27 Open Science Indicators

R. Heller · Date: 2026-05-09 · Last changed: 2026-10-31

27.1 Learning objectives

After completing this chapter, you will be able to:

- Classify open access types (gold, green, hybrid, bronze) using OpenAlex metadata
- Compute OA rates by year, journal, and discipline
- Track the growth of preprint posting and preprint-to-publication pathways
- Assess data availability statement coverage in a corpus
- Discuss the relationship between open science practices and citation impact

27.2 Setup

```
library(tidyverse)
library(openalexR)
library(glue)
library(gt)

set.seed(20260509)

source(here::here("R", "api_helpers.R"))
source(here::here("R", "utils.R"))
source(here::here("R", "sci_palette.R"))
```

27.3 Conceptual background

The open science movement aims to make research processes and outputs freely accessible. Bibliometric analysis can measure the extent to which these ideals are being realised, track trends over time, and identify disparities across disciplines and geographies.

Open access (OA) is the most measurable dimension of open science. OpenAlex classifies each work's OA status using Unpaywall data: **gold** (published in a fully OA journal), **green** (available in a repository), **hybrid** (OA in a subscription journal via APC), **bronze** (free to read but without an explicit open license), or **closed** (Priem et al. 2022).

Preprints are manuscripts posted publicly before peer review. The growth of preprint servers (bioRxiv, arXiv, medRxiv) has accelerated since 2015, with a dramatic spike during the COVID-19 pandemic. Bibliometric analysis can track preprint posting rates, the time from preprint to journal publication, and whether preprints receive more citations than non-preprint papers.

Data availability statements (DAS) indicate whether the data underlying a study are accessible. Their prevalence is growing as journals adopt mandatory data-sharing policies, but the quality of DAS varies: many state “data available upon request” without providing actual access.

Measuring open science is important for policy. Funders increasingly require OA publication and data sharing; institutions report OA rates as indicators of compliance and impact. However, Hicks et al. (2015) caution that metrics-driven OA mandates can produce perverse incentives (predatory journals, superficial compliance).

27.4 Worked example

27.4.1 OA rates by year

```
works ← fetch_openalex(
  entity = "works",
  primary_location.source.id = "S148561398",
  from_publication_date = "2015-01-01",
  to_publication_date = "2023-12-31",
  type = "article",
  options = list(sample = 600, seed = 42)
)

works_oa ← works ▷
  transmute(
    id, year = year(publication_date),
    oa_status = oa_status,
    cited_by_count
  )

cat(glue("Works: {nrow(works_oa)}\n"))
```

```
#> Works: 600
```

```
oa_by_year ← works_oa ▷
  count(year, oa_status) ▷
  group_by(year) ▷
  mutate(pct = n / sum(n)) ▷
  ungroup()

ggplot(oa_by_year, aes(x = factor(year), y = pct, fill = oa_status)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = palette_sci(n_distinct(oa_by_year$oa_status))) +
  labs(x = "Publication year", y = "Proportion", fill = "OA status") +
  theme_sci()
```

