "black", size = 0.2) +
  scale_fill_manual(values = paper_colors) +
  scale_x_continuous(breaks = seq(2000, 2020, 5), limits = c(2000, 2020)) +
  scale_y_continuous(breaks = seq(0, 140, 20), limits = c(0, 160)) +
  theme_minimal(base_size = 13) +
  theme(
    panel.grid = element_blank(),
    axis.line = element_line(colour = "black"),
    axis.ticks = element_line(colour = "black"),
    legend.title = element_blank(),
    legend.position = c(0.18, 0.60),
    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 publis
```

```
  )

fig2_plot
```



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

Step 10 — Save outputs (for your report + slides)

```
ggsave("outputs/figure2_replication.png", fig2_plot, width = 8, height = 4, dpi = 300)
ggsave("outputs/figure2_replication.pdf", fig2_plot, width = 10, height = 4)

write_csv(clean_data,  "outputs/figure2_clean_data.csv")
write_csv(fig2_counts, "outputs/figure2_counts.csv")
```

Comparison: Original vs. Replication

Original Paper

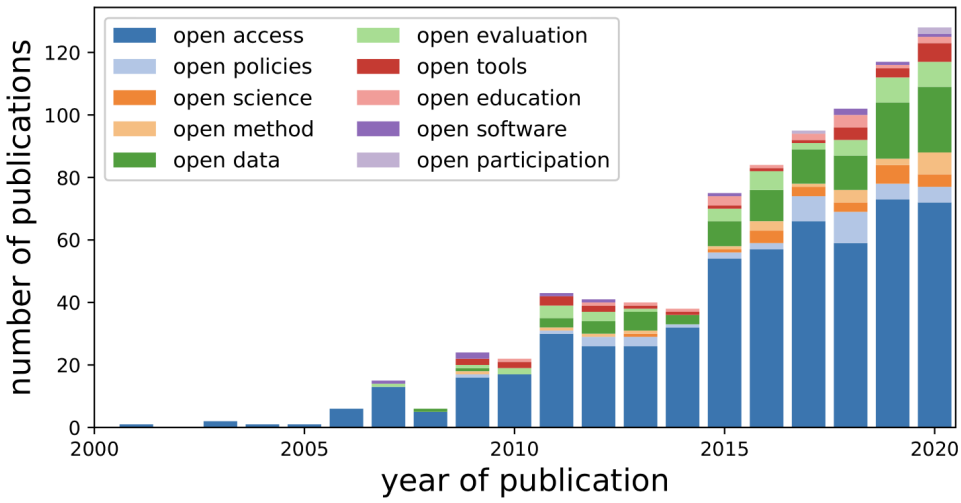


Figure 2. Overview over the number of studies by Open Science subfield published between 2000 and 2020.

Our Replication

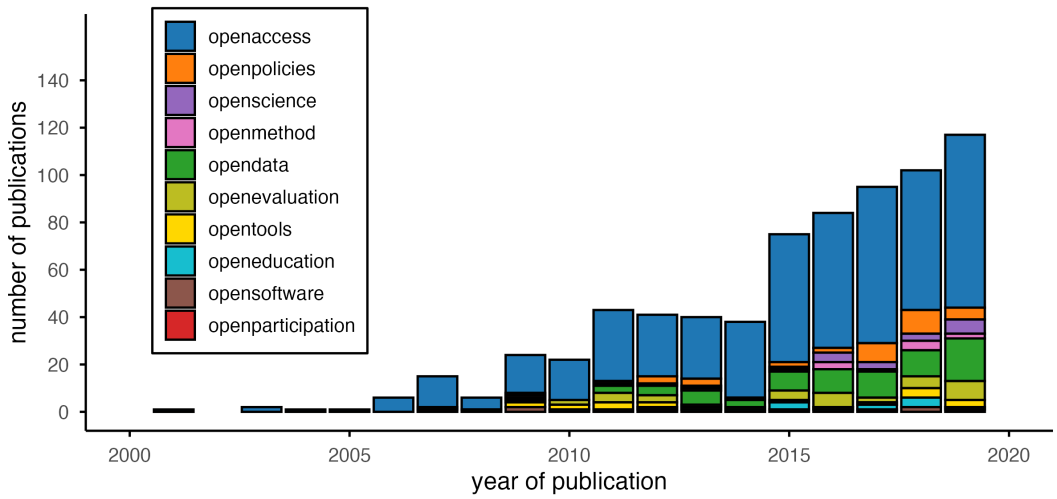


Figure 2. Overview over the number of studies by Open Science subfield published between 2000 and 2020.

Replication Notes)

- The replication is **data-driven**: Figure 2 is fully determined by `Publication Year` and `Action`.
- The only "interpretation" step is **how multi-label `Action` cells are expanded** into multiple rows.
- Matching the paper requires:
  - the **year range filter (2000–2020)**
  - the **factor order** (controls the stack order)
  - the **same color mapping** and legend placement

Deliverables produced by the workflow: final figure (`.png` + `.pdf`) and the intermediate cleaned/count tables in `outputs/`.

Appendix — Original scripts provided

These two scripts correspond to the same steps shown in this presentation:

- Data loading script: `01_load_data.R`
- Cleaning + Figure 2 replication script: `Figure2_clean_data.R`