# Publication tracker reporting

## Table of contents

### Year in Review / Annual Report

Once a year, the ALA produces a "Year in Review" report which encompasses our achievements over the last twelve months. This includes a section providing key summaries of the publication tracker, which differ from the quarterly report.

They are:

1. **The annual (calendar year) numbers of *journal articles* citing the ALA, up until the earliest article in 2007**
2. **Annual (financial year) total publications (all) to journal article ratio**
3. **Top 10 journal articles (grand total + yearly)**
4. **Research domains or categories citing the ALA**

> This code was last updated on 2023-10-23 (yyyy/mm/dd).
>
> Data to run this code can be found here, and within this repo: science - projects - literature-tracking - data (set to .gitignore). You can export the current literature tracking repository by accessing Zotero (how here) and choosing File > Export library (as .csv).

# 1 Step 1 - Data cleaning:

1. Load required packages

```
library(here)
library(tidyverse)
library(janitor)
library(lubridate)
```

2. Load in exported .csv from Zotero

First you will need to export the latest encompassing .csv from Zotero and save it on your local system. You can do this by selecting **File > Export library** in the desktop application.

```
data <- read_csv(here("projects",
                      "literature-tracking",
                      "data",
                      "20 oct 23.csv"))
```

3. Clean column names - it will make it easier to process and interpret the data.

```
dataclean <- janitor ::clean_names(data)
```

4. We also want to make sure the date is the same across every entry so that it processes correctly when we filter it. `as_date` standardises all date data to `yyyy-mm-dd`. We also add a new column that extracts just the month and year for the month by month analysis below.

```
dataclean2 <- dataclean |>
  mutate(
    date_added_clean = as_date(date_added), #date_added clean (remove time stamp)
    publication_date = (date)) |> #rename date to publication_date
  drop_na(publication_date) |> #drop records w. a n/a publication date
   select(title, date, publication_date, place, publication_year, publication_title,
          item_type, url, manual_tags, date_added, date_added_clean) #select relevant colu
dataclean2
```

```
# A tibble: 4,133 x 11
   title         date  publication_date place publication_year publication_title
   <chr>         <chr> <chr>            <chr>            <dbl> <chr>
 1 Australian P~ 2023  2023             <NA>              2023 <NA>
 2 Ficus auricu~ 2022~ 2022-01-07       <NA>              2022 <NA>
 3 Countering e~ 2023~ 2023-08          <NA>              2023 Biological Conse~
 4 Mepimbat ted~ 2023~ 2023-04          <NA>              2023 Austral Ecology
 5 The contempo~ 2023~ 2023-04          <NA>              2023 Journal of Bioge~
 6 Assessing th~ 2023~ 2023-08          <NA>              2023 Applied Geography
 7 Movements, H~ 2021  2021             Aust~             2021 <NA>
 8 Identifying ~ 2019~ 2019-12          Sydn~             2019 <NA>
 9 Conservation~ 2023~ 2023-06-04       <NA>              2023 Alcheringa: An A~
10 Rediscovery ~ 2023~ 2023-04-28       <NA>              2023 Memoirs of the Q~
# i 4,123 more rows
# i 5 more variables: item_type <chr>, url <chr>, manual_tags <chr>,
#   date_added <dttm>, date_added_clean <date>
```

# 2 Step 2 - Summaries:

## 2.1 Annual journal article numbers

In this section we collate the number of **journal articles** mentioning, citing, using, acknowledging etc. (…) the ALA in their given publication year.

> **Note that**: in the past we included the GBIF DOI tag in analyses.
>
> In 2022, it was decided to exclude this tag from the library (held on file) as GBIF's direct downloads are not always related to the ALA's own research impact.