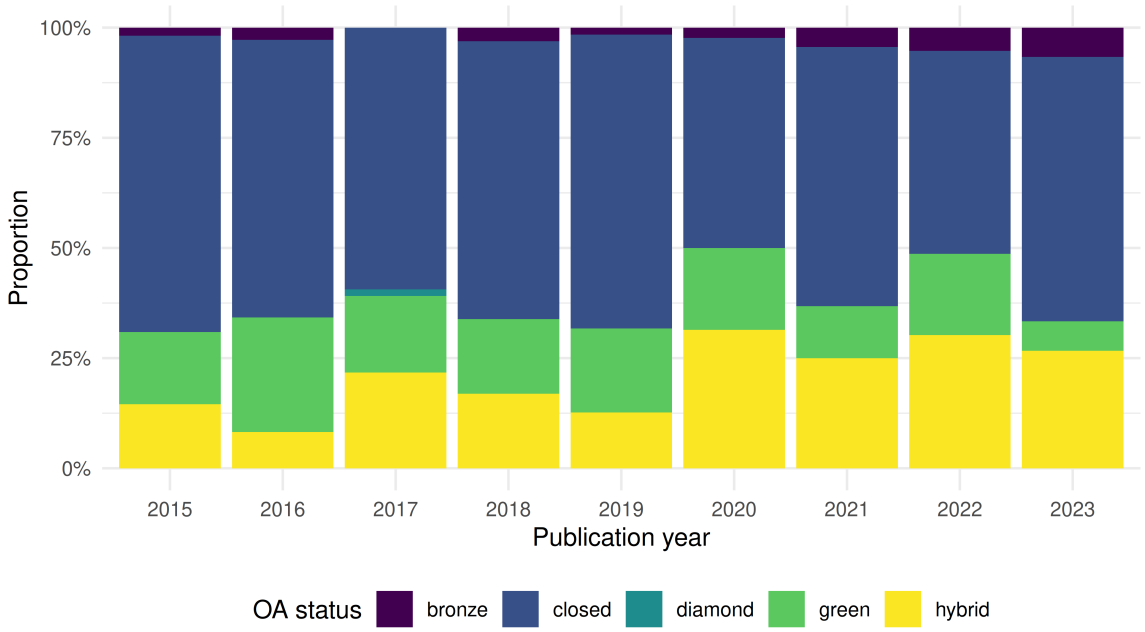


Figure 27.1: Open access rates by year and OA type.

27.4.2 OA citation advantage

```
oa_cites ← works_oa ▷
  filter(!is.na(oa_status)) ▷
  mutate(is_oa = oa_status ≠ "closed") ▷
  group_by(is_oa) ▷
  summarise(
    n = n(),
    mean_cites = round(mean(cited_by_count), 1),
    median_cites = median(cited_by_count),
    .groups = "drop"
  )

oa_cites ▷ gt()
```

is_oa	n	mean_cites	median_cites
FALSE	350	29.5	20.5
TRUE	250	65.3	25.0

27.4.3 Document type breakdown

```
works |>
  count(type, oa_status) |>
  group_by(type) |>
  mutate(pct = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = pct, y = type, fill = oa_status)) +
  geom_col() +
  scale_x_continuous(labels = scales::percent) +
  scale_fill_manual(values = palette_sci(n_distinct(works$oa_status))) +
  labs(x = "Proportion", y = NULL, fill = "OA status") +
  theme_sci()
```

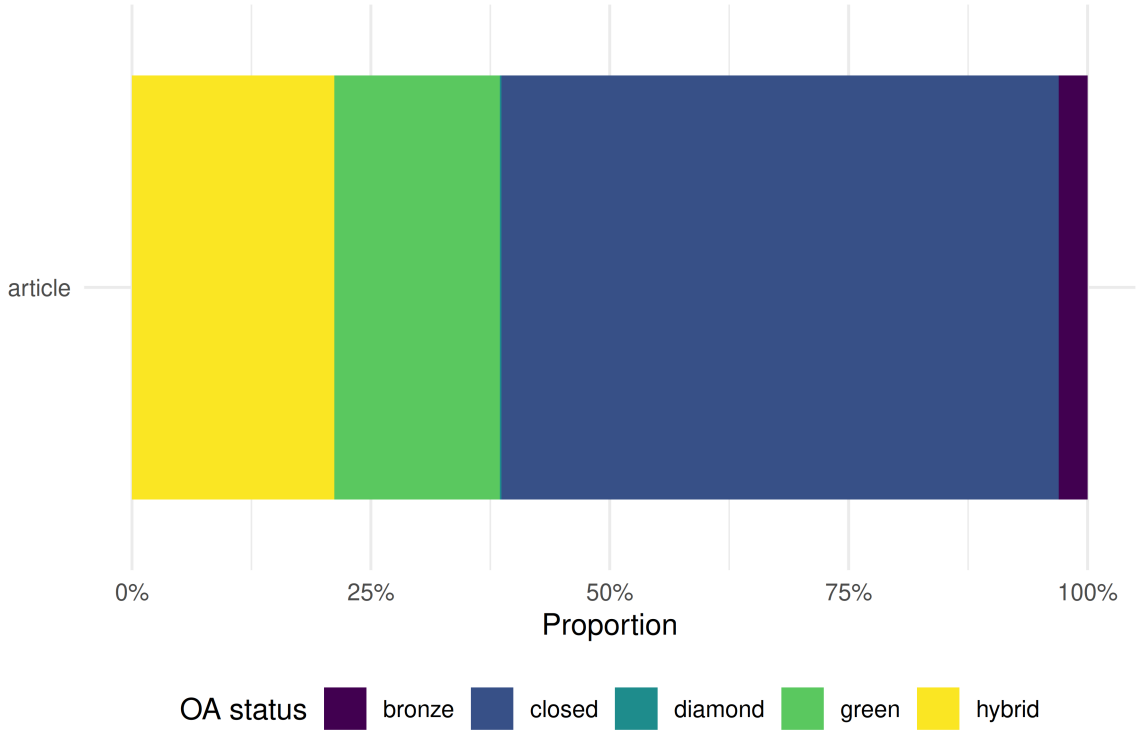


Figure 27.2: OA status by document type.

27.4.4 Summary table

```
works_oa |>
  count(oa_status, sort = TRUE) |>
  mutate(pct = scales::percent(n / sum(n))) |>
  gt() |>
  cols_label(oa_status = "OA Status", n = "Count", pct = "Percentage")
```

OA Status	Count	Percentage
closed	350	58.3%

OA Status	Count	Percentage
hybrid	127	21.2%
green	104	17.3%
bronze	18	3.0%
diamond	1	0.2%

27.5 Diagnostics and interpretation

- **Classification accuracy**: OpenAlex OA classification relies on Unpaywall, which may lag behind publisher changes. Spot-check a sample of "closed" articles to verify they are truly behind a paywall.
- **Bronze ambiguity**: Bronze OA means free-to-read without an explicit license. Publishers can revoke access at any time. Do not treat bronze as equivalent to gold or green.
- **Year effects**: OA rates have risen over time due to funder mandates and policy changes. Compare year-normalised rates rather than pooling across years.
- **Self-selection**: Higher OA rates among highly cited papers may reflect self-selection (better papers are more likely to be made OA) rather than a causal OA effect.

27.6 Limitations and responsible use

27.7 Limitations and responsible use

- **OA ≠ quality.** Open access status is a property of the publication venue and access model, not the research quality. Predatory journals are often gold OA.
- **Coverage gaps.** Green OA via institutional repositories is underreported in some databases. Actual OA rates may be higher than measured.
- **Data sharing claims vs. reality.** A data availability statement saying "data available upon request" often results in no response when requests are made. Measuring DAS presence overestimates actual data sharing.
- **Equity implications.** Gold OA via article processing charges shifts costs from readers to authors, potentially disadvantaging researchers without institutional funding (Hicks et al. 2015).

27.8 Common pitfalls

27.9 Common pitfalls

- **Treating all OA as equivalent.** Gold, green, hybrid, and bronze OA have very different implications for access, sustainability, and cost.
- **Ignoring embargo periods.** Green OA articles may have 6–24 month embargoes before they become freely available. They are "closed" during the embargo.
- **Double-counting.** An article available in both a gold OA journal and a green repository should be counted once. OpenAlex classifies by the "best" OA type.
- **Using OA rate as a performance indicator.** OA rates reflect policy and funding environment, not research quality. Mandated OA inflates rates without changing the underlying work.

27.10 Exercises

1. **Cross-journal OA comparison.** Compare OA rates for three journals in different disciplines. Which discipline has the highest gold OA rate?
2. **Temporal OA growth.** Compute the year-over-year growth rate of gold OA articles in your corpus. Has growth accelerated since 2020?
3. **OA and collaboration.** Is there a relationship between the number of authors on a paper and its OA status? Hypothesise why.

27.11 Solutions

Solutions are provided in E.

27.12 Further reading

- Priem et al. (2022) — OpenAlex OA classification methodology.
- Hicks et al. (2015) — The Leiden Manifesto: principles for responsible metrics, including OA considerations.
- American Society for Cell Biology (2012) — DORA: recommendations for research assessment beyond journal metrics.