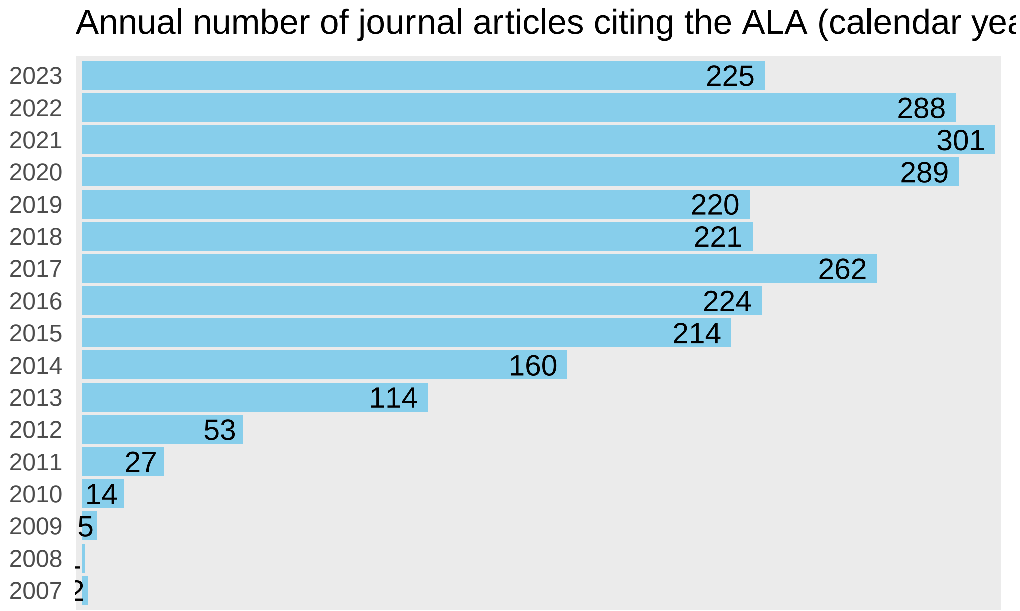
###Calender year:

```
# Annual calendar  (not financial year) journal numbers. Note the current year will show 6
apublicationbyyear <- dataclean2 |>
  filter(item_type == "journalArticle") |>
  group_by(publication_year) |>
  count()
apublicationbyyear
```

```
# A tibble: 17 x 2
# Groups:   publication_year [17]
   publication_year     n
              <dbl> <int>
 1             2007     2
 2             2008     1
 3             2009     5
```

```
4               2010    14
5               2011    27
6               2012    53
7               2013   114
8               2014   160
9               2015   214
10              2016   224
11              2017   262
12              2018   221
13              2019   220
14              2020   289
15              2021   301
16              2022   288
17              2023   225
```

```r
apublicationbyyear$publication_year <- as.character(apublicationbyyear$publication_year)

# Create a horizontal bar plot
ggplot(apublicationbyyear, aes(x = n, y = publication_year)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(aes(label = n), hjust = 1.2, vjust = 0.5) + # Add number annotations
  labs(x = "", y = "") +
  ggtitle("Annual number of journal articles citing the ALA (calendar years)") +
  theme(
    panel.grid = element_blank(),
    axis.text.x = element_blank(),
    axis.line.y = element_blank(),  # Remove y-axis line
    axis.ticks = element_blank()) + scale_x_continuous(expand = c(0, 2))
```



###Financial year:

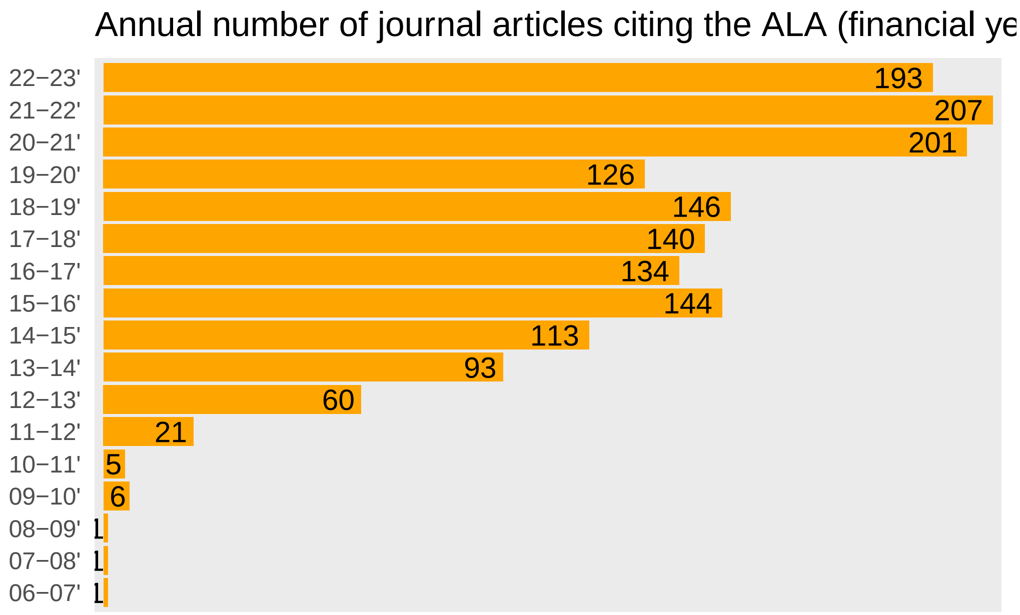
```
library(dplyr)
library(lubridate)
# Financial year journal numbers
# Note: we used publication_date here instead of date_added as a large proportion of artic
dataclean2$publication_date <- ymd(dataclean2$publication_date)
```

```
Warning: 2247 failed to parse.
```

```
# Create an empty data frame to store the results
result_df <- data.frame()

# Iterate through each financial year from 2006-2007 to 2022-2023
for (year in 2006:2022) {
  # Calculate the start and end dates for the financial year
  start_date <- if (year == 2006) {
    ymd(paste(year, "-06-30", sep = ""))
  } else {
    ymd(paste(year, "-07-01", sep = ""))
  }
  end_date <- ymd(paste(year + 1, "-06-30", sep = ""))

  # Create a date range string for the current year
  date_range <- paste(start_date, "to", end_date)

  # Modify your code to filter data for the current financial year using tidyverse
  current_year_data <- dataclean2 %>%
    filter(item_type == "journalArticle") %>%
    filter(publication_date >= start_date, publication_date <= end_date) %>%
    filter(publication_year %in% as.character(year:(year+1))) %>%
    count() %>%
    drop_na() %>%
    mutate(Financial_Year = paste(substr(as.character(year), 3, 4), '-', substr(as.charact


  # Add the data for the current year to the result data frame
  result_df <- bind_rows(result_df, current_year_data)
}

# View the result data frame
result_df
```

```
     n Financial_Year               Date_Range
1    1         06-07' 2006-06-30 to 2007-06-30
2    1         07-08' 2007-07-01 to 2008-06-30
3    1         08-09' 2008-07-01 to 2009-06-30
4    6         09-10' 2009-07-01 to 2010-06-30
5    5         10-11' 2010-07-01 to 2011-06-30
6   21         11-12' 2011-07-01 to 2012-06-30
7   60         12-13' 2012-07-01 to 2013-06-30
8   93         13-14' 2013-07-01 to 2014-06-30
9  113         14-15' 2014-07-01 to 2015-06-30
10 144         15-16' 2015-07-01 to 2016-06-30
11 134         16-17' 2016-07-01 to 2017-06-30
12 140         17-18' 2017-07-01 to 2018-06-30
13 146         18-19' 2018-07-01 to 2019-06-30
14 126         19-20' 2019-07-01 to 2020-06-30
15 201         20-21' 2020-07-01 to 2021-06-30
16 207         21-22' 2021-07-01 to 2022-06-30
17 193         22-23' 2022-07-01 to 2023-06-30
```

```r
# Load the ggplot2 library if it's not already loaded
library(ggplot2)
library(ggthemes)

# Assuming you have a data frame named 'result_df' with columns 'n' and 'Financial_Year'

# Create a horizontal bar plot
ggplot(result_df, aes(x = n, y = Financial_Year)) +
  geom_bar(stat = "identity", fill = "orange") +
  geom_text(aes(label = n), hjust = 1.2, vjust = 0.5) + # Add number annotations
  labs(x = "", y = "") +
  ggtitle("Annual number of journal articles citing the ALA (financial years)") +
  theme(
    panel.grid = element_blank(),
    axis.text.x = element_blank(),
    axis.line.y = element_blank(),  # Remove y-axis line
    axis.ticks = element_blank()) + scale_x_continuous(expand = c(0, 2))
```



## 2.2 Total to journal article ratio

*Total* to *journal only* publications within the current financial year:

NOTE: Publications that are *not* journal articles rarely list a publication `date` in a dd/mm/yyyy format, and there is no streamlined method to correct mass records if they have been missed in the past to this format without a DOI (for journal articles, I have used ZotMeta). For a financial year analysis, data needs to be in a dd/mm/yyyy format. Therefore, for publications other than `journal_articles` we have substituted `date_added` (dd/mm/yyyy format), restricted by the current `publication_year` under the justification that publications are (*most* of the time) added within a similar range of them being published. We use `publication_date` for `journal_articles`.