27.13 Session info

```
sessionInfo()
```

```
#> R version 4.6.1 (2026-06-24)
#> Platform: x86_64-pc-linux-gnu
#> Running under: Ubuntu 24.04.5 LTS
#>
#> Matrix products: default
#> BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
#> LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.26.so;  LAPACK
version 3.12.0
#>
#> locale:
#>  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
#>  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
#>  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
#> [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
#>
#> time zone: UTC
#> tzcode source: system (glibc)
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] uwot_0.2.5                Matrix_1.7-5
#>  [3] word2vec_0.4.1            stm_1.3.10
#>  [5] topicmodels_0.2-17        quanteda.textstats_0.97.2
#>  [7] visNetwork_2.1.4          ggraph_2.2.2
#>  [9] tidygraph_1.3.1           igraph_2.3.4
#> [11] quanteda_4.5.1            pdftools_3.9.1
#> [13] arrow_26.0.1.1            bibliometrix_5.5.0
#> [15] RefManageR_1.4.0          bib2df_1.1.2.0
#> [17] rcrossref_1.2.1           gt_1.4.0
#> [19] tidytext_0.4.3            glue_1.8.1
#> [21] openalexR_3.1.2           lubridate_1.9.5
#> [23] forcats_1.0.1             stringr_1.6.0
#> [25] dplyr_1.2.1               purrr_1.2.2
#> [27] readr_2.2.0               tidyr_1.3.2
#> [29] tibble_3.3.1              ggplot2_4.0.4
#> [31] tidyverse_2.0.0
#>
#> loaded via a namespace (and not attached):
#>   [1] bibtex_0.5.3           RColorBrewer_1.1-3     rstudioapi_0.19.0
#>   [4] jsonlite_2.0.0         magrittr_2.0.5         modeltools_0.2-25
#>   [7] farver_2.1.2           rmarkdown_2.32         fs_2.1.0
#>  [10] vctrs_0.7.3            memoise_2.0.1          askpass_1.2.1
#>  [13] base64enc_0.1-6        htmltools_0.5.9        contentanalysis_1.1.1
#>  [16] curl_8.1.0             janeaustenr_1.0.0      cellranger_1.1.0
#>  [19] sass_0.5.0             bslib_0.12.0           htmlwidgets_1.6.4
#>  [22] tokenizers_0.3.0       plyr_1.8.9             httr2_1.3.0
#>  [25] plotly_4.12.1          cachem_1.1.0           dimensionsR_0.0.3
#>  [28] mime_0.13              lifecycle_1.0.5        pkgconfig_2.0.3
#>  [31] R6_2.6.1               fastmap_1.2.0          shiny_1.15.0
#>  [34] digest_0.6.39          patchwork_1.3.2        shinycssloaders_1.1.0
#>  [37] rprojroot_2.1.1        RSpectra_0.16-2        SnowballC_0.7.1
#>  [40] labeling_0.4.3         urltools_1.7.3.1       timechange_0.4.0
#>  [43] mgcv_1.9-4             polyclip_1.10-7        httr_1.4.9
#>  [46] compiler_4.6.1         here_1.0.2             bit64_4.8.6
#>  [49] withr_3.0.3            S7_1.0.0               viridis_0.6.5
#>  [52] ggforce_0.5.0          MASS_7.3-65            bibliometrixData_0.3.0
#>  [55] tools_4.6.1            otel_0.3.0             stopwords_2.3
#>  [58] zip_3.0.2              httpuv_1.6.17          rentrez_1.2.4
#>  [61] nlme_3.1-169           promises_1.5.0         grid_4.6.1
#>  [64] stringdist_0.9.17      reshape2_1.4.5         generics_0.1.4
#>  [67] gtable_0.3.6           tzdb_0.5.0             rscopus_0.9.0
```

```
#>  [70] ca_0.72                data.table_1.18.6.1    hms_1.1.4
#>  [73] xml2_1.6.0             utf8_1.2.6             ggrepel_0.9.8
#>  [76] pillar_1.11.1          nsyllable_1.0.1        vroom_1.7.1
#>  [79] later_1.4.8            splines_4.6.1          tweenr_2.0.3
#>  [82] lattice_0.22-9         FNN_1.1.4.1            bit_4.6.0
#>  [85] tidyselect_1.2.1       tm_0.7-20              miniUI_0.1.2
#>  [88] downlit_0.4.5          knitr_1.52             gridExtra_2.3.1
#>  [91] NLP_0.3-3              bookdown_0.48          stats4_4.6.1
#>  [94] crul_1.6.0             xfun_0.62              graphlayouts_1.2.5
#>  [97] matrixStats_1.5.0      DT_0.34.0              humaniformat_0.6.0
#> [100] stringi_1.8.9          qpdf_1.4.1             yaml_2.3.12
#> [103] evaluate_1.0.5         httpcode_0.3.0         cli_3.6.6
#> [106] xtable_1.8-8           jquerylib_0.1.4        Rcpp_1.1.2
#> [109] readxl_1.6.0           triebeard_0.4.1        XML_3.99-0.25
#> [112] parallel_4.6.1         assertthat_0.2.1       pubmedR_1.1.0
#> [115] slam_0.1-56            viridisLite_0.4.3      scales_1.4.0
#> [118] crayon_1.5.3           openxlsx_4.2.9         rlang_1.3.0
#> [121] fastmatch_1.1-8
```