1. Updated code (2023-10-23)

```
#non-journal ratio
nonjournalstotalfy <- dataclean2 |>
  filter(item_type != "journalArticle") |>
  filter(date_added_clean >= ymd("2022-06-30")) |>
  filter(date_added_clean <= ymd("2023-06-30")) |>
  filter(publication_year == "2022" | publication_year == "2023") |>
  count() |>
  drop_na()
nonjournalstotalfy
```

```
# A tibble: 1 x 1
      n
  <int>
1   126
```

```
# journal ratio *
journalstotalfy <- dataclean2 |>
  filter(item_type == "journalArticle") |>
  filter(publication_date >= ymd("2022-06-30")) |>
  filter(publication_date <= ymd("2023-06-30")) |>
  filter(publication_year == "2022" | publication_year == "2023") |>
  count() |>
  drop_na()
journalstotalfy
```

```
# A tibble: 1 x 1
      n
  <int>
1   197
```

```
#total pubs *
totalpubsbyfy <- nonjournalstotalfy + journalstotalfy
totalpubsbyfy
```

```
    n
1 323
```

2. Previous code (expandable) (before publication date to (dd/mm/yyyy) data cleaning - see filtering justifications below)

2022 (prior to cleaning, for ’23 report):

```
#total
dataclean2 |>
  filter(date_added_clean >= ymd("2022-06-30")) |>
  filter(date_added_clean <= ymd("2023-06-30")) |>
  filter(publication_year == "2022" | publication_year == "2023") |>
  count() |>
  drop_na()
```

```
#journals only
dataclean2 |>
  filter(item_type == "journalArticle") |>
  filter(date_added_clean >= ymd("2022-06-30")) |>
  filter(date_added_clean <= ymd("2023-06-30")) |>
  filter(publication_year == "2022" | publication_year == "2023") |>
  count() |>
  drop_na()
```

2021:

```
dataclean2 |>
  select(publication_year, title, publication_title, item_type) |>
  filter(publication_year == 2021) |>
  filter(item_type == "journalArticle") |> #remove this line with a # to see total publica
  group_by(publication_title) |>
  count()
```

## 2.3 Top journal articles

This section analyses the ten most common journal articles mentioning, citing, using, acknowledging etc. the ALA **across all time**. It is worth checking that these journals are not preprint repositories (e.g. BioXriv), as, occasionally Zotero auto categorises them as journal articles instead.

These numbers might change subtly over time due to ongoing data cleaning, but will likely remain similar. **They are included in the report**.

```
grandtop10 <- dataclean2 |>
  filter(item_type == "journalArticle") |>
  group_by(publication_title) |>
  count() |>
   drop_na() |>
  arrange(desc(n))
grandtop10
```

```
# A tibble: 819 x 2
# Groups:   publication_title [819]
   publication_title               n
   <chr>                       <int>
 1 Zootaxa                        62
 2 Austral Ecology                56
 3 Ecology and Evolution          51
 4 Austral Entomology             46
 5 PLOS ONE                       46
 6 Biological Conservation        36
 7 Journal of Biogeography        36
 8 Australian Journal of Botany   33
 9 Diversity and Distributions    33
10 PLoS ONE                       32
# i 809 more rows
```

```
grandtop10$publication_title[1:10]
```

```
 [1] "Zootaxa"                       "Austral Ecology"
 [3] "Ecology and Evolution"        "Austral Entomology"
 [5] "PLOS ONE"                     "Biological Conservation"
 [7] "Journal of Biogeography"      "Australian Journal of Botany"
 [9] "Diversity and Distributions"  "PLoS ONE"
```

To find out how many different journal articles the ALA were cited in **the 2022-23 financial year**, run the same code but filter across the financial year dates via date_added (30 June - 30 June) and further restrict to papers published in that year (sometimes old papers are added as they become released online, which aren't relevant to our analysis).

Be sure to remove the NA category (journal articles without an available name in the tracker) using `drop_na()` as well. This might be a chance to work on investigating these entries and see if there is missing data.

```
top10currentyear <- dataclean2 |>
  filter(item_type == "journalArticle") |>
  filter(publication_date >= ymd("2022-06-30")) |>
  filter(publication_date <= ymd("2023-06-30")) |>
  group_by(publication_title) |>
  count() |>
  drop_na() |>
  arrange(desc(n))
top10currentyear
```

```
# A tibble: 134 x 2
# Groups:   publication_title [134]
   publication_title                                   n
   <chr>                                           <int>
 1 Austral Entomology                                  7
 2 PLOS ONE                                            7
 3 Austral Ecology                                     6
 4 Australian Zoologist                                4
 5 Diversity and Distributions                         4
 6 Plants                                              4
 7 Scientific Reports                                  4
 8 Alcheringa: An Australasian Journal of Palaeontology     3
 9 Biodiversity Information Science and Standards      3
10 Diversity                                           3
# i 124 more rows
```

Here, there are 134 rows: meaning 134 unique journal articles that have cited the ALA across this financial year. **This number is included in the report. (p. 189 w date_added)**

The top journals for this financial year listed were:

```
top10currentyear$publication_title[1:10]
```

```
 [1] "Austral Entomology"
 [2] "PLOS ONE"
 [3] "Austral Ecology"
 [4] "Australian Zoologist"
 [5] "Diversity and Distributions"
 [6] "Plants"
 [7] "Scientific Reports"
 [8] "Alcheringa: An Australasian Journal of Palaeontology"
 [9] "Biodiversity Information Science and Standards"
[10] "Diversity"
```

## 2.4 Important filtering justifications

**We chose to filter by `date_added` and *then* restrict to `publication_year` instead of filtering by `date` (date of publication). Note that there are degrees of error in both methods for a few reasons:**

- If filtered by `publication_year` and restricted to the financial year dates `dd/mm/yyyy`, there is a risk of excluding papers that only have a *year* `yyyy` of publication listed (which is common). It is preferred to get the publication `date` of papers entered into the tracker closest to `dd/mm/yyyy` as possible.
- Similarly, a paper may be published online but the listed publication `date` can be some months in advance as it reflects when the physical copy is released. While a paper may have been added and accessible online, it may be excluded from the financial year report if the physical release is outside of this date. As we have done before data cleaning (see below), using `date_added` and restricting to the relevant `publication_year`s, on the other hand, may exclude papers that were added outside of the financial year period but have a publication `date` within it. This seems to be the less likely occurrence and the more logical choice between the two.
- On the 19/10/2023, data were cleaned so that `publication_date` for all records are listed as close to dd/mm/yyyy instead of predominately yyyy. This means a holistic analysis by year (financial year) analysis of literature tracking is now possible. 2006-2016 data were cleaned manually, and 2017-2023 were cleaned using the ZotMeta add-on. Online

publication dates were preferred and used over original/physical publication dates, but in cases where the online publication date only listed year, the original dd/mm/yyyy publication date was used. If neither were available - note that the date was listed as January of the available year (e.g 01/yyyy).

### 2.4.1 2021 YiR error

1. GBIF papers

Note that for the 2021-2022 Annual Report/ Year in review (and those before) our numbers were retrieved with data that *still contained GBIF DOI tag papers* (on file as `30thjune21.csv`). Therefore, the number of publications in the '21-22 report are substantially larger across the entire dataset.

GBIF DOI tagged papers are publications that have retrieved Australian biodiversity data from GBIF (our larger global node) but not the ALA. Technically this data is held in the ALA but as authors have retrieved it from the GBIF interface we decided that this is not a good metric of the ALA being 'used'.

You can find the GBIF papers removed from Zotero on file here.

2. Financial vs. calendar year

In addition to this, we *also* did not restrict the year to papers within the '21-'22 **financial year** specifically for the 21-22' report. We only presented **total to journal article ratios** in the full data **calendar year** of the earlier year (**2021**). The code below is what was used (using the `30thjune21.csv`). It is unclear whether this has always been the method of reporting during YiR as no documentation has been provided for these metrics. From '23 onwards, we have decided to change this to the *financial year* due to it being a more recent metric and allows the more recent year to be equally compared to the others.

```
# 2021 code
dataclean2 |>
  select(publication_year, title, publication_title, item_type) |>
  filter(publication_year == 2021) |>
  filter(item_type == "journalArticle") |> #remove this line to see total publications aga
  group_by(publication_title) |>
  count()
```

### 2.4.2 Conclusion

Thinking about best practice for reporting in the data filtering points above, as well as financial year metrics being a more recent view of publications than the previous full available calendar year (as above, only '21 was provided for the '21-22 report as '21 at that point in time had a full year of publication data), we have made some changes to how tracking data is filtered and reported. Additionally, having experienced issues replicating numbers from previous years we also decided to document all of our code for producing literature tracking metrics so that it can be easily reproduced and our justifications are clearly recorded for future decisions.

## 2.5 Top research domains

We also include a research domain section that quantifies the the top 10 different fields of research the ALA has been utilized across. This involves text mining, and is a little more complex than the last analyses.

To be added - Martin on file (?)